Grenze International Journal of Engineering and Technology, January Issue

Deep Learning Powered Defect Detection and Temporal Localization in CCTV Pipeline Videos

Sumalatha MR[1], Saravanakumar V[2], Ramesh Kanna S[3], Lokesh S[4] and Mageshwar HM[5]
[1-5]MIT Campus, Anna University, Chennai, Tamilnadu, India,
Email: sumalatha.ramachandran, skvel6089, kannaramesh822, sakkaraiplokesh, mageshwarhm@gmail.com

***Abstract*— Automatic vision-based sewer inspection plays a vital role of sewage system in a modern city. Recent advances focus on modelling a deep learning-based methods to realize the sewer inspection system, benefiting from the capability of data-driven feature extraction. Although the acceptable performances of sewer defect classification are achieved, there is still a gap between the emerged methods and actual application scenarios. In this work we focus on implementing a lightweight CNN architecture with reduced number of computations using Ghost Module along with feature fusion technique and attention mechanisms.**

***Index Terms*— Defect Detection, Classification, Ghost Module, SE Attention Block, Feature Fusion, Temporal Localization, Pipeline Defects.**

I. INTRODUCTION

Manual inspection of urban sewage pipeline systems poses potentially fatal physical threats to humans and hence requires prompt intervention of automated approaches. Precisely detecting minute defects and classifying them requires robust Deep Learning algorithms for automation. Alongside, accurately specifying which part of the pipeline is defecting becomes crucial in robotic maintenance-based applications. The proposed software system integrates 3 different modules to perform the pipeline inspection task. The defect detection module focuses on developing a robust deep learning framework for efficiently identifying defects in inner wall images of pipelines captured by CCTV-equipped robots. Additionally, a type classification module (GhostSECNN) is designed to classify six major defect types using deep learning techniques, enhancing classification accuracy. For precise monitoring, the temporal localization module aims to accurately determine the time point and type of defect occurrence in CCTV video footage, ensuring effective pipeline inspection and maintenance.

II. LITERATURE SURVEY

A. Defect Detection in Pipeline CCTV Videos

Zhao et.al [1] Utilizes a Support Clip Module (SCM) for fine-grained defect representation from local-scale video segments and Evidential Deep Learning (EDL) for uncertainty estimation, achieving a 5.39% mAP improvement and a 14.70% boost in mF1 score for multi-label classification. Basireddy et.al [2] employs image preprocessing techniques and deep CNN-based approaches, including MobileNet for initial detection and DenseNet-169 for classification, achieving an F1-score of 90.2% (validation) and 89.4% (test), with an F2CIW score of 63.22% (validation) and 57.68% (test).

Grenze ID: 01.GIJET.12.1.66

Chen et al. [3] propose an innovative convolutional neural network (CNN)-based approach for intelligent sewer defect detection, focusing on an end-to-end classification system directly from raw CCTV images. Xie et al. [4] introduce a hierarchical deep learning framework that utilizes multiple CNN layers integrated with specialized defect detection heads. This architecture allows the model to handle the detection and classification of various sewer defects simultaneously. Rayhana et al. [5] present an automated defect-detection system for water pipelines, utilizing CCTV video footage captured by autonomous robotic platforms. Their system integrates advanced vision-based processing techniques with intelligent filtering and classification methods to detect defects in real-time.

B. Sewer Defect Classification

Hu et.al [6] Implements a Focal Segment Module (FSM) and EDL to enhance multi-label classification, improving mAP by 6.85% and boosting the mF1 score by 12.02%. Haurum et.al [7] Provides a comprehensive benchmark for multi-label classification, introducing the F2CIW metric and achieving an F2CIW score of 55.11% and a normal pipe classification F1 score of 90.94%. Chen et al. [8] investigate the application of Evidential Deep Learning (EDL) to multi-label sewer defect classification, emphasizing the role of uncertainty estimation in improving model reliability.

EDL allows the system to quantify the uncertainty associated with its predictions, which is crucial in scenarios where defect characteristics may overlap or appear ambiguous. Tao et al. [9] introduce the Correlation-Aware Feature Enhancement Network (CAFEN), a novel deep learning architecture designed to improve sewer defect detection by addressing the issue of correlated defects. Haurum et al. [10] extend their earlier work on sewer defect classification by incorporating a multi-task learning framework with a Cross-Task Graph Neural Network (GNN) Decoder.

C. Temporal Localisation of Sewer Defects

Vishwakarma et.al [11] Uses optical flow and perceptual hashing techniques with a Stationary Flow Filter (SFF) for frame classification, achieving 70% accuracy and a 46% recall for motion analysis. Liu et.al [12] Focuses on temporal defect localization, using models such as Video Swin Transformer and ConvNeXt, achieving a high mAP of 72.18% for video classification and 17.65% average mAP for temporal localization. Ma et al. [13] propose a robust multi-defect detection system for sewer pipelines that integrates StyleGAN-SDM with a fusion CNN architecture.

The StyleGAN-SDM module generates synthetic defect samples to address the issue of class imbalance, especially for rare or underrepresented defect types. Fan et al. [14] present a structured light vision system tailored for underwater pipeline inspection and 3D reconstruction. Their method employs structured light projection combined with high-resolution video analysis to build precise spatial models of pipeline surfaces. Bhagat and Basireddy et al. [15] develop an efficient vision-based automated inspection system that combines traditional image processing techniques with lightweight CNNs.

III. OBJECTIVES

One of the critical challenges in pipeline maintenance is detecting minute defects that could lead to structural failures if left unaddressed. The Defect Detection Module aims to develop a robust deep learning framework that can efficiently analyze high-resolution images of inner pipeline walls captured by CCTV-equipped robotic systems. The module will leverage advanced convolutional neural networks (CNNs) and transformer-based models to identify various defect patterns, ensuring a high degree of accuracy in real-world conditions

Pipeline defects can vary significantly in severity and cause, necessitating a specialized approach to classification. The Type Classification Module focuses on identifying and categorizing six major defect types using a novel ensembling technique that combines multiple deep learning architectures. By integrating CNNs and transformer models, this module enhances classification robustness, reducing misclassification errors. The ensemble approach ensures that even visually similar defect types, such as cracks and joint displacements, are accurately differentiated, leading to more effective decision-making in pipeline maintenance operations. In addition to spatial analysis, precise Temporal Localization of defects is essential for proactive maintenance. This module is designed to pinpoint the exact time and location of defects within CCTV video footage, enabling maintenance teams to act swiftly. Utilizing a combination of detection and classification models developed earlier, the system analyses frame sequences to detect anomalies and track defect progression over time. This real-time defect monitoring capability significantly improves predictive maintenance strategies, preventing catastrophic failures in underground sewage and industrial pipeline systems.

IV. Proposed Work

A. Data Collection and Preprocessing

The dataset comprises of images (Fig. 1) collected from various water utility companies. All images have been meticulously annotated by licensed sewer inspectors, ensuring consistent and reliable labelling across all defect classes. This standardized approach enhances the dataset's credibility and usability for defect detection and analysis.

The videos are captured by navigating urban sewer environments, providing a comprehensive view of real-world pipeline conditions. The dataset includes various types of defects, broadly categorized into structural defects such as broken pipes, deformation, misalignment, and disconnection and functional defects, including depositions and obstacles.

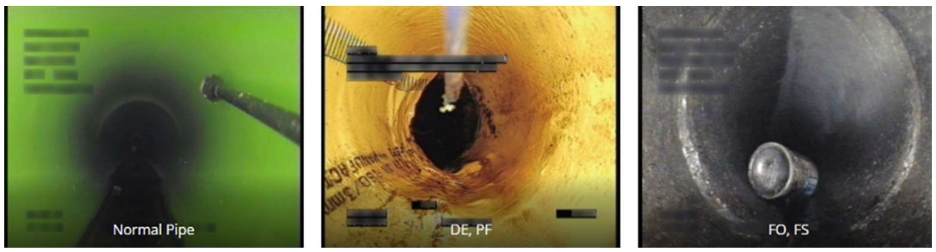

Figure 1. Sample images from the dataset

B. Automated Pipeline Inspection Architecture

The system comprises of two different models one each for detection and classification of defects respectively. DeepLabv3 is a powerful defect detection model that utilizes atrous convolutions for detailed defect detection in pipelines. Its Atrous Spatial Pyramid Pooling (ASPP) module captures multi-scale contextual information, improving segmentation accuracy for defects of varying sizes. Backbone networks such as ResNet and Xception enhance feature extraction, leading to better performance and robustness. This model is particularly effective for defect detection, enabling the identification of cracks, deformations, misalignments, and depositions in pipeline structures. With its high accuracy and efficiency, DeepLabv3 provides precise detection of defects while maintaining robustness to scale variations, making it a reliable choice for industrial inspection and monitoring.

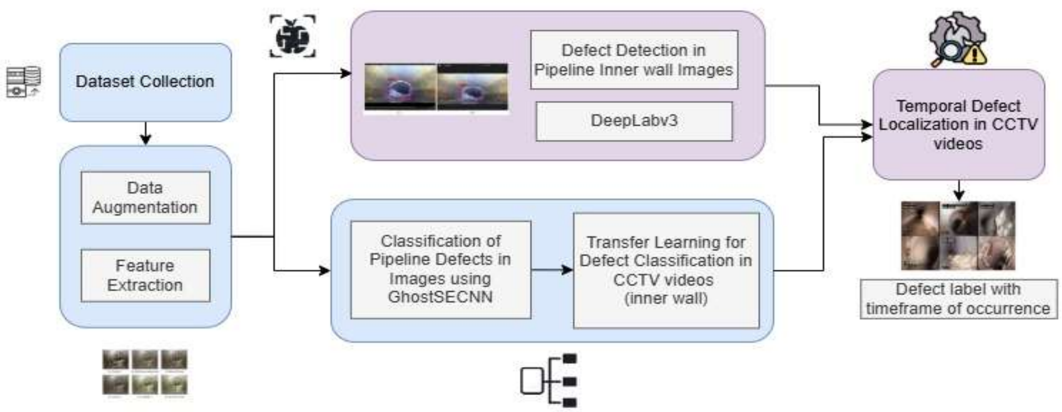

Figure 2. Automated Pipeline Inspection Architecture

Following the fusion stage, the system incorporates Ghost Module to generate more expressive feature maps with fewer parameters by applying cheap linear operations on intrinsic feature maps. This is particularly useful in edge-deployment scenarios like real-time pipe inspection where memory and speed are critical. Additionally, Squeeze-and-Excitation (SE) blocks are used to recalibrate channel-wise feature responses, enhancing important defect-related features and suppressing noise. A spatial attention module further emphasizes localized defect regions. Finally, a global average pooling followed by a fully connected layer predicts the defect class. This architecture is both lightweight and powerful, combining innovations from modern CNN and Transformer paradigms to deliver high accuracy in challenging real-world defect classification tasks.

In the pipeline defect localisation module, DeepLabV3 and the GhostSECNN classifier are integrated in a sequential manner to achieve both localization and identification of defects. First, each video frame is passed through the DeepLabV3 segmentation model, which produces a pixel-wise mask highlighting the regions likely to contain defects.
These masks are then processed to extract contours, from which bounding boxes are generated around the segmented defect areas. Each cropped defect region is resized and preprocessed before being passed to the hybrid classification mode.
This integration enables the system to not only detect where defects occur in the pipeline but also accurately determine what kind of defect is present. The frames are saved into a new video file, and a log is maintained that records the time, estimated camera distance, and the detected defect class.

V. GhostSECNN Model Implementation

A. Feature Fusion

Fusion of features extracted using 2 different backbones - EfficientNet & SwinTransformer. EfficientNet is a CNN-based architecture optimized for high accuracy and efficiency. Given an input image $X \in R$ $(C \times H \times W)$, EfficientNet extracts hierarchical features. It uses depthwise separable convolutions to reduce computations and applies squeeze-and-excitation (SE) blocks to enhance channel-wise attention.
Swin Transformer is a vision transformer (ViT) variant that processes images as a sequence of patches. First, the input image is divided into non-overlapping patches of size $P \times P$ which are then linearly embedded into feature vectors and processed by self-attention mechanisms.

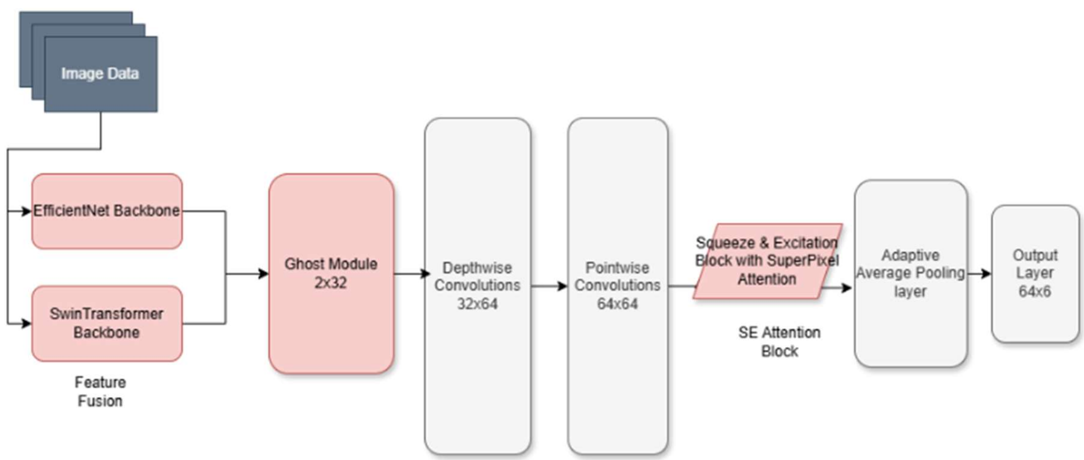

Figure 3. GhostSECNN Model Architecture

B. Ghost Modules

Ghost Modules are an efficient architectural innovation designed to reduce the computational cost of deep neural networks while maintaining high representational power.
Instead of directly computing all feature maps using standard convolutions, Ghost Modules generate a smaller set of intrinsic feature maps and then apply lightweight linear transformations, such as depth wise convolutions, to create additional "ghost" feature maps. This approach significantly reduces the number of parameters and FLOPs (floating-point operations) required for feature extraction, making it particularly useful for mobile and embedded applications.
In Ghost Modules, instead of computing NNN full feature maps, we compute Primary feature maps (Y') using a reduced number of convolutions (e.g., N/r). Ghost feature maps Gi by applying cheap linear transformations Ti on Y':

$$G_i = T_i(Y'), \qquad i = 1,2,\ldots,(r-1) \tag{1}$$

The final output, Y is obtained by concatenating primary and ghost features:

$$Y = \mathrm{Concat}(Y', G_1, G_2, \ldots, G_{r-1}) \tag{2}$$

C. Squeeze-and-Excitation Blocks

Squeeze-and-Excitation (SE) Blocks introduce an advanced attention mechanism that strengthens feature representations by explicitly modeling inter-channel dependencies in convolutional neural networks. By applying a two-step process—squeeze and excitation—SE Blocks first aggregate spatial information across each channel to generate a global context, then use this context to learn adaptive channel importance weights.By dynamically adjusting channel-wise feature responses, SE Blocks enhance the network's ability to capture critical details, making them particularly effective for tasks like defect detection, where fine-grained features play a crucial role in distinguishing different defect types. The squeezed vector S_c, is calculated as follows

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \quad (3)$$

where *H and W* are height and width of the frame respectively

$$z = \sigma(W_2 \delta(W_1 s)) \quad (4)$$

$$\tilde{X}_c = z_c \cdot X_c \quad (5)$$

where δ, σ and X correspond to the Reduction Layer, Expansion Layer and the Excitation output respectively. The SE Attention Mechanism is crucial for defect detection as it selectively enhances important features while minimizing irrelevant details, ensuring that defective regions are more prominently represented in the feature maps. By leveraging a lightweight design that employs simple Multi-Layer Perceptron (MLP) layers, SE Blocks introduce minimal computational overhead while significantly improving feature discrimination. This efficient recalibration process enables convolutional neural networks (CNNs) to focus on defect-related features with greater precision, ultimately boosting classification accuracy. By dynamically adjusting channel-wise feature responses, SE Blocks enhance the model’s ability to detect subtle defects, making them an essential component for high-performance pipeline inspection systems.

D. GhostSENet Classifier

This algorithm outlines the forward pass of a hybrid CNN-based classifier called GhostSENetClassifier, designed for defect classification. It combines EfficientNet and Swin Transformer features, enhances them using Ghost and Squeeze-and-Excitation (SE) mechanisms, and outputs the predicted defect type.

Algorithm 1 GhostSENetClassifier(input)

```
1:  F_eff = EfficientNet(X)                         ▷ Extract global semantic features
2:  F_swin = SwinTransformer(X)                     ▷ Extract hierarchical visual features
3:  Z_I = MSA(LN(Z)) + Z                            ▷ Multi-head self-attention in Swin
4:  Z_II = MLP(LN(Z)) + Z_I                         ▷ Feed-forward network of Swin block
5:  F_fusion = Concat(F_eff, F_swin) ∈ ℝ^((C_eff+C_swin)×H×W)   ▷ Feature fusion
6:  F_reduced = Conv2D(F_fusion)                    ▷ Reduce channel dimension
7:  F_ghost = W_1 * F_reduced + G(F_reduced)        ▷ Generate ghost feature maps
8:  F_se = w_c · F_ghost                            ▷ Channel-wise recalibration (SE)
9:  F_sam = Attention(F_se)                         ▷ Spatial attention for defect saliency
10: y = W_fc F_gap + b_fc                           ▷ Fully connected layer with Global Avg Pool
11: Output: y = defect_type                         ▷ Predicted class
```

D. Generating Localised Defect Logs

To accurately pinpoint the location of defects in sewer pipeline videos, a combination of semantic segmentation and classification is utilized. The DeepLabV3 model is first applied to each frame to segment regions containing possible defects. Once a defect region is identified, its bounding box is extracted and classified. To determine the precise location of each detected defect, a temporal log is maintained that includes the timestamp, camera-travelled distance, and defect type. The location of the defect is calculated using the frame number at which the defect appears, along with the camera's known frame rate and constant speed. The formulas used are:

$$\text{Time (seconds)} = \frac{\text{Frame Count}}{\text{Frames Per Second (FPS)}} \quad (6)$$

$$\text{Distance (meters)} = \text{Time} \times \text{Camera Speed (meters/second)} \quad (7)$$

The first formula computes the time at which the defect appears by dividing the current frame count by the video's frame rate (FPS). The second formula multiplies this time by the speed at which the inspection camera moves through the pipeline to determine the approximate distance from the starting point. This method ensures not only accurate detection and classification of pipeline defects but also their precise localization, which is critical for targeted repair and maintenance.

VI. RESULT ANALYSIS

The following models were evaluated for their effectiveness in detecting pipeline defects, particularly in assessing accuracy and other metrics using image-based analysis. Upon experimenting with various models such as EfficientNet, LGBM Regressor, GhostSECNN, GhostSECNN (with Feature Fusion) the results obtained are presented in comparison with each other as follows

A. Comparison of Performance Metrics

To accurately pinpoint the location of defects in sewer pipeline videos, a combination of semantic segmentation and classification. To evaluate the models, the following performance metrics were considered. Training Accuracy: Measures the proportion of correctly classified samples within the training dataset. A high training accuracy indicates that the model has effectively learned patterns from the training. Testing Accuracy: Evaluates the model's classification performance on a separate test dataset. It reflects the model's ability to generalize to new samples. A significant drop in testing accuracy compared to training accuracy may indicate overfitting.

TABLE I. COMPARISON OF PERFORMANCE METRICS FOR VARIOUS MODELS

Model	Training Accuracy	Testing Accuracy
EfficientNet	0.759	0.483
LGBMRegressor	0.852	0.479
GhostSECNN	0.773	0.735
GhostSECNN (with feature fusion)	0.99	0.88

B. Model Evaluation, Analytics and Insights

The above listed models were implemented for our dataset and the results obtained (Fig. 4 – 11) are as follows:

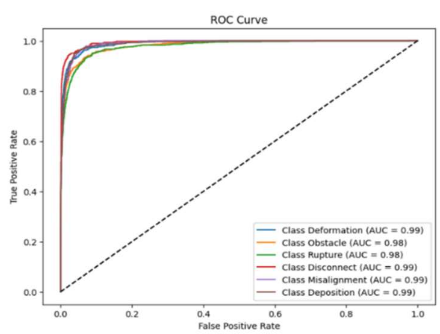

Figure 4. GhostSECNN AUC-ROC curve

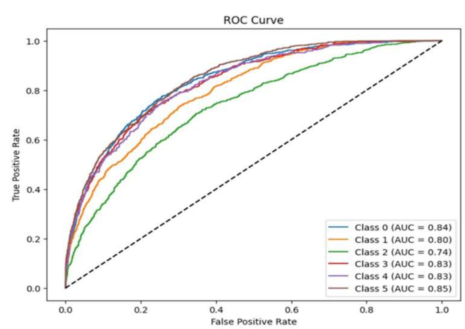

Figure 5. LGBM Regressor AUC-ROC curve

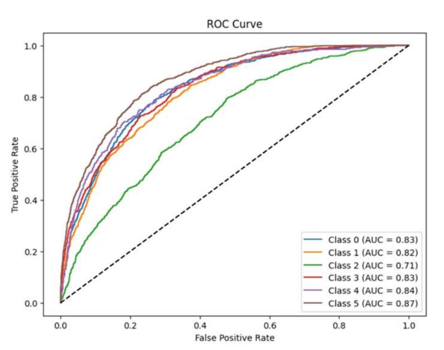

Figure 6. EfficientNet AUC-ROC curve

```
Classification Report:
               precision    recall  f1-score   support

 Deformation       0.92      0.89      0.90       795
    Obstacle       0.88      0.88      0.88       938
     Rupture       0.86      0.85      0.86       823
  Disconnect       0.83      0.94      0.88       484
Misalignment       0.88      0.82      0.85       474
  Deposition       0.90      0.91      0.90       910

    accuracy                           0.88      4424
   macro avg       0.88      0.88      0.88      4424
weighted avg       0.88      0.88      0.88      4424
```

Figure 7. Classification report for the GhostSECNN model

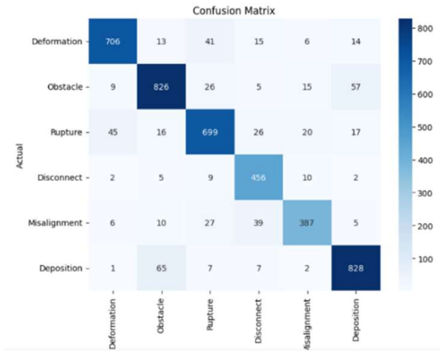

Figure 8. Confusion Matrix for GhostSECNN model

```
Final mAP Score: 0.9482
 Class 0: AP = 0.9603
 Class 1: AP = 0.9489
 Class 2: AP = 0.9336
 Class 3: AP = 0.9622
 Class 4: AP = 0.9233
 Class 5: AP = 0.9609
```

Figure 9. Classwise mAP scores for GhostSECNN

Figure 10. Sample inference image

```
Time: 0.05s | Distance: 0.01m | Defect: Rupture
Time: 0.10s | Distance: 0.01m | Defect: Deformation
Time: 0.15s | Distance: 0.01m | Defect: Misalignment
Time: 0.30s | Distance: 0.03m | Defect: Obstacle
Time: 0.35s | Distance: 0.03m | Defect: Rupture
Time: 0.40s | Distance: 0.04m | Defect: Misalignment
Time: 0.45s | Distance: 0.05m | Defect: Deformation
Time: 0.50s | Distance: 0.05m | Defect: Misalignment
Time: 0.55s | Distance: 0.06m | Defect: Deformation
Time: 0.60s | Distance: 0.06m | Defect: Misalignment
Time: 0.65s | Distance: 0.07m | Defect: Deformation
```

Figure 11. Defect Logs with corresponding timeframes

VII. Conclusion

The proposed model demonstrated superior performance across both the ROC curve, accuracy score and classwise-mAP, outperforming all other models in these key areas. It effectively balances sensitivity (recall) and specificity (precision), which makes it highly suitable for defect type classification. Its ability to generalize to unseen minute defects makes it better suited for automated inspections and is particularly effective for real-time deployment, as it balances accuracy with less parameters than traditional image classification models. The inference time for the proposed model is also observed to be significantly lesser than others.

The results suggest that the proposed GhostSECNN model has the potential to significantly improve real-time defect prediction and classification accuracy and provide valuable insights for avoiding manual overhead and related accidents. The proposed CNN model achieves better AUC-ROC score during training and has a better TPR as compared to FPR at all times.

References

[1] C. Zhao, C. Hu, H. Shao, F. Dunkin and Y. Wang, "Trusted Video-Based Sewer Inspection via Support Clip-Based Pareto-Optimal Evidential Network," in IEEE Signal Processing Letters, vol. 32, pp. 356-360, 2025

[2] A. Bhagat and S. R. Basireddy, "Vision Based Automated Pipeline Inspection," 2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV), Dubai, United Arab Emirates, 2024

[3] K. Chen, H. Hu, C. Chen, L. Chen and C. He, "An Intelligent Sewer Defect Detection Method Based on Convolutional Neural Network," 2018 IEEE International Conference on Information and Automation (ICIA), Wuyishan, China, 2018

[4] Q. Xie, D. Li, J. Xu, Z. Yu and J. Wang, "Automatic Detection and Classification of Sewer Defects via Hierarchical Deep Learning," in IEEE Transactions on Automation Science and Engineering, vol. 16, no. 4, pp. 1836-1847, Oct. 2019

[5] R. Rayhana, H. Yun, Z. Liu and X. Kong, "Automated Defect-Detection System for Water Pipelines Based on CCTV Inspection Videos of Autonomous Robotic Platforms," in IEEE/ASME Transactions on Mechatronics, vol. 29, no. 3, pp. 2021-2031, June 2024

[6] C. Hu, C. Zhao, H. Shao, J. Deng and Y. Wang, "TMFF: Trustworthy Multi-Focus Fusion Framework for Multi-Label Sewer Defect Classification in Sewer Inspection Videos," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 12, pp. 12274-12287, Dec. 2024.

[7] J. B. Haurum and T. B. Moeslund, "Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 13451-13462

[8] C. Zhao, C. Hu, H. Shao, Z. Wang and Y. Wang, "Towards Trustworthy Multi-Label Sewer Defect Classification via Evidential Deep Learning," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5

[9] M. Tao, L. Wan, H. Wang and T. Su, "CAFEN: A Correlation-Aware Feature Enhancement Network for Sewer Defect Identification," 2022 21st International Symposium on Communications and Information Technologies (ISCIT), Xi'an, China, 2022, pp. 204-209

[10] J. B. Haurum, M. Madadi, S. Escalera and T. B. Moeslund, "Multi-Task Classification of Sewer Pipe Defects and Properties using a Cross-Task Graph Neural Network Decoder," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 1441-1452

[11] R. Vishwakarma, R. Vennelakanti, P. Huyn and M. Tamura, "Unsupervised Stationary Flow Filter in CCTV Pipeline Inspection Videos," 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Victoria, Seychelles, 2024, pp. 1-6

[12] Y. Liu et al., "VideoPipe 2022 Challenge: Real-World Video Understanding for Urban Pipe Inspection," 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 4967-4973

[13] Duo Ma, Jianhua Liu, Hongyuan Fang, Niannian Wang, Chao Zhang, Zhaonan Li, Jiaxiu Dong,A Multi-defect detection system for sewer pipelines based on StyleGAN-SDM and fusion CNN, Construction and Building Materials,Volume 312,2021,125385,ISSN 0950-0618

[14] Junfeng Fan, Yaming Ou, Xuan Li, Chao Zhou, Zengguang Hou, "Structured Light Vision Based Pipeline Tracking and 3D Reconstruction Method for Underwater Vehicle", IEEE Transactions on Intelligent Vehicles, vol.9, no.2, pp.3372-3383, 2024

[15] Atul Bhagat, Sandeep Reddy Basireddy, "Vision Based Automated Pipeline Inspection", 2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp.789-794, 2024