Bridging the Annotation Gap: Innovating Sewer Defects Detection with Weakly Supervised Object Localization

Jianyu Yin[1], Xianfei Yin[2], Yifeng Sun[3], and Mi Pan[4]

[134]Department of Civil and Environmental Engineering, University of Macau, Macau, China
[2]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China
dc12714@um.edu.mo, xianfei@ualberta.ca, mc35239@um.edu.mo, mipan@um.edu.mo

Abstract

Urban sewer systems are vital yet often neglected components of modern infrastructure system. Inspecting these systems is expensive due to labour costs and the need for manual examination by professionals. In addition to the challenges posed by traditional methods, developing deep-learning-based automatic defect detection models requires a vast number of bounding box labels, which are challenging to acquire. To address these gaps, our study introduced the application of Weakly Supervised Object Localization (WSOL) for automated defect localization in sewer pipes. WSOL is a technique that allows for the localization of objects within images using only image-level labels, without the need for precise bounding box annotations. We adopted a state-of-the-art WSOL method that mitigates feature directions with class-specific weights misalignment, enabling more accurate and complete localization of defects. By generating heatmaps from Sewer-ML's image-level annotations, bounding box labels are eliminated, rendering our approach scalable and cost-effective. The proposed WSOL-based approach was validated through five distinct classes of defects and one construction feature, demonstrating the promising localization performance. Our method achieved mean MaxBoxAccV2 scores of 64.33% and 56.89% when using ResNet-50 and VGG-16 backbones, respectively, while also attained classification accuracies of 87.00% for ResNet50 backbone and 83.00% for VGG16 backbone. As a pioneering contribution, our work established a new standard for automated sewer system maintenance, offered a benchmark for the application of WSOL methods using solely image-level annotations in defect localization for urban sewer systems, and further expanded the frontier of weakly supervised learning in critical infrastructural applications.

Keywords –

Weakly Supervised Object Localization (WSOL); Sewer Pipelines; Defect Detection; Image-Level Annotation

1 Introduction

Urban sewer system are integral to modern civilization across various urban settings, playing a pivotal role in health, sanitation, and overall well-being. Yet, the inspection and upkeep of these intricate underground networks often go overlooked. Traditional inspection methods are labour-intensive, costly, and hinge on manual evaluations by specialists. While these methods are effective, they aren't scalable for expansive urban sewers. The push towards deep-learning for defect detection introduces another challenge: the need for numerous bounding box labels. Acquiring such labels is not only tedious but also consumes significant resources. Therefore, an effective automatic localization system is urgently needed to overcome these shortcomings.

The core research question this study seeks to address is: "How can feature-level defect detection be realized with image-level annotations in the context of urban sewer system inspection?" The advent of Weakly Supervised Object Localization (WSOL) offers a promising alternative, paving the way for more automated, efficient, and cost-effective solutions as it relies solely on image-level annotations. Yet, while WSOL has seen success in other domains [1-5], its application in the realm of sewer system inspection is still quite new. To the best of our knowledge, this is the first work to apply weakly supervised learning for automatic defect detection in the sewer pipelines with evaluation of localization precision. It's essential to note that existing WSOL methods, which typically utilize Class Activation Maps (CAMs), have their limitations in accurately localizing defects, as they tend to focus on only the most prominent features [1]. This poses a concern in the context of sewer systems where the full shape and scope of defects are crucial for remedial actions. In this study, we adopted a state-of-the-art WSOL method proposed by Kim et al. [2] that mitigates feature directions with class-specific weights misalignment. Our hypothesis is that the proposed WSOL method can achieve reasonable localization accuracy for sewer defects using only image-level labels, ensuring a more holistic and accurate localization of sewer defects.

The main contributions of this study are listed below:

1) Applied weakly supervised learning for automated defect localization in urban sewer systems with evaluation of localization precision;
2) Provided a scalable solution that obviates the need for tedious and resource-intensive bounding box labels with reasonable localization accuracy;
3) Developed A benchmark and methodological framework for future research into weakly supervised defect localization in sewer pipelines.

The rest of the paper is structured as follows: Section 2 delves into the recent advances and the unique challenges of both sewer inspection techniques and defect detection. In Section 3, we detail our methodology, including the dataset used and our experiment set up. Section 4 offers a comprehensive evaluation, presenting our experimental results. Finally, Section 5 wraps up with the study's limitations, prospects for future research, and a concluding remark.

2 Related Work

In recent studies on sewer inspections, deep learning methods have been increasingly utilized to improve performance of defect detection [6]. For example, Cheng and Wang [7] contributed an automated defect detection approach using faster R-CNN, demonstrating how deep CNNs can identify and locate common defects in CCTV images. Li et al. [8] introduced innovations like strengthened region proposals and global context fusion to enable fine-grained defect severity grading, surpassing other methods in this capability. Yin et al. [9] applied state-of-the-art YOLOv3 for real-time automated detection in videos, showing advantages in processing speed and accuracy over manual review. And Yin et al. [10] further developed a CCTV video interpretation algorithm and sewer pipe video assessment (SPVA) system based on their previous developed deep learning-based framework.

However, the major defect inspection models rely on effective supervised learning methods that cost much time in the manual annotation process for training [11]. Therefore, researchers have begun to explore alternative methods to reduce the need for manual annotation in defect inspection models. For example, Zhang et al. [12] used a GAN framework to leverage both labelled and unlabelled images, achieving 79-81% mean IOU on a steel defect dataset with only 1/8 to 1/4 full supervision. Zhang et al. [4] used weak image-level tags rather than detailed pixel-level annotations, extracting spatial information from tags through category-aware convolutions and pooling. In another study, Wu et al. [5] relied solely on image-level labels for training, improving CAM techniques to achieve performance surpassing some fully supervised methods. Manual labelling of feature-level data is both time-consuming and prone to human error. Thereby, our study pioneered the application of WSOL for accurately localizing sewer defects with a methodological framework used for fair comparative evaluation. Sewer pipelines inspection presents unique challenges such as varying lighting conditions, occlusions, and highly irregular object shapes. The adopted method [2] seeks to overcome the limitations of current WSOL methods by mitigating feature directions with class-specific weights misalignment, thereby ensuring more accurate defect detection in sewer pipelines. Furthermore, the completely automated systems that include automatic labelling tools need to be developed for more efficient sewer inspections [6]. The automatic generated bounding boxes from heatmaps can be served as inputs for training a supervised detection model to reduce the redundant annotation process.

3 Methodology

3.1 Dataset

The dataset utilized in this research is Sewer-ML [13], which is the first publicly available dataset dedicated to sewer defect classification. Comprising 1.3 million images sourced from 75,618 videos, the data was aggregated from three Danish water utility companies over a nine-year period. The image-level annotations were performed by licensed sewer inspectors in compliance with the Danish sewer inspection standard, Fotomanualen [14]. This contributes to the high reliability and consistency of the annotations.

For the task of WSOL, it is essential to have images with only one class of object for generating accurate heatmaps. As such, we curated a subset of Sewer-ML specifically designed for WSOL tasks. The customized dataset information is provided in table 1. The images number are the same per class for training, validation and test sets. Our training set consists of 1,000 images per category, focusing on critical defect classes such as settled deposits (AF), roots (RO), attached deposits (BE), displaced joints (FS), and notably, cracks, breaks, and collapses (RB). Among these, RB stands out as the most paramount defect due to its severe implications [15]. These defects were chosen based on both their significant impact on infrastructure integrity and their frequent occurrence. Additionally, while drilled connections (PB) is not a defect but a construction feature, we have incorporated it into our dataset to rigorously test the proposed model's discernment capabilities, given PB's unique characteristics.

For validation, we adopted a set of 10 images per defect class, each with bounding box annotations to assist

in hyperparameter tuning and model selection, but not used for training. This choice of a fixed-size validation set is influenced by Choe et al. [16], advocating that small fixed number of fully annotated images offers a robust yet adaptable baseline for comparative evaluation, thereby ensuring methodological consistency for future research, given that certain level of localization labels are inevitable for WSOL.

Our test set encompassed 50 images per selected class, designed to rigorously evaluate the model's generalization capabilities. By customizing the Sewer-ML dataset in this fashion, we provided a methodological framework that not only catered to the specialized requirements of WSOL but also set a precedent for future research in sewer defect localization.

Table 1. Dataset Information

Dataset	Sewer-ML (Haurum and Moeslund, 2021)
Classes	6
Train	1000
Validation	10
Test	50

*Image numbers are shown per class

3.2 Method

The WSOL method proposed by Kim et al. [2] bridges the gap between image classification and object localization by aligning feature directions to class-specific weights.

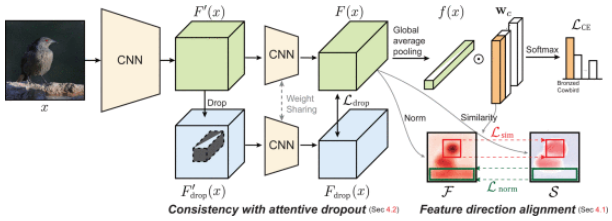

Figure 1. WSOL structure developed by (Kim et al. 2022)

Figure 1 shows the overall of their proposed method. Traditional WSOL methods using CAMs only highlight the most discriminative parts of an object [1]. In contrast, the approach by Kim et al. [2] enhances the alignment of all feature directions to the target class weights through two main strategies: 1) feature direction alignment loss and 2) consistency regularization with attentive dropout. This enables the activation of less discriminative object regions in the CAM, allowing for more accurate and complete localization. The alignment and consistency losses are incorporated into the training of a CNN-based classifier (ResNet-50 or VGG-16) on the Sewer-ML images. This guides the model to align all feature directions to the weights of each defect class.

At inference time, CAMs are generated for each input image to produce a heatmap highlighting potential defect regions. The continuous CAM is then thresholded to obtain a bounding box localization for each predicted defect class. This allows for weakly supervised defect localization using only image-level labels, eliminating the need for manually annotated bounding boxes.

3.3 Experiment Set Up and Training Strategy

We conducted experiments using NVIDIA GeForce RTX 4080 GPU. The WSOL model was implemented in PyTorch, initialized with weights pretrained on CUB-200-2011 [17] for ResNet50 and on ImageNet-1K [18] for VGG16. We evaluated the approach using both VGG-16 and ResNet-50 backbones.

During training, we used a warm-up stage where only the classification and consistency losses are active for the first few epochs. After warm-up, the full training loss was used including the feature direction alignment losses.

We selected the best checkpoint based on localization accuracy on the validation set. The model was then evaluated on the test set to report performance for the six classes.

4 Results and Discussion

The performance of two deep learning models, ResNet50 and VGG16, was evaluated using two primary metrics: Accuracy and Ground Truth Location (GT-LOC). Accuracy measures the percentage of images for which the predicted label matches the true label. GT-LOC, assesses the accuracy of predicting the object's exact location within an image, reflecting the model's ability in object localization.

Table 2. Accuracy and GT-LOC on Validation Set

Backbone	Accuracy	GT-LOC
ResNet50	83.33%	65.00%
VGG16	81.67%	70.00%

Table 3. Accuracy and GT-LOC on Test Set

Backbone	Accuracy	GT-LOC
ResNet50	87.00%	68.33%
VGG16	83.00%	59.67%

Table 2 shows the accuracy and localization performance for both backbones on validation set.

- ResNet50: Exhibits a strong classification accuracy of 83.33%. For object localization, as indicated by the GT-LOC, it achieves 65.00%.
- VGG16: It demonstrates a solid accuracy of 81.67%.

The GT-LOC score of 70.00% shows its capacity in precisely locating objects within images.

Table 3 shows the accuracy and localization performance for both backbones on test set.

- ResNet50: The model has an accuracy of 87.00%, highlighting its consistent performance. The GT-LOC score of 68.33% further emphasizes its reliable object localization capability.
- VGG16: Achieving an accuracy of 83.00%, VGG16 maintains its robustness in object recognition. However, there's a notable drop in GT-LOC to 59.67%, suggesting some variability in its localization performance on the test set.

Additionally, we compared our MaxBoxAccV2 scores [16]. MaxBoxAccV2 measures the models' localization accuracy with ground-truth class based on multiple IOU thresholds. Table 4 and Table 5 indicate that both ResNet50 and VGG16 backbones showed strong results at the loc IOU 30 and 50 thresholds, with scores ranging from 59.67% to 93.00%. Specifically, both ResNet50 and VGG16 have good performances at the loc IOU 30 threshold, achieving as high as 93.00%. However, as the rigor increased to the loc IOU 70 threshold, both models faced challenges, with accuracies dropping to between 18.00% and 31.67%, highlighting the need for potential refinements at higher precision levels. The average localization accuracy across all thresholds placed VGG16 slightly ahead of ResNet50 on the validation set, with scores of 63.89% and 61.67% respectively, however, ResNet50 regained advantages on the test set with a score of 64.33% compared to 56.89% for VGG16. This indicates that while both models are proficient at less stringent thresholds, ResNet50 exhibits a more consistent performance across the board. Our data analysis confirmed that the proposed WSOL method achieved significant localization accuracy thereby validating our hypothesis and addressing the core research question.

Table 4. MaxBoxAccV2 Scores for ResNet50 and VGG16 on Validation Set

Backbone	ResNet50	VGG16
Loc_IOU_30	90.00%	90.00%
Loc_IOU_50	65.00%	70.00%
Loc_IOU_70	30.00%	31.67%
Mean	61.67%	63.89%

Table 5. MaxBoxAccV2 Scores for ResNet50 and VGG16 on Test Set

Backbone	ResNet50	VGG16
Loc_IOU_30	93.00%	93.00%
Loc_IOU_50	68.33%	59.67%
Loc_IOU_70	31.67%	18.00%
Mean	64.33%	56.89%

A confusion matrix visually represents the performance of a classification model by contrasting actual versus predicted classifications. The diagonal elements represent the correct classifications, while off-diagonal elements indicate misclassifications. A well-performing model would have higher numbers on the diagonal and lower numbers off-diagonal.

Figure 2 and Figure 3 show the confusion matrices for both ResNet50 and VGG16 backbones on the test set: ResNet50 showed strong diagonal values for categories BE, PB, and RO, while VGG16 exhibited good results for classes AF, PB and RO. For instance, PB consistently received a high true positive score of 50 for both backbones, indicating that both models have a firm grasp on identifying this class. Similarly, AF, BE, and RO mostly contained high values on the diagonal and minimal off-diagonal interference. However, some classes show room for improvement. For instance, FS and RB had a few off-diagonal values, indicating some misclassifications. RB, in particular, had some misclassifications where it was mistaken for other classes. Figure 4 provides some CAM examples for both backbones for each class.

Figure 2. Confusion Matrix for ResNet50 on Test Set

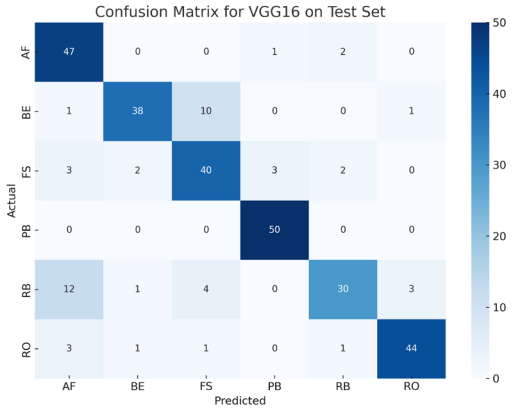

Figure 3. Confusion Matrix for VGG16 on Test Set

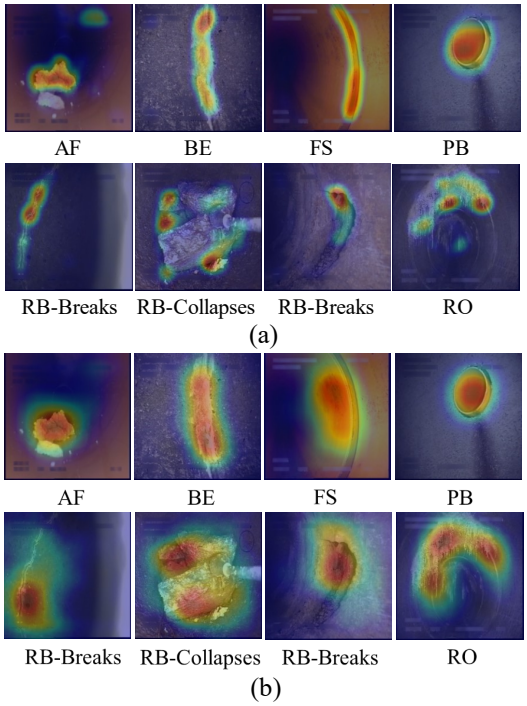

Figure 4. CAM examples for the two backbones for each class: (a) CAM examples for ResNet50; (b) CAM examples for VGG16.

Different backbones are good at localizing different classes. In our study, we found that ResNet50 is relatively better at capturing classes are smaller in scale, like cracks, breaks, and displaced joint, while VGG16 can give a more holistic localization to roots and collapses. Both backbones are good at localizing attached deposits and drilled connection feature, while both fail for settled deposits. A model based on a more recent backbone or even the combination of backbones for different classes should be developed in the future.

Our findings indicated that the VGG16 model tends to produce larger heatmaps for defect localization compared to the ResNet50 model. This difference in heatmap size can have implications on asset management decisions. Larger heatmaps may overestimate the scope of defects, resulting in unnecessary repairs or replacements. Meanwhile, smaller heatmaps could underestimate damage, failing to fully resolve infrastructure issues.

Visualizing CAM to see how and why the method is localizing the defects can be valuable, especially in critical infrastructure inspection like sewer systems where human experts need to trust and understand the model's decision-making.

5 Conclusion and Limitation

This study demonstrated the promising potential of WSOL techniques for automated defect detection in sewer pipelines. The analysis of our results revealed clear trends and relationships that affirm the effectiveness of the proposed WSOL method for sewer pipeline defect detection using image-level annotations. By generating localization heatmaps directly from image-level labels, our approach eliminated the need for tedious bounding box annotations. Across five common sewer defect classes and one construction feature, both ResNet50 and VGG16 achieved reasonable localization performance without any bounding boxes for training.

Specifically, our models attained average MaxBoxAccV2 scores of 64.33% (ResNet50) and 56.89% (VGG16) on the test set. The ResNet50 backbone demonstrated slightly more consistent localization across different IOU thresholds. Additionally, both models showed proficiency in classifying most defect types, with test accuracies reaching as high as 87.00% for ResNet50 backbone and 83.00% for VGG16 backbone.

While these initial results are encouraging, there remain limitations to be addressed. First, the localization accuracy is still far from perfect, with ample room for improvement. Second, certain defect classes like settled deposits proved more challenging for the models to precisely localize. Third, visual analysis revealed that the two backbones had relative strengths and weaknesses in localizing different defect types. An ensemble or multi-backbone approach may help mitigate these class-specific shortcomings. Additionally, transitioning to Weakly Supervised Object Detection (WSOD) approaches would be a valuable future work. WSOD removes the limitation of one object per image class, enabling multi-object defect detection learning directly from image labels. This could better handle real-world sewer images with multiple defects present. Adopting a WSOD approach would require more computational resources but may further advance automated sewer

analysis. Another important future work is to compare our WSOL model against traditional fully supervised models in terms of processing time and accuracy, to understand the trade-offs between annotation efficiency and localization precision in sewer pipeline defect detection. This comparison will help in identifying the most effective approach for automated infrastructure assessment.

References

[1] Zhou B., Khosla A., Lapedriza A., Oliva A. and Torralba A. Learning deep features for discriminative localization. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, Las Vegas, America, 2016.

[2] Kim E., Kim S., Lee J., Kim H. and Yoon S. Bridging the Gap between Classification and Localization for Weakly Supervised Object Localization. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14258-14267, New Orleans, Louisiana, America, 2022.

[3] Wang H., Li Y., Dang L.M., Lee S. and Moon, H. Pixel-level tunnel crack segmentation using a weakly supervised annotation approach. *Computers in Industry*, 133, 103545, 2021.

[4] Zhang J., Su H., Zou W., Gong X., Zhang Z. and Shen F. CADN: A weakly supervised learning-based category-aware object detection network for surface defect detection. *Pattern Recognition*, 109, 107571, 2021.

[5] Wu X., Wang T., Li Y., Li P. and Liu Y. A CAM-Based Weakly Supervised Method for Surface Defect Inspection. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-11, 2022.

[6] Li Y., Wang H., Dang L.M., Song H.-K. and Moon H. Vision-Based Defect Inspection and Condition Assessment for Sewer Pipes: A Comprehensive Survey. *Sensors*, 22, 2722, 2022.

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

[8] Li D., Xie Q., Yu Z., Wu Q., Zhou J. and Wang J. Sewer pipe defect detection via deep learning with local and global feature fusion. *Automation in Construction*, 129, 103823, 2021.

[9] Yin X., Chen Y., Bouferguene A., Zaman H., Al-Hussein M. and Kurach L. A deep learning-based framework for an automated defect detection system for sewer pipes. *Automation in Construction,* 109, 102967, 2020.

[10] Yin X., Ma T., Bouferguene A. and Al-Hussein M. Automation for sewer pipe assessment: CCTV video interpretation algorithm and sewer pipe video assessment (SPVA) system development. *Automation in construction,* 125, 103622, 2021.

[11] Czimmermann T., Ciuti G., Milazzo M., Chiurazzi M., Roccella S., Oddo C. M. and Dario, P. Visual-Based Defect Detection and Classification Approaches for Industrial Applications—A SURVEY. *Sensors*, 20(5), 1459, 2020.

[12] Zhang G., Pan Y. and Zhang, L. Semi-supervised learning with GAN for automatic defect detection from images. *Automation in Construction*, 128, 103764, 2021.

[13] Haurum, J. B. and Moeslund, T. B. Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13455-13467, 2021.

[14] Dansk Vand og Spildevandsforening (DANVA). Fotoman ualen: TV-inspektion af afløbsledninger. Dansk Vand og Spildevandsforening (DANVA), 6 edition, 2010.

[15] Dansk Vand og Spildevandsforening (DANVA). Fotoman ualen: Beregning af Fysisk Indeks ved TV-inspektion. Dansk Vand og Spildevandsforening (DANVA), 1 edition, 2005.

[16] Choe J., Oh S.J., Lee S., Chun S., Akata Z. and Shim H. Evaluating Weakly Supervised Object Localization Methods Right. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3133-3140. 2020.

[17] Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S. and Perona P. Caltech UCSD birds 200. Technical Report CNS-TR-2010-001, Cali fornia Institute of Technology, 2010.

[18] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A.C. and Li F. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.