Automation in Construction 182 (2026) 106751

Contents lists available at ScienceDirect

Automation in Construction

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

Self-supervised learning for multi-label sewer defect classification

Tugba Yildizli [a,*], Tianlong Jia [a,b], Jeroen Langeveld [a,c], Riccardo Taormina [a]

[a] Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Water Management, Stevinweg 1, 2628 CN Delft, The Netherlands
[b] Karlsruhe Institute of Technology (KIT), Institute of Water and Environment, Karlsruhe, Germany
[c] Partners4UrbanWater, 6532 ZV, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Semi-supervised learning
Computer vision
Sewer defect classification
Asset management
Transfer learning

ABSTRACT

Automated sewer defect detection has advanced through deep learning, particularly supervised methods using CCTV images, but based on large annotated datasets. This paper proposes a semi-supervised learning (SSL) approach to reduce labeling demands. The method comprises self-supervised pre-training on unlabeled images using SwAV (Swapping Assignments between multiple Views) followed by fine-tuning for multi-label classification. Experiments on the Sewer-ML dataset demonstrate that the SSL approach, trained on only 35k labeled images, achieves an F1-score of 69.11%, and $F2_{CIW}$ of 54.22%, surpassing the fully supervised baseline trained from scratch on 1.04 million images. Increasing the unlabeled pre-training data further enhances performance, while ImageNet initialization consistently outperforms training from scratch. Self-supervised learning also helps mitigate the effects of mislabeled data, which is observed to be present even in the Sewer-ML ground truth. Overall, self-supervised learning provides an accurate, scalable, and cost-effective alternative to fully supervised approaches, particularly in data-scarce or imperfectly labeled scenarios.

1. Introduction

The sewerage system plays a vital role in urban drainage, as it serves as a fundamental function in the collection and transportation of wastewater and stormwater [1]. However, as the pipeline network ages and is exposed to external pressures such as urbanization and climate change, deterioration may arise [2]. Undetected early-stage defects (e.g., cracks, infiltration, displaced joints) may develop into major structural failures and consequently lead to environmental pollution and public health risks [3,4]. Therefore, regular and timely condition assessment is critical for both planning maintenance activities to prevent such consequences and for evaluating and ensuring the long-term performance of the system [5]. At the same time, the continuous expansion of urban drainage networks increases the scale and complexity of inspection demands, which is exemplified by an 8% increase in the Dutch gravity network length from 2016 to 2024 [6]. This growth, combined with a tight labor market, poses significant challenges for conducting large-scale and data-driven inspections.

Traditional condition assessment techniques frequently lead to delays; therefore, there is a shift towards proactive methods [7,8]. These approaches can be broadly categorized into visual and non-visual techniques. Visual methods, such as CCTV and zoom camera inspections, provide direct imagery for assessing pipe conditions, while non-visual methods, including electromagnetic, infrared, laser profiling, acoustic, and ultrasonic sensing, detect anomalies through physical signal measurements [3,9]. Among these, CCTV inspection is the most widely used due to its detailed visual information and practicality. However, it still depends on manual evaluation by experienced inspectors, which makes the process time-consuming, subjective, labor-intensive, costly, and highly dependent on expertise [10].

Researchers have therefore explored deep learning-based automated methods to improve the efficiency and consistency of sewer inspection [11,12]. Most of these models, however, are usually based on supervised learning (SL), which requires extensive and well-labeled datasets for an accurate and reliable model. Sewer defect annotation requires specialized domain expertise, making large-scale labeling both costly and time-consuming. Although previous studies have not reported the time required for manually generating labels for a sewer detection dataset, studies from other fields indicate significant time involved. For example, annotating 1000 instances across 91 common categories (e.g., car, people) with pixel-level segmentation masks in the COCO dataset requires more than 22 worker hours [13]. This highlights that fully supervised methods, though effective, rely on considerable human labeling and therefore involve significant resource demands.

The sewer domain is inherently data-scarce, as collecting large volumes of images from underground pipes is both challenging and costly. In addition, the sharing of sewer inspection data is highly restricted,

* Corresponding author.
E-mail address: t.yildizli@tudelft.nl (T. Yildizli).

https://doi.org/10.1016/j.autcon.2025.106751
Received 26 July 2025; Received in revised form 24 December 2025; Accepted 30 December 2025
Available online 2 January 2026

since most datasets are collected by private or municipal utility companies and are not released publicly due to confidentiality and ownership issues [14]. Moreover, the annotation process is labor-intensive and highly dependent on the inspector's expertise and subjective interpretation, which makes the labeling process prone to inconsistencies or mislabeling [15]. Even expert-annotated sewer datasets exhibit labeling uncertainty due to variations in interpretation [16]. These constraints highlight the urgent need for learning approaches that remain robust and accurate even when only limited and potentially noisy data are available.

To overcome the limitations posed by data scarcity, transfer learning (TL) and self-supervised learning have emerged as promising strategies. TL enables models to leverage knowledge from large-scale datasets such as ImageNet [17] and adapt to sewer-specific visual features using smaller labeled datasets. This provides a better starting point compared to training from scratch, thereby improving generalization across different pipe systems [18,19]. On the other hand, self-supervised learning allows models to learn meaningful visual representations directly from unlabeled data, significantly reducing the dependence on costly manual annotations [20–22]. Semi-supervised learning (SSL) is an effective approach that leverages both labeled and unlabeled data to improve model performance [23]. In practice, SSL typically involves self-supervised pre-training on large unlabeled datasets, followed by fine-tuning on a smaller labeled subset. The superiority of the SSL approach compared to supervised baselines has been demonstrated in recent studies [24,25]. Its applicability extends to numerous fields, especially where vast amounts of unlabeled data are available and labeling is impractical [26].

The core motivation of this paper is to reduce the dependence on large, expert-annotated datasets in sewer defect classification by integrating SSL to address the limitations of both conventional human-driven CCTV inspections and fully supervised machine learning methods. By leveraging large quantities of unlabeled CCTV images for self-supervised pre-training and a small, labeled subset for fine-tuning, we achieved competitive performance with 35,360 images, approximately 30 times fewer samples, corresponding to a 96.6% reduction compared to the 1.04 million images in the Sewer-ML dataset. This efficiency demonstrates the potential of self-supervised learning to overcome the annotation bottleneck and offers a practical path for real-world implementation. Furthermore, SSL provides a pathway for utilities to exploit the vast amount of unlabeled CCTV footage they already possess, which is unsuitable for training traditional supervised models, to build strong domain-specific representations prior to fine-tuning. In practical terms, the proposed framework facilitates large-scale and cost-efficient inspections, enhances maintenance prioritization, and supports data-driven condition assessment planning, while minimizing reliance on extensive manual labeling.

2. Related work

Deep learning offers a wide range of methods for learning features from visual data in computer vision, beyond traditional image processing techniques such as edge detection and morphological operations, which are limited to the analysis of individual pixels [27,28]. Especially in the last decade, computer vision techniques have made significant progress in automated defect detection in civil infrastructure, including sewer inspection [29]. Below, key computer vision approaches in sewer inspection are discussed, followed by a review of label-efficient techniques and recent advances in self-supervised learning.

2.1. Sewer defect classification with computer vision

Automated sewer defect classification in CCTV images has advanced with computer vision techniques based on deep learning. Early studies demonstrated the viability of Convolutional Neural Networks (CNNs) for multi-class sewer defect classification. Kumar et al. [12] used multiple binary classification layers, each trained separately for three defect types: root intrusion, deposits, and cracks. Subsequent work by Meijer et al. [1] proposed a single CNN to overcome the drawback of inefficient multiple layers. They used 2.2 million images with 12 defect classes, which reflects the real distribution of defect classes.

Hassan et al. [30] developed an integrated AlexNet-based model for sewer defect classification on CCTV videos and localization from a text recognition module, achieving up to 96.33% accuracy across six defect types. Xie et al. [31] and Li et al. [32] addressed data imbalance using hierarchical CNNs that first distinguished defective from normal images before classifying specific defect types. Despite this progress, these models were limited by their reliance on large labeled datasets and struggled to detect co-occurring defects within a single image.

While the literature primarily focuses on binary and multi-class studies, recent research has also addressed multi-label classification. Haurum and Moeslund [33] introduced a large multi-label classification dataset for sewer defects, "Sewer-ML". This is the first large-scale public multi-label dataset for sewer defect classification. They evaluated the performance of state-of-the-art approaches in both sewer defect classification and multi-label classification domains. However, the Sewer-ML benchmark studies have relied on training models from scratch, without exploiting the potential benefits of transfer learning. There is also severe imbalance in the distribution of defect classes, which presents a challenge for consistent model performance across all categories.

Despite the success of deep learning in sewer defect classification, achieving robust performance remains a significant challenge in the data-scarce sewer domain and for label-dependent models. This highlights the need for label-efficient learning methodologies. Reducing dependency on labeled data has been explored in deep learning using various strategies, such as data augmentation and transfer learning. Nonetheless, SSL offers a more scalable methodology by using unlabeled data to learn domain-specific representations. The following sections review these label-efficient strategies within the sewer domain and discuss current advancements in SSL.

2.2. Label-efficient strategies for sewer domain

Data augmentation and transfer learning can be practical ways to reduce the annotation burden. For example, Zhou et al. [34] proposed a label-efficient approach with a CNN for six defect classes by applying data augmentation and transfer learning. They showed that data augmentation improved prediction accuracy by 15%, while a transferred SqueezeNet achieved slightly higher accuracy but required 13 times more computation time. Similarly, Situ et al. [18] demonstrated the efficiency of transfer learning for improved prediction performance on sewer defect detection with the YOLO network [35]. While they reduced the label dependency indirectly, there is still a need for labeled data for a more generalized model.

A shift towards "Machine Supervision" by Singh et al. [36] demonstrated that machine-level supervision can reduce label dependency, achieving comparable performance with 50% fewer annotations in medical imaging. In light of this, several studies have been conducted with self-supervision strategies for the sewer domain as well. Qiu et al. [37] implemented unsupervised learning at the system level to develop early warning systems for sewerage anomalies, providing a data-efficient solution. Yin et al. [38] developed a cost-effective method for sewage defect localization at the image level. This method utilizes weakly supervised object localization (WSOL) to generate heatmaps from Sewer-ML's [33] image-level labels, eliminating the need for bounding box annotations. Beyond images, Li et al. [39] introduced semi-supervised point-cloud segmentation for sewer defects, showing that leveraging abundant unlabeled data can improve performance while reducing annotation cost. In this regard, approaches that increase label efficiency by integrating this line of research into sewer defect classification are considered.

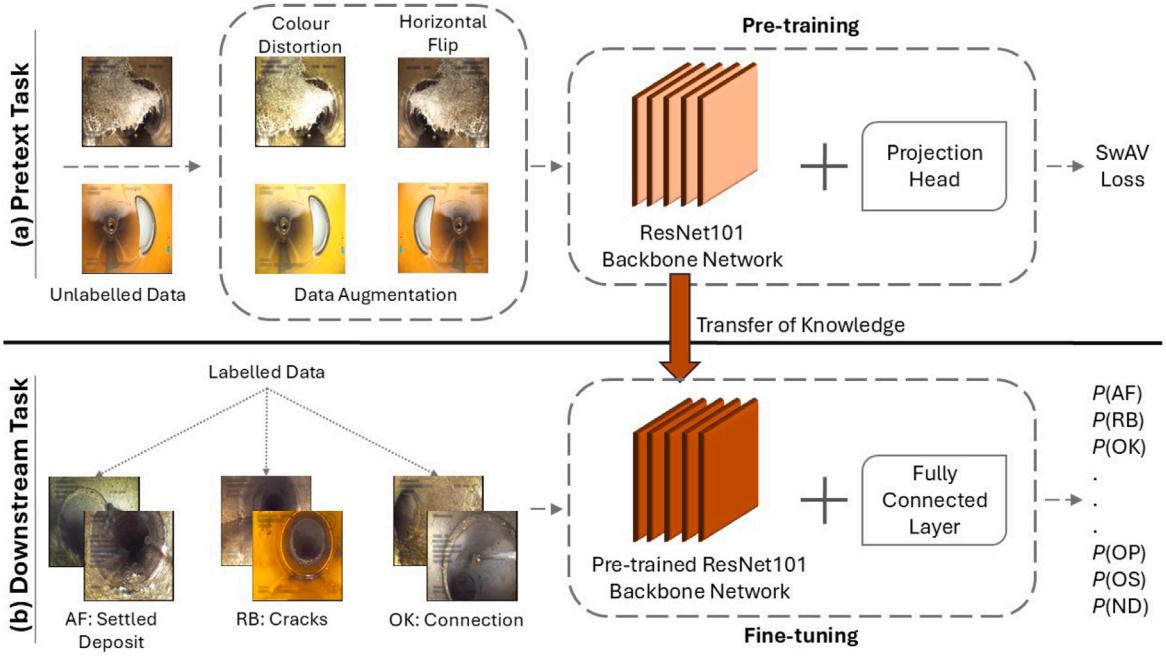


Fig. 1. Illustration of two-part approach. **(a) Stage 1 - Pretext task**: A ResNet101 backbone network is pre-trained with SwAV on a large number of unlabeled images of sewer defects. **(b) Stage 2 - Downstream task**: The learned representations are transferred, and the model is fine-tuned on a limited annotated dataset for supervised multi-label defect classification. The probability of each defect is then calculated, denoted as *P(defect)*.

2.3. Self-supervised learning

Self-supervised learning has emerged as a powerful approach to overcome the limitations of SL, such as data scarcity or difficulties in labeling data. In contrast to supervised learning, which relies on a vast amount of labeled data, self-supervised models can learn underlying structural meaning from the data itself. This enables the model to generalize better on unseen data. Self-supervised learning methods are broadly divided into contrastive and non-contrastive approaches [40]. Contrastive methods, such as Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [41], learn by distinguishing between similar and dissimilar pairs, where similar pairs are created through data augmentations (e.g., cropping, color jitter) of the same image. The model is trained to extract features that are invariant to such transformations. Non-contrastive methods include clustering-based approaches like Swapping Assignments between Views (SwAV) [42], which assign pseudo-labels by grouping similar representations and training the model to predict cluster assignments. This helps reveal semantic structure in unlabeled data. Other non-contrastive strategies include generative methods [43,44], which reconstruct parts of the input, as well as predictive models [45] that infer missing or future data.

Recent applications of self-supervised learning in several domains have shown that the method is highly efficient in capturing representations without relying on extensive labeled datasets. For example, Azizi et al. [46] applied self-supervised learning to medical images and achieved higher classification accuracy compared to supervised baselines. Zabin et al. [47] adopted contrastive learning for defect detection on metal surfaces and demonstrated its generalization capability. In agriculture, Guldenring and Nalpantidis [48] utilized SwAV for plant classification and showed improved accuracy with few annotations. Recently, Jia et al. [25] proposed a SwAV-based two-stage semi-supervised approach for floating litter detection in environmental monitoring, demonstrating that the self-supervised approach improves generalization performance under limited data conditions.

SwAV has achieved superior performance on ImageNet, outperforming SimCLR and MoCo [49] while being memory efficient. Its advantage arises from the online clustering strategy, which allows representations of both positive and negative samples to converge when they share similar structural patterns. This characteristic is particularly relevant for sewer imagery, where repetitive textures and patterns frequently occur within defect classes. In addition, the successful applications of SwAV in the environmental domain [25,48] have further proved its robustness. Driven by these strengths, SwAV is implemented for self-supervision to explore its potential in sewer defect classification.

3. Methods and materials

This section presents the methodological framework of this paper. First, a two-stage semi-supervised learning pipeline is introduced, with the implementation details of both stages presented subsequently. This is followed by a description of the dataset used in the paper and an outline of the methodological approach adopted to analyze labeling inconsistencies.

3.1. Semi-supervised learning for sewer defect classification

A two-stage semi-supervised learning framework is proposed as illustrated in Fig. 1. The first stage involves self-supervised pre-training using a pretext task, enabling the model to learn general and domain-relevant representations from unlabeled images. Specifically, a ResNet101 backbone [50] is trained using the SwAV method [42]. The second stage performs supervised fine-tuning on the downstream classification task, where the pre-trained model is fine-tuned on a small set of labeled images. The learned representations from the self-supervised stage are transferred, and a fully connected layer is added to produce the final multi-label classification output. The subsequent sections provide a detailed description of these phases.

3.1.1. Stage 1: Self-supervised pre-training using SwAV

For pre-training, the SwAV self-supervised learning approach is employed. SwAV is a hybrid self-supervised framework that integrates contrastive learning with clustering-based objectives for representation

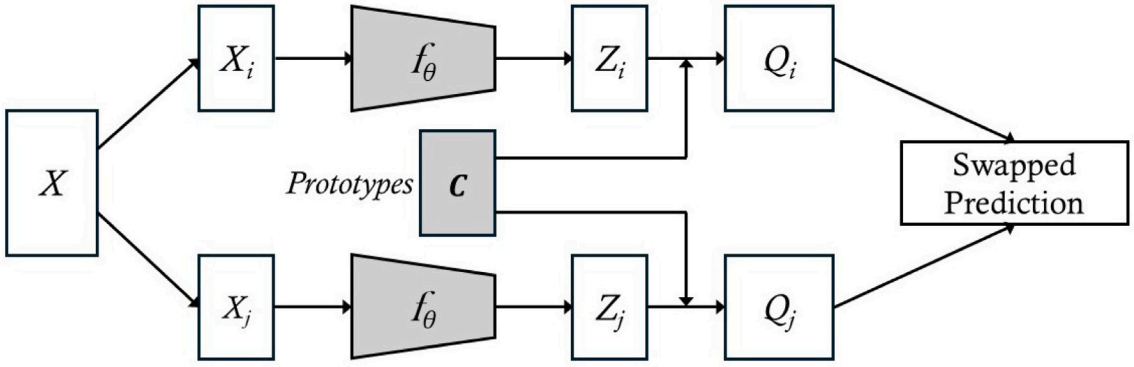


Fig. 2. Illustration of SwAV workflow [42].

learning [42]. SwAV has demonstrated effectiveness in various computer vision tasks, achieving comparable accuracy to state-of-the-art supervised models [51]. In contrast to traditional contrastive learning, the SwAV model can learn representations by performing cluster predictions between multiple augmentations of the same image without distinguishing between positive and negative samples. The SwAV architecture for swapped predictions is shown in Fig. 2.

The procedure starts with a batch of images X, each undergoing different augmentations, resulting in X_i and X_j. For simplicity, only two augmentations are shown in Fig. 2. The encoder network and the subsequent projection head (i.e., fully connected layers) then process these augmented views to create feature vectors Z_i and Z_j. The vectors are mapped to the prototypes designated as C, resulting in their assignment to clusters Q_i and Q_j using the Sinkhorn–Knopp algorithm [52]. The clusters serve as pseudo-labels, and the model is trained to predict Q_j from X_i, a process called swap prediction. The SwAV loss, a swapped prediction loss, is then computed to optimize the encoder. This loss function assesses the degree of similarity between the feature vectors and the corresponding prototype clusters. The mathematical formula for the loss function is given by Caron et al. [42], as follows:

$$L(z_i, z_j) = l(z_i, Q_j) + l(z_j, Q_i) \tag{1}$$

where $l(z, Q)$ calculates the coherence between the cluster predictions Q and the feature z and is computed as follows:

$$l(z_i, Q_j) = -\sum_k Q_j^{(k)} \log p_i^{(k)} \tag{2}$$

$$p_i^{(k)} = \frac{\exp\left(\frac{1}{\tau} z_i^\top c_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau} z_i^\top c_{k'}\right)} \tag{3}$$

where τ is a temperature parameter that indicates the sharpness of the probability distribution and C_k is the mapped prototype vector.

ResNet101 is employed as the encoder network, as it is a deep neural network that works well with complex image data to extract high-level features. Following Jia et al. [25], the network is initialized with ImageNet-1K pre-trained weights to avoid poor local minima and accelerate convergence, rather than using random initialization prior to SwAV training.

3.1.2. Stage 2: Supervised fine-tuning for multi-label classification

In this section, the transfer learning approach is implemented in the context of semi-supervised learning. In transfer learning, knowledge from a related task is used to improve performance on a new task with limited labeled data [53]. A model is first trained on a large data set and then fine-tuned on a smaller, task-specific data set. The knowledge acquired during the pretext phase is transferred to the subsequent multi-label classification downstream task. The ResNet-101 backbone is extended with a multi-label classification head comprising an additional fully connected layer for final class prediction. Full fine-tuning is performed, updating all model parameters such that the pre-trained weights provide effective initialization while the network further adapts the learned representations to the target task.

To optimize the multi-label classification objective, the network is trained using a weighted binary cross-entropy (WBCE) loss to address class imbalance among defect classes. The formulation from the Sewer-ML benchmark [33] is adopted. For each class $c \in \{1, \ldots, 17\}$, the loss is defined as

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^{C} -\left[w_c^{(\text{eff})}\, y_c \log \sigma(x_c) + (1 - y_c) \log(1 - \sigma(x_c))\right], \tag{4}$$

where $C = 17$, $y_c \in \{0, 1\}$ is the ground-truth label, x_c is the raw output of the model, and $\sigma(\cdot)$ denotes the sigmoid activation. Following [33], the per-class weight w_c is defined as the ratio of negative to positive samples.

$$w_c = \frac{N - N_c}{N_c}, \tag{5}$$

where N is the total number of training samples, and N_c is the number of samples containing class c. However, directly applying w_c can yield excessively large weights for rare defects, leading to degraded precision and unstable training. To mitigate this, a moderation strategy is introduced:

$$w_c^{(\text{eff})} = \alpha \sqrt{\text{clip}\left(w_c,\, 1,\, \tau\right)} \tag{6}$$

In Eq. (6), the `clip` function limits excessively large weights to a threshold τ, the square-root transformation reduces their dynamic range, and the scalar α controls the global weighting strength. In the implementation, τ was empirically set to 10 and $\alpha = 1.0$. This moderated weighting preserves the recall improvement for minority classes while preventing false-positive inflation, thus achieving a more stable precision–recall balance.

3.2. Dataset

This paper employs the publicly available Sewer-ML multi-label classification dataset presented by Haurum and Moeslund [33]. The dataset comprises 1.3 million images of sewer pipes with various defects in 17 classes, such as cracks, deformations, obstacles, roots, and infiltrations, as well as images without defects. To facilitate visual understanding, representative image samples from all classes of the Sewer-ML dataset are provided in Figure A.1 of the supplementary material. Sewer-ML includes 1.04 million training images, 130k validation images, and 130k test images. The training set is used for self-supervised pre-training, supervised fine-tuning, and validation. Subsequently, the original validation set is used as the test dataset, as the labels of the Sewer-ML test dataset are not publicly available. Data preprocessing and augmentation for training are explained in Section 4.1.

3.3. Assessment of labeling consistency via feature-space analysis

To support the rationale for adopting a self-supervised learning approach that relies less on carefully labeled data, dataset consistency was examined via feature-space analysis to identify potential labeling noise that could affect model performance. For this analysis, 4202 images were selected with particular attention to minimizing obvious labeling errors. Feature embeddings were extracted from a ResNet-101 model pre-trained on ImageNet to capture high-level visual representations. To facilitate visualization, these features were reduced to 50 dimensions using Principal Component Analysis (PCA) [54] and subsequently projected into two dimensions with t-distributed Stochastic Neighbor Embedding (t-SNE) [55]. The resulting 2D projections were examined to evaluate intra- and inter-class coherence and to identify samples that appeared visually disconnected from their assigned clusters. Such outliers were considered indicative of possible labeling inconsistencies.

4. Implementation details

This section outlines the practical aspects of implementing the proposed SSL framework. Dataset construction and preprocessing are described for both self-supervised pre-training and supervised fine-tuning, followed by architectural configurations. Additionally, computational environment details and evaluation metrics are reported for full reproducibility and transparency of the experimental setup.

4.1. Data preprocessing and augmentation

This section describes the construction of unlabeled and labeled subsets and the resizing, normalization, and augmentation procedures used for SwAV pre-training and subsequent multi-label fine-tuning.

4.1.1. Pre-training

Several experiments were conducted to select the optimal SwAV pre-trained model. These experiments examined the influence of pre-training dataset size and pre-training duration on downstream multi-label classification performance. The training set of Sewer-ML was divided into 14 subfolders so that the class distribution corresponds to the original distribution. The distribution is such that half of the images are normal pipe images, and the rest have one or more defect labels per image. No labels are used during the self-supervised pre-training; the reported proportions are provided only to characterize the data (see Supplementary Table B.1). Three pre-training sets were formed by merging complete subfolders in a fixed order. The first subset contains 104,013 unique images from one subfolder, the second with 208,026 images from two subfolders, and the third with 312,039 images from three subfolders, corresponding to the maximum dataset size permitted by the available computational capacity. These numbers represent the unique base images in each batch prior to any data augmentation.

The image preprocessing included tensor conversion, resizing to 224 × 224 pixels, and standard ImageNet normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). Data augmentation is crucial during pre-training with SwAV, as it forms the starting point of the procedure. The default augmentations from the SwAV framework by Caron et al. [42] were then employed. These include multi-crop with 8 views — 2 images at 224 × 224 pixels and 6 images at 96 × 96 pixels resolution — as well as horizontal flipping, color distortion, and Gaussian blur. Since sewer images have unique visual characteristics, it is essential to preserve feature integrity during augmentation while ensuring generalizability. Therefore, smoother data augmentations are applied by reducing the strength of Gaussian blur to 0.1 and color distortion to 0.5. Each model was trained separately on the corresponding dataset batch using the same augmentation settings. Training was conducted for up to 80 epochs for all experiments.

Table 1
Summary of defective and non-defective images across fine-tuning dataset batches.

Batch	Defective images per class	Non-defective images	Total images
Batch 1	65	1105	2210
Batch 2	130	2210	4420
Batch 3	260	4420	8840
Batch 4	520	8840	17,680
Batch 5	1040	17,680	35,360

4.1.2. Fine-tuning

A limited labeled subset was used to fine-tune the pre-trained model for multi-label classification. Fine-tuning samples were randomly selected from the portion of the Sewer-ML training split not used during self-supervised pre-training, ensuring no overlap between phases. Each image was stored once and annotated with a multi-hot vector over the 17 defect classes. "Non-defective" (ND) was additionally treated as an implicit 18th class for training and evaluation (ND = 1 if all defect entries are 0). No label-based image duplication was performed; images with co-occurring defects contained multiple positive entries.

To evaluate the impact of labeled data volume on model performance, five labeled subsets were constructed by varying the number of labeled images per defect class. Specifically, the training and validation subsets contained 50–15, 100–30, 200–60, 400–120, and 800–240 labeled images, respectively. For instance, Batch 1 included 50 training and 15 validation images per defect class (850 training and 255 validation images in total). To preserve the overall distribution of the original dataset, an equal number of non-defective (ND) images was included relative to the number of defective images, resulting in a balanced composition. Consequently, the total dataset sizes were 2210, 4420, 8840, 17,680, and 35,360 images for Batches 1–5, respectively. The detailed distribution of defective and ND images across batches is reported in Table 1. For each class, images were selected independently; however, since each image may contain multiple labels, all associated defect labels were retained. This approach ensured that the natural multi-label relationships and the overall distribution of the original dataset were largely preserved.

To achieve the same input size as in pre-training, images were resized to 224 × 224 during data preprocessing. Data augmentation techniques such as horizontal flip and color jittering (±0.1) were applied to training images but not to validation or test images. Pixel values were converted to floating-point values in the range [0, 1] and subsequently normalized using the mean and standard deviation of the Sewer-ML training split.

4.2. Experimental setup

The VISSL framework [56] was employed for self-supervised training, implemented in Python 2.20 and PyTorch 1.8.1. Experiments were conducted using an NVIDIA Tesla V100S GPU. Default VISSL hyperparameters for SwAV were adopted, except for "ϵ", which controls the smoothness of cluster assignments in the Sinkhorn–Knopp algorithm. Lower ϵ yields sharper assignments, whereas higher values produce smoother, more uniform distributions. Given the repetitive patterns in sewer imagery, ϵ was reduced from 0.05 to 0.03 to prevent uniform embeddings and improve feature discrimination. In addition, the batch normalization layer was removed from the projection head to improve convergence.

Fine-tuning was performed for 30 epochs using the SGD optimizer with a learning rate of 0.01, momentum of 0.9, and a batch size of 128. The network was trained with the weighted binary cross-entropy loss described in Section 3.1.2, using the moderated class weights defined in Eq. (6). A sigmoid activation was applied to each output node to produce independent probabilities for the 17 defect classes. Predictions were binarized with a probability threshold of 0.5, a commonly adopted

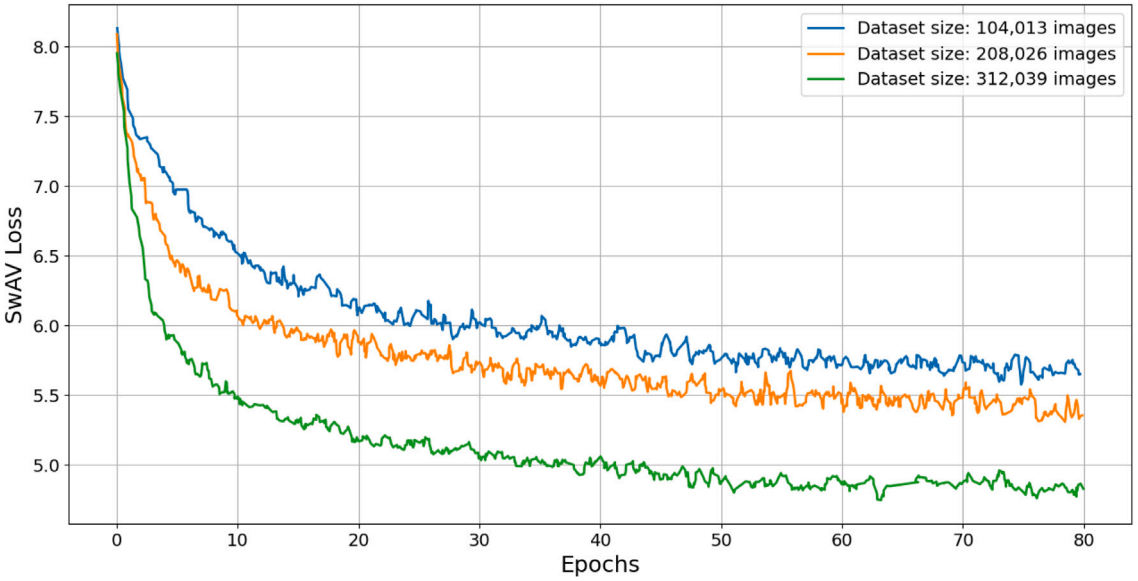


Fig. 3. SwAV pre-training loss curve per epoch, with different amounts of unlabeled images.

value that minimizes missed detections in sewer inspection tasks. An "ND" class was implicitly represented by images with no assigned defect labels. For model evaluation, the full Sewer-ML validation split comprising 130,046 images was used.

4.3. Evaluation metrics

For each class, true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are computed. Metrics are then computed for each class using a one-versus-rest approach (precision, recall, and F1-score, from TP/FP/FN). Precision is the ratio of accurately detected defects to all predicted positives, whereas recall is the percentage of actual defects that are successfully predicted by the model. The F1-score represents a harmonic mean of precision and recall. The overall ("micro") precision/recall/F1-score is then presented by aggregating TP, FP, and FN over all images. This aligns with the Sewer-ML "overall score across all samples, regardless of class" and allows for a fair comparison. In addition to these metrics, a recall-weighted $F2_{CIW}$ presented by Haurum and Moeslund [33], which measures the economic impact of the defect class based on their CIW values (Class Importance Weights), is employed for fair comparison. Furthermore, the $F1_{Normal}$ score is reported to evaluate model performance on non-defective pipe frames, which are excluded from the $F2_{CIW}$ computation. These evaluation metrics are explained in the supplementary material (see Supplementary C).

5. Results and discussion

First, the impact of pretext-task dataset size is presented, followed by the effect of pre-training duration. Second, performance is compared against a fully supervised baseline. Third, results are benchmarked against sewer-specific and general architectures reported by Haurum and Moeslund [33]. To further investigate the results, an exploratory analysis of the dataset was conducted to assess its limitations, with a focus on mislabeling issues and label inconsistencies within the Sewer-ML dataset.

5.1. SwAV model selection

ResNet101 with SwAV was trained with different dataset sizes, starting with 104,013, 208,026, and up to a maximum of 312,039 images. For all three pre-trainings, the SwAV loss values representing the learned features were plotted. The loss curves for these pre-trainings are shown in Fig. 3. The SwAV loss value initially decreased rapidly for each model, as basic image elements such as edges and corners had already been captured by ImageNet weights. All training runs were performed for up to 80 epochs, as the loss curve stabilized by this point. A batch size of 64 was used for models trained on 104,013 and 208,026 images and increased to 128 for 312,039 images to better leverage the larger dataset and available GPU resources and to obtain a more stable optimization trajectory. In this setting, the loss decreased significantly to below 5, reaching the lowest loss value among the evaluated models. Nevertheless, further evaluation on the downstream task is required to assess SwAV's feature extraction capabilities across different dataset scales.

Table 2
Impact of dataset size in pre-training on downstream multi-label classification task performance.

Dataset size for pre-training	Precision	Recall	F1-Score
104,013 images	60.61	73.98	66.63
208,026 images	61.04	74.97	67.29
312,039 images	**62.80**	**76.03**	**68.78**

Note: All metrics represent overall values as percentages.

The models pre-trained on different dataset sizes were subsequently fine-tuned for the downstream multi-label classification task and evaluated on the Sewer-ML validation set. Table 2 lists the evaluation metrics overall precision (OV-P), overall recall (OV-R), and overall F1-score (OV-F1) for the three models mentioned. Increasing the pre-training dataset from 104,013 to 208,026 images led to a modest OV-F1 improvement of 0.46 percentage points, whereas the subsequent increase to 312,039 images yielded a larger gain of 1.49 percentage points. These results support the conclusion that increasing the dataset size enhances downstream performance. The bigger jump from 208,026 to 312,039 can also relate to a change in batch size (from 64 to 128), which contributed to more stable optimization. Overall, the model pre-trained on 312,039 images achieved the highest precision, recall, and F1-score, demonstrating the positive effect of larger pre-training datasets on overall performance. This improvement suggests that increasing the volume of unlabeled data enables the model to learn more diverse and transferable representations.

In the next phase, the effect of pre-training duration on multi-label classification performance was examined. For subsequent experiments,

Table 3
Impact of pre-training duration on multi-label classification performance, using 312k images.

Pre-trained model	Precision	Recall	F1-Score
SwAV - 50 epochs	62.36	75.93	68.48
SwAV - 80 epochs	62.80	76.03	68.78
SwAV - 100 epochs	63.36	76.01	69.11
SwAV - 150 epochs	63.19	**76.38**	69.16
SwAV - 200 epochs	**63.87**	76.35	**69.55**

Note: All metrics represent overall values as percentages.

the model pre-trained on 312k images was used to ensure that evaluations relied on the strongest representations. To identify the optimal pre-training checkpoint for downstream performance, pre-training was run for up to 200 epochs, with model weights saved every 50 epochs. Training for 200 epochs required approximately 200 h on a single GPU, indicating a significant but reasonable computational cost for acquiring robust representations.

The saved checkpoints were fine-tuned and evaluated on the downstream task with the original validation set of Sewer-ML. As shown in Table 3, the performance metrics gradually improve with more pre-training epochs. While recall reaches its peak at 150 epochs (76.38%), the F1-score, which balances precision and recall, continues to increase, reaching its highest value (69.55%) at 200 epochs. This trend indicates that longer pre-training allows the backbone to learn more stable and discriminative visual representations from the unlabeled sewer images, which transfer effectively to the supervised fine-tuning stage. Similar findings have also been reported by Jia et al. [57], where longer pre-training improved downstream task performance.

The gains from increasing pre-training epochs are modest and smaller than those from enlarging the pre-training dataset (Table 2). This suggests that data scale is the main factor influencing downstream performance, while longer pre-training offers secondary but beneficial improvements. In practical terms, with limited computational resources, focusing on a larger unlabeled dataset is likely to yield a greater return on investment than significantly increasing the number of training epochs on a fixed dataset.

5.2. Multi-label sewer defect classification with self-supervised learning

Following self-supervised pre-training of the backbone, semi-supervised performance for multi-label sewer defect classification is evaluated. The analyses are organized into two sections. First, internal performance is assessed by examining the effect of labeled fine-tuning set size on downstream results and by comparing against an ImageNet-initialized fully supervised baseline. This analysis emphasizes the label efficiency and robustness of the proposed strategy across varying levels of supervision. Second, performance is benchmarked against Sewer-ML by comparing with prior studies and state-of-the-art methods [33].

5.2.1. Data scale impact on self-supervised vs. Fully supervised learning

A controlled scaling study was conducted by varying the amount of labeled data for fine-tuning. Five dataset batches were considered (Section 4.1.2), with total sizes of 2260, 4420, 8840, 17,680, and 35,360 images, respectively. The architecture and training protocol were kept fixed across all experiments. Two initialization strategies were compared: SwAV pre-training for semi-supervised learning (SSL) on domain data and a fully supervised (FS) ImageNet-initialized baseline. Performance was reported using precision, recall, F1-Score, recall-weighted $F2_{CIW}$, and $F1_{Normal}$. This analysis shows the effect of the label volume and enables a direct assessment of label efficiency and operating-point differences between SSL and FS.

Table 4 shows that the SSL model outperforms the FS baseline and achieves higher recall, F1-Score, and $F2_{CIW}$ values across all dataset scales. The largest relative gains were observed under limited supervision, where the model fine-tuned on only 2210 labeled images achieved

Table 4
Comparison of Semi-Supervised (SSL) and Fully Supervised (FS) models across dataset scales.

Dataset size	Model	Precision	Recall	F1-Score	$F2_{CIW}$	$F1_{Normal}$
2210 images	SSL	56.7	71.61	63.29	41.27	85.94
	FS	60.21	63.21	61.67	28.62	82.93
4420 images	SSL	56.75	72.74	63.76	44.20	85.87
	FS	58.72	67.45	62.78	38.67	84.58
8840 images	SSL	59.74	74.10	66.15	48.01	86.83
	FS	60.90	69.13	64.76	43.71	85.22
17,680 images	SSL	61.93	75.06	67.87	51.87	87.60
	FS	61.61	71.59	66.23	48.06	86.22
35,360 images	SSL	63.87	**76.35**	**69.55**	**54.83**	**87.99**
	FS	**65.53**	71.92	68.58	50.84	86.91

Note: All metrics represent overall values as percentages.

an increase of +8.4 percentage points in recall and +12.7 percentage points in $F2_{CIW}$. This indicates that pre-training on unlabeled domain data enables the model to learn more transferable and defect-sensitive representations, reducing dependence on labeled data.

FS exhibits slightly higher precision, which means a more conservative decision boundary. In contrast, SSL maintains a better overall balance between precision and recall and consistently achieves higher $F1_{Normal}$ scores. As the amount of labeled data increases, the gap in $F2_{CIW}$ between the two models narrows, as shown in Fig. 4, yet SSL consistently outperforms FS in both $F2_{CIW}$ and F1 across all data scales. These findings confirm the label efficiency and generalization advantages of SSL. For an extended evaluation of dataset scaling effects, see Supplementary Section D.

While SSL requires additional pre-training and more computational resources, its overall cost-effectiveness is visible with annotation efficiency. Since the pre-training stage relies on unlabeled images, it is free from manual annotation costs. As shown in Table 4, the proposed SSL model trained on only 2210 labeled images achieves an $F2_{CIW}$ of 41.27%, outperforming FS trained on 4420 images ($F2_{CIW}$ of 38.67%) and performing comparably to FS trained on 8840 images ($F2_{CIW}$ of 43.71%). This demonstrates that SSL substantially improves label efficiency, achieving competitive or superior performance with up to four times fewer annotations.

As reported in the ImageNet paper [17], labeling one image requires approximately one minute; therefore, annotating the additional 2210 images would demand around 37 h of annotation effort. At an average rate of €20 per hour for domain-specific labeling tasks, this amounts to roughly €740. In comparison, one-time pre-training of the proposed SSL model (200 h) would cost about €300, assuming an average GPU rate of €1.5 per hour (e.g., Google Cloud). The resulting reduction in labeled data directly translates to significant savings in human and financial resources, outweighing the one-time computational cost of pre-training.

5.2.2. Comparison against Sewer-ML benchmark

In this section, SSL and FS models are compared with existing methods reported in the Sewer-ML paper [33]. This comparison aims to evaluate the effectiveness of the proposed approach, fine-tuned on a much smaller dataset, relative to both sewer-specific and general multi-label architectures trained on the full Sewer-ML dataset of 1.04 million images.

The Sewer-ML benchmark studies in Table 5 include four sewer-specific models [30,31,58,59] and five general multi-label architectures, such as ResNet101, KSSNet, and the TResNet family. These models were trained from scratch on the full Sewer-ML training set comprising 1.04 million images with a strong positive weighting strategy to address severe class imbalance. In contrast, the proposed models were trained on a substantially smaller subset (35,360 images) with moderated positive weighting to avoid overemphasizing minority

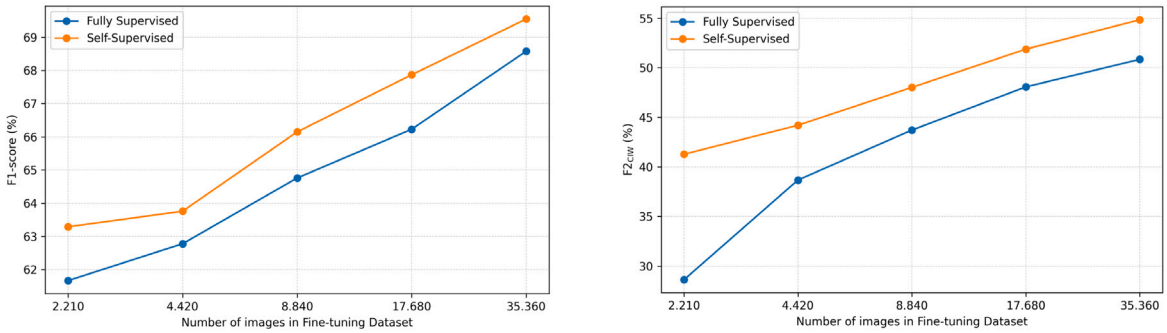


Fig. 4. Dataset scale impact on (a) Overall F1 Score and (b) $F2_{CIW}$.

Table 5
Comparison of SSL with the supervised learning benchmarks.

Model	Precision	Recall	F1-score	$F2_{CIW}$	$F1_{Normal}$
Sewer-specific					
Xie et al. [31]	46.31	82.52	59.33	48.57	**91.08**
Chen et al. [58]	26.38	47.60	33.94	42.03	3.96
Hassan et al. [30]	7.44	44.86	12.76	13.14	0
Myrans et al. [59]	3.19	17.27	5.39	4.01	26.03
General Architectures [33]					
ResNet-101	40.63	82.62	54.47	53.26	79.55
KSSNet	42.52	82.77	56.18	54.42	80.6
TResNet-M	41.22	83.88	55.27	53.83	81.23
TResNet-L	42.09	83.69	56.01	54.63	81.22
TResNet-XL	41.82	**83.98**	55.83	54.42	81.81
Our Work					
ResNet-101-FS	**65.53**	71.92	68.58	50.84	86.91
ResNet-101-SSL	63.87	76.35	**69.55**	**54.83**	87.99

Note: All metrics represent overall values as percentages.

classes. All results reported in the table correspond to evaluations conducted on the Sewer-ML validation set.

The precision of both proposed models is considerably higher than those of the Sewer-ML baselines, indicating a reduced rate of false positives despite training with fewer images. However, the large-scale Sewer-ML models obtain higher recall due to their large-scale training and stronger weighting. This distinction is particularly relevant for interpreting the results, as stronger class weighting tends to boost recall but may reduce precision. The proposed models exhibit a more balanced precision–recall behavior, demonstrating better generalization and robustness with limited labeled data, as demonstrated by higher F1-score and $F1_{Normal}$.

As summarized in Table 5, both SSL and FS models outperform previous sewer-specific methods and are competitive with the general architectures. SSL-based ResNet-101 achieves the highest overall F1-score (69.55%) and the highest $F2_{CIW}$ (54.83%), substantially outperforms the fully supervised ResNet-101 in the Sewer-ML benchmark (53.26% in $F2_{CIW}$). Furthermore, it slightly surpasses the best-performing general architecture, TResNet-L (54.63% in $F2_{CIW}$). These results show that domain-specific self-supervised pre-training is an efficient approach for downstream defect classification. It significantly enhances performance and achieves comparable results to models trained on extensive, fully labeled datasets.

Leveraging pre-trained weights, even with limited labeled data, can outperform models trained from scratch. This suggests that effective representation learning is more impactful than data volume alone. Although ImageNet initialization yields performance enhancements despite being out-of-domain, self-supervised approach, pre-trained on in-domain unlabeled images, provides further advancements beyond fully supervised learning. These findings highlight the efficacy of SSL in domain-specific representation learning. Scaling up the pre-training dataset can further improve performance over supervised baselines, as more diverse unlabeled data can lead to stronger representations. Thereby, it serves as a strong alternative to fully supervised techniques, especially in fields where high-quality labels are scarce or costly to obtain. For completeness, Supplementary Section E reports a detailed class-wise evaluation and representative prediction examples that qualitatively illustrate these results.

5.3. Labeling errors in Sewer-ML dataset

The 2D projections, shown in Fig. 5, reveal the distribution of images across classes. This analysis allowed to assess intra-class coherence and identify outliers, many of which indicate possible labeling errors or underlying dataset biases. From these projections, six representative samples were selected for qualitative inspection, as shown in Table 6. These examples were chosen because they were spatially isolated from other samples in their assigned class and instead positioned closer to clusters of different defect types. Visual inspection confirmed that several of these outliers exhibit inconsistencies with their assigned labels.

Specifically, IDs ID1, ID2, and ID3 are labeled as ND but appear outside the ND clusters in Fig. 5. Visual inspection confirms that each image exhibits a visible defect, corresponding to “PB”, “OP”, and “RB”, respectively. Their location in feature space aligns with typical clusters for these classes. Similarly, examples ID4, ID5, and ID6 are labeled as defect-present frames, yet visual inspection reveals no observable defects. These images appear as outliers within the clusters of their assigned defect classes and are instead aligned with the rightmost ND cluster. External factors such as continuous labeling or variations in camera settings can influence their misclassification; ID6 is an example of this.

The demonstrated label inconsistencies within the Sewer-ML dataset may compromise the models’ overall capability and robustness [14]. Therefore, relying on a smaller, carefully curated set of labeled data combined with pre-trained weights may help mitigate this risk. This observation highlights the importance of data quality over quantity. Nevertheless, further investigation of the entire dataset is necessary to improve the quality of the dataset by eliminating or relabeling the mislabeled images.

6. Limitations

Despite the promising results, several limitations of this paper should be acknowledged. First, the existence of labeling errors in the Sewer-ML dataset, resulting from inaccurate human annotations and

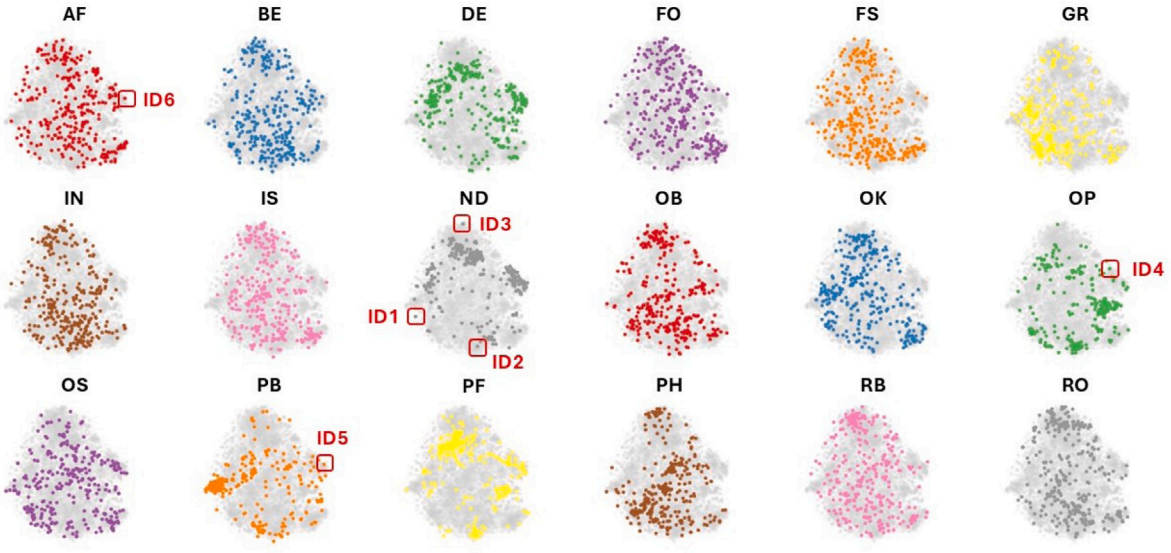
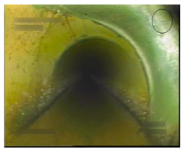
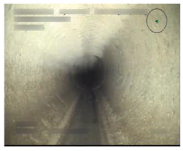
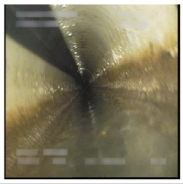


Fig. 5. Visualization of t-SNE clusters of data features of 4k selected images from Sewer-ML for each class. Red boxes with IDs denote the outlier examples within the class that are explained in Table 6.
Note: **AF**: Settled Deposits, **BE**: Attached Deposits, **DE**: Deformation, **FO**: Obstacle, **FS**: Displaced Joint, **GR**: Branch Pipe, **IN**: Infiltration, **IS**: Intruding Sealing Material, **ND**: Non-Defect, **OB**: Surface Damage, **OK**: Connection with Construction Changes, **OP**: Connection with Transition Profile, **OS**: Lateral Reinstatement Cuts, **PB**: Drilled Connection, **PF**: Production Error, **PH**: Chiselled Connection, **RB**: Cracks, Breaks, and Collapses, **RO**: Roots.

Table 6
Details of mislabeling cases and related conditions.

ID	Image file	Photo	Sewer-ML label	True label
ID1	00477562.png		No defect (ND)	Drilled connection (PB)
ID2	00941265.png		No defect (ND)	Connection with transition profile (OP)
ID3	00058381.png		No defect (ND)	Crack (RB)
ID4	00548172.png		Connection with transition profile (OP)	No defect (ND)
ID5	01265565.png		Drilled connection (PB), Obstacle (FO)	No defect (ND)
ID6	00554337.png		Settled Deposit (AF)	No defect (ND)

the automatic aggregation of multi-label entries, may have affected the model's learning and assessment results. Although self-supervised learning can mitigate vulnerability to label noise, increased efforts to improve label quality in datasets remain essential. Second, the Sewer-ML dataset demonstrates significant class imbalance, with some defect classes being underrepresented. While class weighting mitigates this problem to some extent, achieving a balanced trade-off between recall and precision is crucial for reliable performance on such highly imbalanced datasets. Third, this paper relies solely on SwAV as the self-supervised learning strategy; however, alternative SSL approaches (e.g., MoCo [49], DINO [60], SimCLR [41]) and more advanced architectures such as Vision Transformers [61] could be explored to assess whether they yield superior representations for this domain.

Moreover, self-supervised pre-training was conducted using a limited number of unlabeled images due to computational resource constraints. Future research may expand pre-training data to better highlight the different characteristics of the various sewer environments. Expanding the dataset improves the model but also increases the demand for storage, computational capacity, and training duration, which may limit practical feasibility. Furthermore, evaluating the transferability of the SSL pre-trained backbone across different sewer inspection datasets represents an important direction for future work, as such cross-dataset analyses would provide deeper insight into the generalizability and robustness of the learned representations.

7. Conclusion

This paper explored the use of self-supervised learning for multi-label sewer defect classification. A semi-supervised framework is proposed that combines domain-specific pre-training on unlabeled CCTV footage with supervised fine-tuning on a limited labeled subset. This approach aims to reduce dependence on extensive manual annotation while maintaining competitive performance across 17 sewer defect classes. The main findings are summarized as follows:

- **Representation learning over data volume:** Compared to models trained from scratch, models that use pre-trained weights — whether from ImageNet or self-supervised learning — show better overall performance, even when fine-tuned with limited labeled data.
- **Efficiency of SwAV pre-trained weights:** Compared to a fully supervised model, the proposed approach demonstrates that domain-specific representations offer a valid alternative for sewer defect classification. This highlights SwAV's ability to extract meaningful representations from unlabeled data in the pre-training stage.
- **Less dependence on extensive labeled data:** Layered data scarcity in sewer CCTV inspections (acquisition, sharing, and labels) necessitates label-efficient pipelines. The proposed two-stage approach addresses this constraint while maintaining competitive classification performance with only 35,360 labeled images.
- **More unlabeled data and longer pre-training improve self-supervision:** The findings indicate that both increased data diversity and longer pre-training lead to more effective and transferable self-supervised representations. This property of SSL is particularly valuable for developing scalable models that generalize well across tasks while requiring less labeled data.
- **Mislabeling as a challenge for label-dependent models:** The analysis reveals the presence of mislabeled samples in the Sewer-ML dataset, which may limit the effectiveness of models that rely heavily on accurate annotations. This finding underscores the importance of prioritizing data quality over quantity. By leveraging self-supervised learning, resources can be redirected towards carefully curating a smaller, high-quality labeled subset.

In terms of economic benefits, the proposed approach substantially reduces annotation effort and cost. By fine-tuning on 35,360 labeled images rather than the original 1.04 million training set of Sewer-ML, the labeling effort is reduced by approximately 30×(96.6% fewer labels). Since annotation time and cost scale almost linearly with the number of labeled images, this results in commensurate reductions in annotator time and associated costs. In the experiments, this label reduction is achieved through SSL while delivering competitive downstream performance relative to fully supervised baselines. Based on the findings, self-supervised pre-training is a promising technique for a cost-effective alternative to fully supervised learning in sewer inspection.

The results motivate three future research directions: (1) Developing strategies to address label noise and severe class imbalance through noise-aware training, (2) Exploring alternative self-supervised architectures to improve representation transfer, and (3) Assessing cross-dataset transferability across diverse sewer environments. From an operational perspective, given limited inspector capacity and continuously expanding sewer networks, the SSL framework enables the use of archived inspection videos with minimal additional labeling. A practical next step is to deploy the Sewer-ML pre-trained backbone as a generic feature extractor and fine-tune it on a small local dataset to adapt the model to new utilities, camera systems, or sewer environments. This can accelerate inspection workflows and ensure greater consistency in defect reporting. Consequently, the SSL model is not only cost-effective but also capable of addressing real-world operational requirements in sewer asset management.

CRediT authorship contribution statement

Tugba Yildizli: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Tianlong Jia:** Writing – review & editing, Software, Methodology. **Jeroen Langeveld:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Riccardo Taormina:** Writing – review & editing, Supervision, Project administration, Methodology, Data curation, Conceptualization.

Code and data availability

The code repository for this paper is available at https://github.com/tubayildizli/MultiLabel_SewerDefect_SSL. The dataset used in this paper and the model weights can be found at: https://doi.org/10.4121/1c21ce33-715f-4ca0-89fa-c170b30801ff.v2.

Declaration of competing interest

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

Acknowledgments

The first author, Tugba Yildizli, acknowledges the financial support provided by the Turkish Ministry of National Education for the postgraduate scholarship. The utilization of the DelftBlue supercomputers from the Delft High Performance Computing Centre is acknowledged. We initially presented the preliminary results of this paper at the 16th International Conference on Urban Drainage (ICUD'24). The insightful comments from attendees were instrumental in refining the analysis. The work of Tianlong Jia was supported by the China Scholarship Council (No. 202006160032) and the Directorate-General for Public Works and Water Management of The Netherlands (Rijkswaterstaat).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.autcon.2025.106751.

Data availability

The link to GitHub repository and to the data have been shared in the manuscript.

References

[1] D. Meijer, L. Scholten, F. Clemens, A. Knobbe, A defect classification methodology for sewer image sets with convolutional neural networks, Autom. Constr. 104 (2019) 281–298, http://dx.doi.org/10.1016/J.AUTCON.2019.04.013.

[2] J.C. Cheng, M. Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171, http://dx.doi.org/10.1016/J.AUTCON.2018.08.006.

[3] M. Lepot, N. Stanić, F.H. Clemens, A technology for sewer pipe inspection (Part 2): Experimental assessment of a new laser profiler for sewer defect detection and quantification, Autom. Constr. 73 (2017) 1–11, http://dx.doi.org/10.1016/j.autcon.2016.10.010.

[4] M. Wang, S.S. Kumar, J. Cheng, Automated sewer pipe defect tracking in CCTV videos based on defect detection and metric learning, Autom. Constr. 121 (2021) 926–5805, http://dx.doi.org/10.1016/j.autcon.2020.103438.

[5] S. Moradi, T. Zayed, F. Golkhoo, Review on computer aided sewer pipeline defect detection and condition assessment, Infrastructures 4 (2019) 10, http://dx.doi.org/10.3390/INFRASTRUCTURES4010010.

[6] Stichting RIONED, Monitor gemeentelijke watertaken 2024: Werk aan de winkel, Tech. Rep., Stichting RIONED, 2024, https://rioned-web-prod.azurewebsites.net/media/lq2ek3dm/monitor-gemeentelijke-watertaken-2024-webversie-maart-2025.pdf. (Accessed 12 March 2025).

[7] Pipebots, Pervasive sensing of buried pipes, 2019, https://pipebots.ac.uk/. (Accessed 02 May 2025).

[8] A. Hawari, F. Alkadour, M. Elmasry, T. Zayed, A state of the art review on condition assessment models developed for sewer pipelines, Eng. Appl. Artif. Intell. 93 (2020) 103721, http://dx.doi.org/10.1016/J.ENGAPPAI.2020.103721.

[9] Y. Yu, A. Safari, X. Niu, B. Drinkwater, K.V. Horoshenkov, Acoustic and ultrasonic techniques for defect detection and condition monitoring in water and sewerage pipes: A review, Appl. Acoust. 183 (2021) 108282, http://dx.doi.org/10.1016/j.apacoust.2021.108282.

[10] X. Yin, Y. Chen, A. Bouferguene, H. Zaman, M. Al-Hussein, L. Kurach, A deep learning-based framework for an automated defect detection system for sewer pipes, Autom. Constr. 109 (2020) 102967, http://dx.doi.org/10.1016/J.AUTCON.2019.102967.

[11] W. Guo, L. Soibelman, J.H. Garrett, Automated defect detection for sewer pipeline inspection and condition assessment, Autom. Constr. 18 (2009) 587–596, http://dx.doi.org/10.1016/J.AUTCON.2008.12.003.

[12] S.S. Kumar, D.M. Abraham, M.R. Jahanshahi, T. Iseley, J. Starr, Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks, Autom. Constr. 91 (2018) 273–283, http://dx.doi.org/10.1016/J.AUTCON.2018.03.028.

[13] T. Lin, M. Maire, S.J. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft COCO: common objects in context, 2014, CoRR abs/1405.0312, http://arxiv.org/abs/1405.0312.

[14] F. Tscheikner-Gratl, N. Caradot, F. Cherqui, J.P. Leitão, M. Ahmadi, J.G. Langeveld, Y.L. Gat, L. Scholten, B. Roghani, J.P. Rodríguez, M. Lepot, B. Stegeman, A. Heinrichsen, I. Kropp, K. Kerres, M. do Céu Almeida, P.M. Bach, M.M. de Vitry, A.S. Marques, N.E. Simões, P. Rouault, N. Hernandez, A. Torres, C. Werey, B. Rulleau, F. Clemens, Sewer asset management – state of the art and research needs, Urban Water J. 16 (2020) 662–675, http://dx.doi.org/10.1080/1573062X.2020.1713382.

[15] T. Czimmermann, G. Ciuti, M. Milazzo, M. Chiurazzi, S. Roccella, C.M. Oddo, P. Dario, Visual-based defect detection and classification approaches for industrial applications—A SURVEY, Sensors 20 (5) (2020) http://dx.doi.org/10.3390/s20051459.

[16] J. Dirksen, F.H. Clemens, H. Korving, F. Cherqui, P.L. Gauffre, T. Ertl, H. Plihal, K. Müller, C.T. Snaterse, The consistency of visual sewer inspection data, Struct. Infrastruct. Eng. 9 (2013) 214–228, http://dx.doi.org/10.1080/15732479.2010.541265.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, 2015, http://dx.doi.org/10.48550/arXiv.1409.0575.

[18] Z. Situ, S. Teng, W. Feng, Q. Zhong, G. Chen, J. Su, Q. Zhou, A transfer learning-based YOLO network for sewer defect detection in comparison to classic object detection methods, Dev. Built Environ. 15 (2023) 100191, http://dx.doi.org/10.1016/j.dibe.2023.100191.

[19] J. Yin, X. Yin, M. Pan, L. Li, Scalable and transparent automated sewer defect detection using weakly supervised object localization, Autom. Constr. 174 (2025) 106152, http://dx.doi.org/10.1016/J.AUTCON.2025.106152.

[20] Y.C.A.P. Reddy, P. Viswanath, B.E. Reddy, Semi-supervised learning: a brief review, Int. J. Eng. Technol. 7 (2018) 81–85, http://dx.doi.org/10.14419/IJET.V7I1.8.9977.

[21] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners, 2020, http://dx.doi.org/10.48550/arXiv.2006.10029.

[22] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE Trans. Knowl. Data Eng. 35 (2023) 857–876, http://dx.doi.org/10.1109/TKDE.2021.3090866.

[23] X. Zhai, A. Oliver, A. Kolesnikov, L. Beyer, S4L: Self-supervised semi-supervised learning, 2019, http://dx.doi.org/10.48550/arXiv.1905.03670.

[24] J. Huang, X. Yang, F. Zhou, X. Li, B. Zhou, S. Lu, S. Ivashov, I. Giannakis, F. Kong, E. Slob, A deep learning framework based on improved self-supervised learning for ground-penetrating radar tunnel lining inspection, Comput.-Aided Civ. Infrastruct. Eng. 39 (6) (2024) 814–833, http://dx.doi.org/10.1111/mice.13042.

[25] T. Jia, R. de Vries, Z. Kapelan, T.H. van Emmerik, R. Taormina, Detecting floating litter in freshwater bodies with semi-supervised deep learning, Water Res. 266 (2024) 122405, http://dx.doi.org/10.1016/j.watres.2024.122405.

[26] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 4037–4058, http://dx.doi.org/10.1109/TPAMI.2020.2992393.

[27] S.K. Sinha, P.W. Fieguth, Automated detection of cracks in buried concrete pipe images, Autom. Constr. 15 (2006) 58–72, http://dx.doi.org/10.1016/J.AUTCON.2005.02.006.

[28] T.-C. Su, M.-D. Yang, Application of morphological segmentation to leaking defect detection in sewer pipelines, Sensors (Basel, Switzerland) 14 (2014) 8686, http://dx.doi.org/10.3390/S140508686.

[29] D. Ai, G. Jiang, S.-K. Lam, P. He, C. Li, Computer vision framework for crack detection of civil infrastructure—A review, Eng. Appl. Artif. Intell. 117 (2023) 105478, http://dx.doi.org/10.1016/j.engappai.2022.105478.

[30] S.I. Hassan, L.M. Dang, I. Mehmood, S. Im, C. Choi, J. Kang, Y.-S. Park, H. Moon, Underground sewer pipe condition assessment based on convolutional neural networks, Autom. Constr. 106 (2019) 102849, http://dx.doi.org/10.1016/j.autcon.2019.102849.

[31] Q. Xie, D. Li, J. Xu, Z. Yu, J. Wang, Automatic detection and classification of sewer defects via hierarchical deep learning, IEEE Trans. Autom. Sci. Eng. 16 (2019) 1836–1847, http://dx.doi.org/10.1109/TASE.2019.2900170.

[32] D. Li, A. Cong, S. Guo, Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification, Autom. Constr. 101 (2019) 199–208, http://dx.doi.org/10.1016/J.AUTCON.2019.01.017.

[33] J.B. Haurum, T.B. Moeslund, Sewer-ML: A multi-label sewer defect classification dataset and benchmark, 2021, http://dx.doi.org/10.48550/arXiv.2103.10895.

[34] Q. Zhou, Z. Situ, S. Teng, G. Chen, Convolutional neural networks–based model for automated sewer defects detection and classification, J. Water Resour. Plan. Manag. 147 (7) (2021) 04021036, http://dx.doi.org/10.1061/(asce)wr.1943-5452.0001394.

[35] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2016-December, CVPR, IEEE, 2016, pp. 779–788, http://dx.doi.org/10.1109/CVPR.2016.91.

[36] P. Singh, R. Chukkapalli, S. Chaudhari, L. Chen, M. Chen, J. Pan, C. Smuda, J. Cirrone, Shifting to machine supervision: Annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification, Sci. Rep. 14 (1) (2024) 10820, http://dx.doi.org/10.1038/s41598-024-61822-9.

[37] C. Qiu, G. Shao, Z. Zhang, C. Zhou, Y. Hou, E. Zhao, X. Guo, X. Guan, Unsupervised real time and early anomalies detection method for sewer networks systems, IEEE Access 12 (2024) 21698–21709, http://dx.doi.org/10.1109/ACCESS.2024.3359302.

[38] J. Yin, X. Yin, Y. Sun, M. Pan, Bridging the annotation gap: Innovating sewer defects detection with weakly supervised object localization, in: V. Gonzalez-Moret, J. Zhang, B. García de Soto, I. Brilakis (Eds.), Proceedings of the 41st International Symposium on Automation and Robotics in Construction, International Association for Automation and Robotics in Construction (IAARC), Lille, France, 2024, pp. 669–674, http://dx.doi.org/10.22260/ISARC2024/0087.

[39] C. Li, H. Li, K. Chen, Z. Bao, Semi-supervised point cloud semantic segmentation via cross-learning for sewer inspection, Adv. Eng. Inform. 66 (2025) http://dx.doi.org/10.1016/j.aei.2025.103399.

[40] R. Balestriero, Y. LeCun, Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods, 2022, https://arxiv.org/abs/2205.11508.

[41] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, 2020, http://dx.doi.org/10.48550/arXiv.2002.05709.

[42] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, 2021, http://dx.doi.org/10.48550/arXiv.2006.09882.

[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, http://dx.doi.org/10.48550/arXiv.1810.04805.
[44] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, 2021, http://dx.doi.org/10.48550/arXiv.2111.06377.
[45] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) https://api.semanticscholar.org/CorpusID:160025533.
[46] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, M. Norouzi, Big self-supervised models advance medical image classification, 2021, http://dx.doi.org/10.48550/arXiv.2101.05224.
[47] M. Zabin, A.N.B. Kabir, M.K. Kabir, H.J. Choi, J. Uddin, Contrastive self-supervised representation learning framework for metal surface defect detection, J. Big Data 10 (2023) http://dx.doi.org/10.1186/s40537-023-00827-z.
[48] R. Güldenring, L. Nalpantidis, Self-supervised contrastive learning on agricultural images, Comput. Electron. Agric. 191 (2021) 106510, http://dx.doi.org/10.1016/j.compag.2021.106510.
[49] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, 2020, https://arxiv.org/abs/1911.05722.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, http://dx.doi.org/10.48550/arXiv.1512.03385.
[51] A. Jaiswal, A.R. Babu, M.Z. Zadeh, D. Banerjee, F. Makedon, A survey on contrastive self-supervised learning, Technologies 9 (2020) http://dx.doi.org/10.3390/technologies9010002.
[52] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013, http://dx.doi.org/10.48550/arXiv.1306.0895.
[53] T. Jia, A.J. Vallendar, R. de Vries, Z. Kapelan, R. Taormina, Advancing deep learning-based detection of floating litter using a novel open dataset, Front. Water Volume 5 - 2023 (2023) http://dx.doi.org/10.3389/frwa.2023.1298465.
[54] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), Comput. Geosci. 19 (3) (1993) 303–342, http://dx.doi.org/10.1016/0098-3004(93)90090-R.
[55] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605, http://jmlr.org/papers/v9/vandermaaten08a.html.
[56] P. Goyal, Q. Duval, J. Reizenstein, M. Leavitt, M. Xu, B. Lefaudeux, M. Singh, V. Reis, M. Caron, P. Bojanowski, A. Joulin, I. Misra, VISSL, 2021, https://github.com/facebookresearch/vissl.
[57] T. Jia, R. Taormina, R. de Vries, Z. Kapelan, T.H. van Emmerik, P. Vriend, I. Okkerman, A semi-supervised learning-based framework for quantifying litter fluxes in river systems, Water Res. 289 (2026) 124833, http://dx.doi.org/10.1016/J.WATRES.2025.124833.
[58] K. Chen, H. Hu, C. Chen, L. Chen, C. He, An intelligent sewer defect detection method based on convolutional neural network, in: 2018 IEEE International Conference on Information and Automation, ICIA, 2018, pp. 1301–1306, http://dx.doi.org/10.1109/ICInfA.2018.8812445.
[59] J. Myrans, R. Everson, Z. Kapelan, Automated detection of faults in sewers using CCTV image sequences, Autom. Constr. 95 (2018) 64–71, http://dx.doi.org/10.1016/J.AUTCON.2018.08.005.
[60] O. Siméoni, H.V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, P. Bojanowski, DINOv3, 2025, http://arxiv.org/abs/2508.10104.
[61] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021, https://arxiv.org/abs/2010.11929.