Automated Subway Tunnel Leakage Segmentation: A Data-Efficient Fine-Tuning Strategy Based on Large Vision Foundation Models

Longfei Dai[1], Zhiyao Tian[2*], Shunhua Zhou[1]

[1]Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University, China
[2]Department of Architecture and Civil Engineering, College of Engineering, City University of Hong Kong, China

2310828@tongji.edu.cn, zhiytian@cityu.edu.hk, zhoushh@tongji.edu.cn

Abstract –

Leakage detection in subway tunnel linings remain critical yet challenging tasks. Conventional deep learning models typically require tens of thousands of training samples that are scarce in engineering practice. Large Vision Models (LVMs) possess impressive general-purpose capabilities but fall short when applied to specialized downstream tasks. To address this gap, this study proposes a tailored framework to fine-tune LVMs for leakage segmentation. Built upon Segment Anything Model 2 (SAM2), this framework comprises four dedicated components to address the complex visual characteristics of subway tunnels: (i) a tunable module for enhancing leakage edge delineation; (ii) embedding generators for projecting visual features into a leakage-adapted representation space; (iii) adaptors for learning leakage-specific representations; and (iv) a fully fine-tuned decoder for pixel-level segmentation. These components collectively activate SAM2's rich prior knowledge, facilitating high-accuracy leakage segmentation. Experiments demonstrate that the proposed method improves untuned SAM2's accuracy by 3 to 4 times. Remarkably, it achieves performance comparable to DeepLabV3+ (a state-of-the-art deep learning model trained on 2000 samples) using only 50 samples.

Keywords –

Data-Efficient Fine-tuning; Large Vision Model; Leakage Segmentation; Shield Tunnels; Intelligent Infra-structure Inspection

1 Introduction

With the rapid expansion of subway networks, water leakage remains a critical threat to operational safety [1-2]. Current manual detection is labor-intensive and subjective, creating a pressing need for efficient, high-precision, and fully automated solutions.

To overcome the limitations of traditional inspection methods, researchers have increasingly explored deep learning-based computer vision techniques for tunnel leakage detection. In recent years, convolutional neural networks (CNNs) have demonstrated remarkable advantages in automated leakage detection. For instance, Chen et al. [3] developed Leakage-YOLO, an enhanced version of YOLOv8, while Xue et al. [4] modified the Faster R-CNN model to further improve detection accuracy. However, these object detection methods cannot achieve pixel-level segmentation. To address this, Xue et al. [5] employed a Cascade Mask R-CNN framework to achieve segmentation of tunnel leakage. Zhao et al. [6] integrated ResNet-101 with Mask R-CNN for classification and segmentation of leakage images, while Feng et al. [7] enhanced the encoder structure of U-net and U-net++ to develop a U-shape model. These deep learning-based approaches have demonstrated promising results on specific datasets. However, these models require large annotated datasets (often comprising tens of thousands of images) for multiple rounds of training and validation [6]. Acquiring and labeling such large-scale datasets is labor-intensive and time-consuming, a difficulty compounded by the intrinsic scarcity of seepage data in engineering practice [8].

In recent years, Vision Transformers (ViTs) and Large Vision Models (LVMs) have attracted significant attention and have been widely adopted across a broad spectrum of visual domains [9-10]. Notably, one representative model, the Segment Anything Model (SAM) demonstrates strong zero-shot segmentation capability across diverse objects such as humans, vehicles, and animals [11]. Extending this potential to tunnel inspection, Wei et al. [12] extracted the encoder module of SAM and integrated it with an enhanced U²-Net to construct the DeU²-Net model for leakage segmentation. Nevertheless, as their method remains built upon traditional backbone architectures, it still requires extensive training data.

An emerging approach that has attracted increasing attention is the fine-tuning of LVMs with domain-specific knowledge, enabling these models to exhibit data-efficient learning behavior. However, tunnel leakage typically appears as a surface-attached, weak-edged, and low-contrast pattern on concrete, which can easily be mistaken for background noise and thus overlooked. Moreover, subway tunnels often exhibit complex visual conditions, including low illumination, texture interference, and strong visual clutter, making reliable segmentation even more difficult. Existing frameworks largely rely on uniform architectural fine-tuning strategies [13]. Fine-tuning an LVM for leakage segmentation, specifically considering the distinctive visual and structural characteristics of leakage patterns, remains largely unexplored.

To address this gap, we propose a specialized fine-tuning framework for tunnel leakage segmentation. Built on SAM2, this data-efficient learning strategy leverages pre-trained priors instead of complex training paradigm, ensuring high accuracy through a simpler, more robust training process practical for infrastructure maintenance. A Leakage Information Enhancement Module first amplifies leakage boundaries and suppresses background noise in the frequency domain. The Embedding Generator then projects general encoder features into a leakage-adapted latent space, enriching the model's semantic understanding of leakage patterns. These enhanced and projected features are integrated by Adaptors, which inject task-specific prompts into the SAM2 encoder to guide attention toward leakage-relevant regions. Finally, a fully fine-tuned Mask Decoder performs pixel-level segmentation, producing high-precision leakage masks with sharp boundaries and coherent spatial consistency.

2 Methodology

2.1 The Original SAM2 Architecture

SAM2 is a large vision foundation model specifically developed for image semantic segmentation [14]. Building upon the powerful segmentation capabilities of its predecessor SAM, SAM2 further enhances segmentation accuracy, boundary delineation, and computational efficiency. The model is based on the Vision Transformer (ViT) architecture and is specifically designed for Promptable Visual Segmentation, a segmentation paradigm that performs target segmentation through prompts such as points, boxes, or masks. As illustrated in Figure. 1 [14], Image Encoder, Prompt Encoder, and Mask Decoder constitute the core structure of SAM2.

Although the original SAM2 model possesses strong prior knowledge and demonstrates excellent generalization capacity, preliminary experiments (further discussed in Section 3.4) reveal several limitations when it is directly applied to tunnel leakage segmentation: (i) Dependency on user prompts. SAM2 requires user-provided prompts for each image, rendering it impractical for large-scale, automated workflows. (ii) Low leakage segmentation accuracy. Due to the lack of domain-specific training, SAM2 struggles with complex tunnel backgrounds, often misidentifying structural features (Figure. 1), while overlooking actual leakage regions.

2.2 The Proposed Framework

2.2.1 Framework overview

To address the suboptimal performance of SAM2 in tunnel environments, a fine-tuning framework is proposed (Figure 2) based on a parameter-efficient fine-tuning (PEFT) strategy. To minimize computational overhead, the heavy Image and Prompt Encoders are frozen; consequently, only 3.94 M parameters (representing approximately 1.76% of the 224.51 M total parameters) are updated. These trainable parameters comprise the fully fine-tuned Mask Decoder and three modules: the Leakage Information Enhancement Module, the Embedding Generator, and the Adaptors. This configuration ensures a lightweight architecture that maintains inference speeds and memory usage nearly identical to the original SAM2 while enabling effective domain knowledge learning.

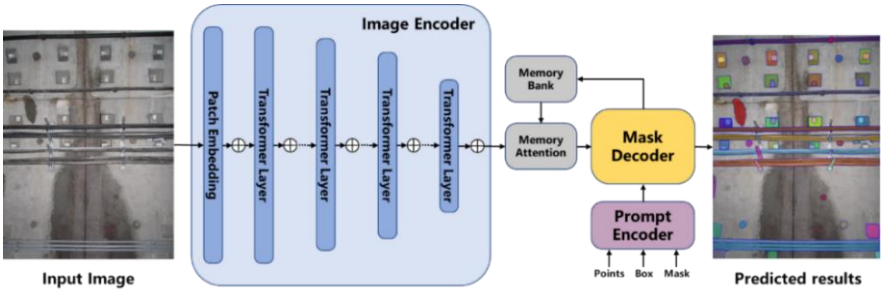

Figure 1. Architecture of the SAM2 large model and segmentation results on tunnel leakage.

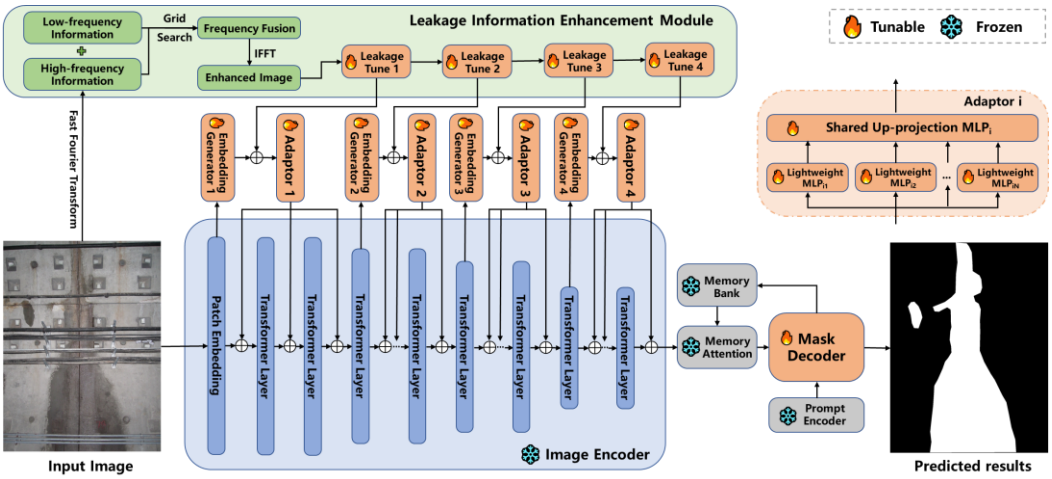

Figure 2. Architecture of the proposed fine-tuning framework.

2.2.2 Leakage Information Enhancement Module

The leakage recognition task in shield tunnels exhibits distinct challenges: (i) leakage boundaries are often blurred and gradual, lacking clear structural features; (ii) the tunnel background contains strong visual noise, including pipelines, bolts, grouting holes, and segment joints, all of which have sharp edges. Targeting these issues, an information enhancement module is designed.

The raw image I (of size $H\times W$) is first transformed into the frequency domain using the Fast Fourier Transform (FFT): z=FFT(I). The low-frequency components are shifted to the spectrum center (H/2, W/2), with frequency increasing radially outward. Two binary masks are constructed to extract different frequency bands:

(i) High-frequency mask $M_{\mathrm{h}}(\tau)$ for extracting rapidly varying edge and texture information:

$$M_{\mathrm{h}}^{i,j}(i,j,\tau)=\begin{cases}0, & \dfrac{4\,|(i-H/2)(j-W/2)|}{HW}<\tau\\ 1, & \text{otherwise}\end{cases} \tag{1}$$

(ii) Low-frequency mask $M_{\mathrm{l}}(\tau)$ for capturing slowly varying structural information and overall brightness:

$$M_{\mathrm{l}}^{i,j}(i,j,\tau)=\begin{cases}0, & \dfrac{HW-4\,|(i-H/2)(j-W/2)|}{HW}<\tau\\ 1, & \text{otherwise}\end{cases} \tag{2}$$

The corresponding high- and low-frequency components are then obtained as:

$$z_{\mathrm{h}}=zM_{\mathrm{h}}(\tau) \tag{3}$$

$$z_{\mathrm{l}}=zM_{\mathrm{l}}(\tau) \tag{4}$$

where the high-frequency component z_{h} captures sharp pixel variations, i.e., edges and texture details, including the boundaries of leakage, pipelines, and bolt holes. The low-frequency component z_{l} represents slowly varying regions, such as the main leakage body and background concrete surface. To highlight leakage characteristics while suppressing interference from irrelevant features, this module performs weighted fusion of the two components:

$$z_{\mathrm{e}}=\alpha z_{\mathrm{h}}+\beta z_{\mathrm{l}} \tag{5}$$

where α and β are learnable hyperparameters. After task-specific fine-tuning, the learned weights yield two key effects: (i) Leakage edges becomes clearer, while isolated non-target structures are attenuated; (ii) The gradient information of the main leakage region is retained, whereas smooth concrete backgrounds are weakened.

Finally, the enhanced image is reconstructed via the inverse FFT (IFFT): I_e=IFFT(z_e), and subsequently passed through four trainable Overlapping Patch Embedding layers, referred to as Leakage Tune modules. These modules are specifically designed to process enhanced leakage features corresponding to the four resolution stages of the SAM2 Image Encoder. This process generates multi-scale enhanced features F_{E1}, …, F_{E4}, which are then fed into the subsequent Adaptor modules for further feature learning.

2.2.3 Embedding Generator

Embedding Generators are further introduced to enhance the model's auto-prompting capability for leakage detection. Each Embedding Generator serves as an auxiliary input source for the Adaptor modules. Following the four-resolution stages of the SAM2 Image Encoder, four trainable Embedding Generator modules are constructed accordingly: When the raw image I enters the Image Encoder, it first passes through the Patch Embedding layer for patch extraction and embedding, resulting in the initial embedded feature I_p. Subsequently, Embedding Generator 1 projects this feature into a task-adaptive feature space:

$$F_{G1}=L_{EG1}(I_p) \tag{6}$$

The remaining three Embedding Generators follow a

similar process; however, their inputs are taken from the first Transformer layer of each corresponding resolution stage (see Figure 2). $L_{EGi}(.)$ (i = 1, 2, 3, 4) denotes a learnable linear transformation layer. By fine-tuning its internal weight matrices and bias terms, the generic visual features produced by the pretrained Encoder are reprojected into the leakage-specific task space. After projection, the resulting features F_{G1}, …, F_{G4}, are obtained sequentially and serve as additional inputs to the subsequent Adaptor modules.

2.2.4 Adaptor

Adaptors are further introduced to fuse the previously obtained leakage-enhanced features F_{Ei} and embedding features F_{Gi} (i=1, 2, 3, 4), guiding task-specific knowledge into the original SAM2 network based on the concept of prompt injection. This allows fine-grained adaptation and enhancement for downstream tasks without compromising the ability of the original model [15].

Four adaptors are constructed, corresponding to the four resolution stages of the Image Encoder (see Figure 2). For each Adaptor, F_{Ei} and F_{Gi} (i=1, 2, 3, 4) are fused to form the input embeddings F_{Ini}, which is subsequently processed multiple times within the internal structure of Adaptor. As shown in Figure 2, each Adaptor adopts a lightweight two-stage multilayer perceptron (MLP) architecture.

F_{Ini} first enters the initial linear layer of the first stage MLP_{i1}, passes through a *GELU* activation function, and is then fed into the shared projection layer MLP_i, producing the output prompt vector:

$$P_{i,1} = MLP_i(GELU(MLP_{i,1}(F_{INi}))) \tag{7}$$

Subsequently, F_{Ini} is sequentially processed by MLP_{i2}, …., MLP_{iN}. By fine-tuning the linear weight matrices and bias terms of the MLP layers, the model generates a set of leakage task-specific prompt embeddings $\{P_{i,1}, ..., P_{i,N}\}$, which are injected layer by layer into the outputs of the corresponding network layers in the i-th stage of the Image Encoder.

2.2.5 Fully fine-tuned Mask Decoder

The Mask Decoder is fully fine-tuned due to its pivotal role in mask generation and its parameter efficiency. Specifically, the Mask Decoder receives the embedding features through the previous framework, and processes them via a combination of multi-layer Transformers, upsampling modules, and Hypernetwork structures to produce multi-scale candidate masks along with their corresponding confidence scores. During fine-tuning, manually annotated leakage regions are used as supervision signals, and the model parameters are updated. Through the synergistic enhancement of the preceding modules, the model ultimately generates leakage masks with sharp boundaries, coherent structures, and high semantic consistency.

3 Experiment

3.1 Datasets

An extensive dataset was provided by Xue et al. [5], comprising 4755 images of shield tunnel leakages collected from Shanghai Metro tunnels. After verifying the one-to-one correspondence between images and labels and removing 196 invalid samples, we obtained a final set of 3359 annotated images for model training and 1000 annotated images to evaluate model performance. Representative examples from this dataset are shown in Figure 3.

3.2 Evaluation Metrics

To quantitatively evaluate the performance of the proposed method, two commonly used segmentation metrics are employed : Intersection over Union (*IoU*), Dice coefficient (*Dice*). The definitions of these metrics are as follows:

$$IoU = \frac{1}{N}\sum^{N}\frac{Int}{Uni} \tag{8}$$

$$Dice = \frac{1}{N}\sum^{N}\frac{2\times Int}{Int+Uni} \tag{9}$$

where, N denotes the total number of testing samples, *Int* and *Uni* refer to the intersection and union between ground truth mask and predicted mask generated by the model, respectively .

Figure 3. Representative samples from the used dataset, adapted from [5].

3.3 Experimental Settings

Among the four official SAM2 checkpoints, we adopted hiera-large as the backbone to fully leverage its strong prior knowledge and superior pretrained performance. During fine-tuning process, a joint loss function combining Binary Cross-Entropy (BCE) and *IoU* loss was employed to optimize the network. The batch size and number of training epochs were set to 2 and 40, respectively. To fine-tune the hyperparameters α, β, both defined within [0,1], a systematic grid search method was introduced. Candidate values for α and β were evaluated at intervals of 0.25. The grid-search results are presented in Figure 4, which identifies α=0.25 and β=0.75 as the optimal hyperparameter combination across all evaluation metrics; these values were therefore adopted for the final model configuration.

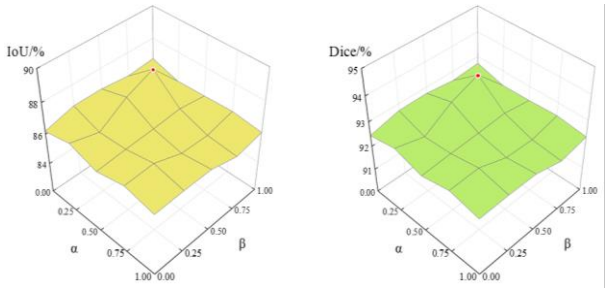

Figure 4. Grid search results for α and β.

3.4 Experimental Results

The validation set, i.e., the remaining 1,000 annotated images [5], was used to evaluate the performance of the fine-tuned model. For comparison, the original SAM2 model was also tested on the same dataset. Because SAM2 requires a manual prompt as an additional input, we simulated this process by randomly selecting one point from the ground-truth mask of each image as the prompt, which actually overestimates the original SAM2 model performance. The results are shown in Table 1.

The fine-tuned model substantially outperforms the original SAM2, achieving improvements by a factor of approximately 3–4. Specifically, the fine-tuned model attains an IoU of 87.01% and a Dice (F1-score) of 92.90%, with Precision and Recall reaching 92.34% and 93.89%, respectively. Representative segmentation results for the two models are shown in Figure 5.

Table 1. Performance comparison on the annotated validation dataset

Model	IoU(%)	Dice(%)
Proposed Framework	**87.01**	**92.90**
Original SAM2	21.49	30.75
UNet	83.07	90.75
SegFormer	85.88	92.42
Mask R-CNN	83.05	90.74
DeepLabV3+	85.96	92.45

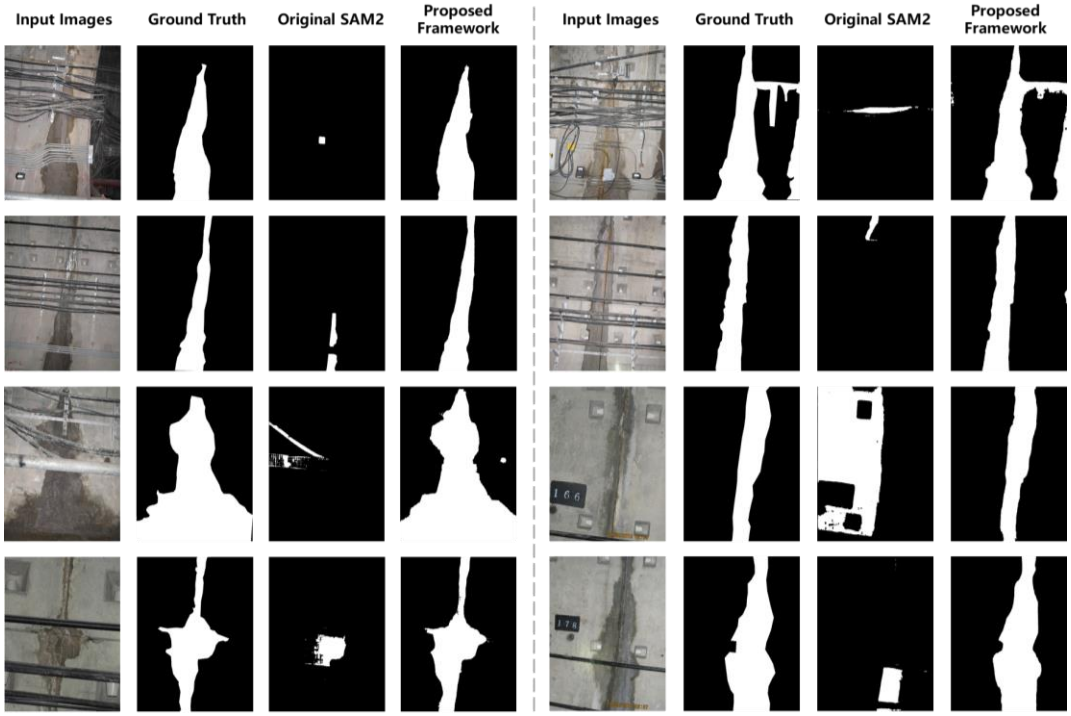

Figure 5. Representative leakage segmentation results on the annotated validation dataset.

4 Discussion

4.1 Comparison with the State-of-the-Art Model

The proposed approach is evaluated against state-of-the-art conventional deep learning models widely utilized for tunnel leakage segmentation. As summarized in Table 1, the framework is compared with several mainstream architectures, including UNet, SegFormer, Mask R-CNN, and DeepLabV3+, all of which are surpassed by the proposed method in terms of accuracy. Notably, DeepLabV3+ emerges as the strongest baseline and is therefore selected for a more rigorous comparative analysis under data-scarce conditions.

Consequently, DeepLabV3+ with an Xception backbone serves as the primary benchmark for evaluating model performance across increasingly constrained data regimes. To systematically investigate data efficiency and model robustness, experiments are conducted across eight training subsets , ranging from 3,000 images down to 10 randomly sampled images.

Table 2. Performance comparison between the proposed framework and DeepLabV3+

Training Set Size	Proposed Framework		DeepLabV3	
	IoU(%)	Dice(%)	IoU(%)	Dice(%)
3359	**87.01**	92.90	**85.96**	92.45
3000	**86.73**	92.72	**80.64**	88.28
2000	**85.37**	91.90	**71.31**	80.48
1000	**83.90**	90.85	**36.68**	47.05
500	**83.66**	90.73	**20.87**	30.44
100	**73.64**	83.71	**1.370**	2.540
50	**69.72**	80.95	**1.170**	2.140
20	**47.46**	63.15	**0.600**	1.160
10	**2.450**	4.780	**2.020**	3.740

The experimental results are summarized in Table 2, and the corresponding trend of IoU performance with respect to training set size is illustrated in Figure 6. The results show that when the training set contains only 10 images, both the fine-tuned SAM2 and DeepLabV3+ perform extremely poorly. Notably, the performance of the fine-tuned SAM2 improves steadily as the training set size increases. When the training set reaches 50 images, the IoU of the large model approaches 70%, which represents a reasonably competitive performance for applications with moderate segmentation requirements. In contrast, DeepLabV3+ shows almost no improvement at this scale, achieving an IoU of only 1.17%, suggesting that it requires substantially more data to learn effectively.

When the training set size reaches 500 images, the fine-tuned SAM2 achieves metrics exceeding 80% across all evaluation criteria, with Dice surpassing 90%, which corresponds to a high-accuracy segmentation performance commonly reported across applied imaging domains. As the dataset size increases further, the accuracy of the fine-tuned large model continues to improve, although the performance gains become less pronounced. Interestingly, DeepLabV3+ requires approximately 2000 training images to achieve a performance comparable to that of the fine-tuned large model trained with 50 images. This finding highlights the strong data-efficient learning capability of the proposed framework. When both models are trained with 3000 images, their performances become broadly similar; however, DeepLabV3+ still exhibits slightly lower overall accuracy (Figure6).

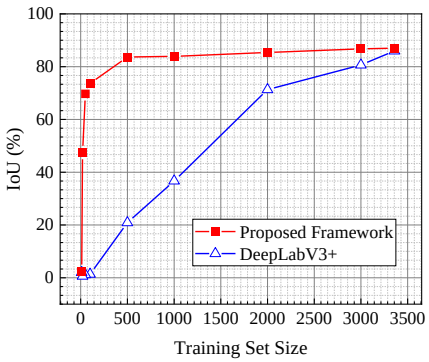

Figure 6. IoU performance versus training set size for the proposed framework and DeepLabV3+.

In summary, conventional deep learning models require large quantities of annotated images to achieve accurate shield-tunnel seepage segmentation. In contrast, although a large vision model initially lacks domain-specific knowledge, it exhibits a fundamentally different learning behavior. Once provided with even a small amount of task-specific supervision, the large model can rapidly activate its latent knowledge and extensive pretrained representations, enabling it to perform the previously unfamiliar task effectively. Importantly, the amount of data required for such fine-tuning is significantly smaller than that demanded by conventional deep learning approaches.

This property is particularly valuable for engineering applications, where annotated data are often scarce or difficult to obtain. For example, tunnel leakage images are typically controlled by transportation authorities or research institutions, resulting in inherently limited datasets. In such scenarios, the ability to fine-tune a large model is transformative, making effective machine learning workflows feasible even under limited data conditions. To further demonstrate this advantage, an engineering case study will be presented in the following subsection.

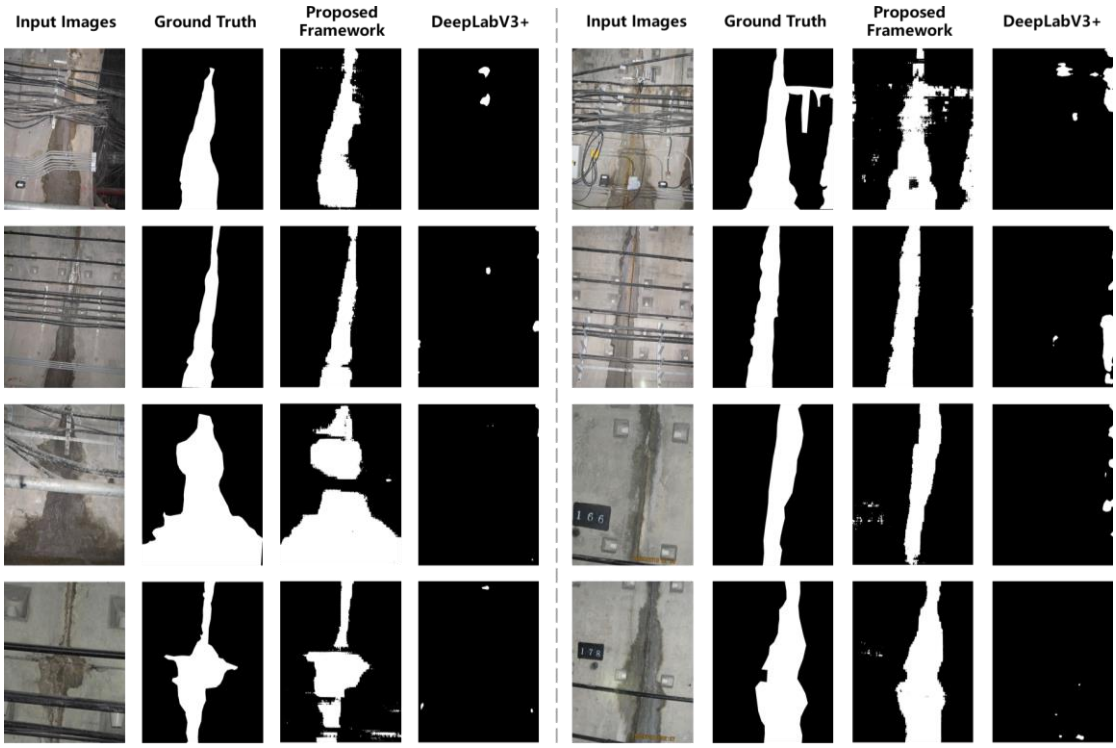

Figure 7. Representative leakage segmentation results on the annotated validation dataset using 50 training samples.

4.2 Ablation Study

In the proposed framework, the Leakage Information Enhancement Module and the Embedding Generator are architecturally coupled, as they generate the synergistic inputs (F_{Ei} and F_{Gi}) required by the Adaptors. These Adaptors serve as the central mechanism for injecting task-specific knowledge into the frozen SAM2 backbone through a prompt-injection strategy. Due to this functional interdependence, the Adaptors are consistently activated in our experimental configurations whenever either the Leakage Information Enhancement Module or the Embedding Generator is enabled to ensure the effective transmission of domain knowledge through the injection chain. To further investigate how this specialized fine-tuning transforms a general large model into a high-precision domain-specific one, four experimental configurations were designed to isolate the contribution of these key components. All configurations were trained on the same dataset and evaluated on the validation set, with the quantitative results summarized in Table 3.

It can be observed in Table 3 that fine-tuning only the Mask Decoder yields a remarkable improvement, increasing IoU from 21.49% to 78.47%. This substantial gain arises from the fact that, fine-tuned Mask Decoder refines the model's understanding of specific leakage cues, allowing it to distinguish weak-edged, low-contrast patterns from the generic objects emphasized in SAM2's pretraining. Nevertheless, the model still exhibits limitations in accuracy, falling short of the requirements for high-precision engineering applications.

Table 3. Results of the ablation study

Fine-tuning configurations	IoU(%)	Dice(%)
Original SAM2	21.49	30.75
Only Mask Decoder	78.47	87.51
Embedding Generator + Mask Decoder	85.89	92.17
Leakage Information Enhancement Module + Mask Decoder	85.61	92.01
Complete Framework	**87.01**	**92.90**

After introducing the Embedding Generator, the performance is further improved, achieving 85.89% IoU. The improvement arises from domain adaptation performed by the Adapter, together with the more effective prompt guidance provided by the Embedding Generator.

Similarly, adding the Leakage Information Enhancement Module produces gains comparable to those achieved with the Embedding Generator, attaining 85.61% IoU. Notably, the Leakage Information Enhancement Module strengthens the high-frequency boundary components of leakage regions in the frequency domain while suppressing background interference.

Finally, the full configuration achieves the optimal performance of 87.01% IoU.

5 Conclusion

This study adapts Segment Anything Model 2 into a specialized tunnel leakage segmentation model through a parameter-efficient fine-tuning framework. The main conclusions are as follows:

(i) The proposed framework overcomes SAM2's original lack of domain knowledge, transforming it into a high-precision tool for tunnel leakage detection.

(ii) The model exhibits remarkable data-efficient capability, where only 50 annotated samples achieve segmentation accuracy comparable to state-of-the-art model trained on 2,000 samples.

(iii) The framework's superiority stems from leveraging SAM2's pretrained priors through the synergy of the Mask Decoder and the proposed Leakage Information Enhancement Module, Embedding Generator, and Adaptors, which collectively enhance boundary delineation and generalization.

Preliminary tests on a Nanjing Metro dataset achieved 83.12% IoU through direct validation without retraining, demonstrating robust generalization. Future research will utilize multi-scale ablations and multi-city studies to further validate cross-domain adaptability and probe theoretical limits.

References

[1] Wang, L., Chen, H., Liu, Y., Li, H., & Zhang, W. Application of copula-based Bayesian network method to water leakage risk analysis in cross river tunnel of Wuhan Rail Transit Line 3. Advanced Engineering Informatics, 57, 102056, 2023.

[2] Tian, Z., Xu, P., Gong, Q., Zhao, Y., & Zhou, S. Health-degree model for stagger-joint-assembled shield tunnel linings based on diametral deformation in soft-soil areas. Journal of Performance of Constructed Facilities, 37(3), 04023019, 2023.

[3] Chen, C. & Liu, W. Leakage-YOLO: Real-Time Object Detection Algorithm for Crack and Leakage in Tunnel Scenarios. Journal of Computer Engineering & Applications, 61(6), 2025.

[4] Xue, Y., Gao, J., Li, Y., & Huang, H. Optimization of Shield Tunnel Lining Defect Detection Model Based on Deep Learning. Journal of Hunan University, 47(7), 137-146, 2020.

[5] Xue, Y., Cai, X., Shadabfar, M., Shao, H., & Zhang, S. Deep learning-based automatic recognition of water leakage area in shield tunnel lining. Tunnelling and Underground Space Technology, 104, 103524, 2020.

[6] Zhao, S., Shadabfar, M., Zhang, D., Chen, J., & Huang, H. Deep learning‐based classification and instance segmentation of leakage‐area and scaling images of shield tunnel linings. Structural Control and Health Monitoring, 28(6), e2732, 2021.

[7] Feng, S. J., Feng, Y., Zhang, X. L., & Chen, Y. H. Deep learning with visual explanations for leakage defect segmentation of metro shield tunnel. Tunnelling and Underground Space Technology, 136, 105107, 2023.

[8] Yu, X., He, B. G., Xu, X., Zhou, Y., Diaz, M. A., Chen, J., & Camacho, D. Application of Artificial Intelligence in Rock Tunnel Engineering: A Survey on Where and How. Expert Systems, 42(7), e70080, 2025.

[9] Wei, X., Li, J., Zhang, X., Gu, H., Van de Weghe, N., & Huang, H. An innovative framework combining a CNN-Transformer multiscale fusion network and spatial analysis for cycleway extraction using street view images. Sustainable Cities and Society, 106384, 2025.

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022), 2021.

[11] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4015-4026), 2023.

[12] Wei, Z., Ji, A., Xiao, Z., & Zhang, L. Segment anything model enhanced dual encoder segmentation for seepage quantification in operational tunnels. Advanced Engineering Informatics, 69, 104030, 2026.

[13] Zhou, Y., Sun, G., Li, Y., Xie, G. S., Benini, L., & Konukoglu, E. When sam2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation. Visual Intelligence, 3(1), 10, 2025.

[14] Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., ... & Feichtenhofer, C. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.

[15] Chen, T., Lu, A., Zhu, L., Ding, C., Yu, C., Ji, D., ... & Zang, Y. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. arXiv preprint arXiv:2408.04579, 2024.

[16] Wang, D., Hou, G., Chen, Q., Li, W., Fu, H., Sun, X., & Yu, X. Segmentation of tunnel water leakage based on a lightweight DeepLabV3+ model. Measurement Science and Technology, 36(1), 015414, 2024.