Lightweight Semantic 3D Mapping in Sewer Pipes Leveraging Cylindrical Geometry

Alex George[1], Lyudmila Mihaylova[1] and Sean R. Anderson[1]

***Abstract*— In this paper, we present a lightweight method for 3D semantic mapping in sewer pipes from Closed-Circuit Television (CCTV), leveraging the cylindrical pipe geometry. The use of a cylindrical prior overcomes the problems with monocular cameras related to depth ambiguity, scale drift, and the need to perform computationally intensive 3D reconstruction for mapping. An additional contribution is that we use weakly supervised semantic segmentation for landmark detection, which avoids the need for pixel-level labelled datasets. The complete system uses separate extended Kalman filters (EKFs) for landmark tracking in the 2D image plane, which avoids problems due to low parallax for distant objects in pipes. Results on real-world, public sewer inspection data demonstrate that the perception module achieves an AUROC above 0.93, while the mapping backend operates in real-time at 27.63 FPS with centimetre-scale longitudinal accuracy (0.12 m).**

I. INTRODUCTION

Sewer networks are critically important infrastructure but one that is aging, so requiring constant inspection to detect defects. The standard inspection tool used in industry is Closed-Circuit Television (CCTV) systems mounted on a tethered mobile crawler. It would be highly advantageous to be able to create 3D semantic maps of the pipes from this standard system because it would enable asset mapping, defect localization, and the ability to monitor the degradation of defects over time. A major barrier to using CCTV for this is that it is a monocular inspection system, lacking depth perception, which is the research challenge we address here.

Typical research in this domain attempts to overcome the limitations of monocular CCTV by using additional sensors like structured lighting [1], which enables spatial perception from single-camera images. Another alternative that has been popularised is by leveraging the known cylindrical geometry of the pipe, which brings multiple benefits in solving the depth perception problem, scale drift and ambiguity, and errors introduced by low parallax of distant objects [2]. Both these techniques are viable but the latter, on using pipe geometry has the advantage that it can be used with standard CCTV crawlers.

The most accurate systems developed in research for sewer inspection go beyond pure CCTV to use multiple sensors such as LiDAR, ring lasers and inertial measurement units (IMUs) as well as cameras [3]. This results in submillimetre level accuracy. However, the use of these additional sensors does mean that the approach is not directly usable on standard industry platforms without extensive modification. Another drawback is that they create dense data that must undergo heavy computational processing to create 3D maps of the sewer pipe environment. This is acceptable for offline use but makes it more difficult to implement online in real-time, especially on hardware constrained crawlers.

An additional problem for creating semantic maps of the sewer pipe environment is that it is difficult to apply AI techniques simply due to the scarcity of labelled data. A number of research papers have applied detection-recognition systems like YOLO to sewer pipes [4], [5], or semantic segmentation networks [6], [7] but the datasets are typically private arising from industry and/or relatively small for the application of deep learning. Major public datasets, such as Sewer-ML [8] and the Water Research Centre (WRc) dataset [9], only contain image labels and so ostensibly are only suited to classification. However, we propose the use of weak semantic segmentation techniques here, based on interpretable anomaly detection, which can make use of these types of dataset without pixel-level labels [10], [11].

In summary, in this paper we propose a novel, lightweight, computationally efficient solution for 3D semantic mapping in sewer pipes, which leverages the cylindrical geometry of the pipe and weakly supervised semantic segmentation. The method makes uses of separate extended Kalman filters (EKFs) for tracking detected landmarks in real-time in the 2D image plane, which overcomes problems associated with low parallax for distant objects. We exploit the known tether odometry to solve the localisation problem, which is standard on CCTV crawlers. The main contributions of the paper are:

- A lightweight system for semantic mapping in pipes, where the use of the cylindrical prior enables fast 3D reconstruction.
- The development of a weakly supervised semantic segmentation system for object detection and localisation in the pipes, overcoming the need for large, pixel-level labelled datasets.
- The development of an efficient EKF tracking scheme with global nearest neighbour data association in the 2D image plane for landmarks.
- Demonstration of the separate components and integrated system on real-world, publicly available datasets, demonstrating centimetre-grade mapping accuracy.

This approach potentially enables the conversion of both historical, current and future industry-standard CCTV inspections into 3D semantic maps of the sewer pipe environment, without hardware modification. The rest of the paper

*This work is supported by the European Union's Horizon Europe Research Programme under Grant Agreement No. 101189847 - Pipeon.

[1]Alex George, Lyudmila Mihaylova and Sean Anderson are with the School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, UK ageorge4@sheffield.ac.uk.

is organised as follows. Section II describes the problem formulation and proposed method. The performance of the proposed semantic 3D mapping method is evaluated over real, publicly available data from sewer networks and results are presented in Section III. Section IV discusses the results and outlines directions for future work and Section V concludes the paper.[2]

II. METHODS

A. Problem Formulation

The problem we address here is to estimate a sparse, 3D metric map of N discrete landmarks $\mathscr{M} = \{\mathbf{m}_i\}_{i=1}^{N}$ within a cylindrical pipe of known radius R. We assume that the longitudinal state of the robot x_r is provided by a calibrated tether encoder. We represent the landmarks in the pipe using a 2D state vector (which is efficient compared to standard SLAM that would often use a 6D representation),

$$\mathbf{x}_i = [l_{x,i}, \theta_i]^\top, \tag{1}$$

where $l_{x,i}$ denotes the absolute axial distance from the pipe starting point and θ_i represents the angular orientation. This formulation exploits the known geometry to decouple the high-uncertainty longitudinal mapping from the high-precision angular tracking.

B. Cylinder-to-Image Projection Model

The transformation of a landmark vector $\mathbf{P}_w = [X_w, Y_w, Z_w]^\top$ from the global pipe frame to the image plane $\mathbf{p}_{img} = [u, v]^\top$ is modelled via the pinhole camera projection,

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{K}$ is the intrinsic matrix, and $[\mathbf{R} \quad \mathbf{t}]$ represents the extrinsic transformation from the world origin to the camera centre. The intrinsic matrix is defined as,

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}, \tag{3}$$

where f_x, f_y are respective focal lengths and (c_x, c_y) is the principal point.

In our framework, we simplify this general projection by leveraging the known cylindrical geometry. A landmark at state $\mathbf{x}_i = [l_{x,i}, \theta_i]^\top$ corresponds to a 3D world coordinate

$$\mathbf{P}_w = [R\cos\theta_i, R\sin\theta_i, l_{x,i}]^\top. \tag{4}$$

We assume that the camera is primarily aligned with the longitudinal axis of the pipe and localised at x_r via the tether, which means that the relative depth is $Z_c = l_{x,i} - x_r$. The 3D-to-2D projection then simplifies to

$$\hat{\mathbf{z}}_t = \begin{bmatrix} c_{x,t} + f_x \frac{R\cos\theta_i}{l_{x,i} - x_r} \\ c_{y,t} + f_y \frac{R\sin\theta_i}{l_{x,i} - x_r} \end{bmatrix}, \tag{5}$$

where $\hat{\mathbf{z}}_t = [\hat{u}, \hat{v}]^\top$.

In our implementation, the principal point $\mathbf{c} = [c_{x,t}, c_{y,t}]^\top$ at time step t is not a static calibration constant but a value representing the pipe vanishing point, or Focus of Expansion (FOE). For a tracked pixel (u_i, v_i) with optical flow velocity $(V_{x,i}, V_{y,i})$, the line of motion must geometrically intersect the FOE, yielding the linear constraint

$$-V_{y,i}c_{x,t} + V_{x,i}c_{y,t} = V_{x,i}v_i - V_{y,i}u_i. \tag{6}$$

Evaluating this across the flow field leads to the overdetermined system $\mathbf{Ac} = \mathbf{b}$, where the i-th row of $\mathbf{A}$ is $[-V_{y,i}, V_{x,i}]$ and $b_i = V_{x,i}v_i - V_{y,i}u_i$. To handle noise in the optical flow vectors, we use RANSAC to isolate a valid inlier subset $(\mathbf{A}_{in}, \mathbf{b}_{in})$ [12]. The optimized FOE is then resolved as

$$\mathbf{c}^* = \underset{\mathbf{c}}{\operatorname{argmin}} \|\mathbf{A}_{in}\mathbf{c} - \mathbf{b}_{in}\|_2^2. \tag{7}$$

This dynamic extraction of the vanishing point compensates for minor extrinsic misalignments and robot drift without the need for full 6-DOF pose estimation.

C. Weakly Supervised Semantic Perception

To generate the semantic detections $\mathscr{Z}$ we use a multi-type Fully Convolutional Anomaly Detection (MultiTypeFCDD) system, which avoids the need for semantic pixel-level labels [10], [11]. This front-end transforms input images into class-specific anomaly heatmaps using only image-level binary labels.

1) Multi-Channel Anomaly Heatmaps: Let $\phi(X_t; W) : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{u\times v\times M}$ represent a fully convolutional network with weights W, where M is the number of defect classes. For an input image X_t, the network produces a low-resolution anomaly heatmap $A_k \in \mathbb{R}^{u\times v}$ for each class $k \in \{1, \ldots, M\}$ via a pseudo-Huber transformation,

$$A_k(X_t) = \sqrt{\phi(X_t; W)^2 + 1} - 1. \tag{8}$$

The model is trained using a multi-type objective function that minimises the anomaly score $z_{ik} = \frac{1}{uv}\|A_k(X_i)\|_1$ for an image X_i when defect k is absent ($y_{ik} = 0$) while maximizing the log-likelihood when the defect is present ($y_{ik} = 1$),

$$\mathscr{L} = \frac{1}{MN} \sum_{i=1}^{N} \sum_{k=1}^{M} \left[(1 - y_{ik}) z_{ik} - y_{ik} \log(1 - e^{-z_{ik}}) \right]. \tag{9}$$

2) Detection Extraction: The heatmaps A_k are upsampled to the original image dimensions using bilinear interpolation to produce A'_k. Unlike fully supervised methods that output bounding boxes, this approach generates localisations where the pixel intensity represents the anomaly probability. To interface with the EKF backend, we extract discrete detections $\mathbf{z}_j = [u, v]^\top$ by identifying the centroids of connected components in the thresholded heatmap $A'_k > \tau$. This allows the system to operate as a weakly supervised semantic segmentor, identifying the 2D image-plane coordinates of landmarks without requiring expert-labeled bounding boxes during training.

[2]Code available at https://github.com/alexgeorge13/Sewer-Pipe-3D-Mapping.

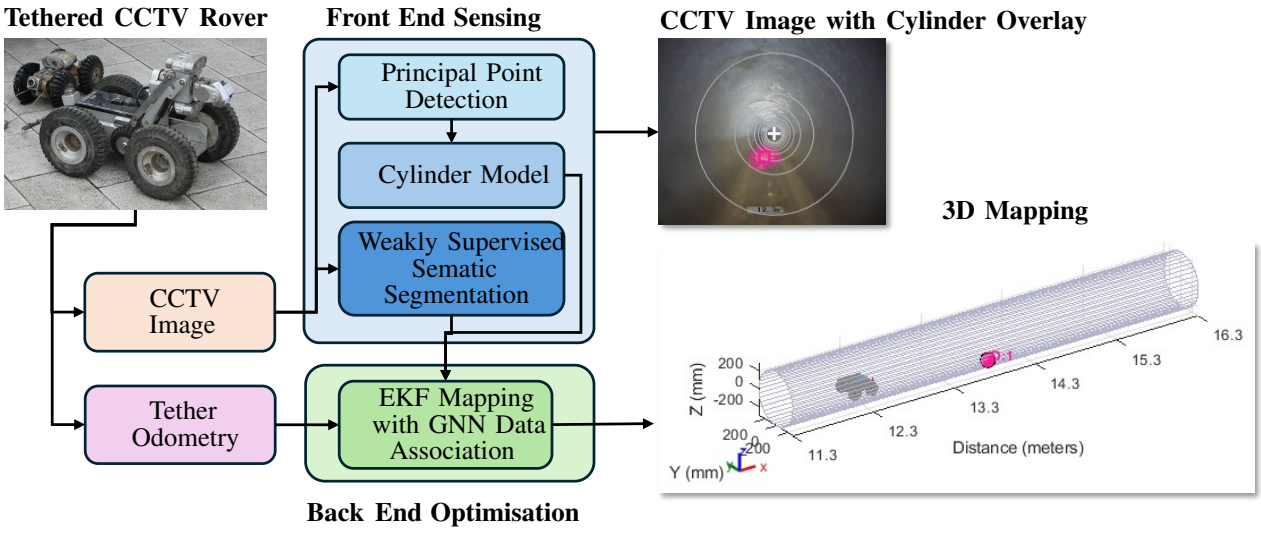

Fig. 1: System Architecture.

D. Structure-Only EKF Mapping

To estimate landmark locations, we use a bank of independent Extended Kalman Filters (EKFs), where each EKF tracks the location for a single landmark i. The EKF only estimates the landmark locations and not the location of the robot x_r because we assume that this is known from the robot tether; this is significantly more computationally efficient than full 6-DOF vSLAM.

1) State-Space Representation: We approach the mapping problem as a stationary landmark estimation task. For each landmark i, the state and measurement equations are defined as

$$\mathbf{x}_{i,t} = \mathbf{F}\mathbf{x}_{i,t-1} + \mathbf{w}_t, \tag{10}$$

$$\mathbf{z}_{i,t} = h(x_{r,t}, \mathbf{x}_{i,t}) + \mathbf{v}_t, \tag{11}$$

where the state vector is $\mathbf{x}_i = [l_{x,i}, \theta_i]^\top$, $\mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q})$ is state noise and $\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{R})$ is measurement noise arising from the camera observations. We assume landmarks are physically static, hence the state transition matrix $\mathbf{F}$ is the identity matrix $\mathbf{I}$, and the process noise $\mathbf{Q}$ is minimal. The non-linear measurement function $h(\cdot)$ is the cylinder-to-image projection derived in Eq. 5.

2) Measurement Linearisation: To update the posterior estimate, we linearise $h(x_{r,t}, \mathbf{x}_{i,t})$ by computing the Jacobian matrix $\mathbf{H}_{i,t}$ with respect to the landmark state $\mathbf{x}_i$. Let $\Delta x = l_{x,i} - x_{r,t}$ represent the relative axial distance. The Jacobian is given by

$$\mathbf{H}_{i,t} = \begin{bmatrix} -\frac{f_x \cdot R\cos\theta_i}{\Delta x^2} & -\frac{f_x \cdot R\sin\theta_i}{\Delta x} \\ -\frac{f_y \cdot R\sin\theta_i}{\Delta x^2} & \frac{f_y \cdot R\cos\theta_i}{\Delta x} \end{bmatrix} \tag{12}$$

This matrix maps the structural uncertainty in the 3D cylindrical frame into the 2D image coordinate space.

3) Estimation and Update: When a detection $\mathbf{z}_{i,t}$ is associated with landmark i, the filter executes a recursive update. Given the previous mean estimate $\hat{\mathbf{x}}_{i,t-1}$ and error covariance $\mathbf{P}_{i,t-1}$, the following steps are performed:

1) **Innovation Calculation:** Compute the residual between the detected and predicted image-plane position,

$$\mathbf{y}_{i,t} = \mathbf{z}_{i,t} - h(x_{r,t}, \hat{\mathbf{x}}_{i,t-1}) \tag{13}$$

2) **Innovation Covariance:** Project the state uncertainty and sensor noise into the measurement space,

$$\mathbf{S}_{i,t} = \mathbf{H}_{i,t}\mathbf{P}_{i,t-1}\mathbf{H}_{i,t}^\top + \mathbf{R} \tag{14}$$

3) **Kalman Gain:** Calculate the optimal weighting factor,

$$\mathbf{K}_{i,t} = \mathbf{P}_{i,t-1}\mathbf{H}_{i,t}^\top \mathbf{S}_{i,t}^{-1} \tag{15}$$

4) **Posterior Update:** Update the landmark estimate and its associated covariance,

$$\hat{\mathbf{x}}_{i,t} = \hat{\mathbf{x}}_{i,t-1} + \mathbf{K}_{i,t}\mathbf{y}_{i,t} \tag{16}$$

$$\mathbf{P}_{i,t} = (\mathbf{I} - \mathbf{K}_{i,t}\mathbf{H}_{i,t})\mathbf{P}_{i,t-1} \tag{17}$$

E. Data Association and Track Management

The system associates incoming semantic detections $\mathscr{Z} = \{\mathbf{z}_j\}_{j=1}^{M}$ with existing map tracks $\mathscr{M}$ using a Global Nearest Neighbor (GNN) strategy. For each potential detection-track pair, we compute the Mahalanobis distance to account for the coupled uncertainty of the landmark estimate and sensor noise,

$$d_{i,j}^2 = (\mathbf{z}_j - \hat{\mathbf{z}}_{i,t})^\top \mathbf{S}_{i,t}^{-1} (\mathbf{z}_j - \hat{\mathbf{z}}_{i,t}), \tag{18}$$

where $\hat{\mathbf{z}}_{i,t}$ and $\mathbf{S}_{i,t}$ are the predicted measurement and innovation covariance for track i, respectively. An assignment is considered valid only if $d_{i,j}^2 < \gamma$, where γ is a χ^2 threshold for a 2-DOF system (typically $\gamma = 9.21$ for 99% confidence). The optimal assignment that minimises the total Mahalanobis distance is obtained by solving the linear assignment problem using the Duff-Koster algorithm [13].

F. Geometric Back-Projection for Initialisation

Detections that remain unassociated after the GNN step are used to initialise new landmark candidates via direct geometric back-projection. Given an unassociated 2D detection $\mathbf{z}_j = [u, v]^\top$ and the current principal point $\mathbf{c}_t$, we calculate the radial displacement ρ as

$$\rho = \sqrt{\frac{(u - c_{x,t})^2}{f_x^2} + \frac{(v - c_{y,t})^2}{f_y^2}}. \tag{19}$$

The initial state $\hat{\mathbf{x}}_{new} = [\hat{l}_x, \hat{\theta}]^\top$ is then computed by inverting the cylindrical projection

$$\hat{l}_x = x_r + \frac{R}{\rho}, \tag{20}$$

$$\hat{\theta} = \text{atan2}\left(\frac{v - c_{y,t}}{f_y}, \frac{u - c_{x,t}}{f_x}\right). \tag{21}$$

To ensure map robustness, new tracks are held in a candidate state and only promoted to confirmed status in the global map $\mathscr{M}$ after N_{min} persistent detections. This pruning logic, combined with a temporal decay for inactive tracks, effectively filters out transient visual anomalies and false positives from the deep perception layer.

G. Integrated Tracking Algorithm

The complete mapping pipeline is summarised in Algorithm 1. The system maintains a dynamic map $\mathscr{M}$ by coupling real-time perception with the cylindrical EKF backend. By solving the linear assignment problem at each timestep, the system ensures that landmark tracks are updated with the statistically most likely detections, while the geometric back-projection allows for the immediate initialisation of new features as they enter the camera's field of view.

Algorithm 1: Lightweight Cylindrical Semantic Mapping

Init: Global Map $\mathscr{M} = \emptyset$, Gate γ, Min Hits N_{min}
for each time step t **do**
  Predict: Update axial pos. $x_{r,t}$ via tether encoder
  Percept: Detections $\mathscr{Z} = \{\mathbf{z}_j\}_{j=1}^{M}$, centre $\mathbf{c}_t$
  1. Data Association (GNN):
  Cost Matrix $\mathbf{C} \in \mathbb{R}^{M \times |\mathscr{M}|}$
  for detection j and track $i \in \mathscr{M}$ **do**
    Predict $\hat{\mathbf{z}}_{i,t}$, covariance $\mathbf{S}_{i,t}$ (Eq. 12, 15)
    Mahalanobis dist: $d_{i,j}^2 \leftarrow (\mathbf{z}_j - \hat{\mathbf{z}}_{i,t})^\top \mathbf{S}_{i,t}^{-1} (\mathbf{z}_j - \hat{\mathbf{z}}_{i,t})$
    Update cost: $\mathbf{C}_{j,i} \leftarrow d_{i,j}^2$ if $d_{i,j}^2 < \gamma$ else ∞
  end for
  Match pairs (j, i) minimising $\sum d_{i,j}^2$
  2. EKF Update:
  for matched pair (j, i) **do**
    Compute Jacobian $\mathbf{H}_{i,t}$ & Kalman Gain $\mathbf{K}_{i,t}$
    Update $\hat{\mathbf{x}}_{i,t}$, $\mathbf{P}_{i,t}$ via EKF (Eq. 17-18)
    $hits_i \leftarrow hits_i + 1$
  end for
  3. Track Management:
  for unassociated detection j **do**
    $\rho \leftarrow \sqrt{((u_j - c_{x,t})/f_x)^2 + ((v_j - c_{y,t})/f_y)^2}$
    $\hat{l}_x \leftarrow x_{r,t} + R/\rho$
    $\hat{\theta} \leftarrow \text{atan2}((v_j - c_{y,t})/f_y, (u_j - c_{x,t})/f_x)$
    $\mathscr{M} \leftarrow \mathscr{M} \cup \{[\hat{l}_x, \hat{\theta}]^\top, \mathbf{P}_{init}, hits = 1\}$
  end for
  Prune: Drop track i if unseen for T frames & $hits_i < N_{min}$
end for

III. Experimental Results

We evaluated the proposed lightweight mapping method with a series of experiments using two publicly available data sources: real-world CCTV video recorded from a tethered crawler deployed in a sewer pipe [14] and the WRc dataset [9] of sewer defects with image-level labels only.

A. Dynamic Principal Point Estimation

We observed during visual inspection of the CCTV video that the principal point was often off-centre of the image frame due to misalignment of the robot with the pipe. This motivated the development of the dynamic principal point estimation. In order to quantitatively assess the benefit of this estimation, we compared the approach to simply taking the image centre as the principal point using frames extracted from the CCTV video [14]. We evaluated the performance of both methods using the Mean Drift Error (Euclidean Distance), Normalised Mean Error (NME), and the Area Under Percentage of Correct Keypoints (PCK AUC) at a normalised pixel threshold of 0-5% along the image diagonal. The results shown in Table I (reduction in NME, increase in PCK AUC) demonstrate that estimating the principal point from optical flow greatly improves over the assumption of a static point, which is important to obtain an accurate and robust cylinder model.

TABLE I: Ablation Study: Dynamic Centre Estimation

Configuration	Mean Drift Error (px)	NME (%)	PCK AUC (0-5%)
Static Centre	32.74	3.55	0.29
Dynamic	12.56	1.36	0.73

B. Semantic Perception Performance

In this section, we evaluate the performance of the front-end weakly supervised semantic segmentation system based on MultiTypeFCDD [11]. This system was trained using 5,000 images from the publicly available WRc dataset [9], on five commonly occurring defects in sewer pipes: Connection, Deposit, Displaced Joint, Fracture, and Roots. We used the same training split as in [15], of 70:10:20 for training, calibration, and testing. The MultiTypeFCDD system was setup with an Inception-v3 backbone. The network was trained using the Adam optimizer for a maximum of 500 epochs, utilising a mini-batch size of 32 and an initial learning rate of 10^{-4}.

As shown in Table II, the MultiTypeFCDD front-end demonstrates robust discriminative performance, yielding AUROC above 0.93 across all classes. Visual examples are provided in Fig. 3, which illustrates class-specific semantic heatmaps overlaid onto representative images from each class.

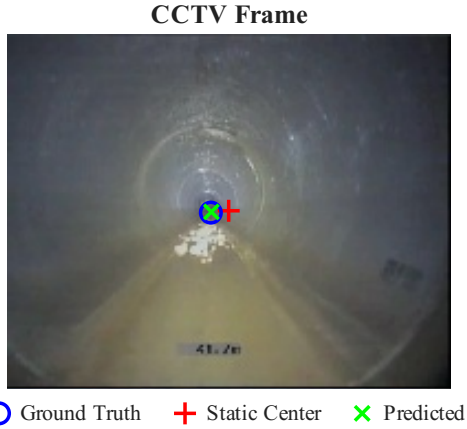

Fig. 2: Impact of principal point adaptivity on mapping geometry. The dynamic estimation compensates for robot pitch and lateral offset, ensuring landmarks adhere to the cylindrical manifold.

TABLE II: Semantic Perception Performance

Class	Precision	Recall	F1-Score	AUROC
Connection	0.71	0.88	0.79	0.96
Deposit	0.70	0.74	0.72	0.93
Displaced Joint	0.69	0.94	0.80	0.97
Fracture	0.74	0.74	0.74	0.93
Roots	0.75	0.87	0.81	0.97

C. Integrated System: Real-World Sewer Inspection

Finally, we evaluated the full integrated system on CCTV video recorded from a tethered crawler deployed in a 600mm diameter concrete sewer pipe [14]. The aim was to reconstruct a 3D semantic asset map over a 28-meter segment containing six connections and two patches of deposits. Ground truth was obtained by manual inspection of the CCTV video. As detailed in Table III, the EKF-based mapping system successfully localized four of the six ground-truth connections (FP = 1, FN = 2) and both deposits (TP = 2). Both classes were accurately oriented along the pipe circumference, yielding mean radial errors of just 16.14° for connections and 10.13° for deposits. Longitudinally, discrete features like connections were mapped with a mean error of 0.12 m. Deposits were distributed over a wide area, so their longitudinal accuracy was evaluated based on their spatial centroids, which led to a mean error of 0.44 m. Overall, the reconstructed 3D semantic map aligns well with the ground-truth, capturing spatial arrangement and orientation of assets within the pipe (Fig. 4). When evaluated on an Intel Core i5-14400F processor, the proposed framework processes each frame in just 36.19 ms, achieving an operational throughput of 27.63 frames per second (FPS).

IV. Discussion

This work developed a 3D semantic mapping system that only made use of CCTV and tether odometry, as is standard in sewer industry inspection systems. We demonstrated each component of the system on real-world data to show that the task could be successfully accomplished. The main strengths of the method are that it uses a cylinder model to efficiently perform the mapping and weak semantic segmentation to avoid the need for large-scale labelled datasets for training the front-end perception system. The main issue that we encountered when transferring the system from the offline training images to evaluation on CCTV was the occasional false positive. This reflects the fact that video inspection in sewer pipes is a challenging problem due to the low-texture environment and visual noise.

TABLE III: 3D Semantic Mapping Performance in Real-World Deployment.

Class	TP	FP	FN	Longitudinal Error (m)	Radial Error (deg)
Connection	4	1	2	0.12	16.14
Deposit	2	0	0	0.44	10.13
Roots	0	3	0	-	-

The main advantage of the proposed framework is its computational efficiency. Alternatives like VILL-SLAM have the advantage that they can produce sub-millimetre accuracy [3], but they do incur higher cost in terms of number of sensors, complexity and computational processing. Methods based on Structure-from-Motion (SfM) and Dense Reconstruction [4] are inherently $O(K^3)$ or $O(K^2 \log K)$ relative to the number of tracked keypoints K. The approach proposed here, based on the cylinder-informed EKF, scales linearly as $O(M \times N)$, where M is the number of semantic detections per frame and N is the number of active tracks in the map. This reduced computational complexity directly enables real-time execution. We found that the execution time was on the order of 27 FPS, but this was dominated by the deep semantic segmentation network (79% of the computation time). However, this could be improved because the network does not need to execute on every frame due to the slow-moving nature of the crawler and the model size could be reduced using model compression techniques.

Future work will focus on improving the robustness of front-end detection to reduce false positives. We also plan to extend this framework from structure-only mapping to a full Graph-SLAM implementation, which will allow us to account for tether uncertainty and incorporate loop-closure constraints. Additionally, we aim to integrate topographical geographic information systems (GIS) priors to align 3D cylindrical segments within a global coordinate system.

V. Conclusions

This paper presented a lightweight semantic 3D mapping framework specifically engineered for the structural geometry of sewer pipes. The approach used a cylindrical projection model to achieve real-time 3D asset mapping. The components of the system were evaluated on separate real-world datasets to show the benefit of each part. The results from real-world sewer pipe CCTV inspection demonstrated effective 3D semantic mapping with centimetre-scale accuracy.

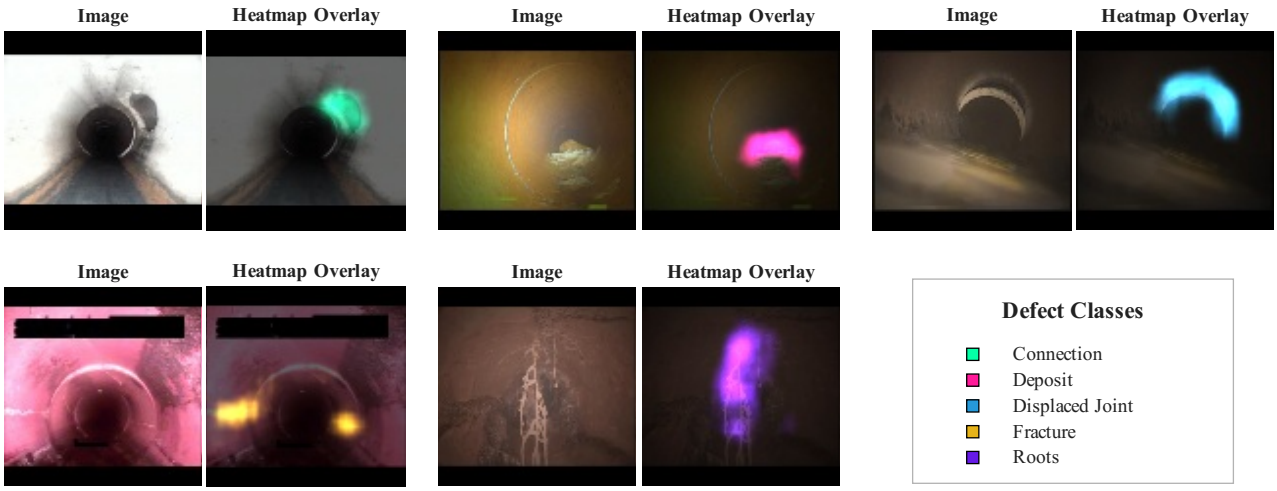

Fig. 3: Weakly Supervised Semantic Segmentation: 3D semantic perception of WRc images.

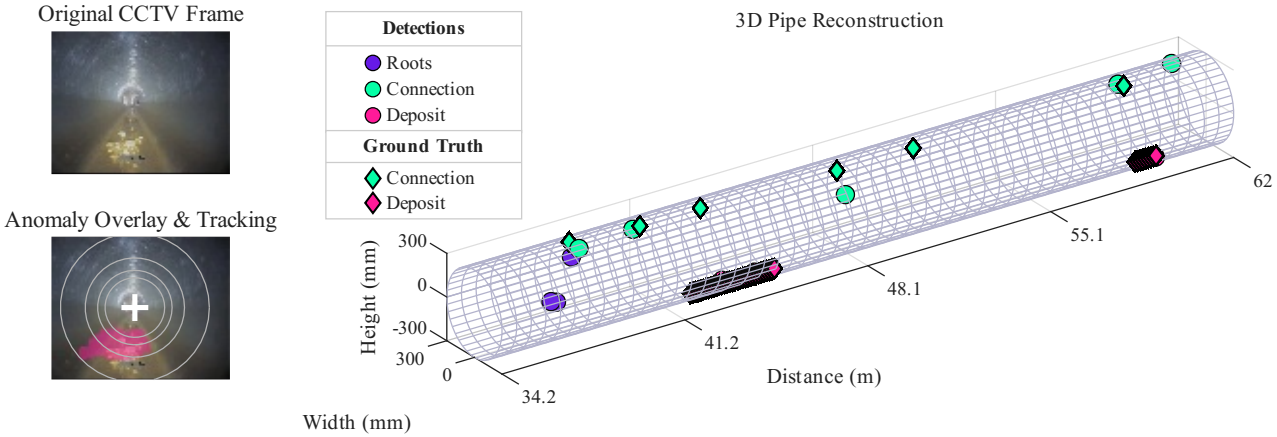

Fig. 4: Full system output: 3D semantic reconstruction of a 28 metre sewer segment.

REFERENCES

[1] P. Hansen, H. Alismail, P. Rander, and B. Browning, "Visual mapping for natural gas pipe inspection," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 532–558, 2015.

[2] R. Zhang, R. Worley, S. Edwards, J. Aitken, S. R. Anderson, and L. Mihaylova, "Visual simultaneous localization and mapping for sewer pipe networks leveraging cylindrical regularity," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3406–3413, 2023.

[3] T. Tian, L. Wang, X. Yan, F. Ruan, G. J. Aadityaa, H. Choset, and L. Li, "Visual-Inertial-Laser-Lidar (VILL) SLAM: Real-Time Dense RGB-D Mapping for Pipe Environments," in *Proc. of the IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS)*, 2023, pp. 1525–1531.

[4] R. Liu and Z. Shao, "Defect Detection and 3D Reconstruction of Complex Urban Underground Pipeline Scenes for Sewer Robots," *Sensors*, vol. 24, no. 23, p. 7557, 2024.

[5] J. T. Jung and A. Reiterer, "Improving sewer damage inspection: Development of a deep learning integration concept for a multi-sensor system," *Sensors*, vol. 24, no. 23, p. 7786, 2024.

[6] 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, p. 103383, 2020.

[7] 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, p. 104494, 2022.

[8] J. B. Haurum and T. B. Moeslund, "Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 451–13 462.

[9] Spring, "Spring Innovation - Accelerating Water Sector Transformation," https://spring-innovation.co.uk, last accessed 2026/03/28.

[10] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K.-R. Müller, "Explainable deep one-class classification," *arXiv preprint arXiv:2007.01760*, 2021.

[11] A. George, L. Mihaylova, and S. Anderson, "Explainable deep convolutional multi-type anomaly detection," *arXiv preprint arXiv:2511.11165*, 2025.

[12] N. Van der Stap, R. Reilink, S. Misra, I. Broeders, and F. Van der Heijden, "The use of the focus of expansion for automated steering of flexible endoscopes," in *Proc. of the 4th IEEE RAS & EMBS International Conf. on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2012, pp. 13–18.

[13] I. S. Duff and J. Koster, "On algorithms for permuting large entries to the diagonal of a sparse matrix," *SIAM Journal on Matrix Analysis and Applications*, vol. 22, no. 4, pp. 973–996, 2001.

[14] S. Edwards, R. Zhang, R. Worley, L. Mihaylova, J. Aitken, and S. R. Anderson, "A robust method for approximate visual robot localization in feature-sparse sewer pipes," *Frontiers in Robotics and AI*, vol. 10, p. 1150508, 2023.

[15] A. George, W. Shepherd, S. Tait, L. Mihaylova, and S. Anderson, "A deep learning benchmark analysis of the publicly available WRc dataset for sewer defect classification," in *Proceedings of 21st Computing & Control for the Water Industry Conference (CCWI 2025)*. Sheffield, 2025.