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. Automation 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, \( F_2^{\text{CIW}} \), reflecting the economic impact of each class, used together with the normal pipe F1 score, \( F_1^{\text{Normal}} \). The benchmark algorithm achieves an \( F_2^{\text{CIW}} \) score of 55.11% and \( F_1^{\text{Normal}} \) 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.

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, F2CIW.
- 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. KSS-Net 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].
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.

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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>552,820</td>
<td>68,681</td>
<td>69,221</td>
<td>690,722</td>
</tr>
<tr>
<td>Defective</td>
<td>487,309</td>
<td>61,365</td>
<td>60,805</td>
<td>609,479</td>
</tr>
<tr>
<td>Total</td>
<td>1,040,129</td>
<td>130,046</td>
<td>130,026</td>
<td>1,300,201</td>
</tr>
</tbody>
</table>

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 “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.

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

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 over-
laided 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 train-
ing 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 im-
gages in the dataset, and the detected text is removed by ap-
plying a Gaussian blur kernel with a radius of 51 pixels.
While this is not a perfect metric score, looking at the de-
tections tells another story. We find that the model detects
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 de-
fect 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 dis-
cuss the per-class performance (Section 4.6).
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 $224 \times 224$, 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 $224 \times 224$ and normalized. For the InceptionV3 network used by Chen et al., the images are resized to $299 \times 299$ [68]. For the GIST features the images are converted to grayscale and resized to $128 \times 128$ [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.

$$L(x, y) = \frac{1}{C} \sum_c w_c y_c \log(\sigma(x_c)) + (1 - y_c) \log(1 - \sigma(x_c))$$

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 $c$, $\sigma$ 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.

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

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 $F_\beta$ metric \cite{f2}:
\[
F_\beta = \frac{1 + \beta^2 \cdot \text{Prc} \cdot \text{Rcll}}{\beta^2 \cdot \text{Prc} + \text{Rcll}}
\]  
where Prc and Rcll are the precision and recall of the classifier, respectively, and $\beta$ is a weighting of recall, such that the recall $\beta$ 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 \cite{f2, f5, f6}. The per-class F2-scores are averaged using a novel, class-importance weighted F2-score, $F_{\text{CIW}}$. 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. $F_{\text{CIW}}$ is calculated as shown in Equation 5.
\[
F_{\text{CIW}} = \frac{\sum_{c=1}^{C} F_{2c} \cdot \text{CIW}_c}{\sum_{c=1}^{C} \text{CIW}_c}
\]  
where CIW$_c$ and $F_{2c}$ 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 $F_{\text{CIW}}$ 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 $F_{\text{CIW}}$ score by predicting one or more classes at all times, we use the F1-score for the normal pipes, $F_{1\text{Normal}}$.

### 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 $F_{\text{CIW}}$ test score of 54.75%. However, the simple two-stage approach by Xie et al. achieves the highest $F_{1\text{Normal}}$ 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 $F_{1\text{Normal}}$ score. Therefore, it appears there is value in using a small CNN for the first stage.

### 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...
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 $F_2^{CIW}$ and $F_{1\text{Normal}}$ 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 $F_2^{CIW}$ scores by approximately 0.5-1 percentage points, and the $F_{1\text{Normal}}$ by up to 10-12 percentage points. For the sewer domain methods their $F_2^{CIW}$ scores are increased by 5-13 percentage points, and the $F_{1\text{Normal}}$ 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 $F_2^{CIW}$ score of 55.11% and $F_{1\text{Normal}}$ 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 $F_2$ score for each defect with the corresponding CIW scores, see Figure 4. We find that each of the classes with a high $F_2$ 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.

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 $F_2$ score, $F_2^{CIW}$, which incorporates the economic impact of each class, and the $F_1$ score for pipes with no annotated classes, $F_{1\text{Normal}}$. 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 $F_2^{CIW}$ of 55.11% and $F_{1\text{Normal}}$ 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


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


