This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SFR-Net: Steering-Fusion-Refining Network in Multi-label Zero-Shot Sewer Defect Detection

Zhao-Min Chen[1] Xinjian Huang[1] Yisu Ge[1,*] Yu Li[2]
[1]Wenzhou University, China
[2]Zhejiang College of Security Technology, China
chenzhaomin123@gmail.com, supercan.cs@gmail.com, ysg@wzu.edu.cn, 18096160@zjcst.edu.cn

Abstract

Due to the prohibitive cost of data annotation and the inability to obtain sufficient sample data for all defect categories, municipal sewer pipe defect detection poses significant generalization challenges for traditional models. Multi-Label Zero-Shot Learning (ML-ZSL) offers a viable solution to address this challenge. However, existing methods struggle to establish robust and fine-grained visual-semantic alignment between the complex visual environment inside the pipes and the often sparse semantic descriptions, leading to a critical issue: Alignment Ambiguity. To mitigate this, we propose a novel Steering-Fusion-Refining Network (SFR-Net) that follows a three-stage paradigm to progressively dissolve this ambiguity. This is achieved as the Representation Steering (RS) module first integrates a parameter-efficient feature steering mechanism to continuously adapt the representation to the pipe scene; the Multi-Granularity Evidence Fusion (MEF) module subsequently aggregates unambiguous multi-granularity visual evidence through decoupled global and local paths; and the Generalized Relational Score Refining (GR) module ultimately learns and transfers relational logic from seen defects to gain a universal score correction ability, directly refining preliminary prediction scores and significantly boosting the model's zero-shot generalization and prediction consistency. Extensive experiments on the public Sewer-ML dataset and our private WZ-Pipe dataset demonstrate that the proposed SFR-Net achieves state-of-the-art (SOTA) performance in multi-label zero-shot learning task [1].

1. Introduction

As highlighted in research on the sustainability of urban utilities, robust sewerage infrastructure projects are essential for sustainable development [2, 9, 30]. Due to long-term deterioration, sewer pipes inevitably develop various defects (e.g., deformation, blockage, leakage), making automated sewer pipe defect detection a quintessential multi-label image recognition task [10]. Despite its criticality, the specialized nature of the pipe environment incurs extremely high data collection and annotation costs. Furthermore, because certain rare defects typically manifest only after decades of operation, the extreme scarcity of early training samples for these anomalies significantly escalates the complexity of the task [28]. Consequently, directly training models on such data inevitably subjects them to severe long-tail distribution biases, or even renders them completely incapable of learning specific anomaly categories due to missing samples.

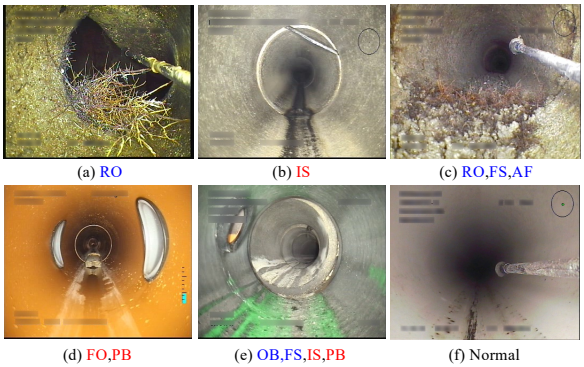


Figure 1. An illustrative example of multi-label zero-shot learning in sewer scene defect detection. Unseen classes are denoted in red font, while Seen classes are shown in blue font.

Multi-Label Zero-Shot Learning (ML-ZSL) [15, 25] provides a new paradigm for resolving this dilemma (refer to Fig. 1). By leveraging knowledge transfer, ML-ZSL equips the model to directly identify unseen defect categories, overcoming the non-exhaustive nature of defect enumeration. However, existing ML-ZSL methods face significant hurdles. Attribute-based methods [4, 5, 19] suf-

*Corresponding author.
[1]Code and dataset: https://github.com/infinitycan/SFR-Net

fer from the immense manual effort required for defining comprehensive attribute matrices, severely hindering their scalability and application to industrial fine-grained defect detection. Word-embedding-based methods, particularly those leveraging large Vision-Language Models (VLMs) [1, 16, 17, 22] are suboptimal due to the significant domain gap between sewer pipe data and pre-training data. Subsequent prompt-tuning approaches (e.g., CoOp [33], CoCoOp [34]) still fail, as their frozen image encoders lack the domain adaptability and feature extraction capability needed to bridge the general semantic space with the specialized, fine-grained visual characteristics of sewer pipes (e.g., dim lighting, subtle cracks). The critical challenge thus remains bridging this domain gap and overcoming visual-semantic alignment ambiguity to achieve robust ML-ZSL.

To address the severe challenge of visual-semantic alignment ambiguity in specialized domains, we propose the novel Steering-Fusion-Refining Network (SFR-Net), which systematically resolves this issue through a three-stage multi-modal paradigm. First, the Representation Steering (RS) module guides feature representations throughout the encoder, achieving parameter-efficient, domain-specific adaptation for the harsh pipe environment. Second, the Multi-Granularity Evidence Fusion (MEF) module enhances diagnostic reliability by decoupling the task into global context judgment and local key region localization, blending features from dual paths for fine-grained alignment. Finally, the Generalized Relational Score Refining (GR) module enforces logical consistency by learning a universal score refining function from seen category relationships via Graph Convolutional Networks (GCN), transferring this correction skill seamlessly to unseen defects for robust predictions. Furthermore, to support this domain, we construct and plan to open-source the WZ-Pipe dataset, the first dedicated dataset for multi-label zero-shot learning in sewer pipes.

Our main contributions can be summarized as follows:

- We propose a novel framework, SFR-Net, this framework systematically resolves the issue of visual-semantic alignment ambiguity through a three-stage paradigm. **To the best of our knowledge, this is the first successful application of multi-label zero-shot learning to the sewer pipe domain, demonstrating significant performance gains.**
- We introduce three innovative modules to systematically dissolve visual-semantic alignment ambiguity: RS for domain-adaptive feature steering, MEF for multi-granularity evidence fusion, and GR for transferring logical correction skills via GCN.
- We present the **WZ-Pipe** dataset, which provides a novel data source explicitly designed for multi-label zero-shot learning in industrial sewer pipe inspection, addressing key challenges in novel defect identification.
- Extensive results on the Sewer-ML and WZ-Pipe datasets validate the effectiveness of SFR-Net, showing it significantly outperforms existing multi-label zero-shot learning methods in addressing the domain's complex generalization challenges.

2. Related Work

2.1. Multi-Label Zero-Shot Learning (ML-ZSL)

ML-ZSL leverages auxiliary information (like word embeddings) to transfer knowledge from seen to unseen categories, while the more challenging Multi-Label Generalized Zero-Shot Learning (ML-GZSL) requires simultaneous identification of both, addressing bias towards seen categories. Existing methods fall primarily into attribute-based approaches [4, 5, 19], which suffer from limited scalability due to intense manual attribute definition, and word-embedding-based approaches [1, 16, 17, 22]. The emergence of CLIP [22] streamlined the latter by creating a powerful shared semantic space. However, subsequent prompt learning methods (e.g., CoOp [33], CoCoOp [34]) that rely on a frozen image encoder still suffer from insufficient domain adaptability and feature extraction capability when faced with specialized, fine-grained industrial scenarios like sewer pipes. This huge domain gap escalates the difficulty of ML-ZSL in this context. Furthermore, advanced methods like MKT [12] and RAM [29] also struggle, mainly because their reliance on the original CLIP image encoder leads to limited feature extraction effectiveness and hampers their ability to encode complex domain knowledge in visually intricate environments.

2.2. Defect Detection in Sewer Inspection

Sewer inspection is formulated as a Multi-Label Classification (MLC) problem, aiming to identify co-existing defects (e.g., cracks, breaks) in complex internal pipe environments. Traditional methods [8], [32], [27], [20] relied on hand-crafted features, but the field has rapidly shifted towards Deep Learning following the release of the Sewer-ML dataset [10]. Recent studies widely adopted CNN-based multi-label classification models. To capture complex defect co-occurrence, researchers began exploring Graph Neural Networks (GNNs). Methods like CT-GNN [11] and LMT-GCN [18] utilize GNNs to model contextual and semantic dependencies, significantly enhancing closed-set robustness.

Despite the excellent closed-set performance of these deep learning methods, their critical limitation is their reliance on large-scale, fully annotated datasets [24, 31]. Crucially, they fundamentally lack the mechanism to transfer knowledge to unseen classes, preventing them from addressing the dynamic emergence of new defect types in

[Role] You are an expert in sewer pipe defects detection scenario, highly proficient with domain standards.
[Task] Generate text inputs for a multi-label zero-shot defects detection task.
[Input] I will provide the list of defect codes from the {**Sewer-ML**} dataset.
[Reference] Base the definitions on the context and terminology found in the {**Fotomanualen: TV-inspektion af afløbsledninger**} standard.
[Required Outputs] For each defect, generate two distinct text representations:**[Detailed Description]**: A comprehensive, robust, and domain-specific sentence (MUST NOT exceed 77 tokens, including start/end tokens, for CLIP context length.)**[Explicit Phrase]**: A concise, unambiguous noun phrase (for GR module adjacency matrix building).
[Preference] Ensure all generated sentences are **concise, unambiguous, and highly domain-specific**.
[Format] Answer the request in a **markdown table** with columns for 'Defect Code', 'Detailed Description', and 'Explicit Phrase'.

Answer:

| **Defect Code** | **Detailed Description** | **Explicit Phrase** |

DE	A deformation of the pipe's cross-section where the circular shape is noticeably altered due to external load, commonly resulting in vertical or horizontal ovality.	Cross-sectional Pipe Deformation
AF	The accumulation of soft solid matter, grease, or sludge on the pipe's invert, partially or completely obstructing the normal flow of wastewater.	Sedimentation and Obstruction
IN	Water entering the pipe system from the surrounding soil through defective joints, cracks, or structural failures in the pipe wall.	Groundwater Infiltration
...

Figure 2. **Customized prompt for LLM-based description generation.**

practical industrial applications.

3. Methodology

3.1. Problem Formulation and CLIP Baseline

We denote the set of seen defect labels as Y^{S} and unseen defects as Y^{U}, with $Y^{\mathrm{S}} \cap Y^{\mathrm{U}} = \emptyset$. The available training dataset is $D_{\mathrm{tr}} = \{x_i, y_i\}_{i=1}^{M}$, where $y_i \in Y^{\mathrm{S}}$. Both the ML-ZSL and ML-GZSL tasks are evaluated on the same test set $D_{\mathrm{te}}^{\mathrm{S}\cup\mathrm{U}}$, which contains samples from the entire union set $Y^{\mathrm{S}} \cup Y^{\mathrm{U}}$. The standard ML-ZSL objective is to train a model on D_{tr} and evaluate its performance only on the unseen subset Y^{U} within $D_{\mathrm{te}}^{\mathrm{S}\cup\mathrm{U}}$, learning a classifier $X \to Y^{\mathrm{U}}$. The more challenging ML-GZSL objective requires simultaneous identification of defects across the entire union set $Y^{\mathrm{S}} \cup Y^{\mathrm{U}}$ on $D_{\mathrm{te}}^{\mathrm{S}\cup\mathrm{U}}$. This latter task demands robust generalization capability while effectively resolving the bias problem toward seen defects.

Our model is based on CLIP [22], a dual-encoder model consisting of a Visual Encoder $f(\cdot)$ and a Text Encoder $g(\cdot)$. CLIP projects image features $f(X)$ and text features $g(T)$ into a shared embedding space where the similarity score $S(X,T)$ is computed via cosine similarity:

$$S(X,T) = \cos(f(X), g(T)) = \frac{f(X) \cdot g(T)}{\|f(X)\|_2 \, \|g(T)\|_2} . \quad (1)$$

3.2. LLM-based Sewer Defects Description Generation

While most existing ML-ZSL methods use only simple template prompts (e.g., "A photo of a [Category]"), this leads to text features with poor semantic robustness, especially in specialized domains like sewer inspection that involve copious domain-specific terminology. To achieve accurate visual-semantic alignment, we leverage the rich knowledge of Large Language Models (LLMs) [6] to generate two types of customized text descriptions (Fig. 2). We generate detailed descriptions and explicit phrases for robust semantic feature generation using the CLIP Text Encoder, and for constructing the category relationship matrix in the GR module, respectively. This customized prompting strategy, which incorporates strict constraints and domain-specific information, effectively guides the LLM to produce highly specialized text descriptions tailored for the distinct requirements of different modules, thereby bridging the semantic gap and enhancing alignment accuracy.

3.3. Representation Steering Module (RS)

The core challenge in applying the CLIP model to zero-shot defect detection is bridging the domain gap without incurring catastrophic forgetting [3], a limitation observed with all-finetune strategy (see Table 5). Drawing inspiration from Parameter-Efficient Fine-Tuning (PEFT) methods [7], [13], [14], we propose the Representation Steering (RS) module to achieve domain-specific feature adjustment with minimal parameter overhead. As shown in Fig. 3, independent RS Blocks are inserted in parallel into each TransformerLayer of the Image/Text encoders. Formally, given an input feature X_{in}, the RS Block computes the domain-specific correction X_{rs}:

$$X_{\mathrm{rs}} = \mathrm{Linear}_2(\mathrm{Dropout}(\mathrm{GeLU}(\mathrm{Linear}_1(X_{\mathrm{in}})))) , \quad (2)$$

where Linear_1 and Linear_2 efficiently adjust the feature dimension via a bottleneck layer ($D \to d \to D$). This correction is added to the original Transformer layer output:

$$X_{\mathrm{out}} = \mathrm{TransformerLayer}(X_{\mathrm{in}}) + X_{\mathrm{rs}} . \quad (3)$$

The residual domain-feature learning mechanism superimposes continuous, progressive domain steering onto the model's intermediate representations. This approach not only effectively prevents catastrophic forgetting but also forges robust, specialized features tailored for the harsh pipe environment, thereby successfully bridging the visual domain gap.

3.4. Multi-Granularity Evidence Fusion Module (MEF)

The CLIP Image Encoder extracts a global feature vector (CLS token) for alignment. While this vector effectively captures overall image context, it limits the encoding of local details crucial for identifying subtle defects due to information aggregation. In complex industrial environments, robust detection of small, fine-grained defects remains a critical challenge. To fully leverage local information and enhance zero-shot defect detection, we propose the Multi-Granularity Evidence Fusion Module.

As shown in Fig. 3, the MEF module utilizes two independent, complementary paths: the global path provides

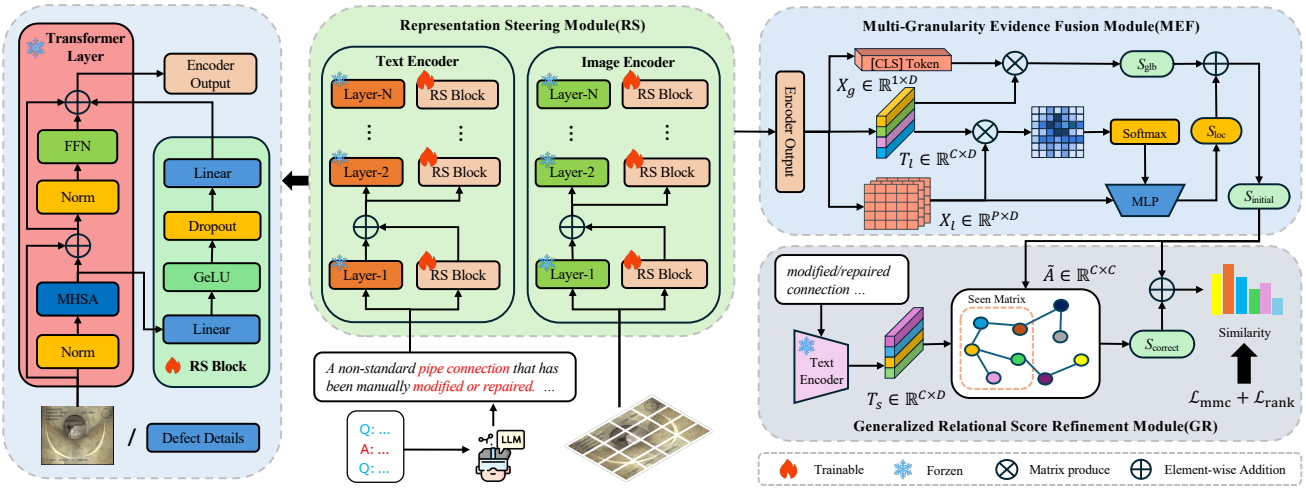


Figure 3. **Overview of the proposed SFR-Net framework.** RS efficiently fine-tunes multi-layer features from CLIP to bridge the domain gap. The MEF fuses global and local image evidence to generate an initial prediction. Moreover, GR learns a universal score correction skill from category semantic structures, directly refining the initial prediction score to ensure robust zero-shot generalization.

stable overall decision-making, while the local path enhances details to form a robust prediction mechanism. After passing through the RS module, we obtain the global image feature $X_g \in \mathbb{R}^{1\times D}$, local image features $X_l \in \mathbb{R}^{P\times D}$, and text features $T_l \in \mathbb{R}^{C\times D}$, where C is the number of categories to be predicted. All features are L2-normalized to ensure stable similarity computation. First, the global path computes the matching scores $S_{\text{glb}} \in \mathbb{R}^C$ via a simple dot product:

$$S_{\text{glb}} = cos(X_g,\, T_l^\top)\,, \tag{4}$$

where S_{glb} represents the overall matching score between the image and each defect category embedding.

Next, the local path computes the matching scores. We first calculate the element-wise matching affinity W_{affinity} between the local image features X_l and the text features T_l, performing a Softmax normalization across the local token dimension P to obtain a set of location weights for each defect category:

$$W_{\text{affinity}} = \text{Softmax}(cos(X_l,\, T_l^\top))\,. \tag{5}$$

Subsequently, we use the transpose of the affinity matrix to perform weighted aggregation of the local tokens X_l, effectively generating a category-specific local evidence feature for each defect. This feature is fed into an MLP to derive the local matching scores $S_{\text{loc}} \in \mathbb{R}^C$:

$$S_{\text{loc}} = \text{MLP}(W_{\text{affinity}}^\top \cdot X_l)\,. \tag{6}$$

Finally, we perform an element-wise summation of the global and local matching scores to obtain the initial prediction scores S_{initial}:

$$S_{\text{initial}} = S_{\text{glb}} + S_{\text{loc}}\,. \tag{7}$$

This fusion of multi-granularity evidence effectively mitigates information loss from global aggregation, fully utilizing the image's fine-grained information to achieve accurate multi-label zero-shot detection of varied sewer pipe defects.

3.5. Generalized Relational Score Refining Module (GR)

The Generalized Relational Score Refining (GR) module tackles the ML-ZSL challenge by learning a transferable skill from seen defects. Our inductive hypothesis is that the logical pattern of prediction score interactions is universal across categories. This enables the GR module to acquire relational logic from seen relationships and transfer it to unseen categories during inference.

We first construct a semantic similarity matrix. Explicit phrases generated by LLM are encoded by CLIP's original Text Encoder to obtain the semantic vectors $T_s \in \mathbb{R}^{C\times D}$. Given CLIP's superior generalization ability, these high-quality vectors form a robust semantic foundation that is essential for the effective operation of the GR module. After L2-norm normalization, we calculate the cosine similarity matrix M between these vectors:

$$M = cos(T_s,\, T_s^\top)\,. \tag{8}$$

Next, we binarize the similarity matrix by setting a threshold γ to construct the category adjacency matrix A. Specifically, A_{ij} is set to 1 if $M_{ij} > \gamma$ and $i \neq j$, indicating a semantic relationship between defect category i and category j:

$$A_{ij} = \begin{cases} 1 & \text{if } M_{ij} > \gamma \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}\,. \tag{9}$$

To use the adjacency matrix A in the GCN, we perform normalization, which includes adding self-loops and symmetric normalization:

$$\tilde{A} = \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} , \tag{10}$$

where $\hat{A} = A + I$ (I is the identity matrix), and $\hat{D}$ is a diagonal matrix whose diagonal elements are $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$.

This matrix is used for information propagation within the GCN, the GCN uses the initial prediction scores $S_{\text{initial}} \in \mathbb{R}^{C\times 1}$ as node feature representations. It leverages the graph structure to learn a rectification score $S_{\text{correct}} \in \mathbb{R}^{C\times 1}$. The message passing process is represented as:

$$S_{\text{correct}} = \tilde{A} \left[\text{ReLU} \left(\tilde{A} S_{\text{initial}} W^{(0)} \right) \right] W^{(1)} , \tag{11}$$

where $W^{(0)} \in \mathbb{R}^{1\times 64}$ and $W^{(1)} \in \mathbb{R}^{64\times 1}$ are learnable weight matrices.

Finally, we sum the rectification scores S_{correct} obtained from GR module with the initial prediction scores S_{initial} to get the final, more logically consistent prediction scores S_{final}:

$$S_{\text{final}} = S_{\text{initial}} + S_{\text{correct}} . \tag{12}$$

3.6. Loss Function

To train the proposed SFR-Net model, we designed a composite loss function, the Synergistic Contrastive Loss $\mathcal{L}_{\text{scl}}$, which combines contrastive learning and ranking constraints.

We first employ the Multi-Matching Contrastive (MMC) Loss $\mathcal{L}_{\text{mmc}}$ to constrain prediction scores against ground-truth labels. This loss maximizes the similarity between the image and its positive category labels while minimizing similarity with all other negative labels within the batch (batch-wise contrastive constraint). Following the design of RAM [29], assuming the final prediction score matrix $S_{\text{final}} \in \mathbb{R}^{B\times C}$ and labels $Y \in \mathbb{R}^{B\times C}$, $\mathcal{L}_{\text{mmc}}$ is defined as:

$$\mathcal{L}_{\text{mmc}} = -\frac{1}{|\mathcal{P}_B|} \sum_{(b,i)\in\mathcal{P}_B} \log \frac{\exp(S_{\text{final}}^{(b,i)}/\tau)}{\sum_{b'=1}^{B}\sum_{j=1}^{C} \exp(S_{\text{final}}^{(b',j)}/\tau)} , \tag{13}$$

where $\mathcal{P}_B = \{(b,i) | Y_i^{(b)} = 1\}$ is the index set of all positive pairs, $S_{\text{final}}(b,i)$ is the score for the i-th category of the b-th sample, and the denominator sums over all scores in the entire batch, serving as the comprehensive negative set. The temperature τ is set to 0.1.

Given the inherent complexity of the sewer pipe scenario, we introduce a Rank Loss $\mathcal{L}_{\text{rank}}$ to enhance the model's ranking capability. This loss ensures the score of a positive label $S_{\text{positive}}^{(i,j)}$ exceeds the score of the hardest negative label $S_{\text{negative}}^{(i)}$ by a predefined margin m:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} max(0, m + S_{\text{negative}}^{(i)} - S_{\text{positive}}^{(i,j)}) , \tag{14}$$

where $|\mathcal{P}|$ is the total number of positive instances, and the margin m is set to 0.2.

The final loss function is the weighted sum:

$$\mathcal{L}_{\text{scl}} = \mathcal{L}_{\text{mmc}} + \lambda \mathcal{L}_{\text{rank}} , \tag{15}$$

where the hyperparameter λ is set to 10 to balance the two loss terms. This composite design combines global contrastive constraints and localized ranking benefits, effectively improving the learning outcome in complex ML-ZSL tasks.

4. Experiments

4.1. Datasets

Sewer-ML Dataset: For fair comparison, we utilize the large-scale public multi-label sewer defect dataset, Sewer-ML [10] (1.3M images spanning 17 categories). As the test set labels are unavailable, we adopt the validation set as the test set. For the ML-ZSL task, we select the five least-frequent categories as the unseen defects, ensuring no corresponding samples exist in the training set.

WZ-Pipe Dataset: Additionally, we introduce a private sewer pipe dataset, named WZ-Pipe. WZ-Pipe totals 63,978 image samples, covering 17 defect categories. We provide detailed introduction for the WZ-Pipe dataset in the **Supplementary Material**. Consistent with the Sewer-ML dataset, we designate the five least-frequent defect categories as the unseen defects for evaluating performance in the ML-ZSL task.

4.2. Evaluation Metrics

For ML-ZSL and ML-GZSL tasks, we adopt the ranking protocol [15]. Performance is evaluated using Mean Average Precision (mAP) and Top-K Ranking Metrics Precison@K, Recall@K, F1-Score@K (P@K, R@K, F1@K) to assess ranking quality in limited label recommendation scenarios. Specifically, we report Top-1/3 metrics for the large-scale Sewer-ML dataset and Top-3/5 metrics for the smaller WZ-Pipe dataset to ensure robust evaluation in ML-GZSL task.

4.3. Implementation Details

We utilize CLIP pre-trained ViT-B/16 as the backbone. Both image and text encoders are frozen during training. Gemini-2.5 Flash is employed as the LLM to generate defect descriptions and explicit phrases. Training utilizes the AdamW optimizer (learning rate: 1×10^{-3}, weight decay: 1×10^{-4}). Input images are resized to 224×224 (yielding $14 \times 14 = 196$ patches post-projection). The RS module's hidden dimension d is set to 128. Key hyperparameters include the similarity threshold $\gamma = 0.85$ for the adjacency matrix A, rank loss weight $\lambda = 10$, and rank margin $m = 0.2$. We train the ML-(G)ZSL task for 8 epochs with

Table 1. **Performance comparison of ML-(G)ZSL methods on Sewer-ML dataset.** The **best** result is shown in **bold**, and the second best result is underlined.

Method	Zero-shot Learning (ML-ZSL)							Generalized Zero-shot Learning (ML-GZSL)						
	P@1	R@1	F1@1	P@3	R@3	F1@3	mAP	P@1	R@1	F1@1	P@3	R@3	F1@3	mAP
CLIP [22]	4.72	23.78	7.88	4.00	60.46	7.51	4.94	6.62	28.45	10.74	6.28	53.99	11.25	9.92
BiAM [21]	3.04	15.12	5.01	3.99	60.17	7.48	5.80	22.46	19.30	20.76	19.97	51.49	28.78	17.96
DualCoOp [26]	4.11	20.69	6.86	3.98	60.09	7.47	6.58	18.22	15.66	16.84	20.39	52.57	29.38	30.67
ML-Decoder [23]	4.93	24.80	8.22	4.47	67.52	8.39	6.24	38.48	33.07	35.57	22.86	58.94	32.94	38.57
MKT [12]	4.06	20.45	6.78	3.86	58.21	7.23	4.86	39.30	33.78	36.33	23.70	61.11	34.16	23.14
RAM [29]	4.12	20.75	6.88	3.99	60.31	7.49	4.01	37.21	31.98	34.40	25.21	65.01	36.33	23.88
Ours	**8.14**	**40.99**	**13.59**	**5.26**	**79.50**	**9.88**	**12.58**	**40.65**	**34.94**	**37.58**	**26.68**	**68.80**	**38.45**	**43.28**

Table 2. **Performance comparison of ML-(G)ZSL methods on WZ-Pipe dataset.**

Method	Zero-shot Learning (ML-ZSL)							Generalized Zero-shot Learning (ML-GZSL)						
	P@1	R@1	F1@1	P@3	R@3	F1@3	mAP	P@3	R@3	F1@3	P@5	R@5	F1@5	mAP
CLIP [22]	1.44	11.45	2.57	1.89	44.85	3.62	2.58	4.45	13.35	6.67	5.75	28.79	9.59	5.83
BiAM [21]	3.69	29.24	6.56	2.98	70.71	5.72	2.47	21.38	64.13	32.07	15.52	77.58	25.86	13.44
DualCoOp [26]	4.33	34.24	7.68	3.11	73.87	5.97	3.48	23.09	69.26	34.63	15.76	78.81	26.27	15.91
ML-Decoder [23]	2.44	19.29	4.33	2.35	55.70	4.50	4.07	23.06	69.19	34.59	15.50	77.49	25.83	22.35
MKT [12]	2.10	16.62	3.73	2.13	50.53	4.09	3.33	19.54	58.61	29.30	13.57	67.87	22.62	26.99
RAM [29]	4.09	32.40	7.27	3.06	72.70	5.87	3.04	23.55	70.66	35.33	15.95	79.76	26.58	20.56
Ours	**5.17**	**40.91**	**9.18**	**3.17**	**75.37**	**6.09**	**9.88**	**27.20**	**81.61**	**40.80**	**17.69**	**88.48**	**29.49**	**37.42**

Table 3. **Ablation(%) on the proposed modules.**

RS	MEF	GR	$\mathcal{L}_{rank}$	Task	mAP	F1@1	F1@3
×	×	×	×	ZSL	4.94	7.88	7.51
				GZSL	9.92	10.74	11.25
✓	×	×	×	ZSL	8.36	9.94	8.42
				GZSL	39.62	35.98	36.96
✓	✓	×	×	ZSL	9.81	10.61	8.75
				GZSL	40.81	36.25	37.87
✓	✓	✓	×	ZSL	12.25	11.91	8.56
				GZSL	41.54	36.29	38.37
✓	✓	✓	✓	ZSL	**12.58**	**13.59**	**9.88**
				GZSL	**43.28**	**37.58**	**38.45**

Table 4. **Ablation(%) on the semantic description.** We compared three different text generation strategies: Base Template (the fixed template "A photo of a [Category]"), Manual Description (brief defect category descriptions provided by human annotators), and LLM-Enhanced Description (professional and rich defect text descriptions generated via an LLM).

Strategy	Task	mAP	F1@1	F1@3
Base Template	ZSL	7.13	9.33	9.06
	GZSL	41.10	37.07	37.44
Manual Description	ZSL	6.90	9.10	8.25
	GZSL	40.76	36.76	36.62
LLM-Enhanced Description	ZSL	**12.58**	**13.59**	**9.88**
	GZSL	**43.28**	**37.58**	**38.45**

a batch size of 896. All experiments were conducted on NVIDIA RTX 3090 GPUs (24GB) and implemented using PyTorch framework.

4.4. Comparison with State-of-the-Art Methods

To demonstrate the efficacy of SFR-Net, we benchmarked it against SOTA methods from general domains, all reproduced under an identical environment (Tables 1 and 2). SFR-Net achieves comprehensive SOTA superiority across both the ML-ZSL and ML-GZSL tasks on the challenging Sewer-ML and WZ-Pipe datasets. On Sewer-ML, SFR-Net's mAP reaches 12.58% in the ML-ZSL setting, surpassing the robust DualCoOp (6.58%) by 6%, while its F1 metrics further demonstrate superior ranking capability compared to global alignment methods. The advantage is even more pronounced in ML-GZSL, where our mAP (43.28%) leads ML-Decoder by 4.71%. Furthermore, on the highly challenging WZ-Pipe dataset, SFR-Net demonstrates a significant performance gain in the ML-ZSL task, achieving an mAP of 9.88%, more than 2.4 times higher than the second best method, ML-Decoder (4.07%). This strongly validates the model's exceptional generalization capability for recognizing unseen defect categories, crucial due to the dataset's long-tail nature. In the ML-GZSL task, our method simi-

Table 5. **Ablation (%) on encoder fine-tuning strategies.** We compare Train-Vision (image encoder finetuning), Partial-Finetune (last three layers of both encoders), and All-Finetune (all layers of both encoders), conducted solely with the RS module.

Strategy	Task	mAP	F1@1	F1@3
Train-Vision	ZSL	6.35	6.61	8.41
	GZSL	**49.27**	**38.47**	**38.73**
Partial-Finetune	ZSL	**10.15**	6.92	7.06
	GZSL	35.05	34.45	36.17
All-Finetune	ZSL	8.36	**9.94**	**8.42**
	GZSL	39.62	35.98	36.96

Table 6. **Ablation(%) on the adjacency matrix**. We compared three different types of adjacency matrices: Identity Matrix (containing only self-connections), Random Matrix (randomly initialized), and Similarity Matrix (constructed based on semantic information, which is our proposed method).

Type	Task	mAP	F1@1	F1@3
Identity Matrix	ZSL	7.74	6.03	6.27
	GZSL	40.82	37.09	35.93
Random Matrix	ZSL	8.28	10.53	9.36
	GZSL	41.23	37.09	37.75
Similarity Matrix	ZSL	**12.58**	**13.59**	**9.88**
	GZSL	**43.28**	**37.58**	**38.45**

larly leads with an mAP of 37.42%, nearly 38% better than MKT. These robust results fully validate the proposed SFR-Net architecture's ability to effectively address the complex semantic and generalization challenges inherent in the sewer pipe domain, successfully mitigating the common bias problem by balancing the recognition of both seen and unseen categories.

4.5. Ablation Study

To comprehensively validate the effectiveness of our proposed method, we conducted a series of ablation studies. All ablation experiments were conducted based on the public Sewer-ML dataset and the ViT-B/16 backbone network, unless specifically noted otherwise in a particular analysis. Additional results, including supervised multi-label classification results and hyperparameter analysis, are provided in the **Supplementary Material** to offer a comprehensive overview of our methodology.

Effects of proposed modules. Table 3 validates the necessity and synergistic effectiveness of our components. The initial poor adaptability of the CLIP baseline is addressed by the Representation Steering (RS) module, which significantly boosts ML-GZSL mAP from 9.92% to 39.62% via parameter-efficient guidance. The Multi-Granularity Evidence Fusion (MEF) module further enhances fine-grained perception, yielding an additional 1.45% gain in ML-ZSL mAP. Crucially, the Generalized Relational Score Refining (GR) module applies our transferable correction hypothesis to dynamically rectify scores, pushing ML-ZSL mAP to 12.25%. Finally, the Rank Loss ($\mathcal{L}_{rank}$) strengthens the embedding space's ranking capability, resulting in the final performance leap.

Table 7. **The numbers of learnable parameters (L-Params) and computational overhead (FLOPs) comparison.** Note that the ML-Decoder [23] method inherently excludes the text encoder, which accounts for its substantially lower FLOPs. The mAP column reports results for ML-ZSL and ML-GZSL tasks, displayed as $\mathbf{mAP}_{\text{ML-ZSL}}/\mathbf{mAP}_{\text{ML-GZSL}}$.

Method	Backbone	L-Params	FLOPs	mAP
CLIP [22]	ViT-B/16	-	106.74G	4.94/9.92
BiAM [21]	VGG-19	119.55M	239.30G	5.80/17.96
DualCoOp [26]	ResNet-101	38.20M	221.08G	6.58/30.67
ML-Decoder [23]	ViT-B/16	94.47M	**36.06G**	6.24/38.57
MKT [12]	ViT-B/16	112.15M	136.56G	4.86/23.14
RAM [29]	ViT-B/16	13.02M	114.70G	4.01/23.88
Ours	ViT-B/16	**6.05M**	110.62G	**12.58/43.28**

Comparison of Semantic Descriptions. We evaluated three text input modalities (Table 4). Manual descriptions (concise phrases) performed worst due to insufficient fine-grained semantics. The base template (fixed template with defect name) surprisingly outperformed manual descriptions, validating CLIP's inherited structural alignment. Crucially, our LLM-enhanced description strategy achieved the best performance across all tasks. This proves that generating detailed, domain-specific category descriptions via an LLM is highly necessary, as it provides the highest semantic density for significantly enhancing generalization and prediction robustness.

Comparison of Fine-Tuning Strategy. We verified the efficacy of the fine-tuning strategy using only the RS module (Table 5). The train-vision strategy suffered from catastrophic forgetting, leading to the worst ML-ZSL performance. The partial-finetune strategy preserved generalization well (highest ML-ZSL mAP), but struggled with F1 and ML-GZSL metrics due to insufficient robust feature learning. Therefore, we adopted the all-finetune strategy (RS Block on every layer). This approach successfully enhances predictive performance while mitigating catastrophic forgetting, achieving the highest F1@1 and F1@3 in ML-ZSL and maintaining excellent ML-GZSL performance, proving it is the most balanced and practical scheme.

Comparison of Adjacency Matrix. We investigated the necessity of the graph structure within the GR Module (Ta-

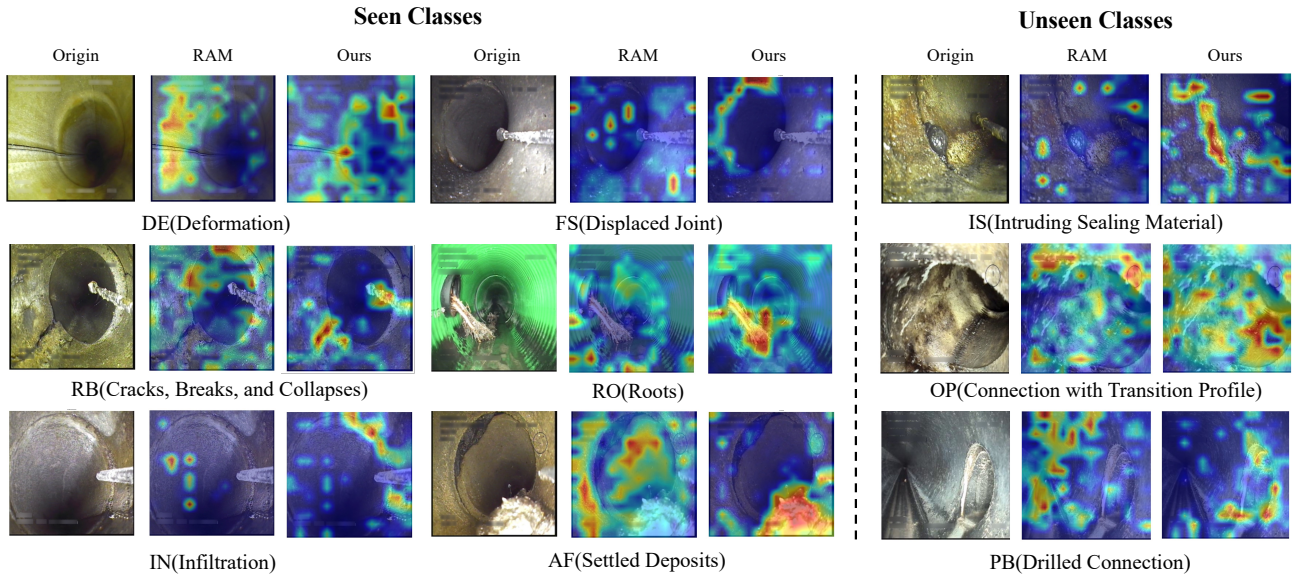


Figure 4. **Grad-CAM visualization.** We performed activation visualization on both seen defects and unseen defects in the Sewer-ML dataset to analyze the model's attention distribution across different scales of defects. Simultaneously, we compared the activation maps of our proposed method with those generated by the RAM [29] method, which uses a similar fine-tuning strategy.

ble 6). Replacing the adjacency matrix (A) with an identity matrix ($A = I$) degrades the GCN into an MLP, causing a severe performance drop with ML-ZSL mAP falls from 12.58% to 7.74%. A random matrix surprisingly outperformed the identity matrix, suggesting regularization benefits from random exchange. Nevertheless, our semantic similarity-based adjacency matrix maintains a significant performance gap over both alternatives. This proves that the explicit modeling and high quality of genuine category relationships are critically important to unlock the full zero-shot generalization potential of the GR Module.

Comparison of Method Performance and Complexity. Analysis in Table 7 highlights the critical trade-off between accuracy and computational overhead. While the CLIP baseline has high base FLOPs, and BiAM/DualCoOp/ML-Decoder/MKT introduce substantial overhead, our method achieves the optimal balance. SFR-Net introduces only 6.05M of learnable parameters, comparable to RAM's overhead, yet achieves superior downstream performance. This indicates that SFR-Net's efficient multi-modal design significantly enhances model effectiveness while remaining lightweight.

4.6. Qualitative Assessment

We compared Grad-CAM defect activation visualization with RAM (a strong ML-ZSL benchmark) on six seen and three unseen defects (Fig. 4). SFR-Net showed significantly superior feature localization. It accurately and completely focuses on the defect region, even for critical or varied-scale defects (e.g., RB long crack, large RO/AF, small IN/IS). In stark contrast, RAM's attention was scattered and unfocused, performing poorly on regional defects (RO, AF, IS). This discrepancy indicates that RAM suffers from ineffective visual-semantic alignment, particularly when applied to the specialized sewer pipe domain. The visualization powerfully demonstrates that SFR-Net's efficient multi-modal interaction captures robust and discriminative visual features across different defect scales, which is vital for complex sewer pipe defect detection.

5. Conclusion

The limitations of existing methods in municipal multi-label zero-shot sewer defect detection, specifically concerning prohibitive annotation costs and insufficient generalization, are systematically addressed by adopting the ML-ZSL paradigm and introducing the novel Steering-Fusion-Refining Network. The network's core innovation lies in its highly synergistic architecture, which establishes robust visual-semantic alignment in harsh pipe environments. Extensive experiments on the Sewer-ML and WZ-Pipe datasets confirmed that the proposed SFR-Net achieves SOTA performance in multi-label zero-shot learning task.

Our future work will focus on two major directions: enhancing model robustness by mitigating spurious attention on background features, including privacy-obscured regions and inspection equipment; and addressing more challenging fine-grained defect grading tasks to provide practical severity assessments, moving beyond simple classification.

Acknowledgements
This research was supported by the Zhejiang Provincial Natural Science Foundation of China (Grant No.LMS26F020041, LQ24F020016), in part by the National Natural Science Foundation of China (Grant No. 62401398), and in part by the Wenzhou Major Scientific and Technological Innovation Project (Grant No.ZF2024001). We would like to thank Wenzhou Water Group Co., Ltd. and Zhejiang Xinyu Technology Group Co., Ltd. for their assistance in data collection.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Ali Alnoaimi and Anisur Rahman. Sustainability assessment of sewerage infrastructure projects: A conceptual framework. *International Journal of Environmental Science and Development*, pages 23–29, 2019.

[3] Reza Akbarian Bafghi, Nidhin Harilal, Claire Monteleoni, and Maziar Raissi. Parameter efficient fine-tuning of self-supervised vits without catastrophic forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3679–3684, 2024.

[4] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 330–338, 2022.

[5] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z Pan, and Huajun Chen. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 405–413, 2023.

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024.

[8] Mahmoud R Halfawy and Jantira Hengmeechai. Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine. *Automation in Construction*, 38:1–13, 2014.

[9] Sabine Hartmann, Raquel Valles, Annette Schmitt, Thamer Al-Zuriqat, Kosmas Dragos, Peter Gölzhäuser, Jan Thomas Jung, Georg Villinger, Diana Varela Rojas, Matthias Bergmann, et al. Digital-twin-based management of sewer systems: Research strategy for the kasytwin project. *Water*, 17(3):299, 2025.

[10] Joakim Bruslund Haurum and Thomas 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*, pages 13456–13467, 2021.

[11] Joakim Bruslund Haurum, Meysam Madadi, Sergio Escalera, and Thomas B Moeslund. Multi-task classification of sewer pipe defects and properties using a cross-task graph neural network decoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2806–2817, 2022.

[12] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xiujun Shu, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 808–816, 2023.

[13] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022.

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[15] Seung-Hwan Lee, Sungeun Kim, and Yong Man Ro. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1576–1585, 2018.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[18] Xinxin Li, Rui You, Mingxin Yu, Ting Sun, Wenshuai Lu, and Shijie Yu. A multi-label sewer defects classification methodology using attention-based dual branch network. 2024.

[19] Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, and Yao Zhao. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15337–15346, 2023.

[20] Joshua Myrans, Richard Everson, and Zoran Kapelan. Automated detection of faults in sewers using cctv image sequences. *Automation in Construction*, 95:64–71, 2018.

[21] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8731–8740, 2021.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[23] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 32–41, 2023.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[25] Ubai Sandouk and Ke Chen. Multi-label zero-shot learning via concept embedding. *arXiv preprint arXiv:1606.00282*, 2016.

[26] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022.

[27] Wided Ben Tagherouit, Saad Bennis, and Jamal Bengassem. A fuzzy expert system for prioritizing rehabilitation of sewer networks. *Computer-Aided Civil and Infrastructure Engineering*, 26(2):146–152, 2011.

[28] Saeed Talebi, Song Wu, Arijit Sen, Nazanin Zakizadeh, Quanbin Sun, and Joseph Lai. Infrastructure automated defect detection with machine learning: a systematic review. *International Journal of Construction Management*, pages 1–12, 2025.

[29] Hao Tan, Zichang Tan, Jun Li, Ajian Liu, Jun Wan, and Zhen Lei. Recover and match: Open-vocabulary multi-label recognition through knowledge-constrained optimal transport. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4650–4660, 2025.

[30] Franz Tscheikner-Gratl, Nicolas Caradot, Frédéric Cherqui, Joao P Leitão, Mehdi Ahmadi, Jeroen G Langeveld, Yves Le Gat, Lisa Scholten, Bardia Roghani, Juan Pablo Rodríguez, et al. Sewer asset management–state of the art and research needs. *Urban Water Journal*, 16(9):662–675, 2019.

[31] Stefanos Tsimenidis. Limitations of deep neural networks: a discussion of g. marcus' critical appraisal of deep learning. *arXiv preprint arXiv:2012.15754*, 2020.

[32] Xiangyang Ye, Jian'e Zuo, Ruohan Li, Yajiao Wang, Lili Gan, Zhonghan Yu, and Xiaoqing Hu. Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern chinese city. *Frontiers of Environmental Science & Engineering*, 13(2):17, 2019.

[33] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.