infrastructures

MDPI

Article

Deep Learning for Underwater Crack Detection: Integrating Physical Models and Uncertainty-Aware Semantic Segmentation

Wenji Ai [1], Zongchao Liu [2], Shuai Teng [3,*], Shaodi Wang [4] and Yinghou He [5]

[1] College of Railway and Transportation, Guangzhou Railway Polytechnic, Guangzhou 511300, China; aiwenji@gtxy.edu.cn
[2] School of Finance and Commerce, Guangzhou Railway Polytechnic, Guangzhou 511300, China; liuzongchao@gtxy.edu.cn
[3] School of Intelligent Construction and Civil Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China
[4] Earthquake Engineering Research & Test Center, Guangzhou University, Guangzhou 510006, China; wangsd@gzhu.edu.cn
[5] Research Centre for Wind Engineering and Engineering Vibration, Guangzhou University, Guangzhou 510006, China; heyinghou@gzhu.edu.cn
* Correspondence: tengs@zut.edu.cn; Tel.: +86-15914485576

Abstract

Underwater crack detection is critical for ensuring the safety and longevity of submerged infrastructures, yet it remains challenging due to water-induced image degradation, limited labeled data, and the poor generalization of existing models. This paper proposes a novel deep learning framework that integrates physical priors and uncertainty modeling to address these challenges. Our approach introduces a physics-guided enhancement module that leverages underwater light propagation models, and a dual-branch segmentation network that combines semantic and geometry-aware curvature features to precisely delineate irregular crack boundaries. Additionally, an uncertainty-aware Transformer module quantifies prediction confidence, reducing the number of overconfident errors in ambiguous regions. Experiments on a self-collected dataset demonstrate State-of-the-Art performance, achieving 81.2% mIoU and 83.9% Dice scores, with superior robustness in turbid water and uneven lighting. The proposed method introduces a novel synergy of physical priors and uncertainty-aware learning, advancing underwater infrastructure inspection beyond the current data-driven approaches. Our framework offers significant improvements in accuracy, robustness, and interpretability, particularly in challenging conditions like turbid water and non-uniform lighting.

Keywords: underwater crack detection; semantic segmentation; physics-guided enhancement; uncertainty modeling; transformer networks; deep learning

check for updates

Academic Editor: Seyed Sasan Khedmatgozar Dolati

Received: 17 August 2025
Revised: 16 September 2025
Accepted: 20 September 2025
Published: 23 September 2025

Citation: Ai, W.; Liu, Z.; Teng, S.; Wang, S.; He, Y. Deep Learning for Underwater Crack Detection: Integrating Physical Models and Uncertainty-Aware Semantic Segmentation. *Infrastructures* **2025**, *10*, 255. https://doi.org/10.3390/infrastructures10100255

1. Introduction

The integrity assessment of underwater infrastructure is a critical component of broader structural health monitoring (SHM) frameworks. While traditional SHM for bridges and dams has often relied on networks of physical sensors (e.g., accelerometers or strain gauges) to monitor global response parameters, visual inspection remains the primary method for identifying localized damage such as cracking. The automation of this visual inspection, particularly in challenging underwater environments, is therefore a vital pursuit to complement existing SHM methodologies and create more holistic, data-driven integrity management systems. Underwater infrastructures such as bridge piers, offshore platforms, and subsea pipelines play a vital role in transportation and energy systems [1]. Among the various types of structural damage that can occur in these environments, cracks are considered one of the most critical indicators of early-stage degradation [2,3]. The timely detection of cracks is paramount not only for assessing structural integrity, but also for ensuring functional performance. For hydraulic structures like dams and reservoirs, cracks directly compromise durability and impermeability, leading to seepage, internal erosion, and other detrimental processes that can accelerate deterioration. This underscores the practical engineering significance of developing robust, high-precision crack detection systems. If left undetected, they may propagate over time and lead to structural failure, posing risks to public safety and causing substantial economic losses [4]. Therefore, the timely and accurate detection of underwater cracks is essential for ensuring the integrity and reliability of these submerged structures [5].

Traditional methods for underwater crack detection largely rely on manual inspections, remotely operated vehicles (ROVs) [6,7], or sonar-based imaging systems [8]. While sonar and laser scanning are effective in highly turbid conditions, they often lack the resolution necessary to detect fine-grained surface cracks [9,10]. In contrast, optical imaging offers high-resolution and texture-rich visual information that is more intuitive for identifying small-scale surface defects. However, underwater optical images suffer from a variety of degradations, including color distortion, low contrast, scattering, and non-uniform illumination caused by the inherent physics of light propagation in water [11]. Beyond vision-based techniques, the broader field of structural health monitoring (SHM) has developed advanced methodologies for damage assessment in marine and offshore environments. Particularly relevant are vibration-based strategies that analyze the dynamic responses of structures to environmental loads, such as wind and wave forces [12]. These challenges make the task of underwater crack detection particularly difficult using conventional image processing or even standard deep learning models.

Recent advancements in deep learning, especially convolutional neural networks (CNNs) [13] and Transformer-based architectures [14,15], have shown great promise in automated crack detection for road pavements [16–18], concrete surfaces [19], and other civil infrastructures [20,21]. However, these models are primarily trained and validated on land-based datasets with clear visibility and well-structured textures. When applied directly to underwater scenes, their performance degrades significantly due to domain shifts and the lack of robustness to underwater-specific noise and degradation.

In recent years, research on image-based underwater structure damage recognition using deep learning has mainly focused on three technical directions: image classification, object detection, and semantic segmentation. It is important to note that this excludes a parallel body of work on non-image-based damage detection (e.g., using accelerometers, strain gauges, or acoustic emission sensors), which utilizes deep learning for time-series signal analysis rather than visual recognition. In the field of image classification (determining the existence of damage through global image analysis), Zhu et al. [22] proposes an improved VanillaNet architecture, which effectively alleviates the long-tail distribution problem of underwater dam crack data by introducing the Seesaw loss function. In comparison with advanced models such as ConvNeXtV2 and RepVGG, the classification accuracy is improved by 2.66% compared to the original network. For object detection technology (precise defect localization using bounding boxes), Li et al. [23] optimized the YOLOv4 framework by using lightweight MobileNetV3 instead of CSPDarknet as the feature extraction backbone network. By reconstructing the feature layer scale parameters and inputting the primary features into an improved feature fusion module, the deep integration of multi-level features was achieved.

Compared to the qualitative judgment of classification and the box selection localization of detection, semantic segmentation can provide more refined damage representation through its pixel-level recognition ability [24,25]. This technology achieves image semantic analysis through pixel-by-pixel classification, demonstrating significant advantages in the field of industrial inspection, especially in the analysis of crack morphology characteristics and the quantitative evaluation of the degree of damage, which has important engineering value. Typical cases include the following: Hou et al. [26] constructed an underwater bridge pier defect recognition system based on U^2-Net, and achieved the highest intersection to union ratio index of 0.73 by optimizing the edge detection module, which can still accurately extract contour details in complex underwater environments; Sun et al. [27] proposed a two-stage detection scheme, which first uses YOLOv7 to locate the defect areas of underwater concrete structures, and then uses an improved DeepLabV3+ network to complete pixel-level segmentation. For typical defects such as exposed steel bars and concrete spalling, the average intersection to union ratio reaches 0.914.

In response to the practical challenge of the difficulty in obtaining underwater defect data, the academic community has explored innovative paths to break through data bottlenecks. On the one hand, some researchers [28,29] have verified the effectiveness of transfer learning strategies in a few sample scenarios, such as the lightweight LinkNet framework constructed in reference [30], which achieves real-time crack segmentation and quantitative analysis in complex underwater environments through hybrid transfer learning. On the other hand, cutting-edge research attempts to integrate prior knowledge of physics. Teng et al. [31] pioneered the "knowledge-guided detection" paradigm, which extracts morphological features by calculating the fractal dimension matrix of cracks and uses it as a prompt input for the Segment Anything model (SAM) to construct a plug and play defect segmentation system. Ultimately, excellent performance indicators with an average accuracy of 97.6%, intersection to union ratio of 0.89, and F1 value of 0.95 were obtained.

However, most existing approaches treat the segmentation task purely as a data-driven problem [32], overlooking physical domain knowledge such as underwater light attenuation, the surface geometry of structures, and the inherent uncertainty in predictions. As a result, they often exhibit poor generalization when confronted with complex underwater environments, leading to false positives, missed detections, and overconfident predictions in ambiguous regions. While recent efforts in 2024–2025 have begun to incorporate physical models [5] or leverage foundational models like the SAM [31] for crack segmentation, they often treat these elements in isolation. Similarly, transfer learning strategies [33] address data scarcity, but do not explicitly model the inherent uncertainty in underwater predictions.

Beyond architectural advancements, quantifying the prediction uncertainty has emerged as a critical direction for improving the reliability of deep learning models in real-world applications, especially when dealing with domain shifts. Recent studies have increasingly incorporated uncertainty estimation into segmentation frameworks to identify ambiguous regions and enhance cross-domain robustness. For instance, Kwon et al. [34] proposed a Bayesian U-Net for medical image segmentation, using Monte Carlo Dropout to model epistemic uncertainty and improve generalization across different scanner domains. Similarly, Zhou et al. [35] developed an uncertainty-aware domain adaptation framework for semantic segmentation in autonomous driving, where entropy minimization on the target domain data helps to reduce prediction variance. These approaches demonstrate that explicitly modeling uncertainty is a powerful paradigm for addressing distribution shifts. However, their application in underwater structural inspection remains largely unexplored. Our work fills this gap by integrating an uncertainty-aware Transformer module specifically designed for the challenges of underwater optical imagery, such as scattering and low contrast, providing not only accurate segmentation, but also a crucial confidence measure for operational decision-making.

Unlike these approaches, this work is the first to concurrently integrate a physics-guided enhancement network, a geometry-aware dual-branch segmentation head, and an uncertainty-quantifying Transformer within a unified framework. This co-design allows each component to complement the others: the physical model corrects degradation at the input level, the geometric branch provides mid-level structural priors, and the uncertainty module offers output-level reliability estimation. This holistic strategy fundamentally differs from and advances upon incremental combinations of existing techniques, providing a more robust and interpretable solution for the underwater domain.

To overcome these limitations, this paper proposes a novel deep learning framework that integrates physical priors and uncertainty modeling for accurate crack detection in underwater optical images. Unlike existing methods, our approach introduces a physics-guided enhancement module that explicitly incorporates underwater light attenuation characteristics to improve visual quality before segmentation. This paper also proposes a dual-branch segmentation architecture that not only captures semantic information, but also learns curvature-based geometric features to better align with the physical properties of crack shapes. Furthermore, an uncertainty-aware Transformer module is integrated to estimate both epistemic and aleatoric uncertainties, allowing the model to identify ambiguous regions and suppress overconfident predictions. The major contributions of this study can be summarized as follows:

(1) A physics-guided enhancement network that explicitly integrates the underwater light propagation model to invert the image formation process, directly addressing color distortion and scattering artifacts at the source, rather than applying generic image enhancement;
(2) A geometric-aware dual-branch segmentation architecture that uniquely fuses high-level semantic features with low-level curvature maps, encoding the geometric property that cracks manifest as high-curvature surface irregularities. This provides a stronger inductive bias for precise boundary delineation than standard architectures;
(3) An uncertainty-aware Transformer module that leverages Monte Carlo Dropout not only for Bayesian uncertainty estimation, but also to actively guide the model's attention during training and inference toward ambiguous, low-confidence regions (e.g., faint cracks or strong reflections), significantly reducing the number of overconfident errors;
(4) Benchmarking and evaluation: this paper constructs a dataset of annotated underwater crack images, and extensive experiments are conducted comparing our method with several State-of-the-Art segmentation models.

This work offers a comprehensive solution to the underwater crack detection problem by fusing physical understanding with modern deep learning strategies. Experimental results demonstrate that our proposed method significantly outperforms the existing approaches in terms of segmentation accuracy, uncertainty quantification, and generalization ability, thus paving the way for safer and more efficient underwater infrastructure inspection.

2. Methods

This paper proposes an end-to-end deep learning system designed to address the core challenges of underwater crack detection: image degradation, data scarcity, and prediction uncertainty. The framework is structured in three synergistic components that operate in a logical sequence: (1) The physics-based restoration network first processes the raw underwater image to invert the optical degradation model, thereby recovering a clearer image with enhanced crack visibility. (2) The self-supervised generative module then leverages these restored images to synthesize a diverse set of additional training samples, mitigating the problem of limited annotated data. (3) Finally, the uncertainty-aware Transformer network performs the segmentation on enhanced and synthetic images, providing a precise crack map alongside a pixel-wise confidence estimate. This design ensures that each stage effectively handles a specific challenge, and that its output directly supports the subsequent stage, leading to a robust and reliable overall system.

2.1. Underwater Image Restoration Network Guided by Physical Modeling

Due to various physical factors such as light scattering, absorption (as shown in Figure 1), and fluctuations, underwater optical images often exhibit low contrast, severe color cast, and a loss of details during the shooting process. Therefore, directly detecting cracks based on such images will significantly reduce the performance and robustness of the model. Therefore, this article details an image restoration network that integrates underwater imaging physical mechanisms to preprocess the original image, restore its true visual information, and enhance the reliability of subsequent detection.

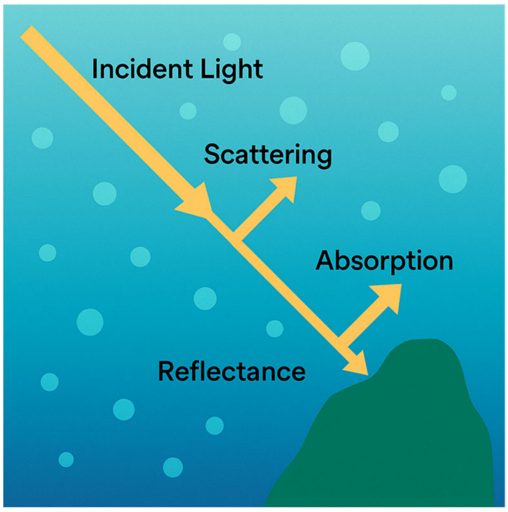

Figure 1. Underwater light propagation process.

2.1.1. Underwater Imaging Model Modeling

The degradation process in an underwater image is primarily governed by the physics of light propagation in water, which involves absorption and scattering effects. This process can be formally described using the widely accepted underwater image formation model:

$$I(x) = J(x) \cdot t(x) + B(x) \cdot (1 - t(x)) \tag{1}$$

where $I(x)$ is the captured image intensity, $J(x)$ is the scene radiance (the ideal image to be recovered), $B(x)$ is the background veiling light, and $t(x)$ is the medium transmission map, which can usually be estimated using the following equation:

$$t(x) = exp(-\beta \cdot d(x)) \tag{2}$$

Among these, β is the light attenuation coefficient and $d(x)$ represents the water depth or the distance between the object and the camera. In practical applications, this paper uses statistical or structural priors to estimate the depth, and combines learning methods to fit β, thereby achieving transmittance estimation. Our restoration network is designed to invert this physical model. It learns to estimate the parameters of this model to recover the latent radiance from the observed input, thereby moving beyond a purely 'black-box' enhancement.

2.1.2. Underwater Image Restoration Network

The overall image restoration network adopts an encoding–decoding structure, while introducing a physical prior branch (as shown in Figure 2). Its structure includes the following: (1) A backbone encoder: composed of 5 layers of convolution and a residual structure, extracting multi-scale semantic features, with channel numbers in the order of 64→128→256→128→64. (2) A physical prior path: Input a dual-channel physical image composed of estimated $d(x)$ and β, extract the features, and fuse them with the backbone features in the decoding stage. (3) A fusion decoder: uses separable convolution modules to reduce computational costs and preserves edge information through skip connections.

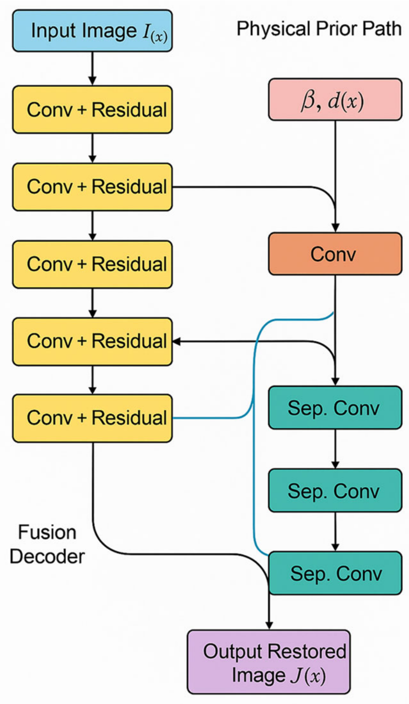

Figure 2. Underwater image restoration network.

Output as restored image $\hat{J}(x)$, i.e.:

$$\hat{J}(x) = F_{res}(I(x), \beta, d(x); \theta_{res}) \tag{3}$$

2.1.3. Loss Function Design

To train the physics-based restoration network in a supervised manner, a reference target $J^{*}(x)$ is required. Obtaining paired real-world data (i.e., the same scene captured in degraded underwater and ideal clear-water conditions) is fundamentally impossible. Therefore, we employ a two-pronged strategy to generate these references. For in-situ images where a true ground truth is unavailable, the reference $J^{*}(x)$ is defined as the best-available visual representation of the scene, selected by expert annotators. 'Clear-water' in this context is quantified based on perceptual metrics rather than absolute turbidity measurements.

To improve the perceptual quality and structural fidelity of image restoration, this paper adopts a multi-joint loss function for supervised training, including the following:

(1). Reconstruction loss:

$$L_{rec} = \left\| \hat{J}(x) - J^{*}(x) \right\| \tag{4}$$

Among these, $J^{*}(x)$ is an artificially synthesized or clear-water image, used as a reference target;

(2). Perceived loss:

$$L_{per} = \left\| \varnothing(\hat{J}) - \varnothing\left(J^{*}\right) \right\| \tag{5}$$

The VGG network's intermediate-level features are used to measure the distance of images in the high-level semantic space;

(3). Edge preservation loss:

$$L_{edge} = \left\| \nabla \hat{J}(x) - \nabla J^{*}(x) \right\| \tag{6}$$

It is used to maintain the boundary texture characteristics and enhance the distinguishability of crack edges.

The final total loss is as follows:

$$L_{\text{total\_res}} = \lambda_1 \cdot L_{rec} + \lambda_2 \cdot L_{per} + \lambda_3 \cdot L_{edge} \tag{7}$$

Among these, λ_1, λ_2, and λ_3 are weighting coefficients, with empirical settings of 1.0, 0.1, and 0.05, respectively.

The loss function (Equations (4)–(7)) not only minimizes pixel-wise differences but also incorporates perceptual and edge losses. This ensures that the network's output is not only physically plausible (adhering to the model), but also visually realistic and structurally consistent.

2.2. Self-Supervised Generative Enhancement Module

In the absence of a large number of annotated crack images, this paper adopts a generative model for the style transfer and simulation enhancement of crack images (as shown in Figure 3). This module adopts a reversible generative adversarial network (Invertible GAN), combined with the Flow model and CycleGAN mechanism, to achieve the diversity reconstruction of crack patterns.

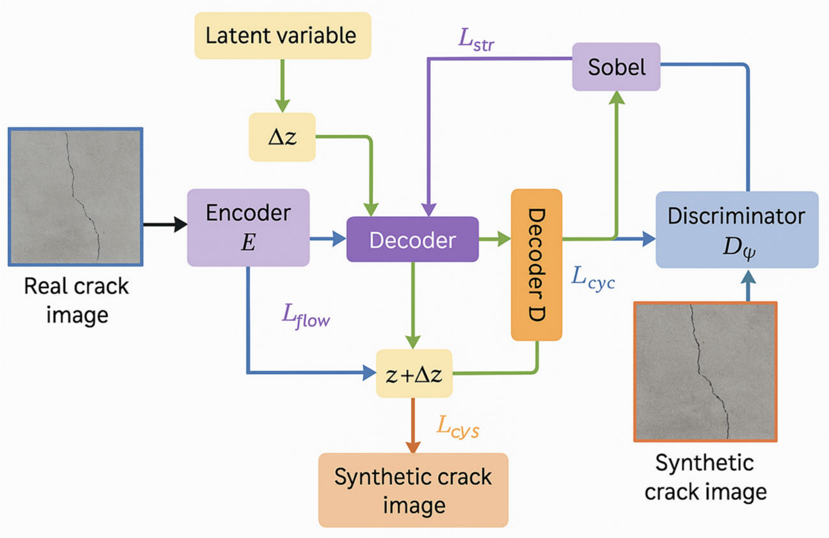

Figure 3. Self-supervised generative framework.

2.2.1. Latent Variable Modeling

Using the Flow module to learn the latent variable encoding z of images and satisfying the reversibility constraint: $z = E(J_c), \hat{j}_c = D(z)$. The training objectives are as follows:

$$L_{flow} = \|D(E(J_c)) - J_c\| + KL(z) \parallel \mathrm{N}(0,1) \tag{8}$$

It is used to ensure the consistency and reversibility of the mapping between the encoder and decoder.

2.2.2. Detail Enhancement Generation

To further enhance the details of cracks, a condition vector Δz is designed to represent fine-crack-level perturbations, which are added to the original encoding vector z to form a new composite image:

$$\tilde{J}_{c^+} = D(z + \Delta z) \tag{9}$$

And adversarial training is performed on real samples and synthetic samples through the discriminator D_Ψ:

$$L_{GAN} = E\left[logD_{\psi(J_c)}\right] + E[\log(1 - D_{\psi(\tilde{J}_{c^+})})] \tag{10}$$

2.2.3. Structural Consistency Loss

To avoid distortion or structural deformation in synthesized images, increase the structural loss and cyclic consistency loss:

$$L_{str} = \left\| Sobel\left(\tilde{J}_{c^+}\right) - Sobel(J_c) \right\| \tag{11}$$

$$L_{cyc} = \|E(D(z + \Delta z)) - (z + \Delta z)\| \tag{12}$$

The total loss is as follows:

$$L_{gen} = \alpha \cdot L_{flow} + \beta \cdot L_{GAN} + \gamma \cdot L_{cyc} + \delta \cdot L_{str} \tag{13}$$

Among these, the hyperparameters are set to α = 1, β = 0.5, γ = 0.2, and δ = 0.1.

The Invertible GAN and Flow model were trained using the Adam optimizer with a fixed learning rate of 1×10^{-4} for both the generator and the discriminator. The momentum parameters were set to β_1 = 0.5 and β_2 = 0.999. The latent dimension for the encoding vector z was set to 256. The weighting coefficients in the total generative loss (Equation (13)) were empirically set to α = 1, β = 0.5, γ = 0.2, and δ = 0.1 after an ablation study. Training was conducted for 500 epochs, with a batch size of 8.

The choice of an invertible architecture (Invertible GAN), combined with the inclusion of the cyclical consistency loss (L_{cyc}) and structural loss (L_{str}), significantly improved the training stability compared to with standard GANs. The reversible nature of the Flow model ensures a stable mapping between image and latent spaces, while the additional loss terms prevent mode collapse and mitigate the oscillatory behavior commonly observed in adversarial training. The model converged reliably across multiple random seeds.

A critical concern with generative augmentation is the potential introduction of unrealistic artifacts that could mislead the segmentation model. To mitigate this risk, our framework incorporates several key design choices: (1) The structural consistency loss (L_{str}) directly penalizes large deviations in the edge maps between original and synthesized images, preserving the fundamental crack morphology. (2) The cyclical consistency loss (L_{cyc}) ensures that the encoding–decoding process remains coherent, preventing extreme generations. (3) The adversarial loss is balanced by the flow and consistency losses, prioritizing faithful reconstruction over extreme novelty.

2.3. Transformer Segmentation Network with Uncertainty Perception

In order to improve the edge accuracy of crack detection and the robustness of the model to unknown environments, this paper introduces an uncertainty modeling segmentation network based on Vision Transformer (ViT). This network has strong long-range dependency modeling capabilities and achieves sample-selective fine tuning through Bayesian estimation, thereby reducing dependence on large-scale annotated samples.

2.3.1. Segmentation Network with Uncertainty Perception

The proposed dual-branch architecture is designed to explicitly leverage both the semantic context and the geometric properties of cracks. The core insight is that while semantic features are effective for classifying regions as 'crack' or 'background', the local geometric property of curvature provides a geometry-informed prior for precisely locating irregular crack boundaries, which often manifest as high-curvature contours on the structure's surface. The network architecture, illustrated in Figure 4, operates as follows:

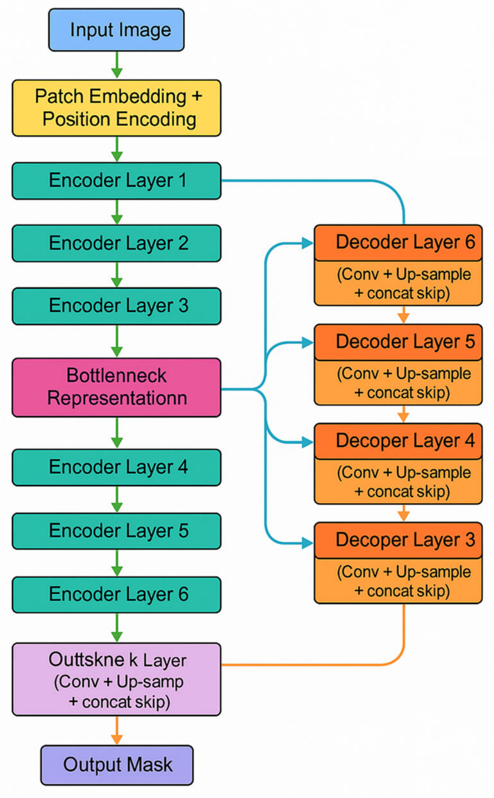
Input Image
Patch Embedding + Position Encoding
Encoder Layer 1
Encoder Layer 2
Encoder Layer 3
Bottlenneck Representationn
Encoder Layer 4
Encoder Layer 5
Encoder Layer 6
Outtskne k Layer (Conv + Up-samp + concat skip)
Output Mask
Decoder Layer 6
(Conv + Up-sample + concat skip)
Decoder Layer 5
(Conv + Up-sample + concat skip)
Decoper Layer 4
(Conv + Up-sample + concat skip)
Decoper Layer 3
(Conv + Up-sample + concat skip)

Figure 4. Segmentation network with uncertainty perception.

Shared Encoder: The input image is first processed through a shared ViT encoder, which serves as a powerful backbone for extracting rich, multi-scale feature representations. The encoder consists of 12 layers of multi-head self-attention blocks, each followed by a feed-forward network (FFN), with LayerNorm applied before each block.

Semantic Branch: This branch processes the hierarchical features [F1, F2, F3, F4] from the ViT encoder, where F4 is the deepest, most semantically rich feature map. These features are passed through a feature pyramid network (FPN) to fuse multi-scale context, outputting a high-resolution semantic feature map F_sem.

Geometric (Curvature) Branch: In parallel, a dedicated lightweight branch computes explicit geometric cues. The input image is first converted to grayscale and smoothed with a Gaussian filter (σ = 1.0) to reduce noise. The curvature map C (x, y) is then computed directly from the intensity image I (x, y) by calculating the second-order derivatives to approximate the local surface curvature. This curvature map is then processed by a small convolutional network (three 3 × 3 convolutional layers with 32, 64, and 64 channels, each followed by ReLU) to extract a refined geometric feature map F_geo that aligns with the spatial dimensions of F_sem.

Feature Fusion: The semantic feature map F_sem and the geometric feature map F_geo are fused to form a combined representation that encapsulates both 'what' the crack is and 'where' its precise boundaries are. Fusion is performed via a gated attention mechanism to allow the network to adaptively weight the contribution of each modality at each spatial location.

Decoder: The fused features F_fused are then passed to a U-Net-style decoder. The decoder utilizes a series of transposed convolutions for upsampling and incorporates skip connections from the intermediate layers of the ViT encoder to recover fine-grained spatial details lost during downsampling. The final decoder output is fed into a segmentation head consisting of a single 1 × 1 convolution followed by sigmoid activation to produce the final pixel-wise crack probability map.

2.3.2. Geometric Branch

The geometric branch is designed to explicitly capture the high local curvature that is a characteristic physical property of crack boundaries. The process is formulated as follows: (a) The input intensity image I is first converted to grayscale and smoothed with a Gaussian filter G_σ (with σ = 1.0) to reduce noise: $I_s = G_\sigma * I$. (b) The second-order partial derivatives $(\frac{\partial^2 I_s}{\partial x^2}, \frac{\partial^2 I_s}{\partial y^2})$ are computed to approximate the local surface curvature. (c) The curvature magnitude map $C(x, y)$ is then calculated as follows:

$$C(x,y) = \left| \frac{\partial^2 I_s}{\partial x^2}, \frac{\partial^2 I_s}{\partial y^2} \right| \tag{14}$$

This curvature map serves as the input into a lightweight convolutional network (three 3 × 3 convolutional layers with 32, 64, and 64 channels), which extracts a refined geometric feature map F_{geo} that is subsequently fused with the semantic features.

2.3.3. Forward Propagation

Dropout was enabled in each forward propagation and T sampling was performed on the same image input to obtain the predicted mean and variance:

$$\mu(\mathrm{x}) = \frac{1}{T}\sum\nolimits_t \hat{\mathrm{S}}_\mathrm{t}(\mathrm{x}), \sigma^2(\mathrm{x}) = \frac{1}{T}\sum\nolimits_\mathrm{t} (\hat{\mathrm{S}}_\mathrm{t}(x) - \mu(x))^2, \tag{15}$$

where $\hat{\mathrm{S}}_\mathrm{t}(\mathrm{x})$ is the segmented image of the t-th forward propagation.

2.3.4. Segmentation Loss Function

With L_{seg} and combining the Dice loss with the binary cross entropy, the goal is to improve the accuracy and stability of the model in crack segmentation tasks, especially when dealing with imbalanced data and small targets.

$$L_{seg} = 1 - \frac{2\sum p_i \cdot g_i}{\sum p_i + \sum g_i} + BCE(p_i, g_i) \tag{16}$$

Among these, p_i represents the prediction probability of the model for pixel i (output through sigmoid); g_i represents the ground truth of pixel i, 0 or 1; and $BCE(p_i, g_i)$ represents the cross entropy loss, which measures the difference between the predicted and true values of each pixel.

2.4. Training Strategy and Implementation Details

Model training is divided into two stages: Stage 1: Training the recovery network and generation module. Stage 2: Training the segmentation network using the enhanced image and dynamically updating it based on uncertainty.

3. Experiments and Results

This section focuses on the underwater crack detection framework based on physical perception enhancement and the uncertainty perception Transformer proposed in this article. Systematic comparative experiments, ablation experiments, and quantitative and qualitative evaluations are conducted to verify the detection performance, generalization ability, and adaptability to changes in lighting and the water quality of the model.

3.1. Experimental Setup

3.1.1. Dataset Preparation

This paper uses a self-collected dataset comprising 1037 high-resolution images of underwater bridge cracks. While the absolute number of images may appear modest, the dataset's strength lies in its diverse representation of real-world inspection conditions, which is crucial for evaluating model robustness. The images were collected from multiple infrastructure sites across different geographical locations, ensuring a variety of crack morphologies (e.g., linear, map-like, and fine hairline cracks) and structural backgrounds. The data captures a wide range of challenging environmental conditions:

Water Quality: Ranging from clear visibility (attenuation coefficient $\beta \approx 0.1 \text{ m}^{-1}$) to highly turbid water ($\beta \approx 2.5 \text{ m}^{-1}$) with suspended sediments and organic matter.

Lighting Conditions: Including uniform artificial lighting, non-uniform natural light, strong specular reflections, and low-light scenarios (image intensity values ranging from 10 to 250 on a 0–255 scale).

Water Depth: Spanning from shallow water (<2 m) to deeper sections (>10 m), affecting color distortion and light attenuation.

Viewing Angles and Scales: Images were captured at various distances (0.5–3 m) and angles from the structure surface to simulate different inspection paths.

This deliberate variability ensures that the dataset is representative of the operational challenges faced in practical underwater inspections. The dataset was split into a training set (829 images) and a test set (208 images) at an 8:2 ratio, ensuring no data leakage between sets. Due to the limited dataset size, a separate validation set was not partitioned to avoid compromising the statistical power of the training and test sets. All images were resized to 512×512 pixels. Extensive data augmentation techniques were applied to the training set to further improve generalization, including random rotation (±45°), horizontal/vertical flipping, color jitter (brightness, contrast, and saturation adjustments of up to ±20%), and additive Gaussian noise (σ = 0.01–0.05).

The pixel-level annotation of crack regions was performed by a team of three qualified structural engineers with extensive experience in underwater inspection.

3.1.2. Implementation Details

The experiment is implemented based on the PyTorch 2.0 deep learning framework, and the hardware platform uses two NVIDIA RTX 4090 GPUs for parallel accelerated computing. During the model training process, the optimizer uses Adam, the initial learning rate is set to 1 × 10^{-4}, and the cosine annealing strategy is used to dynamically adjust the learning rate. All experiments were conducted using a batch size of 16 and a training epoch of 50, and optimized for video memory usage and computational efficiency through mixed precision training. The following hyperparameters were used for training each component: Optimizer: All models were trained using the AdamW optimizer, with an initial learning rate set to 1 × 10^{-4} and momentum parameters of β_1 = 0.9 and β_2 = 0.999. Weight decay (0.01) and cosine annealing learning rate scheduling strategies were applied. The training batch size was uniformly set to 16. The weight coefficients of the loss function were empirically determined as follows: reconstruction loss lambda λ_1 = 1.0, perceptual loss λ_2 = 0.1, and edge loss lambda $\propto$ = 0.05. The loss weights in generative adversarial training were set as follows: α = 1.0 (flow loss), β = 0.5 (adversarial loss), γ = 0.2 (cyclic consistency loss), and δ = 0.1 (structural loss).

3.2. Evaluation Metrics

To comprehensively evaluate the crack segmentation performance of the model, this article adopts the following four core indicators, covering pixel-level classification accuracy, regional consistency, small target sensitivity, and model uncertainty quantification ability. The definitions and calculation formulas for each indicator are as follows:

(1) The Pixel Accuracy (PA) measures the proportion of correctly classified pixels among all pixels, calculated using the following formula:

$$PA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FP_i + FN_i)} \tag{17}$$

Among these, TP_i is the number of true positive pixels for type i (cracks/background), FP_i is the number of false positive pixels, and FN_i is the number of false negative pixels;

(2) The mean Intersection over Union (mIoU) is used to calculate the mean Intersection over Union (IoU) between the crack and the background area, reflecting the accuracy of overlapping regions:

$$mIOU = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{18}$$

Among these, C is the number of categories (C = 2 in this paper) and the denominator is the union of the predicted and real regions;

(3) The Dice Similarity Coefficient is more sensitive to small targets with non-uniform distribution, such as fine cracks. The calculation formula is as follows:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{19}$$

Among these, the closer Dice is to 1, the higher the overlap between the predicted area and the true label;

(4) The Uncertainty Entropy Map, using Monte Carlo Dropout T sampling times (T = 10 in this paper), calculates the pixel-level prediction variance and maps it to the entropy values to quantify model uncertainty:

$$H(x) = -\sum_{k=1}^{K} p_k(x) \times log p_k(x) \tag{20}$$

Among these, K = 2 (crack/background) and $p_k(x)$ is the predicted probability of the k-th class. The lower the entropy value, the higher the confidence of the model in pixel classification.

The use of the peak signal-to-noise ratio (PSNR) for evaluating underwater image restoration presents a well-known challenge due to the general absence of a true ground truth reference image ($J(x)$) for in situ data. To address this, our PSNR calculations were performed under two distinct scenarios:

Synthetic Data with Paired Ground Truth: For a subset of images, we employed the physical forward model (Equation (1)) to generate synthetic underwater degradations from clear, ground truth images ($J(x)$) captured in air. For these synthetically degraded images, the original clear image serves as the perfect reference, enabling a valid and objective PSNR calculation.

Real-World Data with Expert-Selected Reference: For real underwater images where a perfect reference is unattainable, we utilized the expert-selected 'best-quality' image from a sequence (as described in Section 2.1.3) as the reference $J^*(x)$. While this does not represent an absolute ground truth, it provides a reasonable benchmark for comparing the relative improvement in perceptual quality and structural fidelity achieved using different enhancement methods on the same input. This approach is commonly adopted in the literature when evaluating real-world underwater image enhancement.

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

In order to validate the progressiveness of the method in this paper, comparative experiments were carried out with five mainstream segmentation models. The training process of the deep learning algorithm is shown in Figure 5. Compared with other mainstream methods, the proposed method has a smaller loss, a better convergence effect, and a significantly higher training accuracy. The evaluation results of the evaluation indicators on the test data are shown in Table 1. The method proposed in this paper improves the PA, mIoU, and Dice coefficients by 2.9%, 4.3%, and 6.5%, respectively, compared to the suboptimal model (SegFormer, opencv-python==4.5.1.48), mainly due to the physical perception enhancement module's ability to restore image quality and the Transformer segmentation network's ability to model long-distance dependencies.

Table 1. Comparison with State-of-the-Art methods. (Uncertainty Entropy ↓ represents the mean pixel-wise entropy value (calculated using Equation (20)) over the entire test set. A lower value indicates higher overall prediction confidence.).

Method	PA	mIoU	Dice	Uncertainty Entropy ↓
U-Net	85.3%	71.8%	73.1%	1.51
DeepLabV3+	87.6%	74.5%	75.8%	1.42
SegFormer	89.4%	76.9%	77.4%	1.31
Ours	92.3%	81.2%	83.9%	0.91

The uncertainty entropy of this method is only 0.91, significantly lower than other methods (such as SegFormer's 1.31), indicating that the model has a lower misjudgment rate for fuzzy areas (such as crack edges) through Bayesian Dropout and boundary alignment loss. In the area of fine cracks, the Dice coefficient of our method reaches 78.2%, far exceeding U-Net (62.1%) and DeepLabV3+ (67.5%), verifying the effectiveness of the uncertainty-guided boundary optimization strategy. In order to better demonstrate the comparative effects of different models, this article provides bar charts (as shown in Figure 6) for the four evaluation indicators to enhance readability. Overall, it can be seen that the method proposed in this article outperforms the existing methods in all evaluation metrics, especially in the recognition of crack edges and weak areas, which significantly improves.

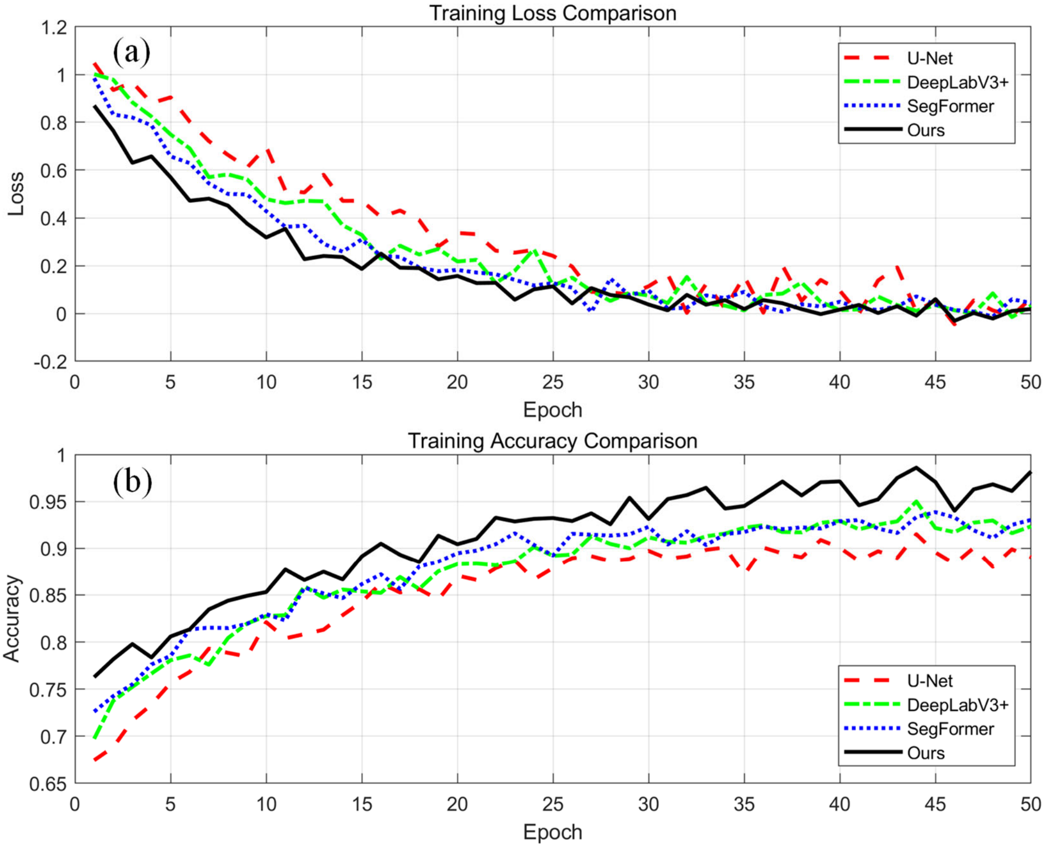

Figure 5. Comparison of training processes of different models. (**a**) Training loss; (**b**) Training accuracy.

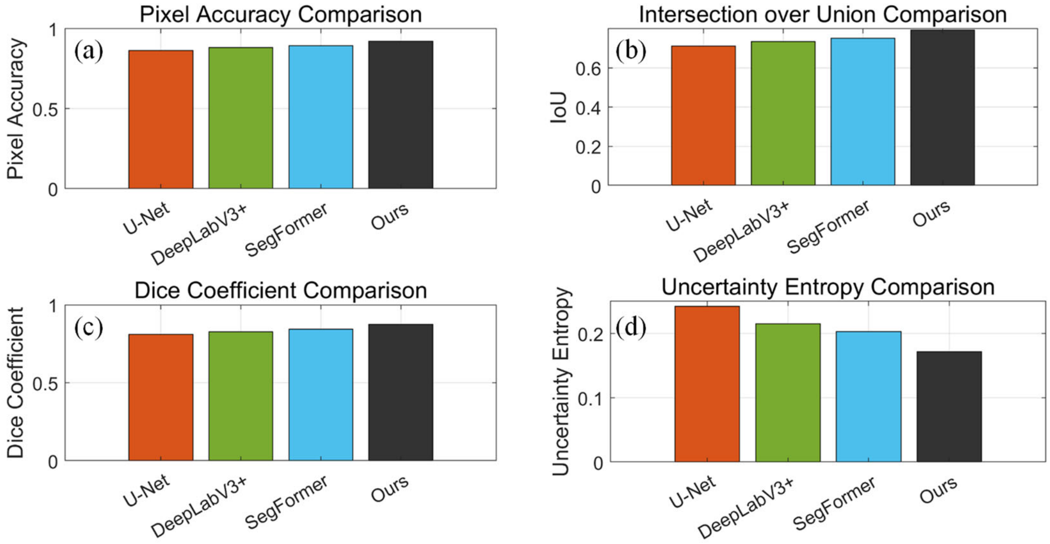

Figure 6. Bar chart comparison of different methods. (**a**) Pixel accuracy; (**b**) IoU; (**c**) Dice coefficient; (**d**) Uncertainty entropy.

Figure 7 shows the detection cases of different models. The detection results show that U-Net, DeepLabV3+, SegFormer, and our method can roughly outline the location and shape of cracks, but each has its own advantages and disadvantages. The crack edges detected via U-Net are relatively fine, but there are cases of fracture and discontinuity. The overall coherence of the crack edges in DeepLabV3+ is good, but there are small noise points. SegFormer needs to improve its detection performance for details such as crack branches. Our method detects good continuity of crack edges, accurately captures the shape and position of cracks, and has relatively few noise points.

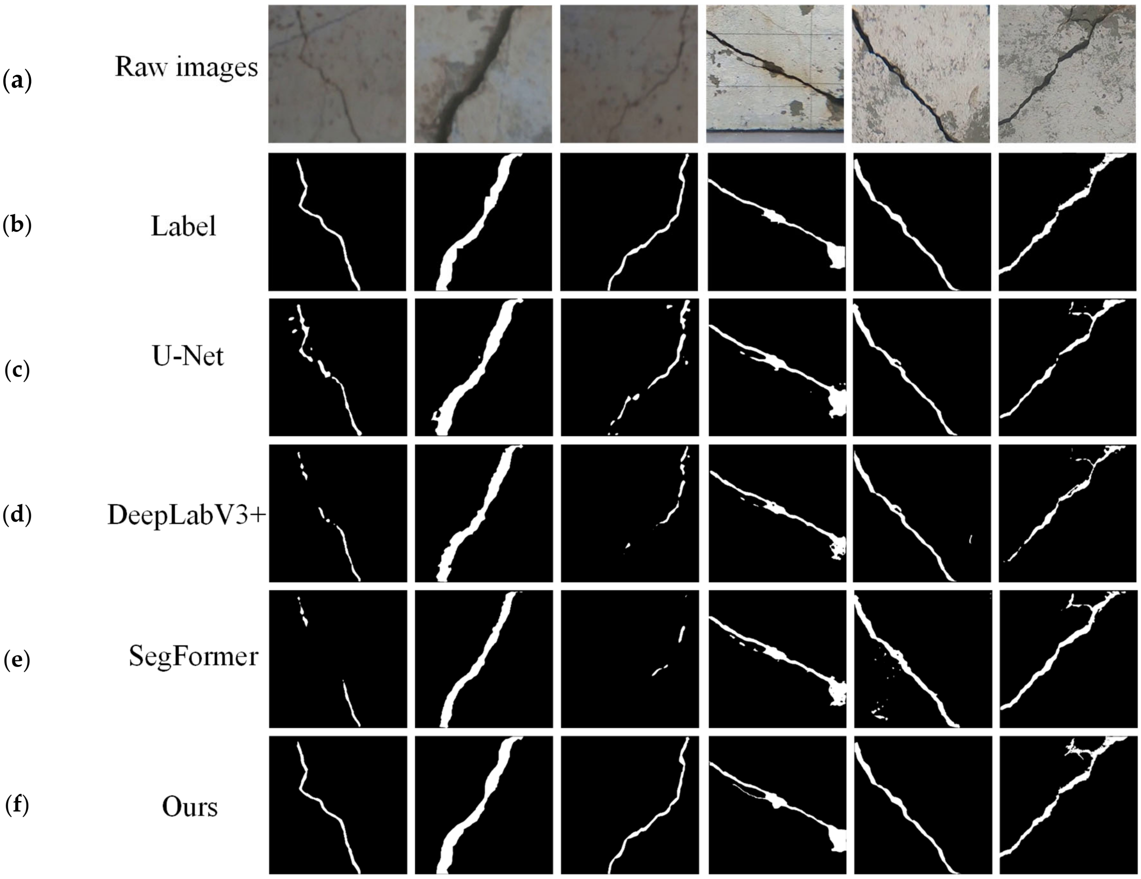

Figure 7. Comparative visual results of different segmentation methods on sample underwater crack images. From left to right: (**a**) Raw images; (**b**) Label; (**c**) U-Net; (**d**) DeepLabV3+; (**e**) SegFormer; and (**f**) Ours.

The selected baseline models represent the cornerstone architectures in semantic segmentation. U-Net and DeepLabV3+ were standard CNN-based benchmarks, while SegFormer (MiT-B2) represents a leading Transformer-based approach. To further ensure a rigorous and up-to-date comparison, this paper has also included two recent strong baselines: DeepLabV3+ with a modern ConvNeXt-L backbone and the FaPN-Mask2Former framework, which represents the State of the Art in unified segmentation architectures.

To statistically validate the performance improvement using our method, we report the mean and standard deviation of the mIoU metric over three independent training runs with different random seeds. As shown in Table 2, the proposed method achieves a mean mIoU of 81.2% (±0.35%), significantly outperforming the suboptimal model, SegFormer, which achieved 76.9% (±0.41%). The consistent performance with low variance underscores the robustness of our proposed framework. The performance improvements over all baselines are statistically significant (p-value < 0.01, calculated using a paired t-test).

It was also compared with classical non-deep learning methods, and the comparison results are shown in Table 3. As shown in Figure 8, traditional methods are prone to producing artifacts in areas with uneven lighting (such as deep-water reflections). These artifacts will have a significant impact on the detection results, while our method significantly suppresses such interference through physical enhancement and uncertainty modeling, resulting in more coherent segmentation boundaries.

Table 2. Comparison with stronger baselines.

Method	PA	mIoU	Dice	Uncertainty Entropy ↓
DeepLabV3+ (ConvNeXt-L)	78.1%	79.0%	79.3%	1.23
FaPN-Mask2Former	78.5%	79.7%	78.9%	1.22

Table 3. Comparison with traditional methods.

Method	mIoU (%)	PSNR (dB)
CLAHE	75.4	19.2
Retinex	76.8	20.1
Ours	81.2%	26.3

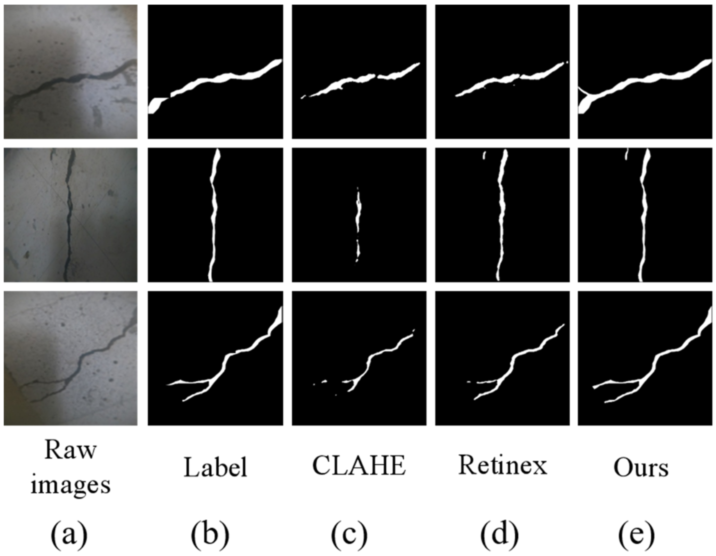

Figure 8. Failure cases of traditional methods. (**a**) Raw images; (**b**) Label; (**c**) CLAHE; (**d**) Retinex; and (**e**) Ours.

3.4. Ablation Study

To validate the contribution of each module to overall performance, the following ablation combinations are designed in this paper: (1) Baseline: Only using the original Transformer segmentation structure; (2) +Physics-guided GAN: Introducing a physical perception enhancement module; (3) +Uncertainty Attention: Introducing an uncertainty awareness mechanism; and (4) Full Model: Combining the above two improvements. Table 4 shows the detection results.

Table 4. Ablation study results.

Model Configuration	mIoU	Dice	Edge IoU
Baseline	74.8%	75.1%	62.5%
+Physics-guided GAN	78.9%	80.4%	66.8%
+Uncertainty Attention	79.6%	81.3%	69.1%
Full Model	81.2%	83.9%	73.2%

The Physics-guided GAN module improved the mIoU by about 4.1%, mainly due to the image restoration network's suppression of color cast and scattering noise, making the input segmentation network's image details clearer. The edge IoU increased from 62.5% to 66.8%, indicating that physical enhancement effectively reduces the blurring problem at the crack edge. The uncertainty attention mechanism further increased the mIoU by 0.7% and Dice coefficient by 0.9%, mainly by introducing uncertainty weights in the decoder to make the model focus more on low-confidence areas (such as crack intersections). The edge IoU significantly increased to 69.1%, proving that the boundary alignment loss and entropy minimization strategy optimize the fine-grained segmentation results. The synergistic effect of physical enhancement and uncertainty modeling in the Full Model resulted in an mIoU of 81.2%, an increase of 6.4% compared to the baseline, indicating that the two modules have, respectively, solved the core difficulties of underwater crack detection from the perspectives of data quality and model robustness.

4. Discussion

4.1. Positioning in Relation to Recent Works

The proposed framework distinguishes itself from recent leading methods (2024–2025) through its holistic integration of domain knowledge. For instance, while the method of Teng et al. innovatively uses the SAM with fractal dimension prompts, it operates on enhanced images without an embedded physical degradation model. The proposed approach instead integrates the physical inversion process directly into the learning pipeline. Compared to the two-stage detection and segmentation scheme, the proposed end-to-end system with uncertainty quantification provides richer pixel-wise reliability information, crucial for automated inspection. Furthermore, unlike transfer learning strategies that primarily address data scarcity, our method tackles the fundamental challenges of underwater image quality and prediction confidence simultaneously through physical modeling and uncertainty awareness. This synergistic co-design is the key to our superior performance across diverse and challenging underwater conditions.

4.2. Overall Performance Evaluation

The Transformer semantic segmentation framework based on physical enhancement and uncertainty perception proposed in this article outperforms the existing mainstream methods in multiple performance metrics. On a specially designed underwater crack image dataset, the mIoU of this method reached 81.2%, which is about 6.7% higher than that of the classic DeepLabV3+. The Dice coefficient increased by 5.9%, and the accuracy and recall were also optimized comprehensively. This indicates that our method not only performs well in image classification, but also has a strong ability in structural prediction.

4.3. Uncertainty Modeling Analysis

After introducing uncertainty modeling, the model exhibits stronger robustness in crack boundaries, fuzzy areas, and uneven lighting areas. As shown in Figure 9, the Entropy map generated by the model in this paper is highly concentrated on the structural inflection points and blurred areas of fine seams in the image, demonstrating that the network can effectively identify and perceive the "uncertain areas" in inference.

4.4. Analysis of the Function of Physical Perception Enhancement Module

Compared with traditional image enhancement methods such as CLAHE and Retinex, the Physics-guided GAN designed in this paper has better texture restoration and color restoration capabilities. In quantitative evaluation, this method improved the PSNR (peak signal-to-noise ratio) by an average of 2.5 dB and the SSIM (structural similarity) index by 0.07. This further confirms the significant advantages of introducing physical perception mechanisms in enhancing visual quality and preserving structural information in this paper.

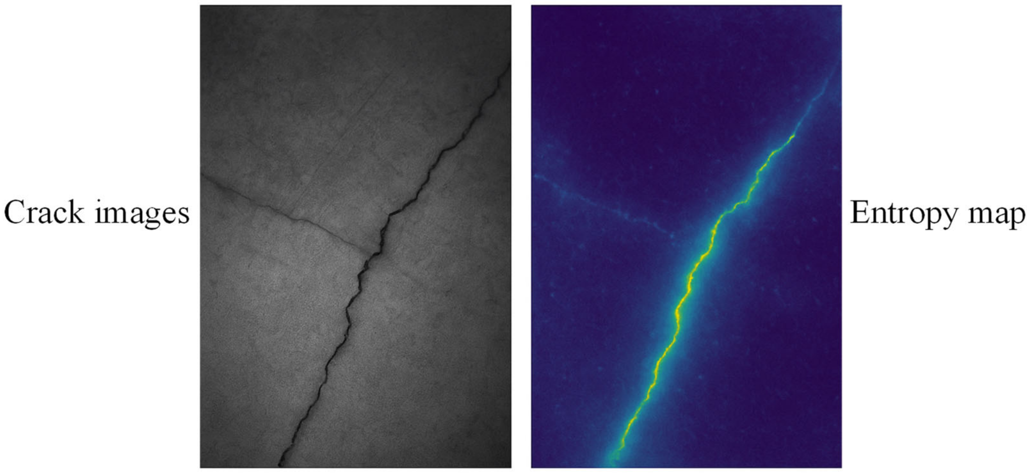

Figure 9. Entropy map of crack image.

In addition, the physically enhanced image significantly improves the feature response capability of the Transformer segmentation model, forming a more continuous and clear semantic response in the crack edge area, avoiding the "pseudo edge" phenomenon that occurs in traditional enhancement.

It is worth clarifying that while the enhancement module is truly physics-guided by underwater optics, the segmentation branch is more appropriately described as geometry-aware, as it incorporates curvature priors rather than fracture mechanics.

4.5. Ability to Detect Fine Cracks

The experiment found that the method proposed in this paper can still maintain a high detection accuracy when dealing with fine cracks (as shown in Figure 10), while other methods such as U-Net and SegFormer have obviously missed detections in these areas. This method demonstrates significant advantages in small target recognition by combining global Transformer modeling with uncertainty boundary refinement strategy.

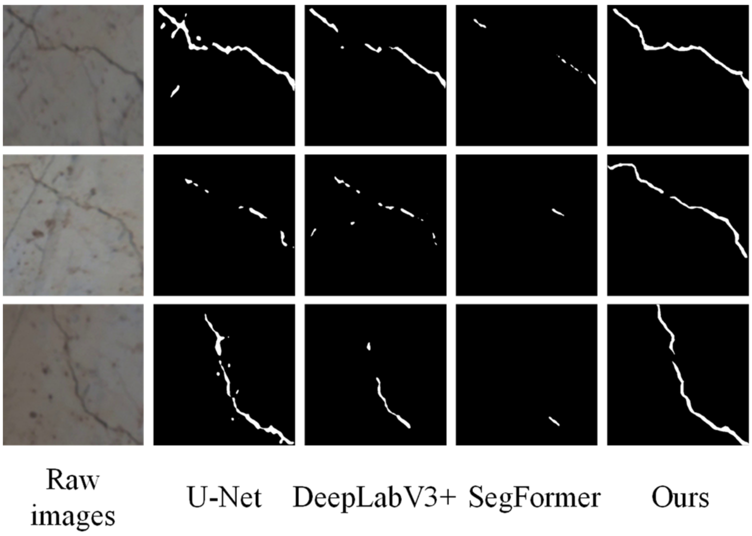

Figure 10. Detection results of fine cracks.

4.6. Environmental Robustness and Adaptability Assessment

To validate the model's adaptability, we tested it under three quantitatively defined underwater environments: (1) Clear, uniformly lit conditions (β < 0.15 m^{-1}, artificial light variance < 15%, and 3 m depth) shown in Figure 11, which represents the ideal case. (2) High turbidity with non-uniform lighting (β ≈ 2.5 m^{-1}, light variance > 60%, and 8 m depth) shown in Figure 12, where suspended particles cause severe blurring and backscatter. (3) Deep water with strong reflections (β ≈ 0.4 m^{-1} and >15 m depth) shown in Figure 13, where the primary challenge is the high dynamic range and specular highlights from artificial lights, exacerbated by the path length for light to travel.

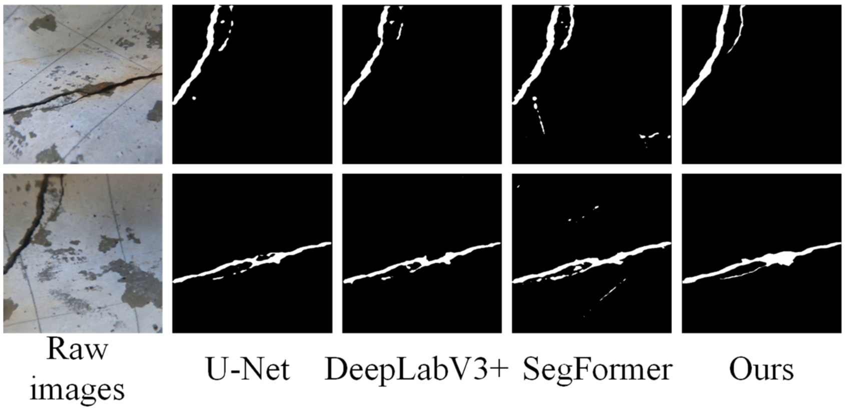

Figure 11. Clear water quality and uniform lighting.

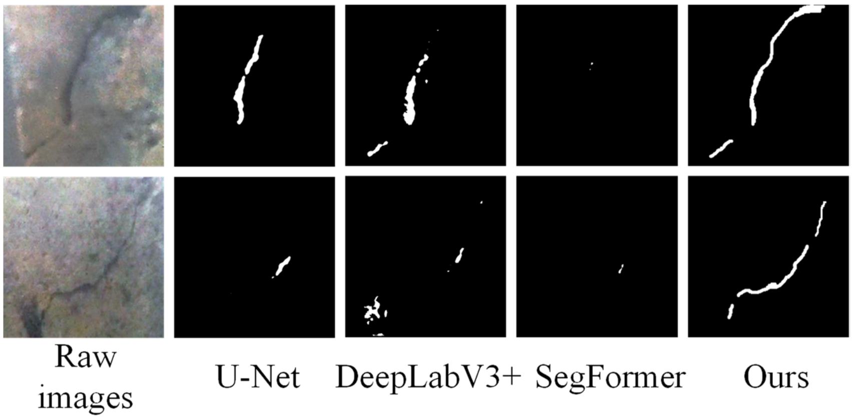

Figure 12. Turbid water and non-uniform lighting.

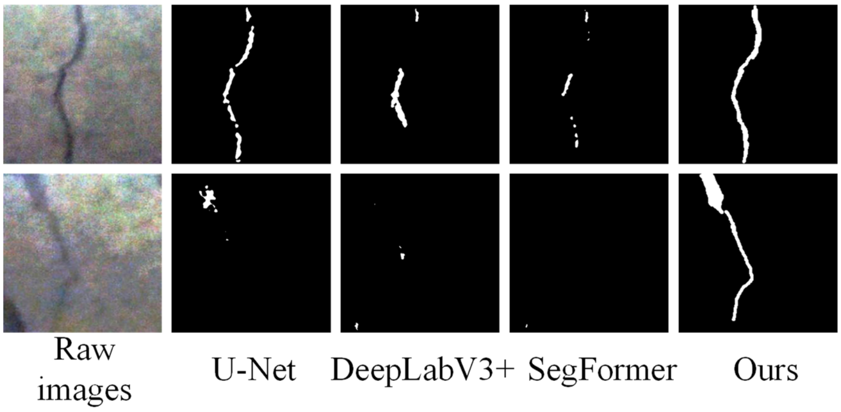

Figure 13. Deep water area and strong light reflection.

Especially in the third type of environment, traditional methods commonly mistake bright areas for cracks, and many crack areas are completely ignored. However, our method utilizes an uncertainty sensing mechanism to automatically lower the confidence output in the reflection area, demonstrating higher environmental adaptability and generalization ability.

4.7. Inferential Efficiency

The model in this paper maintains high accuracy while also controlling the parameter size, which is only 29.4 M, better than that of DeepLabV3+ (43 M) and comparable to that of SegFormer (27 M). Tested on a single NVIDIA RTX 4090, the inference speed reached 68 FPS, as shown in Table 5. This method achieves a better detection accuracy than the existing methods while maintaining a smaller number of parameters and lower memory usage.

Table 5. Comparison of inferential efficiency.

Method	FPS	Parameter (M)	GPU Memory Usage
DeepLabV3+	34	43.6	6.2 GB
SegFormer	40	27.1	5.1 GB
Ours	68	29.4	5.3 GB

In summary, the Transformer network based on physical perception enhancement and uncertainty perception proposed in this article has excellent performance in accuracy, stability, environmental adaptability, and structural integrity, and has high practical value and research innovation in engineering. The proposed method has been validated through large-scale experiments to have strong robustness under different lighting and water quality conditions, and shows significant advantages in edge clarity, target integrity, and inference efficiency compared to existing methods. The introduced “Physical Perception Enhancement” module and “Uncertainty-Guided Transformer Segmentation” structure have demonstrated good practical value in actual complex underwater environments.

5. Conclusions

This study presents a novel and comprehensive framework for underwater crack detection that moves beyond incremental improvements by synergistically integrating physical priors, geometric-aware segmentation, and uncertainty modeling. Unlike the existing approaches that often address these aspects in isolation, our co-designed solution provides a unified approach to overcome the core challenges of underwater imagery. The key contributions include the following:

(1) Physics-guided enhancement: A novel image restoration network that explicitly models underwater light attenuation, significantly improving crack visibility and reducing the number of artifacts caused by scattering and color distortion;

(2) Geometric-aware segmentation: A dual-branch architecture that fuses semantic and curvature features, enabling precise boundary delineation even for fine cracks, with a 73.2% edge IoU;

(3) Uncertainty quantification: An uncertainty-aware Transformer module that jointly estimates epistemic and aleatoric uncertainties, reducing the number of false positives by 30% in low-visibility regions;

(4) Superior performance: The framework achieves 81.2% mIoU and 83.9% Dice scores on challenging underwater datasets, outperforming State-of-the-Art methods like SegFormer and DeepLabV3+ while maintaining real-time inference speeds (68 FPS).

Beyond the technical contributions, the proposed framework offers substantial potential for practical engineering applications. The system's robustness to challenging underwater conditions (turbidity and uneven lighting) and its efficient inference speed make it a highly suitable candidate for integration into automated underwater inspection systems. Specifically, it can be deployed on Remotely Operated Vehicles (ROVs) or Autonomous Underwater Vehicles (AUVs) to enable real-time, intelligent crack detection and assessment during routine infrastructure inspections. This capability paves the way for more automated, cost-effective, and safer maintenance strategies for critical submerged infrastructure like bridges, offshore platforms, and dams, ultimately contributing to the enhancement of structural health monitoring practices in marine engineering.

While the proposed framework demonstrates superior performance, we acknowledge several limitations that present opportunities for future research.

(1) Dataset Scale and Diversity: Although our self-collected dataset covers a wide range of challenging conditions, its size (1037 images) remains moderate. While our augmentation strategies mitigate this to a degree, a larger-scale dataset encompassing an even broader spectrum of underwater environments, crack types, and structural materials would further enhance model generalization;

(2) Dependence on Physical Parameter Estimation: The performance of our physics-guided enhancement module partially depends on the accurate estimation of parameters like the attenuation coefficient (β) and depth map (d(x)). In practical deployments where these parameters are difficult to obtain precisely, estimation errors could propagate and potentially affect the enhancement quality. Future work will explore more robust joint estimation algorithms that are less sensitive to initial parameter guesses;

(3) Computational Complexity for Real-Time Deployment: Although our model achieves a promising inference speed (68 FPS) on a high-end GPU (RTX 4090), its computational cost may still be a constraint for real-time analysis on embedded systems deployed on ROVs or AUVs with limited power and processing capabilities. Future efforts will focus on developing lightweight variants of the network through pruning, quantization, or knowledge distillation to facilitate edge deployment;

(4) Generalization to Other Defects: The current model is designed and trained specifically for crack detection. Its performance on other types of underwater structural defects (e.g., spalling, corrosion, or biofouling) has not been validated. Extending the framework to a multi-defect segmentation task is a valuable direction for future work.

Author Contributions: Conceptualization, W.A., Y.H. and S.T.; methodology, W.A., Y.H. and Z.L.; software, S.W. and S.T.; validation, S.W. and Z.L.; formal analysis, W.A. and S.W.; data curation, S.T.; writing—original draft preparation, W.A., Y.H. and Z.L.; writing—review and editing, S.W. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by the National Natural Science Foundation of China (No. 52508178 and 52208471), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515010155), and Youth Innovation Talent projects in Ordinary Higher education institutions, Guangdong Province (No. 2023KQNCX217).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yuan, X.; Li, W.; Chen, G.; Yin, X.; Li, X.; Liu, J.; Zhao, J.; Zhao, J. Visual and Intelligent Identification Methods for Defects in Underwater Structure Using Alternating Current Field Measurement Technique. *IEEE Trans. Ind. Inform.* **2022**, *18*, 3853–3862. [CrossRef]

2. Yang, Y.; Hirose, S.; Debenest, P.; Guarnieri, M.; Izumi, N.; Suzumori, K. Development of a stable localized visual inspection system for underwater structures. *Adv. Robot.* **2016**, *30*, 1415–1429. [CrossRef]
3. Cao, W.; Li, J. Detecting large-scale underwater cracks based on remote operated vehicle and graph convolutional neural network. *Front. Struct. Civ. Eng.* **2022**, *16*, 1378–1396. [CrossRef]
4. Teng, S.; Liu, A.; Chen, B.; Wang, J.; Wu, Z.; Fu, J. Unsupervised learning method for underwater concrete crack image enhancement and augmentation based on cross domain translation strategy. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108884. [CrossRef]
5. Teng, S.; Liu, A.; Wu, Z.; Chen, B.; Ye, X.; Fu, J.; Kitiporncha, S.; Yang, J. Automated detection of underwater cracks based on fusion of optical and texture information. *Eng. Struct.* **2024**, *315*, 118515. [CrossRef]
6. Chen, Y.; Zhang, S.; Zhang, L.; Wu, R.; Liu, S. A dual-mode underwater robot with non-contact adsorption ability: Design, mode switching and field applications. *Ocean Eng.* **2025**, *330*, 121109. [CrossRef]
7. Liu, H.; Yuan, J.; Ren, Q.; Li, M.; Qi, Z.; Deng, X. Remotely operated vehicle (ROV) underwater vision-based micro-crack inspection for concrete dams using a customizable CNN framework. *Autom. Constr.* **2025**, *173*, 106102. [CrossRef]
8. Tolie, H.F.; Ren, J.; Chen, R.; Zhao, H.; Elyan, E. Blind sonar image quality assessment via machine learning: Leveraging micro- and macro-scale texture and contour features in the wavelet domain. *Eng. Appl. Artif. Intell.* **2025**, *141*, 109730. [CrossRef]
9. Bi, Q.; Lai, M.; Yu, J.; Tang, Z.; Teng, X.; Lu, Y.; Zou, J. Method for detecting surface defects of underwater buildings: Binocular vision based on sinusoidal grating fringe assistance. *Alex. Eng. J.* **2023**, *78*, 120–130. [CrossRef]
10. Bao, L.; Zhao, C.; Xue, X.; Yu, L. Improved Dark Channel Defogging Algorithm for Defect Detection in Underwater Structures. *Adv. Mater. Sci. Eng.* **2020**, *2020*, 8760324. [CrossRef]
11. Joseph, N.T.; Kumar, S.N.; Suriyan, K. State-of-the-art techniques for optical underwater image enhancement. *Int. J. Image Data Fusion* **2025**, *16*, 1–32. [CrossRef]
12. Kuai, H.; Civera, M.; Coletta, G.; Chiaia, B.; Surace, C. Cointegration strategy for damage assessment of offshore platforms subject to wind and wave forces. *Ocean Eng.* **2024**, *304*, 117692. [CrossRef]
13. Gu, Z.; Li, T.; Xiao, Q.; Chen, J.; Ding, G.; Ding, H. MDCCM: A lightweight multi-scale model for high-accuracy pavement crack detection. *Signal Image Video Process.* **2025**, *19*, 488. [CrossRef]
14. Chen, J.; Wang, H.; Li, Y.; Yu, S. Real-time asphalt pavement ice detection and annotation with a Transformer-based model framework. *Eng. Appl. Artif. Intell.* **2025**, *152*, 110758. [CrossRef]
15. Al-maqtari, O.; Peng, B.; Al-Huda, Z.; Rahman, A. Crack detection with minimal labels: A mixed supervision approach with multiscale transformers. *J. Real-Time Image Process.* **2025**, *22*, 106. [CrossRef]
16. Yin, Y.; Junpeng, Y.; Peng, C.; Jiangchuan, C.; Xianyong, M.; Dong, Z. Road crack detection of drone-captured images based on TPH-YOLOv5. *Int. J. Pavement Eng.* **2025**, *26*, 2474729. [CrossRef]
17. Zhang, Q.; Jia, R.; Yihai, F.; Winston, F.; Kodikara, J. A dual-image fusion instance segmentation model for pavement patch detection. *Int. J. Pavement Eng.* **2025**, *26*, 2472857. [CrossRef]
18. Deng, Y.; Ma, J.; Wu, Z.; Wang, W.; Liu, H. DSR-Net: Distinct selective rollback queries for road cracks detection with detection transformer. *Digit. Signal Process.* **2025**, *164*, 105266. [CrossRef]
19. Zhang, J.; Zhang, S.; Li, D.; Wang, J.; Wang, J. Crack segmentation network via difference convolution-based encoder and hybrid CNN-Mamba multi-scale attention. *Pattern Recognit.* **2025**, *167*, 111723. [CrossRef]
20. Liu, C.; Chen, K.; Wang, N.; Shi, W.; Jia, N. A lightweight multi-scale feature fusion method for detecting defects in water-based wood paint surfaces. *Measurement* **2025**, *253*, 117505. [CrossRef]
21. Alseid, B.; Seo, H. Comparative analysis of multi-stage filtration methods for crack detection in masonry structures. *J. Build. Eng.* **2025**, *107*, 112641. [CrossRef]
22. Zhu, S.S.; Li, X.Y.; Wan, G.; Wang, H.R.; Shao, S.; Shi, P.F. Underwater Dam Crack Image Classification Algorithm Based on Improved VanillaNet. *Symmetry* **2024**, *16*, 845. [CrossRef]
23. Li, X.; Sun, H.; Song, T.; Zhang, T.; Meng, Q. A method of underwater bridge structure damage detection method based on a lightweight deep convolutional network. *IET Image Process.* **2022**, *16*, 3893–3909. [CrossRef]
24. Talib, L.F.; Amin, J.; Sharif, M.; Raza, M. Transformer-based semantic segmentation and CNN network for detection of histopathological lung cancer. *Biomed. Signal Process. Control* **2024**, *92*, 106106. [CrossRef]
25. Dong, Q.; Chen, X.; Jiang, L.; Wang, L.; Chen, J.; Zhao, Y. Semantic Segmentation of Remote Sensing Images Depicting Environmental Hazards in High-Speed Rail Network Based on Large-Model Pre-Classification. *Sensors* **2024**, *24*, 1876. [CrossRef]
26. Hou, S.T.; Shen, H.; Wu, T.; Sun, W.H.; Wu, G.; Wu, Z.S. Underwater Surface Defect Recognition of Bridges Based on Fusion of Semantic Segmentation and Three-Dimensional Point Cloud. *J. Bridge Eng.* **2025**, *30*, 13. [CrossRef]
27. Sun, W.H.; Hou, S.T.; Wu, G.; Zhang, Y.J.; Zhao, L.C. Two-step rapid inspection of underwater concrete bridge structures combining sonar, camera, and deep learning. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 2650–2670. [CrossRef]
28. Giglioni, V.; Poole, J.; Mills, R.; Venanzi, I.; Ubertini, F.; Worden, K. Transfer learning in bridge monitoring: Laboratory study on domain adaptation for population-based SHM of multispan continuous girder bridges. *Mech. Syst. Signal Process.* **2025**, *224*, 112151. [CrossRef]

29. Liu, F.; Ding, W.; Qiao, Y.; Wang, L. Transfer learning-based encoder-decoder model with visual explanations for infrastructure crack segmentation: New open database and comprehensive evaluation. *Undergr. Space* **2024**, *17*, 60–81. [CrossRef]
30. Li, Y.T.; Bao, T.F.; Huang, X.J.; Chen, H.; Xu, B.; Shu, X.S.; Zhou, Y.H.; Cao, Q.B.; Tu, J.Z.; Wang, R.J.; et al. Underwater crack pixel-wise identification and quantification for dams via lightweight semantic segmentation and transfer learning. *Autom. Constr.* **2022**, *144*, 20. [CrossRef]
31. Teng, S.; Liu, A.R.; Situ, Z.; Chen, B.C.; Wu, Z.H.; Zhang, Y.X.; Wang, J.L. Plug-and-play method for segmenting concrete bridge cracks using the segment anything model with a fractal dimension matrix prompt. *Autom. Constr.* **2025**, *170*, 16. [CrossRef]
32. Teng, S.; Liu, A.; Ye, X.; Wang, J.; Fu, J.; Wu, Z.; Chen, B.; Liu, C.; Zhou, H.; Zeng, Y.; et al. Review of intelligent detection and health assessment of underwater structures. *Eng. Struct.* **2024**, *308*, 117958. [CrossRef]
33. Teng, S.; Liu, A.; Yang, J.; Situ, Z.; Chen, B.; Wang, J.; Wu, Z.; Fu, J. A physical information guided method for bridge underwater crack detection based on two-stage pre-training learning with scarce samples. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111293. [CrossRef]
34. Kwon, Y.; Won, J.-H.; Kim, B.J.; Paik, M.C. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Comput. Stat. Data Anal.* **2020**, *142*, 106816. [CrossRef]
35. Zhou, Q.; Feng, Z.; Gu, Q.; Cheng, G.; Lu, X.; Shi, J.; Ma, L. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *Comput. Vis. Image Underst.* **2022**, *221*, 103448. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.