The research presented was financially supported by the National Key R&D Program of China (No. 2022YFC3801700), the National Natural Science Foundation of China (No. 52378289), Research Fund for Advanced Ocean Institute of Southeast University (No. KP202407).

Towards Crack Localization of Hollow Slab Bridges: A Global Visual Representation Approach with Matrix Encoding and Dual Lightweight Algorithms

Abstract: Hollow slab bridges are widely utilized in highway infrastructure due to their compact structural form and ease of construction. However, the constrained space and concealed nature of defects on the underside of the girders pose significant challenges for traditional inspection methods, which often fail to meet the efficiency and accuracy demands of modern maintenance practices. To address this issue, this study proposes an intelligent defect identification and localization framework based on panoramic visual representation. The proposed method comprises three core components: (1) A high-resolution image acquisition system mounted on an unmanned aerial vehicle (UAV), integrated with a dual-feature matching strategy driven by texture-awareness. In texture-sparse regions, the Gim-LightGlue algorithm is employed, achieving a 100% matching accuracy while reducing feature extraction time to only 15.3% of that required by the traditional SIFT algorithm. In texture-rich regions, the XFeat-LightGlue algorithm enhances processing speed by 35.45%; (2) To mitigate image distortion in large-scale stitching, a geometric rectification approach is developed by leveraging the linear features of girder undersides. A matrix encoding strategy is further introduced to grid the stitched images, enabling the embedding of spatial positional information for each subregion; (3) The TransUnet model was trained on a total of 2,993 images from the crack dataset (2,694 for training and 299 for validation), achieving an accuracy of 77.62%. By further integrating a spatial-index-based synchronous damage recognition and localization method, precise mapping of bridge soffit defects within the global field of view was realized. Validation conducted on an in-service bridge demonstrates that the proposed framework achieves high accuracy and efficiency in defect identification and localization, offering reliable data support for bridge maintenance decision-making.

Keywords: Bridge; Image stitching; Deep learning; Defect identification and localization

1. Introduction

Hollow slab bridges are widely used in small- to medium-span highway systems due to their ease of construction and cost efficiency. However, prolonged exposure to environmental factors and inadequate maintenance often lead to progressive defects such as cracks, spalling, and corrosion on the girder undersides. These hidden and slowly developing damages, if not promptly detected, pose significant risks to structural integrity and traffic safety [1]. As a result, efficient and accurate inspection of girder undersides has become a pressing challenge in bridge maintenance. Traditional inspection methods, such as manual patrols or vehicle-mounted platforms—are labor-intensive, costly, and susceptible to omissions due to limited visibility and operator subjectivity. Recent advancements in UAVs, high-resolution imaging, and deep learning have driven the adoption of image-based intelligent inspection techniques [2]. UAVs, in particular, offer excellent mobility and access to confined or complex girder spaces, making them well-suited for such applications. However, real-world deployment remains constrained by factors such as limited space, physical obstructions, and unreliable GPS signals beneath girders, which hinder precise localization and reduce system robustness.

To address the challenges of UAV-based bridge inspection, various integrated localization strategies have been proposed, including multi-sensor systems combining IMUs, stereo vision, and LiDAR, as well as image-based methods utilizing feature extraction and structured light reconstruction. For example, Xie et al. [3] developed a robotic inspection platform equipped with multi-modal sensors to enhance localization and coverage of girder undersides, though its bulk, dependence on emergency lane access, and blind spots limit practical deployment. Chen et al. [4] employed a distributed camera chain mounted on a rail to acquire panoramic images and localize defects on large-span bridges; however, the high cost and complexity restrict its use in small- to medium-span structures. Compared to ground-based systems, UAVs offer enhanced mobility and access to hard-to-reach areas without scaffolding or lifts, enabling safer and more efficient inspections. Nonetheless, challenges persist, including unstable positioning due to weak GPS signals beneath girders and scale inconsistencies caused by variable camera-to-target distances. To mitigate these issues, Jiang et al. [5] proposed a vision-guided UAV combining stereo vision and IMU data for improved localization. Wang et al. [6] advanced this approach with the FMC-SVIL method, integrating stereo visual-inertial odometry and an enhanced AprilTag2 algorithm for precise navigation. Bolourian et al. [7] further demonstrated LiDAR-equipped UAVs for 3D mapping of girder undersides. Despite promising results, multi-sensor fusion often increases computational complexity and reduces system robustness under field conditions. To address scale consistency, Jiang et al. [8] introduced a wall-climbing UAV capable of high-resolution imaging, though its reliance on hovering and narrow field of view limits inspection efficiency and coverage, making it less suited for large-scale or rapid assessment tasks.

Recent advancements in three-dimensional (3D) model-based localization have spurred interest in their application to structural inspection. Jahanshahi et al. [9] leveraged multi-view image matching to retrieve spatial depth and eliminate perspective distortion. Expanding on this, Hu et al. [10] developed a deep learning-based 3D reconstruction approach for cable-stayed bridges, combining voxel-based models with geometric primitives and employing multi-view CNNs and point cloud networks to extract features from both 2D images and 3D data. Liu et al. [11] applied structure-from-motion (SfM) to model concrete surfaces and locate cracks, enabling precise defect identification and repair guidance. However, these techniques often involve time-consuming workflows and operational complexity, limiting their suitability for real-time field applications. To address efficiency concerns, Feng et al. [12] proposed a crack assessment method integrating SLAM with image super-resolution, reducing the computational load of 3D modeling. Despite these improvements, high hardware requirements, elevated costs, and complex operation continue to constrain the practical adoption of 3D model-based methods. For bridge structures, generating high-fidelity 3D models remains time-intensive and highly sensitive to environmental conditions, posing challenges to the development of rapid, portable, and scalable inspection solutions.

Compared to complex multi-sensor localization systems, panoramic image stitching offers a more lightweight and practical solution for visual inspection of girder undersides. By leveraging UAVs to capture multi-view images and reconstructing them into a unified panoramic view, comprehensive visual mapping can be achieved without reliance on high-precision positioning hardware. For example, Xie et al. [3] developed a vehicle-mounted robotic system that acquired thousands of seabed images and combined 2D point and 3D line features to reduce stitching drift, yielding seamless panoramas. Similarly, Wang et al. [13] employed a rail-mounted camera setup to capture fixed-angle images of girder undersides, while Hou et al. [14] evaluated SIFT, Harris, and SURF algorithms for bridge image stitching performance. Most stitching systems rely on fixed viewpoints—via robotic arms or rail-mounted cameras—to minimize distortion and alignment errors. However, these systems often require specialized infrastructure, limiting flexibility and operational efficiency. In contrast, UAV-based imaging introduces variability in flight posture, leading to angular deviations and scale inconsistencies between adjacent images. Under such dynamic conditions, conventional stitching techniques, which assume stable viewpoints, are prone to artifacts such as misalignment, ghosting, and geometric distortion. Achieving accurate stitching and localization from UAV images under these constraints remains a key technical challenge.

High-precision detection algorithms are essential for closing the loop in damage localization. Recent advances in deep learning-based computer vision have significantly improved automated defect identification. Cha et al. [15] applied convolutional neural networks (CNNs) with a sliding window approach for large-scale concrete crack analysis, and later employed Faster R-CNN to detect and classify multiple defect types [16]. Yu et al. [17] utilized Mask R-CNN to enable automated crack detection and instance-level segmentation. For single-stage detection, Wan et al. [18] implemented the SSD framework for bridge damage recognition, further enhancing accuracy through an eight-neighborhood post-processing algorithm. Peng et al. [19] extended the YOLO framework to develop YOLO-lump and YOLO-crack for real-time detection of diverse surface defects on bridges. In fine-grained segmentation, Zhang et al. [20] introduced CrackNet, a fully convolutional network without pooling layers, enabling pixel-level predictions. Ni et al. [21] proposed a dual-scale CNN for detecting both wide and hairline cracks and employed Zernike moments for sub-pixel width measurement of microcracks. To address CNN limitations in capturing long-range dependencies, Kang et al. [22] incorporated Transformer-based multi-head self-attention, enhancing global feature extraction. Zhu et al. [23] adopted a U-Net architecture that fused shallow and deep features, improving both structural detail retention and width prediction. Liang et al. [24] developed MU-Net, a lightweight FCN that integrates MobileNet's depthwise separable convolutions and correction modules, achieving high detection accuracy and speed for concrete surface cracks.

To address the above challenges, this study proposes a UAV-driven framework for intelligent damage localization on hollow slab bridges via lightweight panoramic reconstruction. Designed to operate without high-precision localization sensors, the framework offers a low-cost, scalable solution suitable for frequent inspections of small- to medium-span bridges. It introduces three core innovations: (1) a texture-aware feature matching strategy that adaptively switches between Gim-LightGlue and XFeat-LightGlue based on surface texture complexity, ensuring robust and efficient image alignment; (2) a structural line-constrained stitching and spatial encoding method that leverages girder geometry to correct stitching errors and establish a matrix-based mapping between image segments and physical locations; and (3) integration of the TransUNet model for automated defect detection, combined with the encoding system to achieve precise spatial localization. This sensor-independent approach reduces hardware reliance and operational complexity, offering strong potential for practical deployment in bridge inspection and maintenance.

The paper is organized as follows. Section 2 outlines the proposed framework. Section 3 details the feature detection and matching strategy, panoramic stitching with geometric correction, and matrix-based image encoding. Section 4 introduces the TransUNet-based crack extraction method. Section 5 presents field experiment validations. Section 6 concludes the study and discusses future research.

2. Proposed Methodology

To address the challenges of confined spaces, localization difficulties, and weak-texture features in the inspection of hollow slab bridge undersides, this study presents an integrated framework combining UAV-based image acquisition, lightweight texture-guided stitching, and deep learning-based defect recognition. The system architecture is shown in Fig. 1. Considering the limited clearance and potential obstructions beneath girders, a lightweight UAV equipped with a high-resolution camera captures images along a matrix-patterned flight path to ensure high overlap. However, variations in UAV posture introduce tilt and perspective distortions, adversely affecting stitching quality. To counter this, the Segment Anything Model (SAM) is utilized for instance segmentation of structural edges. A homography matrix is then constructed based on the parallelism of side boundaries, projecting images to a standardized viewpoint orthogonal to the girder surface, thereby providing geometric consistency for stitching. To accommodate varying texture conditions, a texture-aware dynamic feature matching strategy is introduced. In weak-texture regions, the

Gim-LightGlue matcher, based on the SuperPoint-LightGlue architecture and fine-tuned for low-texture scenarios offers accurate and efficient performance. For texture-rich regions, the XFeat-LightGlue algorithm is used, which supports parallel keypoint detection and compact descriptor generation through a shallow network, ensuring a balance between speed and feature robustness. A built-in texture classification module dynamically selects the optimal matcher per region, enhancing efficiency while minimizing reliance on computational resources. For robust panorama construction, the RANSAC-USAC-FM-8PT algorithm filters outliers and estimates the fundamental and homography matrices. To mitigate cumulative distortion during

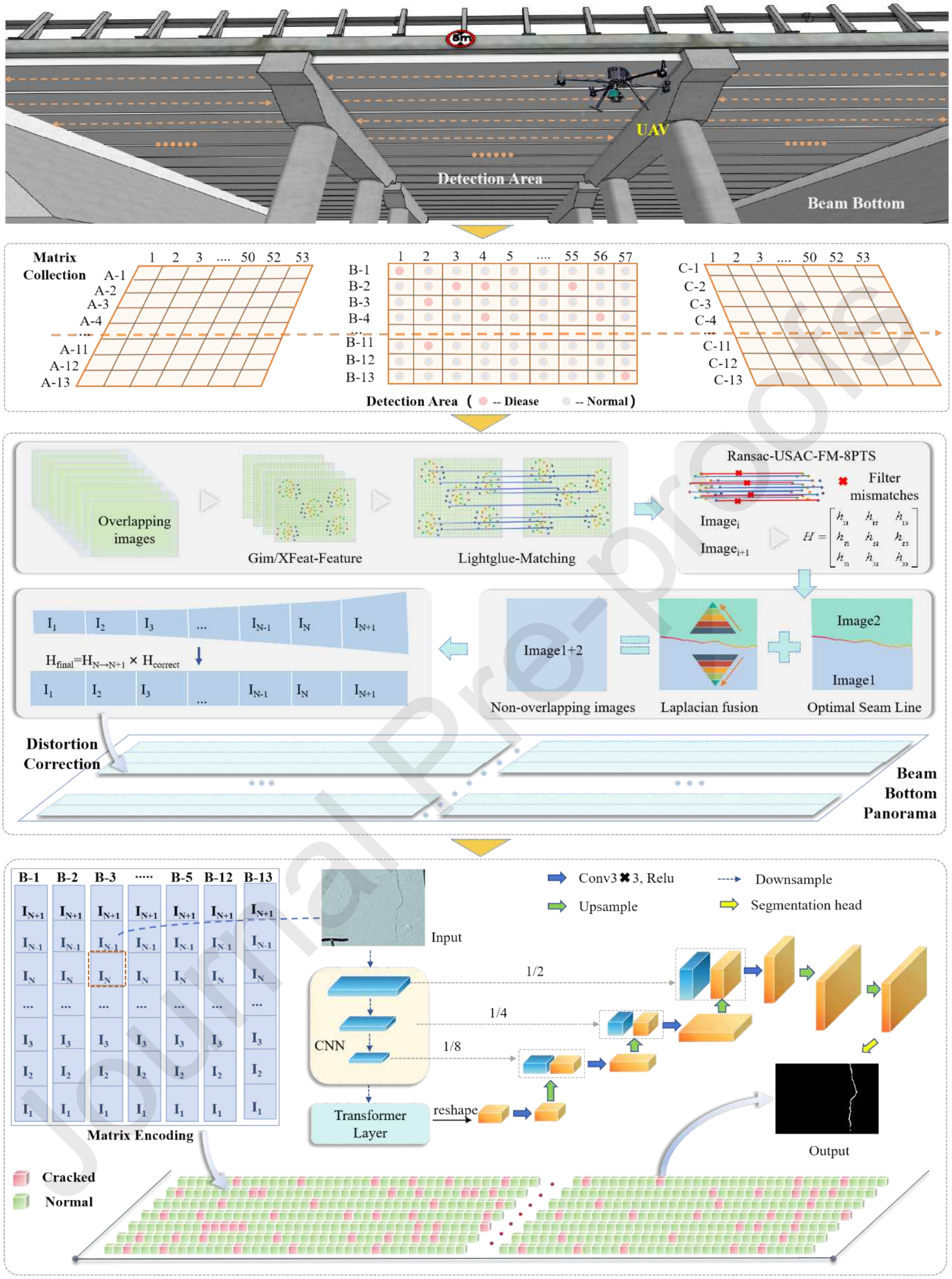


Fig. 1. Framework of Defect Detection and Localization Strategy.

iterative stitching, a structural line-constrained correction method is employed. By enforcing parallel constraints from girder edges and constructing a nonlinear rectification matrix from corner-to-structure point mappings, geometric deformations are effectively suppressed. Seam optimization is achieved via a dynamic energy-density seam search strategy, which determines optimal stitching paths based on color and gradient energy functions. Final image fusion employs Laplacian pyramid blending for consistent luminance and smooth transitions, resulting in seamless, high-resolution, and structurally aligned panoramas. To associate panoramic images with real-world bridge geometry, a matrix-based encoding scheme is proposed. The panorama is divided into non-overlapping blocks, each assigned a unique 2D spatial index to enable precise mapping to physical locations. For defect detection, the TransUNet segmentation model is applied to the stitched images. The results are reverse-mapped using the spatial encoding, establishing a one-to-one correspondence between identified defects and their actual positions. This enables accurate maintenance planning and provides a geometric foundation for targeted interventions.

3. Multi-Scale Texture-Aware Panoramic Reconstruction and Spatial Localization Framework

In response to the spatial constraints and complex operational conditions beneath hollow slab girders, a lightweight multi-rotor UAV equipped with a high-resolution imaging system is employed to systematically scan the underside of the structure. A predefined matrix-based flight path is executed to ensure comprehensive visual coverage with high image overlap and multi-view redundancy. This setup enables the construction of a dense image dataset characterized by strong continuity and diverse perspectives. To facilitate global visual modeling and spatial scene reconstruction of the girder underside, this study proposes a lightweight image reconstruction framework incorporating multimodal texture perception. The framework is designed to address key limitations of conventional stitching methods, particularly those related to viewpoint inconsistencies, weak texture representations, image warping distortions, and spatial mapping discontinuities. The proposed methodology comprises three primary components: (1) high-robustness feature extraction and dynamic matching; (2) sequential image stitching and distortion correction constrained by structural geometry; and (3) matrix-based image partitioning combined with a spatial position encoding strategy.

3.1 Texture-Aware Feature Extraction and Contextual Matching Strategy

During UAV-based data acquisition, it is often infeasible to maintain a strictly orthogonal or parallel imaging posture relative to the underside of the girder. As a result, the captured images typically exhibit varying degrees of tilt and non-linear perspective deviations, which significantly undermine the robustness and accuracy of conventional image stitching algorithms. These inconsistencies frequently lead to the accumulation of stitching errors and the occurrence of local geometric distortions. To effectively mitigate geometric distortions caused by inconsistent viewpoints, this study employs the Segment Anything Model (SAM) [25] proposed by Meta AI to perform precise semantic instance segmentation of bridge beams in the original images, thereby extracting salient edge contours and structural regions. Since beams may contain other linear structures (e.g., manually annotated lines), SAM is pre-trained on a beam segmentation dataset to ensure high-precision soffit segmentation across different environments. Based on the parallel alignment of structural elements on both sides of the girder, a standardized projection transformation model is constructed. By solving for the homography matrix, the original oblique images are projected onto a normalized viewing plane that approximates an orthogonal perspective relative to the underside of the bridge. This process produces a geometrically aligned image sequence suitable for panoramic reconstruction. This geometric rectification strategy not only enforces spatial consistency across the image set but also provides a unified reference plane for subsequent feature matching and stitching operations, thereby enhancing both the stitching accuracy and the computational stability of the overall reconstruction framework.

The core of image stitching lies in the coordinated integration of robust feature detection and precise feature matching mechanisms. Specifically, the feature detection phase aims to identify and localize a set of salient keypoints within the image that exhibit high response values and strong distinctiveness, while simultaneously constructing descriptors that are both discriminative and stable under varying imaging conditions. The subsequent feature matching stage seeks to establish geometrically consistent correspondences across multi-view image pairs, enabling accurate alignment between adjacent images. Conventional feature detection approaches can be broadly categorized into two groups. The first includes hand-crafted descriptors such as Scale-Invariant Feature Transform (SIFT) [26] and Speeded-Up Robust Features (SURF) [27], which offer sub-pixel localization accuracy and invariance to scale and rotation. Despite their robustness, these methods are computationally intensive and exhibit limited suitability for real-time applications. The second group comprises binary descriptors such as Oriented FAST and Rotated BRIEF (ORB), which are designed for computational efficiency and fast processing. However, their performance in terms of feature distinctiveness and noise robustness is generally inferior [28], limiting their reliability in complex or low-texture environments.

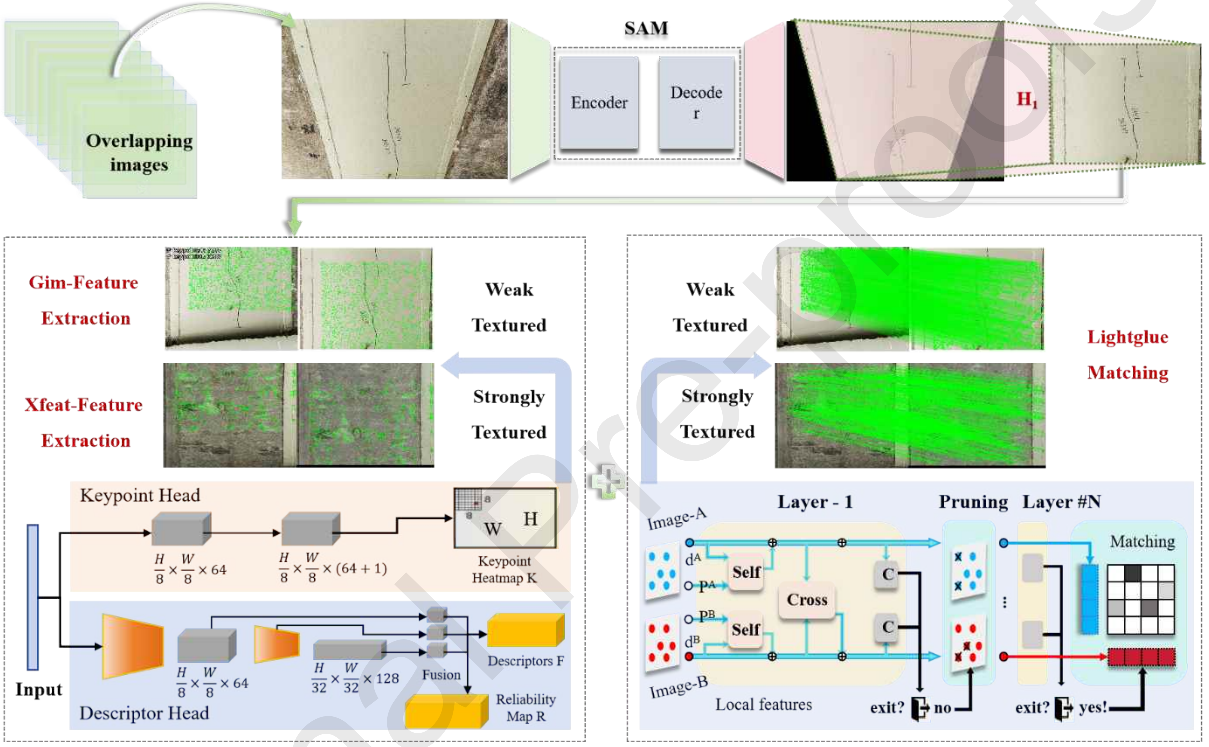


Fig. 2. Architecture of the Dual-Modal Feature Detection and Matching Mechanism.

With the rapid advancement of deep learning in visual feature modeling, a growing number of neural network-based feature detectors have been proposed, including SuperPoint [29], DISK [30], ALike [31], and Gim [32]. These methods, characterized by end-to-end training frameworks and strong feature representation capabilities, have demonstrated superior robustness and adaptability in complex real-world scenarios compared to traditional handcrafted approaches. Informed by field investigations and texture condition assessments of bridge girder underside images, this study develops a texture-aware feature detection framework tailored to heterogeneous surface characteristics. Specifically, for regions with low-texture content, the Gim-LightGlue algorithm is employed to enhance feature detection reliability and matching precision. In contrast, high-texture areas are processed using a lightweight XFeat-LightGlue network, which leverages efficient keypoint detection and compact descriptor generation to maintain a balance between computational efficiency and representational robustness. A dynamic selection mechanism is introduced to adaptively switch between these two strategies based on local texture classification results, as illustrated in Fig. 2. For the dynamic switching threshold, we first employ XFeat-LightGlue to match feature points across four adjacent images; if the average number of matched feature points is fewer than 800, the algorithm switches to Gim-LightGlue for feature matching.

The Gim feature extraction module, constructed on the SuperPoint-LightGlue backbone, adopts a video-sequence-based training paradigm that incorporates temporal continuity constraints to enhance response sensitivity under low-texture conditions. This design is particularly effective in regions with sparse texture distributions or ambiguous structural edges. Gim is capable of producing stable sets of keypoints along with high-confidence descriptors. To further optimize processing efficiency in regions characterized by high texture density, this study integrates the XFeat algorithm [33], which generates a keypoint heatmap (K), a 64-dimensional compact descriptor map (F), and a confidence heatmap (R) in parallel. The architecture leverages early-stage downsampling and shallow convolutional layers to reduce computational overhead, while deep convolutional layers in the later stages enhance the expressiveness of extracted features. Unlike traditional monolithic architectures, XFeat decouples keypoint detection and descriptor extraction into separate modules. A 1×1 convolution operation is employed over 8×8 image patches to enable rapid, localized feature inference, thus achieving both feature decoupling and accelerated reasoning.

In the actual visual data collected from the underside of bridge structures, low-texture regions account for only approximately 7% of the total image area. Therefore, the adoption of a texture-aware model selection strategy enables dynamic deployment of feature matching algorithms based on local texture classification. This not only ensures robust matching accuracy in texture-deficient regions but also significantly improves overall computational efficiency, reduces dependency on hardware resources, and enhances the applicability of the proposed system to resource-constrained platforms scenarios. Concurrently, feature matching algorithms are undergoing a paradigm shift from traditional distance-based heuristic approaches to data-driven frameworks that leverage contextual modeling. Classical methods such as BruteForce and FLANN [34] primarily rely on minimizing Euclidean distances between descriptors. However, their performance degrades in visually challenging environments characterized by weak texture contrast, substantial scale variations, or inconsistent illumination, leading to poor robustness and low matching stability. In contrast, recent deep learning-based approaches, such as SuperGlue [35] and LightGlue [36], incorporate graph neural network architectures to model contextual dependencies among local features. These methods significantly enhance geometric consistency and matching reliability. The LightGlue algorithm employed in this study integrates a hierarchical attention mechanism comprising both self-attention and cross-attention modules at each processing layer. Positional encoding is also embedded to modulate visual descriptors dynamically, thereby strengthening spatial awareness.

Furthermore, a built-in confidence classifier is embedded within the architecture to adaptively control the inference process. If the current set of tentative correspondences yields low confidence scores, the algorithm proceeds to the next matching iteration; otherwise, if high-confidence correspondences are detected, a candidate selection module is triggered. This module finalizes the correspondence prediction by jointly evaluating pairwise similarity and one-to-one feature compatibility. By combining this “hierarchical confidenceearly termination mechanism” with “context-aware inference,” LightGlue achieves a favorable balance between computational efficiency and matching precision. It effectively suppresses redundant operations while maintaining high-quality inlier ratios and robust correspondence accuracy, particularly under complex bridge inspection conditions.

3.2 Smooth Seamless Stitching and Geometric Distortion Correction

Following the establishment of feature correspondences between image pairs, an enhanced robust matching algorithm RANSAC-USAC-FM-8PTS was employed to estimate the underlying geometric transformation while suppressing potential mismatches. This algorithm integrates the robustness of the classical RANSAC framework, the adaptive model selection strategy of Universal Sample Consensus (USAC), and a normalized eight-point algorithm for fundamental matrix estimation. Through this hybrid design, the method significantly improves the stability and accuracy of transformation matrix computation, particularly in the presence of outliers and inconsistent feature correspondences.

Specifically, considering a pair of matched points (x,x') between two images, they are required to satisfy the epipolar geometric constraint governed by the fundamental matrix F:

$$x^{'T}Fx = 0(1)$$

Here, $x = [u,v,1]^T$ and $x' = [u',v',1]^T$denote the normalized homogeneous coordinates of the matched image points. By constructing a linear matrix $A \in R^{N\times 9}$ from $N \geq 8$ pairs of correspondences, where each row corresponds to the expanded form of Equation (1), an initial estimate of the fundamental matrix $F$is obtained via Singular Value Decomposition (SVD) of A. Subsequently, a rank-2 constraint is imposed on F to enforce the geometric consistency, yielding the homography transformation matrix between the images. After obtaining the homography transformation matrix between the image pair, traditional stitching methods often suffer from artifacts and misaligned seams, particularly in regions with sparse feature distributions. To overcome these limitations, an adaptive seam search algorithm based on dynamic energy optimization is proposed to construct locally optimal stitching seams, thereby effectively suppressing stitching artifacts. The core of this method involves the formulation of a two-dimensional energy density function $E(x,y)$, which quantifies pixel-level color and structural inconsistencies between images I_1 and I_2:

$$E(x,y) = E_c(x,y)^2 + \lambda \cdot E_g(x,y)(2)$$

Here, $E_c(x,y)^2$ represents the color difference energy, designed to maintain a natural color transition across the stitching seam. Meanwhile, $E_g(x,y)$ corresponds to the absolute gradient difference, ensuring continuity of texture and structural edges. Specifically, cumulative energy is computed row-by-row from the top to the bottom of the image, where each pixel selects the minimum energy path among three upstream candidates left, middle, and right directions. Subsequently, an optimal seam path is generated by backtracking from the pixel with the lowest cumulative energy at the bottom. Pixels on either side of this seam are sampled from the respective source images, effectively mitigating seam-related artifacts. Following the determination of the optimal stitching seam, although the seam itself has been refined, visible discontinuities may still arise due to global illumination differences and local color temperature shifts at the boundary. To address these challenges and achieve more seamless image blending, a Laplacian pyramid fusion algorithm [37] was employed. This approach fundamentally relies on multi-scale image processing to ensure smooth transitions while preserving fine details. Initially, Gaussian filtering and downsampling are applied to the input images I_1 and I_2 respectively, resulting in the construction of multi-level Gaussian pyramids $\mathrm{G_k}$:

$$\mathrm{G_k = \downarrow(GB(G_{k-1}))(k = 1,2,...,n)(3)}$$

In the above notation, the symbol ↓ denotes the downsampling operation, while GB represents the Gaussian filtering function. Subsequently, the low-resolution images are upsampled back to their original dimensions and subtracted from the corresponding level of the original image to extract the detail layers L_k, which capture the high-frequency information. Here, ↑ denotes the upsampling operation.

$$L_k = G_k - \uparrow(G_{k+1})(k = 0,1,...,n-1)(4)$$

Simultaneously, a mask M, with values constrained within the range (0,1), is defined and its Gaussian pyramid is constructed to facilitate multi-scale smooth transitions. This procedure ensures that transitions between images at different scales remain as seamless as possible, effectively preventing the emergence of harsh edges and discontinuities. Subsequently, at each pyramid level k, the Laplacian coefficients are blended according to the weighted contributions determined by the mask

$$\begin{cases} M_k = \downarrow(GB(M_{k-1})) \\ L'_k = M_k \odot L_{1,k} + (1 - M_k) \odot L_{2,k} \end{cases}(5)$$

Here, the operator $\odot$ denotes element-wise multiplication. This operation enables the fusion of high-frequency details across multiple scales, thereby preserving the edges and textures of both images.

Subsequently, during the reconstruction phase, the process begins from the top-level Laplacian coefficients L'_k, which are progressively upsampled and combined with the Laplacian details at each lower level to restore the original image resolution. This multi-scale reconstruction strategy not only effectively mitigates visual inconsistencies caused by illumination differences or color temperature shifts but also ensures the richness and clarity of the final fused image.

$$R_k = \uparrow(R_{k+1}) + L'_k(k = n-1, n-2, \ldots, 0)(6)$$

Ultimately, the application of the aforementioned multi-scale fusion approach at the original image resolution yields a smoothly blended and seamless stitched image. This technique not only effectively addresses the visibility issues at the stitching seams but also enhances the overall image quality, resulting in a more natural and visually coherent output. In the context of continuous image stitching, to construct the global transformation matrix H_{is} between any two images, a cumulative multiplication method based on the homography matrices of adjacent image pairs is employed:

$$H_{is} = \prod_{i=s-1}^{i=1} H_{ii+1} \quad (7)$$

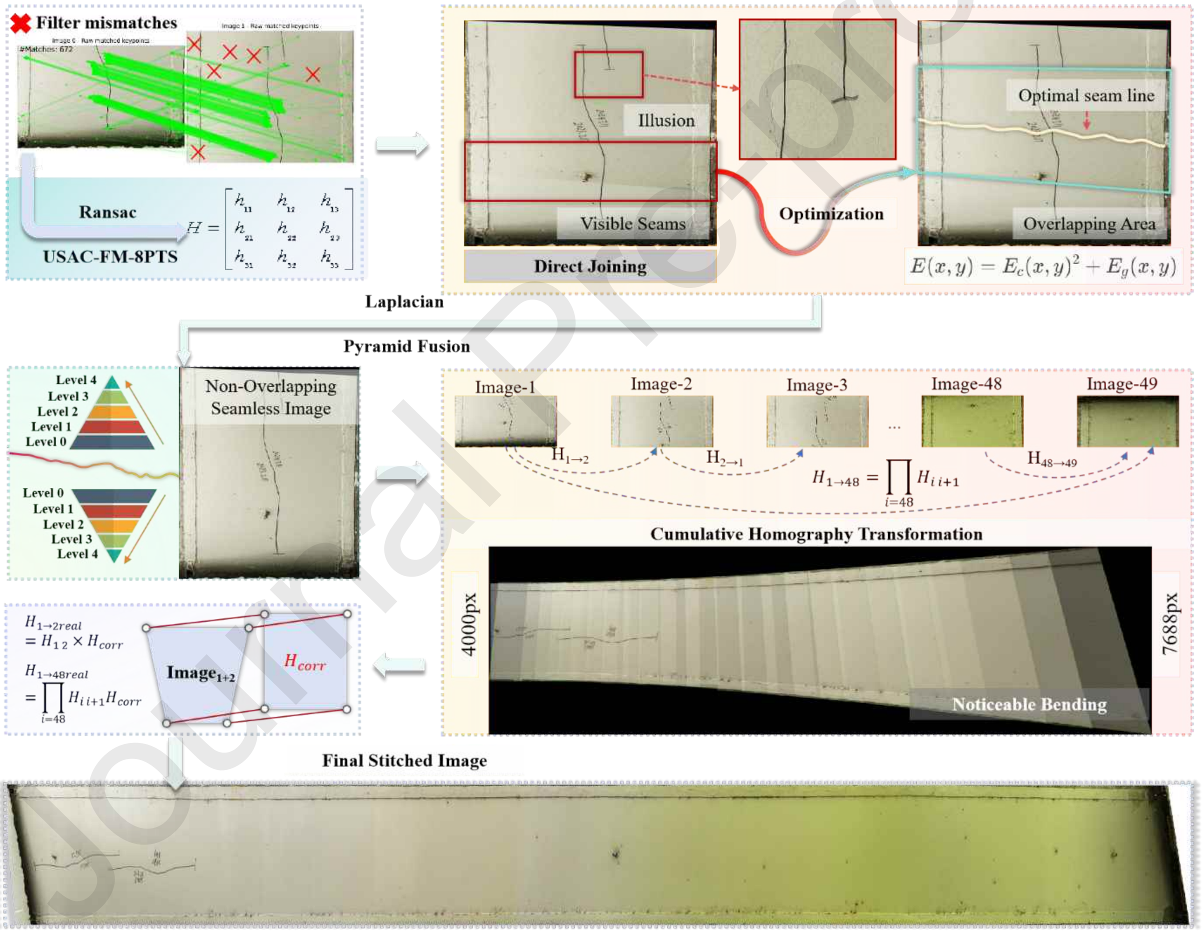


Fig. 3. Seamless and Smooth Panoramic Stitching and Curvature Correction Method for the Underside of Bridge Girders.

While this method establishes the coordinate mapping between the first image and the subsequent ones, it presents a critical limitation: errors in the homography matrix are progressively accumulated and amplified throughout the stitching process, ultimately leading to pronounced geometric distortions in the final composite. Such distortions are particularly evident in the loss of straightness and parallelism of beam structural features. As illustrated in Fig. 3, the stitching of 20 uncorrected images demonstrates that the curvature becomes increasingly severe with more images involved, as the image width expands from 4000 px to 7688 px, resulting in substantial deformation. To address these issues, the present study proposes an innovative enhancement by introducing an optimization mechanism constrained by structural features. Specifically, the linear characteristics of the beam edges are exploited to impose parallelism constraints. A correction matrix is constructed based on the mapping relationship between the source points $P_s^{(i)}$ located at the four corners of the original image and the corresponding target points $P_D^{(i)}$ lying on the beam's edge lines. This yields a correction matrix H_{corr}, which is applied to perform nonlinear refinement of the cumulative transformation matrix during the global optimization stage, as expressed in Equations (8):

$$\begin{cases} H_{is} = \prod_{i=s-1}^{i=1} H_{ii+1} H_{\text{corr}}^{(i)} \\ P_D^{(i)} = H_{\text{corr}}^{(i)} \cdot P_{\text{s}}^{(i)} \end{cases} \quad (8)$$

This strategy not only effectively mitigates the propagation of errors but also ensures the geometric accuracy of critical structural features, thereby providing a robust and reliable solution for large-scale image stitching in complex scenarios such as the underside of bridges. Furthermore, the proposed method enhances the overall quality and visual consistency of the stitched imagery, which is crucial for improving the accuracy and efficiency of subsequent analytical tasks. A schematic illustration of the implementation is presented in Fig. 3.

3.3 Matrix-Encoding-Driven Defect Spatial Localization Mechanism

To achieve precise mapping between detected defects and the physical locations on bridge structures, while ensuring the spatial localization and retrievability of subsequent identification data, this study proposes a structured matrix segmentation and encoding mechanism tailored for panoramic images. This mechanism introduces a two-dimensional spatial coordinate system to partition the image into discrete blocks, each assigned a unique spatial identifier, thereby establishing a bidirectional mapping framework bridging the image domain and the actual structural domain of the bridge girder.

For each panoramic image of the bridge girder bottom, which has been fully stitched and subjected to perspective correction, an initial structural segmentation is performed by treating each individual girder as a fundamental unit. This ensures that each segmented region exclusively encompasses a continuous section of the girder bottom. Subsequently, each soffit region of the beam was further divided into multiple non-overlapping matrix image units U_1, U_2...U_n, each with a basic size of 500 × 500 pixels, forming a spatial grid. This size was chosen for two reasons: first, the input and output of the subsequent crack segmentation model TransUnet are both 512 × 512, making 500 × 500 close to the training scale; second, the vertical dimension of the panoramic stitched image is 4000 × 53333, which can be neatly divided into 500-pixel units, facilitating processing. To guarantee the invertibility and uniqueness of the spatial correspondence between the image and the actual structure, each image block is assigned a two-dimensional integer coordinate index (x,y), where x denotes the horizontal segmentation number and y the vertical segmentation number. These image units collectively provide a complete, non-overlapping coverage of the stitched image within the image coordinate system.

Following this encoding and labeling procedure, each individual image unit is independently processed by the defect recognition model for segmentation. The identified defect areas are annotated within their respective matrix blocks as binary masks. Due to the unique spatial labeling of each image block, the global position of any detected crack can be accurately back-projected from the block's index (x,y) combined with the local pixel coordinates (u,v) within that block. A schematic illustration is provided in Figure 10 of Section 5. Each detected defect is thus explicitly linked to a unique spatial location, enhancing the interpretability of recognition results in the structural context. By incorporating the linear scaling relationship between pixel coordinates in the stitched image and the actual physical coordinates of the bridge, the physical position of cracks within the real bridge structure can be precisely indexed. Mathematically, this approach realizes a tripartite mapping sequence: from image space to encoded matrix space, and subsequently to structural spatial domain, thereby ensuring high consistency and traceability of defect localization. The proposed matrix-based image segmentation and spatial encoding framework not only establishes a rigorous mapping between visual data and structural geometry but also facilitates standardized and structured data organization and information representation for bridge defect detection systems. This foundation is critical for enabling intelligent, spatiotemporal bridge lifecycle monitoring and decision support in maintenance and operation.

4. Deep Learning-Based Automated Defect Analysis

4.1 TransUnet: A Structure for Vision Tasks

After constructing the matrix-formatted images of the bridge beam underside incorporating spatial encoding information, a high-performance semantic segmentation model, TransUNet[38], was employed to achieve both high-precision identification of defect regions and accurate localization at the structural scale. This architecture combines the global modeling capability of the Transformer module, which excels in capturing long-range dependencies, with the multi-scale local feature extraction efficiency of the U-Net framework. As a hybrid CNN-Transformer deep neural network, TransUNet is particularly well-suited for segmentation tasks that demand simultaneous global semantic consistency and fine-grained boundary delineation, making it an ideal choice for pixel-level crack segmentation on bridge structures.

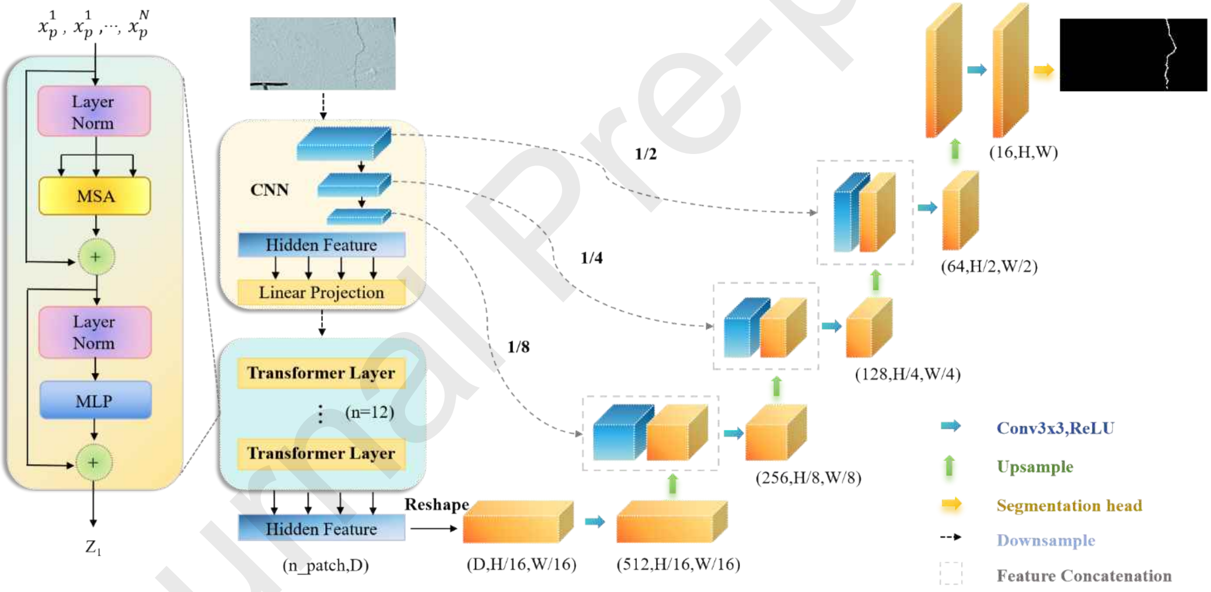


Fig. 4. Structure of TransUnet network.

The overall architecture of TransUNet can be delineated into three fundamental stages: encoding, global modeling, and decoding. Initially, during the encoding stage, a convolutional neural network (CNN) module based on a ResNet backbone is employed to perform hierarchical multi-scale downsampling on the input images. This process progressively extracts local spatial structural information and edge details, while preserving feature maps at each scale for subsequent utilization in skip connections during the decoding phase. By compressing the spatial dimensions of the image, this stage effectively retains essential semantic information, thereby laying the groundwork for the ensuing global modeling module. Subsequently, the low-resolution feature maps produced by the encoder are fed into the Transformer module. Within this module, the feature maps are first flattened into a one-dimensional sequence and augmented with learnable two-dimensional positional encodings to compensate for spatial location information lost during convolution. The core Transformer consists of multiple stacked encoder layers, each employing multi-head self-attention mechanisms to capture contextual relationships across distant regions of the image. This facilitates the integration and enhancement of semantic representations across scales and spatial domains. Such a design significantly expands the receptive field and bolsters the model's capacity to represent macroscopic semantic structures, including crack connectivity and branching topology. The decoding stage follows a symmetric path consistent with the U-Net architecture, where spatial resolution is gradually restored through successive upsampling operations. At each decoding level, the global feature maps generated by the Transformer module are concatenated along the channel dimension with the corresponding local feature maps retained from the encoder via skip connections. This fusion of multi-scale information ensures the preservation of global semantic consistency alongside local boundary details. Ultimately, several convolutional layers compress the channel dimension to produce the final segmentation probability map, enabling pixel-level precise delineation of crack regions.

To address class imbalance in crack detection, we employed a weighted combination of Cross-Entropy Loss and Dice Loss. Cross-Entropy Loss guides pixel-level classification, while Dice Loss is sensitive to small target regions. Their integration improves overall accuracy and enhances learning for fine crack structures.As shown in Fig. 5, the weighted loss achieves smooth convergence: rapid initial decline due to Cross-Entropy Loss, followed by Dice Loss-driven optimization of structural similarity, stabilizing the training and reaching a low loss (~0.045). Notably, IoU results demonstrate the effectiveness of this combination: 62.9% with Cross-Entropy alone, 76.1% with Dice alone, and 77.6% with the weighted loss.

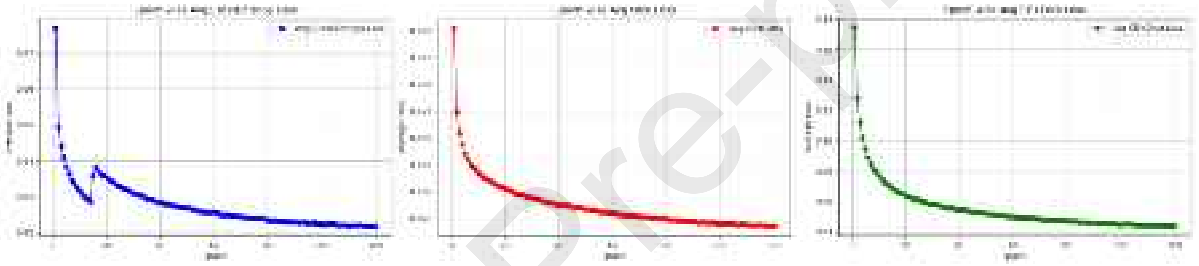


Fig. 5. Loss curve

4.2 Implementation Details

To ensure the practical applicability of the crack segmentation model, this study constructed a comprehensive, finely annotated, and structurally standardized crack image dataset that reflects the typical defect characteristics of hollow slab bridge soffits. A systematic training framework and computational environment were configured accordingly to enhance model convergence, generalization, and portability. The defect dataset was collected from multiple bridges, including the Humen Bridge, laboratory beam segments, and highway bridges, forming a self-constructed repository of defect images. All defect regions were manually annotated frame by frame using the Labelme tool, with the study rooted in real-world engineering problems and involving collaborations with several inspection companies. For quality control, five field engineers were invited to independently verify all annotations, ensuring the reliability of the dataset. The finalized dataset consists of 2,993 images, each standardized to a resolution of 1024 × 512 pixels, with consistent image scales and label formats. Following the mainstream paradigm of supervised learning, the dataset was partitioned into training and validation sets at a 9:1 ratio, comprising 2,694 training images and 299 validation images, thereby balancing sufficient model learning with robust evaluation of generalization to unseen data. The defect statistics are as follows: 2,632 images contain cracks and 361 images are crack-free, with all cracks corresponding to joints less than 1 mm in width.

The training environment was constructed on the PyTorch 2.0.0 deep learning framework, with model building and parameter optimization performed under this platform. Computational acceleration was facilitated by an NVIDIA GeForce RTX 3090 GPU equipped with 24 GB of video memory. The operating system employed was Windows 11, with the runtime environment configured using Python version 3.9.7 and

CUDA platform version 11.8.

To comprehensively evaluate the applicability and performance of the proposed TransUNet model in bridge crack segmentation tasks, a systematic training and comparative experimental framework was established based on the constructed high-quality crack image dataset. Intersection over Union (IoU) was selected as the primary evaluation metric due to its capability to precisely quantify the spatial overlap between predicted segmentation regions and ground truth annotations, thereby directly reflecting the model’s geometric accuracy in target region delineation. Considering the prevalent issue of severe class imbalance inherent in crack detection tasks, additional metrics including recall and precision were incorporated to assess the model’s ability to minimize false negatives and false positives, respectively. The F1 score was further computed to provide a balanced measure that integrates both recall and precision, thereby offering a more comprehensive reflection of the model’s discriminative effectiveness in practical scenarios. Moreover, inference time was introduced as a critical metric to evaluate computational efficiency, quantifying the model’s real-time responsiveness during the testing phase. This holistic set of performance indicators ensures a thorough assessment of the model’s segmentation accuracy and operational viability within bridge inspection applications.

In terms of training strategy, a progressive learning rate decay schedule was adopted to optimize convergence performance. The initial learning rate was set to 0.0005, with a batch size of 12. The total number of training epochs was configured to 120, and the model parameters were updated using the Adam optimizer throughout the training process.

Table 1

Performance comparison of different models

Crack Segmentation					
Network	IOU(%)	Precision (%)	Recall(%)	F1(%)	Inference Time(s)
Unet	74.256	79.449	91.991	85.168	0.032
Unet++	56.408	58.850	85.721	69.170	0.055
Swinunet	68.672	71.822	88.232	79.010	0.030
Deeplabv3	59.904	57.395	98.578	72.408	0.022
TransUnet	77.616	81.345	91.152	85.863	0.042

Notes: (1) #Inference Time: The time (in seconds) a model takes to process an input sample; the lower the value, the better the model's real-time performance;

To thoroughly evaluate the performance of TransUNet in crack segmentation tasks, a comparative study was conducted against four representative deep learning-based semantic segmentation models: U-Net [39], U-Net++ [40], Swin-Unet [41], and DeepLabv3 [42]. All models were trained and tested under identical data conditions and hyperparameter settings to ensure fairness in comparison. The experimental results are summarized in Table 1. As evidenced by the evaluation metrics, TransUNet consistently outperformed the baseline models across several key indicators, demonstrating a marked advantage in overall segmentation quality. Specifically, TransUNet achieved an average Intersection over Union (IoU) of 77.62%, representing an improvement of 3.35% and 6.71% over U-Net and U-Net++, respectively. In terms of precision, TransUNet reached 81.35%, significantly surpassing U-Net++ (58.85%) and Swin-Unet (71.82%). While its recall rate of 91.15% was marginally lower than that of DeepLabv3 (98.57%), TransUNet yielded an F1 score of 85.86%, substantially higher than DeepLabv3’s 72.41%. This result indicates that TransUNet achieved a more desirable balance between recall and precision, which is critical in reducing both missed and false detections in practical applications. Although TransUNet's inference time (0.042 seconds per image) was slightly slower than that of DeepLabv3 (0.022 seconds), the superior segmentation accuracy of

TransUNet offsets this minor difference, enhancing its overall practical applicability and reliability for automated crack detection in structural inspections.

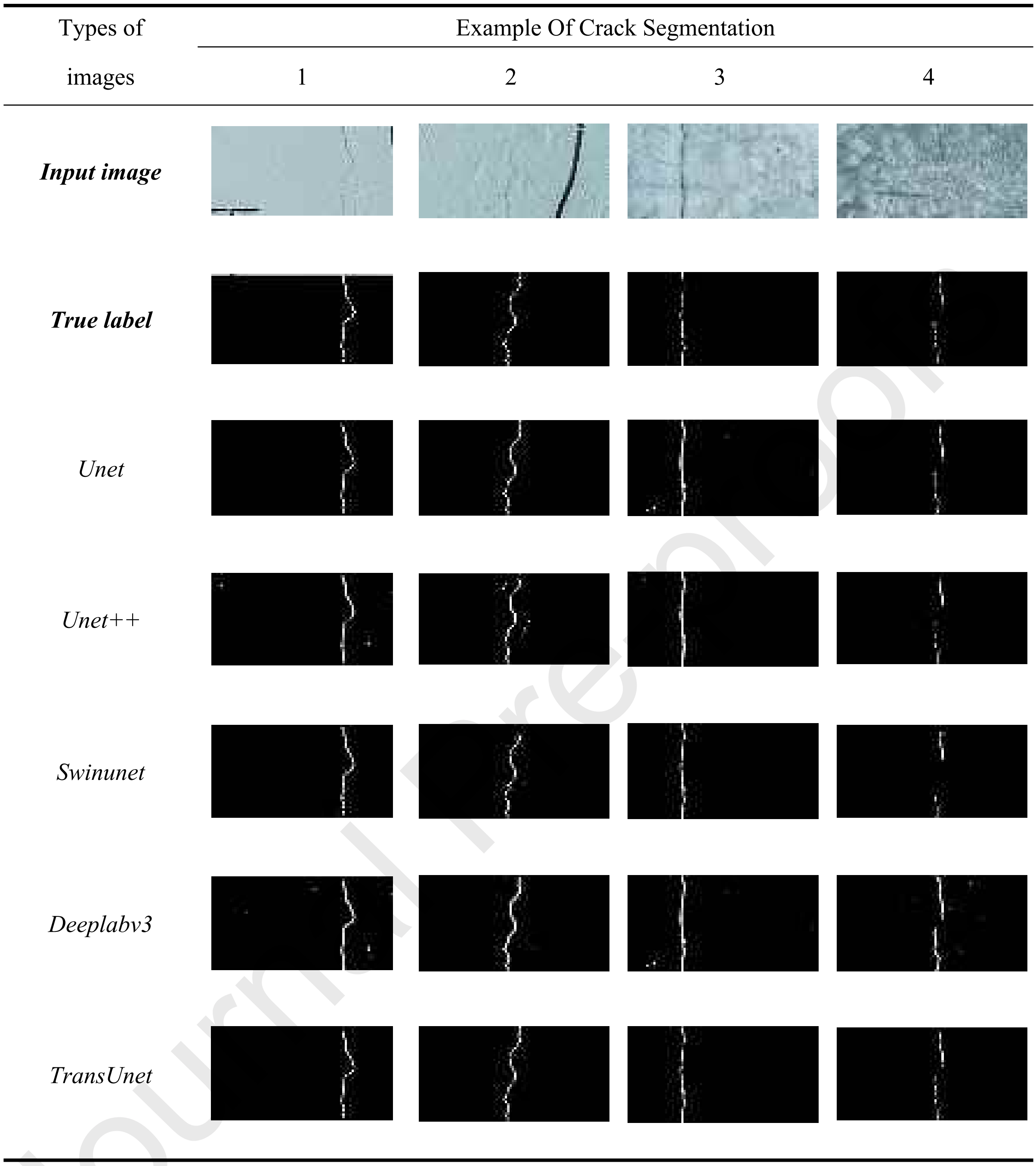


Fig. 6. Visualized results of different networks**.**

As illustrated in Fig. 6, the visualized segmentation results of different models on representative crack images reveal notable variations in their performance across diverse crack scenarios. Overall, TransUNet demonstrated superior accuracy in delineating crack boundaries, maintaining structural continuity, and preserving fine-grained details. The predicted segmentation maps closely aligned with the ground truth in both spatial distribution and morphological characteristics, indicating the model's robust capability in detecting subtle features and modeling semantic context. For prominent and well-defined cracks, the U-Net model was also able to produce reasonably accurate segmentation contours. However, its performance declined in terms of boundary completeness and the identification of thinner or fragmented cracks, where TransUNet showed a clear advantage. In more challenging cases involving low-contrast or visually ambiguous cracks, TransUNet maintained consistent segmentation quality, whereas the other models exhibited a higher rate of false positives or false negatives. These errors were often manifested as over-segmentation of background regions or the disconnection of crack structures. Notably, although DeepLabv3 achieved the highest recall (98.578%) as reported in Table 1, its segmentation maps exhibited a pronounced tendency toward over-expansion, frequently encompassing non-crack areas beyond the actual defect boundaries. While such over-segmentation contributed to elevated recall values, it simultaneously compromised spatial precision and overall segmentation reliability. In contrast, TransUNet not only delivered superior quantitative performance but also achieved enhanced visual coherence, structural fidelity, and robustness against noise and background interference. These findings collectively validate the practical effectiveness and engineering applicability of TransUNet in fine-grained bridge crack segmentation tasks, especially under complex real-world conditions.

5. Field Validation Experiments

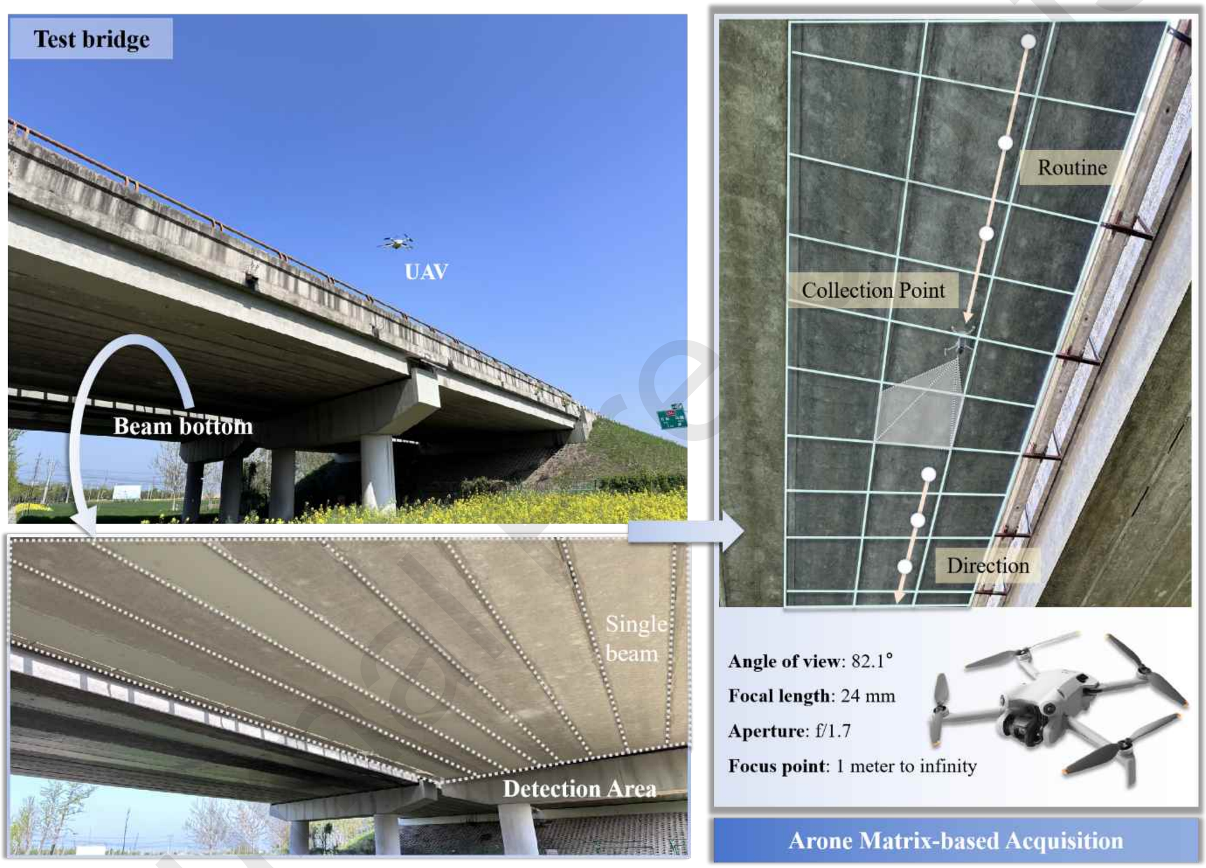


Fig. 7. Schematic diagram of bridge overview.

To systematically validate the adaptability and effectiveness of the proposed visual inspection and defect localization method for the underside of bridge girders in real-world engineering scenarios, a representative field experiment was conducted on a typical hollow slab highway bridge located at the intersection of the S49 Xinyang Expressway and the X101 Xuma Road. The experimental setup is illustrated in Fig. 7. During the data acquisition phase, a DJI Mini 4 Pro lightweight multi-rotor UAV was deployed as the aerial platform. Following a predefined matrix-style flight path, the UAV performed systematic image scanning of the bridge’s underside with a high degree of overlap. This acquisition strategy not only ensured comprehensive coverage of the inspection area but also provided abundant multi-view redundant information, which proved beneficial for subsequent image stitching and feature extraction processes. Based on the collected image set, a comparative analysis was conducted to evaluate the performance of several mainstream feature matching models in terms of detection accuracy, matching robustness, and computational efficiency, with the aim of assessing their suitability for bridge defect identification tasks. Subsequently, using the proposed multi-stage image stitching and geometric correction reconstruction algorithm, the acquired images were fused into a high-resolution panoramic representation of the bridge’s underside. The final stitched image extended to a length of 260 meters, offering full geometric consistency and complete coverage of all structural components beneath the target bridge. Following successful defect recognition, the identified damage regions—such as cracks—were accurately mapped from image coordinates back into the physical coordinate system of the bridge structure. This reverse mapping was achieved through the matrix-based spatial encoding mechanism introduced earlier in this study, thereby enabling precise localization of the defects within the real-world spatial context. Field test results indicate that the panoramic stitching for a single beam takes approximately 10 minutes, whereas the 3D modeling process based on Colmap requires over 2 hours, demonstrating a significant improvement in efficiency. In terms of localization accuracy, the Colmap method achieves an average precision of 0.25 mm, while the proposed corrected panoramic stitching method attains 0.375 mm. Although there is a slight decrease in accuracy, the substantial gain in efficiency underscores the method's strong engineering applicability. Furthermore, the proposed beam soffit crack detection and structure-level localization approach exhibits consistent recognition stability and spatial consistency even in complex structural scenarios, validating its feasibility and practical value for bridge maintenance applications.

To evaluate the adaptability and stability of various feature matching algorithms in the context of underside bridge girder imagery, a standardized testing dataset was constructed to account for the surface smoothness variations commonly encountered during actual inspection scenarios. Given that certain bridge girder undersides are treated with specialized coatings to enhance durability, thereby significantly diminishing surface texture saliency and complicating feature extraction—the dataset was stratified based on texture intensity. Specifically, the image samples were divided into two subsets: low-texture and high-texture, each comprising 42 images, yielding a total of 84 image pairs. All images were uniformly resized to a resolution of 2000 × 2000 pixels to ensure experimental consistency and result comparability across all evaluations. In terms of experimental configuration, feature detection parameters were standardized across all images, with the number of features (nfeatures) set to 4000. RANSAC-based geometric verification was applied using fixed parameters: a reprojection threshold of 1.0 pixels, a confidence level of 0.99, and a maximum of 10,000 iterations. All inference tasks were executed on an NVIDIA RTX 4060 GPU equipped with 8 GB of memory, providing a consistent computational environment for fair benchmarking. The feature matching pipeline was built upon the unified LightGlue architecture to maintain algorithmic consistency. For algorithms without pretrained LightGlue-compatible models, such as ORB native matching implementations were employed instead. This controlled experimental design facilitated a systematic and reproducible comparison of different feature matching strategies under challenging low-texture conditions typically observed in real-world bridge inspection environments.

The performance of mainstream feature detection algorithms, including ORB, SIFT, SuperPoint, ALike, DISK, GIM, XFeat Sparse, and XFeat Dense was comprehensively evaluated based on three key metrics: average inference speed (measured in frames per second, FPS), the number of inliers obtained through RANSAC (reflecting feature matching density), and inlier ratio (representing matching accuracy). These criteria collectively capture the trade-offs among accuracy, computational efficiency, and robustness of each method when applied to bottom-view bridge imagery. As illustrated in Fig. 8, GIM consistently demonstrated superior stability under weak-texture conditions, achieving an average RANSAC inlier count of 1839 and a perfect matching success rate of 100%, significantly outperforming the other models. This indicates that GIM maintains strong response capability even in scenarios with sparse texture information. Although its inference speed averaged 7.87 FPS slightly lower than XFeat’s 10.43 FPS its ability to preserve high-precision features in low-texture regions proved critical for stable geometric model estimation, thereby making it the preferred option for such challenging conditions. In contrast, while XFeat excelled in computational speed, its performance in precision was less satisfactory, with an average inlier count of 1105 and a matching success rate of only 63.60%, rendering it less suitable for tasks requiring high geometric accuracy. ALike and DISK exhibited moderate performance, with ALike achieving 1255 inliers and 66.67% matching accuracy at 7.99 FPS, and DISK achieving 1246 inliers and 65.18% at 4.21 FPS. However, both remained notably inferior to GIM, with a mean inlier count gap exceeding 580.

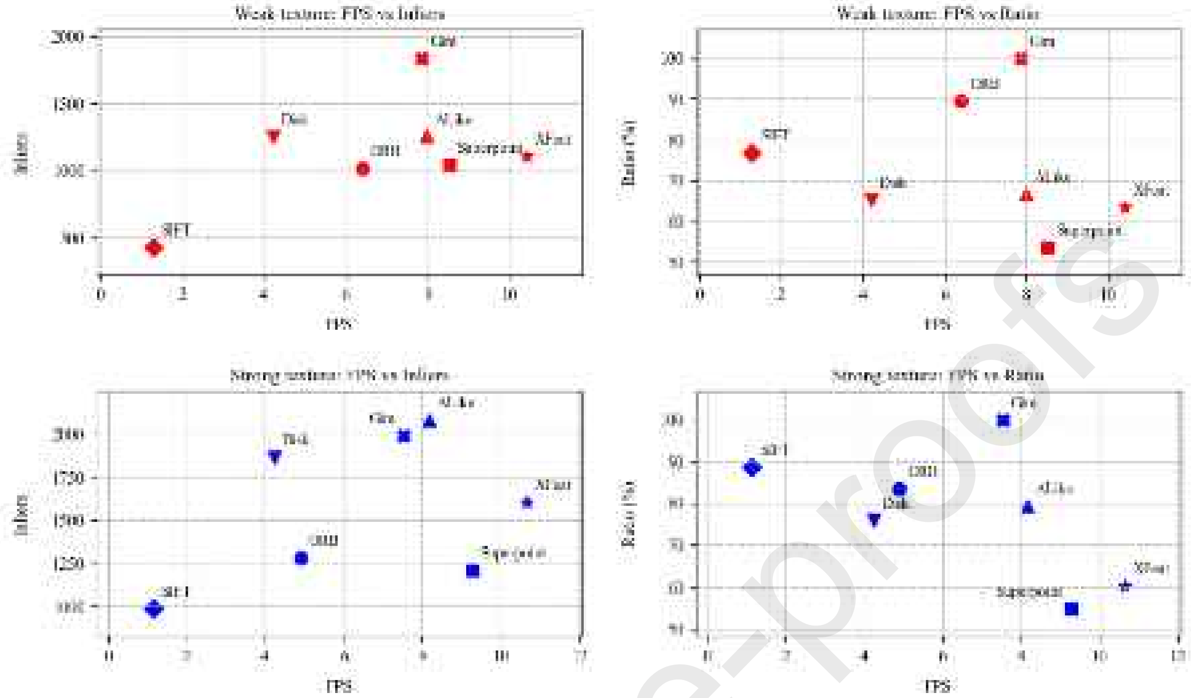


Fig. 8. Performance Comparison of Feature Detectors in Terms of Accuracy and Inference Time.

In high-texture image subsets, all models showed improved matching density due to enhanced surface structural information. However, differences in computational efficiency remained relatively unchanged. Under these conditions, GIM, ALike, and DISK all maintained high matching precision, albeit with slower inference speeds. Notably, XFeat emerged as the optimal choice for texture-rich scenarios, achieving the highest processing speed (10.66 FPS) while maintaining sufficient matching density and a favorable inlier ratio, thus making it well-suited for real-time panoramic stitching tasks. Moreover, the experimental results underscored a substantial impact of image resolution on algorithmic performance. Hand-crafted methods such as ORB and SIFT experienced dramatic increases in computation time with higher resolutions. For instance, ORB achieved an average inference speed of 25.79 FPS at 640×640 resolution, significantly outpacing the second-fastest XFeat by approximately 15 FPS, highlighting its efficiency in low-resolution contexts. However, ORB's inlier count exhibited large fluctuations under high-resolution conditions, ranging from over a thousand to fewer than a hundred, indicating poor consistency across samples. In contrast, deep learning-based approaches such as GIM and XFeat demonstrated more stable behavior, with inlier count deviations typically confined within ±200. This consistency indicates superior robustness and generalization capacity, making them more suitable for deployment in real-world structural inspection scenarios with variable image quality.

To provide a more intuitive comparison of the feature matching performance of various detection algorithms under weak-texture and strong-texture conditions, two representative image samples were selected from the standardized test dataset for visual analysis, as illustrated in Fig. 9. In each image, green lines denote valid inlier matches filtered by the RANSAC algorithm, while the total number of inliers is annotated below the corresponding image to quantitatively reflect the feature matching density. In the weak-texture scenario, the ORB algorithm yielded only 83 valid matches, confirming a marked decline in its feature extraction capability in regions with limited texture information. By contrast, algorithms such as SIFT (232 inliers), SuperPoint (1000 inliers), and DISK (831 inliers) demonstrated significantly improved performance.

Notably, GIM achieved 1443 RANSAC inliers in the same image, with dense and uniformly distributed matching lines, clearly outperforming all other methods. These results underscore GIM's superior adaptability and robustness in texture-deficient environments. Under strong-texture conditions, all algorithms exhibited improved matching performance due to the increased availability of structural features. GIM continued to deliver strong results, achieving 1865 inliers, thereby demonstrating its stability even in highly textured scenes with redundant information. Meanwhile, XFeat achieved a favorable trade-off between accuracy and computational efficiency, with 1289 inliers and the highest processing speed among the evaluated models. This balance makes XFeat particularly suitable for real-time applications where both responsiveness and precision are essential.

Furthermore, the B-2 beam segment which has the fewest feature points among the 13 tested beam segments was selected for detailed comparison. Its 41 images were processed using three individual image matching algorithms: SIFT, Gim-LightGlue, and XFeat-LightGlue, as well as the proposed dynamic switching strategy. The feature matching time and accuracy of each method were recorded (see Table below). In the dynamic switching strategy, when the number of feature points falls below 800 and the matching accuracy is lower than 60%, the system automatically switches to the GIM model, thereby ensuring reliable matching accuracy even under extreme conditions. Experimental results show that, compared with the traditional SIFT method, the dynamic switching strategy improves matching speed by 81.65% while maintaining comparable accuracy; compared with the GIM method, it achieves a 40.73% speed improvement. This strategy not only demonstrates superior performance but also significantly enhances matching stability under challenging conditions, greatly improving algorithm robustness and providing a reliable foundation for subsequent image stitching.

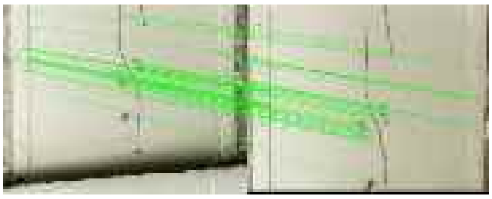
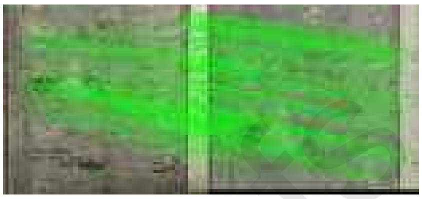
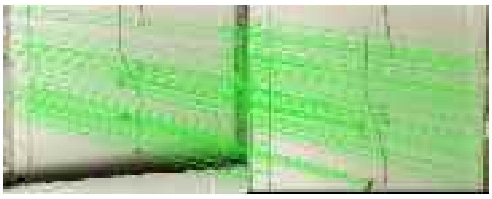
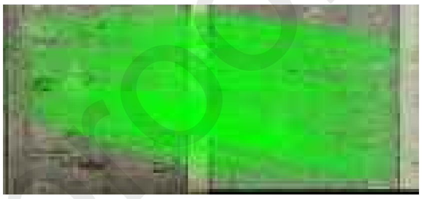
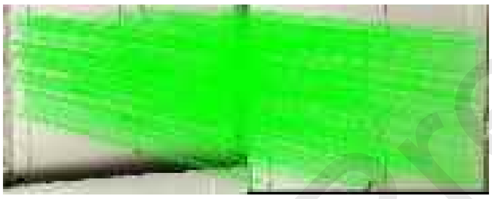
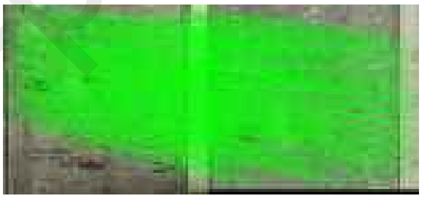
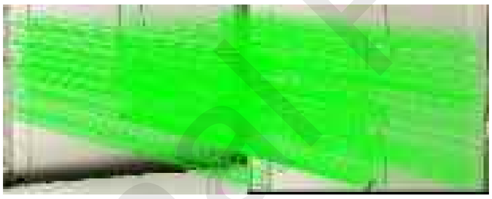
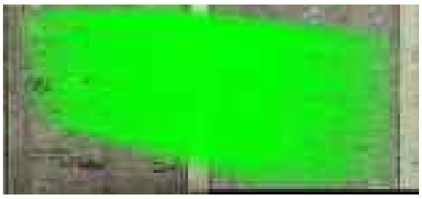
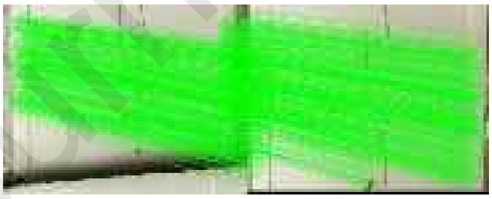
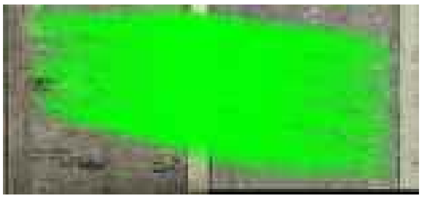
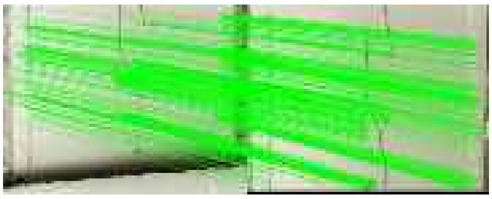
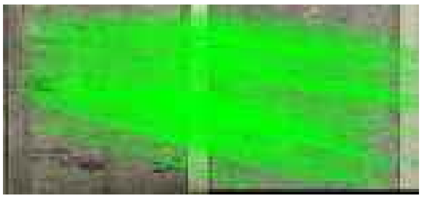
Table 2

Comparative experiments were conducted using the dynamic switching strategy.

Image/Model	SIFT Time(s)/Ratio	XFeat Time(s)/Ratio	Gim Time(s)/Ratio	Dynamic strategy Time(s)/Ratio
0	0.64/0.596	0.08/0.613	0.3/1.00	0.08/0.613
1	0.67/0.672	0.08/0.587	0.31/1.00	0.41/1.00
2	0.77/0.681	0.12/0.651	0.08/1.00	0.12/0.651
3	0.91/0.655	0.24/0.692	0.33/1.00	0.24/0.692
4	0.88/0.752	0.05/0.701	0.24/1.00	0.05/0.701
……	……	……	……	……
39	0.69/0.701	0.08/0.607	0.32/1.00	0.08/0.607
40	0.75/0.792	0.13/0.624	0.25/1.00	0.13/0.624
Avg	0.801/0.701	0.118/0.635	0.248/1	0.147/0.6843

Following the initial feature extraction and coarse matching using the XFeat-LightGlue combination model, a systematic evaluation was conducted to enhance the robustness and geometric accuracy of feature alignment. Specifically, this study assessed the performance of various robust estimation algorithms based on the RANSAC framework, focusing on their behavior under different inlier filtering strategies and model refinement mechanisms. The evaluation aimed to identify trade-offs between estimation precision and computational efficiency when applied to the proposed inspection dataset. All experiments were conducted under a unified parameter configuration to ensure consistency and comparability. The reprojection error threshold was set to 8.0 pixels, with a confidence level of 0.99 and a maximum of 10,000 iterations. Eight representative algorithms were selected for comparative analysis, including the classical RANSAC method and several variants within the Universal Sample Consensus (USAC) framework—namely, USAC-DEFAULT, USAC-FM-8PTS, USAC-PROSAC, USAC-FAST, USAC-ACCURATE, USAC-

PARALLEL—as well as the MAGSAC-enhanced version, USAC-MAGSAC. For each algorithm, 1,000 matched feature point pairs were used as input to perform inlier identification and homography matrix estimation. This standardized input ensured the fairness and reproducibility of performance comparisons. Experimental results

	Examples Of Feature Matching	
Method	Weak Texture	Significant Texture
ORB	 Ransac Matches：83	 Ransac Matches：557
SIFT	 Ransac Matches：232	 Ransac Matches：912
Superpoint	 Ransac Matches：1000	 Ransac Matches：1053
Alike	 Ransac Matches：965	 Ransac Matches：2015
Disk	 Ransac Matches：831	 Ransac Matches：1813
Xfeat	 Ransac Matches：679	 Ransac Matches：1289

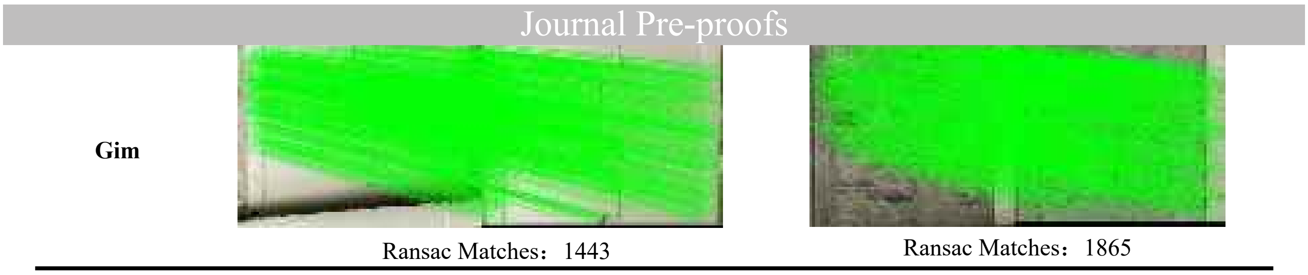
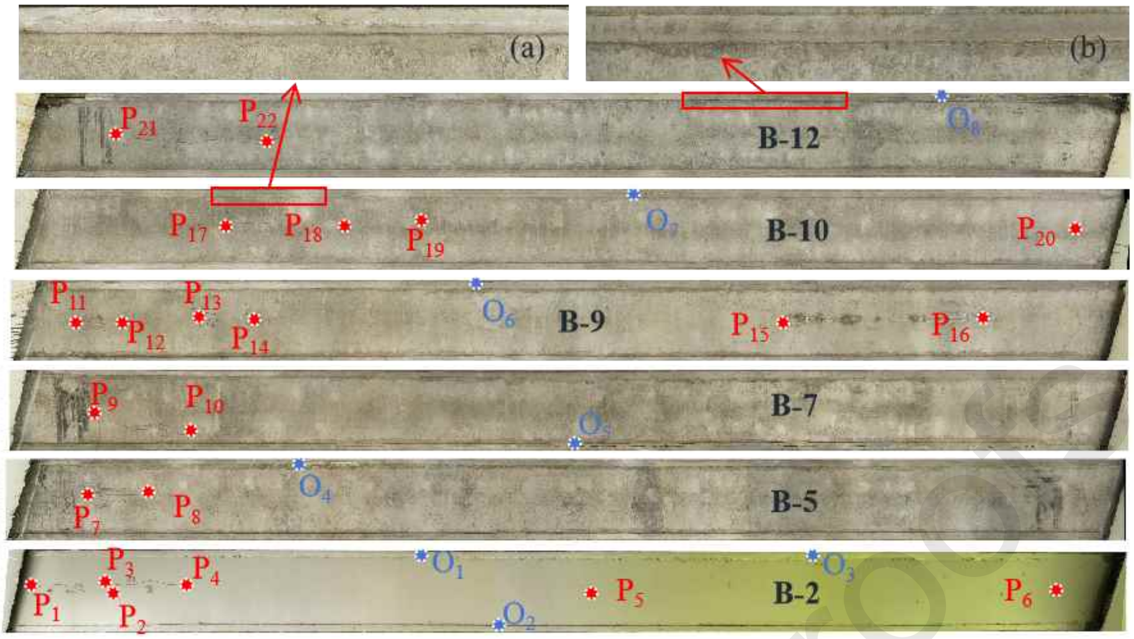


Fig. 9. Visual comparison of feature matching results across different detection algorithms.

demonstrated that the USAC-FM-8PTS algorithm achieved superior performance in both accuracy and efficiency. Among all evaluated methods, it yielded the highest inlier ratio, reaching 91.81%, while also reducing the average computational time by 9.5 milliseconds compared to the classical RANSAC. These results highlight its suitability for real-time geometric model estimation in high-resolution visual inspection tasks.

Building on this foundation, regardless of the algorithm used, the number of valid and repeatable feature points consistently falls below approximately 800 in low-texture regions such as smooth concrete surfaces or shaded areas, whereas typical well-textured regions contain 1000–2000 stable points. This consistent statistical boundary was adopted as the lower confidence limit for reliable feature extraction. In addition, when the inlier ratio of geometrically verified correspondences drops below 60%, the likelihood of mismatch propagation increases significantly, leading to local alignment errors during stitching. Accordingly, the dynamic switching module is designed to activate the GIM model automatically when both criteria, feature point count below 800 and inlier ratio below 60% are simultaneously satisfied. This adaptive mechanism ensures geometric consistency and localization reliability across varying visual conditions without compromising computational efficiency, as validated in subsequent sections.



Measurement result /mm	Calculation results/mm	Vertical Error/mm	Horizontal Error/mm
P_1(409.700, 811.010)	(409.875, 811.125)	0.115	0.175
P_2(2103.550, 460.380)	(2103.750, 460.500)	0.120	0.200 ....
P_{15}(14083.935, 728.120)	(14084.250, 728.250)	0.130	0.315
P_{16}(17396.309, 678.245)	(17396.625, 678.375)	0.13	0.316 ....
P_1-P_{22}（Avg）	**P_1-P_{22}（Avg）**	**0.128**	**0.257**
O_1(7367.785,1371.994)	(7368.000, 1372.125)	0.131	0.215 ....
O_8(16476.431,1398.241)	(16476.750, 1398.375)	0.134	0.319
O_1-O_8（Avg）	**O_1-O_8（Avg）**	**0.136**	**0.224**

Fig. 10. Illustration of the localization accuracy achieved by the proposed approach.

To achieve high-precision and wide-coverage visual reconstruction of the underside structures of bridges, a representative beam section was reconstructed based on the panoramic stitching strategy proposed in Section 3. The selected specimen was a prestressed hollow slab beam measuring 1.5 m in width (transverse) and 20.0 m in length (longitudinal). After undergoing perspective correction and homography-based image warping, the output image reached a spatial resolution of 4000 × 53,333 pixels, corresponding to a physical ground sampling distance (GSD) of 0.375 mm/pixel. This resolution meets the stringent requirements for millimeter-level crack detection in structural health monitoring. To quantitatively evaluate the performance of the stitching and reconstruction system in terms of spatial resolution and positional mapping consistency, a series of high-precision manual measurements were conducted on representative crack regions located on Beam No. 2 during the image acquisition phase. In the subsequent post-processing stage, inverse geometric decoding was applied to extract the pixel coordinates and matrix-coded positions of these cracks from the reconstructed images. The estimated spatial values were validated against ground-truth measurements obtained on site, as shown in Figure 10. Taking the lower-left corner of the beam as the reference point, we measured 22 points at various locations on different beams and compared the computed coordinates with the corresponding measured coordinates. Additionally, eight extra points were placed along the beam edges to further verify the stability of edge localization. The mapping errors between all selected crack feature points in the image space and their counterparts in the physical bridge space remained within an acceptable range, with an average positioning accuracy better than 0.3 mm. The results demonstrate that even in complex geometric scenarios, the proposed image stitching, encoding, and solving framework maintains high consistency and exhibits strong engineering feasibility.

Table 3

Comparison Table of Grid Size, Detection Time, and Accuracy.

Patch-size	Time (s)	Iou(%)
100	768.8416	43.52%
200	192.6928	53.48%
300	90.7056	68.24%
400	49.1616	80.67%
500	31.6976	79.88%
600	23.6992	72.37%
700	17.9328	65.54%
800	13.3136	59.87%
900	12.2544	56.89%
1000	9.0336	51.25%

During the panoramic image generation phase, the bottom surfaces of 13 individual bridge segments were sequentially reconstructed using the previously described image stitching and rectification algorithms. This process resulted in the creation of a comprehensive bottom-view panoramic image set covering the entire bridge, as illustrated in Fig. 11. Each stitched image corresponds uniquely to a specific structural segment of the bridge, ensuring unambiguous spatial localization. The cumulative length of the stitched images reaches 260.0 meters, thereby achieving full coverage of critical inspection zones beneath the bridge and establishing a foundational dataset for cross-segment crack detection and spatial tracking. To mitigate the computational burden associated with high-resolution imagery during defect localization and to enhance the accuracy of spatial mapping, a matrix-based localization framework was introduced. This system employs a hierarchical spatial encoding scheme, whereby each bridge segment is globally identified using structural identifiers (e.g., B-1, B-2, ..., B-13). Within each segment, the stitched image is uniformly partitioned along the horizontal (X-axis) and vertical (Y-axis) dimensions into a grid of 500 px × 500 px subregions. Each subregion is assigned a unique two-dimensional position code (x, y), thereby constructing a reversible and traceable spatial indexing structure. This approach enables a block-structured spatial representation of the image content, providing clear geometric semantic labels for subsequent defect detection outputs. As shown in Fig. 11, the encoded subregions are arranged in a non-overlapping manner within the pixel space, forming a complete and continuous matrix suitable for spatially explicit damage annotation and localization.

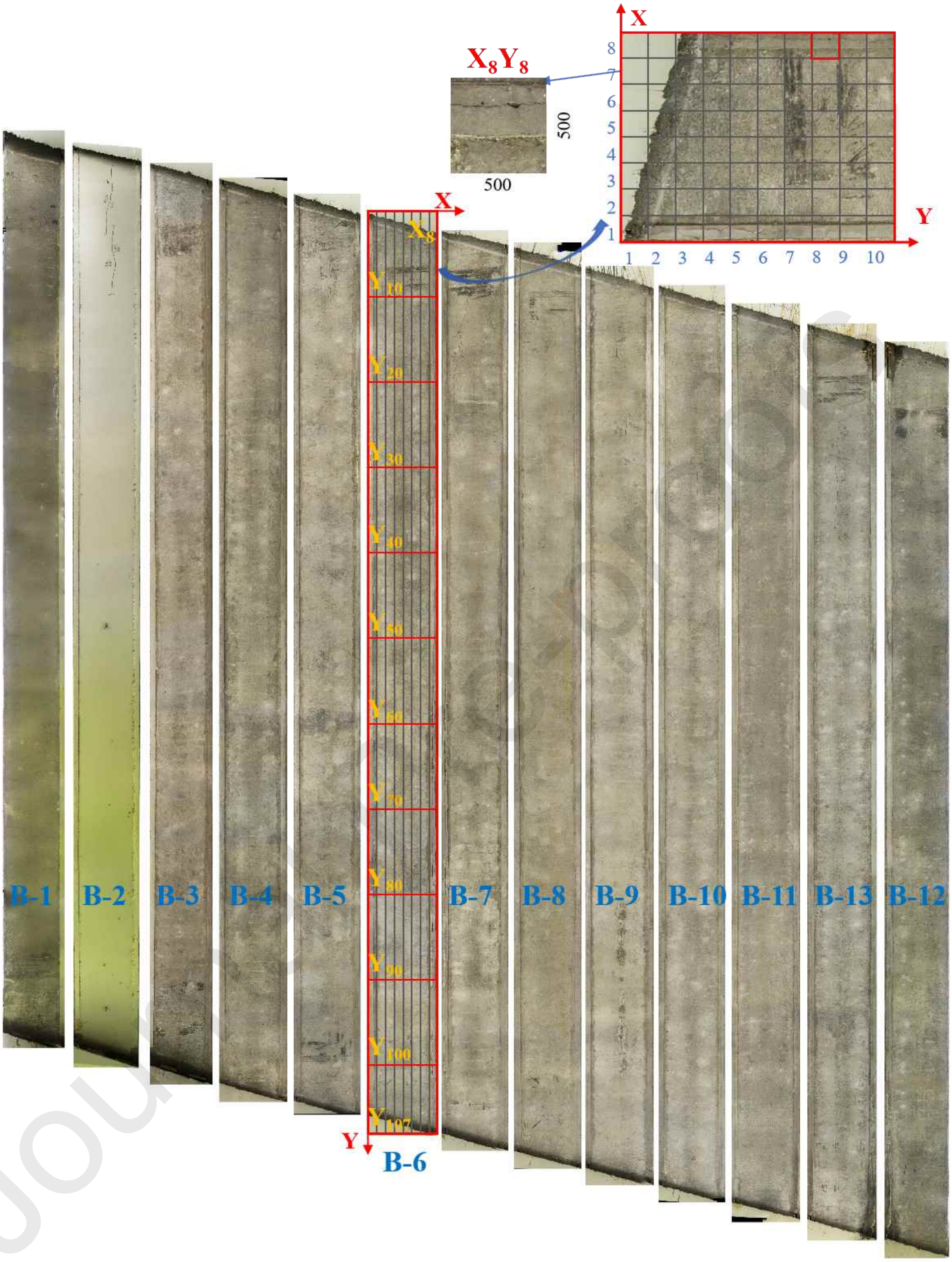


Fig. 11. Visualization of spatial encoding and localization on the underside of the bridge girder.

Regarding the grid segmentation size, to further demonstrate the rationale for using a Patch-size of 500, we conducted a comparative analysis of the average disease recognition accuracy (IoU) and detection time for each beam under different Patch-sizes. As shown in the table, with increasing grid size, the detection time decreased significantly, while the recognition accuracy initially increased and then declined. The analysis indicates that both Patch-sizes of 400 and 500 achieve the highest recognition accuracy; specifically, increasing the Patch-size from 400 to 500 reduces the detection time by approximately 35%, while the IoU decreases only slightly by about 0.79%. Considering both accuracy and efficiency, a Patch-size of 500 is optimal, as it maintains high recognition accuracy while substantially improving detection efficiency.

Following the completion of spatial encoding, each subregion image was independently input into the TransUNet-based crack detection model developed in Section 4.1 for semantic segmentation. The model outputs consisted of a binary crack mask along with a corresponding confidence map. By extracting the pixel coordinates of crack regions within the encoded subimages and combining them with their associated spatial encoding indices (x,y), the localized pixel positions were transformed into global coordinates within the stitched panoramic image.

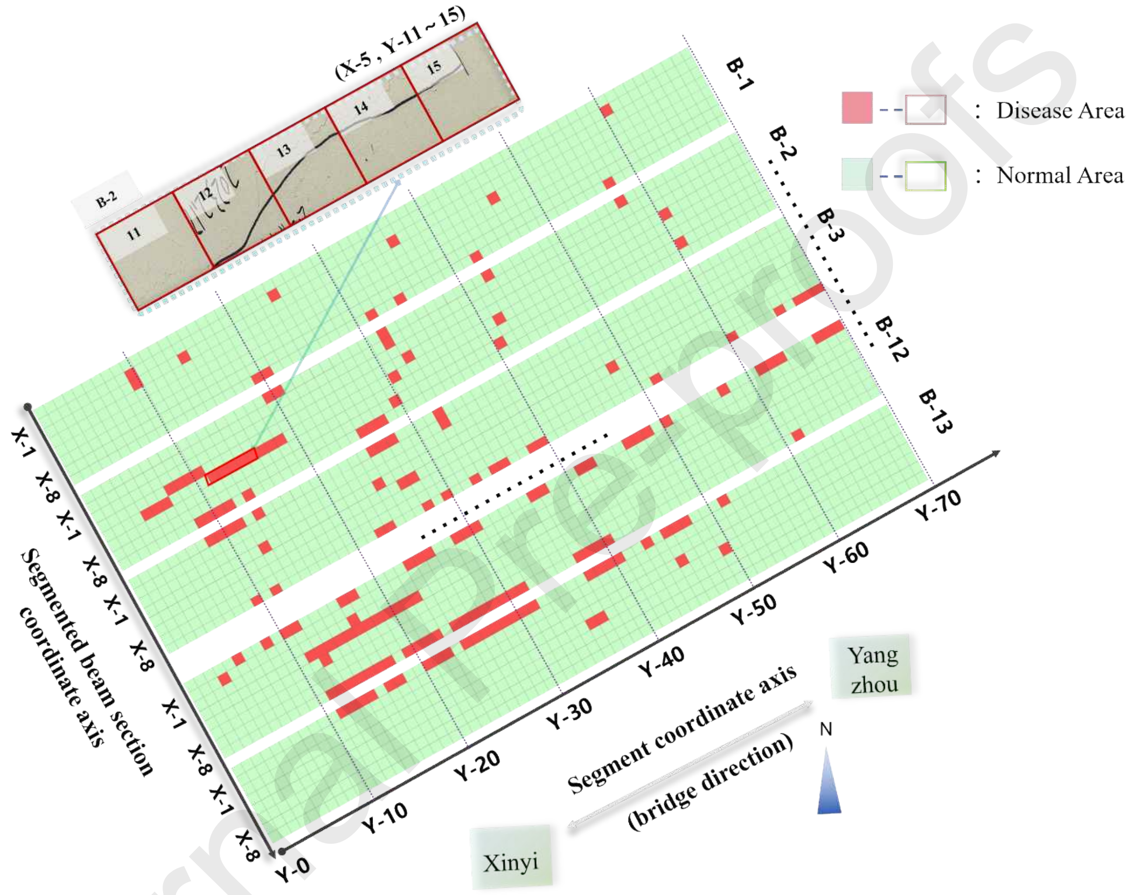


Fig.12. Spatial Distribution of Cracks Across the Entire Underside of the Bridge.

To further enhance spatial interpretability, a local coordinate system was established within each encoded unit, where the top-left corner of the subimage was defined as the origin. The cracks' local pixel coordinates (u,v) were then mapped through a hierarchical transformation: from local coordinates to encoded indices, and finally to global image coordinates. Leveraging this multilevel mapping strategy and the linear scaling coefficient between image and physical space, the crack detection results were accurately projected onto the real-world geometry of the bridge structure. This enabled precise localization and visualization of structural defects at the component level. Fig. 12 presents the panoramic crack visualization map of the bridge underside generated using the proposed matrix-based spatial encoding framework. The entire map is organized by 13 individual bridge segments, each reconstructed using high-resolution panoramic imaging.

Diseased region	*Images*	*Location/mm*	*Area*

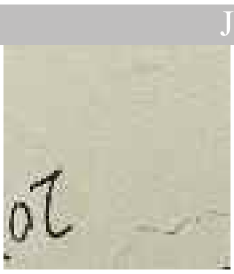
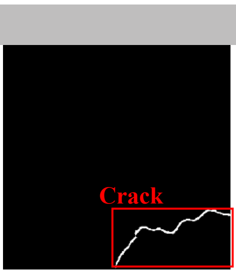
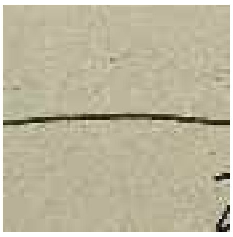
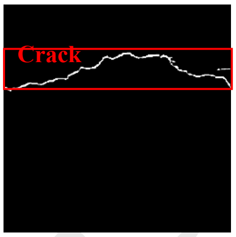
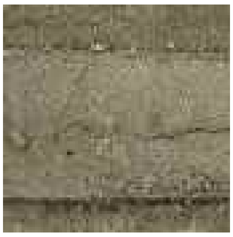
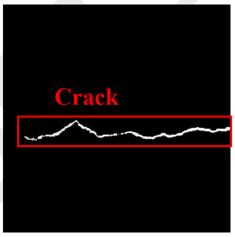
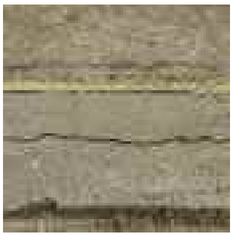
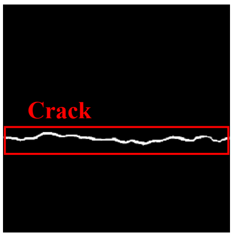
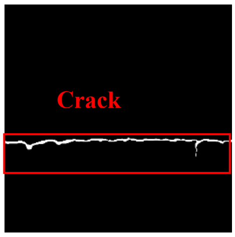
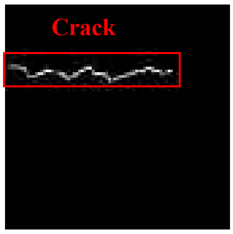
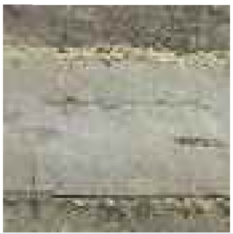
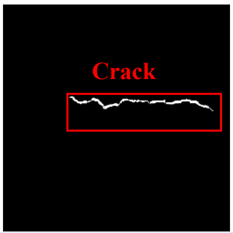
B2 – (X4,Y8)		**Crack**: (747.38, 1404.75), (717.00, 1455.00), (704.62, 1499.25).	
B2– (X5,Y13)		**Crack**: (817.50, 2250.00), (790.88, 2343.75), (817.50, 2436.38).	
B4 – (X8,Y44)		**Crack**: (1420.12, 8079.75), (1418.25, 8171.25), (1413.00, 8250.00).	
B6 – (X8,Y24)		**Crack**: (1421.62, 4312.50), (1424.25, 4406.25), (1422.00, 4500.00).	
B8 – (X8,Y73)		**Crack**: (1425.00, 13500.00), (1423.12, 13593.75), (1424.25, 13687.50).	
B12 – (X1,Y23)		**Crack**: (50, 4141), (59, 4202), (54, 4263).	
B13 – (X1,Y23)		**Crack**: (75.75, 4179.00), (79.12, 4240.50), (87.38, 4297.88).	

Fig. 13. Visualization of Crack Detection and Spatial Localization Results in Typical Defect Areas.

Both the horizontal and vertical directions were uniformly partitioned based on the predefined encoding scheme. In the visualized grid, green cells indicate areas where no damage was detected, while red cells represent encoded units identified by the TransUNet model as containing cracks or structural anomalies. The figure establishes a robust spatial correspondence between the detection results and the actual physical layout of the bridge through a two-dimensional grid mapping approach. It not only specifies the precise location of each detected crack but also allows for immediate referencing to the associated bridge segment (e.g., B-1 through B-13) and its internal image encoding coordinates (X,Y). The inset highlights an enlarged view of localized crack regions within segment B-2, particularly focusing on subregions 4 through 9. This visual representation demonstrates the model's strong capacity for accurate localization and information conveyance, providing both visual and quantitative support for regional damage distribution analysis, targeted maintenance planning, and long-term spatiotemporal structural health monitoring.

Building upon the panoramic crack spatial atlas illustrated in Figure 11, several representative damage-coded regions were further selected to demonstrate localized identification results and to validate their visualization, as shown in Figure 13. The chosen subregions were extracted from different coded blocks within beam segments B2, B3, B4, and B12, where the corresponding spatial encoding enabled direct mapping onto the panoramic stitched image. From left to right, the figure presents the original sub-block images, the spatial coordinate sets of cracks within each block, and the binary crack segmentation outputs generated by the TransUNet model. The crack localization points correspond to the geometric centers and edge coordinates extracted by the model, which, after encoding mapping and pixel-scale conversion, were precisely projected onto the structural domain to achieve millimeter-level crack tracking accuracy. This matrix-based localization strategy not only enables lossless reconstruction and visualization of localized crack detection results within the global structural atlas but also significantly streamlines subsequent processes including damage annotation, quantitative statistics, and maintenance management. Consequently, it exhibits strong structural interpretability, robust engineering applicability, and favorable system scalability.

6. Conclusion

Addressing the challenges of spatial constraints, geometric distortion, and concealment of defects during inspection of the undersides of highway prestressed hollow slab bridges, this study developed and validated an intelligent detection and localization system integrating visual perception, lightweight image processing, and high-precision damage segmentation algorithms. Supported by a synergistic hardware-software architecture, the proposed system realizes fully automated workflows from unmanned aerial vehicle (UAV) data acquisition to accurate defect localization, providing a practical technical pathway for efficient identification and structural-level mapping of underside bridge defects. The principal contributions of this research are summarized as follows:

(1) UAV image acquisition with texture-aware feature matching: A lightweight multirotor UAV equipped with high-resolution imaging and matrix-based flight path planning was employed to capture comprehensive multi-view data. A dynamic feature matching strategy was devised, selecting GIM-LightGlue for low-texture regions to ensure accuracy and stability, and XFeat-LightGlue for high-texture areas to enhance efficiency. This adaptive mechanism balances precision and real-time performance, improving robustness while reducing computational demands.

(2) High-resolution panoramic reconstruction with geometric consistency: A homography correction method leveraging Segment Anything instance segmentation and edge parallelism constraints was introduced to correct image tilt and distortion caused by UAV attitude changes. Linear structural constraints and nonlinear correction matrices controlled cumulative errors during stitching, preserving geometric and structural coherence over extended scales. Combined with energy-optimized seam blending and multi-scale Laplacian pyramid fusion, the approach produces visually consistent, smoothly blended panoramas with uniform brightness and aligned details.

(3) Matrix-based spatial encoding and deep crack segmentation for accurate localization: A structured matrix encoding scheme partitions stitched images into unique spatial units, enabling bidirectional mapping between image and physical structural domains. Using the TransUNet segmentation model, cracks were identified within encoded images and precisely projected onto corresponding physical locations via reverse mapping. This facilitates engineering-relevant, traceable defect localization, supporting informed

maintenance and repair decisions.

At present, this study is dedicated exclusively to the detection and localization of cracks—particularly micro-cracks in hollow slab girder bridges—without the inclusion of other critical damage types such as concrete spalling, steel corrosion, or water leakage. While the proposed geometric stitching framework has proven effective for small-span girder segments, its application to large-span bridges still faces notable challenges, including cumulative alignment errors, geometric degradation over long stitching sequences, and substantial computational demands associated with high-resolution panoramic imagery. These factors limit real-time processing capabilities and affect the precision of subsequent defect analyses. Future research will address these limitations through several key directions: (1) Multi-defects and multi-scale detection. Extending the framework beyond cracks to encompass other major defects such as spalling, corrosion, and leakage, employing multi-task learning architectures, multi-scale feature representation, and cascaded detection modules to capture both large-scale damage and fine-grained cracks with high fidelity. (2) Scalable and lightweight algorithms. Optimizing stitching and perspective correction for large-span bridges to effectively control cumulative error growth while reducing computational complexity. (3) Advanced fusion and error compensation. Leveraging multi-view fusion, adaptive weighting, and error compensation strategies to enhance geometric accuracy under variable imaging conditions. (4) High-performance heterogeneous processing. Designing GPU–CPU–edge collaborative pipelines to allow accelerated inference and near real-time deployment in field inspections. (5) Broadened applicability and environmental robustness. Extending the framework to diverse bridge types (steel girders, composite structures, heavily painted or coated surfaces) and rigorously assessing its robustness under challenging environmental conditions such as rainfall, direct sunlight, heavy shadows, and surface efflorescence.

References

[1] H. Liu, X. Wang, G. Tan, X. He, G. Luo, System reliability evaluation of prefabricated RC hollow slab bridges considering hinge joint damage based on modified AHP, Applied Sciences-Basel. 9 (2019) 4841, https://doi.org/10.3390/app9224841.

[2] T.-Y. Wang, D. Li, J. Zhang, Research on lightweight health-monitoring strategies and technological development for bridge structures, China Civil Engineering Journal. (2024) 1 – 18, https://doi.org/10.15951/j.tmgcxb.23100849.

[3] R. Xie, J. Yao, K. Liu, X. Lu, Y. Liu, M. Xia, Q. Zeng, Automatic multi-image stitching for concrete bridge inspection by combining point and line features, Automation in Construction. 90 (2018) 265 – 280, https://doi.org/10.1016/j.autcon.2018.02.021.

[4] W. Chen, B. Yuan, D. Chen, Y. Hu, F. Wang, J. Zhang, Synchronized identification and localization of defect on the bottom of steel box girders based on a dynamic visual perception system, Computers in Industry. 169 (2025) 104291, https://doi.org/10.1016/j.compind.2025.104291.

[5] S. Jiang, Y. Cheng, J. Zhang, Vision-guided unmanned aerial system for rapid multiple-type damage detection and localization, Structural Health Monitoring—An International Journal. 22 (2023) 319 – 337, https://doi.org/10.1177/14759217221084878.

[6] F. Wang, Y. Zou, C. Zhang, J. Buzzatto, M. Liarokapis, E. del Rey Castillo, J. B. P. Lim, UAV navigation in large-scale GPS-denied bridge environments using fiducial marker-corrected stereo visual-inertial localisation, Automation in Construction. 156 (2023) 105139, https://doi.org/10.1016/j.autcon.2023.105139.

[7] N. Bolourian, A. Hammad, LiDAR-equipped UAV path planning considering potential locations of defects for bridge inspection, Automation in Construction. 117 (2020) 103250, https://doi.org/10.1016/j.autcon.2020.103250.

[8] S. Jiang, J. Zhang, Real-time crack assessment using deep neural networks with wall-climbing unmanned aerial system, Computer-Aided Civil and Infrastructure Engineering. 35 (2020) 549 – 564, https://doi.org/10.1111/mice.12519.

[9] M. R. Jahanshahi, S. F. Masri, G. S. Sukhatme, Multi-image stitching and scene reconstruction for evaluating defect evolution in structures, Structural Health Monitoring—An International Journal. 10 (2011) 643 – 657,

https://doi.org/10.1177/1475921710395809.

[10] F. Hu, J. Zhao, Y. Huang, H. Li, Structure-aware 3D reconstruction for cable-stayed bridges: A learning-based method, Computer-Aided Civil and Infrastructure Engineering. 36 (2021) 89 – 108, https://doi.org/10.1111/mice.12568.

[11] Y.-F. Liu, X. Nie, J.-S. Fan, X.-G. Liu, Image-based crack assessment of bridge piers using unmanned aerial vehicles and three-dimensional scene reconstruction, Computer-Aided Civil and Infrastructure Engineering. 35 (2020) 511 – 529, https://doi.org/10.1111/mice.12501.

[12] C.-Q. Feng, B.-L. Li, Y.-F. Liu, F. Zhang, Y. Yue, J.-S. Fan, Crack assessment using multi-sensor fusion simultaneous localization and mapping (SLAM) and image super-resolution for bridge inspection, Automation in Construction. 155 (2023) 105047, https://doi.org/10.1016/j.autcon.2023.105047.

[13] D. Wang, Y. Zhang, Y. Pan, B. Peng, H. Liu, R. Ma, An automated inspection method for the steel box girder bottom of long-span bridges based on deep learning, IEEE Access. 8 (2020) 94010 – 94023, https://doi.org/10.1109/ACCESS.2020.2994275.

[14] S. Hou, H. Shen, G. Wu, T. Wu, W. Sun, H. Jiang, X. Fan, G. Liu, Comparative analysis of image stitching algorithms for bridge inspection imaging systems, KSCE Journal of Civil Engineering. 29 (2025) 100071, https://doi.org/10.1016/j.kscej.2024.100071.

[15] Y.-J. Cha, W. Choi, O. Buyukozturk, Deep learning-based crack damage detection using convolutional neural networks, Computer-Aided Civil and Infrastructure Engineering. 32 (2017) 361 – 378, https://doi.org/10.1111/mice.12263.

[16] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, O. Buyukozturk, Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types, Computer-Aided Civil and Infrastructure Engineering. 33 (2018) 731 – 747, https://doi.org/10.1111/mice.12334.

[17] J. Y. Yu, F. Li, X. Xue, P. Zhu, X. Wu, P. S. Lu, Intelligent identification of bridge structural cracks based on unmanned aerial vehicle and Mask R-CNN, China Journal of Highway and Transport. 34 (2021) 80 – 90, https://doi.org/10.19721/j.cnki.1001-7372.2021.12.007.

[18] C. Wan, X. Xiong, B. Wen, S. Gao, D. Fang, C. Yang, S. Xue, Crack detection for concrete bridges with image-based deep learning, Science Progress. 105 (2022) 4128487, https://doi.org/10.1177/00368504221128487.

[19] Y. N. Peng, M. Liu, Z. Wan, W. B. Jiang, W. X. He, Y. N. Wang, A dual deep network based on the improved YOLO for fast bridge surface defect detection, Acta Automatica Sinica. 48 (2022) 1018 – 1032, https://doi.org/10.16383/j.aas.c210807.

[20] A. Zhang, K. C. P. Wang, Y. Fei, Y. Liu, S. Tao, C. Chen, J. Q. Li, B. Li, Deep learning-based fully automated pavement crack detection on 3D asphalt surfaces with an improved CrackNet, Journal of Computing in Civil Engineering. 32 (2018) 04018041, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000775.

[21] F. Ni, J. Zhang, Z. Chen, Zernike-moment measurement of thin-crack width in images enabled by dual-scale deep learning, Computer-Aided Civil and Infrastructure Engineering. 34 (2019) 367 – 384, https://doi.org/10.1111/mice.12421.

[22] D. H. Kang, Y.-J. Cha, Efficient attention-based deep encoder and decoder for automatic crack segmentation, Structural Health Monitoring. 21 (2022) 2190 – 2205, https://doi.org/10.1177/14759217211053776.

[23] S. Zhu, J. Du, Y. Li, X. Wang, Method for bridge crack detection based on the U-Net convolutional networks, Journal of Xidian University. 46 (2019) 35 – 42, https://doi.org/10.19665/j.issn1001-2400.2019.04.006.

[24] D. Liang, W. Zhang, Y. Yu, Crack identification method of concrete bridge based on MU-Net, Journal of Beijing Jiaotong University. 46 (2022) 105 – 112, https://doi.org/10.11860/j.issn.1673-0291.20210081.

[25] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., Segment Anything, Proc. IEEE/CVF Int. Conf. Computer Vision. (2023) 4015 – 4026, https://doi.org/10.48550/arXiv.2304.02643.

[26] E. N. Mortensen, H. Deng, L. Shapiro, A SIFT descriptor with global context, Proc. IEEE Conf. Computer Vision and Pattern Recognition. 1 (2005) 184 – 190, https://doi.org/10.1109/CVPR.2005.45.

[27] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-Up Robust Features (SURF), Computer Vision and Image Understanding. 110 (2008) 346 – 359, https://doi.org/10.1016/j.cviu.2007.09.014.

[28] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, Proc. IEEE Int. Conf.

Computer Vision. (2011) 2564 – 2571, https://doi.org/10.1109/ICCV.2011.6126544.

[29] D. DeTone, T. Malisiewicz, A. Rabinovich, SuperPoint: Self-supervised interest point detection and description, Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops. (2018) 224 – 236, https://doi.org/10.1109/CVPRW.2018.00060.

[30] M. Tyszkiewicz, P. Fua, E. Trulls, DISK: Learning local features with policy gradient, Advances in Neural Information Processing Systems. 33 (2020) 14254 – 14265, https://doi.org/10.48550/arXiv.2006.13566.

[31] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Chen, Z. Li, ALiKE: Accurate and lightweight keypoint detection and descriptor extraction, IEEE Transactions on Multimedia. 25 (2022) 3101 – 3112, https://doi.org/10.1109/TMM.2022.3155927.

[32] X. Shen, Z. Cai, W. Yin, M. Müller, Z. Li, K. Wang, et al., GIM: Learning generalizable image matcher from internet videos, arXiv. (2024) https://arxiv.org/abs/2402.11095.

[33] G. Potje, F. Cadar, A. Araujo, R. Martins, E. R. Nascimento, XFeat: Accelerated features for lightweight image matching, Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. (2024) 2682 – 2691, https://doi.org/10.1109/CVPR52733.2024.00259.

[34] M. Muja, D. G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration, Proc. Int. Conf. Computer Vision Theory and Applications (VISAPP). 2 (2009) 331 – 340, https://doi.org/10.5220/0001787803310340.

[35] P. E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, SuperGlue: Learning feature matching with graph neural networks, Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. (2020) 4938 – 4947, https://doi.org/10.1109/CVPR42600.2020.00499.

[36] P. Lindenberger, P. E. Sarlin, M. Pollefeys, LightGlue: Local feature matching at light speed, Proc. IEEE/CVF Int. Conf. Computer Vision. (2023) 17581 – 17592, https://doi.org/10.1109/ICCV51070.2023.01616.

[37] H. Li, J. Wang, C. Han, Image mosaic and hybrid fusion algorithm based on pyramid decomposition, Proc. Int. Conf. Virtual Reality and Visualization (ICVRV). (2020) 205 – 208, https://doi.org/10.1109/ICVRV51359.2020.00049.

[38] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., TransUNet: Transformers make strong encoders for medical image segmentation, arXiv. (2021) https://arxiv.org/abs/2102.04306.

[39] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, Lect. Notes Comput. Sci. (MICCAI). 9351 (2015) 234 – 241, https://doi.org/10.1007/978-3-319-24574-4_28.

[40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, Lect. Notes Comput. Sci. (DLMIA/ML-CDS). 11045 (2018) 3 – 11, https://doi.org/10.1007/978-3-030-00889-5_1.

[41] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-UNet: U-Net-like pure transformer for medical image segmentation, Lect. Notes Comput. Sci. (ECCV Workshops). 13669 (2022) 205 – 218, https://doi.org/10.1007/978-3-031-25066-8_9.

[42] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv. (2017) https://arxiv.org/abs/1706.05587.

Highlights

- A two-stage inspection framework for defect identification in hollow slab bridges.

- High-resolution UAV image system with dual-feature matching strategy.

- Geometric rectification and matrix encoding for image stitching.

- TransUNet-based synchronized defect detection and localization.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Acknowledgments

The research presented was financially supported by the National Key R&D Program of China (No. 2022YFC3801700), the National Natural Science Foundation of China (No. 52378289), Research Fund for Advanced Ocean Institute of Southeast University (No. KP202407).