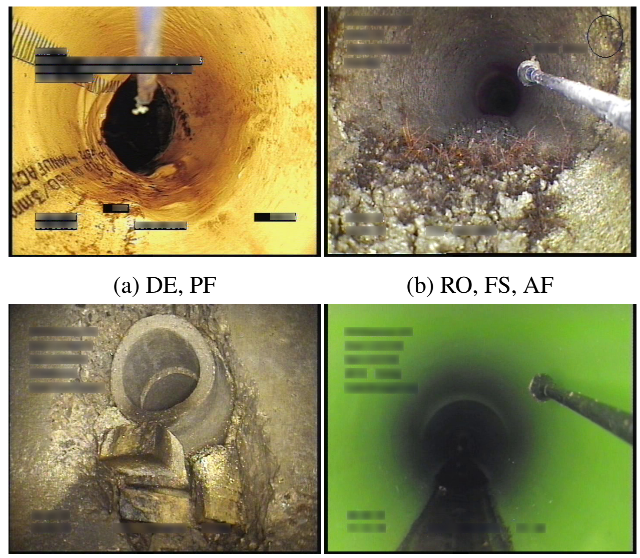
Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark

Joakim Bruslund Haurum   Thomas B. Moeslund
Visual Analysis and Perception (VAP) Laboratory, Aalborg University, Denmark
joha@create.aau.dk, tbm@create.aau.dk

Abstract

Perhaps surprisingly sewerage infrastructure is one of the most costly infrastructures in modern society. Sewer pipes are manually inspected to determine whether the pipes are defective. However, this process is limited by the number of qualified inspectors and the time it takes to inspect a pipe. Automatization of this process is therefore of high interest. So far, the success of computer vision approaches for sewer defect classification has been limited when compared to the success in other fields mainly due to the lack of public datasets. To this end, in this work we present a large novel and publicly available multi-label classification dataset for image-based sewer defect classification called Sewer-ML.

The Sewer-ML dataset consists of 1.3 million images annotated by professional sewer inspectors from three different utility companies across nine years. Together with the dataset, we also present a benchmark algorithm and a novel metric for assessing performance. The benchmark algorithm is a result of evaluating 12 state-of-the-art algorithms, six from the sewer defect classification domain and six from the multi-label classification domain, and combining the best performing algorithms. The novel metric is a class-importance weighted F2 score, $F2_{CIW}$, reflecting the economic impact of each class, used together with the normal pipe F1 score, $F1_{Normal}$. The benchmark algorithm achieves an $F2_{CIW}$ score of 55.11% and $F1_{Normal}$ score of 90.94%, leaving ample room for improvement on the Sewer-ML dataset. The code, models, and dataset are available at the project page http://vap.aau.dk/sewer-ml



(a) DE, PF
(b) RO, FS, AF
(c) PH, RB
(d) Normal pipe

Figure 1: **Sewer-ML data examples.** Images showcasing a subset of the classes and the visual variation in the dataset. The class codes below each image are described in Table 1.

1. Introduction

The sewerage infrastructure is an important but often unnoticed infrastructure. 240 million US citizens are serviced by 1.28 million kilometers of public sewer pipes and 800,000 kilometers of privately owned pipes [3]. In order to maintain public health and sanitation, and avoid *e.g.* unintentional sewer overflows, a 271 billion dollar investment is needed within the next 10 years in order to service an additional 56 million US citizens [3]. Additionally, all of these sewer pipes have to be regularly inspected to avoid sudden pipe collapse or reduced sewer capabilities.

Sewer inspections are currently performed on location by a professional inspector, who simultaneously maneuvers a remote controlled vehicle with a movable camera through the sewer pipe. This is hard and tiresome work, as the inspectors must look at a video feed for a prolonged amount of time. This can lead to flawed inspections, which in the worst case can result in damage to the sewerage infrastructure. Furthermore, the variance in visual appearance within sewer pipes further complicates the task, see Figure 1.

Therefore, the field of automated sewer inspection has been researched by industry and academia for the last three decades, through the development of different robot platforms and specialized algorithms [31]. However, there are at the moment no means to determine which method is the best. Haurum and Moeslund [31] found that there are no open-source benchmark datasets, little to no open-source code, and no agreed upon metrics or evaluation protocol. Instead, many researchers utilize their own datasets from different countries and follow different inspection guides. This leads to stagnation in the field when compared to other computer vision fields and a lack of reproducibility in the automated sewer inspection field.

For these reasons, we present the open-source *Sewer-ML* multi-label defect dataset, containing 1.3 million images annotated by professional sewer inspectors. The dataset is collected from three different Danish water utility companies over a period of nine years. Our contributions are fourfold:

- A publicly available multi-label sewer inspection dataset with 1.3 million annotated images.
- An open-source comparison of state-of-the-art methods using the new dataset.
- A novel, class-importance weighted F2 metric, $\mathrm{F2}_{\mathrm{CIW}}$.
- A benchmark algorithm combining knowledge from sewer defect and multi-label classification domains.

The paper is structured as follows. In Section 2, we review the related works within the multi-label image classification and automated sewer inspection fields. In Section 3, the proposed dataset is introduced and described in detail. In Section 4, we introduce our novel metric, test several state-of-the-art methods on the new dataset, and conduct an ablation study on the obtained results leading to our benchmark algorithm. Finally, in Section 5, we summarize our findings and conclude the paper.

2. Related Works

Multi-label Image Classification. Through the years, the field of multi-label classification has experienced several different trends. Classically, the naive way to approach the problem has been to use an ensemble of binary classifiers and ignore label correlations [88]. This approach has been replaced by methods consisting of a single model incorporating the label correlations into the method itself. These trends have included ranking the label predictions [25, 37, 49], utilizing object localization techniques and attention mechanisms [22, 23, 27, 48, 53, 76, 77, 81, 87, 89], or incorporating a recurrent sub-network to encode label dependencies [8, 9, 48, 72, 76, 82].

Current state-of-the-art networks focus on utilizing the inherent graph nature of the multi-label problem [10, 11, 12, 19, 45, 75, 85], by using the co-occurrence matrix between labels in combination with graph convolutional networks (GCNs) [40]. Chen *et al*. [12] proposed the ML-GCN method, which combines the output of a two-layer GCN with the last feature map of a ResNet-101 [32] network to achieve a well performing multi-label classifier. Wang *et al*. [75] built upon this idea in their KSSNet model. KSSNet improves the performance over ML-GCN by fusing features from a GCN into the final feature map of each residual block in a ResNet-101 model, using a novel lateral connection module. Furthermore, the GCN adjacency matrix is created by combining the label correlation matrix with a label knowledge graph. Lastly, it is also possible to simply take a network which has been proven to work well on a multi-class classification task and instead train it with a relevant loss objective, often the binary cross-entropy loss. This is the case with the recent work of Wu *et al*. [78] who utilized the ResNet-101 architecture and Ridnik *et al*. [66] who proposed a modified variation of the ResNet architecture, called TResNet. The TResNet network has outperformed several models designed for the multi-label task.

The multi-label image classification field has classically worked on smaller datasets such as PASCAL VOC [20], NUS-WIDE [14], and COCO [52], each containing between 5 to 80 thousand training images and 20-80 classes. Therefore, the applied methods have often relied on pretraining the backbone network on ImageNet [67]. However, recently the Tencent-ML [78] and Open Images datasets [44], each containing between approximately 6 and 12 million training images and 11 to 20 thousand classes, have been proposed. These datasets allow for training methods directly on the multi-label task and not pretraining on ImageNet. All of these datasets focus on natural scene images with “common” objects. Furthermore, these datasets are often severely imbalanced, as *e.g*. the class “person” occurs more frequently than the class “sheep”. Therefore, there have been attempts to counteract the data imbalance through weighting the loss objective. This has classically been achieved by utilizing some variant of the inverse class frequency, though custom loss objectives have been proposed specifically for the imbalanced data problem [5, 15, 51, 79].

Automated Sewer Inspection. For several decades there has been an increasing industrial and academic interest in automating the sewer inspection process. This line of research builds heavily upon the general computer vision field, though the current state-of-the-art does not fully utilize the recent advances within computer vision.

Several different types of sensors [18, 54], such as acoustic sensors [35, 38, 39], laser scanners [46, 69], and depth sensors [1, 2, 29, 33], have all been utilized for sewer pipe reconstruction and detecting specific defects but have not seen widespread usage in more generalized tasks. Conversely, image and video based approaches have been utilized to detect, segment, and classify a wide variety of sewer defects. Traditionally, hand-crafted features and small, model based classifiers or heuristic decision rules have been utilized [58, 59, 83]. However, in recent years deep learning based methods have gained traction within the field. This has led to advances within video processing [21, 56, 57, 74], water level estimation [30, 36], defect detection [13, 43, 84], segmentation [61, 63, 73], and classification in multi-class and multi-label settings [7, 28, 42, 47, 55, 58, 80]. For a full review of the field we refer to Haurum and Moeslund [31].

Within sewer defect classification there has been a recent increase in interest, focused on three different system settings: a single end-to-end classifier, a two-stage approach consisting of a binary classifier and a multi-class/label classifier, and an ensemble of binary classifiers. Kumar *et al.* [42] utilized an ensemble of binary classifiers to categorize four types of defects using a small, two-layer CNN trained in a one *vs.* all manner. Hassan *et al.* [28] used AlexNet [41] and Li *et al.* [47] a modified ResNet-18 network [32], trained in an end-to-end manner. Similarly, Meijer *et al.* [55] built upon the work of Kumar *et al.* using a small, three-layer CNN for multi-label defect classification, trained end-to-end. Lastly, Xie *et al.* [80], Chen *et al.* [7], and Myrans *et al.* [58] all used two-stage approaches. Xie *et al.* trained two small, three-layer CNNs, where the first CNN determines whether a defect is present, while the second CNN, a fine-tuned version of the first, classifies the defects. Chen *et al.*, on the other hand, use the lightweight SqueezeNet [34] network for the initial binary defect classification and the deeper InceptionV3 [68] network for predicting the defect class. Myrans *et al.* differ from the other recent methods by using the GIST feature descriptor [60] and two Extra Trees [24] classifiers in sequence.

All prior methods utilize separate private dataset with different classes and class distributions, due to the inherent commercial interest involved in the field [31]. The datasets are typically either balanced such that the number of observations per class is balanced, the number of normal and defect observations are balanced, or the dataset is not balanced but inherently skewed. For example, Meijer *et al.* [55] utilized a dataset consisting of 2.2 million images, but only 17,663 of those images contain defects. In order to counteract this large imbalance, Meijer *et al.* increased the number of defective observations by a factor of five through oversampling. Additionally, there are no common metric nor evaluation protocol, making fair comparison between methods impossible [31]. All of these factors severely hinder the reproducibility and progress within the field.

3. The Sewer-ML Dataset

In this section, we present how the data was collected (Section 3.1), how the multi-label ground truth annotations are obtained (Section 3.2), how the dataset is constructed (Section 3.3), and how we redact information which is present in the images (Section 3.4). Further dataset insights are presented in the supplementary materials.

Table 1: **Sewer inspection classes.** Overview and short description of each annotation class [17] and the class-importance weights (CIW) [16].

Code	Description	CIW
VA	Water Level (in percentages)	0.0310
RB	Cracks, breaks, and collapses	1.0000
OB	Surface damage	0.5518
PF	Production error	0.2896
DE	Deformation	0.1622
FS	Displaced joint	0.6419
IS	Intruding sealing material	0.1847
RO	Roots	0.3559
IN	Infiltration	0.3131
AF	Settled deposits	0.0811
BE	Attached deposits	0.2275
FO	Obstacle	0.2477
GR	Branch pipe	0.0901
PH	Chiseled connection	0.4167
PB	Drilled connection	0.4167
OS	Lateral reinstatement cuts	0.9009
OP	Connection with transition profile	0.3829
OK	Connection with construction changes	0.4396

3.1. Data Collection

A total of 75,618 annotated sewer inspection videos were obtained from three different Danish water utility companies from the period 2011–2019. All videos were annotated by licensed sewer inspectors following a common Danish standard [17] containing 18 specific classes listed in Table 1. According to the inspection standard, each class is given a point score representing the economic consequence of the class, which is determined by professionals involved in the sewer inspection field [16]. We normalize the point scores to the interval $[0, 1]$ by dividing all point scores by the largest one, denoting the new values as the *class-importance weight* (CIW). The collected data span a large variety of materials, shapes, and dimensions from both main and lateral pipes. This leads to a large variety in the available data, reflecting the natural variance observed during actual sewer inspections.

3.2. Multi-Label Ground Truth

The dataset is constructed by extracting a single frame at each class annotation in a sewer inspection video. Each annotation corresponds to a ground truth annotation of a single class at a specific second in the video, with an associated location within the pipe. We obtain the multi-label representation by combining annotations close to each other in the pipe. This is a noisy approach as the camera can rotate in a hemisphere and does not guarantee that all annotations will be visible. For each annotation in an inspection video,

Table 2: **Split between defective and normal observations.** Number of images containing normal and defective observations in the three dataset splits.

Type	Training	Validation	Test	Total
Normal	552,820	68,681	69,221	690,722
Defective	487,309	61,365	60,805	609,479
Total	1,040,129	130,046	130,026	1,300,201

we aggregate the annotated class with all other annotated classes which are up to 0.3 meters earlier in the pipe or 1.0 meters ahead in the pipe. These values have been decided through manual inspection as the position measurement can be noisy. This is necessary in order to include nearby and upcoming, visible classes. Lastly, some entries are noted as *continuous*, which means the class occurs frequently within a specified stretch of the pipe, but are not explicitly annotated at each occurrence. We handle this edge case by adding the continuous class to all other annotated class occurrences within the defined pipe stretch.

The 18 classes are not all instances of pipe defects but can also indicate important information such as a change in pipe shape or material, occurrence of a branch pipe or pipe connections. The VA class is a special class, as it is annotated at the start and end of an inspection video, as well as when the water level changes within a 10% step interval. This means all annotations have an associated water level.

Additionally, we obtain observations of cases with no annotated classes, denoted non-defective (ND), using a set of heuristic rules. First, we apply a one meter buffer zone around each annotated class, such that there is at least two meters between annotated classes before ND images can be extracted. If there are any active continuous class between the annotated classes, no ND images are extracted. Furthermore, we enforce that the inspection vehicle may at maximum move 0.25 m/s, calculated based on the time and distance difference between the two classes. This restriction is based on the maximum speed the inspectors are allowed to move the inspection vehicle during an inspection. Lastly, ND images are only extracted when the inspection vehicle is moving forward through the pipe. This condition is checked using the distance information associated with each annotation. If these conditions are met we can extract ND images. In order to avoid duplicate images of the same pipe area, we extract one ND image per meter uniformly sampled between the two annotated classes. The video timestamps of the ND images are calculated using a constant velocity assumption. Examples from the dataset are shown in Figure 1 and the supplementary materials.

Moreover, the VA class is special, as it is a continuous entity throughout the video. The VA annotations are grouped together with the ND class if there are no other co-occurring labels. This leads to a total of 690,722 images of

Table 3: **Sewer dataset comparison**. A comparison of datasets used for sewer defect classification and the proposed Sewer-ML dataset. We report whether the dataset is publibly available (P), the annotations are multi-label (ML), the number of images with defects (DI), images with normal pipes (NI), annotated classes (C), and the Class Imbalance (CI) for each dataset rounded to the nearest integer.

Dataset	P	ML	DI	NI	C	CI
Ye *et al.* [83]			1,045	0	7	13
Myrans *et al.* [58]			2,260	0	13	102
Chen *et al.* [7]			8,000	10,000	5	5
Li *et al.* [47]			8,455	9,878	7	19
Kumar *et al.* [42]			11,000	1,000	3	4
Meijer *et al.* [55]		✓	17,663	2,184,919	12	12,732
Xie *et al.* [80]			22,800	20,000	7	8
Hassan *et al.* [28]			24,137	0	6	3
Sewer-ML	✓	✓	609,479	690,722	17	123

"normal" pipes with no annotated classes and 609,479 images with one or more annotated classes which we call "defective", resulting in a total of 1,300,201 images. Lastly, we pose the multi-label classification problem as predicting the class labels in Table 1, except for the VA class. This means a normal pipe with no class annotations is the absence of any classes. Therefore, it is an **implicit** class.

3.3. Dataset Construction

We construct the dataset by first splitting the data into three splits: training, validation and test. We randomly select videos until 80% of all annotations are in the training split and the remaining 20% equally split between the validation and test splits. This leads to 60,356 videos for training, 7,692 videos for validation, and 7,570 videos for testing. This way it is ensured that no images from the same pipe are present between splits. These splits lead to a near even split of normal and defective observations, see Table 2.

Looking at the distribution of the class occurrences, as shown in Figure 2, the occurrences are evenly represented in each split, suggesting a similar class distribution in each of the splits. Moreover, it is evident that the constructed dataset is skewed towards a few major classes, such as the "Normal" and "FS" classes. This visually shows the large imbalance in the dataset, representative of the real life distribution of the classes. Unlike prior sewer inspection datasets, we do not manually balance the classes. We quantify the class imbalance (CI) in the dataset by calculating the ratio between the largest and smallest class and compare with the previously used sewer datasets, see Table 3. Meijer *et al.* have a large CI due to sampling every five centimeters, resulting in a large number of normal images. Uniquely, Sewer-ML contains a large number of defect images, which are needed to train discriminative classifiers.

Similarly, it is interesting to see how often several classes are present at the same time. For each split, we plot the distribution of the number of labels in the observations in Figure 2. In this plot we count the normal observations as having zero labels as it is an implicit label. We see that there is an equal number of observations with one or two classes and the number of observations reducing as more classes are present. We quantify this using the label cardinality (LC) using Equation 1 [70] for each split. For these measures, we count the normal pipe observations as having one label.

$$\text{LC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C+1} y_c^{(i)} \tag{1}$$

where N is the number of observations in the split, C is the number of annotated classes, and $y_c^{(i)}$ is the ground truth value for class c in observation i.

We find that across splits, the LC is 1.49-1.50, indicating that on average there are 1.5 labels per observation. We cannot compare this with the LC of the datasets in Table 3, as the datasets and ground truth data are not public.

3.4. Data Anonymization

The raw data provided by the water utility companies have all been post-processed by the inspection software to include metadata and annotation text information on the video itself. In order to avoid including ground truth information in the images and any potential privacy issues, the text has been redacted as shown in Figure 3. Since the overlaid information is not static through the inspection video, due to *e.g.* class codes appearing on screen or pipe material changing, a single redacting mask cannot be used. This leads to a large annotation task, which would be long and tiresome to do manually. Instead, inspired by Borisyuk *et al.* [6], we train a Faster-RCNN [65] model on examples from the overlaid text data. 23,044 videos are used, with one frame extracted per video. The data is split into a training split of 20,739 images and a validation split of 2,305 images. All text information is manually annotated with bounding boxes. The Faster-RCNN backbone is a ResNet-50 FPN [50] pre-trained on ImageNet [67]. We fine-tune the last three residual blocks. As the text data is distinctly different from the data present in the COCO dataset, we use custom anchor boxes. We choose to use three anchor box ratios and five scales, based on the bounding box ratio and area information from the training split. Full details are available in the supplementary materials.

Using the COCO metrics [52], we achieve an mAP@[0.75] of 96.39% and mAP@[0.5:0.95] of 89.10%. The Faster-RCNN model is applied on all 1.3 million images in the dataset, and the detected text is removed by applying a Gaussian blur kernel with a radius of 51 pixels. While this is not a perfect metric score, looking at the detections tells another story. We find that the model detects

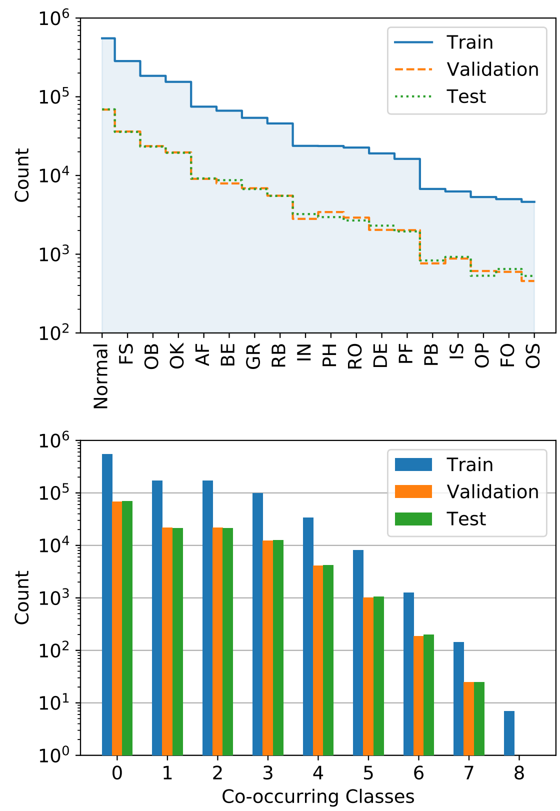
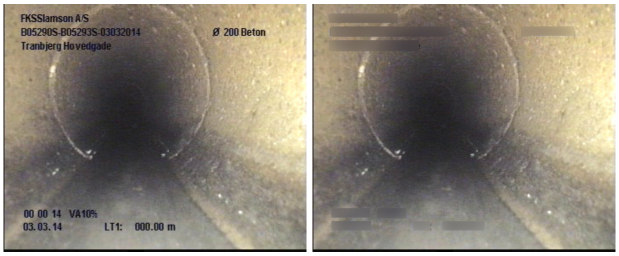


Figure 2: **Dataset label statistics.** The frequencies of the annotated classes and the normal class are shown in the top plot in descending order. The frequencies of the number of labeled classes per split are shown in the bottom plot, where "Normal" pipes have zero labeled classes. Note that the y-axes are log-scaled.

strings of text, annotated with several bounding boxes, as a single bounding box. An example of this is the "Ø 200" in Figure 3. Similarly, text annotated with a single bounding box, are at times detected with several boxes. This leads to a lower metric score even though the redactions are correct. Therefore, we conclude that data leakage is not an issue in the dataset.

4. Benchmark

In this section, we present an approach that can be used as benchmarking for future work on the dataset. To this end, we first select (Section 4.1), train (Section 4.2), and test current state-of-the-art algorithms from the sewer defect classification and the general multi-label classification domains in order to see how they perform on the dataset (Section 4.4), using our novel class-importance weighted F2 metric (Section 4.3). Finally, we conduct an ablation study leading to our benchmark algorithm (Section 4.5), and discuss the per-class performance (Section 4.6).



(a) Before text redaction. (b) After text redaction.

Figure 3: **Effect of the data anonymization process.** By applying our text redaction pipeline the system is capable of detecting and blurring all text information on the images.

4.1. Methods

From the sewer inspection domain we compare the methods proposed by Kumar *et al.* [42], Meijer *et al.* [55], Xie *et al.* [80], Chen *et al.* [7], Hassan *et al.* [28], and Myrans *et al.* [58]. These six methods are chosen to represent the recent advances within sewer defect classification [31]. For the ensemble of binary classifiers and the end-to-end methods, we train using the full dataset. However, for the two-stage classification approach, the first stage is trained on the full dataset to predict the presence of *any* annotated class, while the second stage is trained to predict the classes from a subset of the data containing annotated classes.

From the general multi-label classification domain, we choose four of the current best performing methods on the COCO and VOC datasets [5]. We choose two state-of-the-art graph-based methods, ML-GCN by Chen *et al.* [12] and KSSNet by Wang *et al.* [75], each utilizing a ResNet-101 [32] backbone. Furthermore, we also test the vanilla ResNet-101 model, as used by Wu *et al.* [78], and the TResNet architectures from Ridnik *et al.* [66], where we compare the medium, large, and extra-large versions of the model. All models are trained using an end-to-end approach.

While the normal/defect classification task is intrinsically related to the anomaly detection task, we do not compare with the state-of-the-art anomaly detection methods [62]. This is due to the sewer pipe owners requiring the defect classes in order to correctly manage their assets.

4.2. Training Procedure

Hyperparameters. In order to ensure comparability, we train all networks from scratch using the exact same training procedures. We base our training procedure on the methodology proposed by Goyal *et al.* [26] for efficiently training models on ImageNet. We train each network for 90 epochs, with a batch size of 256 using SGD with momentum. We utilize a learning rate of 0.1, momentum of 0.9, weight decay of 0.0001, and multiply the learning rate by 0.1 at epochs 30, 60, and 80. The ML-GCN and KSSNet networks utilize re-weighted correlation matrices in the GCN subnet, where the hyperparameters stated by the original authors are used. We do not construct a knowledge graph for KSSNet, as the class labels are abbreviations containing little semantic information. We use a one-hot encoding for the initial input to the GCN. For the Myrans *et al.* [58] system, the standard GIST hyperparameters are used and the first and second stage classifiers use 100 and 250 trees, respectively, a maximum depth of 10, and $\log_2(d)$ features when splitting nodes, where d is the dimensionality of the GIST feature vector. We find the hyperparameters through a small grid search, described in the supplementary materials.

Data augmentation. The training data are preprocessed by resizing the images to 224×224, horizontally flipping with a 50% chance, jittering the brightness, contrast, saturation, and hue by $\pm 10\%$ of the original values, and normalizing the data using the training split channel mean and standard deviation. During inference the images are simply resized to 224×224 and normalized. For the InceptionV3 network used by Chen *et al.*, the images are resized to 299×299 [68]. For the GIST features the images are converted to grayscale and resized to 128×128 [58].

Loss objective. We train using the standard binary cross-entropy loss, see Equation 2, which is commonly used in the multi-label image classification domain.

$$L(\mathbf{x}, \mathbf{y}) = \frac{1}{C} \sum_{c}^{C} -[w_c y_c \log(\sigma(x_c)) + (1 - y_c) \log(1 - \sigma(x_c))] \tag{2}$$

where C is the number of annotated classes in the dataset, y_c denotes whether class c is present in the current image, x_c is the raw output of the model for class c, σ is the sigmoid function, and w_c is the weight for class c if it is present in the current image.

As the dataset is imbalanced, we weight each positive class observation by the negative-to-positive class observation ratio, w_c, calculated using Equation 3. This way the loss of minority classes are weighted higher when present in the images, while the loss of majority classes are weighted lower when present. For the InceptionV3 network, a lower weighted loss from the auxiliary classifier is added.

$$w_c = \frac{N - N_c}{N_c} \tag{3}$$

where N is the number of images in the training split, and N_c is the number of images in the split containing class c.

4.3. Metrics

Currently, there is no consensus on how sewer defect classification methods should be evaluated [31]. Commonly, the accuracy is used, but this is a poor metric when working with skewed datasets. Moreover, the metrics do not include domain knowledge. Therefore, we evaluate the model performance using two metrics incorporating domain knowledge, based on the Fβ metric [71],

$$\mathrm{F}\beta = (1+\beta^2)\frac{\mathrm{Prc}\cdot\mathrm{Rcll}}{\beta^2\mathrm{Prc}+\mathrm{Rcll}} \tag{4}$$

where Prc and Rcll are the precision and recall of the classifier, respectively, and β is a weighting of recall, such that the recall β times more important than precision.

When performing sewer inspections, false negatives have a larger economic impact than false positives. This is due to false negatives possibly leading to faulty pipes going unnoticed, whereas a human will verify the predicted classes before a renovation decision is made. Therefore, it is more important to have a high recall than high precision, if both cannot be achieved. To incorporate this domain knowledge into the evaluation, we set $\beta = 2$ when evaluating the annotated classes. This is similar to previous tasks where recall is weighted higher than precision [4, 64, 86]. The per-class F2-scores are averaged using a novel, class-importance weighted F2-score, $\mathrm{F2}_{\mathrm{CIW}}$. The classes are weighted by the associated CIW, see Table 1, as classes with a high CIW will be of larger importance for the pipe owners. $\mathrm{F2}_{\mathrm{CIW}}$ is calculated as shown in Equation 5.

$$\mathrm{F2}_{\mathrm{CIW}} = \frac{\sum_{c=1}^{C}\mathrm{F2}_c\cdot\mathrm{CIW}_c}{\sum_{c=1}^{C}\mathrm{CIW}_c} \tag{5}$$

where CIW_c and $\mathrm{F2}_c$ are the CIW and F2-score for class c, respectively, and C is the number of annotated classes.

However, the normal pipes are not included in the $\mathrm{F2}_{\mathrm{CIW}}$ computation, as normal pipes do not have a CIW. In order to quantify whether the tested methods can handle the absence of classes, and not simply maximize the $\mathrm{F2}_{\mathrm{CIW}}$ score by predicting one or more classes at all times, we use the F1-score for the normal pipes, $\mathrm{F1}_{\mathrm{Normal}}$.

4.4. Model Performances

We report the validation and test split results of each model in Table 4. Unless otherwise noted, a threshold of 0.5 is used to binarize the predictions. For the two-stage approaches, the prediction score from the first stage is used for all classes if the binary classifier detects no classes, and otherwise, the score of the second stage network is used. The results are obtained using the model weights from the epoch with the lowest validation loss. In most cases, the lowest validation loss is obtained after 30-40 epochs, whereafter the networks start overfitting. This indicates that while we utilize a dataset nearly the size of ImageNet, it might not be necessary to train for as long, due to all images being from the same visual domain. We also see that the small CNNs from Kumar *et al*. and Meijer *et al*. immediately diverge during training. This is possibly due to only applying two or three pooling layers before connecting to dense layers, leading to a parameter count of 269 and 135 million, respectively. Comparatively, the small CNN used by Xie *et al*., uses three pooling layers as well as a pixel stride of two in the last two convolutional layers, leading to a parameter count of nine million parameters. Similarly, we observe that the ML-GCN method also diverges immediately, whereas the KSSNet method manages to train. We hypothesize that this is due to the lateral connections in the KSSNet adding stability during training. The loss curves are reported in the supplementary materials. We observe that the methods from the multi-label classification domain are better at classifying the specific classes, with TResNet-L achieving a $\mathrm{F2}_{\mathrm{CIW}}$ test score of 54.75%. However, the simple two-stage approach by Xie *et al*. achieves the highest $\mathrm{F1}_{\mathrm{Normal}}$ score of 90.62%. This indicates that the approach by Xie *et al*. excels at distinguishing whether there are *any* classes, but not which one. The results are not solely due to the two-stage approach. Chen *et al*. also utilize a two-stage approach, but this produces significantly worse results. It is observed that the first stage simply predicts a "defect" in all images, which the later stage cannot properly handle. This is reflected by a low $\mathrm{F1}_{\mathrm{Normal}}$ score. Therefore, it appears there is value in using a small CNN for the first stage.

Table 4: **Performance metrics for each method.** We present the different metrics for each method. The metrics are presented as percentages, and the highest score in each column is denoted in bold. The Kumar [42], Meijer [55] and ML-GCN [12] methods are not shown as they diverged during training. The "Sewer" and "General" identifiers indicate whether the method is from the sewer defect or multi-label classification domains, respectively. The classic multi-label metrics [19] are reported in the supplementary materials.

	Model	Validation		Test	
		$\mathrm{F2}_{\mathrm{CIW}}\uparrow$	$\mathrm{F1}_{\mathrm{Normal}}\uparrow$	$\mathrm{F2}_{\mathrm{CIW}}\uparrow$	$\mathrm{F1}_{\mathrm{Normal}}\uparrow$
Sewer	Xie [80]	48.57	**91.08**	48.34	**90.62**
Sewer	Chen [7]	42.03	3.96	41.74	3.59
Sewer	Hassan [28]	13.14	0.00	12.94	0.00
Sewer	Myrans [58]	4.01	26.03	4.11	27.48
General	ResNet-101 [32]	53.26	79.55	53.21	78.57
General	KSSNet [75]	54.42	80.60	54.55	79.29
General	TResNet-M [66]	53.83	81.23	53.79	79.91
General	TResNet-L [66]	**54.63**	81.22	**54.75**	79.88
General	TResNet-XL [66]	54.42	81.81	54.24	80.42

4.5. Ablation Studies and Benchmark Algorithm

Looking at the results in Table 4, there is merit to both the end-to-end and two-stage approaches. We investigate whether the results can be improved further by combining end-to-end and two-stage methods. In the supplementary materials we report two additional ablation studies focused on getting a better understanding of the two-stage results.

Effect of different second stage classifiers. Based on our results in Table 4, we look into whether combining the general multi-label methods with two-stage approaches

Table 5: **Two-stage classifier permutations.** We evaluate each of the tested multi-label classifiers in a two-stage setup together with the first stage used by Xie *et al.* [80].

	Second Stage	Validation		Test	
		$\mathrm{F2}_{\mathrm{CIW}} \uparrow$	$\mathrm{F1}_{\mathrm{Normal}} \uparrow$	$\mathrm{F2}_{\mathrm{CIW}} \uparrow$	$\mathrm{F1}_{\mathrm{Normal}} \uparrow$
Sewer	Chen [7]	48.67	91.06	48.19	90.60
	Hassan [28]	18.08	91.08	17.89	90.62
	Myrans [58]	27.87	91.08	27.83	90.62
General	ResNet-101 [32]	54.45	91.28	54.01	90.88
	KSSNet [75]	**55.37**	91.30	55.09	**90.95**
	TResNet-M [66]	54.58	91.33	54.32	90.93
	TResNet-L [66]	55.36	91.32	**55.11**	90.94
	TResNet-XL [66]	54.97	**91.37**	54.51	**90.95**

would lead to state-of-the-art performance. Specifically, we combine the first stage of Xie *et al.* with each of the multi-label classifiers in Table 4. The results are shown in Table 5. We observe that by utilizing the first stage of Xie *et al.* both the $\mathrm{F2}_{\mathrm{CIW}}$ and $\mathrm{F1}_{\mathrm{Normal}}$ scores are improved when compared to the best results in Table 4. Moreover, the performance is improved for all tested methods. Specifically, by using the first stage to filter out normal pipes, all general multi-label methods increase their $\mathrm{F2}_{\mathrm{CIW}}$ scores by approximately 0.5-1 percentage points, and the $\mathrm{F1}_{\mathrm{Normal}}$ by up to 10-12 percentage points. For the sewer domain methods their $\mathrm{F2}_{\mathrm{CIW}}$ scores are increased by 5-13 percentage points, and the $\mathrm{F1}_{\mathrm{Normal}}$ by 65-90 percentage points. From these results we can conclude that using a two-stage approach with the binary classifier from Xie *et al.* [80] and the TResNet-L model [66] is the *Benchmark* algorithm on Sewer-ML, with an $\mathrm{F2}_{\mathrm{CIW}}$ score of 55.11% and $\mathrm{F1}_{\mathrm{Normal}}$ score of 90.94%.

4.6. Per-Class Performance

To gain a better understanding of the difficulty of detecting the different defects compared to their economical impact, we compare the F2 score for each defect with the corresponding CIW scores, see Figure 4. We find that each of the classes with a high F2 score exhibit low intra-class and high inter-class variance, as well as more frequently occurring in the dataset. The displaced joint class FS exhibits limited intra-class variance due to limitations in where the defect can occur within the pipe, while being distinct from the other classes. Similarly, the surface damage class OB occurs frequently in the dataset and exhibits high inter-class variance due to the distinct visual appearance of the class.

Contrarily, the lower scoring defect classes exhibit a larger intra-class variance, lower inter-class variance, and are less frequently occurring. The obstacle class FO consists of a wide span of objects, *e.g.* a soda can, a leftover hammer, or another pipe which goes through the main pipe. The RB class exhibits large intra-class variance, due to the class encompassing cracks, breaks, and collapses, and low inter-class variance, due to the similarity in appearance between *e.g.* cracks and the fine roots in the RO class.

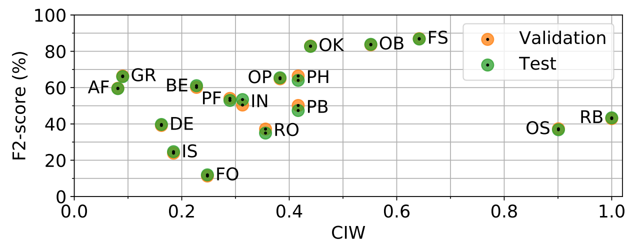


Figure 4: **Per-class performance.** Per-class F2 scores of the *Benchmark* algorithm (TResNet-L + Xie *et al.*), plotted against the corresponding CIW values.

We observe that most of the lower scoring defects do not have a large economic impact. However, the two defects with the highest economic impact, OS and RB, are among the lowest scoring classes. Therefore, in order to improve the performance of the classification system, the detection rate on these two classes should be the main priority.

5. Conclusion

Sewerage infrastructure is a fundamental part of modern society and is continuously expanded. However, current manual inspections are tedious and slow when compared to the immense number of pipes that have to be inspected. Therefore, automated sewer inspection technologies are crucial to ensuring the quality of our sewerage infrastructure. However, current state-of-the-art sewer defect classification methods have not yet adopted recent advances within computer vision. In order to facilitate this transition, we present the first public, multi-label sewer defect classification dataset called Sewer-ML.

Sewer-ML consists of 1.3 million images of a large variety of sewer pipes annotated by professional sewer inspectors. The data is acquired from 75,618 inspection videos conducted over nine years. 12 methods from the sewer defect classification and multi-label classification domains are compared on Sewer-ML. Methods are evaluated using a novel, class-importance weighted F2 score, $\mathrm{F2}_{\mathrm{CIW}}$, which incorporates the economic impact of each class, and the F1 score for pipes with no annotated classes, $\mathrm{F1}_{\mathrm{Normal}}$. We present a benchmark algorithm by combining the best two-stage approach from the sewer domain with the best classifier from the multi-label domain, achieving a state-of-the-art performance with an $\mathrm{F2}_{\mathrm{CIW}}$ of 55.11% and $\mathrm{F1}_{\mathrm{Normal}}$ of 90.94%. The code, data, and trained models are open-sourced in order to lower the barrier of entry and encourage further development within sewer defect classification.

Acknowledgments. This research was funded by Innovation Fund Denmark [grant number 8055-00015A] and is part of the Automated Sewer Inspection Robot (ASIR) project.

References

[1] David Alejo, Fernando Caballero, and Luis Merino. Rgbd-based robot localization in sewer networks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4070–4076, 2017. 2

[2] David Alejo, Gonzalo Mier, Carlos Marques, Fernando Caballero, Luís Merino, and Paulo Alvito. *SIAR: A Ground Robot Solution for Semi-autonomous Inspection of Visitable Sewers*, pages 275–296. Springer International Publishing, Cham, 2020. 2

[3] American Society of Civil Engineers. 2017 infrastructure report card - wastewater. https://www.infrastructurereportcard.org/wp-content/uploads/2017/01/Wastewater-Final.pdf, 2017. Accessed: 08-11-2020. 1

[4] James Atwood, Pallavi Baljekar, Parker Barnes, Anurag Batra, Eric Breck, Peggy Chi, Tulsee Doshi, Julia Elliott, Gursheesh Kour, Akshay Gaur, Yoni Halpern, Henry Jicha, Matthew Long, Jigyasa Saxena, Richa Singh, and D. Sculley. Inclusive images challenge. https://www.kaggle.com/c/inclusive-images-challenge/overview/evaluation, 2018. Accessed: 16-11-2020. 7

[5] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv:2009.14119*, 2020. 2, 6

[6] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 71–79, New York, NY, USA, 2018. Association for Computing Machinery. 5

[7] Kefan Chen, Hong Hu, Chaozhan Chen, Long Chen, and Caiying He. An intelligent sewer defect detection method based on convolutional neural network. In *2018 IEEE International Conference on Information and Automation (ICIA)*, pages 1301–1306, Aug 2018. 3, 4, 6, 7, 8

[8] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Order-free rnn with visual attention for multi-label classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[9] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[10] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 522–531, 2019. 2

[11] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627, 2019. 2

[12] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5172–5181, 2019. 2, 6, 7

[13] Jack C.P. Cheng and Mingzhu Wang. Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques. *Automation in Construction*, 95:155 – 171, 2018. 3

[14] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009. 2

[15] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019. 2

[16] Dansk Vand og Spildevandsforening (DANVA). *Fotomanualen: Beregning af Fysisk Indeks ved TV-inspektion*. Dansk Vand og Spildevandsforening (DANVA), 1 edition, 2005. 3

[17] Dansk Vand og Spildevandsforening (DANVA). *Fotomanualen: TV-inspektion af afløbsledninger*. Dansk Vand og Spildevandsforening (DANVA), 6 edition, 2010. 3

[18] Olga Duran, Kaspar Althoefer, and Lakmal D. Seneviratne. State of the art in sensor technologies for sewer inspection. *IEEE Sensors Journal*, 2(2):73–81, Apr 2002. 2

[19] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 647–657, 2019. 2, 7

[20] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, Sept. 2009. 2

[21] Xu Fang, Wenhao Guo, Qingquan Li, Jiasong Zhu, Zhipeng Chen, Jianwei Yu, Baoding Zhou, and Haokun Yang. Sewer pipeline fault identification using anomaly detection algorithms on video sequences. *IEEE Access*, 8:39574–39586, 2020. 2

[22] Bin-Bin Gao and Hong-Yu Zhou. Multi-label recognition with multi-class attentional regions. *arXiv:2007.01755*, 2020. 2

[23] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018. 2

[24] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, Apr 2006. 3

[25] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv:1312.4894*, 2013. 2

[26] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch

sgd: Training imagenet in 1 hour. *arXiv:1706.02677*, 2017. 6

[27] Hao Guo, Kang Zhang, Xiaochun Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–739, 2019. 2

[28] Syed Ibrahim Hassan, L. Minh Dang, Irfan Mehmood, Suhyeon Im, Changho Choi, Jaemo Kang, Young-Soo Park, and Hyeonjoon Moon. Underground sewer pipe condition assessment based on convolutional neural networks. *Automation in Construction*, 106:102849, 2019. 3, 4, 6, 7, 8

[29] Joakim Bruslund Haurum, Moaaz M. J. Allahham., Mathias S. Lynge., Kasper Schøn Henriksen, Ivan A. Nikolov., and Thomas B. Moeslund. Sewer defect classification using synthetic point clouds. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 891–900. INSTICC, SciTePress, 2021. 2

[30] Joakim Bruslund Haurum, Chris H. Bahnsen, Malte Pedersen, and Thomas B. Moeslund. Water level estimation in sewer pipes using deep convolutional neural networks. *Water*, 12(12), 2020. 3

[31] Joakim Bruslund Haurum and Thomas B. Moeslund. A survey on image-based automation of cctv and sset sewer inspections. *Automation in Construction*, 111:103061, 2020. 1, 2, 3, 6

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Jun 2016. 2, 3, 6, 7, 8

[33] Kasper Schøn Henriksen, Mathias S. Lynge, Mikkel D. B. Jeppesen, Moaaz M. J. Allahham, Ivan A. Nikolov, Joakim Bruslund Haurum, and Thomas B. Moeslund. Generating synthetic point clouds of sewer networks: An initial investigation. In Lucio Tommaso De Paolis and Patrick Bourdot, editors, *Augmented Reality, Virtual Reality, and Computer Graphics*, pages 364–373, Cham, 2020. Springer International Publishing. 2

[34] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*, 2016. 3

[35] Shivprakash Iyer, Sunil K. Sinha, Michael K. Pedrick, and Bernhard R. Tittmann. Evaluation of ultrasonic inspection and imaging systems for concrete pipes. *Automation in Construction*, 22:149 – 164, 2012. Planning Future Cities-Selected papers from the 2010 eCAADe Conference. 2

[36] Hyon Ji, Sung Yoo, Bong-Jae Lee, Dan Koo, and Jeong-Hee Kang. Measurement of wastewater discharge in sewer pipes using image analysis. *Water*, 12(6):1771, Jun 2020. 3

[37] Atsushi Kanehira and Tatsuya Harada. Multi-label ranking from positive and unlabeled data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5138–5146, 2016. 2

[38] Muhammad Safeer Khan and Rajvardhan Patil. Acoustic characterization of pvc sewer pipes for crack detection using frequency domain analysis. In *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–5, 2018. 2

[39] Muhammad Safeer Khan and Rajvardhan Patil. Statistical analysis of acoustic response of pvc pipes for crack detection. In *SoutheastCon 2018*, pages 1–5, 2018. 2

[40] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 3

[42] Srinath S. Kumar, Dulcy M. Abraham, Mohammad R. Jahanshahi, Tom Iseley, and Justin Starr. Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks. *Automation in Construction*, 91:273 – 283, 2018. 3, 4, 6, 7

[43] Srinath Shiv Kumar, Mingzhu Wang, Dulcy M. Abraham, Mohammad R. Jahanshahi, Tom Iseley, and Jack C. P. Cheng. Deep learning-based automated detection of sewer defects in cctv videos. *Journal of Computing in Civil Engineering*, 34(1):04019047, 2020. 3

[44] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, Mar. 2020. 2

[45] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1576–1585, 2018. 2

[46] Mathieu Lepot, Nikola Stanić, and François H.L.R. Clemens. A technology for sewer pipe inspection (part 2): Experimental assessment of a new laser profiler for sewer defect detection and quantification. *Automation in Construction*, 73:1 – 11, 2017. 2

[47] Duanshun Li, Anran Cong, and Shuai Guo. Sewer damage detection from imbalanced cctv inspection data using deep convolutional neural networks with hierarchical classification. *Automation in Construction*, 101:199 – 208, 2019. 3, 4

[48] Peng Li, Peng Chen, Yonghong Xie, and Dezheng Zhang. Bi-modal learning with channel-wise attention for multi-label image classification. *IEEE Access*, 8:9965–9977, 2020. 2

[49] Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1837–1845, 2017. 2

[50] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 5

[51] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2, 5

[53] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *ACM International Conference on Multimedia*, pages 700–708, 2018. 2

[54] Zheng Liu and Yehuda Kleiner. State of the art review of inspection technologies for condition assessment of water pipes. *Measurement*, 46(1):1 – 15, 2013. 2

[55] Dirk Meijer, Lisa Scholten, Francois Clemens, and Arno Knobbe. A defect classification methodology for sewer image sets with convolutional neural networks. *Automation in Construction*, 104:281 – 298, 2019. 3, 4, 6, 7

[56] Saeed Moradi and Tarek Zayed. Real-time defect detection in sewer closed circuit television inspection videos. In *Pipelines 2017*, pages 295–307, 2017. 2

[57] Saeed Moradi, Tarek Zayed, Fuzhan Nasiri, and Farzaneh Golkhoo. Automated anomaly detection and localization in sewer inspection videos using proportional data modeling and deep learning-based text recognition. *Journal of Infrastructure Systems*, 26(3):04020018, 2020. 2

[58] Joshua Myrans, Richard Everson, and Zoran Kapelan. Automated detection of fault types in cctv sewer surveys. *Journal of Hydroinformatics*, 21(1):153–163, Oct 2018. 2, 3, 4, 6, 7, 8

[59] Joshua Myrans, Richard Everson, and Zoran Kapelan. Automated detection of faults in sewers using CCTV image sequences. *Automation in Construction*, 95:64–71, Nov. 2018. 2

[60] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001. 3

[61] Gang Pan, Yaoxian Zheng, Shuai Guo, and Yaozhi Lv. Automatic sewer pipe defect semantic segmentation based on improved u-net. *Automation in Construction*, 119:103383, 2020. 3

[62] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2), Mar. 2021. 6

[63] Claudio Piciarelli, Danilo Avola, Daniele Pannone, and Gian Luca Foresti. A vision-based system for internal pipeline inspection. *IEEE Transactions on Industrial Informatics*, 15(6):3289–3299, 2019. 3

[64] Planet. Planet: Understanding the amazon from space. https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/overview/evaluation, 2017. Accessed: 16-11-2020. 7

[65] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 5

[66] Tal Ridnik, Hussam Lawen, Asaf Noy, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. *arXiv:2003.13630*, 2020. 2, 6, 7, 8

[67] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 5

[68] Christan Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 3, 6

[69] Abbasali Dehghan Tezerjani, Mehran Mehrandezh, and Raman Paranjape. Defect detection in pipes using a mobile laser-optics technology and digital geometry. *MATEC Web of Conferences*, 32:06006, 2015. 2

[70] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. *Mining Multi-label Data*, pages 667–685. Springer US, Boston, MA, 2010. 5

[71] Cornelis J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, USA, 2nd edition, 1979. 7

[72] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294, 2016. 2

[73] Mingzhu Wang and Jack C. P. Cheng. A unified convolutional neural network integrated with conditional random field for pipe defect segmentation. *Computer-Aided Civil and Infrastructure Engineering*, 35(2):162–177, 2020. 3

[74] Mingzhu Wang, Srinath Shiv Kumar, and Jack C.P. Cheng. Automated sewer pipe defect tracking in cctv videos based on defect detection and metric learning. *Automation in Construction*, 121:103438, 2021. 2

[75] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12265–12272, Apr. 2020. 2, 6, 7, 8

[76] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 464–472, 2017. 2

[77] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, 2016. 2

[78] Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. Tencent ml-images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7:172683–172693, 2019. 2, 6

[79] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[80] Qian Xie, Dawei Li, Jinxuan Xu, Zhenghao Yu, and Jun Wang. Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE Transactions on Automation Science and Engineering*, pages 1–12, 2019. 3, 4, 6, 7, 8

[81] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–288, 2016. 2

[82] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartłomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13437–13446, 2020. 2

[83] Xiangyang Ye, Jian'e Zuo, Ruohan Li, Yajiao Wang, Lili Gan, Zhonghan Yu, and Xiaoqing Hu. Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern chinese city. *Frontiers of Environmental Science & Engineering*, 13(2), Jan. 2019. 2, 4

[84] Xianfei Yin, Yuan Chen, Ahmed Bouferguene, Hamid Zaman, Mohamed Al-Hussein, and Luke Kurach. A deep learning-based framework for an automated defect detection system for sewer pipes. *Automation in Construction*, 109:102967, 2020. 3

[85] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12709–12716, Apr. 2020. 2

[86] Chenyang Zhang, Grace Vesom, Jennie Choi, Maria Kessler, and Tallon Loic. imet collection 2019 - fgvc6. https://www.kaggle.com/c/imet-2019-fgvc6/overview/evaluation, 2019. Accessed: 16-11-2020. 7

[87] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. Multilabel image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, 20(10):2801–2813, 2018. 2

[88] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014. 2

[89] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2027–2036, 2017. 2