3D-PipeCLIP: Leveraging Geometric-Language Alignment for Sewer Defect Classification from Point Cloud Data

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

***Abstract*— Automated 3D inspection of sewer pipes using LiDAR or Time-of-Flight sensors offers advantages for defect detection. However, geometric point cloud classifiers often fail to generalize from synthetic training data to noisy real-world scans. To overcome the need for very large, expert-annotated physical datasets, we propose a framework that anchors 3D feature extractors to geometric language descriptions. In this approach, a dynamic graph convolutional neural network (DGCNN) is used to extract embedding features from point clouds. During inference, the system compares these point cloud embeddings to text embeddings of geometric descriptors using cosine similarity. This differs substantially from standard hard classification, which would use a fully connected layer to a softmax head. The system is trained by minimising the contrastive language-image pretraining (CLIP) loss. In a zero-shot transfer setting from synthetic to physical laboratory scans, our method gives an F1-score of 37.86% on the real data, which doubles the performance of a baseline previously published using the same DGCNN without language anchors. When we apply a few-shot fine-tuning protocol based on real data, the model produces a weighted F1-score of 57.81% on independent real data. This places our approach near state-of-the-art methods (62.58%) that have also been trained with cross-domain knowledge of one million plus images. Our approach, therefore, is appealing for its data efficiency and also highlights for the first time the utility of geometric text anchors in this domain.**

I. INTRODUCTION

Sewer networks are an essential part of urban and extra-urban infrastructure but are prone to damage such as cracks and structural defects. Standard inspection for defects is often conducted using vision-based methods with Closed-Circuit Television (CCTV) cameras [1], [2]. One disadvantage of this is that camera images are limited to the 2D image plane with no depth perception. There are a number of defect types, with discernible 3D cues, that would likely benefit from depth sensing, including displaced joints, cracks or blockages. Therefore, it is a subject of some interest to integrate 3D sensors such as LiDAR and Time-of-Flight (ToF) cameras onto pipe crawlers with CCTV [3], [4]. Recent work also demonstrates that low-cost LiDAR can effectively identify joints and blockages in real-time within sewer pipes [5].

Deep learning methods have been used for 3D inspection tasks in sewers to identify defects [4], with standard backbones like PointNet [6] and dynamic graph convolutional neural network (DGCNN) [7]. Researchers have also used Transformer architectures like TransPCNet [8], [9], which leverage attention mechanisms to describe geometrical relationships over long-range. A more recent state-of-the-art method is Sewer-ViT [10]. This approach uses 2D self-supervised learning on large sewer image datasets, with over a million images from the Sewer-ML dataset [11]. It then transfers that visual knowledge over to the 3D point cloud domain.

Labelled data for training machine learning systems is scarce in this domain. This is why researchers often use synthetically generated point clouds to train 3D backbones to solve the problem of limited real-world data [4]. A drawback of using synthetic data is that simulated shapes are noise-free and non-diverse, especially when compared to real pipes. Real-world sensors capture noisy data, creating a challenging synthetic-to-real domain gap to overcome. Standard supervised classifiers trained purely on this type of simulated data tend to perform poorly in this context, achieving F1-scores of less than 25% [4].

These results suggest that standard hard classification is not well-suited to solving the defect classification problem from synthetic point cloud data. If we look elsewhere in the literature for inspiration, in the 2D video domain, the PipeCLIP framework has achieved success through bypassing visual dataset limitations by aligning video features with expert prior descriptions using language-vision models [12]. This approach taught the network to match video frames to specific text-based defect attributes, which gave the model a higher-level semantic understanding of the scene. We suggest that using a similar approach for 3D point cloud data will leverage a similar advantage, which is the research gap we investigate here.

In this paper, we propose a method for sewer defect classification from 3D point clouds, where we align extracted 3D point cloud features directly to custom language descriptions based on the geometry of the pipe, under normal and defect conditions. This enables the network to acquire a physical understanding of pipeline anomalies that translates directly from synthetic simulation to physical laboratory scans.

The paper is structured as follows. In Section II we briefly describe the benchmark dataset used here, which is from a separate study [4]. In Section III we describe our proposed approach for leveraging text prompts with 3D point clouds for defect recognition (Figure 1). In Section IV we present the experimental results using both zero-shot classification and fine-tuning with comparison to baseline performance on a publicly available dataset. In Section V we discuss the results and in Section VI we conclude the paper.

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

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

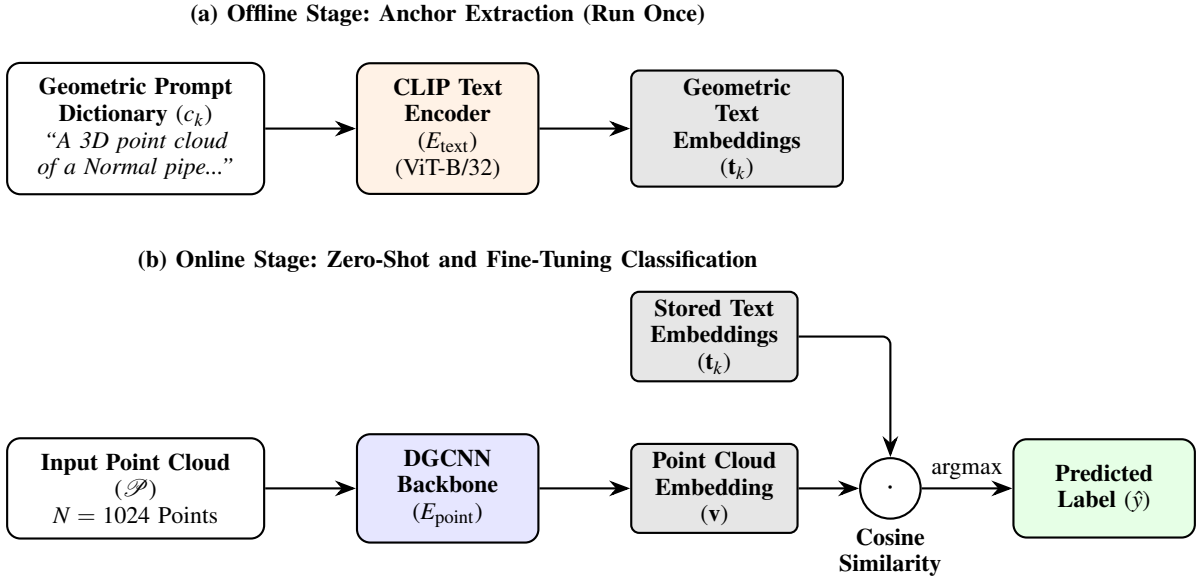

Fig. 1: The 3D-PipeCLIP Framework Overview. The approach operates in two stages: (a) An offline stage where human-written geometric descriptions are passed through the CLIP text encoder to generate static embedding anchors. These are locked and saved to memory just once. (b) An online stage where a system processes real point clouds uses a lightweight DGCNN to extract shape embeddings. These vectors are compared to the stored text embeddings via cosine similarity to classify defects.

II. DATASET AND BENCHMARK ORIGIN

We benchmark our models against the public sewer defect point cloud dataset from Haurum et al. [4] to evaluate the geometric-language alignment framework.[2] This dataset contains both synthetic and physical lab-captured point clouds. It was created because public 3D data for automated sewer inspection is hard to find.

A. Synthetic Data Domain

The synthetic data comes from a simulation environment where random sewer networks were generated. These simulations used perfect, clean PVC pipe geometries without any fluid or sediment. A virtual depth sensor was moved through the synthetic pipes to record the simulated point cloud frames. The simulator included just one defect class in each point cloud to keep the classification target clear.

There are three defect shapes in the dataset in addition to normal pipes (Figure 2):

- **Normal Data:** Cylindrical pipe sections with no features.
- **Displaced Joints:** An offset where two pipe segments meet.
- **Rubber Rings:** Failed or peeling seals at the joints.
- **Bricks:** Blocky shapes sitting on the pipe floor.

B. Physical Data Domain and Sensor Constraints

To test on real data, researchers mapped a physical indoor laboratory setup. They used real PVC sewer pipe segments with a diameter of 400 mm in straight and corner layouts.

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

The physical 3D point clouds were captured with a PMD Pico Flexx depth camera. This is a compact, flash-based 3D Time-of-Flight (ToF) sensor. The sensor floods the scene with 850 nm infrared illumination from a Vertical-Cavity Surface-Emitting Laser (VCSEL) array. It calculates distance by measuring the phase shift of the reflected light.

This phase-shift technology creates unique noise characteristics inside a physical pipe. The benchmark authors noted distance errors and empty holes in the point clouds where the sensor failed to return depth data. These missing points are usually caused by external ambient lighting interference and the shiny reflectivity of the PVC walls. Therefore, moving from clean synthetic models to real data creates a difficult-to-solve domain gap.

C. Evaluation Splits and Distribution

The dataset, as illustrated in Table I, exhibits a heavy class imbalance in both simulated and physical environments. Non-defective, standard pipe cylinders constitute approximately 50% of the entire database. The specific anomaly classes share a nearly uniform distribution of approximately 16% to 17% each. In order to prevent the alignment network from developing a large majority-class bias toward predicting 'Normal' geometries, we calculated inverse class frequencies to weight the loss gradient during all training operations.

In order to make the best use of the large synthetic dataset and smaller real dataset, we split the data so that all the synthetic data was used during the initial training phase, then used the real data for validation, fine-tuning and testing. The corresponding data splits are shown in Table II. The majority of the real data was reserved as an independent test set.

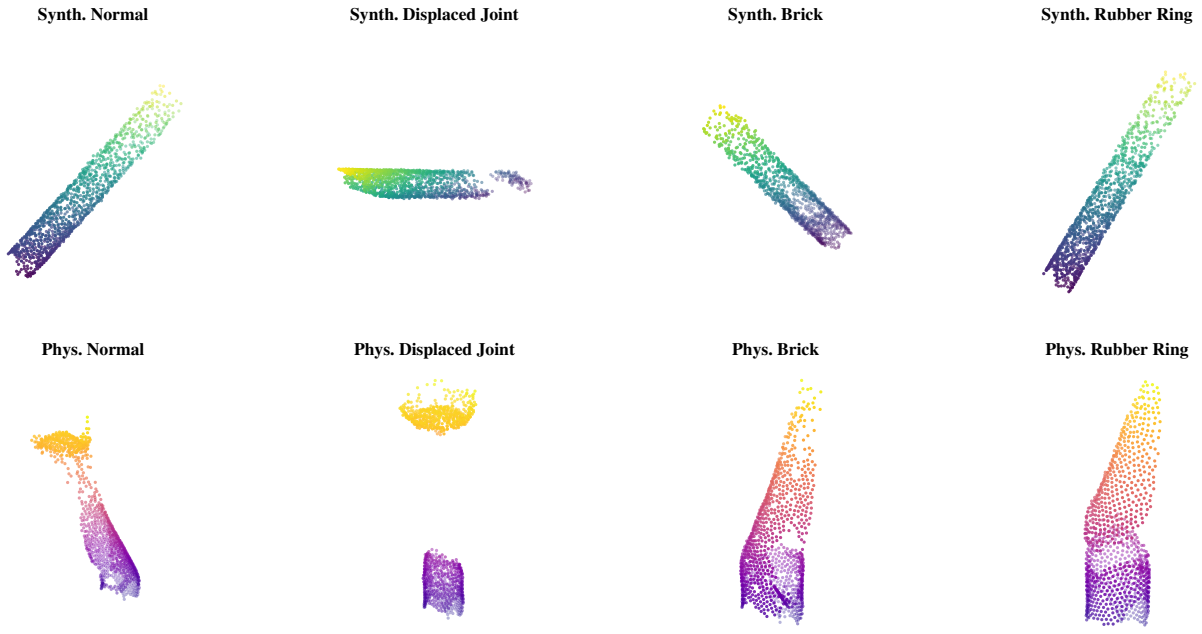

Fig. 2: Visual comparison of 3D point cloud samples across synthetic training data and physical testing data for the four pipeline states.

TABLE I: Distribution of Point Cloud Samples Across Classes.

Defect Class	**Synthetic Data**	**Real Data**
Normal	8,100 (50.00%)	415 (50.2%)
Displaced Joint	2,700 (16.66%)	142 (17.2%)
Brick Obstruction	2,700 (16.66%)	133 (16.1%)
Rubber Ring	2,700 (16.67%)	137 (16.6%)
Total Samples	**16,200** (100%)	**827** (100%)

TABLE II: Distribution of Point Cloud Samples Across Initial Training, Validation, Fine-Tuning and Testing Splits.

Split Phase	**Synthetic Data**	**Real Data**
Initial Training	16,200	0
Validation	0	68
Fine-Tuning	0	274
Testing	0	485
Total Samples	**16,200**	**827**

III. PROPOSED METHODOLOGY

The objective of the 3D-PipeCLIP framework is to map raw 3D point cloud representations directly into a shared multimodal latent space. Instead of using standard integer classification mapping, this method uses language-anchored target vectors. This section details the text prompt engineering protocol, the structural alignment of the point cloud feature extractor, and the inverse-frequency weighted training objective.

A. Geometric Prompt Engineering

We avoid the use of traditional hard-label classification in favour of language-anchored target vectors to help the model understand pipeline anomalies. To do this, we have developed a set of $M = 4$ geometric descriptions, represented as $\mathscr{C} = \{c_k\}_{k=1}^{M}$, which define the specific shape of pipe defects. Table III lists the exact prompts used for each class. For example, a non-defective pipeline is described as *"A 3D point cloud of a Normal pipe..."* whereas a brick defect is described as *"A 3D point cloud of Defective brickwork..."*

The textual descriptions c_k from Table III are passed offline through the pre-trained text transformer from the CLIP (Contrastive Language-Image Pre-training) ViT-B/32 model (Vision Transformer Base with 32x32 patches) [13], denoted as $E_{\text{text}}(\cdot)$. This model produces a normalized text embedding $\mathbf{t}_k \in \mathbb{R}^D$,

$$\mathbf{t}_k = \frac{E_{\text{text}}(c_k)}{\|E_{\text{text}}(c_k)\|_2} \tag{1}$$

where $D = 512$ is the dimension of the CLIP latent space. These extracted embeddings are locked as static geometric anchors (and ViT-B/32 is not used again for online processing).

B. Point Cloud Feature Extraction and Alignment

The input is a normalised point cloud $\mathscr{P} \in \mathbb{R}^{3\times N}$ consisting of $N = 1024$ points. We use a Dynamic Graph Convolutional Neural Network (DGCNN) as the point cloud backbone, denoted $E_{\text{point}}(\cdot)$; note that this is exactly the same DGCNN used in the baseline from [4] obtained from

TABLE III: The Expert Linguistic Geometric Prompts Used to Align the 3D Multimodal Latent Space.

Target Class	Linguistic Geometric Prompt (c_k)
Normal Pipe	A 3D point cloud of a Normal pipe. The walls form a continuous, smooth, uniform cylinder with no internal protrusions or breaks.
Displaced Joint	A 3D point cloud of a Displaced joint. There is a distinct, abrupt step or offset in the pipeline geometry where two cylindrical sections fail to align perfectly.
Rubber Ring	A 3D point cloud of Interface material delamination. A thin, subtle strip or ribbon of material points peeling or hanging away from the internal joint of the cylinder walls.
Brick	A 3D point cloud of Defective brickwork or masonry. A massive, blocky, square-like obstruction of points sitting heavily on the pipe floor.

their online repository[3]. The DGCNN constructs a K-nearest neighbour graph in the feature space for each point. This graph structure lets the network capture local shapes and detailed defect geometries.

The output from the DGCNN backbone is passed through a linear projection layer. This step aligns the dimensions of the point cloud features with the text embedding space, yielding a continuous normalised point cloud embedding $\mathbf{v} \in \mathbb{R}^D$,

$$\mathbf{v} = \frac{E_{\text{point}}(\mathscr{P})}{\|E_{\text{point}}(\mathscr{P})\|_2} \tag{2}$$

In a zero-shot setting, the predicted label $\hat{y}$ for a point cloud is found by calculating the cosine similarity between the point cloud embedding $\mathbf{v}$ and the $M = 4$ text embeddings $\mathbf{t}_k$,

$$\hat{y} = \underset{k\in\{1,\dots,M\}}{\text{argmax}} \ (\mathbf{v}^\top \mathbf{t}_k) \tag{3}$$

C. Weighted Training Objective

In order to train the system, we first performed large-scale training using synthetic data, followed by few-shot fine-tuning on a smaller real-world dataset, where all data was from [4]. One issue we encountered was instability when training initially on the synthetic data, where the model would reduce the loss as intended, but do this by switching accuracies between classes. To account for this, we used an independent subset of the real data as a validation set to monitor and checkpoint the model.

The training objective was based on the standard CLIP Cosine Embedding Loss formulation [13], which measures the cosine distance between the predicted point cloud vector $\mathbf{v}$ from the DGCNN network and the target text vector $\mathbf{t}$,

$$\mathscr{L} = \frac{1}{B}\sum_{b=1}^{B} \omega_b \left(1 - \mathbf{v}_b^\top \mathbf{t}_{y_b}\right) \tag{4}$$

where B is the batch size of the training samples, the DGCNN model predicts a point cloud embedding $\mathbf{v}_b$ for sample b, and $\mathbf{t}_{y_b}$ represents the ground truth geometric text embedding for that sample. The term ω_b is a normalised class-specific weighting that we applied due to the fact that the dataset had a strong class imbalance.

The class weight ω_b was based on inverse frequency, designed to penalise minority class errors more severely. It was defined by taking the total number of occurrences, n_{y_b}, of the defect class y_b present in the training subset, and then computing the raw unnormalized sample weight as $\tilde{\omega}_b = N_{\text{total}}/n_{y_b}$, where N_{total} was the total size of the training split. The raw weights were then normalised to have a mean of one to obtain ω_b.

[3]https://bitbucket.org/aauvap/sewer3dclassification/src/master/

IV. Experimental Results

To evaluate the proposed 3D-PipeCLIP framework, we conducted a series of comparative experiments focusing on zero-shot transferability and few-shot fine-tuning performance. We followed the established benchmark protocols defined by Haurum et al. [4], focusing our analysis on the real-world test split. This split is characterised by physical ToF sensor noise and severe class imbalances, which is a challenging synthetic-to-real domain gap for a machine learning system.

We evaluated our model against the baseline PointNet and DGCNN networks proposed in [4] and also benchmarked against in [10]. We focused on two of the four original data scenarios originally defined in the AAU benchmark [4], which were the ones that used synthetic data for training, and real data for testing:

- **Scenario 1 (S1):** Training on synthetic data with direct, zero-shot evaluation on the physical test split.
- **Scenario 2 (S4):** Pre-training on synthetic data followed by fine-tuning on the available real-world training subset.

A. Zero-Shot Transfer Performance (Scenario 1)

The first stage of evaluation assesses the zero-shot generalization capabilities of the models trained exclusively on synthetic data. It is evident that standard geometric, hard classifiers perform poorly when trained on synthetic data and then test on real data, that will incorporate physical sensor noise (Table IV). The results in Table IV for the baseline hard classifiers PointNet and DGCNN are taken directly from [4] and show that PointNet generates a nominal recall of 15.88% and an F1-score of just 5.25%, effectively failing to identify any defect geometry. The same is true of DGCNN, which although slightly better performing, still produces a low F1-score of 17.65%.

In contrast, the proposed 3D-PipeCLIP framework achieves a balanced weighted precision of 46.52% and a recall of 39.18%, translating to a weighted F1-score of 37.86%, which although in absolute terms may be judged low, is a large improvement on the best baseline zero-shot

TABLE IV: Performance Comparison on the Real Data Test Split. All metrics represent the class-proportional weighted average.

Model	Scenario	Precision	Recall	F1-Score
PointNet [4]	1 (S1)	3.58%	15.88%	5.25%
DGCNN [4]	1 (S1)	29.02%	20.62%	17.65%
PointNet [4]	2 (S4)	23.17%	27.42%	24.24%
DGCNN [4]	2 (S4)	39.69%	26.19%	23.58%
3D-PipeCLIP	1 (S1)	46.52%	39.18%	37.86%
3D-PipeCLIP	2 (S4)	**58.88%**	**57.32%**	**57.81%**

score of 17.65%. It is worth emphasising that this is zero-shot classification with training only on synthetic data but testing on real-world data, and even using the same DGCNN network as the baseline. By anchoring the DGCNN backbone to physical, text-based geometric descriptions rather than closed-set classification labels, the network is far more robust to visual noise that confuses purely geometric hard classifiers. This approach outperforms the zero-shot performance of the PointNet/DGCNN baseline and establishes a new baseline result for zero-shot classification in this domain.

B. Few-Shot Fine-Tuning and Class Balancing (Scenario 2)

The second evaluation stage measures performance after applying a fine-tuning regime using the scarce physical training samples. To correct for dataset imbalance without over-fitting to the majority "Normal" class, we applied the inverse-frequency weighted loss function detailed in Section IV.

As described in Table IV, fine-tuning standard architectures produces discernible but small improvements. Baseline results from [4] shown in Table IV show that PointNet reaches an F1-score of 24.24%, while DGCNN produces a comparable F1-score of 23.58%, which is a more modest increase on the zero-shot performance.

In contrast, fine-tuning the 3D-PipeCLIP framework produces a precision of 58.88% and a recall of 57.32%, resulting in a final weighted F1-score of 57.81% (Table IV). This substantial improvement over the baseline is due to the use of the text prompt linguistic anchors, which effectively preserve the global shape understanding learned during synthetic training, enabling the few-shot physical scans to quickly learn the real-world characteristics of the point cloud sensor distortions, rather than forcing them to relearn the defect classes entirely from scratch. The detailed results of 3D-PipeCLIP for Scenario 2 are shown in Table V.

The normalised confusion matrix in Table VI(a) reveals more detail in the performance of the 3D-PipeCLIP zero-shot classification (Scenario 1). The model has successfully begun to isolate non-defective 'Normal' pipes (18.85%) from 'Brick' obstructions (90.79%) and 'Rubber Ring' delaminations (37.50%). However, the network still displays a substantial amount of confusion between classes, prior to fine-tuning.

Following the execution of the full fine-tuning pipeline (Scenario 2), the model greatly improves its alignment as observed in Table V and Table VI(b), increasing the F1-score of the "Normal" and "Displaced Joint" classes to 61.71% and 38.58%, respectively.

TABLE V: Fine-Tuning Classification Performance Breakdown by Class on the Physical Test Split (Scenario 2).

Class	Precision	Recall	F1-Score	Support
Normal	66.20%	57.79%	61.71%	244
Disp.	33.93%	44.71%	38.58%	85
Brick	83.95%	89.47%	86.62%	76
Ring	39.24%	38.75%	38.99%	80
Macro Avg	55.83%	57.68%	56.48%	485
Weighted Avg	**58.88%**	**57.32%**	**57.81%**	**485**

V. Discussion

A. Review of the Approach

The proposed approach of 3D-PipeCLIP has been demonstrated to successfully perform 3D sewer defect classification from point cloud data. The main advantage of the method is that it leverages pre-trained geometric text embeddings. This bypasses traditional hard-label classification, preventing the model from over-fitting to the mathematically perfect geometries found in synthetic data.

The main novelty of the approach is the use of geometric language descriptions, aligned to the actual point cloud data, inspired by the approach of PipeCLIP in the video domain [12]. The results demonstrate that the model trained on synthetic data only, effectively handles the time-of-flight (ToF) sensor noise and missing point density inherent to real-world inspections. The strength of the method is that, on the basis of this investigation, the approach reduces the need for very large labelled physical training sets. However, a potential weakness lies in the manual creation of the geometric prompt dictionary, which is relatively arbitrary - future work should investigate that aspect.

B. Comparison to Literature

The main advantage of the proposed framework lies in its lightweight training overhead and accessibility. In the original benchmark established by Haurum et al. [4], standard deep geometric extractors like PointNet and DGCNN produced low accuracy results, even when fine-tuning on real data. The F1-scores for the zero-shot performance were as low as 5.25% up to 17.65%. Our method outperformed these baseline metrics, achieving a robust zero-shot F1-score of 37.86%.

Other notable performers in this domain include TransPC-Net [8], which, although it achieved over 70% F1-score, was on a *mixed* dataset of synthetic and real data from [4], so it is not directly comparable to our results (we also score much higher on synthetic data, which is to be expected given the lack of noise, non-diversity of scans and the relatively large amount of data). We chose to focus our evaluation on real data, which we consider to be the main intended application of the method. An additional method proposed by Jing et al. [10] uses Sewer-ViT (a vision transformer network), that

TABLE VI: Normalized Confusion Matrices (%) for 3D-PipeCLIP in the Real Data Test Split. True labels are represented by rows, and predicted labels by columns.

(a) Scenario 1: Zero-Shot Baseline Trained on Synthetic Data and Tested on Real Data

True Label	Normal	Disp.	Brick	Ring
Normal	**18.85%**	37.70%	2.87%	40.57%
Disp.	20.00%	**52.94%**	2.35%	24.71%
Brick	3.95%	0.00%	**90.79%**	5.26%
Ring	28.75%	20.00%	13.75%	**37.50%**
Final Overall Accuracy: 39.18%				

(b) Scenario 4: Fine-Tuned on a Subset of Real Data and Tested on Independent Real Data

True Label	Normal	Disp.	Brick	Ring
Normal	**57.79%**	24.59%	2.05%	15.57%
Disp.	45.88%	**44.71%**	2.35%	7.06%
Brick	5.26%	0.00%	**89.47%**	5.26%
Ring	36.25%	17.50%	7.50%	**38.75%**
Final Overall Accuracy: 57.32%				

produced a high F1-score of 62.58% on the same dataset for the real test split. However, their model relied on a multi-stage training pipeline using cross-modal pre-training on the massive Sewer-ML database consisting of over 1.3 million real-world sewer images [11].

The approach in [10] is the state-of-the-art result and is appealing for its high accuracy. However, the drawback of that method is that it is dependent on a very large, highly curated dataset of pipe images (i.e. over 1 million). The fact that we can get near that score with our approach, 3D-PipeCLIP, with an F1-score of 57.81%, using a much smaller dataset consisting of mainly synthetic point cloud data (numbering in the thousands), with a smaller set of real data (numbering in the hundreds), highlights the extreme data efficiency of our approach.

C. Future Work

Future research will aim to overcome the current zero-shot limits by incorporating multimodal fusion, processing concurrent 2D CCTV video feeds alongside localized point clouds to establish a dense, synchronized feature vector. To better support on-board robotics edge computing, we plan to evaluate the framework's effectiveness when the heavy DGCNN backbone is substituted for a more lightweight point cloud architecture. Additionally, we aim to upgrade the framework from pure static classification to a full dynamic tracking system, integrating temporal smoothing and spatial neighbourhood clustering, for a robotic crawler moving through a pipe.

VI. CONCLUSION

This paper presented 3D-PipeCLIP, a lightweight geometric-language alignment framework specifically engineered to overcome the synthetic-to-real domain gap in automated sewer defect classification. The approach maps 3D point cloud geometries directly to a shared latent space with geometric language descriptions. We achieved a zero-shot F1-score of 37.86% on physical sensor data from only using synthetic data to train the system, an increase up from 17.65% in a previously published zero-shot baseline. This performance rose to an F1-score of 57.81% when fine-tuning on a small set of independent real-world data. In summary, our framework offers a data-efficient approach for 3D defect diagnosis in a domain that often suffers from a lack of high-quality, large, labelled data sets.

REFERENCES

[1] 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.

[2] 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.

[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 *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 1525–1531.

[4] J. B. Haurum, M. M. Allahham, M. S. Lynge, K. S. Henriksen, I. A. Nikolov, and T. B. Moeslund, "Sewer defect classification using synthetic point clouds." in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021)*, 2021, pp. 891–900.

[5] A. Karnezis, R. Worley, A. Blight, S. Anderson, K. Horoshenkov, and L. Mihaylova, "Feature detection and classification in buried pipes using LiDAR technology," in *Proceedings of the The 21st International Computing & Control in the Water Industry Conference, CCWI 2025*, 2025.

[6] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[7] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[8] Y. Zhou, A. Ji, and L. Zhang, "Sewer defect detection from 3d point clouds using a transformer-based deep learning model," *Automation in Construction*, vol. 136, 2022.

[9] 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, 2024.

[10] S. Jing, X. Li, D. A. Beyene, G. Cha, and S. Park, "Classification of sewer defects using point clouds based on a novel sewer vision transformer with cross-modal in-domain knowledge," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[11] 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.

[12] C. Zhao, C. Hu, Z. Cao, Y. Song, and J. Liu, "Pipeclip: Defect-conditioned and cross-focus-driven vision-language model for video-based sewer defect inspection," *IEEE Transactions on Automation Science and Engineering*, 2026.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PmLR, 2021, pp. 8748–8763.