Robotica (2025), **43**, pp. 2480–2503
doi:10.1017/S0263574725101835

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

MVFD-Net: multi-view fusion detection network for occluded underwater dam cracks

Yukai Wu[1], Xiaochen Qin[1] and Lei Cai[2]

[1]School of Mechanical and Electrical Engineering, Henan University of Technology, ZHengzhou, PR China
[2]School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang, PR China
Corresponding author: Lei Cai; Email: cailei2014@126.com

Received: 16 December 2024; **Revised:** 16 May 2025; **Accepted:** 20 May 2025; **First published online:** 24 June 2025

Keywords: semantic segmentation; multi-view fusion network; feature nonuniform scattering; noise reconstruction and fusion; gated adaptive fusion

Abstract

Detecting cracks in underwater dams is crucial for ensuring the quality and safety of the dam. However, underwater dam cracks are easily obscured by aquatic plants. Traditional single-view visual inspection methods cannot effectively extract the feature information of the occluded cracks, while multi-view crack images can extract the occluded target features through feature fusion. At the same time, underwater turbulence leads to nonuniform diffusion of suspended sediments, resulting in nonuniform flooding of image feature noise from multiple viewpoints affecting the fusion effect. To address these issues, this paper proposes a multi-view fusion network (MVFD-Net) for crack detection in occluded underwater dams. First, we propose a feature reconstruction interaction encoder (FRI-Encoder), which interacts the multi-scale local features extracted by the convolutional neural network with the global features extracted by the transformer encoder and performs the feature reconstruction at the end of the encoder to enhance the feature extraction capability and at the same time in order to suppress the interference of the nonuniform scattering noise. Subsequently, a multi-scale gated adaptive fusion module is introduced between the encoder and the decoder for feature gated fusion, which further complements and recovers the noise flooding detail information. Additionally, this paper designs a multi-view feature fusion module to fuse multi-view image features to restore the occluded crack features and achieve the detection of occluded cracks. Through extensive experimental evaluations, the MVFD-Net algorithm achieves excellent performance when compared with current mainstream algorithms.

1. Introduction

Dams are prone to cracking due to prolonged immersion, temperature fluctuations, water chemical corrosion and hydraulic fracturing [1, 2]. The number of cracks increased as the dam was used for longer periods of time. Some of the cracks may extend into the interior of the embankment dam, which affects the dam's structure and load-carrying capacity. Thus, underwater dam crack detection is crucial to ensure the proper functioning and safety of the dam structure.

Dam cracks can be obscured by aquatic vegetation during crack detection in the dam. The traditional single-view image cannot completely capture the characteristic information of the crack, and the use of multi-view image for crack detection can effectively avoid the problem of repeated occlusion in the same area of the crack. The feature information of the occluded region can be complemented by fusing the multi-view images. In recent years, numerous methods for enhancing underwater feature extraction capabilities in underwater object detection have been proposed [3]. Convolutional neural networks (CNNs) are widely used in feature recognition, but they have limitations in dealing with long-range dependencies. Transformer networks can effectively capture global dependencies through positional encoding or attention mechanisms [4, 5]. However, the complexity of the transformer computation increases significantly, especially when computing inter-positional correlations, and the computational resources required grow

exponentially. This leads to the fact that most transformer methods can only run on high-performance servers [6]. The use of gating mechanisms [7] for the problem of too much information can alleviate this problem. However, in order to balance segmentation accuracy and computational complexity, many studies have attempted to combine CNNs with transformers, achieving promising results [8]. Also this method fuses the global information of coarse granularity with the local information of fine granularity facilitates the network to capture features of different sizes [9]. Underwater turbulence leads to an uneven distribution of the suspended sediments, which leads to varying levels of scattered noise in the images. This results in different levels of feature information being obscured from different perspectives. Such differential noise not only affects the fusion of homogeneous feature information in multi-view images but also the completeness and accuracy of crack detection.

To address these challenges, this study introduces a novel multi-view fusion network (MVFD-Net) designed for crack detection in underwater dams. The primary innovations of this network are encapsulated in the following three aspects:

1. The FRI-Encoder is introduced, which facilitates interaction between the multi-scale local features extracted by the CNN encoder and the global features extracted by the transformer encoder. This interaction is achieved through the fusion interaction module (IFM) and the feature reconstruction module (FRM). These design choices enhance the model's ability to capture crack texture features and effectively suppress background noise.
2. The MGAF module is proposed to enable cross-level feature fusion between the encoder and decoder. This module compensates for the semantic loss in low-level features while recovering the details in high-level features, thereby improving the continuity of segmentation results.
3. The MVFF module is proposed to guide the scale-space construction of the SIFT algorithm. This is achieved through multi-view crack masks and unobscured crack masks with dimension-enhancing descriptions. The module effectively mitigates image alignment issues caused by homogeneous feature variability, which is induced by underwater non-uniform scattering noise. Additionally, through adaptive guidance provided by the occlusion masks, the MVFF module restores masked crack features, further improving crack segmentation performance.

2. Related work

2.1. Occluded object detection

Traditional techniques for detecting dam cracks include embedded sensors, ground penetrating radar, and ultrasonic testing. While these methods are effective in traditional environments, their performance is significantly degraded in underwater scenarios due to occlusion, which results in missing target information and poses challenges in feature extraction and crack detection. With the rapid development of deep learning technologies, researchers have been exploring neural networks for hidden object detection. Various approaches have been proposed to address the problem of information loss caused by occlusions. Ke et al. [10] introduced a two-layer convolutional network (BCNet), which is characterized by its two-layer structure: the upper layer detects occluders, while the lower layer infers the occluded parts of the target. This method separates the boundaries of occluders and occluded targets through mask regression, providing an effective solution to the occlusion problem. In the field of multi-object detection, Yuan et al. [11] proposed a generative model that leverages the activation of neural features to accurately localize occluders. By classifying targets based on free areas, the model ensures high detection accuracy. To address the problem of sub-segmentation, Zhang et al. [12] developed OSLPNet, which mitigates the impact of occlusion on feature extraction through multi-scale receptive fields. Additionally, the network leverages the contextual topological relationships of target features to further optimize occluded object detection. Gan et al. [13] improved a two-stage segmentation network by introducing boundary expansion boxes that guide non-modal instance segmentation networks to generate clearer target boundaries. Meanwhile, Wang et al. [14] proposed OccludedInst, a query-based instance segmentation method. By integrating data augmentation techniques and an occlusion correction module, this approach enables

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

robust learning in covert scenarios. For more precise restoration of the appearance of occluded targets, Yan et al. [15] developed an iterative multitasking framework. This framework uses a dual-path structure that includes a 3D model pool and coupled discriminators, which significantly improves the accuracy of target recovery and detection.

2.2. Multi-view object detection

When detecting occluded objects, a single viewpoint image may not accurately identify the target. Multi-view approaches utilize information from multiple perspectives to compensate for the loss of information caused by occlusion in a single view [16]. The biggest challenge in multi-view object detection is effectively merging information from different viewpoints, especially when it comes to occlusions and viewpoint variations. To address these problems, many studies have proposed solutions based on multi-view information fusion [17, 18], generative adversarial networks (GANs) [19, 20], and optical flow learning. Zhou et al. [21] introduced an appearance flow-based method that learns the image flow relationship between the source and target views to reconstruct occluded regions to improve the robustness of multi-view object detection [22]. Choy et al. [23] proposed a recursive neural network framework that recursively merges multi-view information to generate 3D object models with minimal occlusion, thereby mitigating the problems caused by occlusions. Yang et al. [24] developed a Spatiotemporal Graph Convolutional Network (ST-GCN) that integrates both temporal and spatial features, enabling effective video re-identification of pedestrians even in occlusion. Zhang et al. [25] proposed a Multi-View Consistency Generative Adversarial Network (MVCGAN) that successfully generates images from multiple viewpoints through geometric constraints and optimization models and processes complex multi-object scenes using a "decomposition and composition" approach. Overall, these methods cleverly fuse multi-view information and generative models to not only address occlusion problems but also improve the performance of multi-view object detection in complex scenarios, thereby improving the robustness and accuracy of detection systems. Arooj et al. [26] introduced an improved detection network that combines CNNs and SIFT and uses SIFT to extract important feature points from images under different lighting conditions, guiding the network to learn effectively and achieve promising results. Ma et al. [27] proposed a multi-graph matching fusion mechanism that implements a coarse-to-fine matching process and attempts to improve local texture information while preserving the original scene content during the fusion phase.

3. Proposed method

In this paper, we propose aMVFD-Net. The MVFD-Net network structure is shown in Figure 1. MVFD-Net consists of three key components: the feature reconstruction interaction encoder (FRI-Encoder), the multi-scale gated adaptive fusion module (MGAF), and the multi-view feature fusion module (MVFF). First, FRI-Encoder interacts the multi-scale local features extracted by CNN with the global features extracted by transformer, and designs two modules, interaction feature module (IFM) and feature refinement module (FRM), at the middle layer as well as at the end. To solve the problem of feature extraction difficulty caused by underwater non-uniform scattering noise. Second, MGAF performs feature fusion between the encoder and decoder via a pyramid network to further complement the lost feature detail information. Finally, the designed MVFF introduces new perspective features for feature fusion repair to solve the problem of dam cracks obscured by aquatic plants. The following sections provide a detailed analysis of each module and explain how they synergistically improve the overall performance of the MVFD-Net architecture.

3.1. Feature reconstruction interactive encoder (FRI-encoder)

This paper introduces the FRI-Encoder, a solution designed to address the challenges of feature extraction in underwater environments, particularly those caused by non-uniform scattering noise. The FRI-Encoder follows a two-branch architecture. One branch is a lightweight CNN encoder, augmented

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

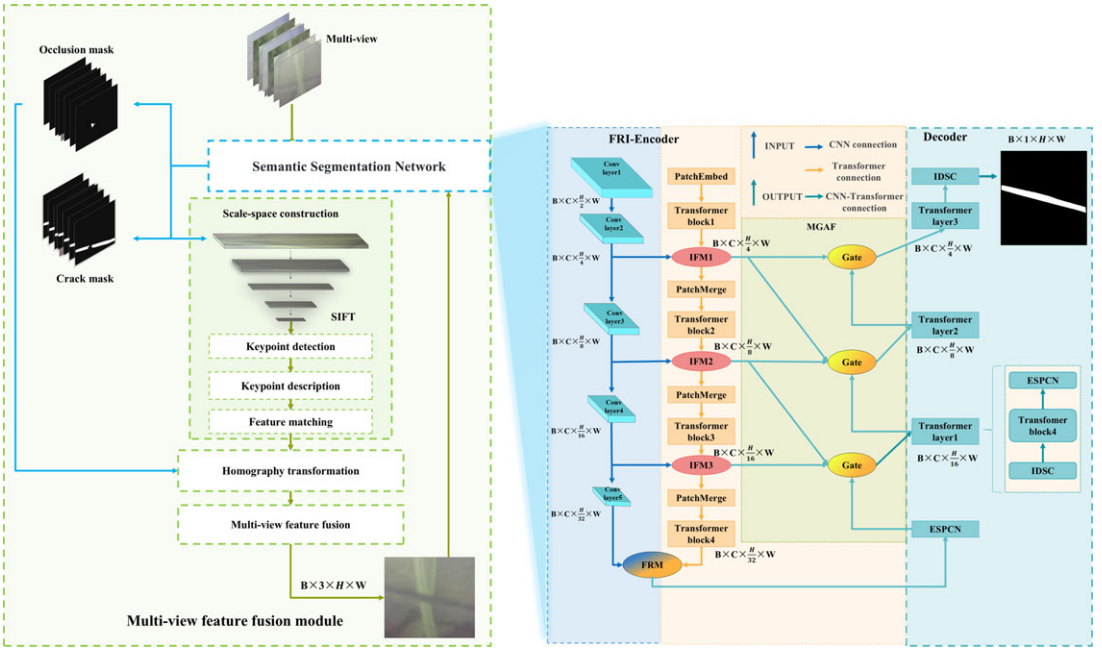

Figure 1. *MVFD-Net Network Framework. The "Conv layer" represents the convolution operations on each layer, "Patch Embed" refers to the embedding layer, and the "transformer block" refers to the transformer blocks used for feature extraction. "ESPCN" indicates the subpixel convolution used for upsampling, while "IDSC" represents depth-wise separable convolutions (PW + DW). Finally, the output provides the predicted results.*

with ResNet34 as the backbone, and incorporates depth separable convolution (DSC) to replace standard convolutional layers [28]. The fundamental unit of the encoder is the Conv block, which includes DSC, batch normalization (BN), and the GELU activation function. By stacking Conv blocks, the FRI encoder effectively extracts local features at each layer, maintaining strong feature extraction capability while minimizing network depth and computational load. The CNN encoder comprises five layers, with the number of channels doubling after every two downsampling operations, leading to a progressively smaller feature map. The second branch, the transformer encoder, consists of four layers. The input image is sequentially processed by the transformer module, with the feature map size reduced by a factor of 1/4 after passing through the embedding layer. To ensure consistent feature map sizes during fusion, features extracted from the second layer of the CNN encoder are merged with those from the first layer of the transformer encoder through the designed IFM module. This process is repeated layer by layer, as illustrated in Figure 2.

Specifically, the IFM aggregates interlayer features of encoders with two branches and further divides them into two subbranches. One of these subbranches is connected to the decoder via jump connections. First, the feature shape is adjusted (from $C \times H \times B$ to $H \times B \times C$), followed by global average pooling (GAP) (reducing to $1 \times 1 \times C$) to calculate the weight for each channel. The weight is then multiplied by the features to adjust the influence of each channel. Finally, the feature shape is restored (back to $C \times H \times B$). After the processed features are multiplied and merged, they are concatenated with the original features to obtain the merged features. Information aggregation is then performed via a convolution module and residual connections, resulting in the complementary fusion features T' that improve feature correlation. The other subbranch generates the fusion feature T, which is then fed back into the transformer encoder. Through this interaction, the local feature information extracted by the CNN encoder can be integrated into the transformer encoder, improving its ability to perceive local details such as edges, shapes, and textures. In this study, the reshaping of inconsistent feature dimensions in

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

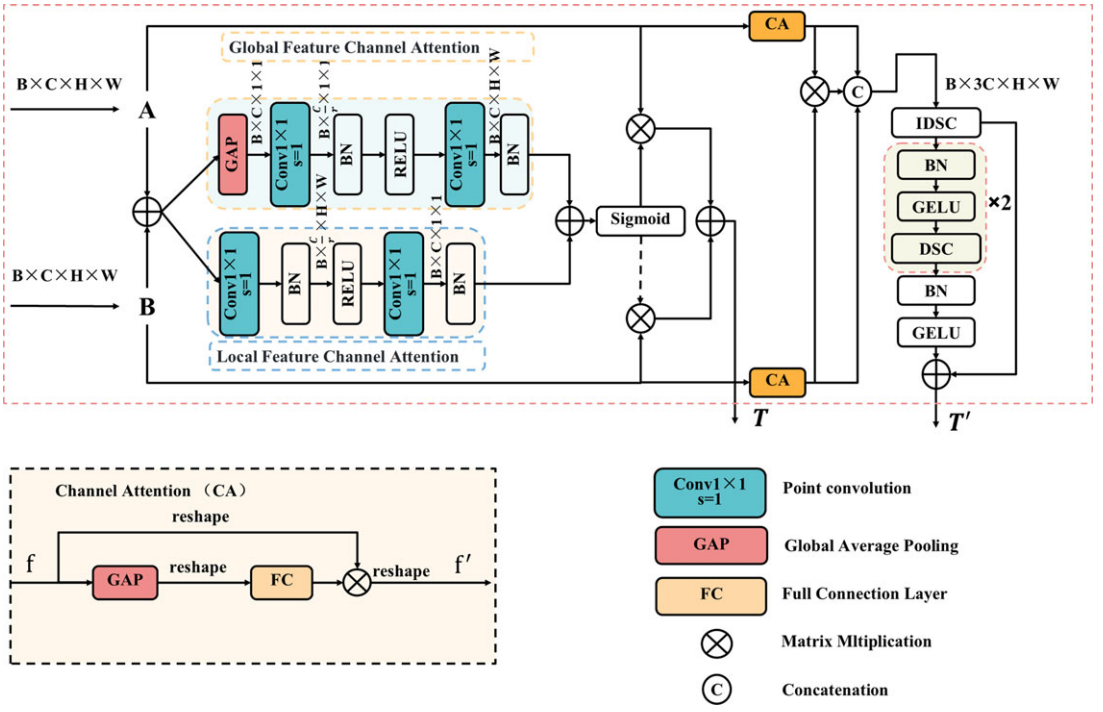

Figure 2. *Illustration of IFM, T denotes the features that are reintroduced into the transformer encoder after fusion, and T′ indicates the features input into the MGAF module after fusion.*

transformer is first aligned with the CNN architecture and summed element-by-element as input. Then, a pixel-wise convolution (1x1 convolution), enhanced by both local and global attention mechanisms, is applied to assign weights and obtain the weighted fused features. The weight calculation process is shown in Table I.

Through the synergy of local and global attention modules, weights can be adaptively assigned to fuse both local and global features. These fused features are then reintroduced into the transformer encoder in an interactive manner, thereby enhancing the transformer's attention to local feature information and generating more representative features. This process can be summarized and represented by the following equation:

$$T = W(A \oplus B) \otimes A \oplus (1 - W(A \oplus B)) \otimes B \tag{1a}$$

where $\oplus$ represents feature integration and we use element-by-element summation, $\otimes$ denotes element-wise multiplication, A denotes the feature maps of the CNN encoder, B denotes the feature maps of the transformer encoder, T represents the fused output features, with $A, B, T \in \mathbb{R}^{C\times H\times W}$. W refers to the processing step described in the pseudo-code in Table I. As illustrated in Figure 2, the dotted line indicates $1 - W(A \oplus B)$. It is important to note that the fusion weight $W(A \oplus B)$ consists of real values between 0 and 1, as does $1 - W(A \oplus B)$, which enables the network to perform a weighted average between A and B.

Although the two-branch coding approach effectively mitigates the noise flooding problem caused by underwater nonuniform noise, it also introduces significant information redundancy. To filter and reconstruct the feature information generated by the FRI-Encoder for fusion, we design the FRM module at the end of the FRI-Encoder. The FRM module is a symmetrical structure, and for ease of presentation, we show only one side of the module as shown in Figure 3.

The FRM module adjusts the shape of the feature maps f_1 and f_2 extracted by the CNN encoder and transformer encoder from (B, C, H, W) to $(B, C, H \times W)$. Next, the global mean and global maximum are computed. After processing by the convolutional layer and ReLU activation function effective global

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table I. *Local and global attention mechanisms.*

Algorithm: Local and global attention mechanisms

Input: X
Output: M
#Local attention mechanism
$L \leftarrow X$
$L \leftarrow \text{Conv1x1}(L, \text{channels}/r)$ # Dimensionality reduction
$L \leftarrow \text{BatchNorm}(L)$
$L \leftarrow \text{ReLU}(L)$
$L \leftarrow \text{Conv1x1}(L, \text{channels})$ # Restore dimensionality
$L \leftarrow \text{BatchNorm}(L)$
#Global attention mechanism
$G \leftarrow X$
$G \leftarrow$ Global Average Pooling(G) # Reduce to 1x1
$G \leftarrow \text{Conv1x1}(G, \text{channels}/r)$ # Dimensionality reduction
$G \leftarrow \text{BatchNorm}(G)$
$G \leftarrow \text{ReLU}(G)$
$G \leftarrow \text{Conv1x1}(G, \text{channels})$ # Restore dimensionality
$G \leftarrow \text{BatchNorm}(G)$
$M \leftarrow \text{Sigmoid}(L + G)$ # Apply Sigmoid normalization to $L + G$
Return: M

feature representations a_1 and a_2 are obtained. This process effectively reduces the amount of post-demand computation. The detailed steps are provided in Table II.

In Table II, i is taken as 1 or 2 corresponding to the features f_1 and f_2 extracted by the input CNN encoder and transformer encoder. "DSC" refers to depthwise separable convolution. The reconstructed feature maps, a_1 and a_2, are extracted by the CNN and transformer encoders, respectively, through feature reconstruction. The cross-attention mechanism is used to reweight f_1 and f_2 by a_1 and a_2 to obtain the feature map $f_{i\text{-cross}}$. The cross-attention weighting process is defined by the following equation:

$$f_{1\text{-cross}} = \text{softmax}(a_1 \times a_2^T) \times f_1 \tag{2a}$$

$$f_{2\text{-cross}} = \text{softmax}((a_1 \times a_2^T)^T) \times f_2 \tag{3a}$$

The cross-attention mechanism captures the correlations between different feature maps, enabling a more accurate representation of information. Next, $f_{i\text{-cross}}$ is reshaped from $(B, C, H \times W)$ to the original shape (B, C, H, W) and fed into the convolutional layer for spatial feature fusion. This process adaptively adjusts the importance of each pixel, preserving richer and more effective spatial structural information. The details are as shown in Table III. $a_{i\text{-spatial}}$ represents the spatial feature weights, and f_i' denotes the feature map obtained after f_i is weighted by $a_{i\text{-spatial}}$. The weighted and fused feature maps, f_1' and f_2', are adjusted in dimensions and then element-wise superimposed. The reshaped feature maps, with shape (B, C, H, W), are used as the final output of the encoder.

FRI-Encoder achieves deep fusion of feature reconstruction and interaction, which not only improves the effectiveness of feature representation but also enhances the model's ability to capture key features of the target. This provides strong support for handling target detection and feature extraction tasks in complex underwater environments [29].

3.2. Multi-scale gated adaptive fusion module (MGAF)

Cracks exhibit complex topological structures, irregular boundaries, and a very small pixel ratio in images. Some of the feature details in cracks often contain important structural information. Nonuniform

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table II. *Feature reconstruction.*

Algorithm: Feature reconstruction

Input: f_i
Output: a_i
$f_i \leftarrow \text{reshape}(f_i, [b, c, -1])$
$avg\_1 \leftarrow \text{Average Pooling}(f_i, \text{dim} = -1, \text{keepdim=True})$
$avg\_1 \leftarrow \text{unsqueeze}(avg\_1, -1)$
$max\_1 \leftarrow \text{Max Pooling}(f_i, \text{dim} = -1, \text{keepdim=True})$
$max\_1 \leftarrow \text{unsqueeze}(max\_1, -1)$
$avg\_1 \leftarrow \text{ReLU}(\text{DSC}(avg\_1))$
$max\_1 \leftarrow \text{ReLU}(\text{DSC}(max\_1))$
$avg\_1 \leftarrow \text{DSC}(avg\_1)$
$max\_1 \leftarrow \text{DSC}(max\_1)$
$a_i \leftarrow avg\_1 + max\_1$
Return: a_i

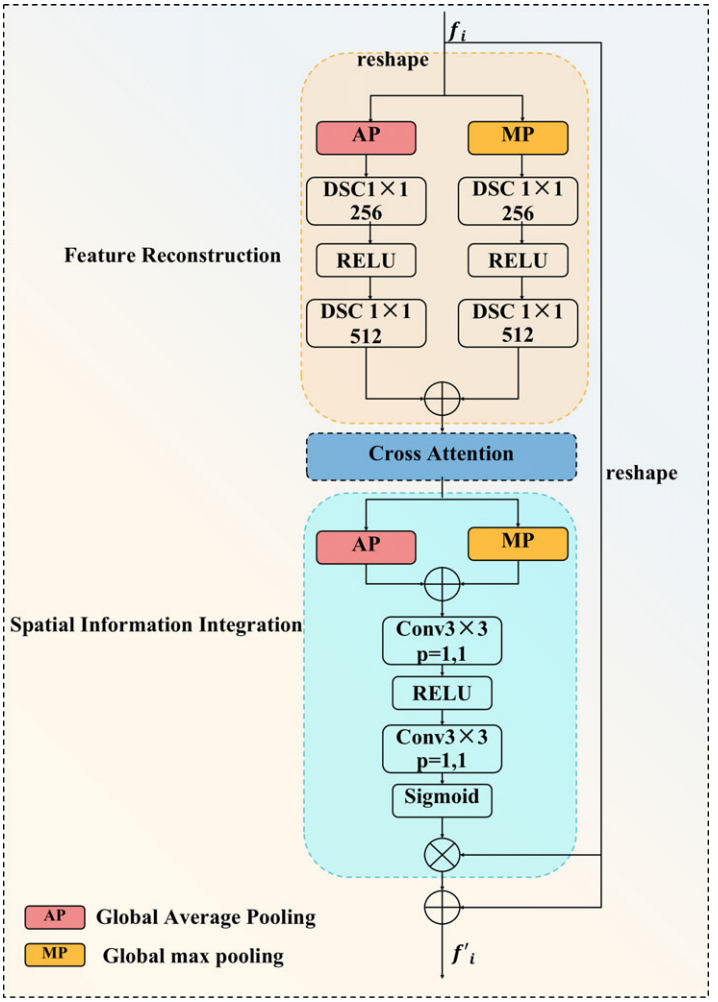

Figure 3. *Illustration of FRM, the feature maps f_i represent the features from the CNN encoder or transformer, while f_i' denotes the reconstructed features.*

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table III. *Spatial information integration.*

Algorithm: Spatial information integration

Input: $f_{i\text{-cross}}$
Output: f_i'
$f_{i\text{-cross}} \leftarrow \text{reshape}(f_{i\text{-cross}}, [b, c, h, w])$
$avg\_out \leftarrow$ Average Pooling($f_{i\text{-cross}}$, dim = 1, keepdim=True)
$max\_out, \_ \leftarrow$ Max Pooling($f_{i\text{-cross}}$, dim = 1, keepdim=True)
$a_1 \leftarrow \text{concat}([avg\_out, max\_out], \text{dim} = 1)$
$a_1 \leftarrow \text{ReLU}(\text{Conv1x1}(a_1))$
$a_1 \leftarrow \text{Conv2x1}(a_1)$
$a_1 \leftarrow \text{reshape}(a_1, [b, 1, -1])$
$a_{i\text{-spatial}} \leftarrow \text{softmax}(a_1, \text{dim} = -1)$
$f_i' \leftarrow f_i * a_{i\text{-spatial}} + f_i$
Return: f_i'

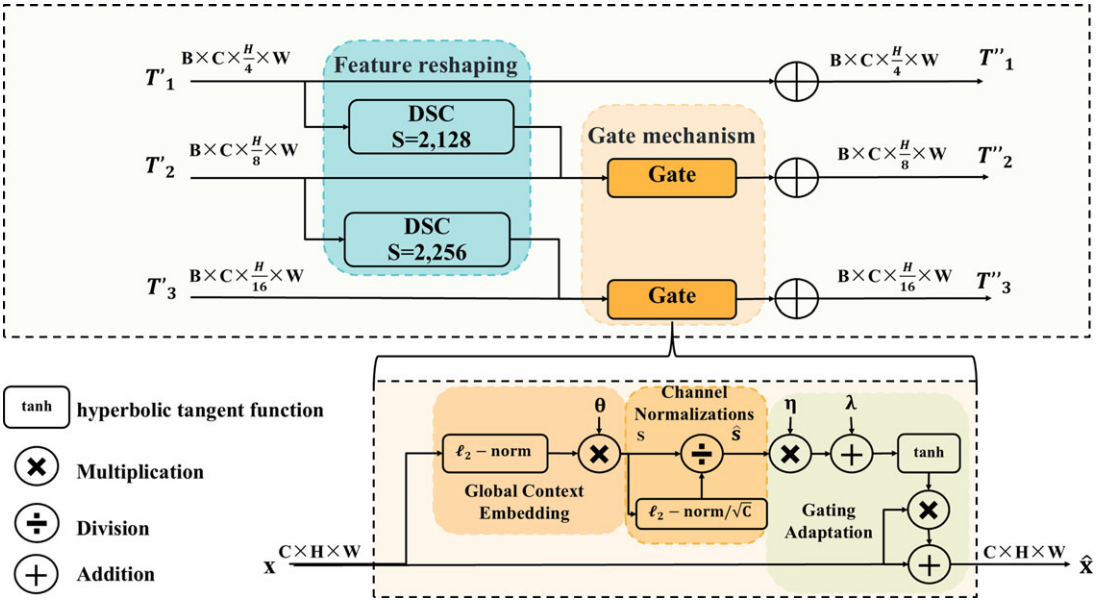

Figure 4. *Illustration of MGAF. T_i' and T_i'' represent the input features and the features processed by the MGAF module, respectively. Gate represents the gating mechanism, and θ denotes the embedding weights that manage the channel weights prior to normalization. The gating weights and bias (η and λ) progressively adjust the input feature proportions x across the channels.*

underwater noise further exacerbates the obscurity of these detailed features. In this article, we propose the MGAF module, which can merge functions from different learning stages, effectively supplementing the detailed information of high-level functions. The MGAF network framework is shown in Figure 4.

The MGAF module also uses DSC instead of standard convolution to process the FRI-Encoder middle layer fusion feature T'. DSC effectively adjusts and aligns features at different scales to achieve more efficient multi-scale feature fusion. Then, these features are concatenated along the channel dimension to obtain the full merged features. Next, the fused feature map is fed into the gating mechanism for processing. We introduce the operator $\boldsymbol{\theta}$ to embed the global context and control the weights of each channel before normalization. Then, the gated adaptation operators $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ are introduced. This operator adjusts the input features line by line based on the normalized outpu, and $\boldsymbol{\theta}$ is responsible for adjusting the

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

embedding output. Gating weights $\boldsymbol{\lambda}$ and bias $\boldsymbol{\eta}$ control the activation of the weight coefficients. These weights determine the behavior of the gating mechanism in each channel. In this paper, let $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ represent an activation feature in a convolutional network, where H and W are the spatial height and width, and C is the number of channels. The total equation of the gating mechanism is given as follows:

$$\hat{\mathbf{x}} = F(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\lambda}), \quad \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\lambda} \in \mathbb{R}^C \tag{4a}$$

where $\hat{\mathbf{x}}$ represents the result processed by the adaptive gating mechanism. For each channel, the receptive field of the convolutional neural network is enhanced by designing a global context embedding module. This enables the network to better aggregate and utilize global information. Given the embedding weight operator $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_C]$, the approach avoids ambiguities that may arise from a limited receptive field. Let $\mathbf{x} = [x_1, \ldots, x_C]$, where $x_c = \left[x_c^{(i,j)}\right]_{H\times W} \in \mathbb{R}^{H\times W}$ with $c \in \{1, 2, \ldots, C\}$. The global feature representation g_c is then obtained as follows, as shown in the equation:

$$g_c = \theta_c \|x_c\|_2 = \theta_c \left(\left[\sum_{i=1}^{H} \sum_{j=1}^{W} (x_c^{(i,j)})^2 \right] + \epsilon \right)^{1/2} \tag{5a}$$

where x_c represents the feature map of each channel in x. In this paper, x_c is ℓ_2 normalized to retain more detailed feature information compared to GAP. θ_c represents the trainable parameters that control the weight of each channel. The proposed channel normalization method employs ℓ_2 normalization across channels to enhance the training stability and model performance while significantly reducing computational complexity. Let $G = [g_1, \ldots, g_C]$ represent the normalized feature representations, as shown in the following equation:

$$\hat{g}_c = \frac{\sqrt{C} s_c}{\|G\|_2} = \frac{\sqrt{C} g_c}{\left(\sum_{c=1}^{C} g_c^2 + \epsilon \right)^{1/2}} \tag{6a}$$

where ϵ is a small positive constant, and $\sqrt{C}$ is introduced to prevent $\hat{g}_c$ from becoming too small when C (the number of channels) is large. To further control the feature expression of each channel, this study employs a gating mechanism to dynamically regulate the control gates on the channels by designing trainable gating weights $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_C]$ and biases $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_C]$ to adjust the activations. The equation is as follows:

$$\hat{x}_c = x_c \left[1 + \tanh\left(\eta_c \hat{g}_c + \lambda_c \right) \right] \tag{7a}$$

Where η_c and λ_c represent the gating weight and bias of the c-th channel, respectively.

Following the MGAF module processing, more complex dependencies are captured, and global context normalization is introduced through efficient feature fusion between the encoder and decoder. By tuning the trainable parameters, the model adaptively optimizes global features and enhances feature representation. This approach is particularly well-suited for small and imbalanced datasets, as it not only reduces computation from excessive fused information but also improves segmentation performance and model generalization.

3.3. Multi-view feature fusion (MVFF)

This article proposes a MVFF to solve the problem of occlusion, as shown in Figure 5.

First, the feature point description level of the SIFT algorithm is improved by extending the traditional 128-dimensional descriptor to 180 dimensions. The improvement is to define a circular area with a radius of 30 pixels around the key point, divided into 15 concentric rings. After calculating the gradients of each region with Gaussian weighting, the main direction of the feature point is optimized taking into account the rotation invariance of the circular region. This approach not only increases computing efficiency, but also simplifies operation. To ensure the uniqueness of the feature points, the gradients in the circular area are divided into 12 directions and the values in each direction are accumulated. Finally,

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

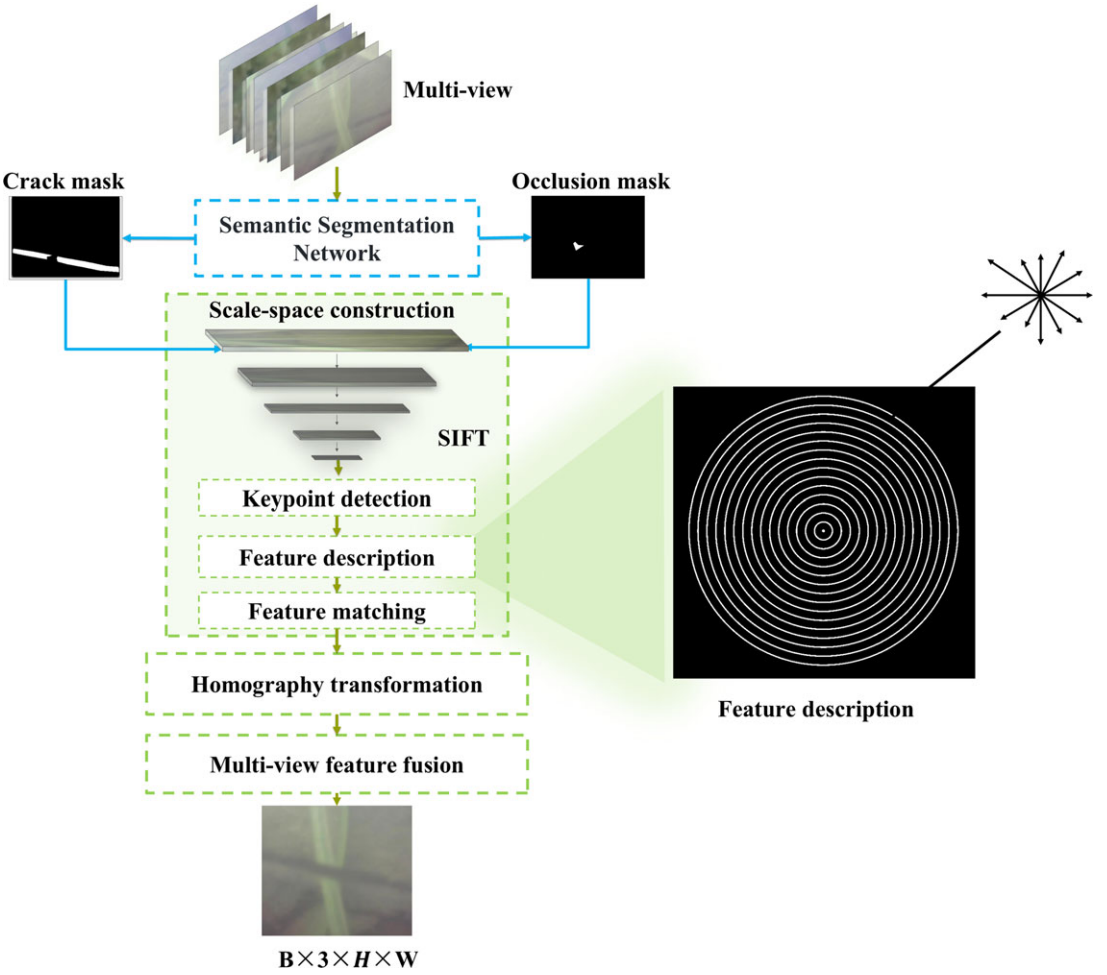

Figure 5. *Multi-view feature fusion module.*

the gradient sums of the regions are arranged in an inside-out order, forming a 180-dimensional feature vector. By expanding the descriptor dimensions, this method enables more precise encoding of the local region details in the image and captures more subtle features in complex underwater environments with uneven noise. In addition, to mitigate the impact of gray level variations, the generated feature vector is normalized.

However, expanding the descriptor also increases computational complexity, especially in feature matching and storage, which may lead to additional burden. To balance precision with computational efficiency and to account for the differences in homogeneity characteristics caused by different noise levels in underwater images from different viewpoints, this paper proposes a network based on the encoder-decoder structure designed in Sections 3.1 and 3.2. This network performs semantic segmentation of both free and blocked cracks in the dam body and generates corresponding masks. The location information provided by the free crack mask is used to guide the construction of the scale space, allowing the network to quickly focus on the regions with significant crack features. This approach enables efficient feature point detection and significantly reduces the amount of unnecessary calculations. Feature point matching is then performed using the SIFT algorithm and based on the matching relationships, a homography matrix transformation is applied to the new viewpoint image. The specific calculation formula is as follows:

$$p' = H \cdot p \tag{8a}$$

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

where $p = [x, y, 1]^T$ represents a point in the original view, expressed in homogeneous coordinates (with the third element being 1), the transformed point is denoted as $p' = [x', y', 1]^T$. The relationship between the original and transformed points is described by the homography matrix H, a 3×3 matrix that governs the perspective transformation between the two views.

Finally, under the guidance of the hidden crack mask, an adaptive weighted multi-feature fusion of hidden features in the original view is performed using the new views. Specifically, there are N views, where V_1 represents the original view and V_i ($i \in [2, N]$) represents the new views. In each view, if a pixel at position p is occluded, the occlusion mask is called $M_i(p) = 1$, otherwise $M_i(p) = 0$. If the position p in V_1 is occluded ($M_1(p) = 1$), we want to use the unoccluded regions in other perspectives to perform the recovery. This means that during the weighted averaging process, only the feature values in corresponding positions in other views that are not occluded should be involved in the recovery. The equation for this is as follows:

$$V_i(p) = \begin{cases} V_i(p), & \text{if } M_i(p) = 0 \\ 0, & \text{if } M_i(p) = 1 \end{cases} \tag{9a}$$

where $V_i(p)$ represents the feature value at position p in the i-th view and W_i is the weight assigned to each view. The less occlusion there is in the new view, the greater its contribution to restoring the original view. Therefore, we introduce a weighting coefficient $W_i(p)$, which is defined as follows:

$$W_i(p) = \frac{1 - M_i(p)}{\sum_{j=1}^{N} (1 - M_j(p))} \tag{10a}$$

where $(1 - M_i(p))$ ensures that only the views in which the position p is not obscured contribute to the weighting. This method effectively solves the problem that hidden and non-hidden features cannot be merged. Finally, for all views, the uncovered feature values at position p are weighted and averaged point by point to generate the repaired feature V. The final repair result can be expressed as follows:

$$V(p) = \begin{cases} V_1(p), & \text{if } M_1(p) = 0 \\ \sum_{i=2}^{N} W_i(p) \cdot V_i(p) \cdot M_i(p), & \text{if } M_1(p) = 1 \end{cases} \tag{11a}$$

where $V(p)$ represents the repaired feature value at position p. In particular, for the visible area in the original view V_1 (where $M_1(p) = 0$), the original feature is retained. For the occluded region (where $M_1(p) = 1$), the occluded feature is adaptively repaired by averaging the unoccluded features from other viewpoints using a weighted fusion approach.

This paper proposes a multi-view image feature fusion method that improves the SIFT algorithm by expanding the descriptor dimension to achieve more precise encoding of local details. Combined with a semantic segmentation network, the method generates occlusion masks to optimize the detection and assignment of feature points. After aligning the views using the homography matrix, the method adaptively repairs the occluded regions using non-occluded features and finally generates a fused feature representation. This approach effectively solves the problem of merging homogeneous feature information affected by different noise levels from different viewpoints, thereby enabling hidden feature recovery.

3.4. Loss function

To ensure that the algorithm can be trained effectively, selecting an appropriate loss function is crucial. In the case of an unbalanced dataset, using Dice Loss can result in significant fluctuations. Prediction errors with small targets can result in sharp loss value changes, leading to severe gradient instability. Given that cracks in images are typically slender and elongated, making them challenging to segment as small targets, a combination of BCE_Loss and Dice_Loss was chosen. This combination enables the network to focus more effectively on foreground regions while mitigating the instability of Dice_Loss during training. The specific formulas are as follows:

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

$$\text{BCE\_Loss} = -(1-m)\log(1-t) - m\log(t) \tag{12a}$$

$$\text{Dice\_Loss} = 1 - \frac{2y\hat{s}+1}{y+\hat{s}+1} \tag{13a}$$

$$\text{Loss} = \text{BCE\_Loss} + \text{Dice\_Loss} \tag{14a}$$

where m represents the ground truth labels and t is the predicted probability in the BCE_Loss function. BCE_Loss measures the discrepancy between the predicted values (t) and the true labels (m). In Dice_Loss, $\hat{s}$ is the model's estimated probability, and y is the label. Dice_Loss assesses the overlap between the predicted and actual regions. Since binary classification is a problem involving 0s and 1s, the final predicted values also fall between 0 and 1. To categorize predictions into two classes, we need to establish a threshold. Since crack segmentation focuses solely on differentiating between crack areas and the background, this represents a binary classification challenge. We utilize Dice_Loss with a threshold of 1 to facilitate this distinction. The combined loss function we constructed takes into account both the accuracy of the model's predictions and the precision of the predicted regions, further enhancing segmentation performance.

4. Experiment setting

The experimental setup in this paper consists of three key components: evaluation metrics, experimental datasets, and model implementation details. First, the evaluation metrics used to assess the performance of the proposed model are outlined. Second, the datasets employed in the experiments are discussed. Lastly, an explanation of the model implementation is provided, including the hardware and software used during the experiments. We adopt the standard evaluation metrics for semantic segmentation [30], including mIoU (MIOU), recall (Re), accuracy (Acc), and F1 score (F1). In addition, in order to evaluate the computational complexity of the model, this paper introduces a comparison between the proposed algorithm and the leading models in terms of the amount of model calculations (FLOPs) and the number of parameters (Params) to comprehensively evaluate the performance of the algorithm.

4.1. Dataset

The self-constructed dataset used in this paper is a crack image of a submerged dam in a reservoir in Zhejiang Province. The images in this dataset present challenges, such as complex underwater non-uniform noise and occlusion caused by aquatic plants. As a benchmark for evaluation, this dataset is highly relevant. To facilitate analysis, the dataset adopts a multi-view setup for each crack, comprising three images: one original view and two additional perspectives. These three images are captured from different viewpoints of the same dam crack. Subsequently, the dataset is categorized into three underwater cases by filtering and collating the images: the Weak Non-Uniform Noise Underwater Occlusion Dataset (UDODW), the Strong Non-Uniform Noise Underwater Occlusion Dataset (UDODS), and the Aquatic Plant Occlusion Dataset (UDPOD). The first two datasets, UDODW and UDODS, use randomly generated black blocks to simulate occlusion from various virtual perspectives, while the UDPOD dataset contains images of underwater dam cracks occluded by actual aquatic plants. Finally, after filtering the images, the three datasets, with 360 image sets in total, are organized into underwater multi-view occlusion datasets. A visual representation of these datasets is shown in Figure 6.

4.2. Experimental environment

The dataset used in this study has a resolution of 256×256. The network model is built using the PyTorch deep learning framework. The configuration for training the network model includes an AMDEPYC 7282 processor, 250 GB of memory, and an NVIDIA A100 80 GB PCIe GPU. The training is carried out using the GPU. In this paper, Adam optimization [31] was employed, the initial learning rate was set to 0.001, and the cosine annealing strategy with thermal restart was applied to dynamically adjust the learning rate. At epoch 20, the initial learning rate was restored for the first time, and each subsequent

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table IV. *Comparison of segmentation models on the UDODW dataset.*

Models	mIoU (%)	Acc (%)	Re (%)	F1 (%)	FLOPs(G)	Params(M)
SegNet [34]	86.25	85.16	87.24	86.46	40.08	167.80
CrackFormer-II [32]	89.82	90.89	91.20	90.77	20.49	4.96
CarNet [33]	86.33	87.45	88.02	87.09	4.79	4.89
OSLPNet [12]	90.75	**93.43**	92.17	91.98	**0.81**	**2.92**
UISS-Net [35]	82.95	85.92	85.82	85.03	42.61	240.47
Proposed	**92.79**	93.36	**93.51**	**93.27**	1.07	4.09

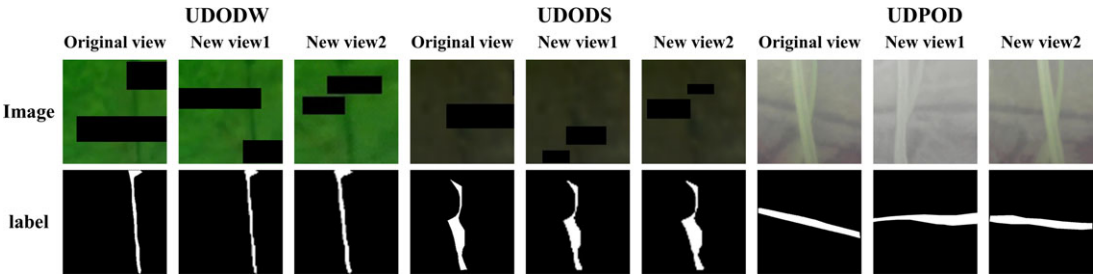

Figure 6. *Illustration of the images in the three datasets.*

restoration doubled the previous one. The network was trained for a total of 120 epochs, with a batch size of 4.

5. Experiments and results

5.1. Comparative experiments

In this section, the UDPOD dataset is randomly divided into a training set, a validation set, and a test set in the ratio of 7:2:1. The proposed algorithm is compared with five other algorithms, namely CrackFormer-II [32], CarNet [33], SegNet [34], OSLPNet [12], and UISS-Net [35]. These compared algorithms are all state-of-the-art segmentation networks in their respective fields, with UISS-Net [35] being a segmentation network designed for underwater scenarios, and OSLPNet [12] being an occlusion-resistant segmentation network. We also use the recommended parameter settings and run the source code provided by the authors to achieve the best results for each method. Comparative experiments are conducted on the UDODW, UDODS, and UDPOD datasets to demonstrate the superiority of the proposed network model. Additionally, ablation experiments are performed on the UDPOD dataset to assess the contribution of each module. The results of these experiments are as follows:

5.1.1 Comparative experimental results analysis

Qualitative analysis: From the comparative evaluation results, it is evident that the algorithm proposed in this paper performs exceptionally well across all three datasets, especially in several key evaluation metrics where it achieves significant improvements. In Table IV, although the accuracy (Acc) of this paper's algorithm on the UDODW dataset is slightly lower than the OSLPNet algorithm, it surpasses it by 2.04% in the mean intersection over union (mIoU). This indicates that the proposed algorithm is better at distinguishing between cracked and non-cracked regions. Additionally, the F1 score of the proposed algorithm is 1.29% higher than the second-ranked algorithm, demonstrating its improved ability to identify cracked regions while minimizing false positives for non-cracked areas. Although the proposed algorithm outperforms the OSLPNet algorithm in FLOPs (G) by 0.26G, it is still far lower than the other compared algorithms, with a total of only 1.07G. Additionally, the number of parameters is just 4.09M, which is sufficient for deployment on mobile devices, such as underwater robots. These

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table V. *Comparison of segmentation models on the UDODS dataset.*

Models	mIoU (%)	Acc (%)	Re (%)	F1 (%)
SegNet [34]	71.93	74.09	77.74	75.25
CrackFormer-II [32]	87.37	87.84	89.04	88.30
CarNet [33]	79.53	90.19	83.39	82.90
OSLPNet [12]	89.49	91.39	91.05	90.71
UISS-Net [35]	82.55	85.43	85.41	84.66
Proposed	**90.93**	**91.58**	**92.07**	**91.64**

Table VI. *Comparison of segmentation models on the UDPOD dataset.*

Models	mIoU (%)	Acc (%)	Re (%)	F1 (%)
SegNet [34]	86.61	86.88	87.84	87.29
CrackFormer-II [32]	93.21	93.39	93.64	93.47
CarNet [33]	86.33	87.45	88.02	87.09
OSLPNet [12]	92.54	93.01	92.89	92.82
UISS-Net [35]	86.20	87.47	87.61	87.22
Proposed	**95.36**	**95.71**	**95.64**	**95.57**

results demonstrate that the proposed network can be efficiently deployed on mobile devices, offering low computational complexity while ensuring excellent recognition performance in the presence of challenges such as occlusion and noise. In the UDODS dataset, shown in Table V, the proposed algorithm improves by 1.45% in mIoU and 0.93% in F1 score. These results indicate that the algorithm exhibits strong robustness in complex environments with significant noise, effectively handling interference from highly noisy scenes. Table VI further confirms the exceptional performance of the proposed algorithm in environments obscured by aquatic plants. All evaluation metrics of the proposed algorithm outperform those of the other five compared algorithms. Specifically, the accuracy (Acc) and mIoU are improved by 2.32% and 2.15%, respectively, over the second-place algorithm, while the F1 score shows an impressive improvement of 95.57%. This highlights the algorithm's excellent balance between segmentation precision and recall. These results provide strong evidence that the proposed network excels in tackling challenges such as occlusion and noise.

Quantitative analysis: Figures 7 and 8 demonstrate that the proposed network shows a consistent improvement throughout the training process, with all evaluation metrics outperforming those of the compared algorithms. The loss decreases smoothly without significant fluctuations, indicating that the network exhibits strong learning ability and convergence. As shown in Figure 9, the proposed algorithm demonstrates robust performance on both the UDODW and UDODS datasets, accurately segmenting the target region. Although the OSLPNet and UISS-Net algorithms also perform well, their segmentation masks display noticeable breakpoints and under-segmentation in regions where cracks are obscured. In contrast, the segmentation masks of the SegNet and CrackFormer-II algorithms are generally larger than the actual crack width, exhibiting significant noise and over-segmentation. This results in a higher number of false negatives (FNs), where non-cracked regions are misclassified as cracked regions. For the UDPOD dataset, the real hydrilla occlusion better reflects the semantic correlation between objects, necessitating stronger semantic reasoning and context-awareness from the algorithm. As shown in Figure 9, the proposed algorithm still effectively segments the crack features in the occluded areas. In contrast, other algorithms can only rely on the visible crack region to predict the occluded crack, resulting in noticeable breakpoints in the segmentation mask and an inability to accurately segment the occluded cracks.

As shown in Figure 9(g), in this crack scene, the homogenization of the background and crack features is even more pronounced due to over-illumination, low image contrast, and the inherent noise

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

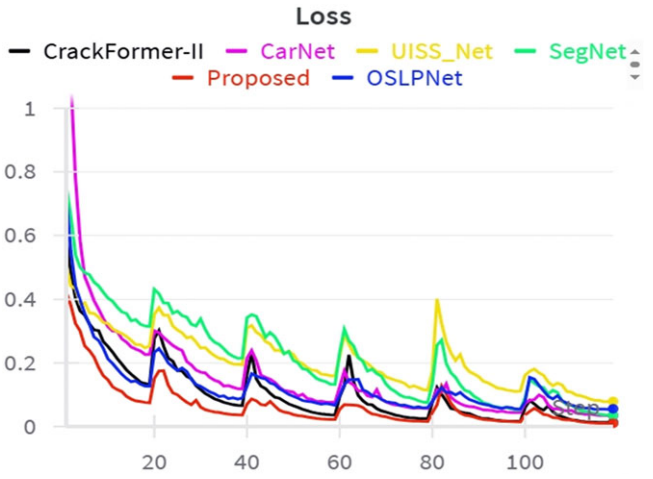

Figure 7. *Illustration of the changes in loss during the training stage.*

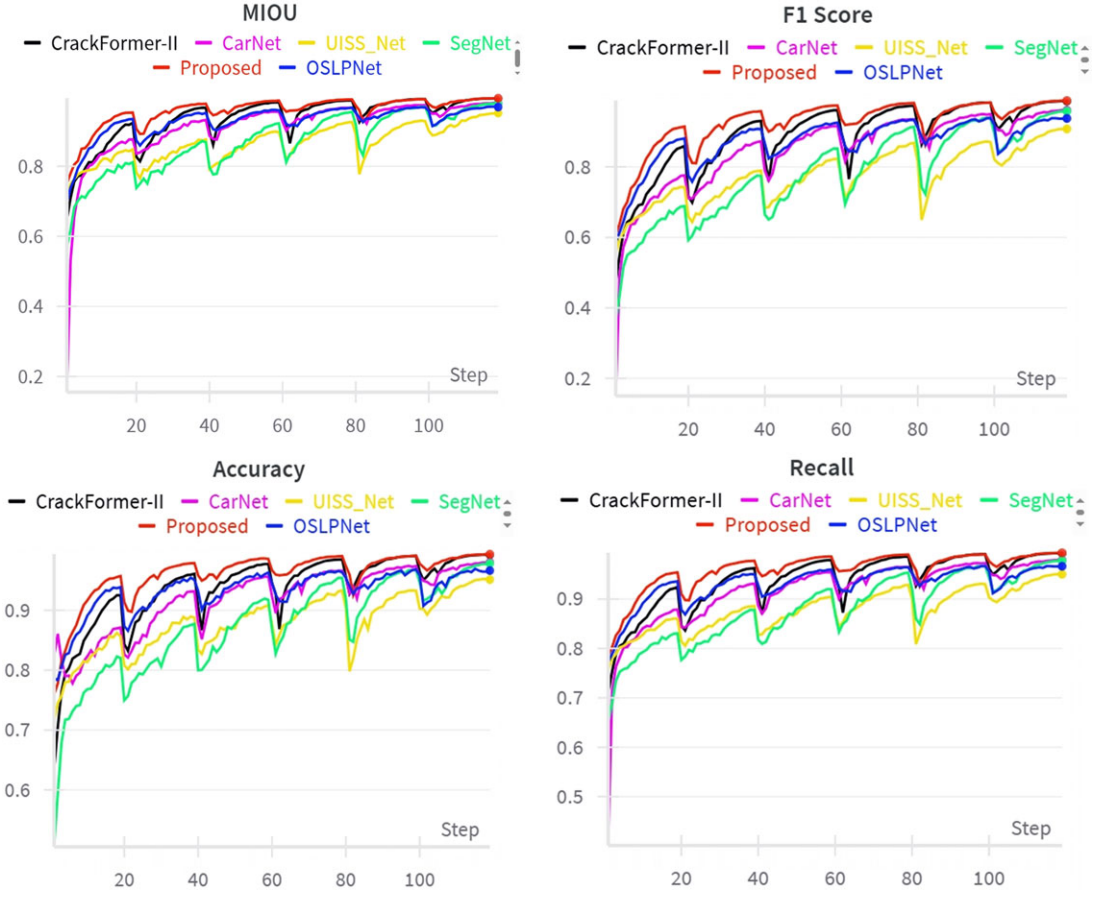

Figure 8. *Illustration of the changes in mIoU, Acc, Re, and F1 during the training stage.*

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

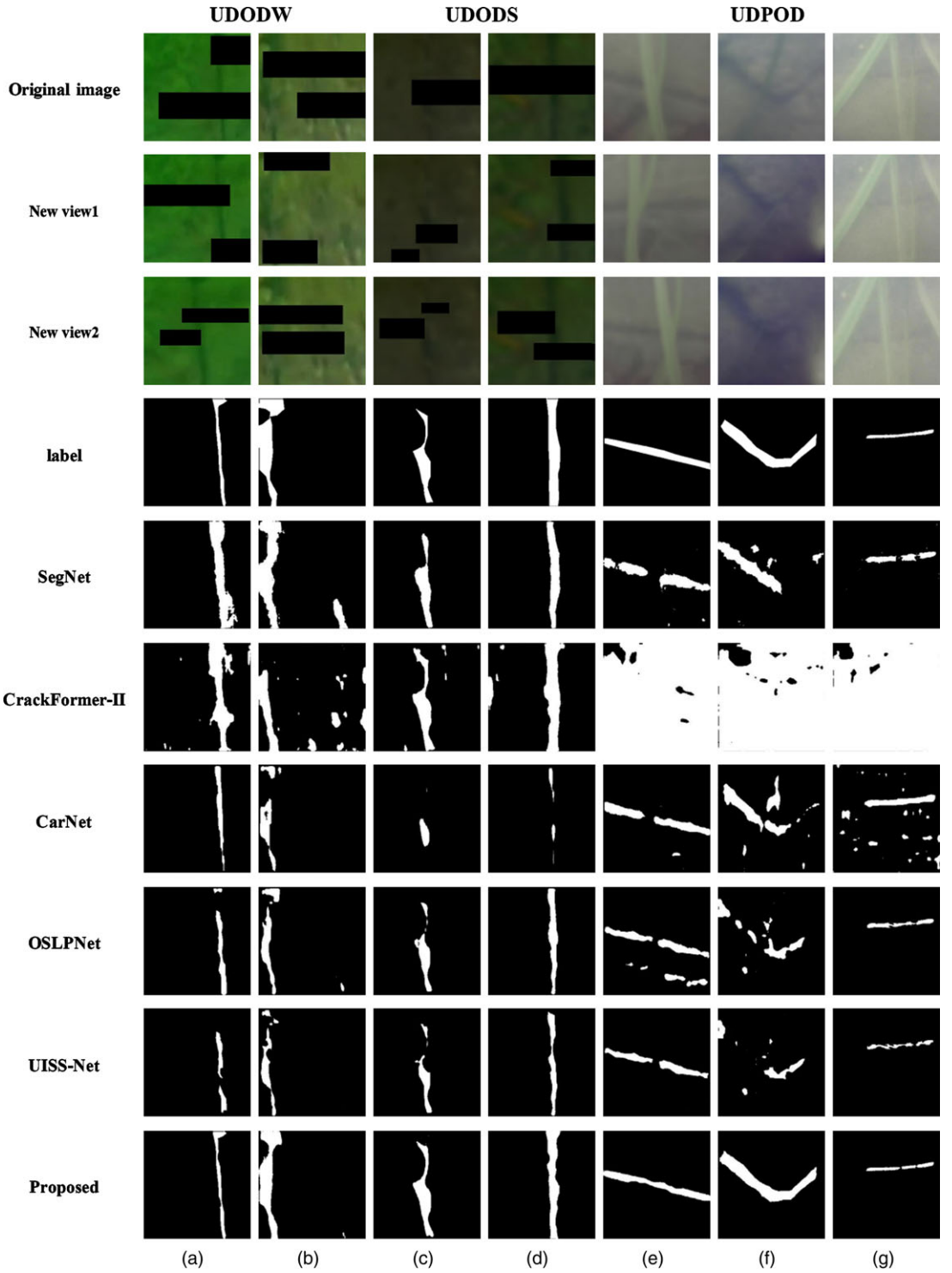

Figure 9. *Illustration of segmentation results of comparative experiments on different datasets.*

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table VII. *Ablation experimental results on UDPOD dataset.*

Ablation Study	w/o Module	mIoU (%)	Acc (%)	Re (%)	F1 (%)
Baseline		90.97	90.37	90.54	90.04
Baseline + MVFF + MGAF	w/o FRI-Encoder	91.07	91.74	92.15	91.71
Baseline + FRI-Encoder + MVFF	w/o MGAF	94.55	94.62	93.89	94.41
Baseline + FRI-Encoder + MGAF	w/o MVFF	94.37	93.89	94.09	94.21
Proposed		**95.36**	**95.71**	**95.64**	**95.57**

effects of the underwater environment. This makes it difficult for the algorithm to accurately distinguish between the occluded and unoccluded parts. Compared to other contrast-based algorithms, the proposed algorithm effectively avoids excessive noise or over-segmentation and performs better. Despite a few disconnected under-segmentations in the occluded areas, the algorithm still segments the overall morphology of the cracks more effectively. This suggests that, in addition to improving the recognition algorithm, the simultaneous optimization of other underwater recognition equipment is also crucial and worthy of further consideration.

In summary, the algorithm proposed in this paper not only demonstrates excellent segmentation performance but also maintains good robustness in complex environments where dam cracks are occluded by aquatic plants, effectively addressing the occlusion problem in the recognition of dam cracks with different morphologies.

5.2. Ablation experiments

To evaluate the effectiveness of the FRI encoder, MGAF module, and MVFF module in segmenting occluded cracks in submerged dams, this paper conducts ablation experiments on the UDPOD dataset. The ablation experiment is designed as follows: Baseline, Proposed + w/o FRI-Encoder, Proposed + w/o MGAF, and Proposed + w/o Multi-view Occlusion Completion Module.

5.2.1 Ablation experimental results analysis

The results of the ablation experiments are presented in Table VII. Since the baseline model is not specifically optimized for the occlusion problem under underwater non-uniform scattering conditions, the algorithm proposed in this paper addresses the target loss issue in occluded areas by incorporating new viewpoint features. Consequently, compared to the baseline, the proposed algorithm demonstrates substantial improvements across all evaluation metrics. Notably, the mean Intersection over Union (mIoU) improves by 4.39%, and the F1 score increases by 5.53%, which underscores the enhanced segmentation stability of the model.

Further analysis of the impact of removing each design module reveals a significant decline in performance metrics compared to the full model. First, when the multi-view feature fusion (MVFF) module is removed, mIoU and accuracy decrease by 1.99% and 2.82%, respectively. This suggests that the MVFF module plays a crucial role in effectively fusing multi-view feature information and enhancing segmentation performance under occlusion. Second, the removal of the feature refinement and integration (FRI) encoder results in a marked decrease in evaluation metrics, with mIoU and F1 dropping by 4.29% and 3.86%, respectively. This indicates that the FRI encoder significantly mitigates the challenges of feature extraction caused by underwater nonuniform noise, thus improving the model's robustness. Finally, the removal of the MGAF module leads to a decrease of 0.81% in mIoU and 1.16% in F1, further demonstrating the contribution of the MGAF module in enhancing model accuracy. Visual results presented in Figure 10 corroborate these findings. When the MVFF module is omitted, the segmentation results shown in Figure 10(e) reveal poor crack identification in the occluded region. Due to the lack of feature

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

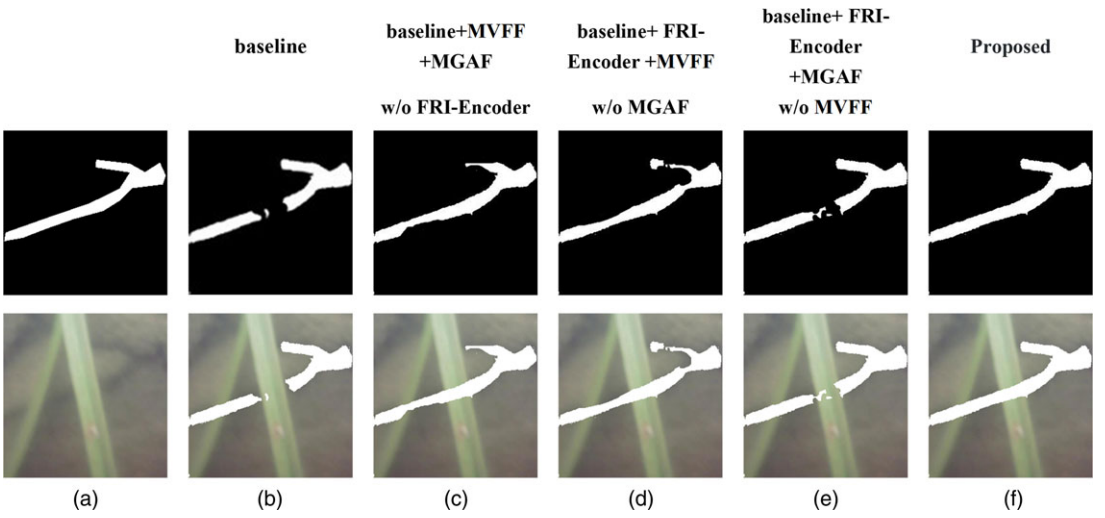

Figure 10. *Illustration of segmentation results of ablation experiments on UDPOD dataset .*

information from the occluded area, the network attempts to infer features based on crack connectivity but is limited in its ability to perform effective recognition. This highlights the critical role of the MVFF module in addressing the occlusion problem by compensating for missing information through multi-view feature fusion. In contrast, the segmentation of local details (e.g., texture information such as edges and shapes) improves significantly when the FRI-Encoder module is introduced, as shown in Figure 10(d) and (e). In Figure 10(c), where the FRI-Encoder module is absent, the crack boundaries are noticeably under-segmented, underscoring the encoder's role in alleviating feature extraction difficulties caused by underwater non-uniform noise. When only the MGAF module is removed, the segmentation results shown in Figure 10(d) reveal obvious breakpoints in the crack boundaries, indicating the loss of information during the up-sampling stage. In contrast, the comparisons in Figure 10(f) and Figure 10(c) demonstrate that the MGAF module effectively compensates for information loss during the up-sampling process, mitigating segmentation discontinuities and improving both segmentation accuracy and coherence.

Overall, the visual and quantitative results of the ablation experiments clearly demonstrate the effectiveness of the FRI encoder, MGAF module, and MVFF module. Furthermore, they highlight that the proposed MVFD-Net model offers a significant advantage in comprehensive segmentation capability, effectively addressing occlusion and noise challenges, and providing stable and accurate segmentation results in complex underwater environments.

5.3. Practical application and result analysis

5.3.1 Introduction to remotely operated vehicles

In this paper, the BlueROV2 underwater robot, shown in Figure 11, is employed for practical application validation. The specific parameters of the robot are provided in Table VIII.

5.3.2 Application validation settings

In this section, we deploy the optimal model weights—trained on the UDPOD dataset—onto an underwater robot that is tethered via cable to an offshore mobile display unit. Under full-power, uniform illumination, the robot conducts field tests along the embankment of Dingguo Lake in Xinxiang City, Henan Province. To evaluate the algorithm's performance, we selected three crack scenarios (Test 1, Test 2, and Test 3). During each scenario, the robot sequentially captures images of the same occluded crack from three positions (P0, P1, and P2), yielding three distinct viewpoints (Image 1, Image 2, and

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Table VIII. *Specific parameters of the BlueROV2 underwater robot.*

Parameter	Specification
Dimensions (L∗W∗H)	460 mm ∗ 560 mm ∗ 255 mm
Weight	12 kg
Max operating depth	300 m
Lighting output	1500 lumens per light source
Sonar localization system	Detection resolution: less than 1cm
Camera type	Wide-angle low-light camera
Image resolution	1080p Full HD
Frame rate	30fps
Gimbal tilt range	±45°

Figure 11. *BlueROV2 underwater robot.*

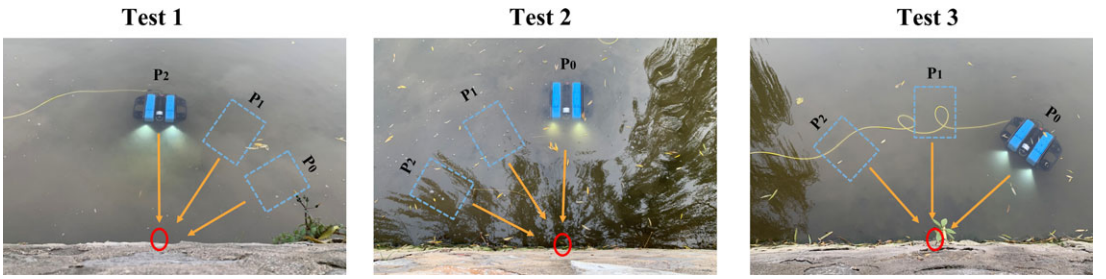

Figure 12. *Illustration of relative positions of the underwater robot. The red circle denotes the location of the dam crack and the yellow guide line represents the direction of the robot's view.*

Image 3). As illustrated in Figure 12, we annotate the robot's acquisition positions corresponding to each image across all test scenarios.

In this study, the underwater robot is remotely operated from an offshore unit and maneuvered to a position at a distance r from the embankment crack area. At this location, the robot autonomously calculates multiple image acquisition viewpoints to capture crack images. To mitigate the impact of uneven viewpoint distribution on testing accuracy, a viewpoint constraint model is established to ensure that the acquisition points for each crack are both accurate and reasonably distributed.

Due to the highly nonlinear motion behavior of the robot in complex underwater environments, the robot's depth and orientation in the vertical direction are adaptively adjusted according to the acquisition

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

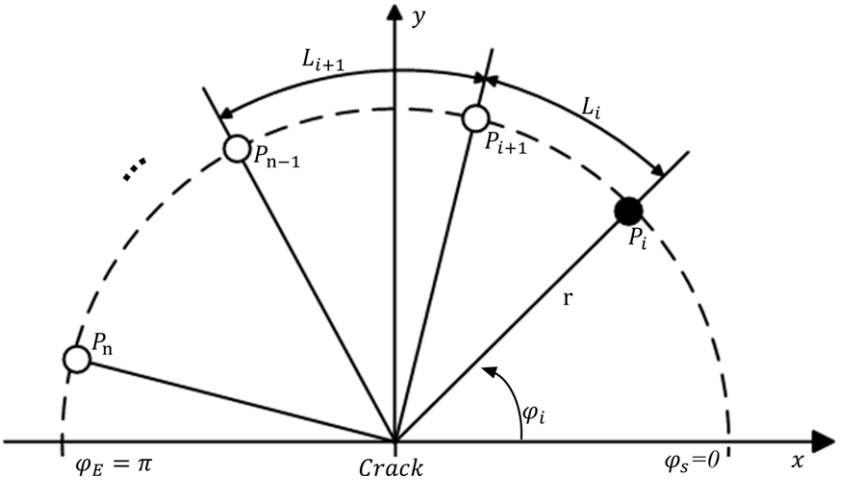

Figure 13. *Schematic layout of underwater robotic imaging positions centered on dam cracks.*

task requirements and terrain constraints. This ensures that the camera's optical axis remains focused on the embankment cracks. On the horizontal plane, the multi-viewpoint positions are uniformly distributed by applying a viewpoint distribution constraint model.

Specifically, the camera position of the underwater robot in space is treated as the origin, while the observable range of the crack is abstracted as a semicircular region centered at the crack center with radius r. A total of n acquisition points are deployed along this semicircular arc to evenly cover the visible range of the crack. Let the initial azimuth angles of the acquisition points satisfy:

$$\phi_S < \phi_0(0) < \phi_1(0) < \phi_2(0) < \cdots < \phi_n(0) < \phi_E, \quad (3 \le n \in \mathbb{N}^*) \tag{15a}$$

where ϕ_S and ϕ_E denote the start and end boundaries of the arc, respectively.

In this section, for illustration purposes, three angle ranges are selected: the first quadrant $[0, \frac{\pi}{2}]$, the second quadrant $[\frac{\pi}{2}, \pi]$, and the combined first and second quadrants $[0, \pi]$, which correspond to the three test scenarios, Test 1, Test 2, and Test 3 shown in Figure 12. The distribution of the multi-viewpoint acquisition points along the semicircular arc with $\phi_S = 0$ and $\phi_E = \pi$ is illustrated in Figure 13.

The discrete evolution equation for the deployment of multi-viewpoint acquisition points is defined as follows:

$$\phi_i(k+1) = \phi_i(k) + u_i(k) \tag{16a}$$

$$u_i(k) = -t \cdot \frac{\partial T(k)}{\partial \phi_i(k)} \tag{17a}$$

where k denotes the iteration step, $\phi_i(k)$ is the azimuth angle of the underwater robot at acquisition point P_i in step k, and t is the learning rate, set to 0.1. $T(k)$ represents the cost function, and $u_i(k)$ is the negative gradient-based control law used to update the robot's acquisition point based on $T(k)$ (detailed later in this section).

Let P_{i+1} denote the left neighbor of acquisition point P_i, and let L_i denote the angular distance (arc length) between P_i and P_{i+1} along the clockwise direction. This distance is computed as

$$L_i(k) = \phi_{i+1}(k) - \phi_i(k), \quad 0 \le i \le n-1 \tag{18a}$$

To ensure both uniform distribution and sufficient angular coverage of the embankment crack from multiple viewpoints, we introduce two constraint metrics: consistency, quantified by the sum of absolute differences between neighboring distances, and coverage, defined by the deviation of the minimum neighbor distance from the ideal uniform spacing.

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

Accordingly, the cost function for the entire multi-viewpoint distribution is defined as:

$$T(k) = \alpha \cdot \sum_{i=0}^{n-1} |L_{i+1}(k) - L_i(k)| + \beta \left(\frac{\phi_E - \phi_S}{n-1} - \min\left(L_i(k)\right) \right) \tag{19a}$$

where $T(k) \geq 0$, and α and β are weighting coefficients corresponding to distribution uniformity and angular coverage, respectively. In this study, both are set to 0.5. When $T(k)$ tends to 0, the viewpoints are optimally and uniformly distributed along the semicircular arc and achieve complete angular coverage. Taking into account environmental constraints, a practical convergence threshold of $T(k) = 0.2$ is used to terminate the iteration and determine the final configuration of the acquisition point.

Meanwhile, the underwater robot is equipped with a high-precision inertial measurement unit, a sonar-based localization and obstacle-avoidance module to realize safe and precise movement from one point to the next. Navigation control is divided into two steps: Path generation and Position revision.

Path generation: The current acquisition angle is ϕ_i and the next is ϕ_{i+1}, so the planar coordinates at P_i are given by $x_i = r\cos\phi_i$ and $y_i = r\sin\phi_i$. We divide the angular distance $L_i = \phi_{i+1} - \phi_i$ into $I = L_i/\sigma$ segments, where σ is the robot's minimum response step. Discrete trajectory points are then:

$$O_j = \left((1-\mu_j)x_i + \mu_j\, x_{i+1},\ (1-\mu_j)y_i + \mu_j\, y_{i+1}\right) \tag{20a}$$

For $j = 0, 1, \ldots, I$ with $\mu_j = j/I$, yielding a smooth sequence $\{O_0, \ldots, O_I\}$ from $O_0 = P_i$ to $O_I = P_{i+1}$. Between each O_j and O_{j+1}, the robot advances at constant speed.

Position revision: After reaching $O_I = (x_{i+1}, y_{i+1})$, the sonar localization provides the actual position $(x_{\text{real}}, y_{\text{real}})$ and the errors :

$$e_x = x_{i+1} - x_{\text{real}}, \quad e_y = y_{i+1} - y_{\text{real}} \tag{21a}$$

The correction step is $\Delta\mathbf{p} = K_p\,(e_x, e_y)^{\mathsf{T}}$ with $K_p = 5$. If $\sqrt{e_x^2 + e_y^2} > \varepsilon_p$ (ε_p denotes the position error threshold, which is set to 0.05 m), the robot iteratively applies $\Delta\mathbf{p}$ until $\sqrt{e_x^2 + e_y^2} \leq \varepsilon_p$. Once positional accuracy is met, the robot adjusts its attitude so the camera's optical axis points precisely at the crack center and completes image acquisition.

Finally, the final multiview dam crack images—Image1, Image2, and Image3—are fused and processed by pretrained weight files deployed on the underwater robot to recognize dam cracks. The entire recognition pipeline operates at 25 fps, satisfying the real-time detection requirements. The recognition results are subsequently transmitted to an offshore mobile computing device via a wired connection.

5.3.3 Application validation results and analysis

Figure 14 illustrates the real-time segmentation results produced by the proposed algorithm for three test scenarios. Figure 14(d) presents the crack labels corresponding to Figure 14(a), while Figure 14(e) displays the predicted segmentation masks generated by the proposed algorithm for Figure 14(a). Figure 14(f) shows the superimposed results of these predicted masks on the original image in Figure 14(a).

In Test 1, the prediction mask generated by the proposed algorithm is compared with the crack labels corresponding to Figure 14(a) for images containing more complex crack shapes. Although minor under-segmentation occurs in regions with complex crack shapes, the overall crack structure is segmented with high accuracy, particularly in masked regions where cracks are well recognized. In Test 2, the algorithm demonstrates strong robustness by achieving accurate segmentation of fine cracks, highlighting its capability to handle detailed complexities effectively. In Test 3, where cracks are significantly obscured by aquatic plants, the predicted segmentation masks exhibit minor noise (i.e., some non-crack areas are misclassified as cracks). Nonetheless, the final segmentation results effectively capture the overall morphology of the cracks.

Experimental results demonstrate that the algorithm proposed in this paper exhibits strong accuracy and robustness in segmenting underwater occluded dam cracks in practical application scenarios.

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

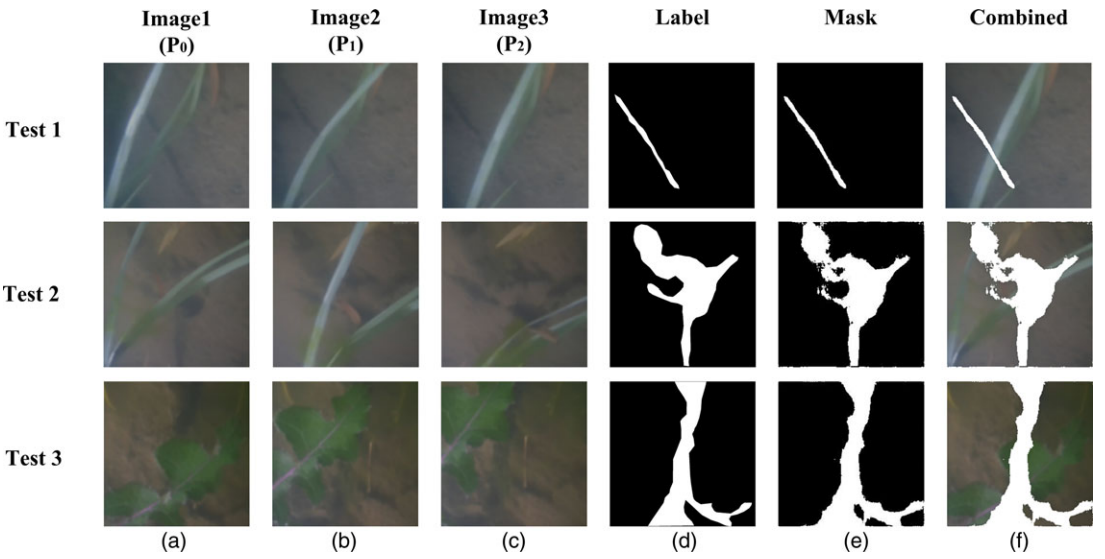

Figure 14. *Illustration of segmentation results of the proposed algorithm in three test scenarios.*

6. Conclusions

In underwater dam crack detection, cracks are often obscured by aquatic plants, and underwater turbulence causes nonuniform diffusion of suspended sediments, resulting in varying degrees of feature submergence across different viewpoints. To address these challenges, we propose a MVFD-Net for occluded underwater dam crack detection,MVFD-Net. First, the FRI-Encoder integrates multi-scale local features extracted by a CNN with global representations from a transformer encoder and performs feature reconstruction fusion at the encoder output to suppress non-uniform scattering noise. Second, we introduce the MGAF module, which employs a pyramid structure to perform gated feature fusion between the encoder and decoder, thereby recovering lost details. Finally, within the segmentation network, we design an MVFF module to enhance crack integrity and recognition accuracy by incorporating features from additional viewpoints to repair occluded regions. We validate MVFD-Net on a self-constructed dataset, demonstrating its superior generalization and significantly improved segmentation performance under aquatic plant occlusion. Future work will focus on optimizing the device functions that accompany the algorithm and developing quantitative crack analysis methods to provide a more scientific basis for crack repair.

Author contributions. Yukai Wu conceived the study, performed the core experiments, and drafted the manuscript; Xiaochen Qin supervised the experimental validation and optimized the article content; Lei Cai guided the direction of the paper, provided the experimental equipment platform, and finalized the academic interpretation.

Funding. This work was supported by Henan Provincial focus on research and development Project (231111220700).

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The authors declare no conflicts of interest exist.

References

[1] L. Mucolli, S. Krupinski, F. Maurelli, S. A. Mehdi and S. Mazhar, "Detecting cracks in underwater concrete structures: An unsupervised learning approach based on local feature clustering," *MTS/IEEE Seattle* **23**(1), 1–8 (2019).

[2] D. Chen, B. Huang and F. Kang, "A review of detection technologies for underwater cracks on concrete dam surfaces," *Appl. Sci.* **13**(6), 3564–3578 (2023).

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

[3] M. Jian, N. Yang, C. Tao, H. Zhi and H. Luo, "Underwater object detection and datasets: A survey," *Intelligent Marine Technology and Systems* **2**(1), 9–22 (2024).

[4] Y. Wu, S. Li, J. Zhang, Y. Li, Y. Li and Y. Zhang, "Dual attention transformer network for pixel-level concrete crack segmentation considering camera placement," *Autom. Constr.* **157**(3), 105166–105178 (2024).

[5] D. A. Beyene, D. Q. Tran, M. B. Maru, T. Kim and S. Park, "Unsupervised domain adaptation-based crack segmentation using transformer network," *J. Build. Eng.* **15**(2), 107889–107990 (2023).

[6] X. Zhang, C. Bai and K. Kpalma, "OMCBIR: Offline mobile content-based image retrieval with lightweight CNN optimization," *Displays* **76**(6), 102355–102368 (2021).

[7] Z. Yang, L. Zhu, Y. Wu and Y. Yang, "Gated Channel Transformation for Visual Recognition," 33st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, Seattle, WA, USA, 2020), pp. 11791–11800.

[8] J. Wang, Z. Zeng, P. K. Sharma, O. Alfarraj, A. Tolba, J. Zhang and L. Wang, "Dual-path network combining CNN and transformer for pavement crack segmentation," *Autom. Constr.* **158**(64), 105217–105239 (2024).

[9] Z. Zhu, S. Huang, J. Xie, Y. Meng, C. Wang and F. Zhou, "A refined robotic grasp detection network based on coarse-to-fine feature and residual attention," *Robotica.* **43**(2), 1–18 (2024).

[10] L. Ke, Y. W. Tai and C. K. Tang, "Deep Occlusion-Aware Instance Segmentation with Overlapping Bilayers," 34st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, Nashville, TN, USA, 2021), pp. 4018–4027

[11] X. Yuan, A. Kortylewski, Y. Sun and A. Yuille, "Robust Instance Segmentation Through Reasoning About Multi-Object Occlusion," 34st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, Nashville, TN, USA, 2021), pp. 11141–11150.

[12] T. Zhang, J. Dai, W. Song, R. Zhao and B. Zhang, "OSLPNet: A neural network model for street lamp post extraction from street view imagery," *Expert Syst. Appl.* **231**(25), 120764–120782 (2023).

[13] H. Gan, F. Menegon, A. Sun, A. Scollo, Q. Jiang, Y. Xue and T. Norton, "Peeking into the unseen: Occlusion-resistant segmentation for preweaning piglets under crushing events," *Comput. Electron. Agric.* **219**(23), 108683–108693 (2024).

[14] H. Wang, S. Zhu, L. Chen, Y. Li and Y. Cai, "OccludedInst: An efficient instance segmentation network for automatic driving occlusion scenes," *IEEE Trans. Emerg. Top. Comput. Intell.* **10**(12), 3414948–3414962 (2024).

[15] X. Yan, F. Wang, W. Liu, Y. Yu, S. He and J. Pan, "Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery," 32st IEEE/CVF International Conference on Computer Vision (ICCV) (IEEE, Seoul, Korea, 2019), pp. 7617–7626.

[16] Y. Guo, H. Yu, S. Xie, L. Ma, X. Cao and X. Luo, "DSCA: A dual semantic correlation alignment method for domain adaptation object detection," *Pattern Recognit* **150**(12), 110329–110345 (2024).

[17] N. Dong, S. Yan, H. Tang, J. Tang and L. Zhang, "Multi-view information integration and propagation for occluded person re-identification," *Inf. Fusion* **104**(22), 102201–102221 (2024).

[18] Z. Xia, M. Liao, S. Di, Y. Zhao, W. Liang and N. Xiong, "Automatic liver segmentation from CT volumes based on multiview information fusion and condition random fields," *Opt. Laser Technol.* **179**(65), 111298–111320 (2024).

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair and Y. Bengio, "Generative Adversarial Nets," 27th International Conference on Neural Information Processing Systems-Volume 2 (NIPS'14) (NIPS, Cambridge, MA, USA, 2014), pp. 2672–2680.

[20] P. Isola, J. Y. Zhu, T. Zhou and A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 30st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, Honolulu, HI, USA, 2017), pp. 5967–5976.

[21] T. Zhou, S. Tulsiani, W. Sun, J. Malik and A. Efros, "View Synthesis by Appearance Flow," Computer Vision–ECCV 2016 (Springer, Amsterdam, The Netherlands, 2016), pp. 286–301.

[22] C. Yang, K. Wang, Y. Wang, Q. Dou, X. Yang and W. Shen, "Efficient deformable tissue reconstruction via orthogonal neural plane," *IEEE Trans. Med. Imaging* **43**(11), 3211–3223 (2024).

[23] C. B. Choy, D. Xu, J. Gwak, K. Chen and S. Savarese, "3D-R2N2: A Unified Approach for Single and Multi-View 3D Object Reconstruction," Computer Vision–ECCV 2016 (Springer, Amsterdam, The Netherlands, 2016), pp. 628–644.

[24] J. Yang, W. S. Zheng, Q. Yang, Y. C. Chen and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," 33st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, Seattle, WA, USA, 2022), pp. 3289–3299.

[25] X. Zhang, Z. Zheng, D. Gao, B. Zhang, Y. Yang and T. Chua, "Multi-view consistent generative adversarial networks for compositional 3D-aware image synthesis," *Int. J. Comput. Vis.* **131**(26), 2219–2242 (2023).

[26] S. Arooj, S. Altaf, S. Ahmad, H. Mahmoud and A. S. N. Mohamed, "Enhancing sign language recognition using CNN and SIFT: A case study on Pakistan sign language," *J. King Saud Univ.-Comput. Inf. Sci.* **36**(23), 101934–101962 (2024).

[27] H. Xu, J. Yuan and J. Ma, "MURF: Mutually reinforcing multi-modal image registration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(23), 12148–12166 (2023).

[28] Y. Hua, X. Huang, H. Li and X. Cao, "Mobile robot tracking control based on lightweight network," *Robotica*, **42**(2), 1–19 (2025).

[29] J. Pan, J. Jia and L. Cai, "Global enhancement network underwater archaeology scene parsing method," *Robotica* **39**(12), 3541–3564 (2023).

[30] S. Diao, J. Su, C. Yang, W. Zhu, D. Xiang, X. Chen and F. Shi, "Classification and segmentation of OCT images for age-related macular degeneration based on dual guidance networks," *Biomed. Signal Process. Control* **84**(35), 104810–104830 (2023).

[31] Y. Liu, H. Wang, Z. Chen, K. Huangliang and H. Zhang, "TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation," *Knowl.-Based Syst.* **256**(23), 109872–109889 (2022).

https://doi.org/10.1017/S0263574725101835 Published online by Cambridge University Press

[32] H. Liu, J. Yang, X. Miao, C. Mertz and H. Kong, "Crackformer network for pavement crack segmentation," *IEEE Trans. Intell. Transp. Syst.* **24**(14), 9240–9252 (2023).
[33] K. Li, J. Yang, S. Ma, B. Wang and S. Wang, "Rethinking lightweight convolutional neural networks for efficient and high-quality pavement crack detection," *IEEE Trans. Intell. Transp. Syst.* **23**(16), 237–250 (2024).
[34] V. Badrinarayanan, A. Kendall and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017).
[35] Z. He, L. Cao, J. Luo, X. Xu, J. Tang, J. Xu and Z. Chen, "UISS-Net: Underwater image semantic segmentation network for improving boundary segmentation accuracy of underwater images," *Aquac. Int.* **32**(12), 5625–5638 (2024).

Cite this article: Y. Wu, X. Qin and L. Cai (2025). "MVFD-Net: multi-view fusion detection network for occluded underwater dam cracks", Robotica **43**, 2480–2503. https://doi.org/10.1017/S0263574725101835