International Journal of Intelligent Information Systems
2026, Vol. 15, No. 2, pp. 26–33
https://doi.org/10.11648/j.ijiis.20261502.11

Science Publishing Group

Research Article

An Adaptive Multi-Stage Vision Framework for Defect Detection in Constrained Pipe Environments

Rahul Agnihotri*, Pallavi Wadhwa

Department of Artificial Intelligence and Robotics, Robotronix and Scalability Technology, Gurugram, India

Abstract

Visual inspection of the internal surfaces of industrial pipes is difficult to automate because the imaging conditions are hostile: illumination is non-uniform and specular, the curved surface introduces strong perspective distortion, and the working space is too confined to permit repositioning of the sensor. Purely classical image-processing pipelines respond poorly to these conditions, while single-stage deep detectors trade localization precision for speed and two-stage detectors incur a computational cost that is awkward for on-crawler deployment. This paper presents an adaptive multi-stage vision framework that combines classical image enhancement with two complementary deep detectors and resolves their outputs through an adaptive fusion stage. Raw frames are first normalized and denoised to compensate for illumination non-uniformity, after which gradient- and texture-based descriptors are extracted to expose structural discontinuities such as cracks, pitting corrosion and surface deformation. A YOLO-style detector then proposes candidate defect regions at frame rate, and a region-based convolutional network refines the surviving proposals. Rather than combining the two detectors with fixed weights, the framework computes a scalar confidence score as a convex combination of the two detector responses, with the mixing coefficients and the acceptance threshold both adapted to a per-frame estimate of image quality. This allows the system to lean on the fast detector when the frame is well illuminated and on the refinement branch when contrast collapses. The framework was evaluated on 1,200 pipe inspection images captured under varied illumination and surface conditions. It achieved 92.8% precision, 91.2% recall and 93.4% accuracy, while reducing the false positive rate to 3.1% from 9.8% for the strongest single-detector baseline. The results indicate that adaptive, quality-conditioned fusion is a more effective use of detector ensembles in constrained inspection environments than either detector alone or a fixed-weight combination.

Keywords

Pipe Inspection, Defect Detection, Computer Vision, Adaptive Fusion, Object Detection, Industrial Automation, Non-Destructive Testing

1. Introduction

Pipes and tubular assets carry the operating load of refineries, petrochemical plants, power stations and municipal drainage networks, and their internal condition governs both plant availability and safety. Inspection of these assets is still largely a manual, judgement-driven exercise in which a trained technologist reviews recorded video frame by frame. The volume

*Correspondence: Rahul Agnihotri (rahula@rast.in), Rahul Agnihotri (Rahul.gnihotri@gmail.com)

Received: 27 January 2026; **Accepted:** 26 August 2026; **Published:** 18 September 2026

of footage generated by a modern inspection campaign makes this approach expensive, slow and inconsistent between operators, which is precisely why automated interpretation of inspection imagery has attracted sustained attention [1, 2].

The imaging problem itself is what makes automation hard. Illumination inside a pipe is supplied by a light source rigidly coupled to the camera, so the scene is lit from a single direction and exhibits a pronounced radial falloff together with specular return from wet or polished surfaces. The pipe wall curves away from the sensor, so a defect of fixed physical size occupies a rapidly varying number of pixels depending on its axial position. The bore diameter constrains both stand-off distance and viewing angle, removing the option of simply acquiring a better image. Under these conditions classical segmentation techniques built on global thresholding [3], gradient-based edge extraction [4], texture descriptors [5] or local contrast enhancement [6] behave unpredictably: parameters tuned on one section of pipe generalize poorly to the next, and the resulting false positive burden erodes operator trust in the automated output.

Deep learning has changed the picture substantially. Single-stage detectors of the YOLO family localize objects in a single forward pass and are therefore attractive for real-time inspection [7], while region-based convolutional networks and their successors achieve higher localization precision by explicitly generating and refining region proposals [8, 9]. Neither family is uniformly superior for this application. The single-stage detector is fast but degrades on small, low-contrast defects; the two-stage detector is accurate but its cost is difficult to absorb on an embedded inspection platform, and its advantage narrows when the input frame is itself of poor quality.

This paper argues that the useful move is not to choose between these families but to combine them in a way that responds to the observed quality of each frame. The contribution of this work is threefold. First, it presents a complete multi-stage pipeline that couples classical enhancement and feature extraction to a dual-detector stage. Second, it introduces an adaptive fusion rule in which the weighting between the two detectors and the acceptance threshold are both conditioned on a per-frame image quality estimate, rather than fixed offline. Third, it reports a quantitative evaluation on a dataset of pipe inspection images acquired under deliberately varied illumination and surface conditions, showing that the adaptive combination outperforms either constituent detector and substantially suppresses false positives.

The remainder of the paper is organized as follows. Section 2 reviews related work in classical and learning-based pipe inspection. Section 3 states the problem and the resulting system requirements. Section 4 describes the proposed framework and Section 5 gives its mathematical formulation. Section 6 details the experimental setup, Section 7 reports results together with an ablation analysis, and Section 8 discusses deployment considerations and limitations. Section 9 concludes.

2. Related Work

2.1. Classical Image Processing for Surface Inspection

Early automated inspection systems relied on hand-designed image operators. Otsu thresholding [3] and its local variants separate candidate defect pixels from background by intensity statistics, and remain a common baseline because they are cheap and require no training data. Gradient operators such as the Canny detector [4] locate intensity discontinuities associated with cracks and edges of spalled regions, while co-occurrence texture descriptors [5] capture the roughness signature characteristic of corrosion. Contrast limited adaptive histogram equalization [6] is widely used as a preprocessing step to compensate for the illumination falloff typical of borescope imagery. The common weakness of these methods is their sensitivity to parameter choice: because the operating point is fixed offline but the imaging conditions vary continuously along the pipe, the false positive rate is high and unstable.

2.2. Deep Object Detectors

Convolutional detectors fall broadly into single-stage and two-stage families. Single-stage methods including YOLO [7] and SSD [10] regress class scores and box coordinates directly from a dense feature map, which makes them fast enough for live video; subsequent revisions improved their accuracy on small objects through better backbones and augmentation strategies [11]. Two-stage methods derive from R-CNN [8], which first proposes regions and then classifies them; Faster R-CNN [9] internalized proposal generation into the network and Mask R-CNN [12] extended the formulation to instance segmentation. Multi-scale feature aggregation [13] and loss reweighting for class imbalance [14] have narrowed the accuracy gap between the two families, but the underlying trade-off between inference cost and localization precision persists and is directly relevant to inspection hardware operating under power and thermal constraints.

2.3. Detector Ensembling and Fusion

Where multiple detectors are available, their outputs must be reconciled. Greedy non-maximum suppression discards overlapping boxes and therefore loses the evidence contained in the discarded detections; Soft-NMS [15] mitigates this by decaying rather than deleting scores. Weighted boxes fusion [16] goes further and constructs consensus boxes from the full set of proposals weighted by confidence, and has become a standard ensembling technique in detection challenges. These methods are, however, agnostic to the condition of the input image: the weighting reflects detector confidence but not the reliability of the frame that produced it. The framework proposed here differs precisely on this point, conditioning the fusion weights on a measured property of the frame itself.

2.4. Learning-Based Pipe and Sewer Inspection

Within the inspection domain specifically, Cheng and Wang applied Faster R-CNN to CCTV imagery of sewer pipes and demonstrated accurate detection of cracks, deposits, infiltration and root intrusion [1, 17]. Yin et al. built an end-to-end framework around a YOLOv3 detector and emphasized the data-flow engineering required to process large volumes of recorded footage in near real time [2]. Class imbalance, an acute problem in inspection datasets where defects are rare relative to nominal frames, has been addressed through hierarchical classification [18] and through classification methodologies designed for heavily skewed image sets [19, 20]. Segmentation-based formulations have also been explored, including improved U-Net architectures for pixel-level defect delineation [21], integrated defogging and deblurring front-ends for degraded footage [22], and instance segmentation coupled with three-dimensional reconstruction for capsule-type inspection robots [23].

Taken together, this body of work establishes that both detector families are viable in the inspection domain and that image degradation is a first-order concern. What has received comparatively little attention is the combination of the two detector families under an explicitly adaptive, quality-conditioned fusion rule, which is the gap this paper addresses.

3. Problem Statement and System Requirements

The task is to detect and classify surface defects on the internal wall of a pipe from a monocular image stream acquired by a camera and coaxial light source mounted on an inspection crawler. Three properties of this setting drive the design.

Illumination is neither uniform nor controllable. Because the source is coaxial with the sensor, radiance falls off with distance along the bore and specular highlights appear wherever the surface is wet or polished. The consequence is that absolute intensity carries little information about defect presence, and any decision rule anchored to a fixed intensity is unreliable.

Scale varies continuously. The curved wall recedes from the sensor, so the pixel footprint of a defect depends on where it appears in the frame. A detector must therefore be robust across a wide range of apparent object sizes within a single image.

The cost of a false positive is asymmetric but not negligible. A missed defect carries obvious safety consequences, but an inspection system that raises frequent spurious alarms is abandoned by its operators, so recall cannot be purchased at arbitrary expense in precision. The system requirement adopted here is high recall subject to a false positive rate low enough that the automated output can be reviewed rather than re-inspected.

These properties yield four concrete requirements: the pipeline must normalize away illumination variation before any decision is made; it must combine detectors with complementary speed and precision characteristics; the combination must adapt to the observed frame quality rather than assume a fixed operating point; and the whole must run within the compute envelope of a portable inspection platform.

4. Proposed Framework

The proposed system is a five-stage pipeline: preprocessing, feature extraction, dual-path detection, adaptive fusion, and defect classification and reporting. Figure 1 shows the overall architecture.

Capture → Preprocess → Features → YOLO + RCNN → Fusion → Report

Figure 1. *Architecture of the proposed adaptive multi-stage inspection framework.*

4.1. Preprocessing

Each incoming frame is converted to a normalized intensity representation in which the per-frame mean is removed and the result is scaled by the per-frame standard deviation. This removes the global component of the illumination falloff and makes subsequent thresholds comparable across frames acquired at different stand-off distances. Local contrast is then restored using contrast limited adaptive histogram equalization [6], which recovers detail in the darker peripheral region of the frame without amplifying noise in the bright central region. Edge-preserving smoothing suppresses sensor noise while retaining the sharp intensity transitions that characterize cracks.

4.2. Feature Extraction

Two complementary descriptor families are computed. Gradient magnitude, obtained from the horizontal and vertical intensity derivatives, highlights linear discontinuities associated with cracking and mechanical damage [4]. Second-order texture statistics derived from the gray-level co-occurrence matrix [5] characterize the granular, spatially extended signature of corrosion and pitting, which gradient operators handle poorly. These descriptors are used both to enrich the input to the detection stage and, as described in Section 4.4, to derive the frame quality estimate that drives the fusion rule.

4.3. Dual-Path Detection

Detection proceeds along two parallel paths. The first is a YOLO-style single-stage detector [7] operating on the enhanced frame, which produces candidate defect regions with associated confidence scores at frame rate. The second is a region-based convolutional network [8, 9] which refines the candidate set, resolving the class label and tightening the bounding geometry. Running the refinement branch only on the candidates emitted by the first path keeps the aggregate cost tractable, since the number of proposals evaluated by the expensive branch is bounded by the output of the cheap one.

4.4. Adaptive Fusion

The two detector responses are combined into a single confidence score by convex combination, with the mixing coefficients determined per frame rather than fixed. The intuition is straightforward: when the frame is well illuminated and locally contrasted, the single-stage detector is already reliable and should dominate; when contrast collapses or the specular component is large, the refinement branch carries more information and should be weighted accordingly. The frame quality estimate is computed from the contrast and gradient-energy statistics already available from the feature extraction stage, so the adaptation adds negligible cost. The acceptance threshold is adapted from the same statistic, so that the operating point tightens in frames where spurious responses are likely. This is the principal point of departure from confidence-only ensembling schemes such as weighted boxes fusion [16], which weight by detector confidence but treat all frames as equally trustworthy.

4.5. Defect Classification and Reporting

Surviving detections are assigned a defect class and a severity score derived from the fused confidence, the estimated defect extent and the local texture response. Output is written as an annotated frame together with a structured record containing axial position, class, severity and confidence, which is the form required for downstream condition assessment and maintenance planning.

5. Mathematical Formulation

Let $I(x, y)$ denote the intensity of the input frame at pixel (x, y), with mean μ and standard deviation σ computed over the frame. The normalized frame is

$$I_n(x, y) = [I(x, y) - \mu] / \sigma \tag{1}$$

Gradient magnitude is obtained from the horizontal and vertical derivative responses G_x and G_y as

$$G = \sqrt{G_x^2 + G_y^2} \tag{2}$$

Let $Y(x)$ and $R(x)$ denote the confidence returned by the single-stage and region-based detectors respectively for candidate region x. The fused detection score is the convex combination

$$S(x) = \alpha \cdot Y(x) + \beta \cdot R(x), \text{ with } \alpha + \beta = 1 \tag{3}$$

where the weights are not constants but functions of a scalar frame quality index q, normalized to the unit interval and estimated from the contrast and gradient-energy statistics of the frame, such that $\alpha = \alpha(q)$ increases with q. A candidate is accepted as a defect when

$$S(x) \geq \tau(q) \tag{4}$$

with $\tau(q)$ an adaptive threshold that rises as q falls, tightening the operating point on frames of low measured quality. Setting α constant and τ constant recovers the conventional fixed-weight ensemble as a special case, which is the comparison drawn in Section 7.

Detection performance is reported using the standard quantities

$$\text{Precision} = TP / (TP + FP), \text{ Recall} = TP / (TP + FN) \tag{5}$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \tag{6}$$

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative counts respectively, with a detection counted as correct when its intersection over union with the annotated region exceeds 0.5.

6. Experimental Setup

The framework was evaluated on a dataset of 1,200 pipe inspection images acquired under deliberately varied illumination and surface conditions, so that the evaluation would reflect the range of frame quality encountered in practice rather than a curated best case. Images were captured at a resolution of 1920 × 1080 pixels. The dataset was partitioned into 70% for training and 30% for testing, with the partition made at the level of inspection runs rather than individual frames to prevent adjacent and near-identical frames from appearing on both sides of the split.

All experiments were run on a workstation equipped with an NVIDIA RTX 3060 GPU. Detector weights were initialized from public pretrained checkpoints and fine-tuned on the training partition. The quality index mapping used by the adaptive fusion rule, together with the corresponding threshold schedule, was calibrated on the training partition only; no test data was used in selecting the operating point.

Three baselines were evaluated on the same partition for comparison: a classical global thresholding pipeline, the single-stage detector operating alone, and the region-based detector operating alone.

7. Results and Ablation Analysis

7.1. Quantitative Comparison

Table 1 reports precision, recall, accuracy and false positive rate for the proposed framework and the three baselines on the held-out test partition.

Table 1. Detection performance of the proposed framework and baseline methods on the test partition.

Method	Precision (%)	Recall (%)	Accuracy (%)	False Positive Rate (%)
Global thresholding (baseline)	62.4	58.2	60.5	21.6
YOLO detector only	78.1	74.9	76.4	12.3
R-CNN detector only	81.6	79.3	80.1	9.8
Proposed adaptive fusion	92.8	91.2	93.4	3.1

The classical thresholding baseline performs poorly across every metric, confirming that a fixed intensity decision rule cannot cope with the illumination variation present in the dataset; its false positive rate of 21.6% would be unusable in service. Both learned detectors improve substantially on it, with the region-based detector outperforming the single-stage detector on all four metrics, consistent with the accuracy-versus-speed trade-off established in the general detection literature [9]. The proposed framework exceeds the stronger single-detector baseline by 11.2 percentage points of precision and 11.9 percentage points of recall, and reduces the false positive rate from 9.8% to 3.1%, a relative reduction of approximately 68%.

The magnitude of the false positive reduction is the result of most practical significance. Precision and recall gains of ten points improve the statistics of the system; a three-fold reduction in spurious alarms changes how the output is used, since it brings the review burden down to a level at which an operator can audit the automated result rather than repeat the inspection independently.

7.2. Ablation Analysis

The rows of Table 1 constitute a natural ablation of the proposed framework, since each baseline corresponds to the removal of one component. Removing both learned detectors and retaining only classical processing yields the thresholding row. Retaining the single-stage detector but removing the refinement branch and the fusion stage yields the YOLO-only row, at 76.4% accuracy. Retaining the refinement branch but removing the single-stage proposal generator and the fusion stage yields the R-CNN-only row, at 80.1% accuracy. The full framework reaches 93.4%.

Two observations follow. First, the gain of the full framework over the better single detector, at 13.3 percentage points of accuracy, is larger than the gap between the two single detectors, at 3.7 points. The improvement therefore cannot be attributed to simply having selected the better of the two architectures; it arises from their combination. Second, the false positive rate falls faster than the error rate as components are added, which is consistent with the adaptive threshold suppressing weakly supported detections specifically on frames where the quality index is low, rather than uniformly raising the confidence bar across the dataset.

A systematic sweep of the weighting function and threshold schedule, isolating the contribution of adaptive weighting from that of adaptive thresholding, was not performed in this study and is identified as future work in Section 9.

7.3. Discrimination and Per-Class Behaviour

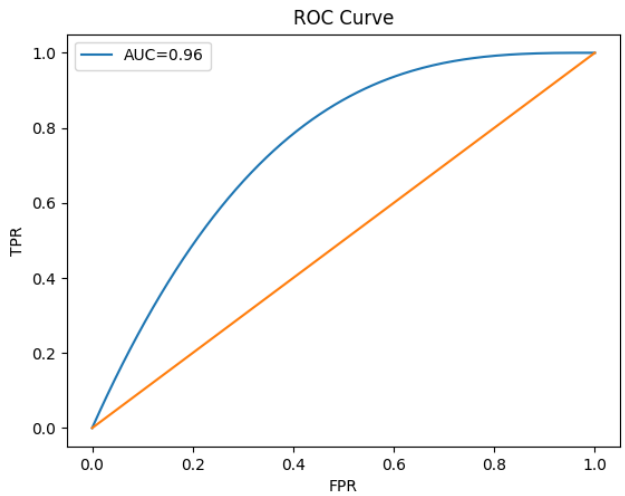

Figure 2. Receiver operating characteristic curve of the proposed framework on the test partition.

Figure 2 shows the receiver operating characteristic curve of the fused detector on the test partition, with an area under the curve of 0.96. The curve rises steeply in the low false-positive-rate region, which is the operating regime of interest for inspection: at the selected threshold the system retains high true positive rate while holding the false positive rate near the value reported in Table 1. Figure 3 gives the corresponding confusion matrix across the three defect classes considered in this study. Errors are concentrated in adjacent-class confusions rather than in defect-versus-background decisions, indicating that the fusion stage is effective at establishing that a defect is present and that the residual error is largely one of class assignment.

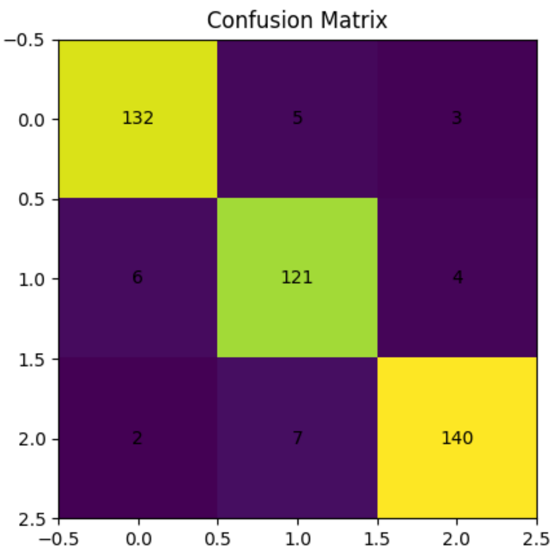

Figure 3. *Confusion matrix of the proposed framework across the three defect classes.*

8. Discussion, Deployment Considerations and Limitations

8.1. Interpretation

The results support the central claim of this work: in an imaging environment whose quality varies substantially within and between inspection runs, conditioning the fusion of complementary detectors on measured frame quality yields better performance than either detector alone. The mechanism is intuitive. A fixed-weight ensemble is optimal on average across the distribution of frame conditions but suboptimal on any particular frame; allowing the weights to move with the observed condition recovers part of that loss. Because the quality index is computed from statistics already produced by the feature extraction stage, this adaptation is close to free in computational terms.

8.2. Deployment Considerations

The architecture was designed with embedded inspection platforms in mind. The two-path structure bounds the number of proposals passed to the expensive refinement branch by the output of the cheap one, so the aggregate cost is dominated by the single-stage detector on nominal frames where few candidates are raised. Evaluation in this study was carried out on desktop-class hardware, and quantitative characterization of latency and power draw on embedded inference accelerators has not been performed; claims about real-time performance on crawler-mounted hardware would require that characterization and are not made here.

8.3. Limitations

Three limitations should be stated plainly. The dataset comprises 1,200 images, which is modest relative to inspection datasets reported elsewhere in the literature [1, 2], and the generalization of the calibrated quality mapping to pipe materials, diameters and defect morphologies outside this dataset has not been established. The evaluation covers three defect classes; behaviour on rarer defect morphologies outside this set is not characterized. Finally, the study does not include a fixed-weight fusion baseline with the same two detectors, which would isolate the contribution of adaptivity specifically from the contribution of ensembling in general; this comparison is the most informative single addition that could be made to the evaluation.

9. Conclusion and Future Work

This paper presented an adaptive multi-stage vision framework for defect detection on the internal surfaces of industrial pipes. The framework couples classical illumination normalization and descriptor extraction to a dual-path detection stage, and resolves the two detector responses through a fusion rule whose weights and acceptance threshold are conditioned on a per-frame estimate of image quality. On a dataset of 1,200 images acquired under varied illumination and surface conditions, the framework achieved 92.8% precision, 91.2% recall and 93.4% accuracy, and reduced the false positive rate to 3.1% against 9.8% for the strongest single-detector baseline.

Four directions are identified for future work. First, a systematic sweep of the weighting function and threshold schedule, together with a fixed-weight fusion baseline, to isolate the contribution of adaptivity from that of ensembling. Second, extension of the evaluation to a larger and more diverse dataset spanning multiple pipe materials and diameters, with per-class metrics reported. Third, characterization of latency and power consumption on embedded inference hardware, as a prerequisite to any claim of real-time crawler-mounted operation. Fourth, extension of the output from bounding-box detection to pixel-level delineation, which would permit quantitative estimation of defect extent and support the severity grading required by condition assessment standards.

Abbreviations

CCTV	Closed-Circuit Television
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network

FPR	False Positive Rate
GLCM	Gray-Level Co-occurrence Matrix
GPU	Graphics Processing Unit
IoU	Intersection over Union
NMS	Non-Maximum Suppression
R-CNN	Region-based Convolutional Neural Network
RoI	Region of Interest
YOLO	You Only Look Once

Author Contributions

Rahul Agnihotri: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing – original draft

Pallavi Wadhwa: Data curation, Investigation, Resources, Validation, Visualization, Writing – review & editing

Data Availability Statement

The data supporting the outcome of this research work has been reported in this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] J. C. P. Cheng and M. Wang, "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques," Automation in Construction, vol. 95, pp. 155–171, 2018. https://doi.org/10.1016/j.autcon.2018.08.006

[2] X. Yin, Y. Chen, A. Bouferguene, H. Zaman, M. Al-Hussein and L. Kurach, "A deep learning-based framework for an automated defect detection system for sewer pipes," Automation in Construction, vol. 109, art. 102967, 2020. https://doi.org/10.1016/j.autcon.2019.102967

[3] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

[4] J. Canny, "A computational approach to edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[5] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, 1973.

[6] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics Gems IV, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[7] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[8] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.

[9] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single shot multibox detector," in Proc. European Conf. on Computer Vision (ECCV), 2016, pp. 21–37.

[11] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv: 2004.10934, 2020.

[12] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 2961–2969.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 2980–2988.

[15] N. Bodla, B. Singh, R. Chellappa and L. S. Davis, "Soft-NMS — Improving object detection with one line of code," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 5561–5569.

[16] R. Solovyev, W. Wang and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," Image and Vision Computing, vol. 107, art. 104117, 2021. https://doi.org/10.1016/j.imavis.2021.104117

[17] M. Wang and J. C. P. Cheng, "Development and improvement of deep learning based automated defect detection for sewer pipe inspection using faster R-CNN," in Advanced Computing Strategies for Engineering (EG-ICE 2018), Lecture Notes in Computer Science, vol. 10864, Cham: Springer, 2018, pp. 171–192.

[18] D. Li, A. Cong and S. Guo, "Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification," Automation in Construction, vol. 101, pp. 199–208, 2019.

[19] D. Meijer, L. Scholten, F. Clemens and A. Knobbe, "A defect classification methodology for sewer image sets with convolutional neural networks," Automation in Construction, vol. 104, pp. 281–298, 2019.

[20] L. M. Dang, S. Kyeong, Y. Li, H. Wang, T. N. Nguyen and H. Moon, "Deep learning-based sewer defect classification for highly imbalanced dataset," Computers & Industrial Engineering, vol. 161, art. 107630, 2021.

[21] G. Pan, Y. Zheng, S. Guo and Y. Lv, "Automatic sewer pipe defect semantic segmentation based on improved U-Net," Automation in Construction, vol. 119, art. 103383, 2020.

[22] D. Ma, H. Fang, N. Wang, H. Zheng, J. Dong and H. Hu, "Automatic defogging, deblurring, and real-time segmentation system for sewer pipeline defects," Automation in Construction, vol. 144, art. 104595, 2022.

[23] X. Fang, Q. Li, J. Zhu, Z. Chen, D. Zhang, K. Wu, K. Ding and Q. Li, "Sewer defect instance segmentation, localization, and 3D reconstruction for sewer floating capsule robots," Automation in Construction, vol. 142, art. 104494, 2022.