https://doi.org/10.71573/qqaxgx55

© Authors. This work is licensed under a Creative Commons Attribution 4.0 International License

# Self-supervised learning approach for automatic sewer defect detection

*Tugba Yildizli* [1, *] https://orcid.org/0000-0003-3916-6684, *Tianlong Jia* [1] https://orcid.org/0000-0001-5142-1321, *Jeroen Langeveld* [1, 2] https://orcid.org/0000-0002-0170-6721 *& Riccardo Taormina* [1] https://orcid.org/0000-0002-1550-504X

[1] *Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Water Management, Stevinweg 1, 2628 CN Delft, The Netherlands*
[2] *Partners4urbanwater, Nijmegen, The Netherlands*
**Corresponding author email: t.yildizli@tudelft.nl*

## Abstract

Automated sewer defect detection has advanced through deep learning, particularly supervised methods using CCTV images, but based on large annotated datasets. This study proposes a semi-supervised learning (SSL) approach to reduce the dependency on annotations. The method includes two stages: self-supervised pre-training on unlabelled images using SwAV (Swapping Assignments between multiple Views of the same Image), followed by fine-tuning on labelled images for multi-label image classification. Experiments on the Sewer-ML dataset show that both ImageNet-pre-trained models -supervised and SwAV- outperform models trained from scratch on 1.04 million images, achieving higher F1-scores with just 13k labelled samples. The proposed SSL approach achieves 64.22% precision, 66.06% recall, and a 65.13% F1 score, surpassing the fully supervised baseline. Additionally, scaling up the pre-training dataset further enhances performance. These findings underscore the importance of ImageNet initialization and highlight self-supervised learning as an accurate, scalable, and cost-effective alternative to supervised methods, particularly in data-scarce scenarios.

## Highlights

- Two-stage semi-supervised learning (SSL) for multi-label sewer defect detection.
- Promising way to reduce the reliance on labelled data.
- Self-supervised pre-training outperforms fully supervised baseline.

## Introduction

The sewage network is one of the most important elements of urban infrastructure and plays an essential roles for the collection and transportation of wastewater and stormwater. However, due to usage and exposure to harsh environmental conditions, these systems deteriorate over time, leading to localized and systemic problems. Major issues include cracks and collapses, infiltration and exfiltration, as well flooding and overflows due to capacity issues. Effective condition assessment is necessary to ensure the longevity of their operation and maintain a healthy and sustainable urban environment. CCTV inspection is widely used for assessing the condition of sewer pipes, joints, and manholes, but presents challenges: It is time-consuming, subjective, labour-intensive, and costly. These drawbacks can lead to delayed responses to critical issues. The limitations and inefficiencies of manual inspection have led researchers to look for ways to automate this system.

In recent years, researchers have demonstrated advances in the automatic detection of sewer defects using computer vision and deep learning due to their ability to interpret images and videos (Cheng & Wang, 2018; Meijer et al., 2019). Most automated models for defect inspection are based on supervised learning methods, which means that models depend on a large number of labelled samples. However, manual labelling is time-consuming and leads to poor generalisation and biased learning due to human-generated annotations (Tscheikner-Gratl et al., 2020).
Self-supervised learning, an emerging and promising direction within deep learning, has begun to address the challenges associated with the need for labelled data in supervised learning (Liu et al., 2023). Its applicability extends to numerous scenarios, especially where vast amounts of unlabelled data are available and the task of labelling this data is impractical. This approach may provide several advantages for detecting defects in sewers, such as reduced reliance on manual annotations, improved ability to process large volumes of images, better generalization and scalability to different sewer environments.

This study aims to demonstrate the potential of self-supervised learning for the development of a sewer defect detection algorithm for multi-label sewer defect classification, with the ambition of enabling more accurate and cost-effective approaches to sewer condition assessment from CCTV imagery.

# Methodology

Our approach consists of two steps: first, self-supervised pre-training to learn visual patterns from unlabelled images; second, transfer learning and fine-tuning of the learned representations on the limited labelled data via supervised learning approach. We assess the effectiveness of our approach by comparing fully supervised baseline pre-trained on ImageNet and commonly used supervised DL methods.

### Pretext Task: Self-supervised Learning

To obtain pre-trained weights for sewer-specific data in a self-supervised context, we employ SwAV (Caron et al., 2020), which aims to learn representations from unlabelled data by performing cluster predictions between several augmentations of the same image. The pre-training includes multi-cropping with 8 views as 2 images with 224x224 pixels and 6 images with 96x96 pixels resolution as well as horizontal flipping, colour distortion and Gaussian blur. A ResNet101 backbone (He et al., 2015) architecture is initialized with ImageNet weights (Russakovsky et al., 2014) and then all layers are fine-tuned with SwAV for detailed feature extraction in a self-supervised manner using unlabelled images. The learning process is observed with SwAV loss function presented in (Caron et al., 2020).

### Downstream Task: Multi-label Image Classification

The knowledge acquired during the pretext phase are applied to the subsequent multi-label classification task. We extend the ResNet101 backbone network with a multi-label classifier head that includes an additional fully connected layer for the final class predictions. We employ full fine-tuning, training all model parameters to adapt pre-trained features for the target task.

### Dataset

This study uses the publicly available Sewer-ML multi-label classification dataset introduced by Bruslund et al., (2021). It consists of 1.3 million sewer pipe images with various sewer defects in 17 categories, such as cracks, deformations, obstacles, roots, and infiltration, as well as images without defect. The dataset includes 1.04 million training images as well as 130k validation and 130k test images. In this study, we use the training set for self-supervised pre-training, supervised fine-tuning and validation. We select hundreds of thousands of unlabelled data for the self-supervised pre-training of the models with SwAV, ensuring that the classes were represented proportionally to the original dataset. Afterwards, we use a much smaller labelled images for fine-tune the pre-trained model for

the multi-label classification task. The dataset consists of 9,809 photos for training and 2,412 for validation, total 12,221 images. Subsequently, the original validation set is used as the test dataset since the labels of the Sewer-ML test dataset are kept private.

## Results and discussion

We conducted a series of experiments to evaluate the performance of SSL for sewer defect detection. These experiments aimed to analyse: (i) the impact of dataset size in the pretext task, (ii) the effect of pre-training duration (number of epochs), and (iii) the comparison with a supervised training baseline and with existing sewer-specific and general architectures. The evaluation can provide information about the model's performance using metrics such as average precision, average recall and average F1 score.

For pre-training, ResNet101 with SwAV was trained with different dataset sizes, starting with 104,000, 218,000, and up to a maximum of 312,000. We fine-tuned these models separately for the downstream multi-label classification task and analysed their performance on the original validation dataset. Table 1 lists the evaluation metrics of these models. A trend is observed: the more images used for self-supervised pre-training, the better the performance of the model. The model that was pre-trained with 312,000 images shows the best performance. This model was then used for the subsequent experiments.

**Table 1.** Impact of dataset size in pre-training for the multi-label classification performance. All metrics represent overall values.

| Dataset Size for Pre-training | **Precision (%)** | **Recall (%)** | **F1 Score (%)** |
|---|---|---|---|
| 104.013 images | 64,32 | 61,40 | 62,82 |
| 208.026 images | 64,29 | 63,14 | 63,71 |
| **312.039 images** | **65,66** | **63,84** | **64,25** |

Pre-training with 312k unlabelled images was performed up to 200 epochs, saving model weights every 50. These checkpoints were fine-tuned and evaluated on the downstream task. Table 2 reports their performance. The 100-epoch checkpoint yielded the highest scores. Shorter runs underfit the representations were not yet fully shaped by the data. Beyond 100 epochs, performance plateaued or declined slightly, indicating the onset of overfitting despite continued optimisation. Determining a suitable duration for pre-training is therefore essential to avoid overfitting while preserving meaningful feature representations.

**Table 2.** The impact of training duration (number of epochs) in SSL for the multi-label classification performance. All metrics represent overall values.

| Pre-trained Model | **Precision (%)** | **Recall (%)** | **F1 Score (%)** |
|---|---|---|---|
| SwAV - 50 epochs | 64,04 | 62,43 | 63,23 |
| **SwAV - 100 epochs** | 64,22 | **66,06** | **65,13** |
| SwAV - 150 epochs | **65,62** | 64,57 | 65,09 |
| SwAV - 200 epochs | 64,49 | 63,41 | 63,95 |

We compared our proposed model against two baselines: (1) a fully supervised ResNet-101 model initialized with ImageNet (Russakovsky et al., 2014) pre-trained weights and fine-tuned on the same labelled set, and (2) a ResNet-101 model trained from scratch on the entire SewerML training (Haurum & Moeslund, 2021). The evaluation results on the validation dataset are presented in Table 3.

Our approach, ``ResNet-101-SSL" achieves the highest precision (64.22%) and F1-score (65.13%), outperforming the fully supervised models. Leveraging pre-trained weights, even with limited labelled

data, can outperform models trained from scratch. This suggests that effective representation learning is more impactful than data volume alone. Our self-supervised approach, pre-trained on in-domain unlabelled images, provides further advancements beyond fully supervised learning with ImageNet initialization. These findings highlight the efficacy of SSL in domain-specific representation learning.

**Table 3.** The comparison of SSL with the supervised learning benchmarks. All metrics represent overall values.

| Model | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|
| ResNet101 - SL (Scratch) | 40,63 | **82,62** | 54,47 |
| ResNet101 -SL (ImageNet) | 59,35 | 65,23 | 62,15 |
| **ResNet-101-SSL** | **64,22** | 66,06 | **65,13** |

## Conclusions and future work

This study explored the use of self-supervised learning for multi-label sewer defect classification through CCTV imagery while reducing the dependency on large labelled datasets. The SSL approach offers a strong alternative to fully supervised techniques, especially in domains where high-quality labels are scarce or costly to obtain. Compared to models trained from scratch, those initialized with pre-trained weights—either from ImageNet or SSL—consistently achieve superior performance, even when fine-tuned on limited labelled data. These results underscore the advantages of transfer learning in data-scarce settings.

For future research, more focus is needed for dataset analysis for effective feature extraction with self-supervised learning. In addition, different architectures of self-supervised learning could be implemented to adapt to less computationally intensive environments.

## References

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems*, *2020-December*. http://arxiv.org/abs/2006.09882

Cheng, J. C. P., & Wang, M. (2018). Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques. *Automation in Construction*, *95*, 155–171. https://doi.org/10.1016/J.AUTCON.2018.08.006

Haurum, J. B., & Moeslund, T. B. (2021). *Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark*. 13456–13467. http://arxiv.org/abs/2103.10895

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2023). Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 857–876. https://doi.org/10.1109/TKDE.2021.3090866

Meijer, D., Scholten, L., Clemens, F., & Knobbe, A. (2019). A defect classification methodology for sewer image sets with convolutional neural networks. *Automation in Construction*, *104*, 281–298. https://doi.org/10.1016/J.AUTCON.2019.04.013

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2014). *ImageNet Large Scale Visual Recognition Challenge*. http://arxiv.org/abs/1409.0575

Tscheikner-Gratl, F., Caradot, N., Cherqui, F., Leitão, J. P., Ahmadi, M., Langeveld, J. G., Le Gat, Y., Scholten, L., Roghani, B., Rodríguez, J. P., Lepot, M., Stegeman, B., Heinrichsen, A., Kropp, I., Kerres, K., Almeida, M. do C., Bach, P. M., Moy de Vitry, M., Sá Marques, A., … Clemens, F. (2020). Sewer asset management – state of the art and research needs. *Urban Water Journal*, *16*(9), 662–675. https://doi.org/10.1080/1573062X.2020.1713382