Dataset Optimisation for Enhanced Automated Sewer Damage Detection

Kamil Altinay1 Zehao Ye1, and Jelena Ninic1

1 University of Birmingham, Birmingham B15 2TT, UK

**Abstract.** Starting from the hypothesis “less is more” this research proposes an dataset optimisation for automated sewer damage detection method employing automated data filtering and active learning to reduce the quantity and maximize the utility of existing datasets. The study utilizes the Sewer-ML dataset and proposes data optimization strategy that reduces the data to high-quality/relevance data by applying CLIP-IQA (Contrastive Language-Image Pre-training - Image Quality Assessment) to filter images based on their quality, as well as prediction reliability and label checks for automated filtering of high-quality data. Moreover, and active learning query strategy is employed to select the most informative samples from the data pool. A damage detection model, based on Residual Networks (ResNets), is trained on the refined dataset, achieving a significant increase in performance. The Sewer-ML dataset contains more than 1.3 million images; however, many of these are low-quality or misrepresented, which can negatively impact training outcomes. Our results demonstrate that using less than 3% of the dataset that carefully selected through automated filtering, can reach 4% higher F2ciw results than models trained on the full dataset, validating the effectiveness of the proposed method.

**Keywords:** Sewer Damage Detection, Data Recycling, Computer Vision, Image Quality Assessment.

1. Introduction

The maintenance of sewer systems is a cornerstone of urban infrastructure sustainability, yet traditional inspection methods remain largely reactive and labour-intensive [5]. Manual analysis of CCTV or SSET footage is time-consuming, costly, and prone to human error, delaying the detection of critical defects such as cracks, deformations, and blockages. These delays can escalate risks of catastrophic failures, leading to environmental contamination, public health crises, and exorbitant repair costs. For instance, U.S. utilities spent over $3 billion in 2019 alone to replace nearly 4,700 miles of deteriorating pipelines [9]. Moreover, EU loses 24% of its water consumption due to pipeline leakage [10]. While automated systems powered by machine learning (ML) and computer vision (CV) offer transformative potential, their adoption is hindered by the scarcity of high-quality datasets and the prohibitive costs of data annotation.

Particularly, deep convolutional neural networks (CNNs) [6] and residual networks (ResNets) [2], have enabled automated defect detection. The reality of sewage inspection data, which is frequently noisy, unbalanced, and fragmented, is incompatible with these models' usual requirement for large, precisely labelled datasets. For example, the Sewer-ML dataset, despite containing 1.3 million images, suffers from label inconsistencies, low-quality frames, and severe class imbalance [1], where critical defects (e.g., structural collapses) are vastly underrepresented compared to more frequent but less severe issues.

This research addresses these challenges through a data-centric framework that prioritizes dataset quality over quantity. We integrate CLIP-IQA, a no-reference image quality assessment method, to filter out low-quality images (e.g., motion-blurred or poorly lit frames) [4]. To further mitigate label noise, we implement reliability checks via a pre-trained model in two steps: (1) computing absolute errors between predictions and ground truth to discard ambiguous samples, and (2) retaining only data where the model’s highest-confidence prediction matches the annotated class. Finaly, based on estimated image quality, an active learning strategy has been employed to iteratively design data, leveraging high quality images to establish the optimised size, leading to improved model training and predictability. Our experiments demonstrate that this approach improves model robustness while drastically reducing dataset size.

1. Methodology

This study proposes a systematic framework for training a robust sewer defect classification model by prioritizing dataset quality and label reliability. Our methodology integrates reliability-driven data refinement using a pre-trained model, rigorous quality filtering, an active learning framework to prioritize informative samples and class-aware evaluation to align outcomes with real-world economic priorities.

The proposed framework operates through three sequential phases designed to maximize data utility while minimizing noise and bias:

1. **High-quality data filter:** The pipeline begins by systematically eliminating low-resolution, motion-blurred, or poorly lit images using CLIP-IQA, a vision-language model that quantifies perceptual quality through text-guided semantic comparisons.
2. **Label reliability filter:** A pre-trained model computes prediction errors to identify and discard samples with high discrepancies, mitigating label noise. Subsequently, a categorical consistency check retains only samples where the model’s highest-confidence prediction aligns with at least one annotated label.
3. **Optimizing datasets through active learning:** To counteract persistent class imbalance an active learning loop iteratively selects the most informative samples. Starting with a small high-quality subset that created earlier steps, the method prioritizes underrepresented classes exhibiting low F₂ scores, dynamically augmenting the training data.

By prioritizing dataset quality over quantity, this work aims to demonstrate that strategic data refinement can achieve competitive accuracy with a fraction of the original data, leading to more suitable model learning and aligning with the "less is more" hypothesis.

* 1. Dataset and Preprocessing

The study uses the Sewer-ML dataset, a publicly available multi-label sewer defect classification dataset comprising over 1.3 million images across 17 defect categories. The dataset, derived from video footage collected by three Danish water companies (2011–2019), is structured to include 53% normal (non-defect) images, ensuring balanced representation of defect and non-defect scenarios (see Table 1).

**Table 1.** Dataset splits for normal and defective classes across training, validation, and test sets, along with overall totals.

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

The training split exhibits severe class imbalance (Table 2), with defect frequencies ranging from 4,625 samples for rare defects like lateral reinstatement cuts (OS) to 283,983 samples for prevalent defects like displaced joints (FS). To mitigate bias toward high-frequency classes, Class Importance Weights (CIW) are assigned based on a point system defined by sewer inspection standards [3], which quantifies defects’ economic and structural consequences. For instance, cracks, breaks, and collapses (RB), which require urgent repair to prevent catastrophic failures, receive the highest weight (CIW = 1.0) despite moderate representation (45,821 samples). Conversely, settled deposits (AF), though frequent (74,856 samples), are assigned a low CIW (0.081) due to their minimal operational impact. Rare defects like chiselled connections (PH) (23,685 samples) retain moderate weights (CIW = 0.4167), reflecting their potential to escalate into costly failures under the scoring framework.

**Table 2.** Categories, including label codes, descriptions, CIW values, and corresponding image counts for training split.

|  |  |  |  |
| --- | --- | --- | --- |
| Code | Description | CIW | Counts |
| RB | Cracks, breaks, and collapses | 1.0000 | 45,821 |
| OB | Surface damage | 0.5518 | 184,379 |
| PF | Production error | 0.2896 | 16,254 |
| DE | Deformation | 0.1622 | 19,084 |
| FS | Displaced joint | 0.6419 | 283,983 |
| IS | Intruding sealing material | 0.1847 | 6,271 |
| RO | Roots | 0.3559 | 22,637 |
| IN | Infiltration | 0.3131 | 74,856 |
| AF | Settled deposits | 0.0811 | 74,856 |
| BE | Attached deposits | 0.2275 | 66,499 |
| FO | Obstacle | 0.2477 | 5,010 |
| GR | Branch pipe | 0.0901 | 53,986 |
| PH | Chiselled connection | 0.4167 | 23,685 |
| PB | Drilled connection | 0.4167 | 6,746 |
| OS | Lateral reinstatement cuts | 0.9009 | 4,625 |
| OP | Connection with transition profile | 0.3829 | 5,325 |
| OK | Connection with construction changes | 0.4396 | 154,624 |

Sewer inspection environments introduce challenges such as variable lighting, occlusions, and motion blur from CCTV equipment movement. These factors cause significant quality variability in the Sewer-ML dataset, ranging from sharp, detailed images to poorly resolved or overexposed frames. For example, motion blur can obscure fine cracks, while uneven lighting may cast shadows resembling structural anomalies.

* 1. Image Quality Assessment

Image Quality Assessment (IQA) evaluates an image’s perceptual quality through objective metrics or subjective analysis, focusing on clarity, sharpness, and usability for defect detection. This study employs CLIP-IQA, a method leveraging contrastive language-image pretraining to assess quality via text-guided semantic comparisons. For sewer defect analysis, CLIP-IQA uses two text prompts:

“Good photo”: Characterized by sharpness, balanced lighting, and minimal noise.

“Bad photo”: Exhibiting blur, noise, or poor exposure.

The input image is projected into CLIP’s joint embedding space, and its similarity to these prompts is quantified. A quality score (0–1) is computed based on alignment with the “good photo” prompt, prioritizing critical attributes like sharpness, noise reduction, and lighting consistency. For example, Fig. 1 contrasts a high-quality image (score = 0.9479) with sharp defect visibility against a low-quality example (score = 0.0366) obscured by motion blur and noise, demonstrating CLIP-IQA’s efficacy in filtering unusable samples.



 (a) (b)

**Fig. 1** Comparison of Image qualities from the Sewer-ML dataset processed with CLIP-IQA (a-“Bad photo”, b- “good photo”).

Using an illustrative threshold of 0.1, CLIP-IQA filtering partitioned the Sewer-ML dataset’s 1,040,129 training images into 498,929 high-quality samples (score ≥0.1) and 541,200 low-quality samples (score <0.1), demonstrating that most raw images exhibit substandard quality.

* 1. Data Selection via Reliability Checks

To refine the dataset and ensure robust model training, two complementary reliability-based methods such as Absolute Error Filtering and Label Consistency Verification were applied using a pre-trained defect classification model that trained on the full Sewer-ML dataset. These steps prioritize samples where the model demonstrates high confidence and alignment with ground truth annotations.

Absolute Error Filtering quantifies the discrepancy between the pre-trained model’s predictions and ground truth labels. The absolute error () for each sample is calculated as:

|  |  |
| --- | --- |
|  | (1) |

where - predicted probability.

 – ground truth label.

Samples with errors below a predefined threshold are retained, as lower errors indicate strong agreement between predictions and annotations. This step reduces noise caused by mislabelled or ambiguous samples, ensuring the model trains on reliable examples.

Label Consistency Verification adds categorical validation by ensuring the model’s highest confidence predicted class aligns with at least one annotated ground truth label. For example, if the model predicts “cracks, breaks, and collapses” (RB) with the highest probability for an image labelled [RB, OB], the sample is retained; mismatches (e.g., predicting PH for an [RB, OB] label) are discarded. This addresses label noise by ensuring training only on data where predictions align with human judgment.

By combining continuous error metrics with categorical validation, this approach ensures the final dataset is both visually clear (via CLIP-IQA) and semantically reliable, directly addressing challenges of noise, imbalance, and annotation subjectivity in sewer defect detection.

* 1. Active learning for dataset optimization

Active learning is a machine learning technique that aims to obtain high scores with fewer data samples by progressively changing the training set [7]. In active learning, the entire training dataset is split into two parts: an initial labelled training set and a large pool of remaining data. A supervised learning model is first trained on the initial set. Then, instead of using all remaining data an active class selection query strategy is used to select the most informative samples from the data pool. The classes that require additional samples to improve results were identified using active class selection query strategy [8]. After the subset selected for initial training is removed from the original training set, the remaining images are transferred to a data pool, sorted by their IQA scores and grouped by class labels. The active class selection strategy addresses performance disparities between defect classes by iteratively augmenting the training set with high-quality samples from underperforming categories. For each defect class *c,* the performance gap Δc is computed as the difference in F2 scores between the full-dataset model (Mfull​) and the subset-trained model (Msubset​) (Eq. 2). Classes with Δc>0 means underperformance and these classes are prioritized. From the excluded data pool, the top 200 samples per class, ranked by CLIP-IQA scores to ensure diagnostic utility, are added to the training subset as given in Eq. 3. This number balances reinforcement of critical classes (limiting additions to ~10% of the initial subset size, 28,580 images) while avoiding dilution of data quality or training instability. The model is retrained iteratively on the augmented dataset until performance gaps narrow.

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |

Adding data samples to the training subset and having the model learn them initially works well, but eventually the quality of the additional photos decreases significantly thus the evaluation scores begin to decline. Active learning is therefore stopped if there is no clear improvement in the F2 score in the assessment conducted at the conclusion of the iterations.

The use of active learning allows us to first create a model based on high-quality data and then select the images that are to be included in a high-quality subset based on the model is trained initially. This allows us to fully automate the process of establishing the most valuable data from an already-existing dataset.

 

**Fig. 2** Workflow of the proposed data optimization framework

* 1. Model Training and Evaluation

The refined dataset trains a defect classification model using a ResNet-101 architecture pretrained on ImageNet, fine-tuned to prioritize critical sewer defects. The model is evaluated using class-weighted metrics that reflect the economic severity of defects, including precision, recall, and the Class-Importance Weighted F₂ Score (F2CIW). Fβ Score emphasizes recall over precision and it is defined as:

|  |  |
| --- | --- |
|  | (4) |

Where P - Precision.

 R - Recall.

F2 CIW aggregates per-class F₂ scores weighted by their economic impact (CIW). It is computed as shown in Equation (3):

|  |  |
| --- | --- |
|  | (5) |

Where *CIW* - Class Importance Weight.

 *C*– total number of classes.

1. Results

We applied the proposed strategy to demonstrate the effectiveness of active learning in improving the quality of existing datasets and reducing the need for extensive new data collection. By applying CLIP- IQA with 0.4 threshold value, the training split of the dataset was reduced from around 1 million images to a high-quality subset of less than 50,000 images.

Reliability checks with 5% error rate further reduced the dataset to less than 30,000 images retaining just 3.2% of the original training data. The final refined dataset retained only samples with strong agreement between pre-trained model predictions and ground truth labels, ensuring high semantic and visual reliability for robust model training.

The randomly selected samples in Fig 3 exemplify the efficacy of the data refinement pipeline in retaining high-quality, diagnostically viable sewer inspection images. Both samples exhibit optimal brightness, resolution, and label clarity, critical for accurate defect detection. In Fig 3 (a) the branch pipe (GR) defect is unambiguously visible, with sharp edges and no occlusions, while Fig 3 (b) clearly displays a displaced joint (FS), characterized by misaligned pipe segments and no motion blur. These images validate the success of CLIP-IQA and reliability checks in filtering out low-quality or ambiguous data, ensuring the training set prioritizes visually and semantically reliable samples.



(a) (b)

**Fig. 3.** The randomly selected exemplify the high-quality data retained after rigorous refinement. Both images exhibit optimal brightness, sharp resolution, and unambiguous defect visibility, free from motion blur, occlusions, or lighting artifacts.

 We have tested the model performance on two datasets: (i) Subset I of high-quality images (28,580) and ii) Subset II generated by active learning and have compared the performance with full data set. As illustrated in Fig 4, the (i) subset I exhibited significant class imbalance, with certain defect labels such as FS (Displaced joint) dominating the distribution, while others like OS (Lateral reinstatement cuts) were underrepresented or only appeared alongside other labels. Naturally, certain labels are not sufficiently learnt because of this unbalanced distribution. To ensure that the model learns these classes as well, the active learning approach attempts to include classes that lag in result metrics, like low F2, in the training data.



**Fig. 4.** The bar chart illustrates the distribution of defect labels in the refined validation set, with the y-axis representing the number of images for each label. While some labels, such as GR and PB, are highly represented, others have significantly fewer instances.

Fig 5 presents a comparative analysis of ResNet101 models trained on progressively reduced subsets of the dataset, all evaluated on a fixed validation set to ensure methodological consistency. The baseline model, trained on the original the Sewer-ML dataset, serves as the reference point. To investigate the impact of data quality and quantity on performance, two additional ResNet101 variants were trained. The first variant utilized a reduced subset of 28,580 images (3.2% of the original training dataset), filtered by applying a CLIP-IQA threshold of 0.4 and 5% error threshold. The second variant further constrained the training using active learning strategy. In comparison to the reference results, it was determined that certain labels with low outcomes required additional training instances due to an insufficient quantity for effective generalisation. Therefore, additions were made to the training subset at certain times throughout the new training session. The training subset was expanded after iterations during active learning, as seen in Fig 8, and the final training set comprised 33,031 samples. The retained validation set ensures comparability across models, enabling a systematic evaluation of robustness under varying training conditions.



**Fig. 5.** Bar chart comparing model performance in terms of Accuracy and F2 Score.

Notably, models trained on rigorously created subsets demonstrated comparable or superior performance to the baseline model, trained on over one million images, across critical evaluation metrics. The Fig. 6 clearly demonstrates the effectiveness of the proposed methodology in enhancing model performance across a wide range of defect categories. Notably, the Active Learning approach consistently performs on par with or better than both the Full set and Subset baselines in most classes, despite using significantly fewer training samples. This confirms the core premise of the study that prioritizing data quality through targeted filtering and intelligent sample selection can outperform supervised training on larger but noisier datasets. Particularly in challenging or underrepresented classes such as PF, IS, and OS, Active Learning yields marked improvements, highlighting its strength in addressing noisy or mislabelled samples and imbalance. Moreover, for classes like Normal, OB, OK, and FS, the high F2 scores under all three setups suggest strong generalization, but the fact that Active Learning nearly matches or slightly exceeds the Full set further underscores its data efficiency. Overall, the graph validates the “less is more” hypothesis, showing that a smaller, carefully curated dataset can support robust, economically meaningful defect classification performance. This outcome contradicts conventional assumptions about data volume requirements in deep learning and highlights the potential efficacy of systematic data curation or targeted augmentation strategies in achieving robust model performance. Such approaches not only reduce computational resource demands and training duration but also suggest that prioritization of data quality and representativeness may supersede sheer dataset scale in specific domain applications. While training a model on the original dataset with a single GPU per epoch takes 15–16 hours, training the model with the same structure on generated subset for a single epoch only takes 20 minutes.

.

**Fig. 7.** Bar chart comparing F2 performance across defect labels and normal labels which means non-defect.



**Fig. 8.** Number of images used during training (bar chart) and corresponding F2 scores (line chart) across iterations. The chart illustrates the performance improvement as the dataset size increases.

Fig 8 illustrates the iterative convergence of active learning improving model performance and reducing dataset size. The bar chart shows the progressive expansion of the training set from 28,580 to 38,847 samples, across 8 iterations, while the line chart tracks the corresponding improvement in F₂ scores. As targeted samples from underrepresented classes such as OS, PH are added, the F₂ score increases, demonstrating that strategic data augmentation enhances model accuracy despite minimal growth in dataset size. This inverse relationship between data efficiency and performance underscores the effectiveness of active learning in prioritizing high-impact samples over raw data volume.

1. Conclusions and Future Work

This study proposes a data-centric framework for sewer damage detection that challenges the conventional "bigger is better" paradigm in machine learning. By integrating automated image quality assessment with CLIP-IQA, reliability-driven filtering using absolute error thresholds and label consistency checks, and active learning, the method optimizes existing datasets to retain only high-quality, diagnostically viable samples. The refined dataset, comprising less than 3% of the original Sewer-ML training data, enabled a ResNet-based model to outperform a same structure model trained on the full dataset while drastically reducing computational costs.

The key innovation lies in the systematic fusion of vision-language pretraining (CLIP-IQA) with reliability verification for domain-specific data curation. Unlike prior work focusing solely on model architecture improvements, this framework addresses foundational data quality challenges in sewer inspection: label noise, class imbalance, and visual degradation. The integration of active learning further automates the selection of informative samples, enabling resource-efficient training without compromising performance.

Remarkably, training a model on just 3.2% of the original data (33,031images) achieved a 4% higher F₂CIW score compared to the full-dataset baseline, alongside significant improvements in critical defect detection and more than 95% reduction in training time per epoch. These results validate the “less is more” hypothesis, demonstrating that strategic data refinement can enhance model robustness while minimizing computational costs. However, the framework’s dependency on empirically set thresholds and a pre-trained model introduces potential biases, and its generalizability to diverse sewer systems remains untested.

While the proposed framework demonstrates promising results in reducing training costs and improving defect detection accuracy through strategic data refinement, several limitations should be acknowledged. First, the reliance on empirically defined thresholds for error filtering and label consistency introduces potential subjectivity and may limit adaptability across diverse datasets. Additionally, the framework depends on a vision-language pre-trained CLIP-IQA model, which, although effective, may not generalize well to sewer-specific image features or conditions not represented in its training corpus.

Future work will explore automated data selection using reinforcement learning to optimize real-time data pruning. Such systems could minimize redundant data collection and labelling expenses without compromising performance. Further validation will test these methods on other inspection datasets. These strategies aim to balance efficiency and robustness, particularly for infrastructure monitoring in resource-limited settings Additionally, current limitations such as the reliance on empirically set thresholds, potential domain bias from pre-trained models will be addressed in future studies.

Acknowledgements

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham. The computations described in this paper were also performed using the University of Birmingham’s BlueBEAR (http://www.birmingham.ac.uk/bear) HPC service, which provides a High Performance Computing service to the University’s research community.

References

1. Haurum, J.B., Moeslund, T.B.: Sewer-ML: A multi-label sewer defect classification dataset and benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13456–13467 (2021).
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
3. og Spildevandsforening, D.V.: Fotomanualen: Beregning af Fysisk Indeks ved TV-inspektion. Dansk Vand og Spildevandsforening (DANVA), Denmark (2005).
4. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 2, pp. 2555–2563 (2023).
5. Babanagar, Nandeesh, Brian Sheil, Jelena Ninić, Qianbing Zhang, and Stuart Hardy. "Digital twins for urban underground space." *Tunnelling and Underground Space Technology* 155 (2025): 106140.
6. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).
7. Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. "A survey of deep active learning." *ACM computing surveys (CSUR)* 54, no. 9 (2021): 1-40.
8. Kottke, Daniel, Georg Krempl, Marianne Stecklina, Cornelius Styp von Rekowski, Tim Sabsch, Tuan Pham Minh, Matthias Deliano, Myra Spiliopoulou, and Bernhard Sick. "Probabilistic active learning for active class selection." *arXiv preprint arXiv:2108.03891* (2021).
9. American Society of Civil Engineers: 2021 Infrastructure Report Card: Wastewater. American Society of Civil Engineers (2021). <https://infrastructurereportcard.org/cat-item/wastewater-infrastructure/>
10. Ociepa-Kubicka, Agnieszka, Iwona Deska, and Ewa Ociepa. "Issues in Implementation of EU Regulations in Terms of Evaluation of Water Losses: Towards Energy Efficiency Optimization in Water Supply Systems." *Energies* 17, no. 3 (2024): 633