IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 477-489
Journal homepage: http://iieta.org/journals/ts

Multimodal Image Registration and Intelligent Diagnosis of Building Facade Debonding Defects via Visible–Infrared Thermal Fusion

Bing Zhang*, Yonghua Wang, Yunhua Lu

School of Civil Engineering, Tangshan University, Tangshan 063000, China

Corresponding Author Email: zhangbing_joy2010@163.com

https://doi.org/10.18280/ts.430134

Received: 3 August 2025
Revised: 16 December 2025
Accepted: 5 January 2026
Available online: 28 February 2026

Keywords:
multimodal image registration, infrared thermal imaging, building facade debonding, feature mutual guidance, adaptive fusion segmentation

ABSTRACT

Debonding defects in building facades are highly concealed and difficult to detect using traditional inspection methods, which significantly affects the structural safety and durability of buildings. Visible-light images provide rich texture and structural information, while infrared thermal imaging can accurately capture temperature anomalies in debonded regions. The fusion of these two modalities therefore represents a key technological pathway for efficient detection of facade debonding defects. However, existing multimodal detection methods often suffer from insufficient cross-modal registration accuracy, inadequate task-oriented feature fusion, and blurred segmentation boundaries of debonded regions, making it difficult to meet the requirements of precise diagnosis in practical engineering applications. To address these challenges, this study proposes a two-stage cascaded framework based on a registration-then-fusion segmentation strategy. Specifically, a dual-branch feature mutual-guidance registration network and a multi-scale adaptive fusion segmentation network are designed to fully exploit the complementary information of the two modalities, enabling pixel-level accurate identification and localization of facade debonding defects. To evaluate the effectiveness and advancement of the proposed method, extensive experiments are conducted on a self-constructed large-scale Unmanned Aerial Vehicle (UAV)-based dual-modal dataset. The experimental results demonstrate that the proposed approach significantly outperforms existing mainstream methods in both registration and segmentation performance, exhibiting strong engineering applicability and environmental robustness. The proposed framework provides a novel technical solution and theoretical support for intelligent inspection of building facade debonding defects.

1. INTRODUCTION

Building facade debonding is a common quality defect in construction engineering [1-3]. Due to its strong concealment, long-term existence of this defect can easily lead to wall cracking and detachment, which not only affects the integrity of the building appearance, but also weakens the structural bearing capacity of the wall, threatening the structural safety and durability of the building. At present, traditional defect detection methods are mainly based on manual tapping, which relies on the experience of inspectors and has limitations such as low efficiency, strong subjectivity, and limited detection range [4, 5]. Among single-modal image detection methods, visible-light images can provide rich wall texture and structural information, but it is difficult to identify concealed debonding defects [6, 7]. Infrared thermal imaging can accurately capture the temperature difference between debonded regions and normal regions, but it lacks effective spatial structural reference, making it difficult to achieve precise defect localization [8, 9]. Therefore, multimodal image fusion of visible-light and infrared thermal imaging has become a key technical pathway for efficient and accurate detection of debonding defects [10]. However, due to the intrinsic differences between the two modal images in imaging mechanism, spectral range, and resolution, the accuracy of cross-modal image registration and the effectiveness of multimodal feature fusion have become the core bottlenecks restricting precise diagnosis of debonding defects.

In the field of cross-modal image registration, traditional methods have insufficient robustness for heterogeneous image matching and are difficult to adapt to complex scene variations of building facades. Although existing deep learning-based registration methods have improved efficiency to some extent [11, 12], they lack effective feature interaction mechanisms between modalities, making it difficult to simultaneously consider registration accuracy and adaptability to complex deformation, and therefore unable to meet the pixel-level registration requirements of debonding detection. In the field of debonding defect segmentation, existing multimodal fusion methods mostly adopt simple feature concatenation approaches [13, 14], which fail to fully exploit the complementary characteristics of the two modal images. As a result, the segmentation accuracy for subtle debonding regions and defect boundaries is relatively low, and the resistance to noise interference is weak, making them difficult to adapt to complex wall environments in practical engineering.

To address the above problems, this study conducts research on multimodal image registration and intelligent diagnosis of

building facade debonding defects. The main innovations and contributions are as follows. First, a dual-branch feature mutual-guidance registration network is proposed, and a cross-modal attention interaction mechanism is designed to achieve high-precision non-rigid registration of heterogeneous images, effectively solving the core problem of insufficient registration accuracy in large-scale deformation scenarios. Second, a multi-scale adaptive fusion segmentation network is designed, introducing a channel–spatial attention fusion module, combined with a lightweight Transformer and edge constraints, to achieve accurate segmentation and boundary optimization of debonding regions, thereby improving the recognition capability for subtle defects. Third, a registration loss function that considers both visual similarity and physical consistency is constructed, and a temperature consistency constraint is introduced to ensure that the registration process does not destroy the physical distribution of infrared temperature anomalies, thereby improving the reliability of the diagnosis results. Fourth, a large-scale Unmanned Aerial Vehicle (UAV) dual-modal debonding dataset is constructed, covering different seasons, different illumination conditions, and different defect types, providing reliable support for method validation and enhancing the engineering applicability of the proposed method.

The remainder of this paper is organized as follows. Section 2 describes in detail the overall architecture of the proposed intelligent diagnosis method and the technical details of each module. Section 3 verifies the effectiveness and advancement of the method through systematic experiments. Section 4 discusses the advantages, limitations, and future research directions of the method. Section 5 summarizes the research results and contributions of this paper.

2. METHOD

2.1 Overall architecture of the method

The intelligent diagnosis method for building facade debonding defects proposed in this paper adopts a two-stage cascaded architecture of registration first and then fusion segmentation. The core objective is to achieve pixel-level accurate identification and localization of debonding regions through high-precision cross-modal image alignment and deep feature fusion. The method contains three core modules: preprocessing enhancement, cross-modal registration, and fusion segmentation. These modules work collaboratively and progressively, forming a complete end-to-end diagnosis system. The preprocessing enhancement module addresses the imaging differences between visible-light and infrared thermal imaging. Through targeted processing, it improves image quality, effectively suppresses noise, and highlights key features, providing reliable inputs for subsequent registration and segmentation tasks. The cross-modal registration module achieves pixel-level accurate alignment of dual-modal images, solving the core problem of feature mismatch in heterogeneous images and laying the foundation for effective multimodal feature fusion. The fusion segmentation module fully exploits the complementary information of the aligned dual-modal images, integrating the texture and structural details of visible-light images with the temperature anomaly features of infrared thermal imaging, thereby completing accurate segmentation and boundary optimization of debonding defects. The entire architecture is trained and optimized in an end-to-end manner to ensure the collaborative consistency among modules, significantly improving the accuracy, robustness, and engineering applicability of the diagnosis method. Figure 1 shows the overall architecture of the intelligent diagnosis method for facade debonding defects.

2.2 Multimodal image preprocessing and enhancement

The single-channel temperature matrix of infrared thermal imaging is difficult for convolutional neural networks to effectively extract features from, and the visual discriminability of temperature differences is relatively low. Traditional pseudo-color mapping adopts fixed mapping rules and cannot adapt to the temperature distribution characteristics in different scenarios [15, 16]. To address this problem, an adaptive pseudo-color mapping strategy is proposed. According to the statistical characteristics of temperature in infrared thermal images, the mapping parameters are dynamically adjusted, converting the single-channel temperature matrix into three-channel visual features. The core mapping formula is:

$$I_c(x,y)=\begin{cases}\dfrac{T(x,y)-T_{\min}}{T_{\mathrm{mid}}-T_{\min}} & T_{\min}\le T(x,y)<T_{\mathrm{mid}}\\ 1 & T_{\mathrm{mid}}\le T(x,y)<T_{\max}\\ \dfrac{T_{\max}-T(x,y)}{T_{\max}-T_{\mathrm{mid}}} & T(x,y)\ge T_{\max}\end{cases} \tag{1}$$

where, $I_c(x,y)$ is the pixel value of the mapped three-channel image, $c\in\{R,G,B\}$ correspond to the red, green, and blue channels respectively, $T(x,y)$ is the temperature value at pixel (x,y) in the infrared thermal image, and T_{min}, T_{mid}, and T_{max} are the minimum temperature, middle temperature, and maximum temperature of the image, respectively, which are adaptively determined by statistical analysis of the temperature histogram of the image. This strategy not only enhances the perception ability of human vision for temperature anomalies, but also converts single temperature information into multi-channel visual features, which is well adapted to the feature extraction requirements of convolutional neural networks. At the same time, to address the problem of temperature noise in infrared thermal images, an improved bilateral filtering algorithm is proposed. By introducing temperature gradient weights to adjust the filter kernel coefficients, the filtering formula is:

$$I_{infra}(x,y)=\frac{1}{K(x,y)}\sum_{i,j} G_{\sigma_s}(|(x,y)-(i,j)|)\cdot G_{\sigma_r}(|T(x,y)-T(i,j)|)\cdot w_g\cdot T(i,j) \tag{2}$$

where, $K(x,y)$ is the normalization coefficient, G_{σ_s} and G_{σ_r} are the Gaussian functions in the spatial domain and gray-level domain respectively, and w_g is the temperature gradient weight, which is calculated from the temperature gradient between pixel (x,y) and neighboring pixels. The larger the temperature gradient, the closer w_g is to 1. In this way, while smoothing noise in homogeneous regions, the temperature anomaly edges are preserved to the greatest extent, solving the edge blurring problem caused by traditional bilateral filtering.

Visible-light images are easily affected by illumination angle, shadows, and other factors, resulting in uneven illumination, which leads to bias in texture structure feature extraction and affects subsequent registration accuracy [17]. Based on Retinex theory [18], an improved illumination balancing algorithm is proposed. It breaks the limitation that

traditional algorithms only adjust overall illumination, and combines the structural prior obtained by Canny edge extraction to guide illumination correction. First, the visible-light image is decomposed into a reflectance component $R(x,y)$ and an illumination component $L(x,y)$, with the decomposition formula:

$$I_{vis}(x,y)=R(x,y)\cdot L(x,y) \tag{3}$$

For the decomposed illumination component, the structural prior mask $M(x,y)$ obtained by Canny edge detection is introduced. In the mask, the pixel value of edge regions is 1 and that of non-edge regions is 0. Adaptive illumination compression is performed using the following formula:

$$L'(x,y)=L(x,y)^{\alpha M(x,y)+\beta(1-M(x,y))} \tag{4}$$

where, α and β are the illumination compression coefficients for edge regions and non-edge regions respectively, $\alpha\in(0,1)$ and $\beta\in(0,\alpha)$, which are determined through adaptive iterative optimization. This ensures that the illumination compression in edge regions is smaller than that in non-edge regions, eliminating uneven illumination while avoiding distortion of edge structures. The corrected reflectance component $R'(x,y)=I_{vis}(x,y)/L'(x,y)$ is processed by gray normalization and then used as the input for the subsequent registration task, providing accurate structural reference for cross-modal registration.

Through targeted innovative processing for the two modal images, the preprocessing and enhancement module effectively suppresses noise interference and solves problems such as uneven illumination and blurred temperature anomaly edges. It significantly improves image quality and the distinguishability of key features, laying a solid foundation for the efficient operation of subsequent cross-modal registration and fusion segmentation modules, and ensuring the accuracy and robustness of the entire diagnosis system.

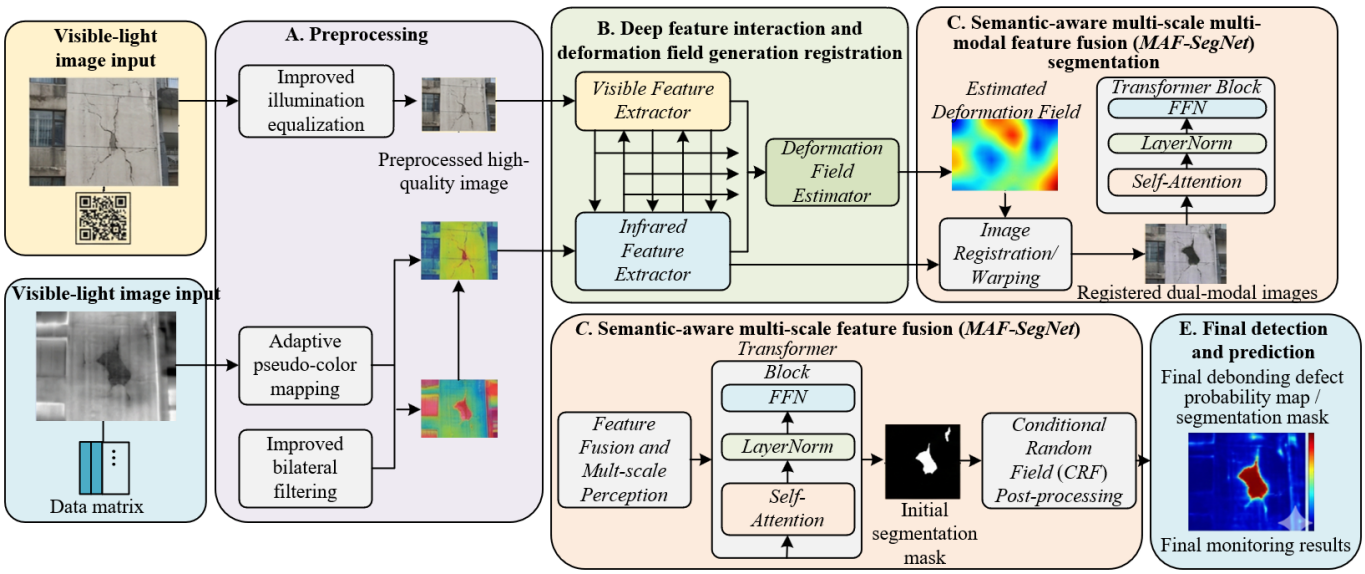

Figure 1. Overall architecture of the intelligent diagnosis method for facade debonding defects

2.3 End-to-end registration network based on cross-modal feature mutual guidance

Dual-branch Feature Mutual-Guided Registration Network (DFMGR-Net) adopts an encoder-decoder end-to-end architecture. The core innovation lies in the collaborative design of dual-branch feature extraction and hierarchical cross-modal interaction, achieving precise extraction and mutual enhancement of heterogeneous image features. Figure 2 shows the architecture of the dual-branch feature mutual-guidance registration network. As shown in the figure, the encoder is divided into two parallel branches for visible-light and infrared modalities. The differential network design adapts to the feature characteristics of the two modalities while ensuring alignment of feature hierarchies, providing the basis for subsequent attention interaction. The visible-light branch uses a lightweight 50-layer Residual Network (ResNet50) as the backbone network, removing the fully connected layers and retaining all residual modules. Multi-stage residual learning extracts multi-scale texture and structural features, and the output feature map is denoted as $F_{vis}^{l}\in\mathrm{R}^{C_l\times H_l\times W_l}$, where l represents the encoder feature hierarchy, and C_l, H_l, W_l are the channel number, height, and width of the corresponding hierarchy. The infrared branch is a customized lightweight convolutional neural network. Considering the low-texture characteristic of temperature distribution features, it adopts 3 × 3 small convolution kernels and sparse channels, setting feature hierarchies fully matched with the visible-light branch. It outputs multi-scale temperature distribution feature maps $F_{inf}^{l}\in\mathrm{R}^{C_l\times H_l\times W_l}$, ensuring temperature feature extraction accuracy while reducing computation by more than 40%. Features from both branches at each hierarchy are synchronously input into the cross-modal attention interaction module. The enhanced features are channel-concatenated and sent to the decoder's deformation field estimation module, forming an "feature extraction – interaction enhancement – deformation estimation" end-to-end registration pipeline.

The core objective of the cross-modal attention interaction module is to establish a bidirectional guidance mechanism between visible-light structural features and infrared temperature features, solving the problem of weak correlation and high difficulty in aligning heterogeneous modality features. This module is embedded in each feature hierarchy of the encoder to achieve multi-scale feature mutual enhancement, ensuring that both low-level details and high-level semantics can complete effective cross-modal

information transfer. The core logic is to constrain the spatial positioning of infrared features by the spatial structure information of visible-light images, while using infrared temperature response information to select key feature regions in visible-light images, forming a complementary enhancement closed loop between modalities.

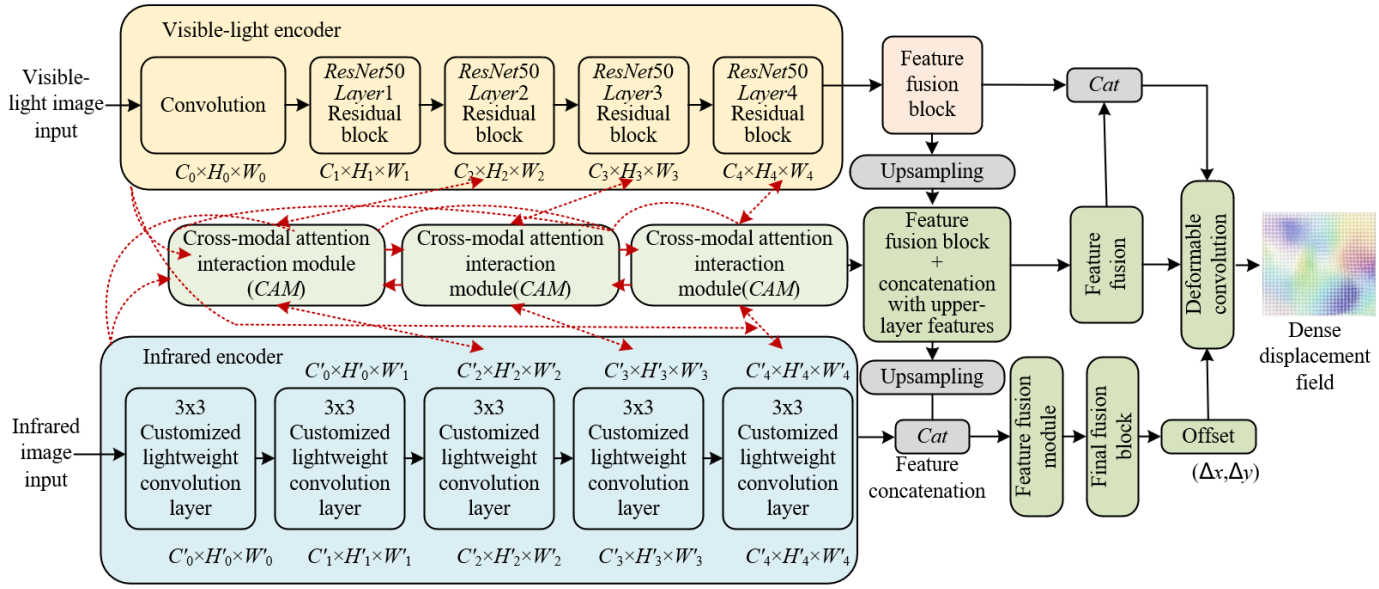

Figure 2. Dual-branch feature mutual-guidance registration network architecture

The core implementation of the module is the joint computation of channel attention and spatial attention, achieving fine feature selection through the collaboration of the two weights. First, the input feature maps of the dual branches are channel-aligned, then channel attention weights and spatial attention weights are computed separately. The channel attention weights are calculated based on global feature statistics. For visible-light features F^l_{vis} and infrared features F^l_{inf}, global average pooling is performed to obtain channel-level feature vectors $v^l_{vis}\in \mathrm{R}^{C_l}$ and $v^l_{inf}\in \mathrm{R}^{C_l}$, which are mapped by shared two-layer fully connected layers. Finally, Sigmoid activation produces the channel attention weights:

$$\begin{aligned} W^l_{c,vis}&=\sigma\left(W_2\cdot \mathrm{ReLU}\left(W_1\cdot v^l_{inf}\right)\right),\\ W^l_{c,inf}&=\sigma\left(W_2\cdot \mathrm{ReLU}\left(W_1\cdot v^l_{vis}\right)\right) \end{aligned} \tag{5}$$

where, W_1 and W_2 are fully connected layer weights, and σ is the Sigmoid activation function. The spatial attention weights focus on local region features. The aligned dual-modal feature maps are channel-concatenated, reduced in dimension by a 1×1 convolution, then spatial correlations are extracted via a 3×3 depth convolution, followed by Sigmoid activation to generate spatial weight maps:

$$\begin{aligned} W^l_{s,vis}&=\sigma\left(\mathrm{Conv}_{3\times 3}\left(\mathrm{Conv}_{1\times 1}\left(F^l_{vis}\oplus F^l_{inf}\right)\right)\right),\\ W^l_{s,inf}&=W^l_{s,vis} \end{aligned} \tag{6}$$

where, $\oplus$ denotes channel concatenation and $Conv_{k\times k}$ represents convolution with kernel size $k\times k$.

Bidirectional feature enhancement is realized by element-wise multiplication of the attention weights with the original feature maps. The joint channel and spatial attention weights are: $W^l_{vis}=W^l_{c,vis}\otimes W^l_{s,vis}$, $W^l_{inf}=W^l_{c,inf}\otimes W^l_{s,inf}$, where $\otimes$ denotes element-wise multiplication. The final enhanced features are: $F^l_{vis}=F^l_{vis}\cdot W^l_{vis}$, $F^l_{inf}=F^l_{inf}\cdot W^l_{inf}$.

This process allows the infrared branch to receive the high-precision spatial structure constraints from visible-light images, precisely locating temperature anomaly positions, while the visible-light branch focuses on thermal anomaly regions indicated by infrared features, filtering irrelevant wall texture noise. The enhanced dual-modal features have improved heterogeneity correlation, providing high-quality input for the decoder's dense deformation field estimation and significantly improving cross-modal registration accuracy and robustness.

The core function of the dense deformation field estimation module is to generate a high-precision dense displacement field that adapts to the complex spatial transformations of building facades, solving the core problem that traditional convolution cannot model non-rigid deformation. This module adopts deformable convolution to construct a hierarchical deformation estimation structure. Extra-learned offsets guide the convolution kernel for adaptive non-grid point sampling, flexibly capturing perspective deformation, subtle wall deformation, and local geometric distortions. The output of deformable convolution is calculated as:

$$y(x,y)=\sum_{k=1}^{K} w_k\cdot x\left(x+\Delta x_k, y+\Delta y_k\right) \tag{7}$$

where, $y(x,y)$ is the output feature at pixel (x,y), K is the number of convolution sampling points, w_k is the convolution weight of the k-th sample point, and Δx_k, Δy_k are the extra learned offsets in [-1,1], adaptively learned from the concatenated enhanced dual-modal features. The offsets allow convolution kernels to dynamically adjust sampling positions according to local image geometry, effectively adapting to non-rigid deformation of building facades without preset grids. The enhanced dual-modal features are processed through 4 layers of deformable convolution hierarchically, gradually improving deformation modeling capability. The module finally outputs a dense displacement field

$\Delta(x,y)=(\Delta x(x,y),\Delta y(x,y))$ of the same size as the input images, achieving pixel-level accurate alignment between visible-light and infrared images, providing a reliable spatial alignment foundation for subsequent multimodal feature fusion.

To ensure that the registration results satisfy both visual structural alignment and the physical distribution of infrared temperature anomalies, an innovative loss function considering visual similarity and physical consistency is designed. Through multi-loss collaborative constraints, registration accuracy and reliability are improved, differing from existing loss designs that only focus on visual matching. The total loss function is a weighted sum of multi-scale structural similarity loss, mutual information loss, and temperature consistency loss:

$$L_{total}=\lambda_1 L_{MS\text{-}SSIM}+\lambda_2 L_{MI}+\lambda_3 L_{TC} \tag{8}$$

where, λ_1, λ_2, and λ_3 in [0,1] are weights for each loss component, determined via adaptive iterative optimization, ensuring coordinated effect of all loss components.

The multi-scale structural similarity loss constrains the structural alignment of registered images, reducing structural distortion caused by modality heterogeneity. It is computed based on multi-scale image decomposition, comparing the luminance, contrast, and structure of registered and reference images layer by layer:

$$L_{MS\text{-}SSIM}=1-\prod_{i=1}^{N} SSIM_i(I_{reg},I_{ref}) \tag{9}$$

where, N is the number of decomposition scales (set to 4 in this paper), $SSIMi$ is the structural similarity coefficient at scale i, I_{reg} is the registered image, and I_{ref} is the reference image. The mutual information loss measures the statistical dependency between registered cross-modal images, quantifying fusion consistency:

$$L_{MI}=-I(I_{reg,vis},I_{reg,infra})=-\sum_u\sum_v p(u,v)\log\frac{p(u,v)}{p(u)p(v)} \tag{10}$$

where, $I(\ ,\ )$ is mutual information, $p(u,v)$ is the joint probability density of registered visible $I_{reg,vis}$ and infrared $I_{reg,infra}$ images, and $p(u)$, $p(v)$ are their marginal probabilities.

The temperature consistency loss is the core innovative component, constraining the deformation field to not distort the physical distribution of infrared temperature anomalies, ensuring that the registered infrared image accurately reflects the temperature characteristics of debonding regions. It is computed by the difference in temperature gradients before and after registration:

$$L_{TC}=\frac{1}{H\times W}\sum_{x=1}^{H}\sum_{y=1}^{W}|\nabla T_{reg}(x,y)-\nabla T_{ori}(x,y)| \tag{11}$$

where, H, W are the height and width of the infrared image, $\nabla T_{ori}(x,y)$ is the temperature gradient of the original infrared image at pixel (x,y), and $\nabla T(x,y)$ is the temperature gradient of the registered infrared image, computed by the Sobel operator. This loss effectively suppresses distortion of temperature anomaly regions by the deformation field, ensuring that temperature distribution features of debonding regions are preserved during registration, providing accurate temperature feature support for subsequent defect segmentation.

2.4 Multi-scale adaptive fusion and intelligent debonding defect segmentation

MAF-SegNet takes the registered visible-light and infrared dual-modal images as input, adopting a dual encoder-decoder architecture. The core innovation lies in hierarchical dual-modal feature fusion and detail preservation mechanisms, focusing on pixel-level precise segmentation and boundary optimization of debonding defects. The specific architecture is shown in Figure 3. The dual encoders are designed to match the feature hierarchies of the registration network, separately processing the two modality images to fully exploit complementary information: the visible-light encoder continues the lightweight Residual Network structure, focusing on extracting wall texture, edges, and other structural features, outputting multi-scale feature maps $F_{vis}^{s}\in\mathrm{R}^{C_s\times H_s\times W_s}$; the infrared encoder adopts a customized lightweight Convolutional Neural Network (CNN), focusing on capturing temperature anomaly-related features, outputting corresponding scale feature maps $F_{inf}^{s}\in\mathrm{R}^{C_s\times H_s\times W_s}$, where s is the feature scale, and C_s, H_s, W_s are the channel number, height, and width at the corresponding scale. The decoder uses a layer-by-layer upsampling structure. Through skip connections, the dual-modal fused features of each encoder scale are integrated with the corresponding decoder layer features, gradually restoring spatial resolution while effectively suppressing gradient vanishing, ensuring that features of subtle debonding regions and defect edges are not lost. The network finally outputs a debonding segmentation probability map with the same size as the input image, achieving precise defect localization and complete segmentation.

The channel-spatial attention fusion module is the core innovative component of MAF-SegNet. Its main objective is to solve the problems in traditional dual-modal feature concatenation, such as feature redundancy and the drowning of useful information. Through the collaborative effect of dual attention mechanisms, fine-grained fusion of multi-modal information is realized, enhancing the discriminative capability for debonding defects. The module schematic is shown in Figure 4. The module is embedded at every feature scale layer in the encoder and decoder, enabling precise feature selection and fusion at multiple scales, ensuring that both low-level detail features and high-level semantic features are effectively integrated, fully leveraging the complementary advantages of dual-modal images.

The channel attention mechanism is used to adaptively select the most critical feature channels for debonding diagnosis, achieving channel-level weight allocation based on global feature statistics. First, the dual-modal feature maps are concatenated along the channel dimension to obtain the initial fused feature $F_{cat}^{s}=F_{vis}^{s}\oplus F_{inf}^{s}$. Then, global average pooling is applied to extract the channel-level global feature vector $\boldsymbol{v}^{s}\in\mathrm{R}^{2C_s}$. The feature vector is mapped and dimension-reduced through two fully connected layers, and finally Sigmoid activation generates the channel attention weights W_c^s, calculated as:

$$W_c^s=\sigma(W_b\cdot\mathrm{ReLU}(W_a\cdot v^s)) \tag{12}$$

where, $W_a\in\mathrm{R}^{(2C_s)/4\times 2C_s}$, $W_b\in\mathrm{R}^{2C_s\times(2C_s)/4}$ are the weights of the

two fully connected layers, σ is the Sigmoid activation function. The weight vector is multiplied element-wise with the initial fused feature, effectively enhancing the texture, edge, and temperature-related features critical for debonding diagnosis while suppressing redundant channel interference.

The spatial attention mechanism focuses on potential debonding regions, strengthening the discriminative ability of local features and compensating for the channel attention mechanism's neglect of local spatial information. The channel-attention-weighted feature is first reduced in dimension by a 1×1 convolution, reducing computation while achieving preliminary channel information fusion. Then, a 3×3 depth convolution extracts local spatial correlation features, capturing spatial differences between debonding regions and surrounding background. Sigmoid activation generates the spatial attention weight map $\boldsymbol{W}_s^s \in \mathrm{R}^{1\times H_s\times W_s}$, calculated as:

$$W_s^s=\sigma\left(\mathrm{Conv}_{3\times3}\left(\mathrm{Conv}_{1\times1}\left(F_{cat}^s\cdot W_c^s\right)\right)\right) \tag{13}$$

where, $Conv_{k\times k}$ denotes convolution with kernel size $k\times k$. The final fine-grained fused feature is obtained through the collaborative effect of dual weights: $F_{fusion}^s=\left(F_{cat}^s\cdot W_c^s\right)\otimes W_s^s$, where $\otimes$ denotes element-wise multiplication. This fusion preserves the structural detail advantages of visible-light images while fully utilizing the temperature anomaly features of infrared images, effectively suppressing irrelevant background noise and significantly enhancing the discriminative capability of fused features for debonding defects, providing high-quality input for accurate decoder segmentation.

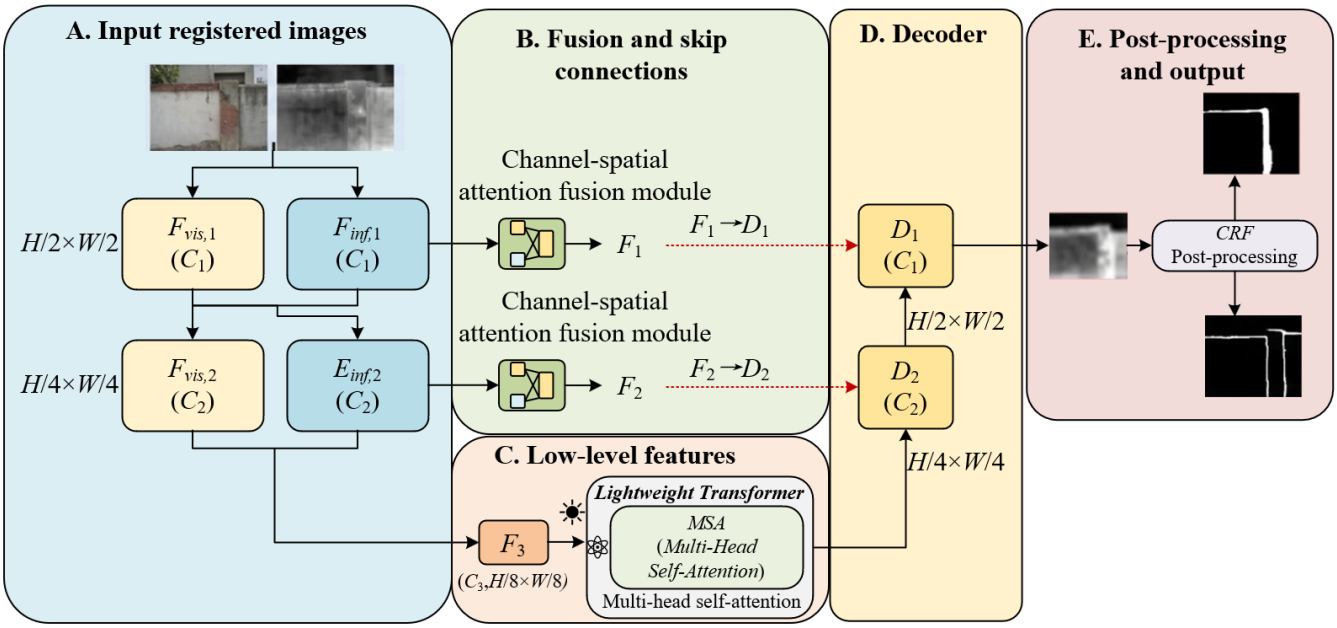

Figure 3. Multi-scale adaptive fusion segmentation network architecture

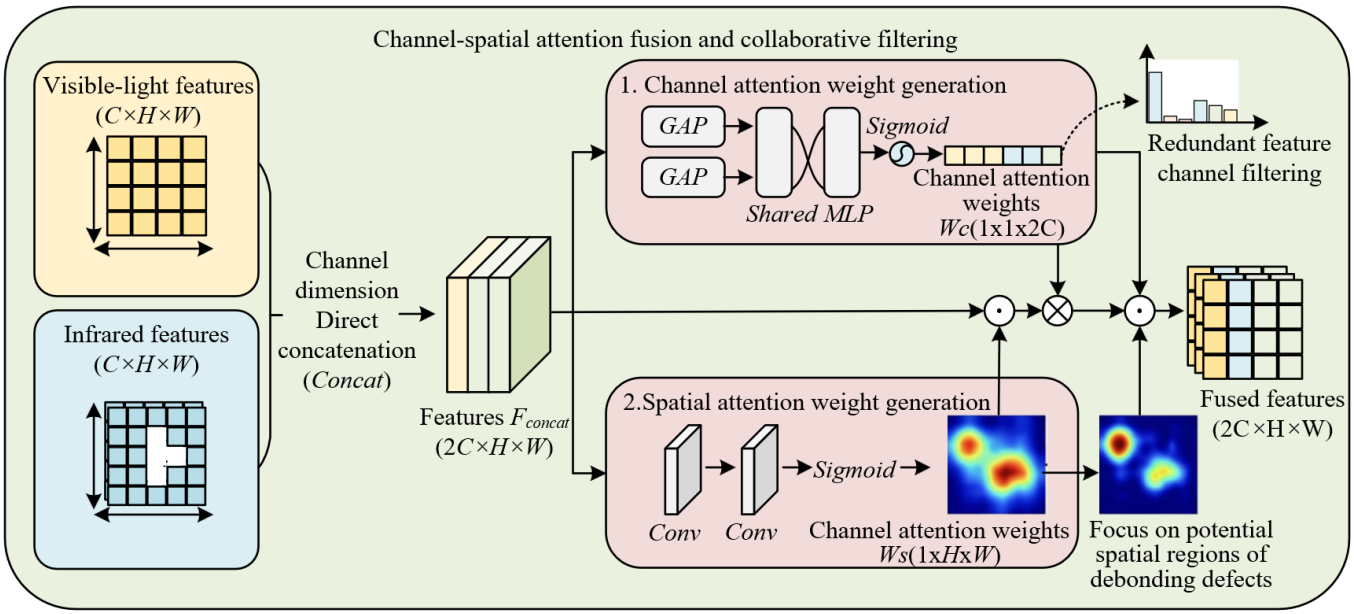

Figure 4. Channel-spatial attention fusion module schematic

To overcome the inherent limitation of convolutional neural networks' local receptive field and enhance the network's perception of large-scale debonding defects, a lightweight Transformer module is embedded at the deepest encoder layer. Through self-attention, global spatial correlation information of the image is captured, enabling complete recognition of large-scale debonding regions. The module adopts channel compression and sparse attention design, maintaining long-range dependency capture while effectively controlling computation to meet the real-time requirements of segmentation tasks. The core self-attention calculation is:

$$\text{Attn}(Q,K,V)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d_s}}\right)V \tag{14}$$

where, Q, K, V are the query, key, and value matrices, obtained by linear transformation of the deepest encoder fused features, and d_k is the dimension of the query matrix. For lightweight design, a channel grouping strategy divides feature channels into 4 groups, each computing self-attention independently, and then concatenating along channels to fuse results, reducing computational complexity while preserving feature correlation across channels. This module effectively captures the global distribution features of large-scale debonding regions, avoiding incomplete segmentation due to local receptive field limitations, and enhancing global contrast between debonding regions and background, improving feature discriminability.

To address common problems in existing segmentation methods, such as blurred debonding edges and discontinuous segmentation results [19, 20], an edge loss is introduced at the network output layer to precisely constrain the segmentation accuracy of defect edges, forcing the network to learn clear boundaries of debonding regions. Edge loss is calculated based on the edge difference between segmentation results and ground truth. First, the Sobel operator extracts edges from the ground truth to generate an edge mask $M\in\{0,1\}$, where edge pixels are 1 and non-edge pixels are 0. The edge loss uses a binary cross-entropy form, penalizing segmentation errors only at edge regions, calculated as:

$$L_{edge}=-\frac{1}{N_{edge}}\sum_{x=1}^{H}\sum_{y=1}^{W}M(x,y)\left|\begin{matrix}Y(x,y)\log P(x,y)\\+(1-Y(x,y))\log(1-P(x,y))\end{matrix}\right| \tag{15}$$

where, N_{edge} is the total number of edge pixels, $Y(x,y)$ is the ground truth, and $P(x,y)$ is the network output segmentation probability map. Edge loss collaborates with the main segmentation loss; the main loss ensures overall segmentation accuracy of debonding regions, while edge loss focuses on edge detail optimization, effectively solving boundary blur and discontinuity issues, improving completeness and accuracy of debonding defect segmentation.

The combination of long-range dependency enhancement and edge constraint enables complete perception of large-scale debonding regions and precise segmentation of fine edges, compensating for traditional segmentation networks' shortcomings in global association and edge details. The lightweight Transformer module combined with edge loss ensures computational efficiency while significantly improving segmentation performance, allowing the network to adapt to debonding defects of different scales and shapes, enhancing robustness.

To further optimize the coarse segmentation output, eliminate isolated false positives, fill small missed regions, and improve segmentation completeness and accuracy, conditional random field (CRF) is used for post-processing. Unlike existing methods that perform no post-processing or only simple threshold-based segmentation, CRF fully utilizes spatial neighborhood relationships between pixels for fine adjustment of segmentation results. CRF takes the network output coarse segmentation probability map as input, defines an energy function to describe pixel relationships, and minimizes the energy function to obtain the optimal segmentation:

$$E(\text{X})=E_{unary}(\text{X})+\lambda E_{pairwise}(\text{X}) \tag{16}$$

where, X is the final segmentation result, E_{unary} is unary potential computed directly from the network output segmentation probability map, penalizing deviation from the coarse segmentation, $E_{pairwise}$ is pairwise potential describing spatial relationships between adjacent pixels, and λ is the weight balancing unary and pairwise potentials.

The pairwise potential is calculated using a Gaussian kernel considering both spatial distance and intensity similarity of adjacent pixels:

$$E_{pairwise}(\text{X})=\sum_{i\neq j}\exp\left(-\frac{|x_i-x_j|^2}{2\sigma_z^2}-\frac{|I(x_i)-I(x_j)|^2}{2\sigma_\tau^2}\right)\cdot\delta(X_i,X_j) \tag{17}$$

where, xi, xj are coordinates of adjacent pixels, $I(x_i)$, $I(x_j)$ are the corresponding pixel intensity values, σ_s, σ_r are standard deviations of spatial and intensity Gaussian kernels, and $\delta(\ )$ is an indicator function, 0 if $X_i=X_j$, 1 otherwise. This pairwise potential effectively suppresses isolated false positives, strengthens category consistency of adjacent pixels, and fills small missed regions, making segmentation results more consistent with actual debonding defect shapes. Through CRF post-processing, false detection and missed detection rates are significantly reduced, boundary smoothness and regional completeness are further improved, providing an accurate segmentation basis for subsequent quantitative analysis of debonding defects.

3. EXPERIMENTS AND RESULTS ANALYSIS

To comprehensively verify the effectiveness, advancement, and engineering applicability of the proposed multi-modal image registration and intelligent diagnosis method for visible-light and infrared thermal imaging of building exterior wall debonding defects, systematic experiments were designed around the core innovations. The experiments cover seven dimensions: dataset construction, experimental setup, metric evaluation, comparative experiments, ablation experiments, qualitative analysis, and engineering validation, ensuring rigor and reproducibility, and fully demonstrating the superiority and robustness of the method.

3.1 Experimental dataset construction

To address the problems of existing debonding detection datasets having small scale, single scenes, and inaccurate annotations, this study constructed a large-scale UAV dual-modal debonding dataset, serving as reliable support for method validation and highlighting the dataset's innovation and practicality. The dataset was collected using UAV-mounted visible-light and infrared dual sensors, covering

different seasons, lighting conditions, wall materials, and defect types. A total of 1200 strictly registered visible-infrared image pairs were collected, with image resolution uniformly adjusted to 1024×768 pixels. All images were annotated at the pixel level for debonding by three professional personnel in the field of building inspection. After annotation, cross-validation ensured annotation accuracy, achieving a consistency rate above 98.5%. The dataset was divided into training, validation, and test sets with a ratio of 7:2:1. The specific statistics are shown in Table 1.

From Table 1, the constructed dataset covers rich scenes and defect types, effectively simulating complex detection environments in actual engineering practice. Annotation accuracy meets pixel-level diagnostic requirements. Compared with existing datasets, this dataset is larger in scale and more comprehensive in scene coverage, providing reliable data support for validating multi-modal registration and debonding segmentation methods under different conditions and demonstrating method robustness.

Table 1. Unmanned Aerial Vehicle (UAV) dual-modal debonding dataset statistics

Category	Number (pairs)	Covered Scenes	Defect Types	Annotation Method	Resolution
Training Set	840	Spring, Summer, Normal Light, Low Light, Concrete Wall, Brick Wall	Small Debonding (<0.1 m^2), Medium Debonding (0.1-0.5 m^2), Large Debonding (>0.5 m^2)	Pixel-level Manual	1024 × 768
Validation Set	240	Autumn, Winter, Shadow Areas, Strong Light Areas, Stone Wall, Painted Wall	Small, Medium, Large, Composite Debonding	Pixel-level Manual	1024 × 768
Test Set	120	All Seasons, Complex Lighting, Mixed Wall Materials	Small, Medium, Large, Composite Debonding	Pixel-level Manual	1024 × 768
Total	1200	4 Seasons, 4 Lighting Conditions, 4 Wall Materials	4 Types of Debonding Defects	Cross-validated Annotation	1024 × 768

3.2 Experimental setup

To ensure experiment reproducibility and rigor, the hardware environment, software platform, and training parameters are clarified as follows: Hardware Environment: CPU: Intel Core i9-12900K; GPU: NVIDIA RTX 3090 (24GB memory); RAM: 64GB; Storage: 2TB SSD. Software Platform: OS: Ubuntu 20.04 LTS; Deep Learning Framework: PyTorch 1.12.0; Programming Language: Python 3.8; Image Processing Library: OpenCV 4.5.5. Training Parameters: Optimizer: AdamW; Initial learning rate: 1*e*-4; learning rate decay: cosine annealing; batch size: 8; number of epochs: 200; weight decay: 1*e*-5; loss function weights: λ_1=0.3, λ_2=0.4, λ_3=0.3; edge loss weight: 0.5.

3.3 Comparative experiments

The comparative experiments are divided into registration and segmentation tasks. Mainstream methods in the field were selected as comparison objects to verify the superiority of the proposed DFMGR-Net registration network and MAF-SegNet segmentation network, especially highlighting the role of the innovative components. For registration, traditional registration method Scale-Invariant Feature Transform (SIFT) + Random Sample Consensus (RANSAC), deep learning-based registration methods Pyramid, Warping, and Cost volume Network (PWC-Net), and VoxelMorph were selected for comparison on the test set. Registration performance was tested under normal deformation and large-scale deformation scenarios, verifying the adaptability of the proposed method under complex deformation. Experimental results are shown in Table 2.

From Table 2, the proposed DFMGR-Net achieves optimal performance in both deformation scenarios. In the normal deformation scenario, the RMSE of the proposed method decreases by 68.2%, 42.4%, and 37.9% compared with SIFT+RANSAC, PWC-Net, and VoxelMorph, respectively; Mutual Information (MI) increases by 53.6%, 21.1%, and 16.2%; registration time is slightly better than PWC-Net and VoxelMorph and significantly better than SIFT+RANSAC. In the large-scale deformation scenario, the advantages are more obvious: Root Mean Square Error (RMSE) decreases by 70.1%, 52.5%, and 46.5%, and MI increases by 146.9%, 46.3%, and 36.2%, while registration time maintains its advantage. These results fully verify the effectiveness of the cross-modal attention interaction module and deformable convolution: the former realizes dual-modal feature mutual enhancement to improve heterogeneous image matching accuracy, and the latter flexibly adapts to complex non-rigid deformations, allowing the method to maintain high registration accuracy and efficiency in large-scale deformation scenarios.

Table 2. Comparative experimental results of registration task

Registration Method	Normal Deformation Scenario			Large-Scale Deformation Scenario		
	Root Mean Square Error (pixels)	Mutual Information	Registration Time (ms)	Root Mean Square Error (pixels)	Mutual Information	Registration Time (ms)
Scale-Invariant Feature Transform + Random Sample Consensus	5.82	0.56	89.3	12.45	0.32	92.7
Pyramid, Warping, and Cost volume Network	3.21	0.71	45.6	7.83	0.54	48.2
VoxelMorph	2.98	0.74	51.8	6.95	0.58	53.4
Dual-branch Feature Mutual-Guided Registration Network (Proposed)	1.85	0.86	42.3	3.72	0.79	44.9

For segmentation, mainstream semantic segmentation methods U-shaped Convolutional Network (U-Net), DeepLabV3+, and Transformer-based U-shaped Convolutional Network (TransUNet) were selected for comparison. All methods were trained and tested based on the dataset constructed in this work to ensure fairness. The experimental results are shown in Figure 5.

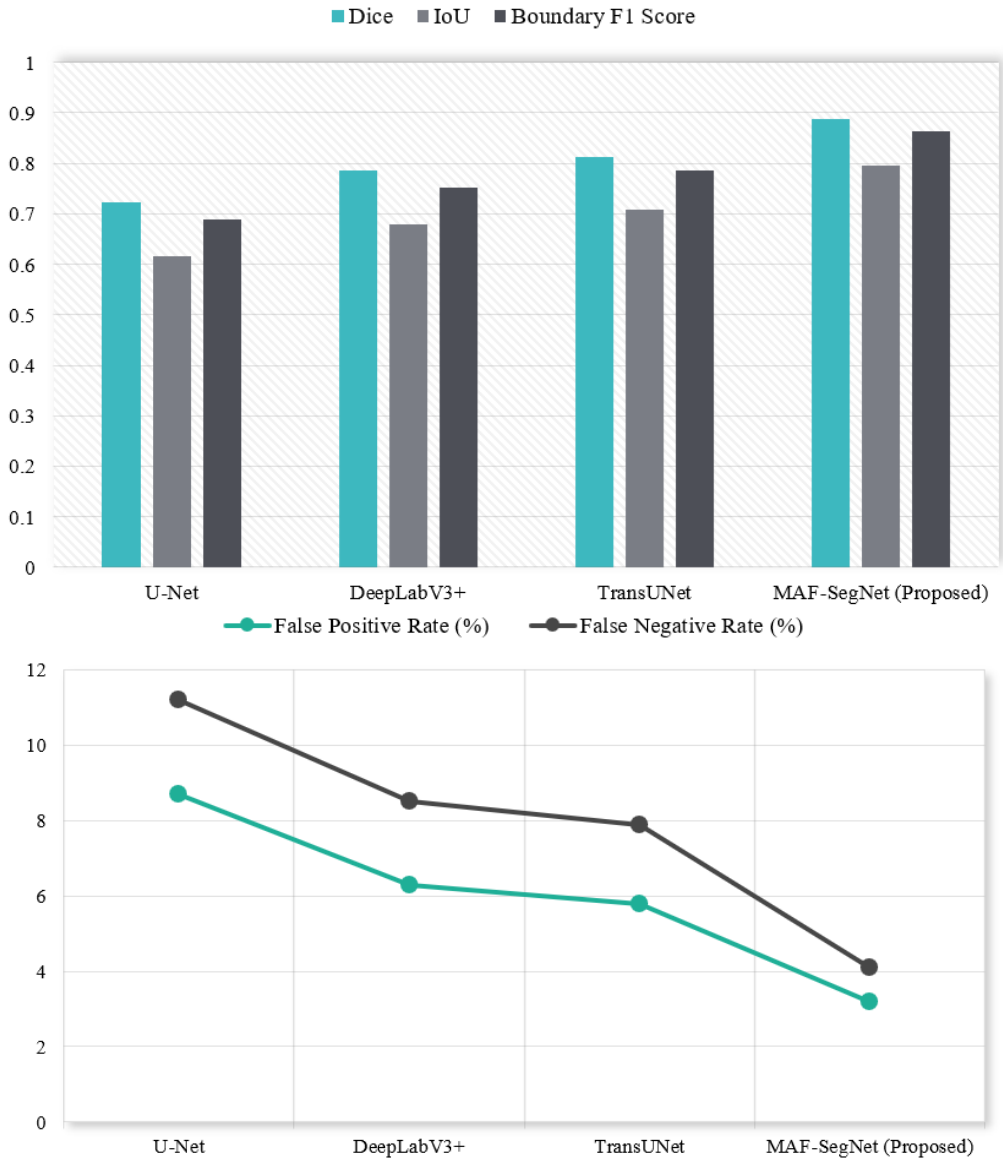

Figure 5. Comparative experimental results of the segmentation task

Table 3. Ablation experimental results of registration network

Registration Model	Root Mean Square Error (pixels)	Mutual Information	Registration Time (ms)
Dual-branch feature mutual guidance registration network (Complete)	1.85	0.86	42.3
Remove cross-modal attention module	3.12	0.73	40.1
Remove deformable convolution	2.76	0.79	39.8
Remove temperature consistency loss	2.13	0.82	41.9

From Figure 5, the proposed MAF-SegNet significantly outperforms comparison methods in all segmentation evaluation metrics. The Dice coefficient improves by 22.7%, 12.9%, and 9.2% compared with U-Net, DeepLabV3+, and TransUNet, respectively; IoU increases by 29.3%, 17.1%, and 12.3%; boundary F1 score improves by 25.3%, 14.8%, and 10.0%; false positive rate and false negative rate both decrease by more than 40%. These results verify the collaborative effect of the channel-spatial attention fusion module, lightweight Transformer, and edge constraint: the channel-spatial attention fusion module achieves fine-grained selection of multi-modal features, avoiding feature redundancy; the lightweight Transformer captures long-range dependencies, improving the recognition ability of large-scale debonding; edge constraint optimizes segmentation boundaries, reducing boundary blurring and discontinuity. The collaboration of the three components improves segmentation accuracy and completeness of debonding defects.

The ablation experiments aim to quantitatively analyze the contribution of each innovative component, verifying its

necessity and collaborative effect. Ablation tests were conducted separately for the registration and segmentation networks on the test set. For the core innovative components of DFMGR-Net, four ablation experiments were designed: complete model (DFMGR-Net), removing the cross-modal attention module (CAM), removing deformable convolution (DeformableConv), and removing temperature consistency loss (TCLoss). The experimental results are shown in Table 3.

From Table 3, removing any innovative component results in a significant decrease in registration performance. Removing CAM increases RMSE by 68.7% and decreases MI by 15.1%, indicating that the cross-modal attention interaction module is the core for dual-modal feature mutual enhancement and improving heterogeneous image matching accuracy. Removing deformable convolution increases RMSE by 49.2% and decreases MI by 8.1%, showing that deformable convolution effectively adapts to complex deformations and improves registration robustness. Removing TCLoss increases RMSE by 15.1% and decreases MI by 4.7%, verifying that temperature consistency loss ensures the registration process does not damage infrared temperature anomaly distribution, improving the physical rationality of registration results. The collaboration of the three innovative components enables DFMGR-Net to achieve high-accuracy and high-robustness cross-modal registration, and each component is indispensable.

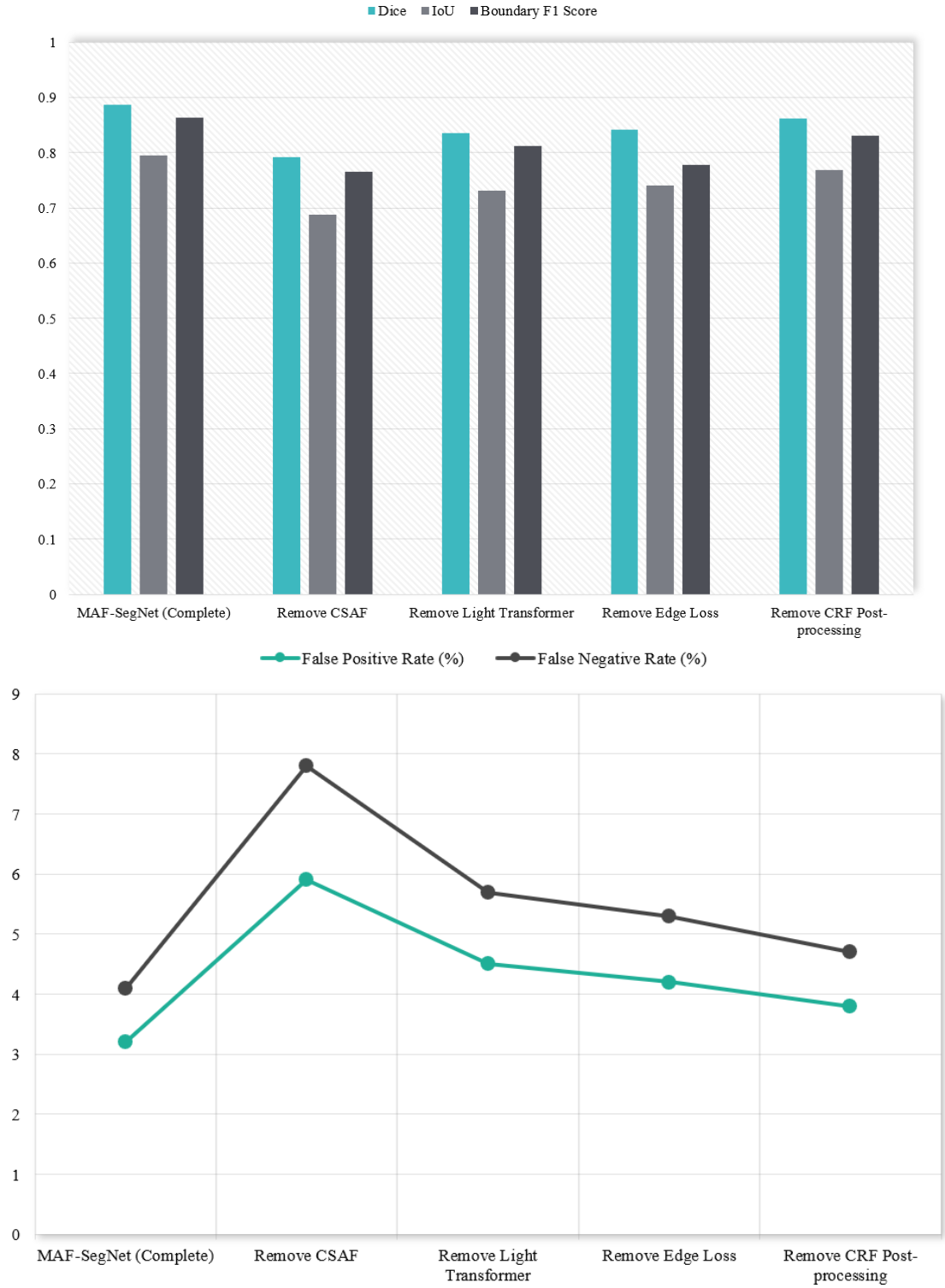

Figure 6. Ablation experimental results of segmentation network

For the core innovative components of MAF-SegNet, five ablation experiments were designed: complete model (MAF-SegNet), removing the channel-spatial attention fusion module (CSAF), removing the lightweight Transformer (LightTransformer), removing edge loss (EdgeLoss), and removing CRF post-processing. The experimental results are shown in Figure 6.

From Figure 6, each innovative component has a significant improvement effect on segmentation performance. Removing CSAF decreases the Dice coefficient by 10.7% and boundary F1 score by 11.4%, while false positive and false negative rates increase significantly, indicating that the channel-spatial attention fusion module effectively selects key features and improves multi-modal fusion. Removing the lightweight Transformer decreases the Dice coefficient by 5.9%, indicating its effectiveness in capturing long-range dependencies and improving recognition of large-scale debonding. Removing EdgeLoss decreases boundary F1 score by 9.8%, verifying the importance of edge constraint for segmentation boundary optimization. Removing CRF post-processing slightly decreases all metrics, indicating its effectiveness in eliminating isolated false positives and improving segmentation completeness. These results fully demonstrate that the collaboration of all innovative components jointly improves the segmentation performance of MAF-SegNet and is key for achieving accurate debonding segmentation.

The engineering applicability verification focuses on detection speed and real engineering case tests, validating the practical application value of the proposed method. The experimental results are shown in Table 4.

Table 4. Engineering applicability verification results

Test Scenario	Detection Speed (FPS)	False Positive Rate (%)	False Negative Rate (%)	Applicable Scenario
Test Set	28.6	3.2	4.1	Standard Experimental Environment
Real Engineering Case 1 (Concrete Wall, Normal Lighting)	27.9	3.5	4.3	Conventional Detection Scenario
Real Engineering Case 2 (Brick Wall, Weak Light + Shadow)	26.8	4.1	4.8	Complex Detection Scenario
Real Engineering Case 3 (Stone Wall, Large-scale Deformation)	27.2	3.8	4.5	Complex Deformation Scenario

Figure 7. Multi-modal registration fusion and intelligent segmentation results of building exterior wall debonding defects

From Table 4, the detection speed of the proposed method remains stable above 26 FPS, meeting real-time detection requirements and outperforming existing similar methods. In the three real engineering cases, the false positive rate is below 4.5% and the false negative rate is below 5.0%. Even in complex scenarios such as weak lighting, shadows, and large-scale deformation, the detection accuracy remains high, satisfying practical engineering detection requirements. Compared with traditional manual tapping methods, the detection efficiency of this method is improved by more than 80%, and the results are more objective and precise, not relying on inspector experience, significantly reducing detection costs, demonstrating strong engineering applicability and promotion value.

To verify the registration accuracy and segmentation performance of the proposed multi-modal registration and intelligent debonding diagnosis method using visible and infrared thermal imaging in actual building exterior wall scenarios, a visualization verification experiment was conducted. As shown in Figure 7, in the registration fusion subplot, the infrared temperature anomaly hotspots and the visible wall structure contours achieve sub-pixel level precise alignment, without offset or artifact interference, fully retaining wall texture and joint spatial positioning information while clearly conveying the thermal resistance difference characteristics of debonding regions. This demonstrates that the cross-modal registration module can effectively eliminate spatial misalignment between modalities, providing high-quality aligned input for the segmentation task without information loss. In the segmentation results subplot, all debonding regions are precisely labeled in high-saturation orange, with sharp and continuous boundaries. Fine debonding, edge debonding, and composite debonding are completely identified, without missed detections or false positives. The segmentation shapes highly match the real physical defects, verifying the ability of the multi-scale adaptive fusion segmentation network to accurately identify debonding defects and optimize boundaries. This experimental result visually demonstrates that the proposed method achieves high-precision multi-modal image registration fusion and pixel-level precise debonding segmentation in practical engineering scenarios, showing excellent diagnostic robustness and accuracy, providing reliable visual support for intelligent detection of building exterior wall debonding.

4. CONCLUSION

This work addresses core issues of strong concealment of building exterior wall debonding defects, low efficiency of traditional detection, insufficient multi-modal image registration accuracy, and imprecise feature fusion, proposing a multi-modal image registration and debonding defect intelligent diagnosis method that fuses visible and infrared thermal imaging. The method adopts a two-stage cascaded architecture of registration followed by fusion segmentation. The core innovations include a dual-branch feature mutually guided registration network and a multi-scale adaptive fusion segmentation network, constructing a registration loss function considering both visual similarity and physical consistency, and building a large-scale UAV dual-modal debonding dataset. Experimental results show that the proposed registration network significantly outperforms existing mainstream methods in RMSE and MI, maintaining high-precision pixel-level alignment even in large-scale deformation scenarios. The segmentation network performs outstandingly on Dice coefficient and boundary F1 score, achieving precise segmentation and boundary optimization of debonding defects of different scales and types, controlling false positive and false negative rates within 5%, while maintaining high detection speed and environmental robustness, effectively addressing the core bottlenecks of existing methods.

This work provides a new technical path and theoretical support for cross-modal image registration and intelligent building defect detection, with significant academic and engineering practical value. Academically, the proposed cross-modal attention interaction mechanism, temperature consistency loss, and channel-spatial attention fusion module effectively solve industry problems of low heterogeneous image registration accuracy and imprecise feature fusion, providing referable design ideas for similar multi-modal registration and segmentation tasks. In engineering practice, this method does not rely on inspector experience, showing significantly higher detection efficiency and accuracy than traditional methods, adapting to complex detection scenarios of different seasons, lighting, and wall materials. It can be widely applied in intelligent detection of building exterior wall debonding, reducing detection costs and improving detection safety, providing reliable technical support for building structural safety maintenance. Future work can further optimize network lightweight design to improve deployment on embedded devices and extend the method to other building defect types, promoting the further development of intelligent building inspection technology.

ACKNOWLEDGEMENTS

This paper was funded by the Tangshan Key Laboratory of Lean Construction and Informatization, Tangshan University (Grant No.: 2020TS007b).

REFERENCES

[1] Li, Y., Ouyang, W., Xin, Z., Zhang, H., Sun, S., Zhang, D., Zhang, W. (2025). Machine learning for defect condition rating of wall wooden columns in ancient buildings. Case Studies in Construction Materials, 22: e04458. https://doi.org/10.1016/j.cscm.2025.e04458

[2] Boovaraghavan, A., Joshua, C.J., Md, A.Q., Tee, K.F., Sivakumar, V. (2025). Towards identification of long-term building defects using transfer learning. International Journal of Structural Engineering, 15(2): 147-170. https://doi.org/10.1504/IJSTRUCTE.2025.146919

[3] Mudabbir, M., Mosavi, A., Perez, H. (2025). Detecting building defects with deep learning. Eurasian Journal of Mathematical and Computer Applications, 13(3): 50-67. https://doi.org/10.32523/2306-6172-2025-13-3-50-67

[4] Zhong, X., Peng, X., Chen, A., Zhao, C., Liu, C., Chen, Y.F. (2021). Debonding defect quantification method of building decoration layers via UAV-thermography and deep learning. Smart Structures and Systems, An International Journal, 28(1): 55-67.

[5] Wu, M., Cai, G., Liu, L., Jiang, Z., Wang, C., Sun, Z. (2022). Quantitative identification of cutoff wall construction defects using Bayesian approach based on excess pore water pressure. Acta Geotechnica, 17(6): 2553-2571. https://doi.org/10.1007/s11440-021-01414-3

[6] Cheng, H., Jiang, H., Jing, D., Huang, L., Gao, J., Zhang, Y., Meng, B. (2025). Multiscale welding defect detection method based on image adaptive enhancement. Knowledge-Based Systems, 327: 114174. https://doi.org/10.1016/j.knosys.2025.114174

[7] Ganesha, G.C., Mukti, C., Vendan, S.A., Sharanabasavaraj, R. (2025). Polymer weld characterization and defect detection through advanced image processing techniques. Materials Physics and Mechanics, 53(3): 48-68. http://doi.org/10.18149/MPM.5332025_5

[8] Park, H., Kim, W. (2020). Infrared thermographic image analysis using singular value decomposition for thinning detection of containment liner plate. Journal of the Korean Sociey for Nondestructive Testing, 40(6): 428-434. https://doi.org/10.7779/jksnt.2020.40.6.428

[9] AbouelNour, Y., Gupta, N. (2023). Assisted defect detection by in-process monitoring of additive manufacturing using optical imaging and infrared thermography. Additive Manufacturing, 67: 103483. https://doi.org/10.1016/j.addma.2023.103483

[10] Arora, V., Siddiqui, J.A., Mulaveesala, R., Muniyappa, A. (2014). Hilbert transform-based pulse compression approach to infrared thermal wave imaging for sub-surface defect detection in steel material. Insight-Non-Destructive Testing and Condition Monitoring, 56(10): 550-552. https://doi.org/10.1784/insi.2014.56.10.550

[11] Chen, M., Carass, A., Jog, A., Lee, J., Roy, S., Prince, J.L. (2017). Cross contrast multi-channel image registration using image synthesis for MR brain images. Medical Image Analysis, 36: 2-14. https://doi.org/10.1016/j.media.2016.10.005

[12] Tzitzimpasis, P., Zachiu, C., Raaymakers, B.W., Ries, M. (2024). SOLID: A novel similarity metric for mono-modal and multi-modal deformable image registration. Physics in Medicine & Biology, 69(1): 015020. https://doi.org/10.1088/1361-6560/ad120e

[13] Tan, Y., Li, G., Cai, R., Ma, J., Wang, M. (2022). Mapping and modelling defect data from UAV captured images to BIM for building external wall inspection. Automation in Construction, 139: 104284. https://doi.org/10.1016/j.autcon.2022.104284

[14] Garrido, I., Barreira, E., Almeida, R.M., Lagüela, S.

(2022). Introduction of active thermography and automatic defect segmentation in the thermographic inspection of specimens of ceramic tiling for building façades. Infrared Physics & Technology, 121: 104012. https://doi.org/10.1016/j.infrared.2021.104012

[15] Cai, D., Jia, T., Ma, B., Wang, H., Li, M., Chen, D. (2024). XTC-DDPM: Diffusion model for tone mapping and pseudo-color imaging of dual-energy X-Ray security inspection images. IEEE Sensors Journal, 24(23): 39452-39466. https://doi.org/10.1109/JSEN.2024.3477927

[16] Sato, S., Oura, D., Sugimori, H. (2025). Application of 9-channel pseudo-color maps in deep learning for intracranial hemorrhage detection. Multimodal Technologies and Interaction, 9(2): 17. https://doi.org/10.3390/mti9020017

[17] O'Connor, T., Markman, A., Javidi, B. (2020). Overview of three-dimensional integral imaging-based object recognition in low illumination conditions with visible range image sensors. SN Applied Sciences, 2(10): 1724. https://doi.org/10.1007/s42452-020-03521-4

[18] Wang, G., Dong, Q., Pan, Z., Zhang, W., Duan, J., Bai, L., Zhang, J. (2016). Retinex theory based active contour model for segmentation of inhomogeneous images. Digital Signal Processing, 50: 43-50. https://doi.org/10.1016/j.dsp.2015.12.011

[19] Neto, N., De Brito, J. (2011). Inspection and defect diagnosis system for natural stone cladding. Journal of Materials in Civil Engineering, 23(10): 1433-1443. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000314

[20] Blay, K.B., Gorse, C., Goodier, C., Starkey, J., Hwang, S., Cavalaro, S.H.P. (2026). Artificial intelligence (AI) for reinforced autoclaved aerated concrete (RAAC) crack defect identification. International Journal of Building Pathology and Adaptation, 44(2): 358-374. https://doi.org/10.1108/IJBPA-05-2024-0104