Construction and Building Materials 325 (2022) 126584

Contents lists available at ScienceDirect

Construction and Building Materials

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

DefectTR: End-to-end defect detection for sewage networks using a transformer

L. Minh Dang [c], Hanxiang Wang [a], Yanfen Li [a], Tan N. Nguyen [b], Hyeonjoon Moon [a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea
[b] Department of Architectural Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea
[c] Department of Information Technology, FPT University, Ho Chi Minh city 70000, Viet Nam

ARTICLE INFO

Keywords:
Deep learning
Transformers
Sewer pipe
Crack detection
Defect analysis

ABSTRACT

The sanitary sewer is a crucial underground infrastructure of any country that collects wastewater and carries it to the treatment plant. The damage triggered by various factors, such as external interference, long-term corrosion, and uneven distribution of pressure, could lead to various types of defects inside the sewer pipe. Previous studies primarily relied on human visual perception to evaluate the sewage system, which was tedious, time-consuming, and costly. As a result, an efficient and robust sewer defect localization framework was proposed in this manuscript. The main contributions include (1) a novel sewer defect detection system motivated by the state-of-the-art detection transformer (DETR) architecture, which views object localization as a set prediction topic; (2) a defect severity analysis approach based on the transformer's self-attention operation to analyze defect zone of influence and defect grade; and (3) a manually validated sewer defect localization dataset that contains 10 types of commonly appeared sewer defects. The experimental results suggested that the proposed system outperformed the previous standard object detection approaches with the highest mean Average Precision (mAP) of 60.2% on the collected dataset.

1. Introduction

The sanitary sewer system is a fundamental public infrastructure used to stimulate economic development. Although each sewer system has an anticipated life span if maintained appropriately, the damage still follows an unexpected trend line, leading to unpredictable incidents [1]. In addition, it is challenging to manage the sewer system when there is an increasing trend for a clean environment and tight budgets. Therefore, periodical maintenance of the sewer pipelines has long been one of the primary municipality issues. The structure of a sewer pipe could deteriorate quickly because of the harsh environments, which lead to aging, damages, and corrosion that significantly affect its functional operation [2]. Not only can this cause severe consequences, but it also requires high repair costs and a huge labor force [3]. Some defect examples are presented in Fig. 1(b).

Previously, a periodical investigation can fall into either in-field coder or office-based coder. In the earlier case, defects are coded during the inspection, whereas in the latter case, they are coded after CCTV videos were completely recorded. This study follows the office-based coder approach, which has four main processes: (1) closed-circuit television (CCTV)-based sewer image/video acquisition, (2) defect detection, (3) in-depth analysis, and (4) rehabilitation [4]. Among those steps, defect detection and in-depth analysis are the primary objectives of this work because they are considered subjective, time-consuming, and costly. In addition, the in-depth defect analysis is still particularly challenging because defect features from any category but with different severity degrees are comparatively similar. As a result, until now, defect inspection and in-depth analysis were performed mainly by trained operators in most sewer inspection companies [5]. Another drawback of CCTV sewer inspections is inconsistency in defect reporting. While these inconsistencies can be alleviated through training and the use of standardized reporting formats, such as the Pipeline Assessment Certification Program (PACP) [6], the operator's skills, experience, and biases can greatly influence the final report. All things considered, it is important to introduce an efficient sewer defect detection framework that offers various solutions to solve the aforementioned problems.

Computer vision (CV) and artificial intelligence (AI) have witnessed immense development, which has been extensively applied in a variety of disciplines, such as agriculture [7], structural inspection [8,9], and autonomous driving [10]. Even though the traditional vision-based approaches, which were usually applied to the small datasets, achieved good performance, they performed poorly during the testing time when

* Corresponding author.
E-mail address: hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.conbuildmat.2022.126584
Received 25 November 2021; Received in revised form 18 January 2022; Accepted 21 January 2022
Available online 10 February 2022

(a) Sewer defect inspection equipment

(b) Some sample sewer defects

Fig. 1. Visualization of the defect inspection process. (a) demonstrates the two main tools that are utilized to collect the inspection videos and (b) shows some defects that appear in a sewer pipe. **Note:** Defect types that are displayed are as follows. Red: protruding lateral, yellow: displaced joint, orange: surface damage, and gray: debris silty. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the dataset was slightly different from the training dataset since the traditional vision-based models selected and extracted the hand-drafted features manually [11,12].

In recent decades, convolutional neural network (CNN) models have demonstrated remarkable performances in most common CV problems, including classification [8], object detection [13,14], and segmentation [15,16], which outperformed the previous machine learning (ML) methods by a significant margin. Moreover, the current deep learning-based object detection approaches usually extract abstract and coarse features, which are semantically robust.

Most recently, a new end-to-end detection transformer, which was named DETR [17], was proposed. It demonstrated better than the latest one-stage and two-stage object detectors, including You Only Look Once (YOLO) [18], region-based convolutional neural networks (RCNN) [19] on common objects detection datasets. DETR streamlined the previous detection pipeline by removing several time-consuming parts, such as anchor creation or non-maximum suppression algorithm that specifically encode the human's former knowledge about the task. The DETR model considers the object detection topic as a direct bounding box prediction task using the encoder–decoder architecture [20]. DETR learns a bipartite match between queries and ground truths (GTs) by introducing a custom loss function, which involves the Bipartite matching loss and Hungarian loss. DETR obtained comparable results to other standard detection models on the Common objects in context (COCO) dataset [21], and notably better performance on the large objects. Compared to the common object detection task, a comprehensive analysis of localized objects is challenging. For instance, it is challenging to analyze the severity of defects that belong to the same class because they are pretty similar.

In this study, we demonstrate a novel way to effectively recognize the defect severity of the localized defects based on customizing DETR. The self-attention weights of the last encoder and decoder layers from the DETR model are extracted, which are then used to construct the mean feature map. Finally, the defect damage degree and zone of influence are calculated using the generated mean feature map. In summary, the main contributions of the study are described below.

- A novel sewer defect localization and defect severity analysis framework, which achieve high object localization accuracy.
- Various module of the original transformer-based DETR structure is changed to facilitate the sewer defect detection topic.
- The experimental results revealed that the proposed model was more stabilized and achieved higher performance than the original model. In addition, it was applied to practical applications for automatic sewer pipe inspection in Korea.
- Defect severity analysis using attention features extracted from the transformer-based model. The suggested framework is the first to apply the transformer architecture to detect and analyze the severity of the sewer defect.

The rest of the manuscript is outlined as follows. Section 2 commits to performing a literature review. Section 3 explains details the sewer pipe defect detection and defect severity analysis. The manually collected and evaluated sewer defect detection dataset is mentioned in Section 4. Several experiments are carried out in Section 5 to thoroughly evaluate the suggested framework. Finally, we summarize the study and discuss the future work in Section 6.

2. Literature review

2.1. Pre-processing

Raw CCTV videos collected by the robot usually have two primary problems, including uneven brightness and fog, so it is crucial to handle them appropriately to reduce the negative impacts and enhance the detection rate [4].

Contrast enhancement is an image processing method that aims to intensify a raw image's contrast in order to cope with the uneven brightness issue. Among the contrast enhancement approaches, histogram equalization is a standard approach that has been applied for a long time to improve image contrast because it is straightforward and can achieve maximum efficiency. However, most previous contrast enhancement approaches showed relatively low structural similarity index measure (SSIM) [22]. As a result, Abdullah et al. suggested a novel image contrast enhancement algorithm named dynamic histogram equalization (DHE) [22], which efficiently improved the image quality and reduced noise without missing any important information.

Various defogging models have been represented to cope with the foggy environment inside the sewer pipes. For instance, a well-known and efficient dehazing approach called dark channel prior (DCP) was usually implemented to perform defogging [23]. However, DCP performed poorly on the low contrast or low-resolution images [24]. Therefore, many deep learning-based defogging approaches have been studied recently. Shao et al. proposed a different dehazing strategy than the existing methods, consisting of an image translation phase and two image dehazing phases [25]. The proposed method offered a good generalization by incorporating the hazy image into the dehazing training. The results collected from various experiments demonstrated that domain adaptation defogging outperformed the current state-of-the-art approaches.

2.2. Sewer defect detection

2.2.1. Traditional approaches

Before the popularity of deep learning, traditional CV and image processing approaches were mainly adopted to deal with the vision tasks, because they delivered good performance on the task under consideration [26,27].

For example, Halfawy et al. suggested a framework for automated tree root intrusion inspection for sewer inspection videos by training the support vector machine (SVM) using the histograms of oriented gradients (HOG) [26]. HOG was first implemented to obtain the potential defect region of interest (ROI) from a list of training sewer images. SVM was then trained on the extracted HOG features to classify whether a test sample is positive (have defect) or negative (no defect). Ye et al. used the features extracted from Daubechies wavelet transform, lateral Fourier transform, texture features, and Hu invariant moment, to train the SVM model to categorize 7 categories of sewer cracks [3]. The experimental results revealed that the overall classification accuracy was 84.1%. Moradi et al. introduced a novel real-time sewer defect detection system from the inspection CCTV videos that extracted the spatio-temporal features to train the hidden Markov model (HMM) [11].

Most of the mentioned research from the traditional approach manually extracted essential features from the training dataset and then fed them into ML models to perform defect analysis. Although each research addressed part of the problem, sewer defects were usually missed because the traditional systems were trained on relatively small datasets, relied entirely on the hand-crafted features, and were affected by the relatively complicated environment inside the sewer pipes. In addition, previously developed features like scale-invariant feature transform (SIFT) and HOG were incapable of comprehensively representing such defects [26]. Consequently, it is a significant challenge for traditional approaches to achieving satisfactory performance and solid robustness.

2.2.2. Deep learning-based approaches

Over the last few years, deep learning has become increasingly known to the research community for its remarkable performance [15,28,29]. In the area of some practical engineering problems that require precise precision, such as civil infrastructure inspection [4], CNN architectures have been proved to offer remarkable performances and excellent robustness as well [30]. As a result, they have been gradually implemented for sewer defect detection [8]. Compared with the traditional approaches, the feature engineering process is performed automatically by the deep learning models during the learning process.

Following the development trend of the object detection topic, one-stage detectors, such as YOLO [18], and two-stage detectors, such as RCNN, are usually utilized to perform sewer defect detection [31]. For instance, Wang et al. introduced a robust sewer defect detection and tracking framework using metric learning [32]. The tracking module was carried out by performing a Kalman filter using two input sources, which include (1) detected objects from the faster-RCNN framework and (2) the extracted appearance features from the metric learning. The experimental results confirmed that the framework could track the sewer cracks in the inspection videos with a robust IDF1 score of about 57%.

Similarly, Cheng and Wang et al. introduced an automated sewer crack detection method based on faster R-CNN, which was trained on 3000 defect images with 4 defect classes [5]. The authors showed that adjusting the hyperparameters, including kernel dimensions and stride values, enhanced the detection rate, with the final mean average precision (mAP) of 83%. Recently, Yin et al. suggested a novel sewer defect detection framework that accepted inspection videos as the input and showed defect frames as the output [28]. The YOLOv3 model was trained and evaluated on a dataset that contained 4056 images for 6 defect types, which include broken, fracture, deposits, root, hole, and crack. The proposed framework showed a high F1 score of 0.882 and mAP of 85.37% on the testing set.

The popular object detection approaches typically involve a hand-crafted post-process, such as non-maximum suppression (NMS) [33], to eliminate the detected bounding boxes with low detection probability. Although many variants of the NMS have been proposed in order to partly address the remaining problems of the NMS, such as soft-NMS [34], harmony search-based NMS [35], NMS must be employed independently and cannot be employed in an end-to-end way because its procedure involves no image and network features.

The drawbacks of NMS and the recent introduction of the transformers, which demonstrated state-of-the-art performances on common natural language processing (NLP) topics [36], motivated the introduction of a novel transformer-based object detection, named DETR [17]. DETR utilized the encoder–decoder structure of the transformer model and could construct context features and discard duplicates essentially. DETR achieved relatively high performance in the COCO object detection benchmark dataset due to the introduction of a set-based Hungarian loss function that requires unique predictions for every ground-truth bounding box via bipartite matching without the need for NMS. As a result, this paper aims to optimize the DETR performance for sewer defect detection.

3. Sewer defect detection dataset

The CCTV videos are recorded at different concrete sewer utility holes across South Korea. The robots are equipped with a high-resolution 1/3-in. SONY Exmor CMOS camera can rotate 360° and support up/down tilt. The robot's head is equipped with a powerful light-emitting diode (LED) bulb to capture the images/videos under the dark environment inside the sewer pipes. The recorded CCTV videos are at 30 frames per second (FPS), and each inspection duration ranges from 3 to 20 min.

Most of the previous studies, even the latest ones, worked on less than 6 types of sewer defects, such as (fractures, root intrusions, and lateral connection) [32], (crack, fracture, collapse, broken, and hole) [16], (crack, deposit, root, crack & deposit, crack & root, and deposit & root) [15]. As illustrated in Fig. 2, this study investigates 10 defect types, which outnumbers most of the previous research in terms of the number of defect types.

A detailed definition of each defect class is defined as follows.

- Broken pipe (BK): a severe type of defect, which indicates that the pipe's internal structure was partly or fully collapsed. As a result, timely maintenance by the experts is required.
- Longitudinal crack (LC): a diagonal or vertical crack appears on the sewer wall that results from settling in the concrete foundation as concrete shrinks during curing.
- Circumferential crack (CC): any damage that appears parallel to the channel axis caused by pressure from the outside of the walls. It is considered more severe than the longitudinal crack, because it can lead to a permanent failure in the pipe foundation.

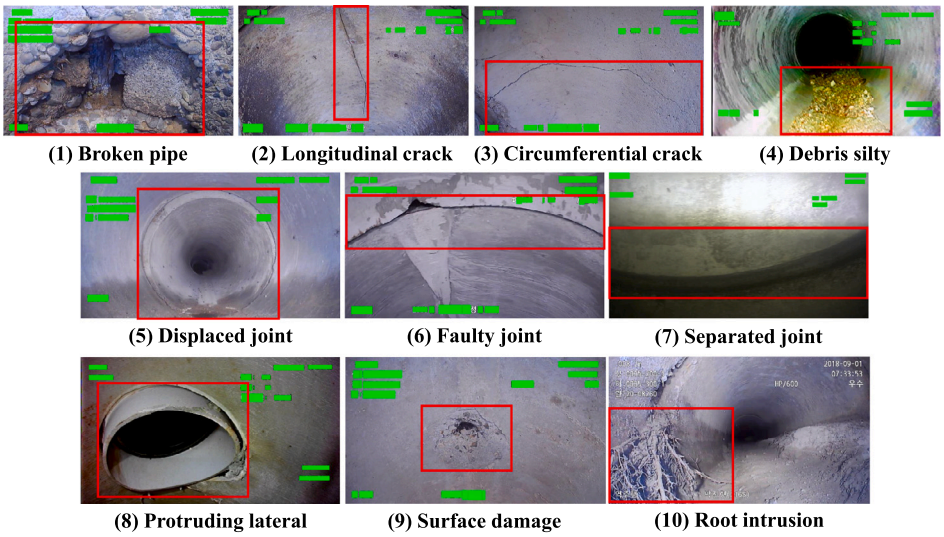


Fig. 2. Sample images for the ten defect types, which are included in the proposed sewer defect detection dataset. **Note:** Defects are highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- Debris silty (DS): gravel and seals left in the sewer pipe. Moreover, the sediments of grease or other obstacles also belong to this type of defect.
- Displaced joint (DJ): minor displacement in the pipe joints.
- Faulty joint (FJ): physical deterioration in the pipe joints.
- Separated joint (SJ): major displacement in the pipe joints.
- Protruding lateral (PL): a connecting pipe part protrudes from the primary sewer pipe.
- Surface damage (SD): minor damage on the sewer's surface, defective pipe, brittleness, erosion by abrasion, or chemical corrosion.
- Root intrusion (RI): roots intrude into a sewer pipe network and cause a sewer line back-up.

A group of 11 experts from a deep inspection company[1] was involved in a two-month labeling process, and each person labeled about 70 images per day on average. The annotation tool used in this process was a standard open-source graphical image annotation tool named labelImg,[2] which was developed using Python and Qt5, allowing the experts to precisely label different types of defects. Fig. 3 describes the number of images, which were labeled for each class. From the collected defect dataset, containing a total of 47,100 images, 80% of the data (37,680 images) was selected randomly as the training dataset, while the other 20% of the data (9420) was used as the testing dataset. Finally, 10% of the training data (3768) was selected as a validation dataset.

4. System overview

Fig. 4 explains the main processes of the automatic sewer pipe defect detection system, called DefectTR.

- Pre-processing phase. The extracted frames from the sewer inspection videos host a set of problems, including uneven brightness and a foggy environment. It is essential to balance the image brightness and remove possible noise to improve the quality of the input data for crack detection and defect severity analysis. As a result, data pre-processing is crucial to ensure the defect detection framework's performance even if the captured video's quality is poor.
- Transformer-based defect detection. The previous object detection models, such as YOLO, single-shot detector (SSD), and RCNN, are complicated, do not generalize well with limited parameters, and do not have a straightforward procedure for the training and testing processes as existing classification models.
  This paper implemented the state-of-the-art detection transformer object (DETR) [17], to perform sewer defect detection. The model does not demand prior knowledge about anchors or handcrafted methods like NMS. In addition, we replace the existing CNN backbone with the backbone trained by [37] to allow the DETR to extract defect-related deep, coarse features effectively. The performance of the mentioned DefectTR is proven by a series of experiments that will be explained in Section 6.
- Defect severity analysis. Sewer defect detection is a common object detection task that localizes defects that appear inside an input image. On the other hand, defect severity analysis, which involves recognizing detailed information related to the detected defects, such as zone of influence and defect grade, is a practical and challenging problem because there exists no reliable source evidence to determine if a crack is minor or severe using the detected bounding boxes.

5. Methodology

5.1. Image pre-processing

First of all, DHE [22] is carried out to improve the image contrast of the collected raw images in order to reduce the uneven brightness issue, owing to its high performance and computational efficiency. DHE has three main processes, which include histogram partitioning based on local minima, sub-histogram gray level ranges assignment, and histogram equalization. Compared to the previous contrast enhancement approaches, DHE enhanced the image without losing image details.

Besides, high humidity and strong water vapor inside the sewer pipes can easily cause the lens of the CCTV cameras to be blurred, which is unavoidable during the data collection process. Therefore, a pretrained encoder–decoder-based denoising model, which was called gated context aggregation network (GCANet) [38], was adopted to appropriately denoise the CCTV videos before feeding them into the proposed framework. All training parameters were set as recommended in the original paper [37]. It predicts the residual feature maps of the hazy and the target images in an end-to-end approach. The encoder path has three convolution blocks, whereas the decoder path contains

[1] http://www.deepinspection.ai/.
[2] https://github.com/tzutalin/labelImg.

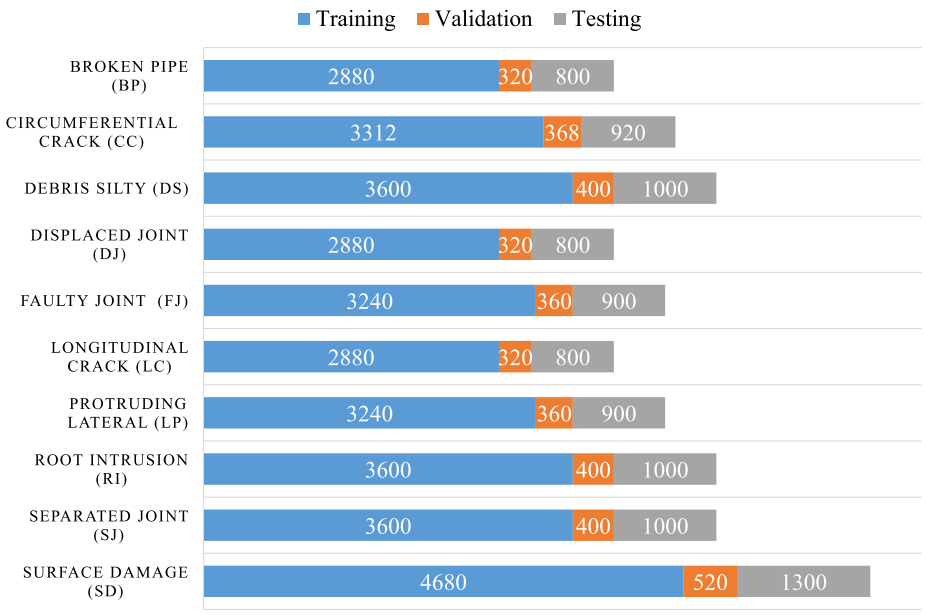


Fig. 3. Descriptions for the number of training/testing images for each class of the defect in the proposed dataset.

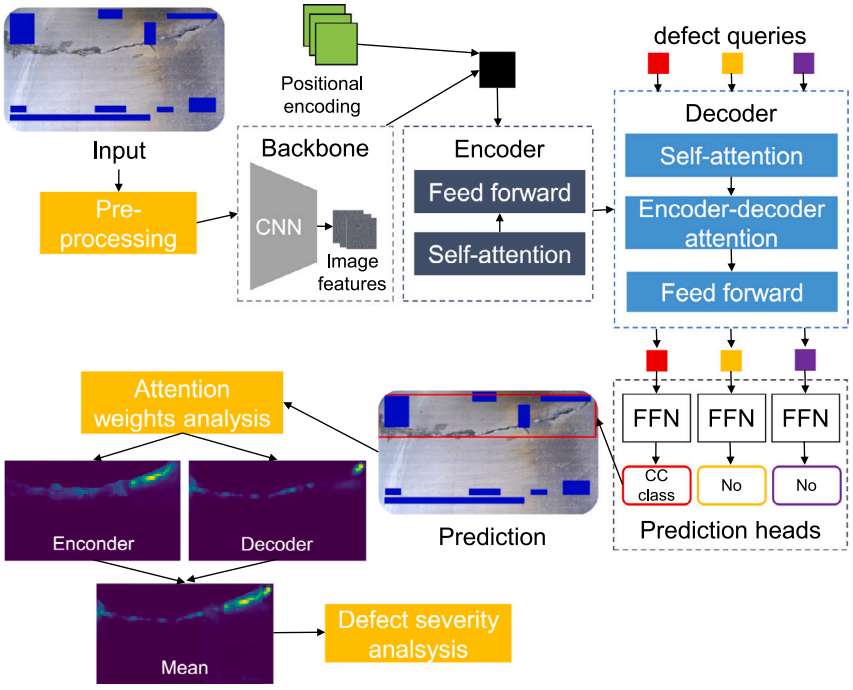


Fig. 4. Detailed description of the proposed sewer defect detection system (DefectTR). **Note:** There are 4 main steps: (1) pre-processing, (2) DefectTR model, (3) attention weights analysis, and (4) defect severity analysis.

two convolution blocks and one deconvolution block. Moreover, multiple smoothed dilated Resblocks were added between the encoder and the decoder to extract more context features without gridding artifacts. Finally, an additional gate fusion subnetwork is implemented to combine the extracted features of distinct levels. GCANet was proved to perform denoising well without prior knowledge and maintained the image's initial brightness [38].

Fig. 5 shows the pre-processing results of some random images selected from the proposed dataset. By feeding the input images into the DHE and GCANet models, the outputs show a huge improvement in the image quality compared to their original version. For instance, it is a challenge to observe the defect in the original low-light example. However, the pre-processed image delivers a significant improvement in the image brightness, which enables the observation of what types of defects appear in the image. On the other hand, the pre-processing module shows that it does not affect images with sufficient brightness and have no noise.

5.2. Transformer-based defect detection

5.2.1. Attention in transformers

For the transformer, although the encoder and decoder have many submodules, the most crucial component is the multi-head attention that contains multiple self-attention heads, which offers the transformer the ability to learn various relationships and variations for

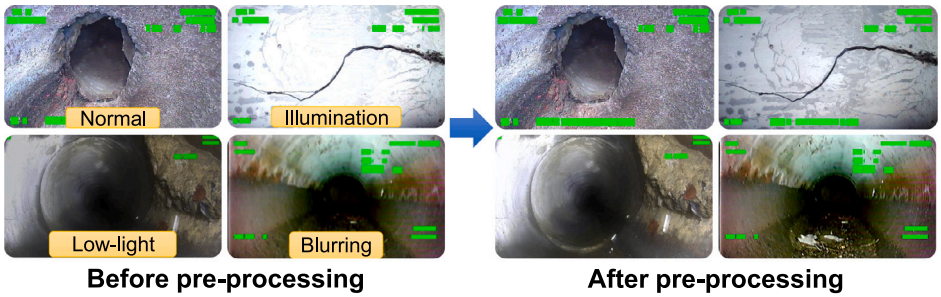


Fig. 5. Visualization of the images before and after applying the pre-processing module, which involves contrast enhancement (DHE model) and dehazing (GCANet model).

each word [36]. The encoder contains the self-attention module, which forces the input sequence to pay attention to itself, whereas the self-attention of the decoder helps the target sequence pay attention to itself. Moreover, the decoder contains an extra encoder–decoder-attention module to enable the target sequence to pay attention to the input sequence.

- Self-attention
  Self-attention's main goal is to assure that any element in a sequence can relate to others while being efficiently computed. If an input sequence has a length of T, the attention of a series of queries, keys, and values can be computed using the scaled dot-product attention as the similarity metric as follows.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

  where Q indicates a matrix consisting of a set of query vectors as columns. K and V are the corresponding matrices of key and value vectors, respectively. With d_k is the hidden dimensionality for keys. Downscaling by $\sqrt{d_k}$ discourages $\text{softmax}(QK^T)$ from taking large values, which may lead to the computation of respective small gradients, and eventually causes the optimization to stop.
- Multi-head (MA) attention
  The transformer structure contains multiple attention heads, each of which refers to the attention module that is repeatedly calculated in parallel. The Q, K, and V are split independently N-ways and fed independently to each separate head. Finally, these attentions are combined to obtain the final attention score, called MA attention. The MA attention module is a crucial part of the transformer structure because it was proved to improve the multiple relationships encoding performance of the transformer. The MA attention can be calculated as follows.

$$\begin{aligned} &\text{MA}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)\,W^O \\ &\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2}$$

  Each head is computed by multiplying Q, K and V with corresponding parameter matrices W_i^Q, W_i^K, and W_i^V.

$$\begin{aligned} &W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_q} \\ &W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k} \\ &W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v} \\ &W_i^O \in \mathbb{R}^{\text{numheads}\cdot d_h \times d_{\text{model}}} \end{aligned} \tag{3}$$

  where d_q, d_k and d_v, is the amount of vectors for Q, K, and V matrices, respectively. Lastly, the computed attention heads are multiplied with W^O, which can be computed as follows.

$$W_i^O \in \mathbb{R}^{\text{numheads}\cdot d_{\text{model}} \times d_h} \tag{4}$$

  where $(d_h = d_v/h)$ indicates the dimension applied to compute the attention heads.

The transformer model has two common issues. Firstly, it requires a significantly long training process before convergence in order to allow the attention weights to concentrate on a particular key. Pixels are usually considered the key elements under concern for CV applications. However, the number of keys N_k can become massive if a high-resolution image is fed into the transformer, leading to slow convergence. Secondly, multi-head attention's computational and memory complexity can be very high on numerous query/key pairs.

5.2.2. DETR

Given the feature vectors $x \in \mathbb{R}^{C\times H\times W}$ obtained using a CNN backbone (e.g., ResNet [17]) with C, H, W represents output channel, width, and height. DETR utilizes the encoder–decoder architecture of the transformer to convert the feature vectors into features of a collection of object queries. The detection head, which uses a typical feed-forward neural network (FFN) that has 3 fully connected layers and a linear projection, is then placed on top of the obtained object query features. The FFN plays the role of a regression model to get the bounding box coordinates $bb \in [0,1]^4$, where $\boldsymbol{bb} = \{bb_x, bb_y, bb_w, bb_h\}$ represents the normalized center coordinates, height, and width of the box. On the other hand, the linear projection is used to perform classification and produce the outputs.

Fig. 6 explains the structure of the proposed DefectTR framework. For the encoder part of DETR, both query and key matrices are set to pixels of the activation maps and encoded positional embedding, which were extracted from the backbone (ResNet). Let H and W indicate the activation map's height and width, the complexity of computing the self-attention is of $O\left(H^2W^2C\right)$, which can rise quadratically when the spatial size increases.

Next, the DETR decoder receives the encoder output and a small restricted number of learned positional embeddings, which are referred to as object queries. The decoder has two different attention modules, which are called cross-attention and self-attention modules. Cross-attention refers to the attention that is performed on queries generated by one embedding sequence and the key–value pairs from another embeddings. On the other hand, self-attention indicates attention that has values and queries generated from the same embeddings.

DETR has a promising structure that can be implemented to perform object detection, because it reduces numerous components that require manual settings. However, some weaknesses exist that can lead to the application of transformer attention in addressing image feature maps as key elements. The issues are listed as follows.

- DETR showed relatively poor results in localizing tiny objects. Although high-resolution feature maps can be extracted to enable a better detection rate of small objects, it would make the computation of the self-attention module grow quadratically.
- DETR demands a significantly longer training scheme in order to converge compared to modern object detectors. The main reason is that the attention computation based on image features as the input is more challenging to train. For instance, the cross-attention modules are initially of average attention on the entire

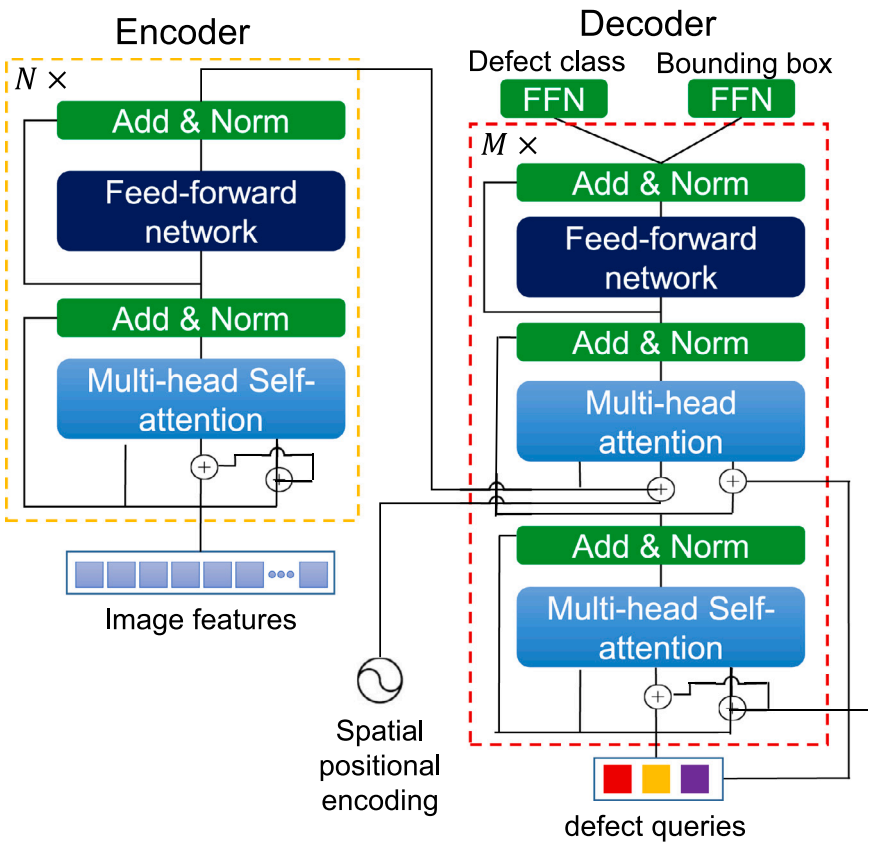


Fig. 6. Full architecture of the defectTR model, which was motivated by the DETR model [17].

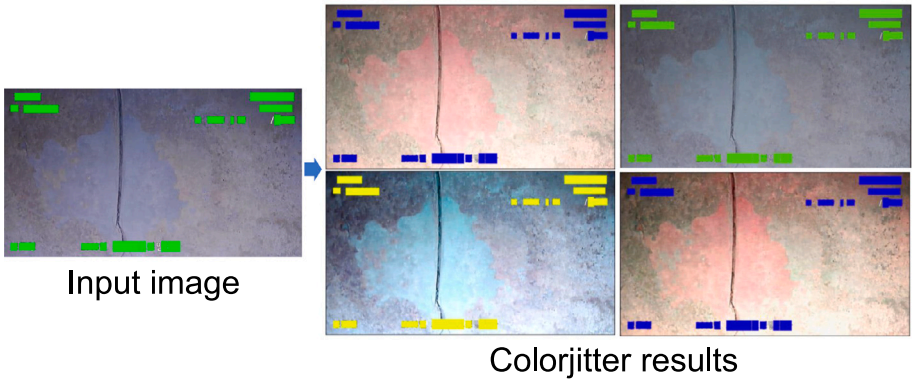


Fig. 7. Four colorjitter samples for an input image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

feature maps. However, if the number of epochs is insufficient, the learned attention maps become sparse and focus solely on the object extremities.

As a result, Table 1 describes several changes are proposed in order to address the drawbacks of DETR. The changes are as follows:

- ColorJitter: Color jitter is an augmentation method that randomly changes the brightness, hue, and saturation of an image. The value range for these parameters cannot be too large, because it can potentially introduce noise to the data. As a result, we set the brightness, contrast, saturation, and hue at the range of [0.6, 1.4], [0.7, 1.3], [0.6, 1.4], and [−0.1, 0.1], respectively. Fig. 7 illustrates some outputs of the colorjitter method.
- LeakyReLU activation function: The rectified linear unit (ReLU) activation usually suffers from the *dying ReLU* issue, which happens when the neuron is stuck on the negative side and always returns 0. Such neurons become unusable because they are not playing any role in learning the features. As a result, LeakyRelu [39] was proposed to cope with this issue by eliminating the zero-slope parts of Relu, which also speeds up the training process.
- LaProp optimizer: Adam optimizer has undesirable coupling between momentum and adaptivity, leading to instability and divergence when there is a mismatch between the momentum and adaptivity parameters [40]. Therefore, LaProp optimizer was introduced to separate momentum and adaptivity [40]. The experimental results showed that LaProp achieved faster speed and better stability than Adam on various benchmark datasets.

Table 1
Changes that are applied to increase the detection rate of the original DETR model.

Model	Augmentation	Activation	Optimizer	Loss	Params ($\times 10^6$)
Original	\\	ReLU	AdamW	GIoU	$\gtrsim$40
DefectTR	Color jitter	LeakyReLU	LaProp [40]	CIoU	4.3

Table 2
Defect grade for each type of defect based on the PACP manual.

Korean manual (Our)	PAPC	Grade
Broken pipe	Broken pipe	5
Longitudinal crack	Crack (Longitudinal)	2
Circumferential crack	Crack (Circumferential)	1
Debris silty	Obstacles/Obstructions	3
Displaced joint	Joint (Displacement)	2
Faulty joint	Joint (Faulty)	2
Separated joint	Joint (Separated)	1
Protruding lateral	Lateral (Protruding)	3
Surface damage	Surface damage	5
Root intrusion	Roots (R)	2

- Complete intersection over union (CIoU) loss: generalized intersection over union (GIoU) loss [41], which is utilized by the DETR model, expands the predicted box to fit the ground truth box. This paper proposes to use CIoU over GIoU because, unlike the GIoU loss, it matches the predicted box precisely on top of the ground truth box. CIoU is calculated using three geometric measures, which include overlapped area, central point distance, and the aspect ratio. CIoU contains additional training signals (aspect and center distance), plus the original GIoU for context. CIoU has been proved to converge faster than GIoU loss [42].

5.3. Defect severity analysis

5.3.1. Attention weights analysis

During the testing process, for each image, the attention weights (averaged over all heads) from the last encoder and decoder layers were extracted and visualized to show which part of the image the model was looking at to predict this specific bounding box and class. The attention weights from the DETR model is a squared matrix of size $[H * W, H * W]$, so it was reshaped to $[H, W, H, W]$ for a more interpretable feature map representation. After that, the mean activation map is calculated based on the encoder activation map and the decoder activation map. Finally, the binarization process is implemented on the mean activation map to get the final binary activation map, which can be used in the following subsection.

5.3.2. Defect severity analysis

The predicted sewer defect's severity can be analyzed by evaluating the zone of influence (ZOI) and the defect grade (from PACP).

Firstly, the mean feature maps can be used for computing the zone of influence (ZOI). The ZOI can be assessed by computing the total number of defect pixels in the mean activation map within the localized bounding box.

After that, a defect grade can be obtained using the PACP scoring system. For PACP, there are 5 degrees of the defect grade ranging from 1–5, which represent excellent, good, fair, poor, and immediate attention, respectively. Our defect grading is based on the standard grading system introduced by PACP [6] for each type of defect, as shown in Table 2. Although there are some differences in the naming of the defect type between our work (Korean manual) and the PACP manual, the defect types are mostly exchangeable.

Finally, a defect severity can be determined using the computed ZOI information and the grading information obtained from Table 2 as follows.

$$\begin{cases} Grade & \text{if } ZOI \le 0.2 \\ Grade + 1 & \text{if } ZOI > 0.2 \text{ and } ZOI \le 0.3 \\ Grade + 2 & \text{if } ZOI > 0.3 \end{cases} \tag{5}$$

The ZOI is a crucial information to decide whether the degree of the defect grade should be increased or kept unchanged. It is inspired by the PACP manual, which assigned higher grade to a defect when it makes up more than 20% of the entire image.[3]

6. Experimental results

In this section, various experiments are carried out on the collected dataset to assess DefectTR's performance under various testing scenarios thoroughly. First of all, Section 6.1 represents the evaluation metrics that were adopted to evaluate different aspects of the introduced model's performance. Next, Section 6.2 describes the hardware and the environment where the model was implemented. Moreover, the hyperparameters of various models are also explained in this section. Section 6.3 offers a group of experiments implemented to assess different aspects of the proposed model. The first experiment in Section 6.3.1 was carried out to check the effectiveness of the pre-processing part in improving the defect detection performance. The proposed DefectTR's performance was then explained in Section 6.3.2. In addition, the detailed comparison between DefectTR and other models was described in Section 6.3.4. Finally, we also show the qualitative evaluation of the proposed model on the ten defect classes and explain some challenging cases in Section 6.3.6.

6.1. Evaluation metrics

During the assessment of the proposed sewer defect detection system's performance, three main elements of the confusion matrix, which include the true positive (TP), false negative (FN), and false positive (FP), are computed to enable the computation of mAP, precision, and recall. For the COCO benchmark dataset, the mAP over various IoU thresholds (the minimum IoU ranges to decide a predicted bounding box is a match) is considered the standard evaluation metric and is calculated for all defect classes. For example, mAP@[.5:.95] points out the average mAP over the IoU values ranging from 0.5 to 0.95 with a default step size of 0.05.

$$mAP = \frac{1}{N_{\text{classes}}} \sum_i AP_i \tag{6}$$

where i is a defect type, and N_{classes} is the total number of defect types, which is 10.

In addition, precision and recall are fundamental metrics for sewer pipe defect localization practically because while precision measures the wrong detection rate, the recall reports the missing detection. The calculations of the mAP, precision, and recall are described as follows.

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$

6.2. Implementation details

The sewer defect detection framework was constructed and trained on PyTorch,[4] an open-source ML library created especially for the Python programming language. In order to guarantee the integrity of the following experiments, all the mentioned deep learning models were trained using the features from the SewerML pre-trained backbone

[3] https://www.oakville.ca/assets/2011%20planning/Pages%20from%20StormSewerMasterPlanPhase1Rep-Appendices.pdf.
[4] https://pytorch.org/.

Table 3
The performance of various models before and after applying the pre-processing module.

Dataset	Approach	Precision (%)	Recall (%)	mAP (%)
Original	SSD [43]	49.4	56.3	41.7
	YOLOv4 [18]	57.3	60.6	46.8
	Faster R-CNN [19]	58.8	62.7	53.4
	CenterNet [44]	59.3	64.3	48.7
	DETR [17]	56.5	66.1	54.5
	DefectTR (Ours)	**58.7**	**68.2**	**55.9**
Pre-processing	SSD [43]	53.1	55.3	43.9
	YOLOv4 [18]	57.9	65.7	47
	Faster R-CNN [19]	57.5	63.2	56.2
	CenterNet [44]	59.7	60.1	51.9
	DETR [17]	61.2	69.3	54.2
	DefectTR (Ours)	**65.4**	**69.7**	**60.2**

Table 4
Ablation study of the DefectTR model.

	ColorJitter	Laprop	LeakyRelu	CIoU	mAP
Model A	✓				56.2
Model B	✓	✓			56.2
Model C	✓	✓	✓		57.8
Model D	✓	✓	✓	✓	60.2

Table 5
Final performance in terms of mAP, precision, and recall of the proposed model on each type of defect.

	BK	LC	CC	DS	DJ	FJ	SJ	LP	SD	RI
mAP	49.7	67.8	70.2	72.7	50.2	57.1	56.1	65.7	52.7	59.3
Precision	58.2	73.9	80.5	78.1	54.2	59.7	59.3	75.4	54.9	60.3
Recall	59.7	80.5	82.2	75.4	60.1	67.4	65.2	79.8	61	66.3

network (ResNet-50) [37]. Notably, the stochastic gradient descent optimization function with the learning rate begins with 0.1, and the momentum of 0.9 is applied to train the backbone. The learning decreased to 0.0001 after 20k iterations. The batch size during the training process of these models was fixed to 4.

For DETR and DefectTR models, the number of encoding and decoding layers was fixed to 6. The number of attention heads inside DETR's attentions was set as 8. Parameters of DefectTR 's encoder were shared among various feature levels. The number of object queries was kept at 100 as the DETR model. Other hyper-parameter settings and training strategies are similar to DETR [17], except that CIoU was used for the matching cost instead of GIoU as reported in the original work. The models were trained for 25 epochs, and the initial learning rate of $1e-4$ was decayed at the 10th epoch by a factor of 0.001. Following DETR, the models were trained using an Adam optimizer with a base learning rate of $2 \times 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$, and weight decay of 10^{-4}. The learning rate of the linear projection are multiplied by a factor of 0.1. The training and inference processes were conducted on an Nvidia Tesla V100 32 gigabyte.

6.3. Defect detection performance analysis

6.3.1. Image pre-processing results

The performances of SSD [43], YOLOv4 [18], Faster R-CNN [19], CenterNet [44], DETR [17], and DefectTR (ours) on the collected dataset with and without the proposed pre-processing process are compared to confirm the crucial role of the pre-processing module in improving the sewer defect detection performance. All of the mentioned models are well-known object detection models. Therefore, they can be directly adapted to the defect detection topic by using the proposed ten-class sewer defect dataset, which was labeled following the scheme of the COCO benchmark.

Table 3 reports the performance of 6 models with and without the pre-processing module. The detection models trained with the pre-processed images achieved higher precision, recall, and mAP than those trained using the raw images. The mAP value increased by 1%–4% for each model after implementing the pre-processing module. For our model, in particular, the pre-processing module advanced the mAP from 55.9 to 60.2. In summary, the pre-processing module is crucial, especially for the videos recorded by CCTV, to enhance sewer defect detection's performance.

6.3.2. DefectTR's performance evaluation

Initially, DefectTR's performance was primarily assessed using the collected dataset.

Fig. 8 presents the class error and mAP values of the training process. The proposed model's training and validation class error reduced significantly after epoch 10th to about 0.15 and 0.10, respectively. The class error values decrease gradually and stop at 0.12 for the training and 0.08 for validation at epoch 25. The mAP value increases sharply to over 0.57 at epoch 10. It then rises steadily and stops at about 0.6 at the end of the training process

6.3.3. Ablation study

This section first performs a thorough ablation study to analyze several parts of the DefectTR model. All the experiments were conducted on the collected sewer defect detection dataset. The experimental results are described in Table 4

The baseline (model A) is configured to be closely similar to the original DETR model without any modification of the model's components, but with an additional ColorJitter augmentation method. Model A achieved the mAP of 56.2%, which is 5% better than the original DETR model. After that, the Laprop optimizer is applied instead of the Adam optimizer (model B). The model showed that it converged 30% faster (150 iterations) than the Adam optimizer (200 iterations), which greatly improved the network training speed. With the utilization of the LeakyRelu activation (model C) and the additional changes from model A and model B, mAP was improved by 1.6 to 57.8 compared to the baseline model A. Finally, mAP reached 60.2 (model D) when the CIoU was implemented to compute the bounding loss instead of the original GIoU. It is concluded that with some chances of the network component, the DefectTR model obtained the highest mAP of 60.2 compared to mAP of 55.7 from the DETR model.

Table 5 describes the mAP, precision, and recall of the proposed model on each class of the defects on the testing dataset. In general, the mAP of over 65% indicates the model works well on LC, CC, DS, and LP defect class. The highest mAP of 72.7% was obtained for the DS class. However, the model performed poorly on two particular classes, BK, DJ, and SD, with low mAP values of 49.7%, 50.2%, and 52.7%, respectively. The reason that led to low performance for BK and SD was that these two classes were similar, and in most cases, BK can be considered SD that was damaged severely. The DJ class is quite similar to FJ and SJ class, so sometimes, the model misrecognizes DJ as FJ or SJ, leading to poor performance detection on the DJ class.

6.3.4. Comparison study for DefectTR

The primary objective of this experiment is to show the advantage of the presented model over existing models, which include SSD [43], YOLOv4 [18], Faster R-CNN [19], CenterNet [44], and DETR [17], using the collected defect dataset. Table 6 summarizes their performances in terms of precision, recall, mAP, and inference time. Among the models, DefectTR appropriately detected different defect types with the best mAP of 60.2%, which was 16.3% higher than the SSD model [36]. As indicated in [17], the inference speed of the transformer-based model was usually lower than that of one-stage detectors. The inference speed of the DefectTR model was 85 ms per image, while the SSD model achieved the fastest inference time at 35 ms per image.

6.3.5. Defect severity analysis

In addition to defect localization, the proposed framework can automatically analyze the severity and ZOI of the detected defects. For this purpose, we extracted the attention weight feature map of the encoder

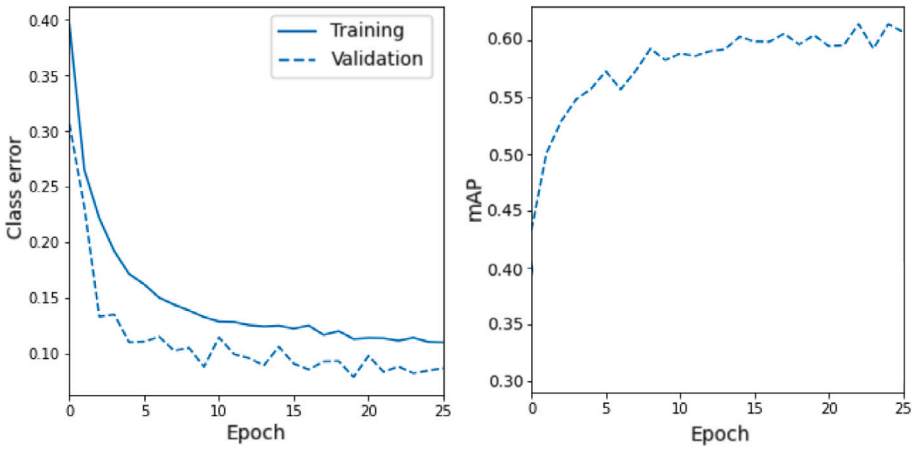


Fig. 8. Class error and mAP curves of the proposed DefectTR model.

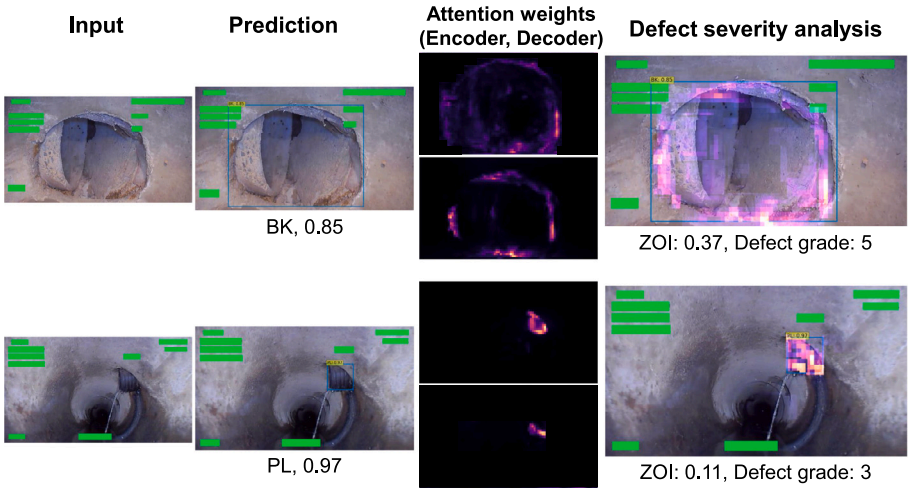


Fig. 9. Sample visualization of the defect severity analysis from the DefectTR system. **Note:** the first column shows input images contained defects; the second column presents the predicted labels, bounding boxes, and their probabilities using the DefectTR model; the third column shows the visualization of the encoder and decoder attention weights; the fourth column offers the ZOI and severity of the detected defect using the attention weights.

Table 6
Performance of DefectTR compared to the other approaches.

Model	Precision (%)	Recall (%)	mAP	Inference time (ms)
SSD [43]	53.1	55.3	43.9	**36**
YOLOv4 [18]	57.9	65.7	47	48
Faster-RCNN [19]	57.5	63.2	56.2	1000
CenterNet [44]	59.7	60.1	51.9	128.2
DETR [17]	61.2	69.3	54.2	83
DefectTR (Ours)	**65.4**	**69.7**	**60.2**	85

and decoder. After that, the mean attention weight feature map was calculated. The ZOI was then calculated based on the mean attention weight feature map. Finally, the defect severity was determined using ZOI. A set of sample cases from this section is described in Fig. 9.

6.3.6. Qualitative evaluations

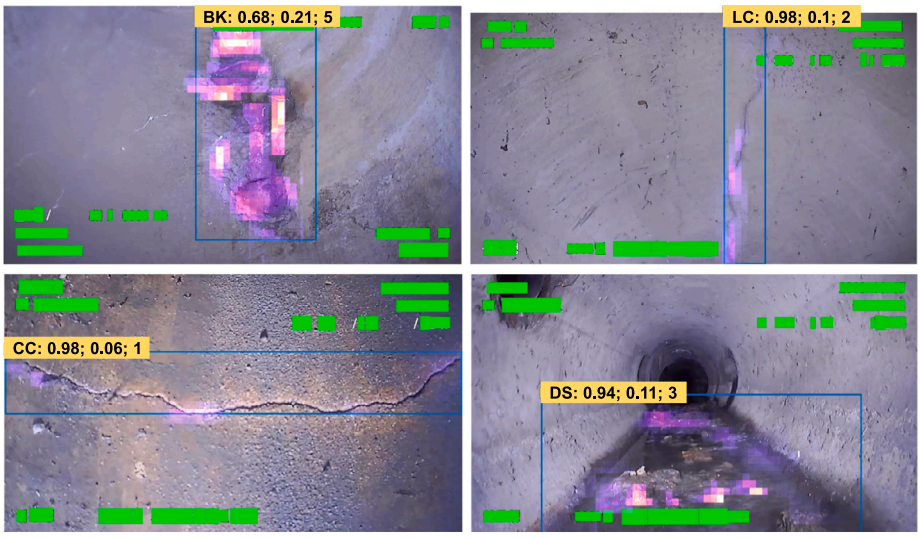
This section quantitatively examines the outputs (detection and defect severity) of the proposed DefectTR structure for each defect type, as displayed in Figs. 10 and 11. Fig. 10 shows that the DefectTR model correctly detected each type of defect and offered additional defect severity information. The mean attention weight feature map overlays the original image to show the whole defect severity analysis process through computing the ZOI. Overall, the model attention focused correctly on the correct part of the defect in order to determine the defect type. Fig. 10b (row 2, column 3) demonstrates that the model correctly predicted two instances of the root intrusion class as two different objects.

Fig. 11 further shows the model's robustness on challenging cases. In the first case, the image contains various defects. The model accurately localized all the defects and provided correct defect severity information. Although the second image is both blurry and contains multiple defects, the model efficiently detected all of them due to the support of the pre-processing module.

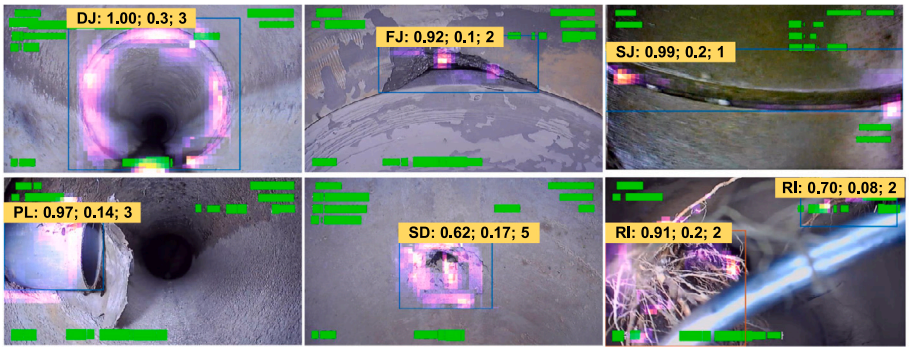
7. Conclusions and future works

This paper introduces a novel transformer-based sewer defect detection for sewer inspection videos. First of all, a total of 47,100 images for 10 types of defects are extracted from a collection of CCTV videos. After that, the corresponding annotations are manually labeled by experts. Various changes, which include colorjitter augmentation, LeakyRELU activation, LaProp optimizer, and CIoU loss, are introduced in order to improve the performance of the original DETR model. In addition, most of the previous defect detection studies can only localize sewer defects. However, this study proved that attention weights, a unique feature of the transformer model, can be utilized to analyze the severity of detected defects to either minor or severe.

The obtained results from various experiments showed that the proposed framework robustly detected 10 types of sewer defects with



(a) Results for BK, CL, CC, and DS



(b) Results for DJ, FJ, SJ, PL, SD, and RI

Fig. 10. The complete output of the proposed framework for each type of defect contains all necessary information, including input image, localized defect, attention weights visualization, ZOI information, and defect grade.

Fig. 11. Outputs of the proposed DefectTR model for challenging cases.

the highest mAP of 60.2%, which outperformed existing standard object detection models. Moreover, mAP value of the proposed model improved significantly from 56.2% to 60.2% compared to the original DETR model. Finally, the proposed framework showed that it could assess the defect severity effectively.

Even though the collected defect dataset proposed in this work contains 10 common types of sewer defects, more defect types can be added for more precise defect label identification, such as lining crack, permanent obstruction, hole, and water intrusion. In addition, the remaining life of a sewer pipe can be predicted if some sensors are utilized along with the CCTV, which would significantly reduce the time required to manually analyze each defect to decide the sewer's remaining life. Finally, the proposed sewer defect detection framework can be optimized in terms of robustness and time efficiency for real-world applications.

CRediT authorship contribution statement

L. Minh Dang: Conceptualization, Methodology, Data curation, Writing – review & editing. **Hanxiang Wang:** Conceptualization, Methodology, Data curation, Writing – review & editing. **Yanfen Li:** Visualization, Investigation. **Tan N. Nguyen:** 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.

Acknowledgments

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

References

[1] E. Vahidi, E. Jin, M. Das, M. Singh, F. Zhao, Environmental life cycle analysis of pipe materials for sewer systems, Sustainable Cities Soc. 27 (2016) 167–174.
[2] S. Madraszewski, F. Dehn, J. Gerlach, D. Stephan, Experimentally driven evaluation methods of concrete sewers biodeterioration on laboratory-scale: A critical review, Constr. Build. Mater. 320 (2022) 126236.
[3] X. Ye, R. Li, Y. Wang, L. Gan, Z. Yu, X. Hu, et al., Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern Chinese city, Front. Environ. Sci. Eng. 13 (2) (2019) 17.
[4] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, P. Fieguth, A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure, Adv. Eng. Inform. 29 (2) (2015) 196–210.
[5] J.C. Cheng, M. Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171.
[6] H. Khaleghian, Y. Shan, P. Lewis, Development of a quality assurance process for sewer pipeline assessment and certification program (PACP) inspection data, in: Pipelines 2017, pp. 360–369.
[7] Y. Li, H. Wang, L.M. Dang, A. Sadeghi-Niaraki, H. Moon, Crop pest recognition in natural scenes using convolutional neural networks, Comput. Electron. Agric. 169 (2020) 105174.
[8] S.I. Hassan, L.M. Dang, I. Mehmood, S. Im, C. Choi, J. Kang, Y.-S. Park, H. Moon, Underground sewer pipe condition assessment based on convolutional neural networks, Autom. Constr. 106 (2019) 102849.
[9] L.M. Dang, S.I. Hassan, S. Im, I. Mehmood, H. Moon, Utilizing text recognition for the defects extraction in sewers CCTV inspection videos, Comput. Ind. 99 (2018) 96–109.
[10] H. Wang, Y. Li, L. Dang, H. Moon, et al., Robust Korean license plate recognition based on deep neural networks, Sensors 21 (12) (2021) 4140.
[11] S. Moradi, T. Zayed, Real-time defect detection in sewer closed circuit television inspection videos, in: Pipelines 2017, 2017, pp. 295–307.
[12] X. Zuo, B. Dai, Y. Shan, J. Shen, C. Hu, S. Huang, Classifying cracks at sub-class level in closed circuit television sewer inspection videos, Autom. Constr. 118 (2020) 103289.
[13] Q. Xie, D. Li, J. Xu, Z. Yu, J. Wang, Automatic detection and classification of sewer defects via hierarchical deep learning, IEEE Trans. Autom. Sci. Eng. 16 (4) (2019) 1836–1847.
[14] D. Ma, J. Liu, H. Fang, N. Wang, C. Zhang, Z. Li, J. Dong, A multi-defect detection system for sewer pipelines based on StyleGAN-SDM and fusion CNN, Constr. Build. Mater. 312 (2021) 125385.
[15] M. Wang, J.C. 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.
[16] M. Wang, H. Luo, J.C. Cheng, Towards an automated condition assessment framework of underground sewer pipes based on closed-circuit television (CCTV) images, Tunn. Undergr. Space Technol. 110 (2021) 103840.
[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[18] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection, 2020, arXiv preprint arXiv:2004.10934.
[19] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.
[20] K. Kaddoura, T. Zayed, An integrated assessment approach to prevent risk of sewer exfiltration, Sustainable Cities Soc. 41 (2018) 576–586.
[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[22] M. Abdullah-Al-Wadud, M.H. Kabir, M.A.A. Dewan, O. Chae, A dynamic histogram equalization for image contrast enhancement, IEEE Trans. Consum. Electron. 53 (2) (2007) 593–600.
[23] Z. Wang, G. Hou, Z. Pan, G. Wang, Single image dehazing and denoising combining dark channel prior and variational models, IET Comput. Vis. 12 (4) (2018) 393–402.
[24] 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.
[25] Y. Shao, L. Li, W. Ren, C. Gao, N. Sang, Domain adaptation for image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2808–2817.
[26] M.R. Halfawy, J. Hengmeechai, Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine, Autom. Constr. 38 (2014) 1–13.
[27] W. Wu, Z. Liu, Y. He, Classification of defects with ensemble methods in the automated visual inspection of sewer pipes, Pattern Anal. Appl. 18 (2) (2015) 263–276.
[28] 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.
[29] 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.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[31] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[32] M. Wang, S.S. Kumar, J.C. Cheng, Automated sewer pipe defect tracking in CCTV videos based on defect detection and metric learning, Autom. Constr. 121 (2021) 103438.
[33] J. Hosang, R. Benenson, B. Schiele, Learning non-maximum suppression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4507–4515.
[34] N. Bodla, B. Singh, R. Chellappa, L.S. Davis, Soft-NMS–improving object detection with one line of code, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5561–5569.
[35] Y. Song, Q.-K. Pan, L. Gao, B. Zhang, Improved non-maximum suppression for object detection using harmony search algorithm, Appl. Soft Comput. 81 (2019) 105478.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[37] J.B. Haurum, 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. 13456–13467.
[38] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, G. Hua, Gated context aggregation network for image dehazing and deraining, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1375–1383.
[39] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, 2015, arXiv preprint arXiv:1505.00853.
[40] L. Ziyin, Z.T. Wang, M. Ueda, Laprop: a better way to combine momentum with adaptive gradient, 2020, arXiv preprint arXiv:2002.04839.
[41] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 658–666.
[42] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 12993–13000.
[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[44] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6569–6578.