Automated Detection and Tracking of Sewer Pipeline Structural Elements

Fabian Wientjes[1,2][0009−0004−6331−2594], Bram Ton[1][0000−0002−9525−5633], and Henry Maathuis[2][0009−0002−5542−0478]

[1] Saxion University of Applied Sciences, The Netherlands
f.a.b.wientjes@saxion.nl, b.t.ton@saxion.nl
[2] HU University of Applied Sciences Utrecht, The Netherlands
henry.maathuis@hu.nl

Abstract. Regular inspection of sewer systems is essential to ensure reliable operation and prevent costly failures or environmental harm. Current practice relies on manual review of camera-based inspection footage, a process that is labor-intensive, subjective, and difficult to scale. While object detection methods offer a promising alternative, conventional frame-based approaches cannot reliably determine whether detections across consecutive frames represent the same object. This limits both detection reliability and the ability to map unique sewer components which are critical steps for condition assessment and maintenance planning. In this work, we present a method for detecting and counting unique sewer objects, specifically inlets and joints, across frames in 360° inspection imagery. Unlike prior approaches focused solely on defect detection, our framework addresses the challenge of object-level consistency across video data. The method achieves high accuracy (96.8% mAP@50 overall, 96.5% for inlets, 97% for joints on the validation set), bridging the gap between frame-level recognition and pipeline-level evaluation. This advances automated sewer inspection toward industry-standard reporting, enabling more efficient, objective, and scalable maintenance planning.

Keywords: Object detection · Computer vision · Civil infrastructure · Sewer pipe inspection · YOLO · Tracking · Class imbalance

1 Introduction

Sewer pipelines are a critical component of urban infrastructure, and are a key component to public health [10]. Therefore maintaining this infrastructure is crucial, but challenging due to its vastness. For instance, a relatively small country like the Netherlands already has approximately 150,000 kilometers of sewer infrastructure. Currently, the Netherlands faces a major civil infrastructure challenge as the majority of these sewer pipelines are approaching the end of their service life [16]. In 2022, Dutch municipalities spent 1.8 billion euros on sewer maintenance, much of it on costly full-pipe renovations [15]. These substantial costs highlight the need for innovative and cost-effective maintenance strategies [11]. Predictive maintenance has therefore become a key strategy in asset management, where failures are anticipated and mitigated based on data-driven insights.

A widely used technology to support predictive maintenance is robot-assisted sewer inspection with closed-circuit television (CCTV). Camera-equipped robots capture high-resolution imagery inside pipes, which human inspectors analyze according to the NEN-EN-13508-2 standard, the European guideline for condition assessment of sewer systems. This process is labor-intensive, subjective, and difficult to scale. Computer vision methods have been explored to support or replace manual inspection [20,21,8,7,3]. However, existing approaches are frame-based and do not reliably determine whether detections of inlets or joints, hereafter referred to as structural elements, across consecutive frames correspond to the same physical object.

This lack of temporal consistency creates a fundamental challenge: while frame-level methods can signal the presence of an observation, they cannot provide instance level information. What is needed is an approach that combines accurate detection with temporal tracking to count the number of unique objects in a pipe segment, in line with inspection standards. Furthermore, identifying individual objects can provide multiple viewpoints of the same object, which can help operators make better judgments of degradation severity.

In this work, we present an end-to-end framework for automated detection and counting of sewer *structural elements*, defined as observable features that define the structural health of the sewer pipe. Our focus is on detecting unique inlets and joints, rather than defects, as specified by NEN-EN-13508-2. Our contributions are threefold. First, we evaluate state-of-the-art object detection on large-scale 360° sewer inspection imagery, where object detection aims to identify and localize individual objects by detecting and framing them within the images. Second, we integrate temporal tracking to achieve detections at instance level. Finally, we demonstrate the effectiveness of our approach on real-world inspection data, showing that it enables more efficient, objective, and scalable object assessment, and supports predictive maintenance of sewer infrastructure.

2 Related Work

Recent advances in computer vision have enabled automated detection of defects and structural elements in sewer imagery. Convolutional neural networks (CNNs), particularly one-stage object detectors such as the YOLO family [14], have been widely adopted due to their balance of speed and accuracy, making them suitable for near real-time video analysis [14,2]. Enhancements such as attention mechanisms [18] and multi-scale feature extraction further improve performance in visually challenging scenarios, including occlusion, cluttered backgrounds, and varying defect scales [5,9]. The review by Moradi et al. confirms that deep neural networks are effective for detecting cracks, joint displacements, infiltrations, and deposits, but also highlight persistent challenges with illumination and generalization across pipelines [13].

Beyond sewer inspection, automation methods have been proposed to improve reporting and maintenance planning within the domain of sewer maintenance. For example, Tang et al. developed a rule-based system to localize sewer defects and calculate their geographic positions from CCTV videos [17]. Other work has combined hidden Markov models with CNNs to identify anomalous frames and classify defects [12]. Such approaches illustrate growing momentum toward predictive maintenance, where reliable quantification of structural elements can support prioritization of repairs, reduce reliance on manual inspection, and integrate with asset management workflows [4].

Although prior work has advanced frame-level detection, relatively few studies systematically integrate these methods to enable robust counting of unique sewer objects such as inlets and joints. Furthermore, most studies focus on general defect categories such as cracks and deposits, while structurally important elements like inlets and critical faults such as joints have received limited attention. Addressing this gap is essential for standards-compliant condition assessment and predictive maintenance, motivating the present study's focus on detecting and counting inlets and joints.

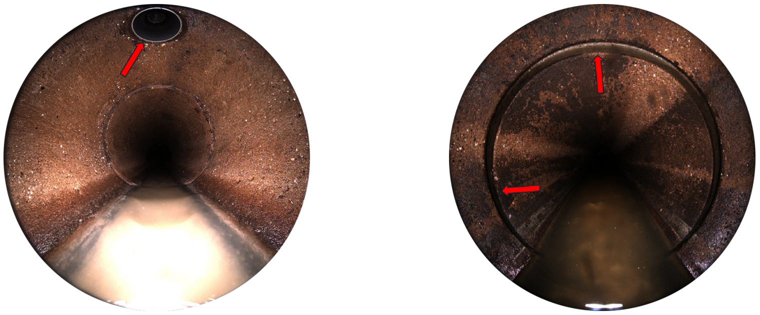
To enable robust counting of structural elements, object tracking has emerged as a complementary approach to detection. Tracking methods such as DeepSORT [19] associate detections across frames using appearance features and motion models, thereby handling occlusion and intermittent visibility. In inspection contexts, tracking has been applied to reduce duplicate counts and better estimate the occurrence of defects across video sequences [17].

3 Methodology

This section outlines our methodology. It covers the dataset and labels, training setup, evaluation of remedies for class imbalance and sparse labels (class weights, focal loss, balanced sampling), plus object tracking and the associated evaluation metrics.

3.1 Dataset

The dataset consists of 1,831 still frames extracted from sewer inspection videos provided by a municipality in the Netherlands. Images are captured using a fish-eye lens and provide a 360° view of the pipe. Images are captured at approximately 5 cm intervals within the pipe. Each image was manually annotated using bounding boxes, based on prior image-level labels conforming to the NEN-EN-13508-2 inspection standard. Two classes were selected related to the original condition codes: Inlet and Joint. Examples of these two classes are provided in Figure 1. These classes were chosen based on the selection criteria established in previous work [1], primarily because the structural elements are relatively uniform in location, appearance, and size, which makes them well-suited for object detection.

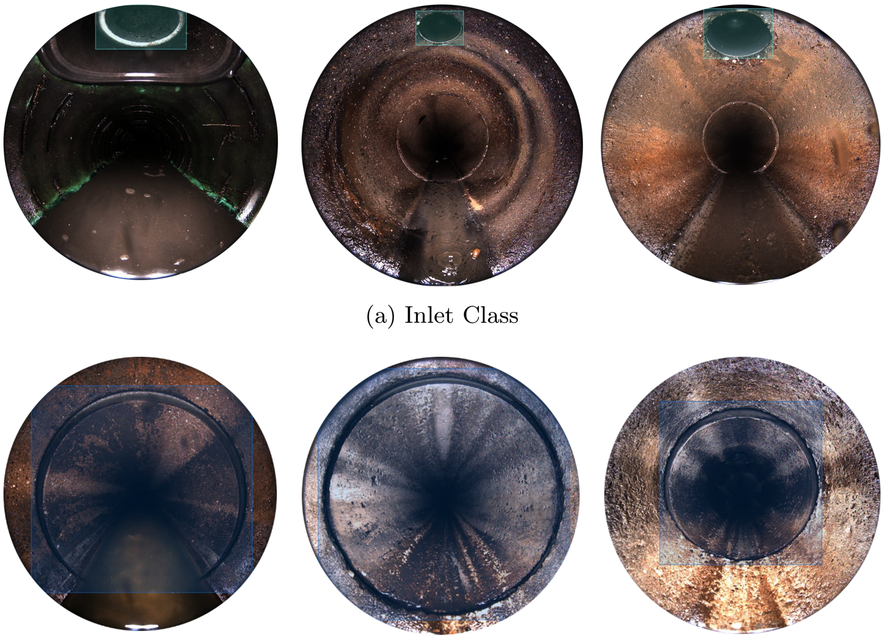


Inlet: A lateral pipe connection is clearly visible.

Joint: Visible as a circumferential ring marking the transition between pipe segments.

Fig. 1: Example image for each of the two selected classes: Inlet and Joint.

The bounding boxes were drawn manually to fully enclose each visible structural element to capture the entire relevant region. Figure 2 shows three examples per class to illustrate the annotation process.



(a) Inlet Class

(b) Joint Class

Fig. 2: Example bounding-box annotations per class.

There is a class imbalance between inlets and joints, with the inlets being in the majority. Full details regarding the class imbalance are provided in Table 1. To mitigate this imbalance, additional balancing strategies were applied during training.

Table 1: Comparison of image-level and bounding-box annotations per class.

Class	**Image-level annotations**	**Bounding-box annotations**
Inlet	1410	1446
Joint	421	421

To prevent data leakage across sets, the dataset was split at pipe level: each sewer pipe appears in exactly one of the splits. In total, the dataset comprises 231 unique sewer pipes. Of these, 180 pipes contained image-level annotations provided by a certified inspector; the remaining 51 unlabeled pipes were excluded from all experiments. The annotated subset was divided into approximately 80% training, 10% validation, and 10% test, resulting in 144 pipes (1,483 images) in the training set, 18 pipes (165 images) in the validation set, and 18 pipes (183 images) in the test set.

3.2 Model Architecture and Training

The YOLOv11 models come in various variants, to choose the best variant for our use-case, we trained a YOLOv11 object detector for each of these variants. Each variant was trained for 100 epochs on images resized to 1024×1024, matching the native dataset resolution. A batch size of 12 was used consistently across all experiments; this choice is discussed later in relation to balanced sampling of the two object classes. All other training parameters followed the default Ultralytics YOLO settings[3], including optimizer, learning rate schedule, and data augmentation.

Balancing Strategies To mitigate class imbalance, three techniques were evaluated: class weighting, focal loss, and balanced sampling with augmentation. For class weighting, the Joints class was assigned a weight three times higher than Inlets, reflecting the approximate 3:1 frequency ratio between Inlets and Joints in the dataset.

Focal loss [9] emphasizes difficult-to-classify examples by modifying the cross-entropy loss:

$$\mathcal{L}_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{1}$$

where γ controls the focus on hard samples. We evaluated $\gamma \in [1.3, 2.1]$, with a step-size of 0.2. The balanced sampling further equalized representation through oversampling and augmentation.

[3] Ultralytics YOLO: https://github.com/ultralytics/ultralytics

Augmentation Strategy To improve model generalization and robustness, eight data augmentations were applied. The augmentations included horizontal flip, simulating viewing the pipe from the opposite direction and accounting for mirror-like symmetries in structural element patterns; brightness adjustment, simulating a variety of lighting conditions including overexposure and underexposure; color jitter, simulating variations in color tone and intensity caused by staining, discoloration, or environmental effects; Gaussian blur, simulating mild defocus or image softness from camera misalignment or lens imperfections; shift-scale-rotate, replicating subtle camera shifts and pipe curvature; coarse dropout, simulating occlusions from debris, sediment, or droplets; contrast limited adaptive histogram equalization (CLAHE), enhancing subtle textures in uniformly dark or obscured regions; and Gaussian noise, simulating sensor interference or water turbidity. These augmentations were incorporated both in the balanced sampling to enrich oversampled samples and in the general training pipeline.

Object Tracking Using a tracker with unique track IDs helps to correctly distinguish structural elements that occur in close temporal proximity (only a few frames apart) and makes it more likely that predictions are assigned to the correct instance. We therefore use DeepSORT, with the following parameters: detection threshold = 0.25, match IoU threshold = 0.9, and track buffer = 90 frames; Kalman filtering and appearance-based re-identification (ReID) were enabled with the default DeepSORT thresholds. We use the per-detection track IDs only to visually disambiguate temporally close predictions in the timeline plot (Fig. 3).

3.3 Evaluation Metrics

Object Detection Performance of the detection models was evaluated using precision, recall, and mean Average Precision (mAP) based on predicted bounding boxes. The Intersection over Union (IoU) between a predicted bounding box B_p and a ground-truth bounding box B_{gt} is defined as

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}. \tag{2}$$

A detection was considered a True Positive (TP) if IoU ≥ 0.3 with a ground-truth box of the same class; otherwise, it was counted as a False Positive (FP). Ground-truth boxes without a matching detection were counted as False Negatives (FN). Precision and recall are computed as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

The mean Average Precision (mAP) is calculated as the mean area under the precision–recall curve across all classes. Given the safety-critical nature of sewer inspection, emphasis was placed on recall to minimize missed inlets and joints.

Frame-level evaluation For each frame, we record the presence or absence of *Inlet* and *Joint*, yielding four labels: *None*, *Inlet*, *Joint*, and *Both*. Table 4 shows the 4×4 GT×Pred contingency table (GT rows, Pred columns). Metrics are computed per class from the corresponding 2×2 counts (TP, FP, FN, TN). Multiple detections of the same class within a frame are counted once.

Unique Object Counting To evaluate unique-object counts (inlets and joints) along a pipeline segment, we linked consecutive detections of the same physical object using the DeepSORT object tracking algorithm. Let U_p and U_{gt} denote the sets of predicted and ground-truth unique objects. We call a pair $(u_p, u_{gt}) \in U_p \times U_{gt}$ admissible if

$$\max_{f \in \mathcal{F}(u_p) \cap \mathcal{F}(u_{gt})} \mathrm{IoU}\big(B^f_{u_p}, B^f_{u_{gt}}\big) \geq 0.3, \tag{4}$$

i.e., over the frames in which *both* u_p and u_{gt} are present, where $\mathcal{F}(u)$ denotes the set of frames in which object u appears, and B denotes the bounding box. Let $E \subset U_p \times U_{gt}$ be the set of all admissible, class-constrained pairs (Joint↔Joint, Inlet↔Inlet). Let $M \subset E$ be a one-to-one maximum-cardinality matching. We take the number of correct predictions to be $|M|$, hence

$$TP = |M|, \qquad FP = |U_p| - |M|, \qquad FN = |U_{gt}| - |M|.$$

Object-level precision and recall are

$$\text{Precision}_{\text{object}} = \frac{|M|}{|U_p|}, \quad \text{Recall}_{\text{object}} = \frac{|M|}{|U_{gt}|}. \tag{5}$$

If multiple pairs are admissible, we select the match with highest confidence (ties by highest maximum IoU over overlapping frames) when constructing M. At object level, true negatives are undefined. The object counting error (Object-CE) is

$$\text{Object-CE} = \big||U_p| - |U_{gt}|\big|.$$

For evaluating object counting and tracking, we used a test set different from the one used to train the model. Ten sewer-pipe inspection videos were annotated frame-by-frame. To define tracks, which correspond to one physical object over time, we first define tracklets. A tracklet is a temporally contiguous set of same-class detections; nearby overlapping tracklets are merged into tracks. We used IoU ≥ 0.3 for assigning objects to tracklets because boxes vary across consecutive frames, especially for inlets; higher thresholds fragmented single structural elements while lower ones merged distinct ones. 0.3 offered a practical balance. In addition, we enforced a maximum frame gap of 2 frames (continuation across brief dropouts), a merge-gap of 3 frames (merging two short track fragments across slightly longer interruptions when spatially consistent), and removed single-frame tracklets. Because the camera progresses linearly along the pipe, genuine structural elements remain visible across multiple adjacent frames;

isolated single-frame detections are therefore unlikely to correspond to real structural elements and are treated as noise. Predictions were processed identically after confidence score filtering (≥ 0.25). The resulting tracks for ground truth and predictions were used for per-pipe counts and the confusion matrix.

4 Results

This section presents the results of our study. We first report the performance of the baseline YOLOv11l detector, which was selected after a comparison of different backbone variants. Next, the findings of the strategies evaluated to address class imbalance are presented. Finally, the results of the object tracker are presented, providing insight into the stability and quantification of detections across frames.

4.1 Model Selection

As presented in Table 2, no single variant dominates all metrics, but YOLOv11l offers the best trade-off on the validation set. It achieves the highest *Inlet* recall (0.951) and the highest *Inlet* mAP@[50:95] (0.821), while maintaining high *Joint* precision (0.975) and competitive mAP@50 for both classes (Joints 0.970, Inlets 0.965). Although other variants peak on individual metrics (e.g., YOLOv11x Joint precision = 1.000; YOLOv11n Joint mAP@50 = 0.977), these gains come with reduced inlet performance on recall. Given these considerations, YOLOv11l is selected as the backbone, as it provides the most consistent and robust overall performance on the validation set.

Table 2: YOLOv11 performance metrics across different variants on the validation set.

	YOLOv11n	**YOLOv11s**	**YOLOv11m**	**YOLOv11l**	**YOLOv11x**
Joints					
Precision	0.967	0.937	0.962	0.975	**1**
Recall	**0.976**	**0.976**	**0.976**	0.971	0.922
mAP@50	**0.977**	0.972	0.97	0.97	0.975
mAP@[50:95]	**0.947**	0.941	0.933	0.914	0.89
Inlets					
Precision	0.983	**0.994**	0.987	0.959	0.915
Recall	0.927	0.933	0.942	**0.951**	0.924
mAP@50	**0.98**	0.977	0.974	0.965	0.934
mAP@[50:95]	0.795	0.81	0.803	**0.821**	0.72

4.2 Class Imbalance Mitigation

In an attempt to improve performance on the joint minority class, several strategies were evaluated. First, a focal-loss function with varying γ values was implemented, balanced sampling per batch was applied, and experiments with adjusted class weights were performed. Table 3 provides a summary of these experiments, focusing on the minority class joints. Among the focal loss configurations evaluated, a γ value of 1.3 yielded the best results for joints and is therefore reported here.

Table 3: Performance of inlet and joint detection under different class imbalance mitigation strategies, measured by mAP@50, recall, and precision on the test set. Focal Loss is applied with $\gamma = 1.3$. The balanced sampling method balances classes by sampling an equal number of examples per class in each batch. The Class Weights method assigns higher importance to joint instances during training.

	Inlets			**Joints**		
	mAP@50	**Recall**	**Precision**	**mAP@50**	**Recall**	**Precision**
YOLOv11l Standard	0.965	**0.951**	0.959	0.970	**0.971**	**0.975**
+ Focal Loss	0.967	0.947	**0.975**	0.970	0.951	0.948
+ Balanced Sampling	0.947	0.933	0.880	0.816	0.847	0.713
+ Class Weights	**0.973**	0.949	0.957	**0.973**	0.948	0.975

The results in Table 3 indicate that different imbalance mitigation strategies have varying effects on performance for inlets and joints. Using class weights yielded the highest mAP@50 for both classes. In contrast, the standard YOLOv11l configuration achieved the highest recall for both inlets and joints, indicating it is more sensitive to detecting all instances.

4.3 Tracking and Counting Objects

Table 4 presents the frame-level confusion matrix (per-frame presence/absence encoding; see Section 3.3). The confusion matrix highlights how often the detector correctly or incorrectly detected the presence of joints and inlets per frame. The detector correctly identifies most joints (446) and inlets (1177), but shows a clear bias toward false positives, particularly for inlets (395). Confusion between classes is limited; the “Both” category remains challenging, with only 69 correct predictions and frequent misdetection as single-object cases. Overall, results indicate good sensitivity but reduced precision, especially for inlets and frames containing multiple observations. Table 5 reports the object-level confusion matrix (unique objects with one-to-one IoU-based matching). At this level, true negatives are not defined and the label *Both* does not arise (each unique object has a single class). Using this protocol, we observe high recall for *Joints* (201/213; 94.4%) and *Inlets* (211/217; 97.2%) with virtually no cross-class confusion; the main limitation is precision, driven by 76 inlet false positives.

Table 4: Frame-level confusion matrix for Joints and Inlets

	None (Pred)	**Inlet (Pred)**	**Joint (Pred)**	**Both (Pred)**
None (GT)	5925	395	158	7
Inlet (GT)	94	1177	10	28
Joint (GT)	115	11	446	12
Both (GT)	1	12	26	69

Table 5: Object-level confusion matrix for Joints and Inlets.

	Joint (Pred)	**Inlet (Pred)**	**None (Pred)**
Joint (GT)	201	0	12
Inlet (GT)	0	211	6
None (GT)	10	76	0

Figure 3 shows timeline visualizations of five representative pipes (out of a total of ten). The horizontal axis denotes the frame index, which corresponds to the spatial progression along the sewer pipe. In the ground-truth (GT) rows, each vertical bar indicates the confirmed presence of a unique object: joints (left panels, red) or inlets (right panels, blue). In the prediction (Pred) rows, the vertical bars represent the model's detections, i.e. its attempt to identify the same objects. By comparing predictions with ground truth, one can assess how accurately the model detects joints and inlets along the length of each pipe, both in terms of their occurrence and their total number.

To move beyond simple occurrence counts, the frame indices belonging to each object were grouped into continuous segments. This representation highlights not only where objects were observed, but also how consistently they were visible or detected across successive frames. Ground-truth (GT) segments correspond directly to unique annotated structural elements and can therefore be counted to obtain the actual number of joints or inlets in a pipe. Prediction (Pred) segments reflect how the detector and tracker perceived those objects: ideally, one GT object yields a single segment, but fragmented tracks, ID switches, interruptions or missed frames within the same physical entity, may produce multiple segments for the same structural element, leading to overcounting. Visualizing these segments as vertical bars thus enables both quantitative object counting and qualitative assessment of temporal consistency between ground-truth and predicted timelines.

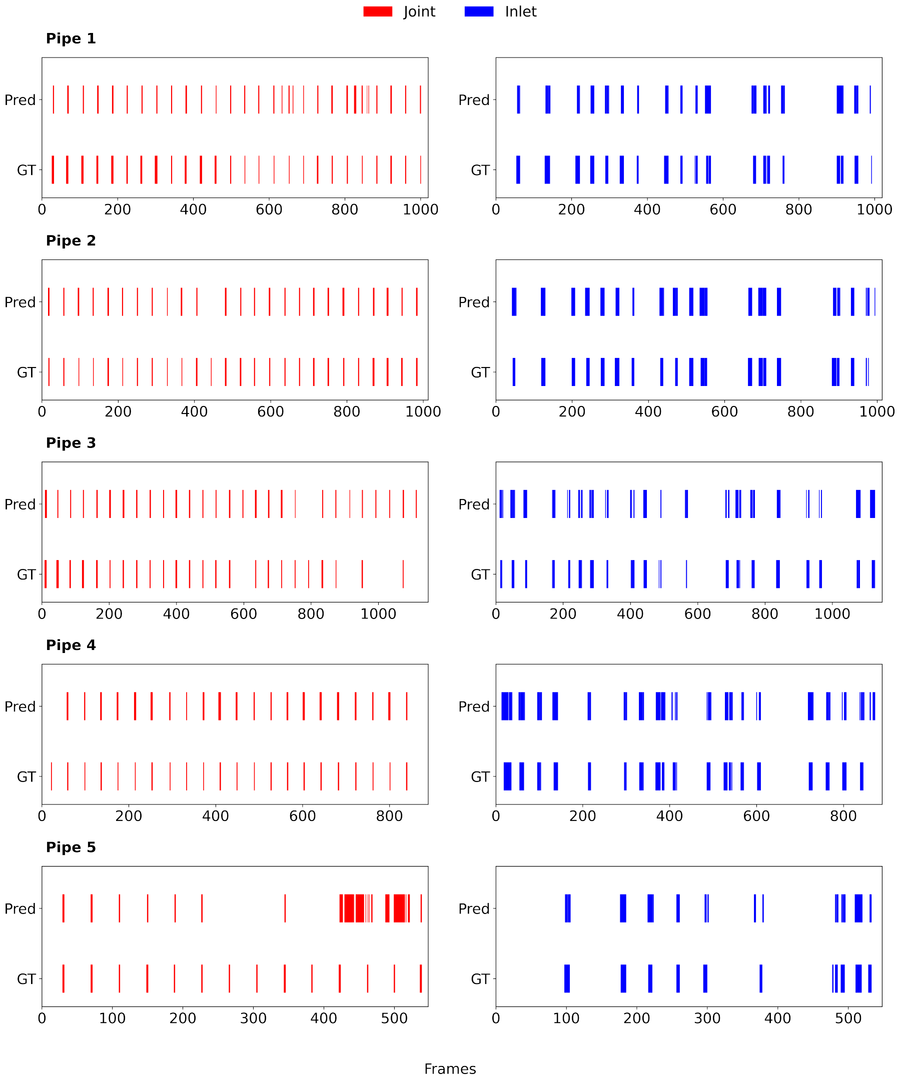


Fig. 3: Timeline visualization of five sewer pipes, where *Pred* denotes model predictions and *GT* refers to the corresponding ground-truth annotations.

4.4 Discussion

This study investigated the feasibility of detecting structural elements of sewer pipes by applying object detection and tracking with a YOLOv11 object detector on 360° still images. The results demonstrate that counting sewer structural elements from video is feasible: by grouping frame-level detections into tracks, unique structural elements can be quantified across entire pipes. Importantly, the model achieves high recall for both joints and inlets, meaning that most actual structural elements are detected at least once. This is crucial in a sewer inspection context, where the cost of missing a defect is often higher than the cost of reporting too many. Discussions with key stakeholders suggested that, in practice, missing an inlet or joint is considered unacceptable, whereas occasional false positives are regarded as tolerable. This perspective underscores the need for models with high recall, even at the expense of lower precision.

At the same time, precision remains a challenge: counts depend strongly on the consistency of the underlying detections, and inconsistent or incomplete detections may inflate false positives or cause missed detections during evaluation. In this study, a relatively high match IoU threshold of 0.9 was used in the tracker to reduce the number of fragmented predictions. The drawback of such a strict threshold is that even small shifts in bounding boxes between consecutive frames may prevent valid associations, leading to premature termination of tracks and the creation of new track IDs for the same structural element. Improving precision will therefore require both more diverse training data, particularly negative examples of common artefacts, and refinements in tracking to better handle fragmented detections. Several class imbalance mitigation strategies were evaluated to enhance performance on the minority class Joints, but none yielded consistent improvements, indicating that annotation quality remains the more critical factor.

In Figure 4, three examples of inlet detection errors are shown: two false positives and one false negative. The first false positive detects the end of a sewer pipe as an inlet because the dark circular opening resembles the contour of a true inlet. The second is caused by foam inside the channel, often seen where one sewer pipe discharges into another, which the model confuses with an inlet. The false negative occurs when the camera lens is heavily obscured by condensation caused by water flowing through the inlet; despite targeted data augmentations, this condition remains underrepresented in training.

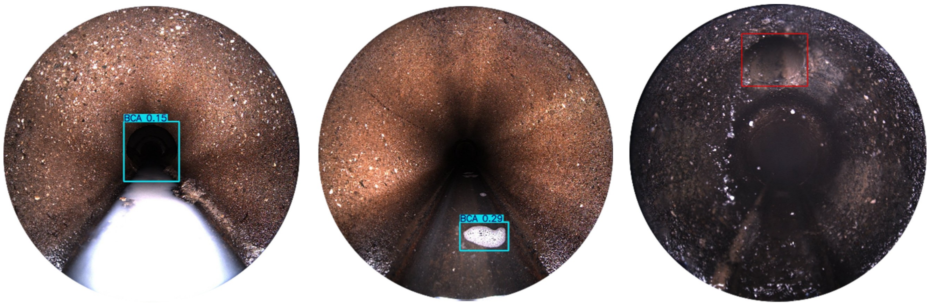


Fig. 4: Three inlet detection errors: Blue boxes show model predictions with class label (BCA, i.e., inlet) and confidence. In the rightmost image, the red box marks the ground-truth inlet missed by the model.

Figure 5 shows three consecutive frames in which the model incorrectly predicts a joint. These are false positives: no structural feature is present at this pipe location. The most likely cause is condensation or moisture on the robot's lens. Because the model is not sufficiently trained on such conditions, it misinterprets these artefacts as structural elements.

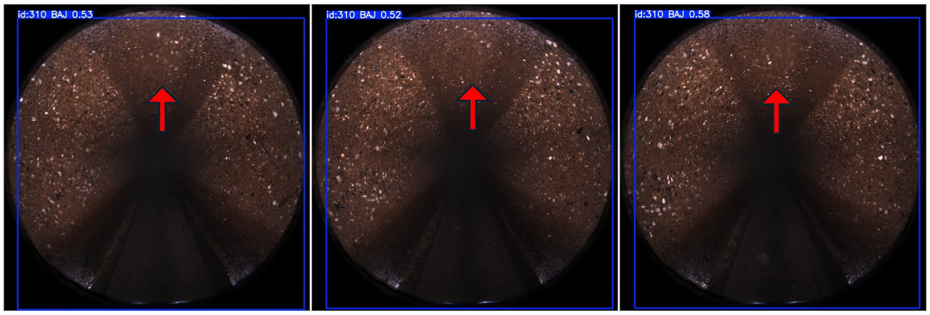


Fig. 5: Three consecutive false positives for joints in Pipe 5 (around frame 425). Blue boxes show model predictions with class label and confidence. The red arrow highlights the moist area on the lens.

The main limitation of this study is annotation quality. All bounding boxes were labeled by a single researcher in the absence of certified guidelines. Another limitation is that the model was trained on only two main structural element classes, inlets and joints, whereas NEN-EN-13508-2 defines a much broader range of structural elements, including subclassifications and severity levels. These were excluded due to limited training data, as several structural element types occur infrequently.

5 Conclusion

This study demonstrated the integration of object detection and temporal tracking for automated sewer inspection, focusing on inlets and joints. We achieved reliable detection of both structural elements, and by incorporating tracking, we were able to quantify unique structural elements across consecutive frames.

This work offers a preliminary contribution to predictive maintenance by producing standards-compliant counts of inlets and joints intended to support condition assessment and resource planning. The study contributes by (1) evaluating state-of-the-art object detection in a real-world sewer inspection context, (2) integrating object clustering to achieve structural element-level quantification, and (3) validating the approach on large-scale, high-resolution 360° imagery.

Overall, the results show that combining detection and tracking provides a scalable, practical solution for automated sewer inspection. While this study focused on two structural elements, the methodology lays the groundwork for extending automated recognition and counting to additional structural element classes defined in NEN-EN-13508-2, enabling broader applicability in predictive maintenance and civil infrastructure management.

6 Future Work

This study focused on detecting and counting inlets and joints, two structural elements that are relatively uniform in location, appearance, and size. Future work could explore extending these methods to a broader range of structural elements as defined in the NEN-EN-13508-2 inspection standard. Many of these defect types, such as cracks, infiltrations or surface damage, present greater variability in shape, size, and visual appearance, which may require adaptations in annotation protocols, model architectures, and training strategies. In addition, collaboration with stakeholders is essential to evaluate the urgency and prioritization of different condition aspects.

One promising direction is the unwrapping of 360° still images to convert circular fisheye projections into flattened panoramic views. Flattened images preserve the pipe wall geometry more consistently, enabling more intuitive and precise annotations. Unwrapping reduces distortion at the edges, allows tighter bounding boxes, and may improve the signal-to-noise ratio during training. Such an approach could be particularly valuable when extending detection to defects that are subtle, elongated, or span large portions of the pipe wall.

Pixel-level segmentation represents another potential avenue for future work. Although more labor-intensive to annotate, semantic and instance segmentation can substantially improve the reliability and accuracy of detection, especially for narrow or irregular defects, such as cracks, that are not well captured by bounding boxes. Recent developments, including the Segment Anything Model (SAM) [6], could help reduce manual effort and enable more efficient pixel-level annotation, making segmentation feasible for additional structural element classes.

Image classification may also facilitate broader structural element detection without requiring precise localization. Classifiers trained on image-level labels can indicate the presence of structural elements even in cases where bounding-box annotation is challenging or inconsistent. This approach could leverage larger datasets while reducing annotation costs and may be particularly useful for structural element types with diffuse or irregular boundaries.

Extending the methods developed in this study, including one-stage object detection, class imbalance handling, and tracking of temporally linked structural elements, to other civil infrastructure contexts, remains a valuable future direction. Bridges, tunnels, and water systems share similar challenges, including variable imaging conditions, temporal dependencies between frames, and the need for scalable predictive maintenance. Adapting these techniques to different structural element types and inspection scenarios could contribute to more generalizable and robust frameworks for infrastructure asset management.

Another promising direction is the integration of predictive perception techniques. By incorporating information from adjacent frames and accounting for camera motion, models can better anticipate object continuity and reduce frame-to-frame inconsistencies. Such temporal reasoning could improve detection stability, especially in cases of motion blur or partial occlusion, and enhance the accuracy of tracking across entire pipe segments.

A key direction for future work is to systematically benchmark model performance against that of human inspectors. Such benchmarking will help define realistic performance targets and reveal where automated detection can complement human judgment in sewer inspection.

Finally, progress in this area will benefit from standardization and collaboration. Clear protocols for bounding-box dimensions, overlap hierarchy, and annotation margins will ensure consistency across structural element classes and support temporal tracking. Joint annotation infrastructures and standardized data formats among municipalities, water authorities, and industry partners could increase dataset scale and diversity, enabling reproducible, safe, and application-ready AI solutions. A larger, more comprehensive dataset will allow systematic study of additional structural element types from the NEN-EN-13508-2 standard, improving model generalization and practical applicability across sewer systems.

References

1. Berkhout, M.: Multi-Label Classification for Sewage Pipe Anomalies. Master's thesis, University of Twente, Enschede, The Netherlands (2024), https://purl.utwente.nl/essays/102015
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Haurum, J.B., Moeslund, T.B.: A survey on image-based automation of CCTV and SSET sewer inspections. Automation in Construction **111**, 103061 (2020). https://doi.org/10.1016/j.autcon.2019.103061
4. Hu, Y., Bai, X., Zhou, P., Shang, F., Shen, S.: Data augmentation imbalance for imbalanced attribute classification. arXiv (2020), https://arxiv.org/abs/2004.13628
5. Huang, J., Kang, H.: Automatic defect detection in sewer pipe CCTV images via improved YOLOv5. IEEE Access **12**, 92797–92825 (2024). https://doi.org/10.1109/ACCESS.2024.3422275
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.: Segment anything. arXiv (2023), https://arxiv.org/abs/2304.02643
7. Kumar, S.S., Wang, M., Abraham, D.M.: Deep learning-based automated detection of sewer defects in CCTV videos. Journal of Computing in Civil Engineering **35**(1), 04020061 (2021). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000866
8. Li, Y., Wang, H., Dang, L.M., Song, H., Moon, H.: Vision-based defect inspection and condition assessment for sewer pipes: A comprehensive survey. Sensors **22**(7), 2722 (2022). https://doi.org/10.3390/s22072722
9. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. ICCV. pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324
10. Mackenbach, J.P.: Sanitation: pragmatism works. BMJ **334**(s17) (2007). https://doi.org/10.1136/bmj.39044.508646.94
11. Ministry of Infrastructure and Water Management: Reports on the condition of infrastructure by rijkswaterstaat and prorail. Tech. rep., Ministry of Infrastructure and Water Management, The Hague, Netherlands (2023)
12. Moradi, S., Zayed, T., Golkhoo, F.: Automated Sewer Pipeline Inspection Using Computer Vision Techniques, pp. 582–587. ASCE (2018). https://doi.org/10.1061/9780784481653.064, https://ascelibrary.org/doi/abs/10.1061/9780784481653.064
13. Moradi, S., Zayed, T., Golkhoo, F.: Review on computer aided sewer pipeline defect detection and condition assessment. Infrastructures **4**(1), 10 (2019)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
15. Riool & Raad: Funding. https://www.rioolenraad.nl/benodigdheden/geld/ (2023)
16. Scheperboer, I.: Structural assessment of aging sewer pipes. https://www.tue.nl/nieuws-en-evenementen/nieuwsoverzicht/14-02-2023-controle-op-sterkte-van-verouderde-rioolbuizen/ (2023)
17. Tang, J., Xia, J., Xie, Z., Li, Z., Zhang, Y.: A method for automatically locating defects in cctv inspection data of sewer pipes. IEEE Access (2025)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st

International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

19. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: Proc. ICIP. pp. 3645–3649. IEEE (2017). https://doi.org/10.1109/ICIP.2017.8296962
20. Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., Kurach, L.: A deep learning-based framework for an automated defect detection system for sewer pipes. Automation in Construction **109**, 102967 (2020). https://doi.org/10.1016/j.autcon.2019.102967
21. Zhou, Q., Situ, Z., Teng, S., Liu, H., et al.: Automatic sewer defect detection and severity quantification based on pixel-level semantic segmentation. Tunnelling and Underground Space Technology **123**, 104403 (2022). https://doi.org/10.1016/j.tust.2022.104403