DETECTION AND LOCALIZATION OF ROOT DAMAGES IN UNDERGROUND SEWER SYSTEMS USING DEEP NEURAL NETWORKS AND COMPUTER VISION TECHNIQUES

by

Muzi Zheng

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer and Information Science at IUPUI

Indianapolis, Indiana

December 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Shiaofen Fang, Chair

Department of Computer & Information Science

Dr. Mihran Tuceryan

Department of Computer & Information Science

Dr. Yao Liang

Department of Computer & Information Science

Approved by:

Dr. Shiaofen Fang

This project is especially dedicated to my academic advisor who helped and guided me in this dissertation. I would love to express my deepest appreciation to my loving family, especially my parents and my dear husband, who have always motivated and supported me to pursue my dreams during graduate school and life challenges.

ACKNOWLEDGMENTS

Words cannot express my gratitude to my professor for his continuous encouragement, guidance, and patience over the past two years. Also, this endeavor would not have been possible without my defense committees, who generously provided expertise and feedback. Additionally, I am deeply indebted to the support from ClearObject Company which financed my research and provided countless assistance in completing this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

B-box	Bounding Box
CHO	Convex Hull Overlap
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
CDD	Camera-Damage Distance
DCNN	Deep Convolutional Neural Network
HCT	Hough Circle Transformation
GPU	Graphic Processing Unit
IoU	Intersection over Union
I&I	Inflow and Infiltration
LF	Last Frame
ML	Machine Learning
DL	Deep Learning
OCR	Optical Character Recognition
R-CNN	Region-based Convolutional Neural Network
RNN	Recurrent Neural Network
ROI	Region of Interest
TST	Triangle Similarity Theorem
YOLO	You Look Only Once

ABSTRACT

The maintenance of a healthy sewer infrastructure is a major challenge due to the root damages from nearby plants that grow through pipe cracks or loose joints, which may lead to serious pipe blockages and collapse. Traditional inspections based on video surveillance to identify and localize root damages within such complex sewer networks are inefficient, laborious, and error-prone. Therefore, this study aims to develop a robust and efficient approach to automatically detect root damages and localize their circumferential and longitudinal positions in CCTV inspection videos by applying deep neural networks and computer vision techniques. With twenty inspection videos collected from various resources, keyframes were extracted from each video according to the difference in a LUV color space with certain selections of local maxima. To recognize distance information from video subtitles, OCR models such as Tesseract and CRNN-CTC were implemented and led to a 90% of recognition accuracy. In addition, a pre-trained segmentation model was applied to detect root damages, but it also found many false positive predictions. By applying a well-tuned YoloV3 model on the detection of pipe joints leveraging the Convex Hull Overlap (*CHO*) feature, we were able to achieve a 20% improvement on the reliability and accuracy of damage identifications. Moreover, an end-to-end deep learning pipeline that involved Triangle Similarity Theorem (*TST*) was successfully designed to predict the longitudinal position of each identified root damage. The prediction error was less than 1.0 feet.

CHAPTER 1. INTRODUCTION

Underground sewer pipelines are utilized for the collection and transportation of wastewater and stormwater. The maintenance of the sewer system is a major challenge partly due to the roots penetration from nearby plants which are hard to notice. These roots would grow through pipe cracks or loose joints to reach the moisture, which may eventually cause serious pipe blockages and collapse without taking actions. According to the EPA (2014), pipe cracks and blocks often deteriorate wastewater pipelines with the conditions of inflow and infiltration (I&I), problematic overflows, or sinkholes thus may cause an extra treatment expense of up to $1.3 per 1,000 L sewage. Therefore, periodical inspections of the sewer pipelines are crucial to detect defects and discover failures for the maintenance of a healthy sewer infrastructure (Cheng & Wang, 2018).

Due to the hidden conditions and complex structures of the underground sewer networks, human access to inspect the sewer pipelines would be difficult and dangerous. Presently, the application of Closed-Circuit Television (CCTV) in sewer systems is considered as the most widespread practice for visual pipeline inspections, which mounts a television camera on a robot that can be navigated through the inner pipelines (Figure 1). During the inspection process, a trained operator is capable of remotely controlling the robot to stop, return left or right for a close look at any suspicious pipe defects, and eventually generate a recorded video for further defect examination and interpretation. Thus, CCTV inspection videos can be reviewed either in operation or offline, to let the operator manually assess and mark the defect conditions as a digital inspection record for each inspected sewer network. Once the defect is identified and localized, the ground at a certain location would be dug to let the operators access and repair/replace the defect pipelines (Abhinaya, 2021). However, manually reviewing hours of inspection videos is time-consuming and labor-intensive (Cheng & Wang, 2018). Other factors such as the operator's skill, experiences, exhaustion, and concentration would also significantly affect the reliability and consistency of inspection interpretation, which may lead to further structural defects in the sewer pipeline system. According to the research from Dirksen et al (2013), operators overlooked approximately 25% of sewer defects in the inspection process. Therefore, a faster, more efficient, and more reliable technique to automatically identify root damages and their positions in the sewer systems would be critical to avoid pipe collapse and emergency repair costs. All the time-consuming and tedious

tasks of manually detecting damages in hours of inspection videos can be empowered by computer processing to save a considerable amount of time.

Figure 1. An example of CCTV inspection in sewer pipelines.

In accordance with the requirements of the National Association of Sewer Service Companies (NASSCO, 2018), it is mandatory to comprise the location information of sewer damages within the report of sewer inspection in the United States. The specific damage locations include the circumferential position relative to the pipe's cross-section, and the longitudinal position away from the pipeline entrance. For the circumferential position of sewer damages, many studies have demonstrated that their machine learning-based models perform excellently not only on the detection and classification of sewer damages but also on the capability to accurately identify the circumferential location of the damages (Kumar et al., 2020, Cheng and Wang, 2018) (Figure 2). In addition, for CCTV inspection videos there is a piece of specific equipment attached to the moving robot to measure the distance traveled, thus the longitudinal position of each sewer damage could be achieved by reading the traveling distance from video subtitles using text recognition techniques (Jahanshahi, M. R. 2011, Hassan et al. 2019, Dang et al. 2018).

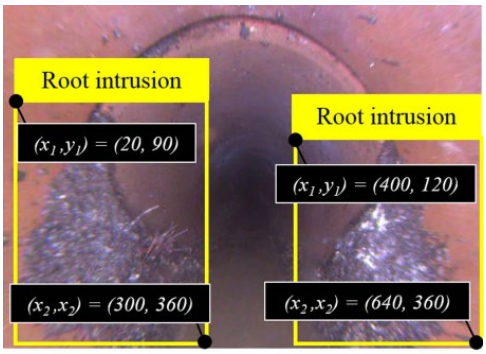
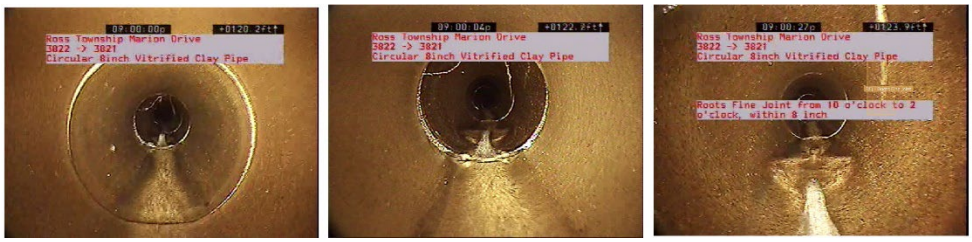


Figure 2. An example of detecting and localizing root damage in sewer pipelines is where the bounding box (b-box) identifies the circumferential position of each damage (Kumar et al., 2020).

Figure 3 indicates how the given root damage (Figure 3a) goes from in the distance to closer (Figure 3b) and finally passes by the robot camera (Figure 3c). Note that in this example, the information of traveling distance appears on the right image corner representing how far the robot traveling forward (in feet). Accordingly, when the robot camera first detects the damage, the camera-damage distance (CDD) might be up to 5 ft depending on the camera's intrinsic parameters, then the value of CDD would gradually decrease to 0 as the robot travels toward the damage and eventually passing it. However, to determine the longitudinal position of each sewer damage, it is not reliable to simply report the traveling distance recognized from the video subtitle in the last frame (LF) where the camera captures the last bit of the root damage. Different factors such as camera shot angles, size of pipe diameters, uneven traveling surface, and other unpredictable conditions within the sewer pipeline would significantly affect the accuracy of localizing the longitudinal position of sewer damages using the LF approach. For instance, if the robot must stop before reaching the damage (due to out of battery, pipe blockage, trapped, etc.), it's impossible to achieve the actual longitudinal position of the sewer damage recognized from the LF, and it would be also difficult to predict the damage position since there is still some distance ahead of the camera to reach the damage. Under this situation, the distance value recognized from the video frame where the robot stops would be smaller than the actual damage position in the sewer pipeline, leading to inappropriate decision-making for damage localization. Therefore, a new approach based on Triangle Similarity Theorem (TST) was proposed in this study, along with the benefits of deep learning-based models to predict the longitudinal position of each sewer damage when the robot first sees it. Although many researchers such as Ahrary, A. et al. (2005) and Draelos et al. (2015) have studied how to localize sewer damages based on the stereo camera which would be

more efficient and accurate, it should be noticed that in this study all of the CCTV inspection videos were collected from different resources and they were recorded by monocular cameras (instead of stereo/depth camera) without knowing camera's intrinsic and extrinsic parameters. Hence, TST is a robust and efficient way to estimate CDDs in this application, and it further leads to the determination of the longitudinal position of each root damage within the sewer pipelines. More details will be discussed in the following sections.



(a) (b) (c)

Figure 3. Inspection images with a certain root damage. (a) when the camera first sees it; (b) when the robot moves closer, and (c) when the robot just passed by the point of the condition.

Overall, this study proposes an end-to-end deep learning (DL) pipeline that first supports damage detection and text recognition in CCTV inspection videos, and later predicts the longitudinal position of each identified root damage in the sewer pipelines. Four main modules were developed to process and train such DL pipelines. The first module was to extract representative keyframes that preserve visual features in each video. Such selected keyframes were combined as a video summarization to eliminate redundant frames to decrease the computational time of video analysis. The second module was applied to recognize the subtitle texts that contain distance values in keyframes, thus generating a detailed distance map for each inspected sewer system; The third model was designed to identify root damages and localize their circumferential positions from keyframes; The last module was used to predict the longitudinal position of each identified sewer damage. In general, this pipeline has the potential to assist existing CCTV systems to improve inspection procedures and notify operators in advance to avoid any missing defects from the extensive sewer networks.

CHAPTER 2. RELATED WORK

2.1 Vision-based Defect Detection using Deep Learning Approach

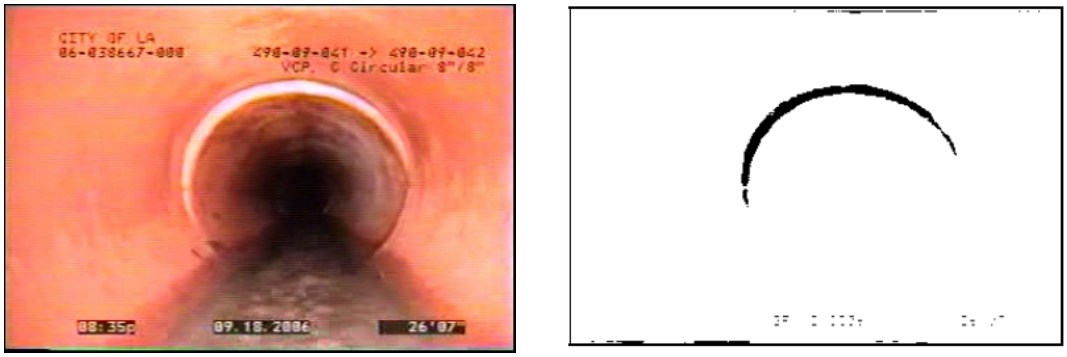
Conventional computer vision methods were commonly applied to extract complex features of underground pipelines, to automatically interpret CCTV inspection videos for various damages in the sewer systems. For example, morphological operation and segmentation were proposed to detect different pipeline defects such as pipe joints and fractures, cracked and damaged pipes (Yang &Su, 2009). In addition, Halfawy & Hengmeechai (2014) confirmed that morphological operations can be successfully used to enhance pipeline cracks with the benefits of applying sober operator and Hough transform as the image pre-processing step, then crack segments would be extracted with a customized filter. Also, with the advantages of video sequences, many techniques were proposed to automatically detect faults from videos based on frame differences combined with various image processing and shape analysis (Guo et al., 2009, Hawari et al., 2018).

However, with the advent of ML/DL approaches applied in various computer vision tasks (object detection and segmentation), there were many vision-based models to detect root damages from the underground sewer systems (Moselhi et al., 2000, Meijer et al. 2019, Jack et al., 2018, Li et al., 2019, Cheng and Wang 2018; Kumar et al. 2020). For instance, the study by Hassan et al. (2019) applied CNNs on the extracted frames of CCTV inspection videos for the classification of sewer defects, and the accuracy can reach up to 90%+. CNNs have also been widely used for object detection to learn abstract features from different images. There are a set of layer components that build up the deep neural network, including convolutional layers, pooling layers as well as fully connected layers with the application of activation functions. A rectangular receptive field slides through input image dimensions to calculate kernel weights, and the generated feature map will be fed directly to the next layer. In addition, the pooling layers (e.g., average pooling, max pooling, and L2-norm pooling, etc.) down-samples the extracted features from the previous layers to reduce total model parameters, accelerate computational speed, and prevent the network from overfitting. Moreover, the fully connected layer is applied as the final layer, to transform the feature map into vectors, thus computing class probabilities with appropriate active functions (e.g., sigmoid, tanh, softmax, and relu) (Moradi, S., 2020). These activation functions provide nonlinearities to the

neural network, to update weights and losses with gradient backpropagation within the network. Therefore, the capability of automatic feature extraction during the training stage leads to new insights into deep neural networks over traditional ML algorithms (Moradi, S., 2020). The input images are fed into the DL algorithms leading the deeper layers of the neural network to extract more complex features of the input images.

The DL-based object detection models can be categorized into three approaches, including sliding-window detection, two-stage object detection as well as single-stage detectors. The sliding-window approach benefits from CNNs over multiple sub-windows of an image and aggregates the classifications (Lee, J., Bang, J., & Yang, S. I., 2017, Sudowe, P., & Leibe, B., 2011). In addition, the two-stage object detector increased the computation speed with the benefit of a CNN-based region proposal (e.g., the R-CNN, Fast R-CNN, and Faster R-CNN) (Ren et al., 2015, Kumar, S. S, 2020). Commonly two-stage detectors of most DL-based techniques are applied for less complex image datasets, which may lead to some restrictions for practical applications within inspection systems. However, single-stage detection approaches such as SSD and YOLO could significantly increase computational speed as compared to two-stage detectors, due to the elimination of the region proposal step. The most common technique YOLO was introduced by Redmon et al. (2016), to detect objects by passing the image once. Applied with such a detection module, the input image can be divided into N x N grid and the object b-boxes and associated confidence scores are predicted in each cell (Moradi, S., 2020). Compared with an older version of YOLO modules, the YOLOv3-based single-stage detector represents a faster and more accurate approach to real-time object detections (Redmon & Farhadi, 2018, Zhang et al., 2020). Studies from Kumar et al. (2020) and Yin et al. (2020) applied YOLOv3 to prove that this technique is efficient to detect various pipeline defects in sewer systems. However, the acquisition of manual annotations for the Yolov3 model is labor-intensive, thus many studies have investigated the detection of sewer defects based on anomaly detection algorithms, which can reduce the workload for manual labeling (Myrans et al., 2018, Fang et al., 2020) since most of the frames from each CCTV inspection video stream don't contain any sewer defects. In 2020, Fang et al. proposed a defect detection approach based on unsupervised anomaly detection algorithms with feature extraction to videos, in which the overall detection accuracy of sewer defects was above 90%.

Moreover, the detection of pipe joints should be considered as a significant step since the majority of root damages grow through the joints causing misalignments and infiltrations. The most robust algorithm that Jahanshahi (2011) evaluated to extract joint features was by applying the morphological operation on the original image without preprocessing and postprocessing needed. This operation is useful since the sewer joints from Jahanshahi's study have brighter intensities than the background (Figure 4). However, the CCTV inspection videos collected in this study have more complex pipe structures, where using the conventional computer vision technique with morphological operation is not robust enough to extract sewer joints.



(a) (b)

Figure 4. a) Original CCTV video frame that contains a sewer joint on a Vitrified Clay Pipe, b) joint extraction using morphological operation on the original frame (Jahanshahi, M. R., 2011).

To identify sewer defects, there are also many state-of-art models for semantic segmentation based on DL methods which localize and distinguish different objects by classifying every pixel. The study of Ronneberger et al (2015) proposed a method for image segmentation called U-net (Figure 5). This model has a symmetric U-shaped architecture that includes encoder and decoder parts where the encoder path using a convolutional neural network aims to extract features and patterns from the image, and the decoder path using transposed 2D convolutional layers to re-create a full binary image. This U-net approach takes the advantage of data augmentation techniques to make more efficient use of the dataset without adding more labeled data in the training process. Also, the architecture contains links between the encoder and decoder paths to combine features from different spatial regions of the image that enables precise localization. Many researchers such as

Pan, G et al (2020) have widely applied the U-net segmentation model for detecting sewer defects with good results.

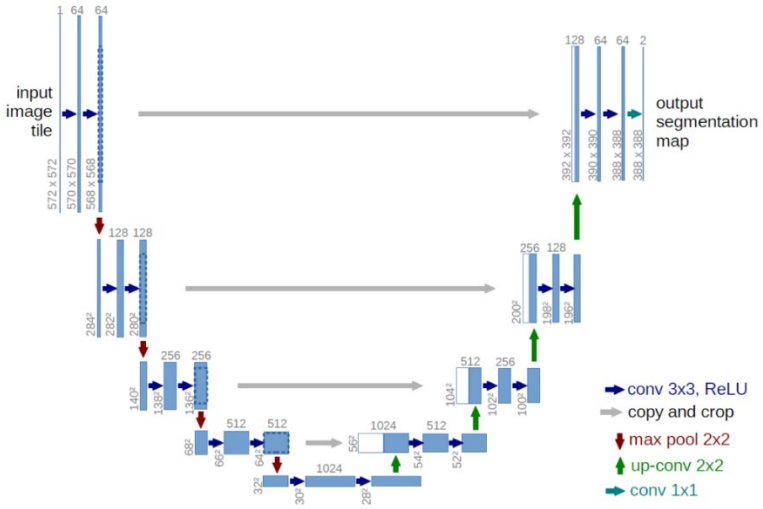


Figure 5. Semantic Segmentation Model U-net Architecture.

Commonly, text subtitles of CCTV inspection videos contain many details regarding the condition of each underground sewer system, such as the traveling distance of inspected robot, pipe diameter, pipe material, etc. Hence, the recognition of text subtitles is crucial to extract valuable information for further video analysis. Most of the applications of deep neural networks such as CNNs are devoted to object detection and classification (Girshick, R. et al., 2014, Krizhevsky, A. et al., 2012). However, visual objects such as scene text and handing writing tend to occur in the form of sequence thus would be better to recognize a series of text characters instead of detecting isolated characters (Shi, B., Bai, X., & Yao, C., 2016). Also, text length usually varies drastically thus some DL techniques that operate on fixed dimensional datasets such as deep convolutional neural network (DCNN) (Ptucha, R. et al., 2019, Krizhevsky, A. et al., 2012) are not capable to apply on these variable-length sequence predictions. The algorithm of DCNN is a character detector trained with labeled character images, which is required to segment each character individually from the original word image precisely and correctly. However, for a more efficient approach to handle such sequence problems, recurrent neural networks (RNNs) were designed and developed with the major advantage of not relying on the character positions in a sequence-like object image. Moreover, some traditional text recognitions outperformed as compared with the previous neural network-based modules. For example, some research applied multi-scale representation and mid-

level features for the recognition of scene text (Yao et al.,2014, Gordo et al, 2015). Others like Rodriguez-Serrano, J. A., Gordo, A., & Perronnin, F. (2015) proposed learning a Euclidean space where word images and text labels are embedded, then treat the text recognition problem to a retrieval problem which is to find the closest word label in this space.

For CCTV inspection videos collected in this study, there were many limitations such as varying illumination, poor lighting conditions, complex natural background, or damaged video clips. For image-based sequence recognition, reading unstructured text is also challenging since these texts are sparse, and usually appear at random places in a natural scene without standard text font, color, and size. Thus, a text recognition engine was utilized to extract the subtitle which is treated as a sequence-like object. The algorithm includes two phases: a) text localization by applying the YOLOv3 module to identify the text lines, with the benefits of text quality enhancement beyond the background. b). text recognition using optical character recognition (OCR) modules to recognize the distance value that indicates how far the robot traveling along the sewer pipelines. There are increasing researchers studied on text recognition, but presently the most popular techniques are Tesseract Engine and CRNN-CTC network.

Tesseract, an open-source OCR engine, was implemented to recognize subtitle text that adapts to over 100 languages. This engine was originally developed by HP and recently taken over by Google. Dang et al. (2018) indicated that Tesseract performs impressively on the text recognition tasks for sewer imaging, but the performance might be poor in the unstructured text that containing with significant noises since Tesseract was initially created to recognize structured text data. In addition, the CRNN-CTC network is a combination of CNN, RNN, and CTC loss (Connectionist Temporal Classification). Not like traditional OCR models, the application of the CRNN network does not perform operations like character segmentation and can recognize text sequences of arbitrary length. Accordingly, CNN is applied to extract visual features from text lines such as color, font, stroke, etc., while deep bidirectional RNN along with LSTM structure is implemented to extract contextual features of texts thus predicting sequential output with some relation between the characters (Chen, L., & Li, S., 2018). Additionally, as shown in Figure 6, the transcription layer attempts to convert the per-frame predictions generated from RNN layers into a label sequence; while the CTC loss function applied in the transcription layer aims to prevent a single

character that spans multiple time-step which needs further processing if without CTC, which is useful in sequence-to-sequence OCR models. This CRNN model performs fast because of the less total number of parameters (8.3 million) as compared with other DCNN models, which is suitable to be deployed on edge devices (Shi, B., Bai, X., & Yao, C., 2016).

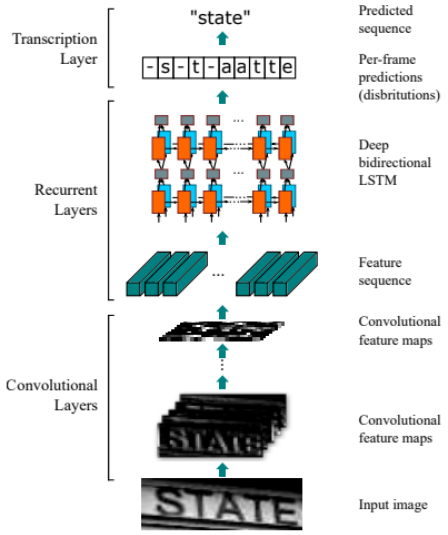


Figure 6. CRNN+CTC network.
(Shi, B., Bai, X., & Yao, C., 2016).

2.2 Longitudinal Position of Sewer Damages

In each CCTV inspection video, the distance values shown in the subtitles changed when the robot moved along the sewer pipeline. In this study, the length of sewer pipelines could be up to 300 ft. The distance value recognized from the last frame where the robot camera captures the last bit of the root damage was usually considered as the actual damage location by other studies. For example, Hassan, S. I., (2019) applied a text recognition model using the last frame approach to identify the longitudinal position of each sewer defect after classifying the defects based on CNN models. Moradi et al. (2020) also used the distance information that was recognized from video subtitles in the last frame to locate the identified anomaly in sewer pipelines. From another perspective, the distance or depth estimation of an object in a sewer pipeline system can also be estimated based on stereo vision with two cameras, which are widely used in the application of intelligent transportation systems, robotics, location identification, and object tracking, etc., (Megalingam et al, 2016). Like the stereo/depth camera, the camera-to-object distance using a

single camera can be estimated by applying two convex mirrors that imitate the use of two cameras to create a disparity map (Murmu, N.; Nandi, D., 2014). Later, In-Sub Yoo and Seung-Woo Seo (2015) proposed a frequency domain analysis-based algorithm for distance estimation, and Kumar M.S.S., et al, (2013) computed the distance based on inverse perspective mapping for face distance estimation (Bharade, A., 2014). Furthermore, Triangle Similarity Theorem (TST) was also widely applied for distance estimation. For instance, Megalingam et al (2016) used TST to accurately estimate the camera-object distance with the actual object width and calibrated camera focal length (Nienaber, S., et al., 2015). Some common ways to obtain focal length are based on camera calibration (using an image such as a chessboard) or checking camera specifications/EXIF data from the image that the camera captured. But Megalingam's study was based on a calibration phase that taking three pictures at different distances from the camera, then the focal length (F) would be calculated with the known pixel width of the object obtained from the image (Wp), known distance values (D), as well as the known width of the object (Equation 1). Therefore, after the calibration with the calculated focal length, camera-object distance can be achieved using the traditional TST equation. However, with the advantage of video sequences, this study proposed a more straightforward algorithm without knowing the object's actual width and the camera's focal length. Details will be brought in Section 3.2.2.5.

$$F = \frac{Wp * D}{W} \tag{1}$$

Therefore, building upon previous applications and algorithms, this study provided a deep neural network-based approach to predict the longitudinal and circumferential positions of each identified root damage within the sewer pipelines. Transfer learning models such as fine-tuning CNNs, Tesseract, as well as CRNN modules were applied with various computer vision techniques.

CHAPTER 3. METHODOLOGY OVERVIEW

3.1 Proposed Framework

This proposed framework intends to improve sewer inspection by detecting the root damages and identifying their circumferential and longitudinal positions from CCTV inspection videos. To reach this goal, there are three phases included: a) keyframe extraction; b) subtitle recognition; c) root damage identification; d) damage position prediction. Specifically speaking, based on the sum of absolute differences in a LUV color space between two consecutive video frames, a certain number of local maxima would be first extracted as keyframes (Schanda, J., 2007, Amanpreet Walia, 2018). As a result, a large image dataset was constructed from each inspection video, and a single-shot detector YOLOv3 was implemented to localize the text region of interest (ROI) in every keyframe. Once the text region was founded, the OCR model was applied to recognize the subtitle text that indicates distance values (in ft), to track the distance covered by the inspection video to reconstruct a detailed map of the sewer pipeline. Moreover, a pre-trained segmentation module was implemented to identify root damages and their circumferential positions. By applying another YOLOv3 module for the detection of pipe joints leveraging the Convex Hull Overlap (CHO) feature, lots of false positive predictions generated from the segmentation model were significantly reduced thus improving the accuracy of the identification for sewer damages. At the last, the longitudinal position of each identified root damage was mathematically derived based on the concept of the Triangle Similarity Theorem (TST) and validated with the ground truth value using the LF method.

Thus, specific requirements in this study include:

1) The DL pipeline should be run automatically to decrease the inference time with the input of raw CCTV inspection videos.
2) All proposed models and algorithms should have the capability to analyze CCTV inspection videos with various illumination, size, and quality.
3) This system should detect and localize root damages as well as recognize subtitle text from CCTV inspection videos, with high consistency and accuracy.

4) The DL pipeline should predict the longitudinal position for each identified root damage (in ft) with errors in an accepted range.

3.2 Methodology and Model Description

3.2.1 Keyframe Extraction

Keyframes decrease the total number of video frames, which are defined as the frames in the video sequence where a sudden change occurs. In this study, the inspection robot usually moves quite slowly through the sewer pipes, causing too many redundant and repeated scenes recorded in consecutive frames. Many CCTV inspection videos also contain irrelevant content (sky, human, manhole cover, etc.), especially at the beginning or end of the videos. To reduce the computation cost for future model development, it's significant to select a set of keyframes that can purely represent the inspection sequences in the sewer networks. In this study, color features of the video frames that used absolute differences in LUV colorspace were applied to check the similarity between consecutive frames. The colorspace "LUV" is a CIE-derived color space that was adopted by the International Lighting Commission on Illumination (CIE) in 1976 (Schanda, J., 2007, Burger, W., & Burge, M. J., 2010). Unlike the most common Red, Green, and Blue ("RGB") colorspace, LUV decouples the "color" (chromaticity values represented by U and V) and "lightness" (illuminance represented by L) from color images and describes the image in a way that is much more perceptual uniform. Thus, LUV colorspace is more suitable for image difference comparisons, and better to represent objects for the applications of object detection (Dollár, P., etc. al., 2014, Rahimzadeganasl, A., & Sertel, E., 2017). Also, it's significant to extract representative keyframes from each inspection video for further object annotations, as well as create training datasets to accelerate and optimize the object detection process in images (Toledo Ferraz, C., 2021).

Converting the image color space from RGB to CIE LUV was considered as the first step, The luminance and chromaticity values in LUV are useful to detect image features with a remarkable precision (Rahimzadeganasl, A., & Sertel, E., 2017). For each CCTV inspection video, a certain number of keyframes were extracted according to the sum of absolute differences in LUV colorspace and local maxima for the nearest frames (Toledo Ferraz, C.,2021). For preserving the information of subtitle text, the dimension of the smoothing window was 5 frames. Therefore,

keyframes acquired from the videos were later applied to recognize the subtitle text, as well as identify the root damages and predict their longitudinal positions within the sewer network.

3.2.2 Dataset Preparation

A total of 200+ CCTV inspection videos were collected from different resources and can be accessed from the Google Cloud Platform (GCP) provided by the company. The dataset applied in this study contains 20 random videos, and the lengths of these videos vary between 10 to 20 minutes, and the diameters of sewer pipelines range from 3 to 5 feet. The frame rate of each video is usually 30 frames per second (fps), and the subtitles embedded in videos typically contain details such as date/time, project ID, sewer address, pipe size, circumference size, traveling distance (in feet), operator's information and so on (Table 1). For different video types, subtitles may appear at different locations with various text fonts, sizes, and colors. All the training and validation from different models were carried out with a GPU using NVIDIA Tesla T4 with TensorFlow 2.7 environment.

Table 1. Subtitle information in CCTV inspection videos.

Subtitle Information	Details
Traveling distance	Distance from the pipe entry
Traveling direction	Forward or backward
Inspection date	Date of the pipe inspection
Inspection time	Time of the pipe inspection
Pipe type	Pipe material and type
Pipe number	Pipe id from start to end
Pipe circumference	Pipe size

All color images were first converted from RGB to grayscale to accelerate the computational time. Due to the natural characteristics of internal sewer pipelines, the image pre-processing stage is essential in recorded videos to remove various artifacts and noises for image enhancement. Especially for sewer inspection videos, there are attempts to have noise due to the illumination variation and camera movement. These noises might be random or generated by the camera system to become coherent noises. Thus, the application of the Gaussian filter in this study is considered as an effective algorithm to reduce such noises and blur images without affecting the brightness of

the image. This filter performs excellent, especially for text recognition since the noise present in video subtitles is similar to gaussian noise (Moradi et al. 2020), and usually filter size of [3,3] with $\sigma = 1.1$ was applied in this study. The median filter is another effective way to enhance image quality, but it works better for salt and pepper noises and more edge preserving as compared to the Gaussian filter. However, for the application of object detection using deep neural networks such as CNNs, the steps of image pre-processing and feature extraction are excluded, as compared with conventional machine learning approaches (Figure 7) (Moradi et al. 2020).

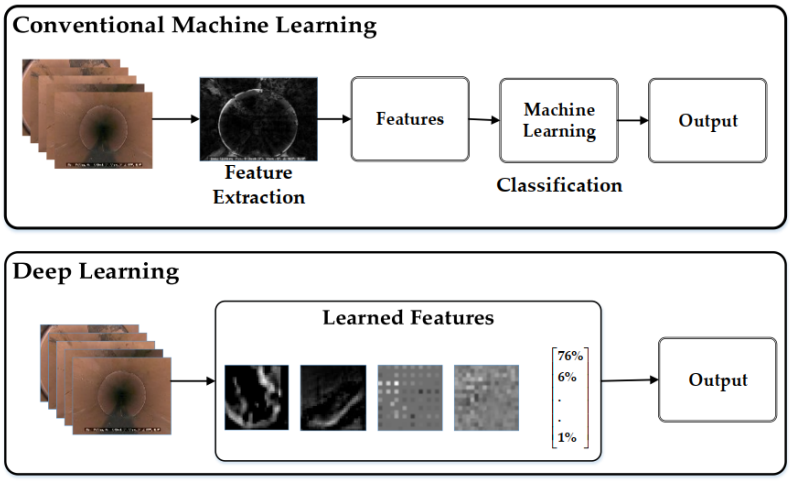


Figure 7. The comparison of conventional ML and DL techniques (Moradi et al. 2019a).

3.2.2.1. Data augmentation

The accuracy of object detection models can be significantly increased due to the application of data augmentation by artificially creating a large training dataset, which could reduce overfitting and improve the generalization capability for the developed models. Various augmentation methods were applied to considerably increase the size of provided dataset, including horizontal flipping, contrast enhancement, sharping, blurring, raising hue values, resizing, shifting, adding Gaussian noises, and so on.

3.2.2.2. Text ROI localization and recognition

To recognize text subtitles in a keyframe, the region of interest (ROI) that contains traveling distance values were required to be identified. In most inspection videos, text regions provide

consistent pixel intensity, scale, and illumination variations, thus having a substantial difference as compared with the background. But several video observations tell that the text varied in sizes, font colors among different videos, and even the location of text ROI may change in a single video which might be due to the camera's internal error when recording. Hence CNN-based model such as YOLOv3 was applied to localize text ROIs and thus prepared for further text recognition. About 5000+ frames from these 20 inspection videos were randomly selected, and these images were annotated with bounding box (b-box) coordinates that indicate the ROI in the image. Accordingly, 500 images were labeled from these 20 videos and then saved in .txt files required by the YOLOv3 module. Among the dataset, 80%, 10%, and 10% were randomly selected for training, validation, and testing process, respectively, with training epochs of 10000. To be more specific, the numbers of batches and subdivisions for the training process are 16, and 4 respectively in the darknet YOLOv3 network. Also, the network would keep learning until reaching max batch times, and the value of max batch is usually equal to the number of classes multiplied by 2000, thus setting it to 4000 in this study. The detection object is one so class = 1 for all the detection tasks using the YOLOv3 network, hence total filters would be 18 which is equal to (classes + 5) * 3 where 3 represents the number of masks. Moreover, the learning rate determines how fast the model converges, according to the weight updates in neural networks or coefficient changes in linear/logistic regression (Ray Smith, 2007, Martin Thoma, 2018). In this study, the learning rate was set to 0.01. With 10000 epochs, the total computational time lasted about 2 hours and the highest accuracies reached from the training and validation process were 99% and 96%, respectively.

For further testing on every new video, about 10 random keyframes in each video were selected for the application of text ROI localization using the YOLOv3 model, and the median b-box prediction returned as the actual position of the text region in case sometimes the ROIs (even in a single video) varies at different locations in images. Then the detected ROI was fed into different OCR modules to recognize the distance value from keyframe subtitles. For the application of Tesseract Engine, it can be directly run on the ROI in keyframes to extract coarse words for each character without taking the extra effort of manual annotations. The main image-preprocessing steps in this recognition phase include interpolation, resizing, and applying gaussian blur along with OTSU binarization. Also, Tesseract has the benefit of a whitelist function that only reads

numbers without other characters to increase recognition accuracy. With a certain number of keyframes extracted from each video, there were 5000+ images (after data augmentation) were applied as the input dataset. For another recognition module, CRNN-CTC, it needs to crop the identified ROI first and prepare the cropped sub-image for the text annotation. There were 250 sub-images, and their ground truth labels were created manually which include English numbers and alphabets. The size of the dataset was enlarged considerably to 1250 images by applying data augmentation techniques such as scaling, adding noise, adjusting brightness, sharing, blurring, etc. Also, this CRNN model was implemented using Keras high-level API (https://keras-ocr.readthedocs.io/en/latest/index.html) with Tensorflow backend and ran on GCP.

3.2.2.3. Root damage segmentation

For the pixel-wised detection of root damages in the sewer system, a pre-trained segmentation model was provided by the company, and it was kept updated with an increasing training dataset (Figure 8). The dataset only contains root damages which were annotated with closed polygons. However, this segmentation model generated lots of false positive predictions that consist of non-damage pixels which are incorrectly identified as damage pixels, mainly due to the presence of similar damage patterns on the pavement surface.

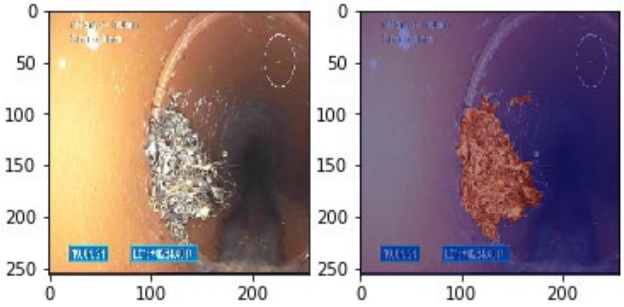
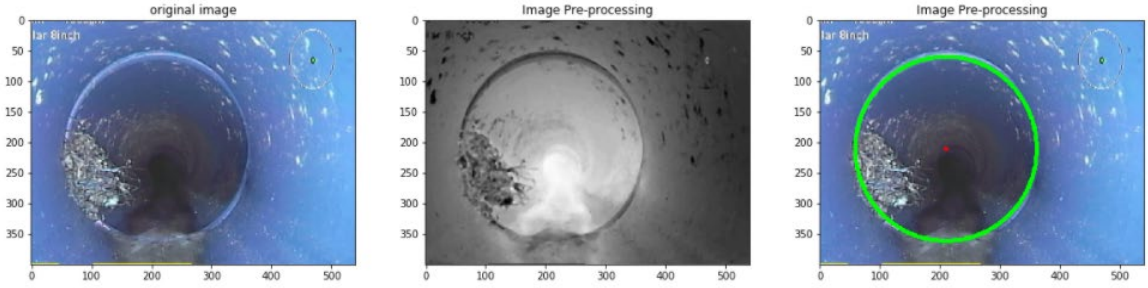


Figure 8. Pixel-wised segmentation for root damages in a CCTV inspection video.

3.2.2.4. Pipe joint detection and the Convex Hull Overlap (CHO) feature

To reduce false positive occurrences of damage segmentation in CCTV inspection videos, leveraging the location advantage of pipe joints for better detecting these damages would be significant. The reason is that the majority of root damages are grown through the pipe joints.

Accordingly, the YOLOv3 module can be first applied to localize and detect pipe joints. However, since CNN-based detection models such as YOLOv3 is supervised learning algorithm which depends on defect labels that are considered as the ground truth for the training stage, thus the biggest limitation is to manually annotate the samples which are considered labor intensive. However, most of the joint shapes are fixed-size circles, thus Hough Circle Transformation (*HCT*) technique can assist to identify and annotate these circular joints as shown in Figure 9. The reason why HCT cannot be used for direct joint detection is firstly due to the complex structure of various pipeline systems. Some of them have a very noisy nature background with vague joints thus HCT is not robust enough to successfully detect such pipe joints. Another reason is due to the difficulties in automatically adjusting HCT parameters (such as minimum circle distance, minimum and maximum radius, etc.) under various types of pipelines or illumination conditions.



(a) (b) (c)

Figure 9. Hough Circle Transformation (HCT) to identify pipe joints. (a) original image; (b) after image pre-processing, and (c) detected pipe joint using the HCT technique.

In total, 13,580 images with corresponding joint annotations were generated by applying the HCT approach. After manually removing incorrect HCT detections for the circular joints, 11,472 outputs were used as the dataset prepared for the YOLOv3 detection model. However, if the application of HCT was not robust enough to correctly detect the joints such as in the example shown in Figure 10, then manual annotations are necessary for this certain scenario. In addition, 80% and 10% of the dataset were used for training and validation purposes, respectively, while another 10% was used for the testing stage. Due to the various resolutions of CCTV inspection videos, all images need to be resized to 416 x 416 pixels which is the standard input resolution for a balanced Yolov3 model. Later, to improve the accuracy for the identification of root damages, a Convex Hull Overlap (*CHO*) feature, which reflects the overlap relationship between the detected

pipe joints and the segmented root damage, was applied to increase the accuracy for the identification of sewer damages. This feature aims to reduce the false positive predictions generated from the previous segmentation model. To be more concrete, a Convex Hull represents the smallest convex set that involves all the pixels of a region in the Euclidean plane (Liang, Y., 2018), which can be applied to segmented root damage. Also, after achieving the b-box coordinate of the pipe join returned by the YOLOv3 module, the segmented pixels of the circular joints can be determined by drawing with the radius (half-length of the b-box width/height) from the center pixel of the b-box detected. Therefore, the CHO feature was computed as the overlapped pixels between the segmentations of the detected pipe joint and root damage. Any segmentation outputs located out of the pipe joints without a CHO feature were considered as false positives and need to be excluded. More results regarding the CHO feature are well demonstrated in Figure 20 of Section 4.2.

Figure 10. The HCT limitations to detect pipe joints with complex pipe structures.

3.2.2.5. Camera-damage Distance (CDD)

This section describes how to estimate the camera-damage distance (CDD) using monocular vision in each CCTV inspection video. First, input keyframes should undergo different detection models along with a series of image processing steps for identifying desired root damages and pipe joints that were mentioned in earlier sections. Then the Triangle Similarity Theorem (TST) was proposed, and Figure 11 demonstrates how to apply this projective geometry approach to estimate the distance between the root damage and the camera in the sewer pipeline. The assumption of this geometric concept is to consider the robot camera as an ideal pinhole camera, and the camera's optical axis and the pipe walls are parallel.

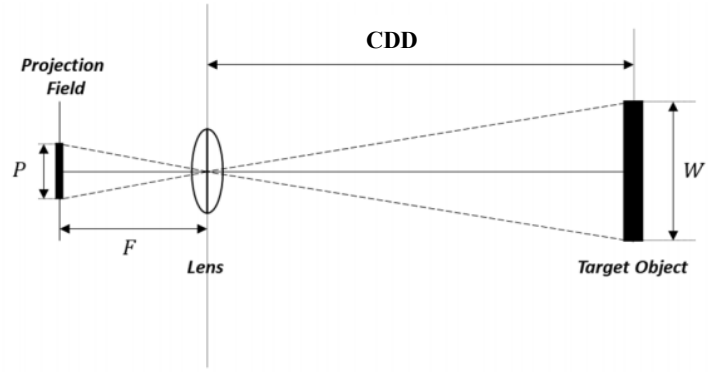


Figure 11. Triangle Similarity Theorem of a Pinhole Camera.

Based on the TST approach, a relationship between the following parameters could be derived. F is considered as the camera's focal length (in pixels), P is the object width (in pixels) from the image, CDD is the camera-to-damage distance when the camera first sees the damage (in feet), and W is the actual width of the target object (in feet). P is known by applying the object detection model to the image. Thus, these parameters created two similar triangles shown in Figure 11 for a pinhole camera, and it applies that:

$$\frac{F}{P} = \frac{CDD}{W} \tag{2}$$

Where P was acquired from the object detection module and returned as the pixel width of the detected root damage. The target object in this work was the pipe joint that has root damage grown on, hence its actual width (W) is assumed to be constant in each inspection video (same-sized pipeline). F was unknown for this practical work due to the unknown camera types thus not being able to calibrate the camera. However, with the sequential advantage of each inspection video, the focal length (F) is considered the same for consecutive frames for a single camera (no zoom-in or zoom-out within the same-sized pipeline). Thus Equation 3 derives the perceived F of the camera by applying the triangle similarity concept. When the robot first sees the damage and moves closer, F can be eliminated and the parameters that only changed are the pixel width of the object (P_{first} to P_{sub}) and camera-to-object distance (CDD_{first} to CDD_{sub}) from the frame where the camera first seen the damage to the subsequent frame when the camera continues to move closer to the damage. P_{first} and P_{sub} can be attained in images by applying the YOLOv3 detection model; While CDD_{sub}

is equal to (CDD_{first} – M) where M indicates how far the robot camera moving forward thus it can be obtained by calculating the difference of OCR readouts between the subsequent frame and the first frame when the camera first sees the damage (OCR_{sub} - OCR_{first}) (Equation 4). M is considered as a positive value because the camera continues traveling further in the sewer pipeline, thus leading to a larger distance value (OCR_{sub}) as compared to the starting distance value (OCR_{first}) when the camera first sees the damage. Accordingly, the CDD_{first} is derived with these known parameters in Equation 5, resulting in a final calculation of the longitudinal position (LP) for root damages. Therefore, Equation 6 determines the LP value by directly adding OCR_{first} readout on the CDD_{first}.

$$F = \frac{P_{first} \times CDD_{first}}{W} = \frac{P_{sub} \times CDD_{sub}}{W} = \frac{P_{sub} \times (CDD_{first} - M)}{W} \quad (3)$$

$$M = OCR_{first} - OCR_{sub} \quad (4)$$

$$CDD_{first} = \frac{P_{sub} \times M}{P_{sub} - P_{first}} \quad (5)$$

$$LP_{damage} = CDD_{first} + OCR_{first} \quad (6)$$

In this study, no ground truth distances were labeled to tell the precise longitudinal position of sewer damages in each inspection video. Thus, the OCR readout using the LF method was considered as the ground-truth value which is also the most widely used approach to localize the longitudinal position of each identified root damage by other researchers.

3.2.3. Evaluation Metrics

As for detecting text ROI and pipe joints using Yolov3 models, four metrics including confusion matrix, intersection over Union (IoU), recall, and precision as well as F score were used to evaluate the model performances. Specifically, the confusion matrix has four attributes:

- True Positive (TP) represents the number of real objects that are correctly detected as an object.

- True Negative (TN) represents the number of non-object images that are predicted as non-object images.
- False Positives (FP) represents the number of non-object images that are predicted as has-object images (Type I Error).
- False Negatives (FN) represents the number of real objects that are detected as non-object (Type II Error).

Precision, also called positive predictive value, indicates how often the model predicts correctly, while Recall indicates how well all the positives can be detected. Also, prediction accuracy was estimated as the ratio of total correct predictions to the total number of input samples. However, instead of using accuracy to evaluate model performance, the F1 score is considered as a better way since it doesn't involve True Negative samples, which is specifically useful for imbalanced datasets. All the equations are shown as follows:

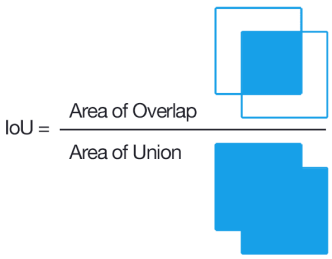
$$precision = \frac{TP}{TP+FP} \tag{7}$$

$$recall = \frac{TP}{TP+FN} \tag{8}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{9}$$

$$F = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision + Recall} \tag{10}$$

IoU (Intersection over Union) specifies the area ratio between the predicted and ground truth b-box (Equation 11). A higher IoU score indicates more precise localization of the predicted b-box, as compared to the ground truth box coordinates.



(11)

To evaluate the performance of the text recognition model, the metric *Text Recognition Performance* (*TRP*) shown in Equation 12 was applied by calculating the number of correctly recognized letters (|C|) and the number of ground truth text (|T|).

$$TRP = \frac{|C|}{|T|} \tag{12}$$

The metric *Error Rate* is used for the comparison of calculated longitudinal position and ground-truth distance (Equation 13), which is presented as the *mean absolute error* between Lc and L. This way, the results would focus on the accuracy of the distance calculation, without taking into account whether the measurement is lower or higher as compared to the ground-truth value.

$$Error\ Rate\ = |Lc - L| \tag{13}$$

Where *Lc* is the calculated longitudinal position of the root damage, and *L* is the ground-truth distance using LF method.

CHAPTER 4. EXPERIMENTAL RESULTS

Different datasets described in earlier sections were trained for different purposes which include: two YOLOv3 models with darknet-53 as the backbone for the tasks of text ROI localization and pipe joint detection; One U-net segmentation model for the damage identification; Two OCR modules in terms of Tesseract Engine and CRNN for the text recognition in video subtitles. Later, the camera-object distance (CDD) was mathematically derived using Triangle Similarity Theorem (TST) approach, thus further determining the longitudinal position of each identified root damage in the sewer pipelines. Specific evaluation metrics in the testing process include precision, recall, F1 score, and accuracy (Table 2). More details will be explained in each sub-section.

Table 2. Prediction evaluation metrics of the proposed modules.

	Model	Precision	Recall	F1 Score	Accuracy
Localization of text ROI	YOLOv3	1	0.99	0.99	99%
Segmentation of root damages	Unet	0.36	0.98	0.53	66%
Detection of pipe joints	YOLOv3	0.98	0.96	0.97	95%

For a better understanding of the proposed approaches applied on video sequences to detect and localize root damages in sewer systems, five unseen CCTV inspection videos with over 80,000+ frames were used. Taking one test video (1,2478 frames) as an example. After computing the sum of absolute differences in LUV color space between two consecutive frames, higher spikes shown in Figure 12 indicate the larger dissimilarity with respect to the nearest frames. Thus, local maxima were selected as keyframes (2270 images in total), which were approximately 20 times less than the number of original frames. Figure 13 shows the selected keyframes and it can be seen that the changes in both sewer pipes and text subtitles were depicted well. In general, these 2270 selected keyframes were prepared as an input dataset for different models at different phases, such as text ROI localization and text recognition using Tesseract Engine and CRNN-CTC network (Section 4.1), identification of root damages (Section 4.2), estimation of camera-damage distances (CDDs)

(Section 4.3) as well as the prediction of the longitudinal position for each identified root damage (Section 4.4).

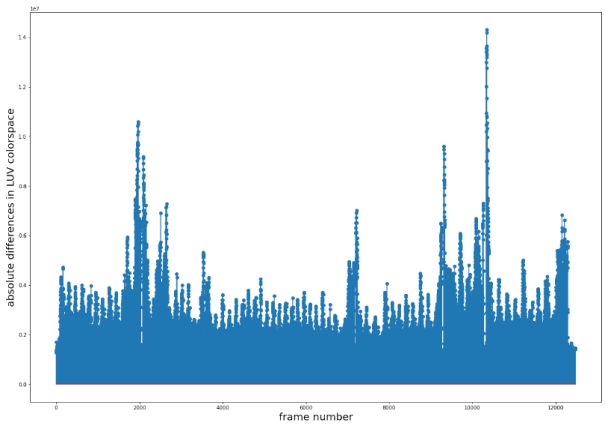


Figure 12. Sum of absolute differences in LUV colorspace between two consecutive frames in a test video.

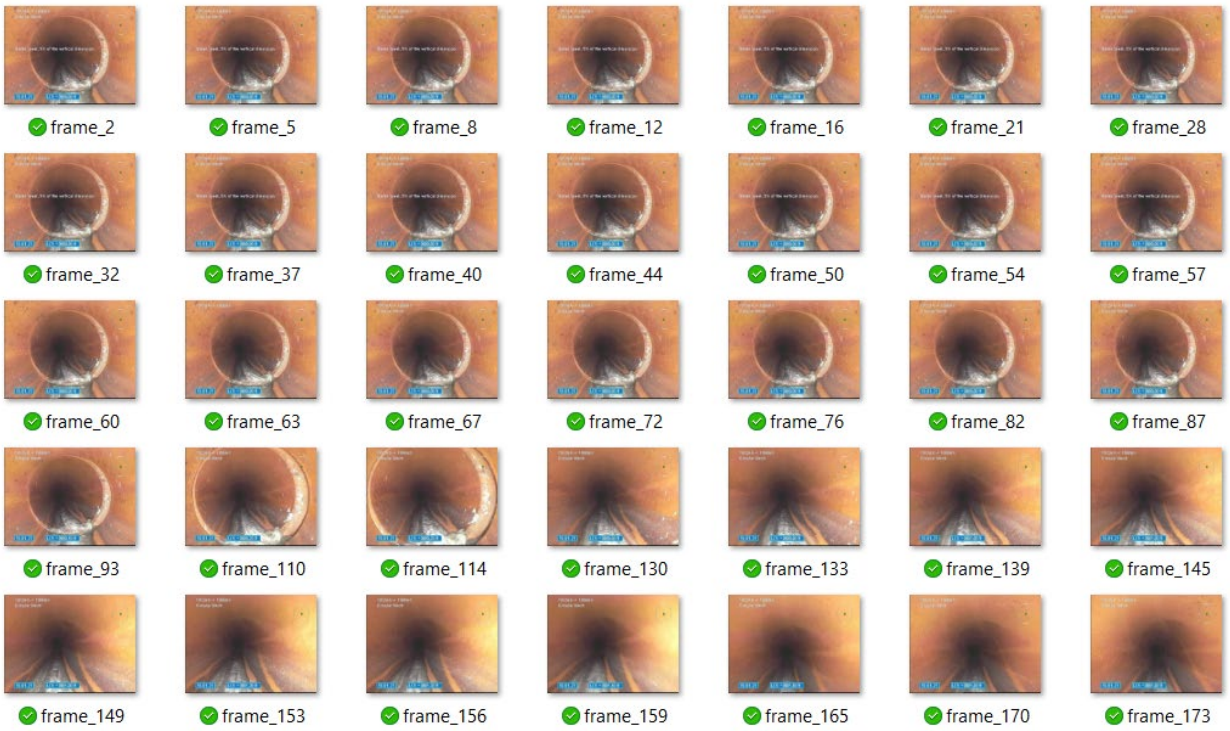


Figure 13. Example keyframes generated from the test video.

4.1 Text ROI Localization and Recognition.

After training the YOLOv3 model that was introduced earlier in Section 3.2.2.2, text ROIs in each inspection video were localized and the b-box coordinate (highlighted in green) was returned (Figure 14) with a high accuracy of 99%.

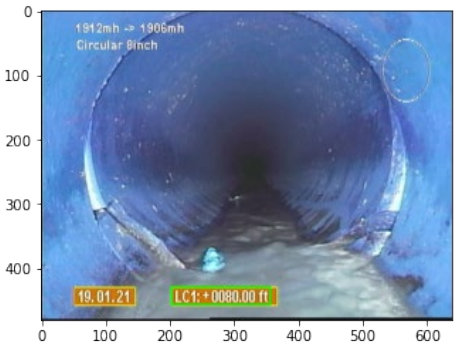


Figure 14. The localization of text ROI using the YOLOv3 model. The detected text region was highlighted in a green b-box at the bottom.

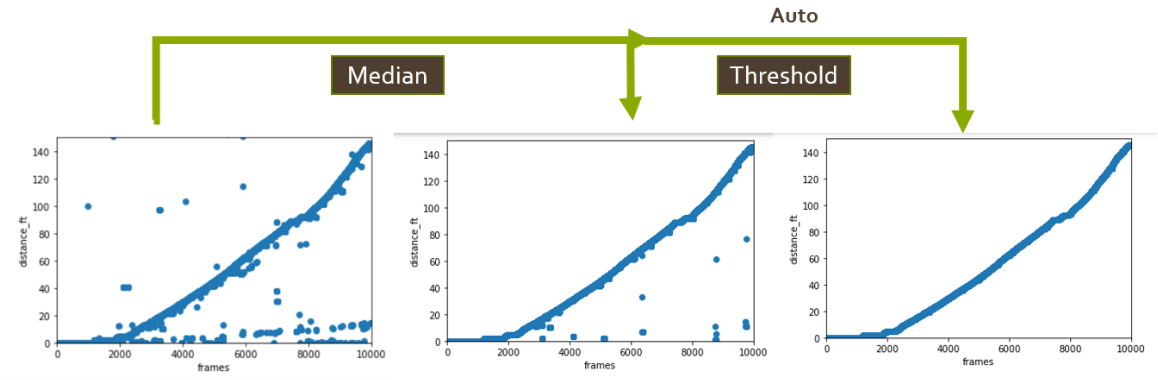
After the model implementation of text ROI localization, OCR modules (Tesseract and CRNN) were applied to further recognize text characters with the aim of extracting the traveling distance information (in feet) from each keyframe. Tesseract Engine followed by image pre-processing performed better on the text region that has a solid background (Figure 15). This approach thus can successfully convert the image to text correctly followed with an ad-hoc post-processing by adding a decimal point if there was a space between text characters. The result indicates that the application of Tesseract Engine can achieve 85% of accuracy for text recognition. Take an example shown in Figure 15, specific text values were recognized by 0127.0 ft and 59.7 ft using the Tesseract Engine.

Figure 15. Tesseract Engine recognizes text values from video subtitles that are located at the upper right corner of the keyframe.

However, the TesseractV3 engine was not supported on the cloud docker that was provided by the company, and it was also sensitive and hard to recognize the text region which had natural and complex backgrounds, or those low-resolution images with small text characters. Thus, the OCR model based on the CRNN module was alternatively applied, and this approach can obtain 81% accuracy for text recognition depending on the image quality (Figure 16). Among the prediction outcomes, negative recognitions were generated because of those characters with a complex background, thus the model was difficult to recognize the right characters and failed (Figure 16b). In order to eliminate the negative impacts of these incorrect predictions for further CDD estimations, a post-processing step was applied to smooth the text readouts, thus improving the accuracy of distance values over time in each video (Figure 17). Figure 17a shows the text readouts initially recognized from the test video using the CRNN model. Many unreasonable readouts were falling below and above the main trend line where the line represents the increasing distance values while the robot camera traveling further in the sewer pipelines. However, each distance readout can be refined by taking the median value for the nearest 15 readouts (Figure 17b), and at the last, all the extreme outliers were further removed by automatically setting up a threshold (Figure 17c). As a result, entire text readouts smoothly increased over time without any outliers, and this proposed refining approach could lead to a 10%-12% increase on the accuracy of text recognition.

(a) (b)

Figure 16. An example of (a) positive OCR predictions; b) negative OCR predictions.



(a) (b) (c)

Figure 17. Refining approach for smoothing distance readouts that are recognized from keyframe subtitles. The x-axis indicates the frame numbers, which can also be considered as traveling time. The y-axis indicates represents the distance values (in feet) by applying the CRNN module.

4.2 Root Damages Identification

A pre-trained segmentation model using U-net for detecting root damages was provided by the company. The IoU score was approximately 85% by comparing the predicted output and the ground-truth mask. A threshold value of 0.7 was applied which means the segmentation with a confidence score of 0.7 or above would be considered as the correct detection. However, due to the video limitation such as illumination variation, nature noises, and light conditions in CCTV inspection videos, this segmentation model generated lots of false positive predictions resulting in a low precision score of 0.36 with an accuracy of 66% and an F1 score of 0.53 (Table 2) after applying this model on a new test video. A low precision indicates that the model incorrectly identified non-damage pixels as root damages (Figure 18). The reason is due to the presence of similar damage patterns on the pavement surface, which might be caused by the constructional noises on the inner environments of the pipe.

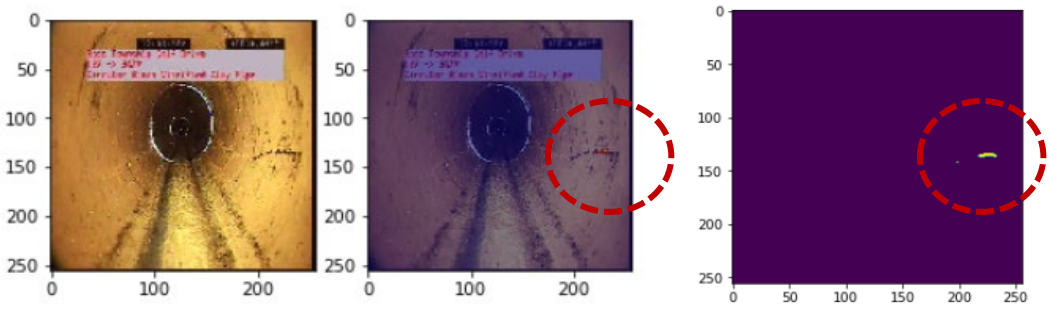


Figure 18. An example of false positive predictions by applying the segmentation model on the root damage.

Therefore, using the pipe joint as a reference aims to reduce false positives detected in the previous step, since the majority of root damages grew through the joints. Thus, the detection of pipe joints using the YOLOv3 model combined with the convex hull overlap (CHO) feature was applied to improve the accuracy of root damage identification. Table 2 shows the detection scores of recalls, precision, and F1 for detecting sewer joints were 0.98, 0.96, and 0.97, respectively, which confirms the model's capability in detecting the joints correctly (Figure 19). Also, Figure 20 shows how the CHO captures the existence and position of the root damage. As the results illustrated, the sewer damage was truly grown on the pipe joint (Figure 20e) due to a convex hull overlap (Figure 20d) between the detected pipe joint (Figure 20c) and the segmented damage (Figure 20b). Thus, such CHO feature is considered as an effective approach to significantly reduce false positive samples that generated from the segmentation model, with the aim of increasing the accuracy of damage identification in the noisy pipeline environment.

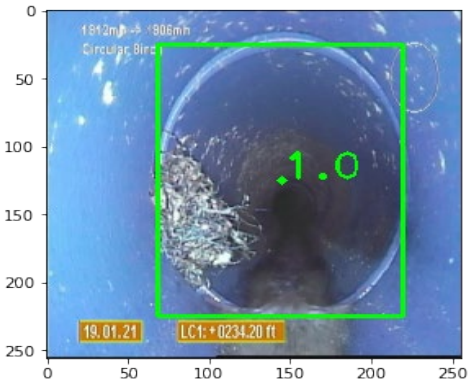
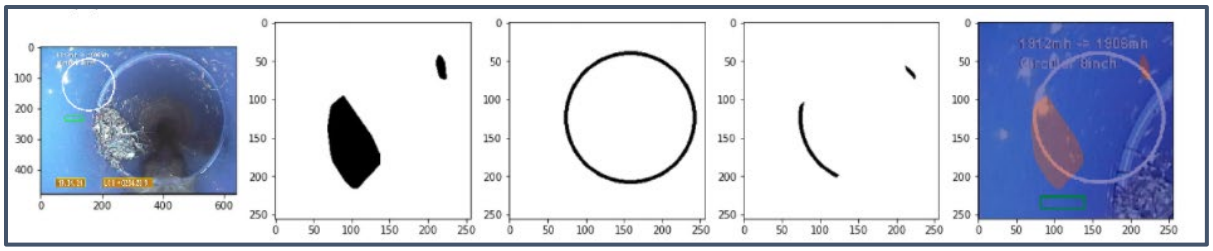


Figure 19. Detection of sewer joints using the YOLOv3 model.
(Highlighted in a green b-box with a confidence level of 100% in this example)



(a) (b) (c) (d) (e)

Figure 20. An illustration of the Convex Hull Overlap (CHO) feature for the identification of root damage. a) the original image, b) the segmented root damage, c) the segmented pipe joint segmentation, d) the CHO feature between the root damage and pipe joint, e) the actual root damage on the original image (highlighted in orange) determined by the CHO feature.

By applying the pre-trained segmentation model for identifying root damages, a detailed map was created to illustrate different stages in that root damages appeared over time in each sewer inspection video (Figure 21). Since the inspection robot is controlled by an inspector thus whenever the robot camera stops or slows down, there is a higher probability that having root damage around. Each green bar represents the existence of the root damages detected by the segmentation model, and each separated stage indicates different root damages identified in the sewer pipeline. Results in Figure 21 show that damage stages in the tested video were from ('frame474' to 'frame487'), ('frame646' to 'frame662'), ('frame1291' to 'frame1361'), ('frame1685' to 'frame1736'), ('frame1858' to 'frame1876'), and ('frame2228' to 'frame2267'). It can be concluded that the wider the damage stage was, the more keyframes it had for the time when the camera first saw the damage until past it. However, after leveraging the impact of pipe joints with the CHO feature which tends to improve the identification of root damages, results shown in Figure 22 demonstrate a more robust and effective approach to correctly detect damages by reducing false positive detections generated from the previous step. This is the reason why the total stages in Figure 22 were fewer, and each stage was narrower as compared with the stages in Figure 21. However, among these damage detections that are visualized in Figure 22, there still might have some false positive predictions that have been discussed in Section 3.2.2.3. For example, the damage stages such as ('frame646' to 'frame658'), ('frame1291' to 'frame1306'), and ('frame2228' to 'frame2265') were still considered as false positive predictions even with the CHO feature (Figure 23). Therefore, re-train and improving the segmentation model for better damage detection is crucial and considered as a priority in further work.

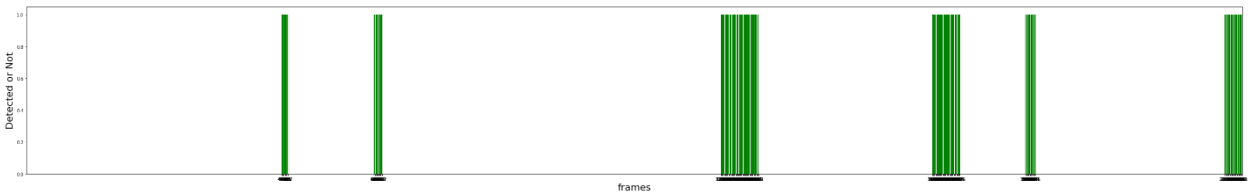


Figure 21. Damage stages determined by the segmentation model. The x-axis indicates frame numbers, and the y-axis indicates whether the root damage was detected or not.

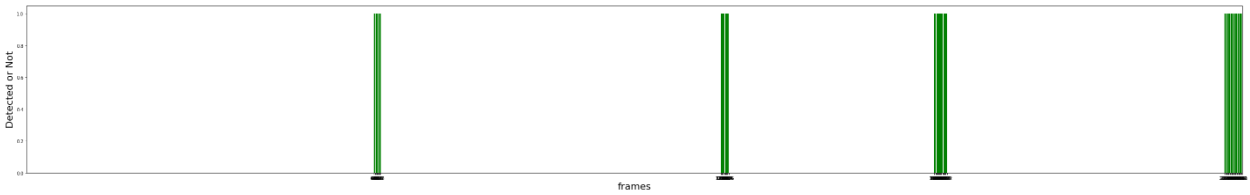


Figure 22. Damage stages determined by the CHO feature. The x-axis indicates frame numbers, and the y-axis indicates whether the root damage was detected or not.

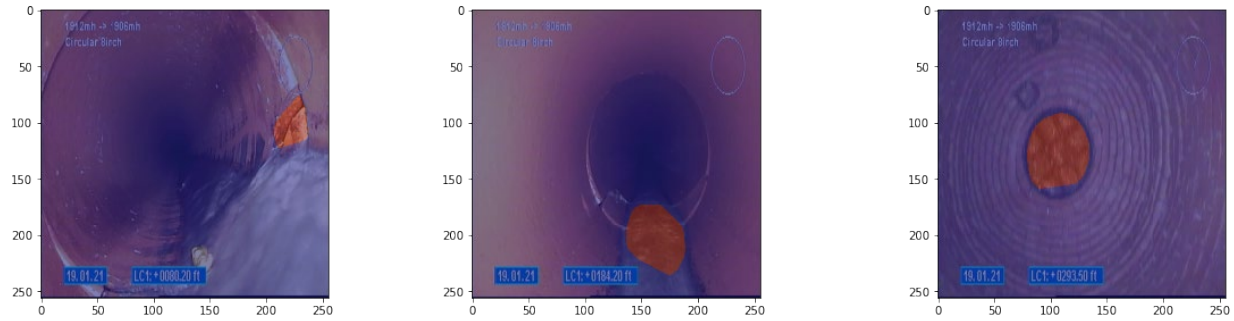


Figure 23. False positive detections of root damages with CHO feature.

Taking another test video as a representative explanation (Figure 24), if there were multiple root damages exist continuously in a part of the sewer pipeline, the segmentation model might constantly detect the pixel segmentations for such damages thus finally returning the longitudinal position of the last damage in such sub-pipeline using the LF approach (Figure 25a, highlighted in a yellow box). However, Figure 26b demonstrated that the benefits of our proposed CHO feature not only helped to reduce the incorrect damage detections, but also separated the entire stage of

the continuous damages into four different sub-stages (Figure 25b, highlighted in a yellow box) between keyframe1146 to keyframe1318.

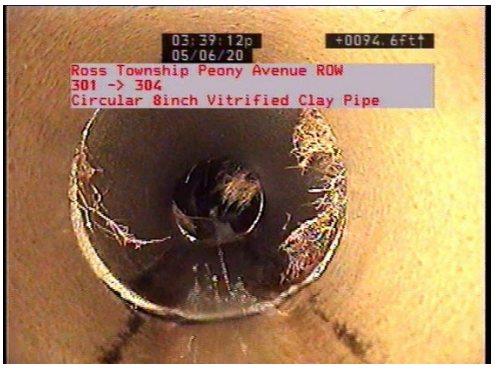
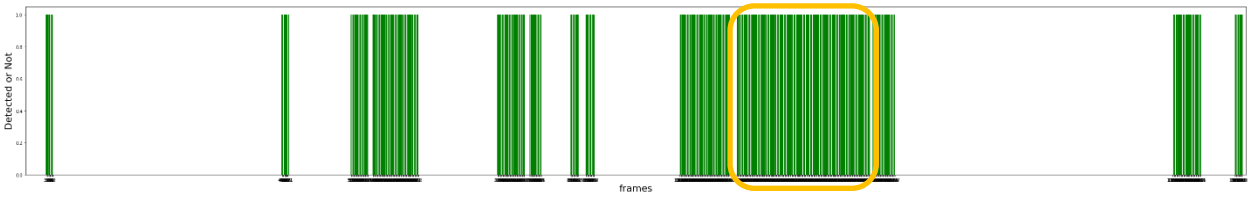
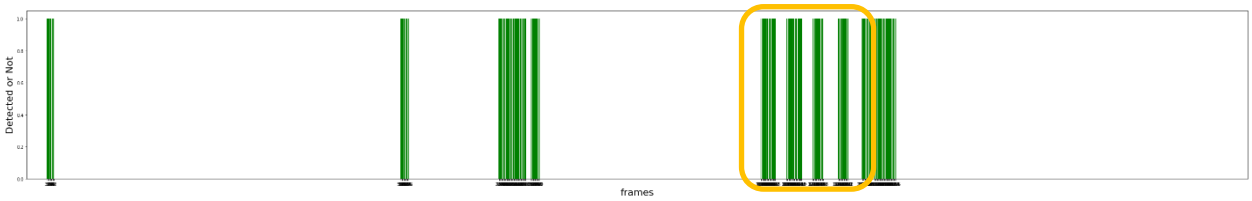


Figure 24. Multiple root damages exist continuously in the sub-pipeline of the sewer system.



(a)



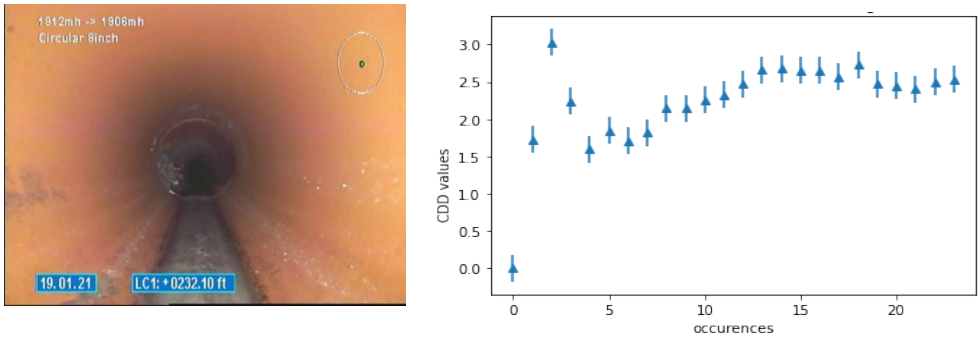
(b)

Figure 25. (Another test video) Damage stages determined by the segmentation model (a) and CHO features (b). The x-axis indicates frame numbers, and the y-axis indicates whether the root damage was detected or not.

4.3 Camera-damage Distances (CDD) Estimation.

When every time the root damage was first seen by the camera, the CDD value was then calculated based on the TST approach that uses the current frame and all the subsequence frames till the camera passed by the point of the damage condition. In theory, all the CDDs in each stage are supposed to be the same, since each CDD calculation in the stage represents the camera-to-damage distance when the camera first sees the damage. To remove outliers and reduce noises, the median

CDD value was selected as the actual camera-to-damage distance (in feet). An example shown in Figure 26b indicates that the CDD values were sensitive and thus jumped around at the beginning of the keyframe sequences when the camera starts to target the root damage. The reason was probably because the damage/joint detection was not accurate enough at the beginning of frame sequences, resulting in enormous variance or non-significant differences between P_{first} and P_{sub} that were used in Equation 5. This result might reveal that this proposed TST approach has increased error when the damage is further away to estimate the CDD value. However, CDDs later tends to be stable around 2.42 inch which indicates that the distance of the root damage in the image was about 2.42 feet (Figure 26a). For a further validation of the CDD estimations using the TST approach, Figure 26 demonstrates that the damage distance to the camera in the image was 0.9 ft (Figure 27a).



(a) (b)

Figure 26. (a) The sewer scene when the camera first detects the root damage; (b) Estimation of camera-damage distance (CDD) using the TST approach when the camera first detects the root damag. The x axis indicates the occurences of calculting the CDD in each damage stage, while the y axis indicates represents the CDD values in feet.

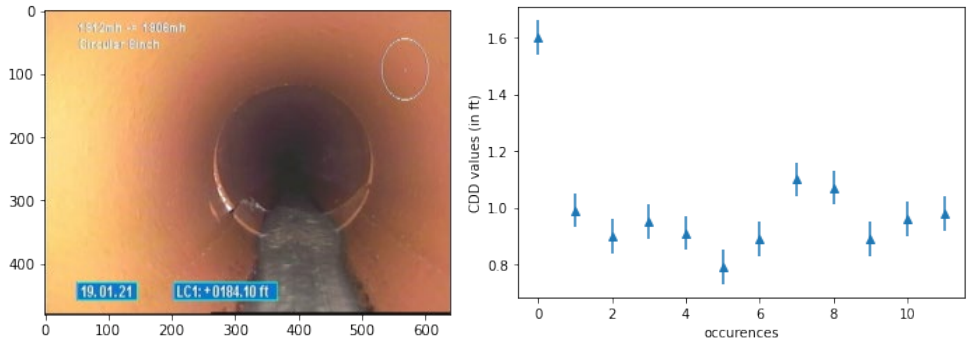


Figure 27. CDD values calculated from another damage stage using the TST approach.

4.4 Longitudinal Position (LP) for Each Identified Root Damage

After obtaining the camera-object distance (CDD) for each root damage in Section 4.3, the longitudinal position of the root damage from the manhole or point of sewer entry can be straightforwardly calculated by Equation 6. Take the test video as an example, Figure 28 indicates the comparison between the calculated LP and ground-truth value in each damage stage. There were 10 damage stages originally identified by using the segmentation model but reduced to 4 stages (highlighted in the red box in Figure 28) after leveraging the advantages of the pipe joint detection and the CHO feature that were described earlier in Section 3.2.2.4. These four damage stages are considered to have actual root damages, and the results indicate that our proposed approach has achieved highly accurate results with 0 to 0.75 ft of error rates. Even for other stages that have false positive damages, most of the error rates to estimate LPs were still quite low which were between 0.13 to 0.87 ft. In another word, the calculated longitudinal position of each identified root damage has a small error (less than 1 foot) as compared to the value using the LF method. With these observations, such errors (less than 1 foot) are considered acceptable, especially for those sewer pipelines that were longer than 200 ft. In Figure 28, it can be noticed that there was one exception at the damage stage starting from keyframe 2044, showing an error rate of 2.73 ft between the calculated and ground-truth LP values. The reason might be that in such a stage, the segmentation model incorrectly detected a non-damage object as root damage (Figure 29a). Another reason might be due to the insufficient occurrences (only four) applied to the CDD calculations (Figure 29b), resulting in a large error rate to predict the longitudinal position of the root damage. Commonly, about 10 keyframes were at least required to ensure sufficient occurrences for the CDD estimation, like the stages in Figure 26 and Figure 27. Therefore,

increasing the accuracy of the segmentation model along with increased occurrences involved would be significant to improve the performance on the estimation of longitudinal positions for each sewer damage.

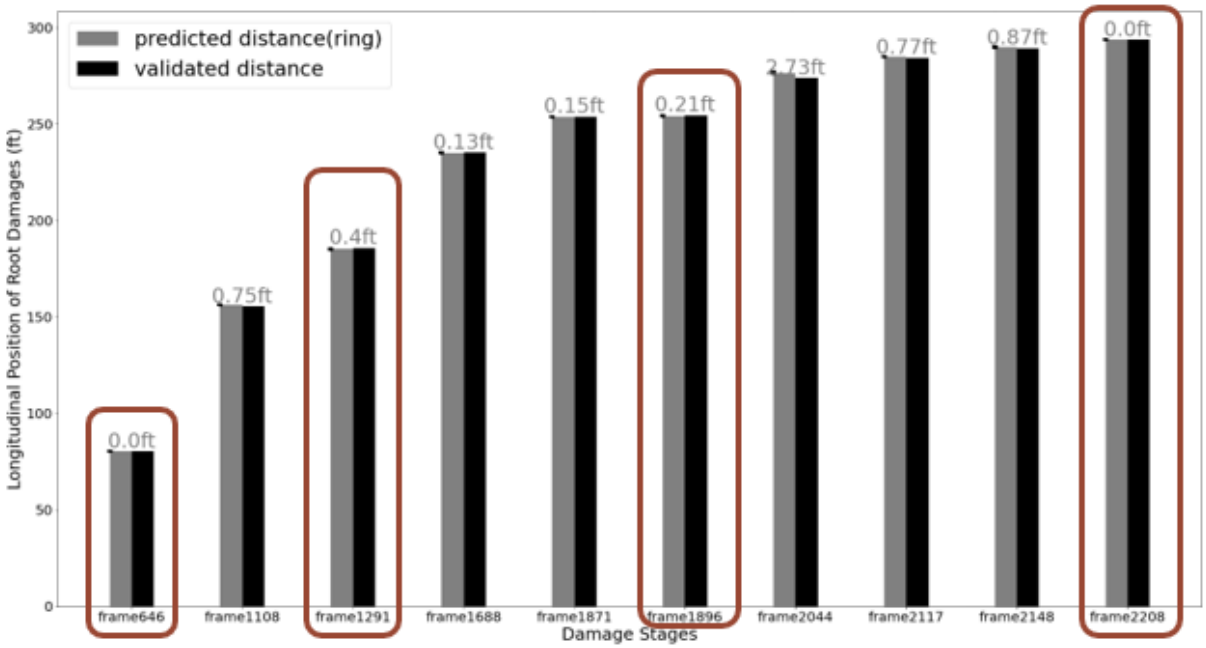


Figure 28. LP Comparisons between our proposed method and the LF approach among different damage stages in the test video. The x-axis indicates damage stages showing with the frame number when the camera first saw the root damage; the y-axis indicates the longitudinal position (in feet) of each identified root damage in the underground sewer pipeline.

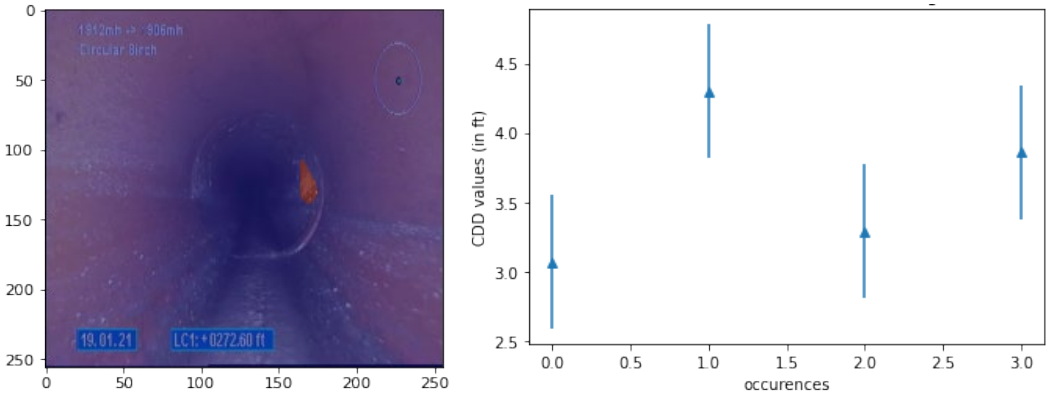


Figure 29. A false positive prediction of root damages leads to a large error in CDD estimations. (a) Root damage (highlighted in an orange region) detected by the segmentation model in the keyframe 2044; (b) CDD values calculated in such damage stage that starting with keyframe 2044 of the test video with 4 subsequent frames (occurrences).

There were also another four testing videos applied to validate LP predictions, which the performance using our proposed approach was great for most root damages as the lowest error rate to the ground truth down to 0.1 ft, and up to 1.03 ft. Even for the situation with continuous root damages indicated in Figure 24, results show that each root damage has its longitudinal position in the sub-pipeline which had a high prediction accuracy of 90% after a manual validation on the test video. With few exceptions of large error rates, the conclusions imply that the TST approach with the benefits of DL-based models might not be suitable for inspection videos that have bad image quality. Since these images had too many noises or naturally complex backgrounds, it might lead to inaccurate object detection and text recognition in the images.

Overall, this study achieved the following conclusions:

(1). Implemented Optical Character Recognition (OCR) model to extract text from different inspection videos for identification of monocular camera moving distance using Tesseract Engine as well as Convolutional and Recurrent Neural Networks (CRNN), with the accuracy of 90% on the task of text recognition after applying the refining approach proposed in this study.

(2). Designed a semi-automatic annotation method for keyframes of inspection videos based on Hough Transform techniques which significantly increased the annotation efficiency by 70%.

(3). Built well-tuned YoloV3 models with Darknet-53 network for pipe damage recognition leveraging the Convex Hull Overlap (CHO) feature that led to a 20% increase on the accuracy of damage identification.

(4). Successfully designed an end-to-end DL pipeline that involved Triangle Similarity Theorem (TST) to predict damage locations which achieved error rates that were less than 1 feet.

CHAPTER 5. CONCLUSION AND FUTURE WORK

The visual inspections that rely on inspectors are time-consuming and labor-intensive, which may cause further deconstruction of the underground sewer systems. Among the possible techniques for the inspection of sewer defects, this study attempts to automatically detect root damages and identify their circumferential and longitudinal positions in sewer networks. This framework can eliminate the time-consuming tasks for sewer inspection and improves the accuracy of identifying the damaged locations, thus providing guidance in support of any further maintenance of the underground sewer networks. By applying CCTV videos from different resources, this study brings innovative approaches to perform complex feature extraction and pattern recognition to overcome the video limitations such as the variation of illumination, subtitle format, lighting conditions, and so on. Also, by taking advantage of the sequential nature of video frames, promising vision-based approaches such as YOLOv3 and OCR models along with the Theorem of Triangle Similarity (TST) were applied to detect and localize the root damages in the videos. An end-to-end DL pipeline was developed to evaluate these inspection techniques and some video examples were presented to demonstrate model effectiveness and efficiency, as well as the limitation of the leading methodologies.

Text recognition integrates feature extraction and text sequence modeling to read the traveling distance which provides important information to the inspector about the relative location of the controlled robot away from the entrance of the manhole. After fully comparing the text recognition modules of TesseractV3 and CRNN and considering the performances of their prediction accuracy and cloud deployment on the recognition task, CRNN was chosen as the preliminary OCR model in this study. Cross-validation results indicated that after refining, both Tesseract and CRNN modules can considerably increase the recognition capability and achieve 90% of accuracy for recognizing distance values. Thus, refining operation is considered a significant improvement to map the detailed traveling map of the robot within the sewer system from each CCTV inspection video. However, even though the TesseractV3 module was easier with pre-trained weights, this version was difficult to be deployed on GCP. Thus, with the extra efforts to manually annotate the subtitle texts for the CRNN model, the result indicated that the application of the YOLOv3 detection model on text ROI localization combined with CRNN still achieved high accuracy on

the task of text recognition. Overall, CRNN can take input images and generate good predictions with variable-length texts., and it also removes the fully connected layers that are used in CNNs thus becoming more compact and efficient. To further increase the accuracy of text recognition using CRNN, more training images should also be added which is part of our ongoing work.

Also, target objects such as the root damage should have less variation in pixel segmentation in image sequences which may have a better result for the CDD calculations. Thus, with the aim of accuracy improvement, the pipe joint that has root damage grown through was considered as a better reference for CDD calculations due to its constant circular dimension. Moreover, this study aims to integrate the convex hull overlap (CHO) feature and the application of CNNs such as the YOLOv3 module on pipe joints to improve the performance of the current vision-based segmentation approach for better identifying root damages. The convexity feature is an optimal representation of the sewer damage, thus effectively reflecting the actual position of the damage by inferring the spatial relationships between detected pipe joints and segmented damage in the image. After increasing the accuracy of damage identification by reducing the false positive predictions, the assumption of a pinhole camera model with a TST approach can be applied to calculate the camera-damage distance (CDD) value when the root damage was first seen by the camera, thus further estimate the longitudinal position of each identified damage in the sewer system. However, the mathematical derivation of the TST approach for the CDD calculation might be sensitive to image quality and the lack of enough subsequent frames that described earlier in Section 4.3.

Therefore, in different underground sewer systems, due to illumination variation and complex pipe features/patterns, it's tough to define a template for detecting root damages and predicting its circumferential and longitudinal positions in the sewer pipes. Sewer damage patterns might be reshaped also because of the sudden movements, camera pose change, and poor lighting conditions. Hence, the generalization capability of the proposed DL-based framework needs to be improved for different environments and pipe materials. Accordingly, creating more comprehensive training samples along with more accurate models would be helpful for the generalization capability. Although the training process in a DL-based framework is computationally expensive, it would be beneficial from various data augmentation techniques and increasingly rapid computational

powder of graphic processing units (GPUs) (Moradi, S., 2020). In addition, the assessment of damage severity deserves more attention. Some studies achieved great progress to quantify the extent of sewer damages based on pixel intensities or saliency models, but the accomplishment is still not enough to generate an automated model in severity evaluation. Another limitation of this study is to not consider the correction of image radial distortion, which could be a significant component to predict the longitudinal position for sewer damages. More research needs to be conveyed on this part for developing a more robust DL-based assessment system.

Overall, this project aims to build a DL pipeline combined with various compute vision techniques, as well as derive a mathematical function to identify the longitudinal position of root damages in sewer pipelines. Results have proved the effectiveness of the pipeline, but more work needs to be done to improve the robustness of entire models, such as text recognition accuracy, segmentation predictions, as well as pipe joint detections. Also, since the application of such a DL-based framework for CCTV inspection is heavily dependent upon the visual data, the image quality needs to be standardized for optimal detection of sewer damages, and the camera angle and position within the sewer pipelines are also significant to predict the longitudinal position of the root damage in the pipeline. To minimize the illumination fluctuation of the inspection videos, lighting conditions should be properly adjusted in the pipe. More future work will be devoted to developing algorithms and computational hardware for massive calculations. Considering all the promising perspectives, automated assessment of sewer damages and positions will be more robust and efficient resulting in a substantial decrease in inspection effort. Then such a DL pipeline will be considered as a guideline to accurately assess and locate root damages for further infrastructure maintenance.

REFERENCES

Abhinaya, A. (2021). Using Machine Learning to detect voids in an underground pipeline using in-pipe Ground Penetrating Radar (Master's thesis, University of Twente).

Ahrary, A., Tian, L., Kamata, S. I., & Ishikawa, M. (2005, November). An autonomous sewer robot navigation based on stereo camera information. In 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05) (pp. 6-pp). IEEE.

Amanpreet Walia, 2018. KeyFramesExtraction. Github Link https://github.com/amanwalia123/KeyFramesExtraction

A_K_Nain, 2020. OCR model for reading Captchas. Github Link. https://github.com/keras-team/keras-io/blob/master/examples/vision/captcha_ocr.py

Atha, D. J., and M. R. Jahanshahi. 2017. "Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection." Struct. Health Monit. 17 (5): 1110–1128. https://doi.org/10.1177/1475921717737051.

Burger, W., & Burge, M. J. (2010). Principles of digital image processing: core algorithms. Springer Science & Business Media.

Chen, L., & Li, S. (2018, November). Improvement research and application of text recognition algorithm based on CRNN. In Proceedings of the 2018 International Conference on Signal Processing and Machine Learning (pp. 166-170).

Chen, F. C., and M. R. Jahanshahi. 2018. "NB-CNN: Deep learning–based crack detection using convolutional neural network and naive Bayes data fusion." IEEE Trans. Ind. Electron. 65 (5): 4392–4400.

Cheng, J. C., & Wang, M. (2018). Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques. Automation in Construction, 95, 155-171.

Christian Wolf, Jean-Michel Jolion, Object count/area graphs for the evaluation of object detection and segmentation algorithms, IJDAR 8 (4) (2006) 280–296.

Dang, L. M., Hassan, S. I., Im, S., Mehmood, I., & Moon, H. (2018). Utilizing text recognition for the defects extraction in sewers CCTV inspection videos. Computers in Industry, 99, 96-109.

Dirksen, J., Clemens, F. H. L. R., Korving, H., Cherqui, F., Le Gauffre, P., Ertl, T., Plihal,H., Müller, K., and Snaterse, C. T. M. (2013). "The consistency of visual sewer inspection data." Structure and Infrastructure Engineering, 9(3), 214–228.

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. IEEE transactions on pattern analysis and machine intelligence, 36(8), 1532-1545.

Draelos, M., Qiu, Q., Bronstein, A., & Sapiro, G. (2015, September). Intel realsense= real low cost gaze. In 2015 IEEE International Conference on Image Processing (ICIP) (pp. 2520-2524). IEEE.

EPA. 2014. "Quick guide for estimating infiltration and inflow." Accessed July 15, 2018. https://www3.epa.gov/region1/sso/pdfs/QuickGuide4 EstimatingInfiltrationInflow.pdf.

Fang, X., Guo, W., Li, Q., Zhu, J., Chen, Z., Yu, J., ... & Yang, H. (2020). Sewer pipeline fault identification using anomaly detection algorithms on video sequences. IEEE Access, 8, 39574-39586.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

Gordo, A. (2015). Supervised mid-level features for word image representation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2956-2964).

Guo, W., Soibelman, L., & Garrett Jr, J. H. (2009). Automated defect detection for sewer pipeline inspection and condition assessment. Automation in Construction, 18(5), 587-596.

Halfawy, M. R., & Hengmeechai, J. (2014). Efficient algorithm for crack detection in sewer images from closed-circuit television inspections. Journal of Infrastructure Systems, 20(2), 04013014.

Halfawy, M. R., and J. Hengmeechai. 2014c. "Optical flow techniques for estimation of camera motion parameters in sewer closed circuit television inspection videos." Autom. Constr. 38 (Mar): 39–45.

Hassan, S. I., Dang, L. M., Im, S. H., Min, K. B., Nam, J. Y., & Moon, H. J. (2018). Damage Detection and Classification System for Sewer Inspection using Convolutional Neural Networks based on Deep Learning. Journal of the Korea Institute of Information and Communication Engineering, 22(3), 451-457.

Hassan, S. I., Dang, L. M., Mehmood, I., Im, S., Choi, C., Kang, J., ... & Moon, H. (2019). Underground sewer pipe condition assessment based on convolutional neural networks. Automation in Construction, 106, 102849.

Hawari, A., Alamin, M., Alkadour, F., Elmasry, M., & Zayed, T. (2018). Automated defect detection tool for closed circuit television (cctv) inspected sewer pipelines. Automation in Construction, 89, 99-109.

Jack C.P. Cheng, Mingzhu Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171

Jahanshahi, M. R. 2011. "Vision-based studies for structural health monitoring and condition assessment." Ph.D. dissertation, Faculty of the USC Graduate School, Univ. of Southern California.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Kumar, S. S., Wang, M., Abraham, D. M., Jahanshahi, M. R., Iseley, T., & Cheng, J. C. (2020). Deep learning–based automated detection of sewer defects in CCTV videos. Journal of Computing in Civil Engineering, 34(1), 04019047.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning." Nature, 521, 436.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Lee, J., Bang, J., & Yang, S. I. (2017, October). Object detection with sliding window in images including multiple similar objects. In 2017 international conference on information and communication technology convergence (ICTC) (pp. 803-806). IEEE.

Li, D., Cong, A., & Guo, S. (2019). Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification. Automation in Construction, 101, 199-208.

Liang, Y. (2018, December). Salient Object Detection with Convex Hull Overlap. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4605-4612). IEEE.

Martin Thoma, Analysis and optimization of convolutional neural network architectures, arXiv preprint arXiv:1707.09725, 2017.

Megalingam, R. K., Shriram, V., Likhith, B., Rajesh, G., & Ghanta, S. (2016, January). Monocular distance estimation using pinhole camera approximation to avoid vehicle crash and back-over accidents. In 2016 10th international conference on intelligent systems and control (ISCO) (pp. 1-5). IEEE.

Meijer, D., Scholten, L., Clemens, F., and Knobbe, A. (2019). "A defect classification methodology for sewer image sets with convolutional neural networks." Automation in Construction, 104, 281–298

Moradi, S., Zayed, T., & Golkhoo, F. (2019). Review on computer aided sewer pipeline defect detection and condition assessment. Infrastructures, 4(1), 10.

Moradi, S. (2020). Defect Detection and Classification in Sewer Pipeline Inspection Videos Using Deep Neural Networks (Doctoral dissertation, Concordia University).

Moselhi, Osama, and Tariq Shehab-Eldeen. "Classification of defects in sewer pipes using neural networks." J. Infrastruct. Syst. 6, no. 3 (2000): 97–104.

Murmu, N.; Nandi, D., "Low cost distance estimation system using low resolution single camera and high radius convex mirrors", 2014 International Conference on Advances in Computing Communications and Informatics (ICACCI), Pages: 998 - 1003, DOI: 10.1109/ICACCI.2014.6968509

Myrans, J., Kapelan, Z., & Everson, R. (2018, July). Using Automatic Anomaly Detection to Identify Faults in Sewers:(027). In WDSA/CCWI Joint Conference Proceedings (Vol. 1).

Murmu, N.; Nandi, D., "Low cost distance estimation system using low resolution single camera and high radius convex mirrors", 2014. International Conference on Advances in Computing Communications and Informatics (ICACCI), Pages: 998 - 1003, DOI: 10.1109/ICACCI.2014.6968509

NASSCO (North American Society of Sewer Service Companies). 2018. “Pipeline assessment and certification program.” Accessed August 1, 2018. https://www.nassco.org/pipeline-assessment-and-certification-program.

Nienaber, S., Kroon, R. S., & Booysen, M. J. (2015, December). A comparison of low-cost monocular vision techniques for pothole distance estimation. In 2015 IEEE Symposium Series on Computational Intelligence (pp. 419-426). IEEE.

Pan, G., Zheng, Y., Guo, S., & Lv, Y. (2020). Automatic sewer pipe defect semantic segmentation based on improved U-Net. Automation in Construction, 119, 103383.

Ptucha, R., Such, F. P., Pillai, S., Brockler, F., Singh, V., & Hutkowski, P. (2019). Intelligent character recognition using fully convolutional neural networks. Pattern recognition, 88, 604-613.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. “Faster R-CNN: Towards real-time object detection with region proposal networks.” In Proc., Advances in Neural Information Processing
Systems, 91–99. Montréal: Neural Information Processing Systems Foundation.

Rahimzadeganasl, A., & Sertel, E. (2017, May). Automatic building detection based on CIE LUV color space using very high resolution pleiades images. In 2017 25th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). “You Only Look Once: Unified, Real-Time Object Detection.” 280–292.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Rodriguez-Serrano, J. A., Gordo, A., & Perronnin, F. (2015). Label embedding: A frugal baseline for text recognition. International Journal of Computer Vision, 113(3), 193-207.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Schanda, J. (Ed.). (2007). Colorimetry: understanding the CIE system. John Wiley & Sons.

Selvakumar, A., Tuccillo, M. E., Martel, K. D., Matthews, J. C., & Feeney, C. (2014). Demonstration and evaluation of state-of-the-art wastewater collection systems condition assessment technologies. *Journal of Pipeline Systems Engineering and Practice*, *5*(2), 04013018.

Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 39(11), 2298-2304.

Sudowe, P., & Leibe, B. (2011, September). Efficient use of geometric constraints for sliding-window object detection in video. In International Conference on Computer Vision Systems (pp. 11-20). Springer, Berlin, Heidelberg.

Tan N. Nguyen, Chien H. Thai, H. Nguyen-Xuan, Jaehong Lee, Geometrically nonlinear analysis of functionally graded material plates using an improved moving kriging meshfree method based on a refined plate theory, Compos. Struct. 193 (2018) 268–280.

Toledo Ferraz, C., Barcellos, W., Pereira Junior, O., Trevisan Negri Borges, T., Garcia Manzato, M., Gonzaga, A., & Hiroki Saito, J. (2021). A comparison among keyframe extraction techniques for CNN classification based on video periocular images. Multimedia Tools and Applications, 80(8), 12843-12856.

Yang, M. D., & Su, T. C. (2009). Segmenting ideal morphologies of sewer pipe defects on CCTV images for automated diagnosis. Expert Systems with Applications, 36(2), 3562-3573.

Yao, C., Bai, X., Shi, B., & Liu, W. (2014). Strokelets: A learned multi-scale representation for scene text recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4042-4049).

Zhang, C., Chang, C. C., & Jamshidi, M. (2020). Concrete bridge surface damage detection using a single-stage detector. Computer-Aided Civil and Infrastructure Engineering, 35(4), 389-409.