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

Supplementary Material

Table 1. **Sewer defects classes of our WZ-Pipe dataset.**

Type	Code	Description
Structural Defects	AJ	Lateral Branch Connection
	BX	Pipe Deformation by Compression
	CK	Joint Misalignment (Lateral)
	CR	Foreign Object Penetrating Pipe
	FS	Pipe Wall Corrosion/Spalling
	PL	Pipe Fracture / Crack
	QF	Joint Misalignment (Longitudinal)
	SL	Infiltration
	TJ	Joint Disconnection
Functional Defects	BJ	Deformed Pipe (Ovality)
	JG	Pipe Wall Deposits
	ZW	Internal Obstruction
	CQ	Test Sealant Residue
	FZ	Surface Debris/FOG Accumulation
	TL	Joint Material Ingress
	SG	Root Intrusion
	CJ	Siltation / Sediment Deposits

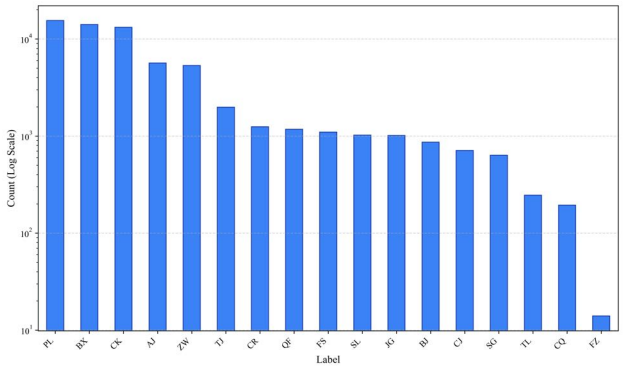

Figure 1. **Category distribution of WZ-Pipe dataset.**

1. WZ-Pipe Dataset: Statistics and Characteristics.

The data acquisition process for this dataset is stringent, with image samples collected via CCTV (Closed-Circuit Television) and subjected to professional video processing and classification annotation. Tab. 1 displays the names and brief explanations of the defect categories we defined, which span two major groups: Structural Defects and Functional Defects. Structural defects include critical issues such as Pipe Fracture / Crack (PL), Pipe Deformation (BX), and various Joint Misalignments (CK, QF, TJ). Functional defects cover common operational issues like Siltation / Sediment Deposits (CJ), Internal Obstruction (ZW), and Root Intrusion (SG). Notably, Fig. 1 shows the category distribution of the dataset, which reveals that WZ-Pipe suffers from a severe long-tail distribution problem. This high diversity and class imbalance reflect the reality of urban pipeline inspection, where defect occurrences are naturally long-tailed. Addressing this imbalance makes the core challenge of our ML-ZSL task to summarize transferable patterns from high-frequency (seen) defects to reliably predict low-frequency (unseen) new defect categories. This ultimately poses a significant challenge to model training and generalization.

2. Multi-Label Classification on Sewer-ML Dataset.

For the fully supervised MLC task (closed-set performance), we report mAP and conventional global metrics: Precision, Recall, and F1-Score [2], ensuring a comprehensive evaluation of both ranking and classification accuracy. As shown in Table 2, we present comparative results for the fully supervised multi-label classification task, including general-purpose SOTA methods [4–6] benchmarked against our model. We note that methods requiring full fine-tuning of the CLIP image encoder (denoted as CLIP*) demonstrate strong performance, with CLIP* achieving a notable 6.49% improvement in $F2_{CIW}$ over AutoSewerNet. Furthermore, Query2Label, by employing a query mechanism, more effectively utilizes image features for label prediction, thus outperforming CLIP* in both mAP and Recall. Conversely, TagCLIP, despite improving the label encoding mechanism, is penalized by its failure to address the domain adaptability of the frozen image encoder, resulting in inferior performance. Our proposed SFR-Net model, however, achieves SOTA performance (mAP 71.03%) while utilizing the Asymmetric Loss (ASL) [7] and only a small number of fine-tuned parameters. Although adopting full fine-tuning (Ours*) yields a marginal further increase in performance to mAP 71.55%, this minimal gain fully validates the superiority of the RS module's feature representation steering in effectively balancing high performance with computational efficiency.

3. Hyperparameter Analysis.

Effects of loss weight. To investigate the impact of the rank loss $\mathcal{L}_{rank}$ on overall model performance, we con-

Table 2. **Fully supervised multi-label classification results on Sewer-ML dataset.** We full-finetune the image encoder of CLIP, called **CLIP***. Ours* denotes the fully fine-tuned variant of our complete SFR-Net model.

Domain	Method	mAP	$F2_{CIW}$	$F1_{Normal}$	Precison	Recall	F1-Score
Sewer	SqueezeNet [1]	62.06	48.67	91.06	26.38	47.60	33.94
Sewer	NDCNN [9]	66.40	48.57	91.08	46.31	82.52	59.33
Sewer	CTGNN-STL [3]	63.09	52.45	91.48	**84.70**	76.63	80.46
Sewer	CTGNN-MTL [3]	64.12	59.55	91.81	83.78	78.38	80.99
Sewer	AutoSewerNet [8]	-	60.07	-	75.32	54.18	62.51
General	TagCLIP [4]	5.85	5.90	69.12	12.59	52.06	20.27
General	CLIP*	64.81	66.56	91.35	76.09	83.45	79.60
General	Query2Label [5]	68.53	66.91	90.99	71.96	85.00	77.94
Sewer	Ours	71.03	70.40	92.82	77.16	85.98	81.34
Sewer	Ours*	**71.55**	**71.12**	**93.17**	77.68	**86.47**	**81.84**

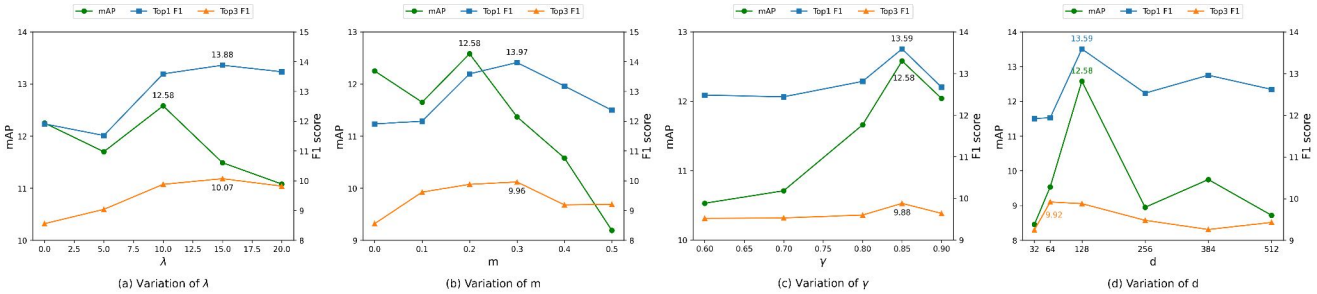

Figure 2. **Sensitivity analysis of hyperparameters in multi-label zero-shot learning task.**

ducted experiments with different settings for λ. Fig. 2(a) illustrates the effect of the hyper-parameter λ on model performance. As λ increases from 0, all metrics—including mAP, Top1-F1, and Top3-F1 exhibit an initial increase followed by a decrease. Specifically, the mAP peaks at 12.58% when λ=10, after which it quickly declines. Although Top1-F1 and Top3-F1 reach their maximum value of 13.88% at λ=15, the model's performance significantly degrades when λ is either too high or too low. This indicates that an excessively large λ causes the model to over-focus on hard negative samples, while a value that is too small fails to adequately leverage their value, with both extremes leading to suboptimal solutions. We ultimately select λ=10 as the weight for the loss function, as this setting yields relatively superior performance across both the mAP and F1 metrics.

Effects of loss margin. To study the effect of the rank margin on overall performance, we conducted experiments with different settings for m. Fig. 2(b) demonstrates the impact of the hyper-parameter m on model performance. As the hyper-parameter m increases from 0 to 0.2, all metrics—including mAP, Top1-F1, and Top3-F1 show an upward trend. The mAP reaches its peak of 12.58% at m=0.2. Once m exceeds 0.2, the mAP performance begins to rapidly decline, falling to around 9.2% at m=0.5. Top1-F1 peaks at 13.97% when m=0.3 but decreases thereafter. Top3-F1 peaks at 9.96% when m=0.2 and then slightly drops and stabilizes. We finally select m=0.2 as the fixed parameter, as this m setting results in robust performance across both mAP and F1 metrics.

Effects of adjacency matrix threshold. To investigate the influence of the category adjacency matrix on the overall prediction scoring capability, we experimented with different settings for γ. Fig. 2(c) shows the impact of the hyper-parameter γ on model performance. As the hyper-parameter γ continuously increases, all metrics—including mAP, Top1-F1, and Top3-F1 show an upward trend. All three metrics peak at γ=0.85. This suggests that this threshold effectively balances the sparsity and density of the category relationship graph, connecting categories with strong relevance while filtering out irrelevant or noisy connections. This enables the GCN to aggregate the most valuable structural information, significantly boosting the model's overall performance. When γ is too small, the threshold is too lenient, leading to an overly dense adjacency matrix where the GCN aggregates excessive noise, thereby damaging the model's discriminative power. Conversely, when γ is too large, the threshold is too strict, making the adjacency matrix overly sparse, causing the GCN to underutilize inter-category relationships, which also leads to a performance drop. Thus, γ=0.85 is the optimal threshold setting for the

category adjacency matrix in this task.

Effects of RS block bottleneck dimension. To explore the influence of the bottleneck layer dimension in the RS Block on the encoder's feature representation capacity, we conducted experiments with different settings for d. Fig. 2(d) illustrates the impact of the hyper-parameter d on model performance. As the hyperparameter d increases, the model's expressiveness is enhanced, and the metrics improve. Both mAP and Top1-F1 peak at d=128. This indicates that moderately increasing the dimension of the bottleneck layer provides the model with a richer feature representation space, thus effectively boosting performance. However, when d exceeds 128 and continues to increase, we observe a decline in model performance. At d=256, the mAP drops to 9.0%, and Top1-F1 decreases to 12.2%. Although performance slightly recovers at d=384, it still does not reach the optimal level achieved at d=128. At d=512, all metrics decrease again. This suggests that an overly large bottleneck layer dimension does not lead to continuous performance gains. Instead, the increased number of model parameters may lead to a risk of overfitting, preventing the RS Block from effectively extracting or fusing critical features from the encoder, thus impairing the model's generalization ability. Based on the comprehensive experimental results, setting the bottleneck layer dimension d to 128 is the optimal choice for the RS Module in this study, as it achieves the best balance between the model's expressive power and parameter efficiency.

References

[1] Kefan Chen, Hong Hu, Chaozhan Chen, Long Chen, and Caiying He. An intelligent sewer defect detection method based on convolutional neural network. In *2018 IEEE International conference on information and automation*, pages 1301–1306. IEEE, 2018.

[2] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[3] 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.

[4] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3513–3521, 2024.

[5] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[6] 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.

[7] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021.

[8] Yu Wang, Jiahao Fan, and Yanan Sun. Classification of sewer pipe defects based on an automatically designed convolutional neural network. *Expert Systems with Applications*, 264: 125806, 2025.

[9] Qian Xie, Dawei Li, Jinxuan Xu, Zhenghao Yu, and Jun Wang. Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE Transactions on Automation Science and Engineering*, 16(4):1836–1847, 2019.