Structures 80 (2025) 110086

Contents lists available at ScienceDirect

Structures

journal homepage: www.elsevier.com/locate/structures

# Lightweight panoramic reconstruction and precise defect localization of bridge undersides based on multi-view pose registration

Jiangfan Zhao [a], Rui Zhao [a,*], Wang Chen [b], WenBin Xu [a], Jian Zhang [a,b,c,**]

[a] College of Civil Engineering and Architecture, Xinjiang University, Urumqi, China
[b] School of Civil Engineering, Southeast University, Nanjing, China
[c] Advanced Ocean Institute of Southeast University, Nantong 226000, China

ARTICLE INFO

*Keywords:*
Bridge engineering
Unmanned aerial vehicle
Multi-view image stitching
Three-dimensional reconstruction
Deep learning
Defect localization

ABSTRACT

Prestressed hollow-slab bridges commonly exhibit underside cracking. Conventional manual inspection is time-consuming and lacks positional accuracy, while the underside's inability to receive GPS signals further hampers precise localization. This study proposes a lightweight framework for panoramic underside reconstruction and intelligent defect localization based on multi-view unmanned-aerial-vehicle (UAV) imagery. The workflow comprises three principal stages: (1) high-resolution underside images are acquired by a UAV camera, and Structure-from-Motion (SFM) rapidly generates a sparse point cloud and camera poses; combined with plane fitting and multithread processing, a 1.5 m × 20 m girder panorama is reconstructed within 6 min at a spatial resolution of 0.25 mm $px^{-1}$ ; (2) A white-balance lookup table (WB-LUT) and a grid-based Laplacian fusion algorithm eliminate illumination and color-temperature discrepancies in 2 min, and a matrix-coding system achieves high-precision mapping between defects and image coordinates; (3) Using the improved Light TransUNet for crack segmentation increased the IoU by 3.35 percentage points to 80.97 %, reduced the parameter count by 86.4 %, and accelerated inference by 35.7 % relative to TransUNet. Finally, the spatial mapping provided by matrix coding enables sub-millimetre-level defect localization, markedly enhancing inspection efficiency and accuracy. Field tests on an in-service bridge confirm clear advantages in detection efficiency, segmentation accuracy, and localization precision, demonstrating the engineering applicability and effectiveness of the proposed method.

## 1. Introduction

The underside of highway bridge girders constitutes the core load-bearing region of the structure; it is not only the primary locus of stress concentration but also a concealed space that poses significant challenges for manual inspection. Prolonged exposure to dynamic loads and environmental degradation commonly initiates micro-cracks in this zone, which, under cyclic stress, gradually propagate into macro-cracks and ultimately lead to severe structural deteriorations such as concrete spalling and reinforcement corrosion. Conventional inspection practice still relies primarily on aerial work platforms combined with visual observation and contact measurements, and faces three major bottlenecks: (i) low efficiency and high risk associated with working at height; (ii) detection accuracy that is strongly dependent on the inspector's experience; and (iii) a high incidence of missed and false detections [1]. Although crack detectors [2], ultrasonic testing instruments [3], and impact-echo devices [4] have been increasingly adopted, heterogeneous operator skill levels, insufficient standardisation of inspection procedures, and the continued requirement for manual target calibration in semi-automated workflows render mis-detections and omissions persistent issues. Recent technological advances exhibit two salient trends: (i) deep-learning-based intelligent recognition algorithms that employ neural networks to achieve automated crack identification and pixel-level segmentation, thereby markedly improving detection speed and accuracy over traditional manual methods [5]; and (ii) the deployment of unmanned aerial vehicles (UAVs) equipped with multi-spectral sensors, which greatly enhance the efficiency of underside inspection [6]. Nevertheless, in GPS-denied environments such as girder soffits, achieving efficient and precise localisation and recognition remains a central challenge in current bridge inspection research.

* Corresponding author.
** Corresponding author at: College of Civil Engineering and Architecture, Xinjiang University, Urumqi, China.
*E-mail addresses:* zhaorui@xju.edu.cn (R. Zhao), jian@seu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.istruc.2025.110086
Received 6 July 2025; Received in revised form 20 August 2025; Accepted 26 August 2025
Available online 31 August 2025

Because Global Positioning System (GPS) signals beneath bridge girders are easily obstructed, current research on soffit defect localization has concentrated on three technical approaches. The first involves developing and deploying hardware tailored specifically for the underside environment. Xie et al. [7] designed a mobile inspection platform equipped with a robotic arm and multiple sensors, capable of stable positioning and comprehensive inspection of the bridge underside; however, its portability is poor, its environmental adaptability limited, and inspection blind spots remain. Wang et al. [8] developed an image-acquisition system comprising an industrial camera, lenses, and a linear guide rail, enabling stable imaging of the girder soffit from a fixed viewpoint. Chen et al. [9] proposed an integrated system combining a distributed camera chain with intelligent algorithms for long-span bridge inspection, where cameras travel along guide rails to capture panoramic images and localize defects. Although dedicated-equipment solutions provide reliable defect identification and localization, their high cost and the requirement to customise equipment for each bridge hinder routine inspection of the large population of small bridges. The second approach employs unmanned aerial vehicles (UAVs). UAVs offer a highly efficient, low-cost means of soffit inspection, yet the absence of GPS signals impedes accurate localization. To overcome this limitation, Lin et al. [10] utilised ultra-wideband (UWB) technology to achieve precise positioning in GPS-denied environments and implemented seamless switching between GPS- and UWB-based methods. Cui [1] experimentally demonstrated that UWB positioning accuracy improves markedly with an increasing number of base stations. Building on this, Jiang et al. [11] developed a vision-guided UAV system that fuses stereo vision with an inertial measurement unit (IMU) to supply more accurate positional information. Although UWB–IMU fusion has been validated experimentally, drift errors in practical settings and the complexity of multi-sensor data fusion remain significant challenges for real-world application.

The third approach employs three-dimensional (3D) reconstruction and panoramic image stitching, mapping the resulting model or imagery to the bridge's actual dimensions to localize defects. Among camera-based 3D reconstruction methods, Structure from Motion (SFM) is the most widely used. Liu et al. [12] successfully applied SFM to obtain 3D models of bridges and their components, whereas Xu et al. [13] decomposed the reconstruction into multiple distributed tasks, enhancing computational efficiency without sacrificing model accuracy. Nevertheless, SFM requires large image overlap and precise flight-path planning, resulting in long processing times and low efficiency. Feng et al. [14] fused multi-sensor simultaneous localization and mapping (SLAM) with image super-resolution for pier inspection, proposing a rapid and accurate crack-assessment method that substantially reduced 3D-modeling complexity. However, their method was validated only in a local scene; when extended to large-scale, high-precision reconstruction, SLAM still suffers from cumulative drift, failures in low-texture regions, and pose jumps. In practice, even for small and medium-sized bridges, generating a stable sub-millimetre model typically requires several hours and exhibits a high failure rate, limiting widespread adoption in large-scale bridge inspection. By contrast, for the planar geometry of girder soffits, panoramic image stitching offers a lighter and more robust localization solution. Recent deep-learning-based feature-detection and matching algorithms—such as SuperPoint [15], ALIKE [16], XFeat [17], and LightGlue [18]—have further improved stitching accuracy and processing speed. Nonetheless, large-scale stitching continues to face challenges of image distortion and cumulative error because two-dimensional feature matching under varying camera poses cannot completely eliminate viewpoint differences. To address this issue, Chen et al. [9] employed a distributed camera chain with rail-mounted cameras to achieve parallel-viewpoint panoramic stitching with minimal parallax, enabling defect identification and localization; Xie et al. [7] developed a mobile platform equipped with a robotic arm and multiple sensors to obtain stable imaging perspectives. Although these methods are effective, their equipment costs and maintenance requirements are high. Unmanned aerial vehicles (UAVs) are highly portable and can collect images efficiently, making them well suited to panoramic stitching for small and medium-span bridges. However, maintaining a viewpoint parallel to the girder soffit during flight is difficult, resulting in substantial parallax among images. Consequently, achieving stable panoramic stitching of multi-view UAV imagery constitutes the primary focus of this study.

With the widespread adoption of deep learning in image recognition and segmentation, deep learning–based methods for bridge defect detection are gradually replacing traditional approaches. Unlike manual visual inspection, deep learning can directly extract multi-level features from images, enabling automatic and precise identification of defect regions, thus providing more reliable data for bridge condition assessment. Early crack detection research relied mainly on convolutional neural networks (CNNs). Zhang et al. [19] first applied CNNs for crack recognition, outperforming traditional methods. Cha et al. [20] combined CNNs with a sliding window approach for large images and used Faster R-CNN [21] for automatic multi-type defect detection. Yang et al. [22] compared AlexNet, VGGNet, and ResNet, showing ResNet's superior feature extraction in crack classification. These studies established the foundation for deep learning in crack detection. As real-time bridge monitoring became more important, single-stage detectors like YOLO [23] and SSD [24] gained popularity for their speed and accuracy. Peng et al. [25] proposed dual networks, "YOLO-lump" and "YOLO-crack," improving real-time detection of multiple surface defects. Zhang et al. [26] optimized network structure and applied lightweight design to reduce complexity without losing accuracy. Luo et al. [27] developed SFW-YOLO with high-resolution feature enhancement, dynamic detection heads, and improved EIOU loss, achieving a mean average precision (mAP) of 90.0 % through network pruning. For fine-grained crack characterization, pixel-level segmentation methods have become a focus. CrackNet removes pooling layers for pixel-wise prediction [28]; dual-scale CNNs detect wide and narrow cracks and measure width with sub-pixel methods [29]; improved U-Net models fuse multi-layer features to retain detail and width information [30]. Hybrid Transformer–U-Net architectures, such as TransUNet, SwinUNet, and MTUNet, have been developed, with TransUNet demonstrating superior convergence speed and higher detection accuracy [31]. Despite high accuracy, challenges remain in achieving lightweight, quantized, and real-time models. Research continues on using lightweight modules, quantization, knowledge distillation, and automated network search to reduce parameters and computation while maintaining detection performance, enabling deployment on embedded or edge devices.

To address the foregoing engineering challenges, this study proposes an integrated framework for refined identification and high-precision spatial localisation of soffit defects that combines multi-view unmanned-aerial-vehicle (UAV) image acquisition, matrix-coding mapping, and an enhanced lightweight Light TransUNet crack-detection model. The core workflow is as follows. First, a UAV collects high-resolution images along a matrix flight path beneath the girder. An incremental Structure-from-Motion (SFM) algorithm rapidly recovers a sparse point cloud and camera poses; a girder-soffit plane is then fitted by singular-value decomposition (SVD) to project all views into a common plane. Second, white-balance lookup tables (WB-LUTs) are applied for colour correction, and a grid-based Laplacian-pyramid fusion method eliminates stitching seams to yield a seamless panorama. Third, an image matrix-coding system embeds positional information into image tiles, which are fed to a lightweight, improved Light TransUNet deep-learning model for fine crack segmentation. Finally, matrix-coding and dimension mapping enable precise spatial localisation of defects. The method requires neither high-precision positioning sensors nor additional fixed equipment, thus offering low cost, strong versatility, and flexible deployment. It is particularly suited to frequent inspection of small- and medium-span bridges and holds considerable promise for engineering application.

This paper is organised into seven sections: Section 2 presents the

overall technical framework; Section 3 details the incremental SFM–based camera-pose estimation and projection stitching, the unified white-balance processing, the grid-based Laplacian fusion algorithm, and the matrix-coding scheme; Section 4 introduces the improved Light TransUNet model for fine crack segmentation; Section 5 validates the proposed method through field experiments on an in-service bridge; and Section 6 discusses directions for future research.

## 2. Proposed methodology

This study addresses the technical challenges of multi-view image stitching and precise crack localisation on bridge girder soffits. Large viewpoint differences prevent traditional stitching from obtaining accurate initial views, causing serious distortion and error accumulation. In addition, the enclosed underside environment precludes stable Global Positioning System (GPS) reception for image georeferencing, while traditional three-dimensional (3D) reconstruction, although accurate, is cumbersome, time-consuming, and unsuitable for high-frequency inspection of short-span bridges. To overcome these obstacles, an intelligent inspection system that integrates multi-view image stitching with an improved deep-learning crack-detection algorithm is proposed; the overall architecture is illustrated in Fig. 1. A UAV equipped with a high-resolution camera first acquires comprehensive, gap-free imagery along a matrix flight trajectory beneath the girder. An incremental Structure-from-Motion (SFM) algorithm then rapidly generates a sparse point cloud and recovers camera poses. The underside reference plane is fitted from the point cloud, and the homography between each image and this plane is computed from the camera poses, enabling unified projection and stitching of all views. The method reconstructs a seamless panorama of a 1.5 m × 20 m girder within six minutes at a spatial resolution of 0.25 mm $px^{-1}$ , markedly improving stitching efficiency and reducing perspective distortion and cumulative error relative to conventional approaches, thus meeting the stringent image-quality and localisation requirements of bridge-defect detection. To mitigate illumination, colour-temperature, and white-balance disparities inherent in multi-view acquisition, a deep-learning-based white-balance lookup-table (WB-LUT) algorithm is employed for rapid global colour correction. Residual differences are removed by grid-based projection and Laplacian-pyramid image fusion, eliminating stitching seams in an additional two minutes and producing a visually continuous, colour-consistent panoramic image. Because the resulting panorama is too large for direct efficient analysis, a matrix-coding system partitions the image into spatial regions, assigning each a unique positional identifier. These coded tiles are input to an improved Light TransUNet model for high-precision crack segmentation. The model replaces the original ResNet backbone with the lightweight StraNet, reduces the number of Transformer layers, and thereby lowers the parameter count by 86.4 % relative to the baseline. During U-Net feature fusion, a MixNet module is introduced to enhance the integration of global and local features, yielding a 3 % increase in mean Intersection-over-Union (mIoU). After crack detection, precise mapping between the panorama and the girder's actual dimensions enables fine spatial localisation of defects, satisfying the practical needs of bridge inspection and maintenance. The proposed method is particularly suited to routine inspection of small- and medium-span bridges, providing efficient crack identification and accurate localisation, and thus offers substantial engineering value for advancing bridge health monitoring and preventive maintenance.

## 3. Pose-based multi-view image stitching

To meet the demand for rapid and accurate localisation beneath bridge girders, a lightweight image-stitching method based on camera poses is proposed. First, underside images are acquired along a matrix flight path, and a deep-learning model is applied to equalise white-balance differences caused by illumination and shooting conditions, thereby mitigating the influence of lighting inconsistencies on stitching. Robust feature detection and matching, combined with an incremental Structure-from-Motion (SFM) algorithm [32], are then employed to align unordered images rapidly. The fundamental and essential matrices are computed, and camera poses are recovered through triangulation. Using the pose information and matched points, a sparse point cloud of the girder soffit is reconstructed; the soffit plane is fitted from this point cloud (this method is suitable for planar soffits and is not applicable to structures with significant curvature or complex geometry), and a homography is derived to project each image accurately onto the plane. White-balance lookup tables (WB-LUTs) [33] are subsequently applied, followed by grid-based projection and Laplacian-pyramid image fusion to suppress ghosting and correct residual colour discrepancies, ultimately yielding a seamless, high-resolution panorama of the girder soffit without overlap. The proposed approach not only delivers fast and accurate image stitching but also simplifies localisation by mapping to physical dimensions, thereby significantly reducing computational load and enhancing overall processing efficiency.

### 3.1. SFM pose estimation and projection stitching

Because images captured by a UAV beneath a girder cannot be guaranteed to align parallel to the soffit, conventional panorama stitching typically requires manual calibration of the first image to a parallel viewpoint or relies on epipolar constraints to estimate its homography. Owing to large manual calibration errors and the limited constraints between only two images, this approach seldom yields a truly parallel rectified view. Even when the first image is accurately calibrated, subsequent feature-based stitching still accumulates matching errors, leading to severe deformation and drift (see the lower-right inset of Fig. 2). Accordingly, an incremental Structure-from-Motion (SFM) stitching scheme is proposed. Reliable poses are obtained by reconstructing an initial image pair, after which ordered image addition iteratively expands and refines the model. All camera parameters and three-dimensional points are jointly optimised under global bundle adjustment, resulting in high-precision, low-distortion stitching of the entire girder segment. The method was implemented on an Intel Core i5–12490F CPU (3.00 GHz) with 16 GB of RAM; its workflow is illustrated in Fig. 2.

First, the Scale-Invariant Feature Transform (SIFT) algorithm [34] is applied to each matrix-acquired image to detect keypoints and extract descriptors. A KD-tree–based nearest-neighbour search, combined with Lowe's ratio test, rapidly establishes preliminary matches within the unordered image set. The number of matches is then counted for every possible image pair, and the pair $I_a$ and $I_b$ that exhibits the greatest number of correspondences and an appropriate baseline is selected as the initial pair. For this pair, the Random Sample Consensus (RANSAC) algorithm together with the classical eight-point method is employed to estimate the fundamental matrix $F$. Specifically, in each RANSAC iteration, eight matched point pairs are randomly sampled, $F$ is solved, and the number of inliers is determined according to the epipolar-geometry constraint:

$$\left(\mathbf{x}^{b}\right)^{T} F \mathbf{x}^{a} \approx 0 \tag{1}$$

The correctness of each correspondence is evaluated. After all iterations, the fundamental matrix containing the largest number of inliers is retained as the final estimate $F$ and the associated outliers are discarded to yield a reliable set of matched inliers.

After the fundamental matrix has been obtained, the camera focal length $f$ is extracted from the image EXIF metadata, and the principal point is assumed to lie at the image centre $\left(c_x, c_y\right)$; the camera intrinsic matrix $K$ is thus directly constructed as :

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

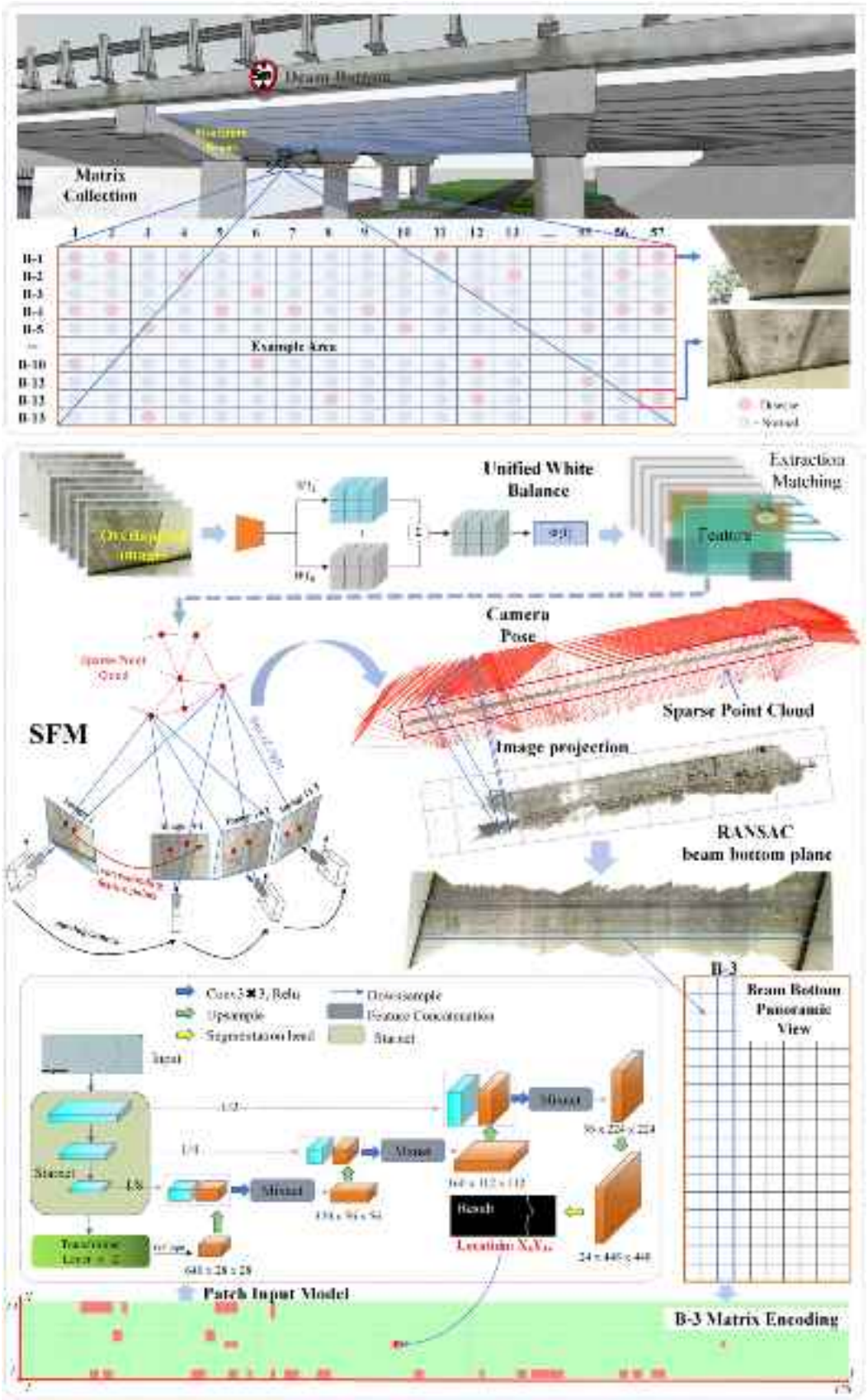


**Fig. 1.** Schematic diagram of the multi-view girder-soffit image-stitching and defect-localisation framework.

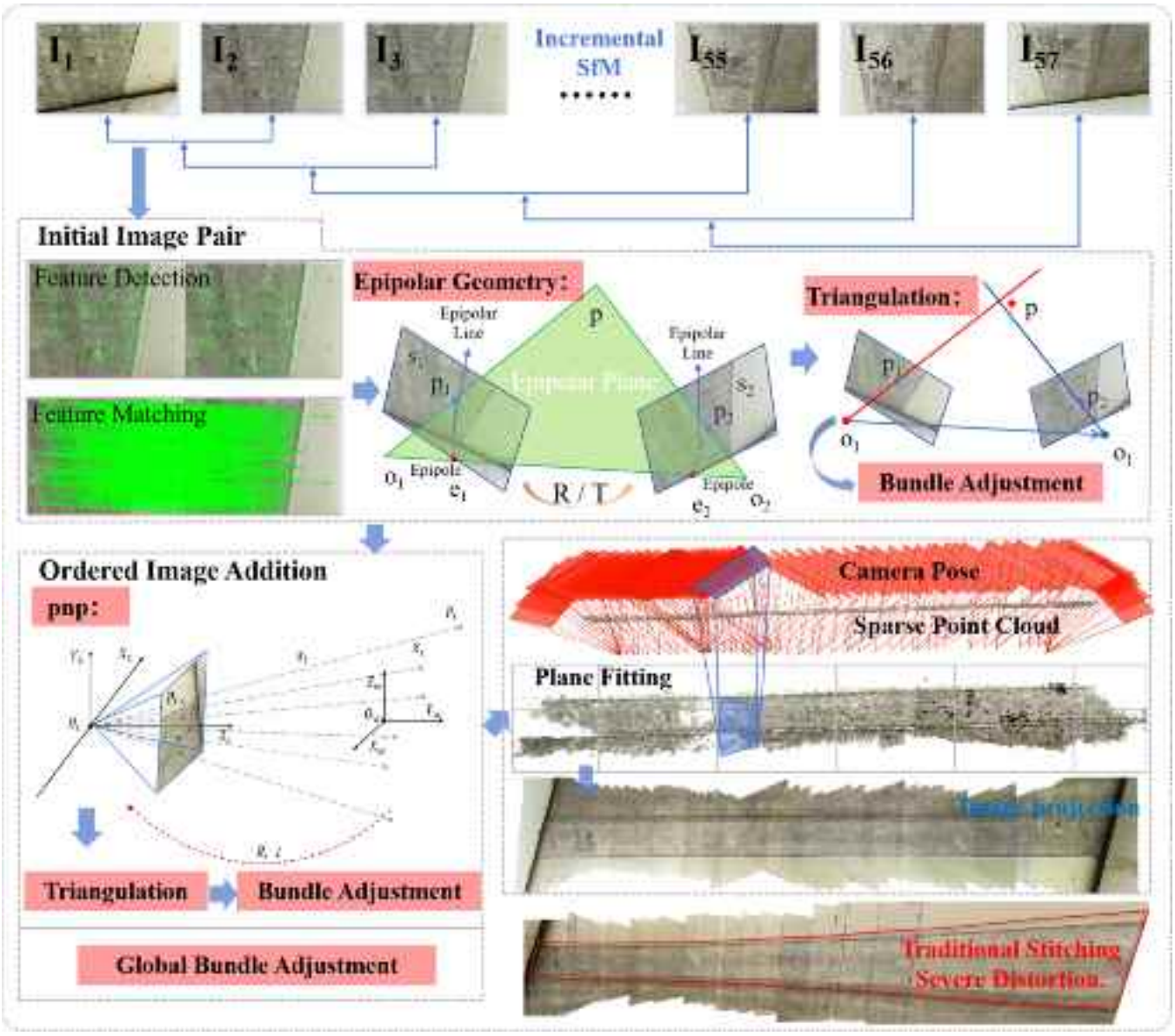


**Fig. 2.** Schematic of SFM pose estimation and projection.

Using the camera intrinsic matrix $K$ obtained above, the fundamental matrix $F$ can be further converted into the essential matrix $E$:

$$E = K^{\top} F K \tag{3}$$

The essential matrix $E$ describes the geometric relationship between two camera coordinate systems; because it contains only the cameras' relative rotation and translation, with the intrinsic parameters eliminated, it can be used directly to recover their relative pose.

To recover the cameras' relative rotation matrix $R$ and translation direction vector $t$, the essential matrix is subjected to singular-value decomposition (SVD) in accordance with Eq. (4):

$$E = U \Sigma V^{\mathrm{T}} \tag{4}$$

Two orthogonal matrices, $U$ and $V$, are thus obtained; $U$ and $V$ define the rotational subspace structure of the essential matrix in the first and second camera coordinate systems, respectively.

Further, with the aid of the template matrices $W$ and Z defined below:

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, Z = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{5}$$

The decomposed matrices $U$ and $V$ allow four candidate solutions for the relative rotation and translation direction to be constructed directly, specifically:

$$R_1 = UWV^{\top}, R_2 = UW^{\top}V^{\top}, t = U[:,3](\pm). \tag{6}$$

According to the epipolar-geometry constraint, each candidate solution can be used to perform an initial triangulation to recover three-dimensional points. The physically valid relative rotation matrix and translation direction are then identified by the cheirality check, that is, by requiring the depth (the Z -coordinate) of each reconstructed point to be positive in both camera coordinate systems. In this way, a unique, physically plausible relative camera pose $(R_{ab}, t_{ab})$ is obtained.

Finally, using the determined camera poses, the projection matrices of the two cameras are constructed as :

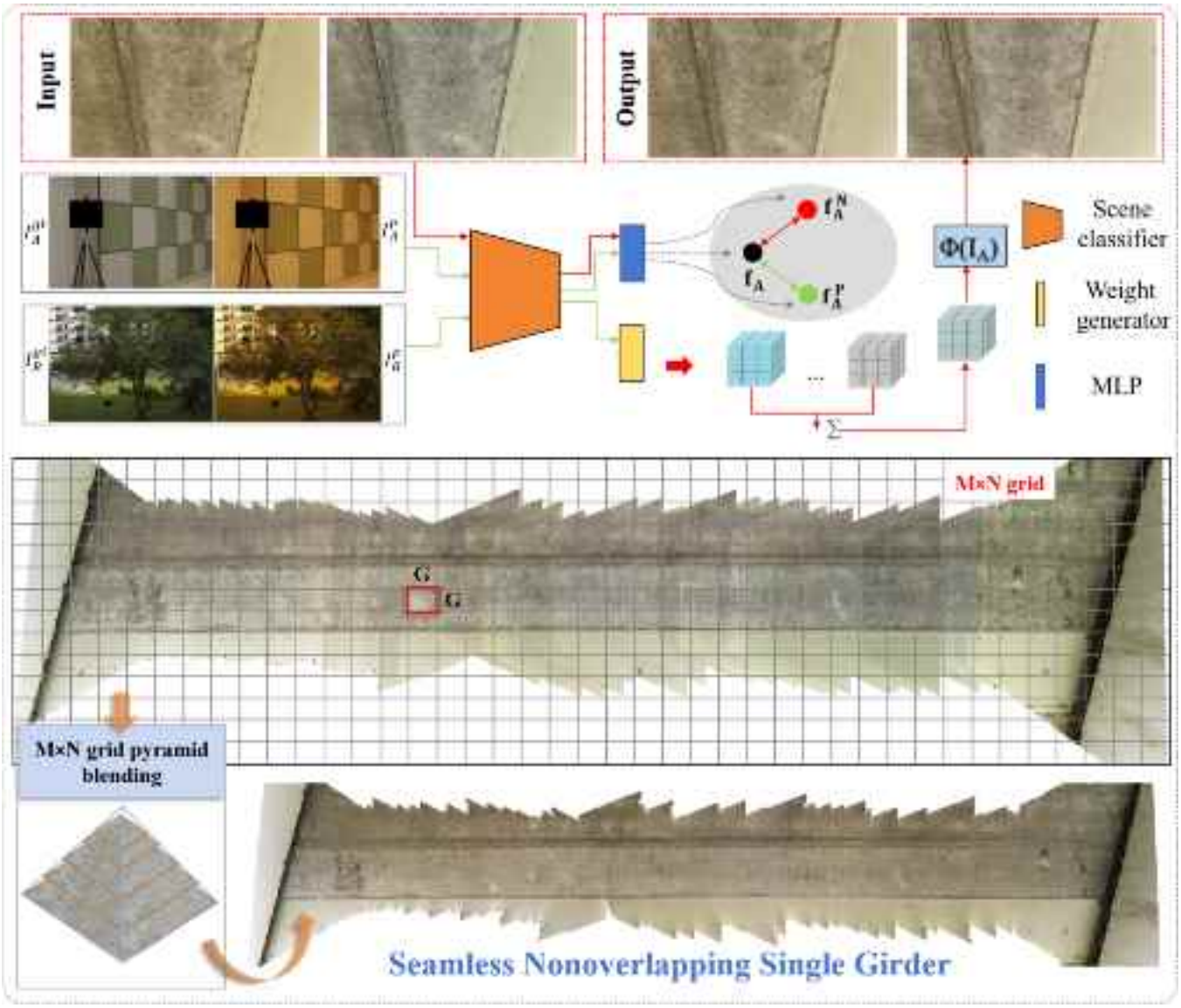


**Fig. 3.** Image white-balance harmonisation and Laplacian fusion method.

$$P_a = [I|0], P_b = [R_{ab}|t_{ab}]. \tag{7}$$

Linear triangulation is then performed for each inlier correspondence to estimate the three-dimensional coordinates:

$$\mathbf{X} = \underset{X}{\operatorname{argmin}} \|x^a - \pi(P_a X)\|^2 + \left\|x^b - \pi(P_b X)\right\|^2 \tag{8}$$

This yields the initial sparse three-dimensional point cloud. The foregoing procedure ensures a reliable reconstruction of the initial image pair and provides high-quality initial estimates of both camera poses and the sparse point cloud for the subsequent incremental SFM process.

After the relative pose of the initial image pair and its accompanying sparse point cloud have been recovered, this "seed" model is incrementally expanded to reconstruct the sparse point cloud and camera poses for the entire girder segment. For each candidate image $I_K$ to be added, the number of feature correspondences with the existing 3D points is tallied and denoted as the covisibility count $N_K$. In general, a larger $N_K$ implies stronger constraints between the image and the current reconstruction, resulting in a more stable and reliable pose estimate. Accordingly, the remaining images are sorted in descending order of $N_K$, and incrementally registered starting with the image that exhibits the highest covisibility, thereby ensuring that the reconstruction proceeds in a stable and efficient manner.

After the image $I_K$ with the highest covisibility has been selected, two-dimensional–three-dimensional correspondences are established between its feature points and the existing sparse point cloud. The camera pose of $I_K$ is then solved using the perspective-nnn-point (PnP) algorithm. Specifically, within a RANSAC framework, the following optimisation problem is iteratively solved to obtain a robust pose estimate:

$$\{\mathbf{R}_k, \mathbf{t}_k\} = \underset{R,t}{\operatorname{argmin}} \sum_{i=1}^{N_k} \left\|\pi(K[R|t]\mathbf{X}_i) - \mathbf{x}_i^k\right\|^2 \tag{9}$$

where $X_i$ denotes an existing 3D point and $\mathbf{x}_i^k$ is its corresponding two-dimensional observation in image $I_K$. During this step, RANSAC automatically discards outliers, thereby ensuring the robustness of the pose estimation.

After the pose of image $I_K$ has been successfully estimated, some feature points in $I_K$ may not yet be observed in the existing sparse point cloud, although they form stable correspondences with features in previously reconstructed images. These new correspondences are therefore subjected to triangulation to enlarge the point cloud and extend scene

coverage. Specifically, one or more images that exhibit the most reliable matches with $I_K$—denoted as reference images $I_{ref}$—are selected, and the matched points are linearly triangulated:

$$\mathbf{X}_{\text{new}} = \text{Triangulate}\left(P_{\text{ref}}, P_k, \mathbf{x}^{\text{ref}}, \mathbf{x}^k\right) \tag{10}$$

thereby generating additional three-dimensional points and further enhancing the completeness of the reconstructed model.

The incremental addition of images inevitably introduces cumulative error. To mitigate this local drift, immediately after each new image $I_K$ is registered and additional 3D points are created, a local bundle-adjustment is performed on a small subgraph comprising $I_K$ and the previously registered images with the highest covisibility. Specifically, the following non-linear optimisation problem is solved:

$$\min_{\substack{R_i,t_i \\ \mathbf{X}_j}} \sum_{i\in\mathscr{V}_k} \sum_{j\in\mathscr{J}_i} ||\pi\left(K[R_i|t_i]\mathbf{X}_j\right) - \mathbf{x}_j^i||^2 \tag{11}$$

where $\mathscr{V}_k$ denotes the set containing the newly added image $I_K$ and the previously registered images with the greatest covisibility, and $\mathscr{J}_i$ is the inlier set of image $I_K$. The optimisation is solved with the Levenberg–Marquardt algorithm, which markedly reduces local error accumulation during incremental reconstruction.

Although local optimisation mitigates part of the error, global drift inevitably accumulates as the number of images increases. Accordingly, once a prescribed number of images has been incorporated, a comprehensive global bundle adjustment is performed over all recovered camera poses and three-dimensional points. The optimisation objective is :

$$\min_{\substack{R_i,t_i \\ \mathbf{X}_j}} \sum_i \sum_j ||\pi\left(K[R_i|t_i]\mathbf{X}_j\right) - \mathbf{x}_j^i||^2 \tag{12}$$

By jointly refining every camera pose and point, the cumulative error of the overall model is effectively reduced, ensuring global consistency and accuracy in the final reconstructed scene.

By means of the foregoing incremental addition strategy, images are incorporated into the reconstruction framework progressively and efficiently, ensuring local accuracy while maintaining high overall precision and consistency in the final sparse point cloud. The resulting high-quality pose and point-cloud data provide a robust foundation for panoramic stitching of the girder.

After the incremental SFM reconstruction, camera poses for all images and a sparse 3D point cloud are obtained. Because the girder soffit can be approximated by a single plane, the sparse cloud must be fitted to an accurate geometric plane to enable unified image projection and stitching. First, 3D points that lie within the soffit region are selected from the reconstructed cloud. These points are then subjected to plane fitting by singular-value decomposition (SVD) to determine the optimal plane normal $n$ and intercept $d$. The procedure begins by computing the geometric centroid $\overline{X}$ of the point set and forming the matrix $Y$ of centred coordinates:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad , \quad Y = \begin{bmatrix} (X_1 - \overline{X})^\top \\ \vdots \\ (X_N - \overline{X})^\top \end{bmatrix}. \tag{13}$$

Denote by $X_i$ the coordinate vector of the $i$-th 3D point. Matrix YYY is then factorized by singular value decomposition (SVD) as follows:

$$Y = U\Sigma V^\top, \Sigma = \text{diag}(\sigma_1, \sigma_2, \sigma_3), \sigma_1 \geq \sigma_2 \geq \sigma_3. \tag{14}$$

The right singular vector $v_3$ associated with the smallest singular value $\sigma_3$ gives the plane normal $n$. After normalisation, the plane equation is

$$d = -n^\top \overline{X}, n^\text{T} X + d = 0. \tag{15}$$

After plane fitting, the original 3D point cloud is rotated about its normal vector so that the plane aligns exactly with the axes of the world coordinate system, thereby establishing a unified, standard projection frame for the girder soffit. Let the target normal be $\text{t} = (0,0,1)^\text{T}$. The rotation matrix $R$ is constructed via the Rodrigues formula:

$$R = I + [\omega]_x + \frac{[\omega]_x^2}{1 + n^\text{T}\text{t}}, \omega = n \times t \tag{16}$$

where $[\omega]_x$ is the skew-symmetric matrix of $\omega$.

For each image $i$, a set of two-dimensional–three-dimensional correspondences $\{(x_k^{(i)}, X_k^{(i)})\}$ is established, where $x_k^{(i)} \in \mathbb{R}^2$ denotes the pixel coordinates undistorted by the intrinsic matrix and $X_k^{(i)} \in \mathbb{R}^2$ represents the planar coordinates obtained after rotation onto the aligned plane. The homography $H_i$ is estimated with RANSAC:

$$x_k^{(i)} \sim H_i X_k^{(i)}, H_i = \underset{H}{\text{argmin}} \sum_k \rho\left(\left\| x_k^{(i)} - H X_k^{(i)} \right\|\right) \tag{17}$$

where $\rho$ is a robust loss on the reprojection error.

To improve efficiency and reduce the computational load of image warping, a localised projection strategy is adopted. After defining the output canvas size and a global translation matrix, region-of-interest (ROI) perspective projections are computed in parallel, with each image's local warp executed on separate threads. This approach accelerates the stitching process by more than an order of magnitude and reconstructs a 1.5 m × 20 m girder panorama at 0.25 mm px$^{-1}$ resolution within six minutes:

$$I_i'(u) = I_i\left(H_i'^{-1}u\right), u \in \text{ROI}_i \tag{18}$$

### 3.2. Deep-learning white-balance unification with grid-Laplacian seamless fusion

In Section 3.1 a panoramic mosaic of a single girder was generated using camera poses and a sparse point cloud. Owing to pronounced differences in illumination and colour temperature among images, however, hard seams appear in the stitched panorama, degrading visual quality and reducing the accuracy of subsequent crack-detection models. To address this issue, the captured soffit image set is pre-processed with the deep-learning-based white-balance normalisation algorithm WB-LUTS [33]. This lightweight network predicts mixture weights for several learnable three-dimensional look-up tables (LUTs) for each image and linearly combines them to achieve rapid pixel-level white-balance correction. Contrastive learning and hard-sample mining are incorporated to maintain robustness under varying lighting conditions, and millisecond-level processing is achieved even at high resolution.Although global white balance is thereby unified, subtle chromatic differences persist between images. Computing a dedicated seam and performing Laplacian blending for every image individually would incur prohibitive computational cost—exceeding that of the stitching process itself. To smooth residual discrepancies without sacrificing speed, the output canvas is partitioned into M×N uniform grid cells, each of size G×G pixels, replacing image-wise, seam-wise blending with an efficient grid-based scheme. For each input image $I_k$, its four corner points are projected onto the canvas with the homography $H_k$, and the centroid $c_k$ of the projected quadrilateral is computed. Each grid cell $(i,j)$ with centre $\mathbf{g}_{ij}$ is then assigned to the image whose centroid is closest in Euclidean distance:

$$\text{owner}(i,j) = \underset{k}{\text{argmin}} \parallel \mathbf{g}_{ij} - \mathbf{c}_k \|_2 \tag{19}$$

To enhance computational efficiency, the contiguous grid cells assigned to each image are aggregated into minimal rectangular regions of interest (ROIs). These ROIs are processed in parallel through multi-threading, rapidly completing grid-based projection across the entire canvas and enabling efficient panorama generation. To eliminate hard seams between adjacent grids, the pixels and masks within each ROI are subjected to L-level Gaussian smoothing and down-sampling, yielding a

multiresolution Gaussian pyramid $\{G^l\}_{l=0}^{L}$. A Laplacian pyramid $L^l$ is then constructed by backward differencing to extract the high-frequency details at every scale:

$$\begin{cases} L^l = G^l - \text{expand}(G^{l+1}), l = 0, \ldots, L-1 \\ L^L = G^L \end{cases} \tag{20}$$

Next, the mask pyramid $M^l$ at the corresponding scale is employed to weight-blend the high-frequency details of adjacent images on each layer of the Laplacian pyramid:

$$R^l(y,x) = M^l(y,x)L_1^l(y,x) + \left[1 - M^l(y,x)\right]L_2^l(y,x) \tag{21}$$

where $L_1$ and $L_2$ denote the high-frequency components of the two neighbouring grid regions at layer lll, and $M^l$ is a smoothed weighting map (range [0,1]) obtained by down-sampling the original binary mask to layer l. Starting from the coarsest level, fusion and up-sampling reconstruction are performed iteratively toward higher resolutions:

$$R^{l-1} = R^l + \text{expand}(R^{l+1}) \tag{22}$$

Upon completion of these steps, the final output $\widehat{R}^0$ constitutes the fully fused panorama with seamless transitions.

### 3.3. Matrix partition coding

By performing precise pose estimation and image projection for each girder segment, an independent panoramic image of every soffit was first constructed. These panoramas were then rescaled to a uniform resolution of 6000 × 80,000 px via a homography grounded in the girder's true dimensions (1.5 m × 20 m), yielding a spatial precision of 0.25 mm $px^{-1}$ . After overlap removal, the 13 regional panoramas were merged into a single underside view of the entire bridge (Fig. 4). Because the bridge is skewed, the global mosaic exhibits considerable inclination; naïvely applying matrix coding would therefore assign indices to extensive blank areas, resulting in redundant computation. To prevent such waste, invalid regions were culled before coding so that only meaningful image areas entered the crack-detection pipeline, thereby improving inference efficiency and resource utilisation. The refined panorama was divided into matrix cells of 488 × 488 px. Each cell was assigned a unique two-dimensional spatial index (X,Y) that corresponds one-to-one with its physical location on the bridge. Crack detection and instance segmentation were then executed within every cell using the optimised Light TransUNet model, and the results were precisely mapped back to physical coordinates via the indexing scheme. This coordinate-encoded segmentation strategy avoids duplicate analysis and redundant computation, secures sub-millimetre crack-localisation

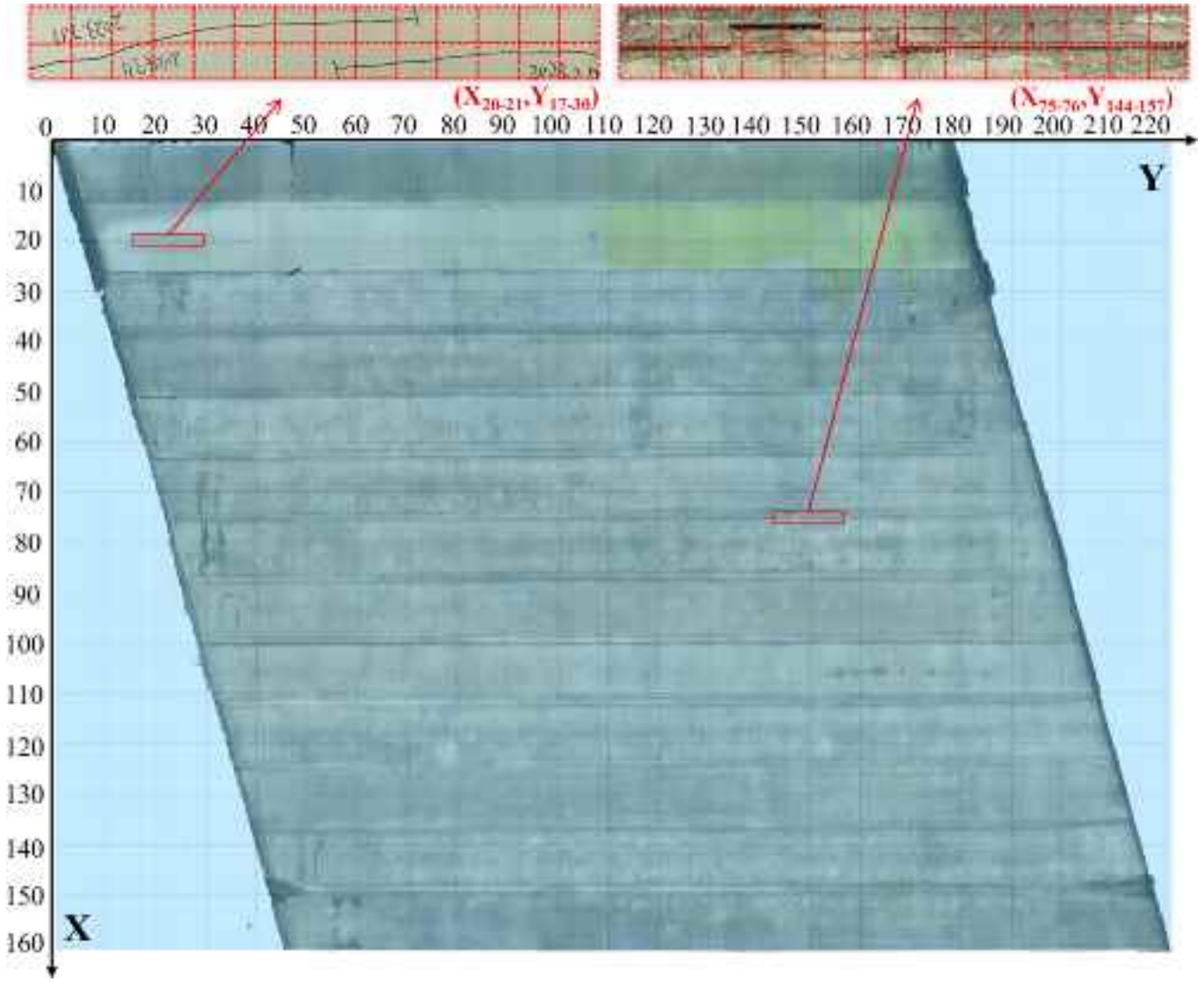


**Fig. 4.** Schematic of matrix coding for the girder-soffit panorama.

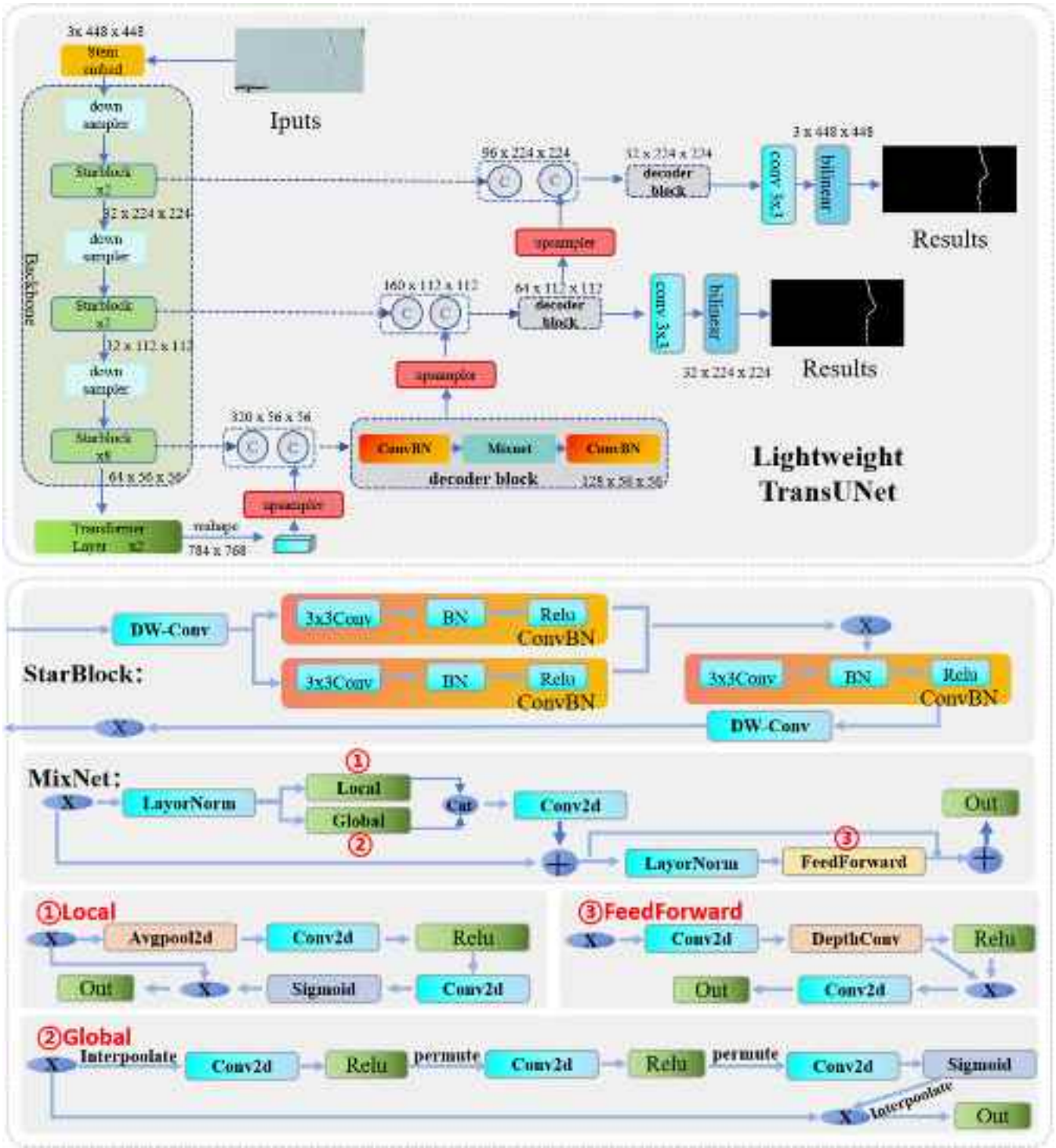


**Fig. 5.** Architecture of the Light TransUNet network.

accuracy, and supplies high-quality data to support digital modelling, intelligent management, and fine-grained assessment of bridge defects.

## 4. Defect analysis

After the matrix-coded soffit sub-images are generated, a lightweight, enhanced TransUNet model is developed for efficient crack instance segmentation. To overcome the large parameter count and slow inference of conventional high-accuracy networks, Stranet [35] is adopted as a compact backbone, reducing channel width and the number of Transformer layers and compressing the model size by ~86.4 % to ≈ 12.671 M parameters. This optimisation not only preserves performance but also accelerates inference, rendering the model suitable for resource-constrained devices. Furthermore, a MixNet module [36] is incorporated into the U-Net feature-fusion stage to integrate local detail with global semantic context, markedly strengthening feature representation. On high-resolution imagery the optimised model doubles inference speed and raises crack-detection accuracy from 77.616 % to 80.970 %, effectively addressing the efficiency limitations of traditional architectures.

### 4.1. Lightweight TransUNet network

In medical image segmentation, the TransUNet model—an archetypal hybrid architecture [37]—successfully integrates the long-range dependency modelling of a Transformer encoder with the multi-scale feature fusion of the U-Net framework via skip connections. However, its use of a ResNet backbone combined with a 12-layer Transformer yields 93.231 M parameters, imposing substantial computational demands that hinder practical deployment. To address this challenge, a three-fold optimisation strategy is proposed to markedly enhance computational efficiency and engineering applicability. First, the heavyweight ResNet backbone is replaced by the lightweight StraNet jointly developed by Northeastern University and Microsoft [35]. StraNet employs an innovative star operation for element-wise, high-dimensional non-linear mapping, coupled with a deeply compressed three-stage downsampling structure, achieving exceptional parameter efficiency. Second, the Transformer module is streamlined: the number of layers is reduced from 12 to 2, and both the multi-head self-attention (MSA) head count and the multilayer perceptron (MLP) dimensionality are lowered, significantly reducing model size. Third, a MixNet module [36] is introduced during the U-Net decoder's feature-fusion phase to compensate for any representational loss due to compression. To integrate MixNet into the training pipeline, its activation functions are changed from GELU to ReLU, selected $1 \times 1$ convolutions are replaced by same-channel $3 \times 3$ convolutions, and convolutional widths are linearly scaled to match the backbone's channel multipliers, without altering MixNet's core mechanisms. Following these optimisations, the Light TransUNet model achieves a 3 % increase in segmentation accuracy, reduces its parameter count from 93.231 M to 12.671 M, and doubles inference speed—effectively overcoming the drawbacks of large model size and slow inference, and greatly enhancing its practical value.

The optimized lightweight TransUNet backbone accepts a (3,448,448) input tensor and constructs a multi-scale feature pyramid via three stages of progressive downsampling. A hierarchical reinforcement design is employed: the first two stages each stack two StraNet modules, while the third stage deepens the stack to eight modules, ultimately producing a high-level semantic feature map of size (64,56,56). Within each StraNet module, the input features are first processed by a $7 \times 7$ depthwise separable convolution (DW-Conv) to expand the receptive field while minimising computational cost. The resulting feature map is then passed into two parallel branches of $1 \times 1$ convolution–batch-normalisation–ReLU (Conv1 $\times$1–BN–ReLU). One branch's output serves as gating weights—after ReLU activation—that are element-wise multiplied with the other branch's output, thereby enhancing salient features. The gated features undergo a further $1 \times 1$ convolution followed by another $7 \times 7$ DW-Conv to produce the module output. A residual SUM operation fuses this output with the original input, blending shallow and deep semantic information and strengthening representational capacity. To further boost expressiveness, a multi-head attention mechanism processes the StraNet outputs: two Transformer layers with eight attention heads each replace the original 12-layer, 12-head configuration, while the MLP dimensionality is reduced from 3072 to 1024. Any loss in feature-extraction power is compensated during fusion by a MixNet module, which comprises three key components: a global feature modulation layer, a local feature modulation layer, and a feed-forward layer. The global layer captures long-range dependencies across spatial dimensions at low complexity, restoring global context; the local layer, with variable receptive fields, enhances fine-detail capture while suppressing noise. Element-wise fusion of global and local outputs implements a "global-guides-local" enhancement pathway, improving both extraction and fusion efficiency. Finally, the feed-forward layer employs skip connections to preserve original features—mitigating gradient vanishing—and accelerates convergence by projecting multi-scale features into a compact representational space, balancing computational efficiency with information integrity.

### 4.2. Dataset

A systematic workflow was adopted for model development and training in soffit crack detection. First, diseased images were meticulously annotated using Labelme to create a bespoke dataset of 2 993 images, each at $1024 \times 512$ px resolution. The dataset was split 9:1 into a training set (2 694 images) and a validation set (299 images) to ensure robust generalisation. During training, stochastic gradient descent (SGD) with momentum and weight decay was employed to achieve smooth and stable gradient updates in complex feature scenarios. Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM) using the PyTorch 2.0.0 framework under Windows 11, with Python 3.9.7 and CUDA 11.8 as the core dependencies, thereby establishing a fully GPU-accelerated training environment.

### 4.3. Network training and testing

The lightweight TransUNet was systematically trained on the preprocessed dataset. A progressive learning-rate schedule was employed, with an initial rate of $5 \times 10^{-4}$, and the batch size was fixed at 12 to balance GPU memory usage with stable gradient updates. Training was conducted for 120 epochs, taking a total of 20.4 h and achieving an optimal trade-off between accuracy and efficiency. Performance was evaluated in terms of segmentation accuracy and computational efficiency. Segmentation accuracy was quantified by the Intersection-over-Union (IoU), which measures the overlap between predicted masks and ground-truth labels. Computational efficiency was assessed by forward FLOPs (G), parameter count (Params, M), and per-sample inference latency (s). Collectively, these metrics form a multidimensional evaluation matrix that characterises the model's adaptability and generalisation across datasets and guides the accuracy–efficiency balance, confirming the effectiveness and practicality of the proposed lightweight TransUNet.

To evaluate the proposed feature-representation framework comprehensively, a benchmark was established encompassing five mainstream crack-segmentation networks—U-Net [38], SwinUNet [39], U2-Net [40], ID-UNet [41], TransUNet—and the present Light TransUNet. Under a unified dataset and training regimen, Table 1 reports results in five dimensions: Intersection-over-Union (IoU), floating-point operations (FLOPs), parameter count, per-image inference time, and model size. Light TransUNet achieved an IoU of 80.97 %, exceeding TransUNet and ID-UNet by 3.35 and 2.70 percentage points, respectively. With only 12.67 M parameters—a reduction of 86.4 % relative to TransUNet and only modestly higher than ID-UNet's 1.14 M—Light TransUNet also offers the lightest computational load at 12.67 G FLOPs. Its inference latency is just 0.027 s per image, the fastest among all models and four times quicker than the next-most-accurate ID-UNet. The model footprint is restrained to 50.8 MB—an 87.4 % compression

**Table 1**
Performance comparison of different models.

| | *Crack Segmentation* | | | | |
|---|---|---|---|---|---|
| Network | IOU(%) | FLOPs (G) | Params (M) | Inference Time(s) | Model size (M) |
| Swinunet | 65.672 | 23.761 | 27.191 | 0.030 | 238.00 |
| U-net | 74.256 | 167.647 | 31.083 | 0.032 | 124.20 |
| U2-net | 75.129 | 115.487 | 44.024 | 0.038 | 168.30 |
| TransUnet | 77.616 | 98.908 | 93.231 | 0.042 | 403.50 |
| ID-unet | 78.266 | 67.745 | 1.143↑ | 0.109 | 4.70↑ |
| **Light TransUnet (ours)** | **80.970↑** | **12.671↑** | **12.671** | **0.027↑** | **50.80** |

Notes: (1) #FLOPs (G): floating-point operations required for one forward pass, expressed in billions; Param: number of learnable parameters, measured in millions (M) where 1 M = 1000,000 parameters; Inference Time: time (in seconds) the model needs to process a single input sample—the lower the value, the better the real-time performance.

compared with TransUNet and considerably smaller than U-Net and SwinUNet. Balancing accuracy, speed, and resource utilisation, Light TransUNet delivers the best trade-off among high precision, real-time performance, and low resource demand, providing an efficient, practical solution for real-time crack detection on UAV or embedded platforms.

Fig. 6 presents visualised segmentation results from each model on representative crack scenarios, allowing direct comparison of detection accuracy. Light TransUNet consistently reproduces crack contours with high fidelity to the ground truth across various crack morphologies. Although U-Net, ID-Unet, U2-net, and TransUNet perform comparably on well-defined cracks, they exhibit some shortcomings in edge detail and crack continuity. For instance, in Scenarios 3 and 4 of Fig. 6—characterised by low-contrast, faint-texture cracks—Light TransUNet still fully identifies the crack, whereas the other models either miss segments of the crack or produce false positives in crack-free areas. These results further confirm the robustness and generalisation advantage of Light TransUNet in fine-grained crack detection.

To evaluate the generalization capability of our model, we conducted experiments on the DeepCrack dataset[42], which contains 537 crack images and is split into 300 training images and 237 validation images for comprehensive assessment. As shown in Table 2, Light TransUnet outperforms other models across multiple metrics, achieving the highest IOU (75.621 %) and Dice score (86.180 %), with a recall of 88.290 %. Notably, despite having only 12.671 M parameters—significantly fewer than TransUnet's 93.231 M—Light TransUnet attains superior segmentation accuracy and more complete crack detection. These results indicate that Light TransUnet achieves an effective balance between model efficiency and performance on the DeepCrack dataset, highlighting its suitability for practical bridge crack detection. Overall, the experiments demonstrate that the proposed model exhibits strong generalization on public datasets, providing reliable support for real-world engineering applications.

### 4.4. Ablation experiment

Table 3 presents the ablation results of Light TransUnet on our dataset. The baseline TransUnet (Exp. 1) achieved an IOU of 77.62 %, with 98.91 G FLOPs and 93.23 M parameters. Different modules had distinct effects on performance and efficiency: Mixnet significantly improved IOU (81.58 %); Layer Reduction markedly reduced FLOPs (43.15 G) and parameters (22.35 M) with a controllable drop in accuracy (IOU 75.49 %); StarNet provided limited computational savings while also leading to a controllable accuracy decrease (IOU 74.39 %). Notably, Mixnet further reduced parameters by ~3.93 % due to structural alignment in feature fusion. Integrating all three modules (Exp. 6, Ours) achieved an IOU of 80.97 %, FLOPs of 23.24 G, parameters of 12.67 M, and inference time of 0.027 s, representing the optimal trade-off between accuracy and efficiency. These results demonstrate that the proposed lightweight model maintains high accuracy while substantially enhancing the real-time performance and deployability of bridge defect detection.

## 5. Field Experiments

At the intersection of the S49 Xinyang Expressway and the X101 Xuma Line, a field validation of the proposed UAV-based intelligent crack-detection system was conducted on a highway hollow-slab bridge. A DJI Mini 4 Pro (takeoff weight ≤249 g; 1/1.3″ CMOS sensor; 48 MP; max wind resistance 10.7 m/s) was flown along a matrix-pattern grid beneath the bridge soffit at a height of approximately 1 m, capturing 1100 ultra-high-resolution images (8064 × 4536 px) that covered all visible undersides of the 13 hollow slabs. For each slab, an incremental Structure-from-Motion (SFM) pipeline recovered camera poses and a sparse point cloud, followed by plane fitting and homography-based orthorectification to generate a local panorama at 6 000 × 80 000 px resolution. Comparison against manually surveyed ground-control points yielded a mean projection error of 0.16 mm (maximum 0.20 mm), confirming sub-millimetre stitching accuracy. The 13 local panoramas were then assembled—according to their true dimensions and relative positions—into a complete underside view of the bridge. This composite was partitioned into 448 × 448 px grid cells to establish a matrix-coded spatial indexing scheme. Using the optimised Light TransUNet model, crack segmentation was performed within each indexed cell at up to 37 fps (448 × 448 px input). Leveraging the known mapping between image coordinates and physical geometry, detected cracks were reprojected onto the actual bridge structure, achieving sub-millimetre localisation precision. The entire workflow—from image acquisition through panorama stitching—was executed on two workstations (Intel Core i5–12490F CPUs at 3.00 GHz, 16 GB RAM, NVIDIA RTX 3090), requiring only 35 min for data collection and stitching, with crack detection and report generation completed in under 15 min. This represents a multiple-fold improvement over traditional 3D reconstruction and analysis methods, which often require several hours, and demonstrates the method's high accuracy, real-time capability, and practical scalability in complex engineering environments.

Using the camera-pose–guided soffit panorama stitching method introduced in Chapter 3, panoramic images were generated for each of the 13 girder segments. Redundant pixels outside each segment's central region were cropped, and perspective rectification via homography mapping was applied to produce final output images of 6000 × 80000 px. Given the actual segment dimensions (1.5 m transverse × 20 m longitudinal), this resolution corresponds to a physical precision of 0.25 mm pixel$^{-1}$ , satisfying the sub-millimetre requirements for crack detection. To validate this accuracy, high-precision ground control measurements were performed on-site, and the pixel coordinates of the surveyed points were extracted from the stitched panoramas and compared against their true positions. The resulting localisation errors for crack feature points were all below 0.2 mm, confirming that the proposed image stitching, spatial coding, and crack-mapping methodology maintain high reliability and practical applicability even in geometrically complex settings.

After generating individual panoramas for the 13 girder segments, these local views were assembled—according to their true dimensions and relative positions—into a complete underside panorama of the bridge (Fig. 9), covering a total area of 360 m$^2$. To reduce the computational complexity of crack localisation, a position-coding scheme was employed. The full panorama was partitioned along the width (X axis) and length (Y axis) into cells of 448 px × 448 px, each assigned a unique two-dimensional code (X,Y). Each segmented cell was then fed into the Light TransUNet model for crack detection. Upon detection, the model returns the encoded cell identifier, enabling rapid localisation of the defect. Using the top-left corner of the panorama as the origin, the pixel coordinates of the segmented crack—together with the cell code—were transformed into global coordinates, precisely mapping each crack to its spatial position on the bridge. This position-coding approach transparently conveys crack locations and allows the individual cells to be reassembled into the full panorama, facilitating seamless transition from local detection to overall inspection. As shown in Fig. 9, the red-shaded cells indicate detected cracks, which are predominantly concentrated near the bridge's hinge regions—a characteristic manifestation of crack development in hollow-slab bridges—thereby providing an intuitive overview of the overall defect distribution.

Based on the panoramic crack map shown in Fig. 9, several representative coded cells were selected for local crack localisation and visual validation (Fig. 10). Each cell's spatial code is directly mapped back to the full panorama. For each region, the original image patch, the set of crack-coordinate points, and the binary segmentation output of Light TransUNet are presented in sequence. The crack centroids and boundary coordinates extracted by the model are transformed via the cell code and pixel-to-metric conversion, then accurately projected into the structural domain, achieving sub-millimetre tracking. This coding-based

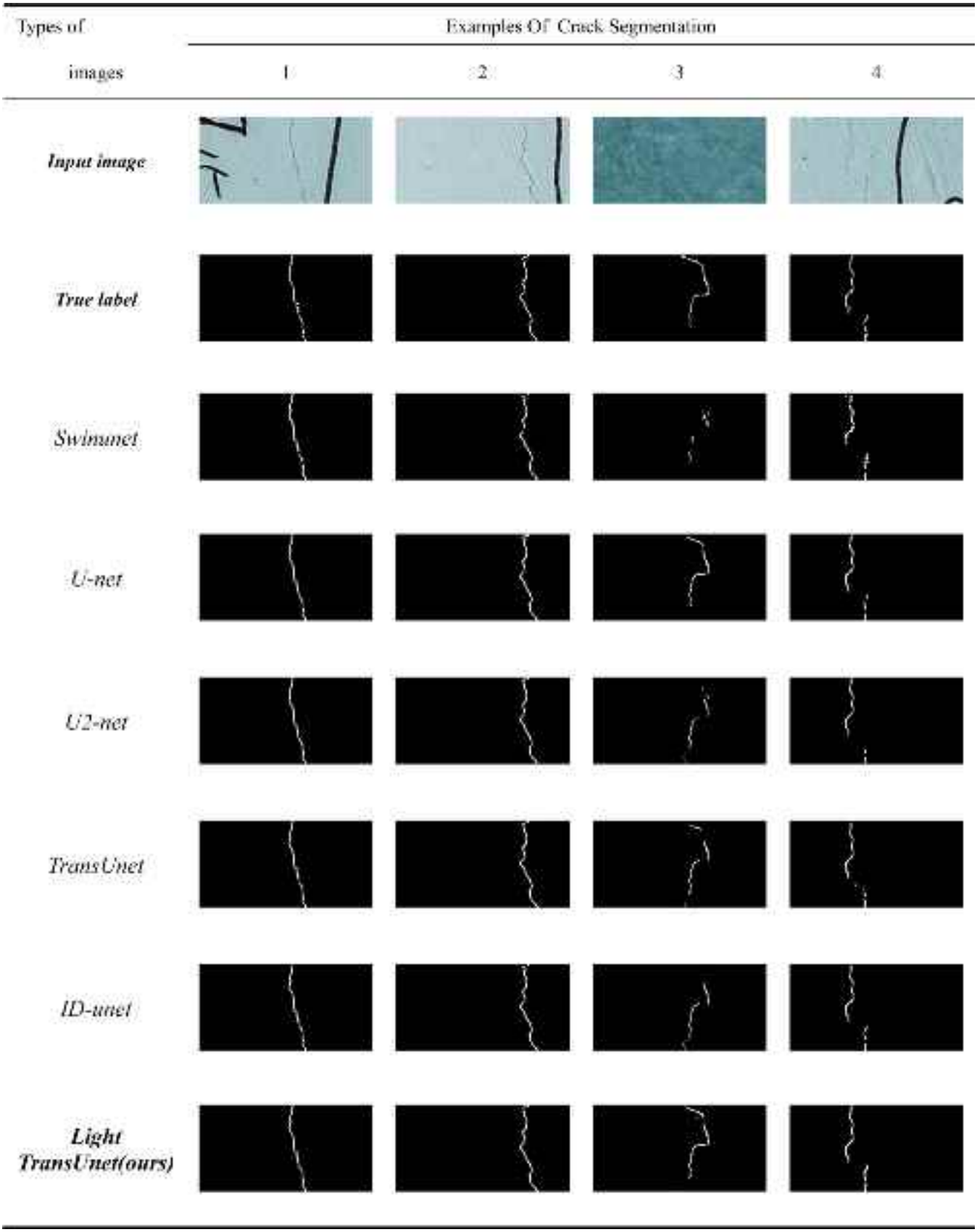


**Fig. 6.** Visual examples of inference results from different models.


**Fig. 7.** Field experiment setup and imagery.

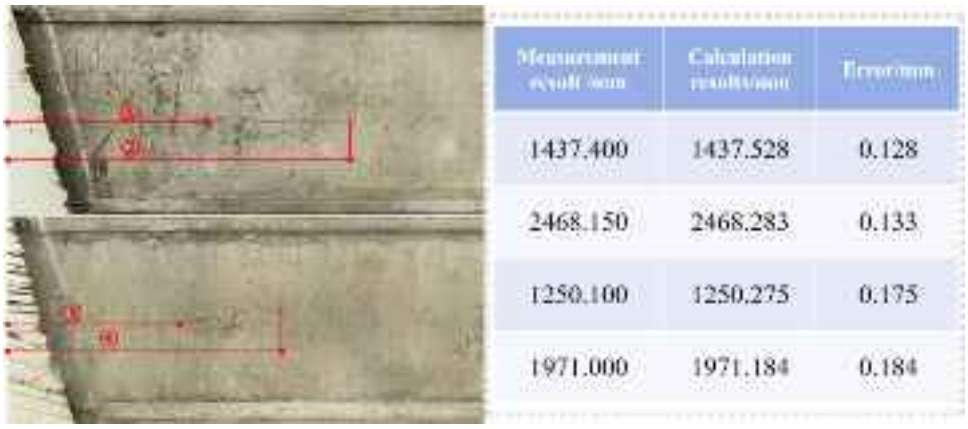


**Fig. 8.** Schematic comparison of girder soffit localisation accuracy.

**Table 2**
Performance comparison of different models on DeepCrack.

| | *Crack Segmentation* | | | | |
|---|---|---|---|---|---|
| Network | **IOU(%)** | **Params (M)** | **Dice(%)** | **Acc (%)** | **Recall (%)** |
| Swinunet | 67.350 | 27.191 | 78.755 | 85.656 | 77.543 |
| U-net | 70.480 | 31.083 | 81.155 | 88.521 | 78.989 |
| U2-net | 71.515 | 44.024 | 82.483 | 81.343 | 87.430 |
| TransUnet | 71.863 | 93.231 | 82.710 | 85.777 | 82.566 |
| ID-unet | 71.929 | 1.143↑ | 82.699 | 89.396 | 79.581 |
| **Light TransUnet (ours)** | **75.621↑** | **12.671** | **86.180↑** | **86.881** | **88.290↑** |

localisation approach preserves lossless reintegration of local detection results into the global panorama while significantly streamlining defect annotation, statistical analysis, and maintenance workflows. It offers structural interpretability, practical applicability, and system scalability, all while enhancing localisation precision.

## 6. Conclusion

To address the complexity and technical challenges of detecting soffit defects in short-span hollow-slab highway bridges, this study has proposed and validated an efficient system integrating unmanned aerial vehicles (UAVs), incremental Structure-from-Motion (SFM) pose estimation, and an enhanced Light TransUNet recognition model for

**Table 3**
Ablation study of Light TransUnet on the ours dataset.

| Experience | StarNet | Layer Reduction | Mixnet | IOU (%) | FLOPs (G) | Params (M) | Inference Time(s) |
|---|---|---|---|---|---|---|---|
| 1(TransUnet) | | | | 77.616 | 98.908 | 93.231 | 0.042 |
| 2 | √ | | | 74.386 | 83.948 | 89.845 | 0.040 |
| | | | | (↓3.32) | (↓14.96) | (↓3.39) | (↓0.002) |
| 3 | | √ | | 75.490 | 43.148 | 22.352 | 0.029 |
| | | | | (↓2.13) | (↓55.76) | (↓70.88) | (↓0.013) |
| 4 | | | √ | 81.58 | 71.927 | 89.561 | 0.032 |
| | | | | (↑3.96) | (↓26.98) | (↓3.67) | (↓0.010) |
| 5 | √ | √ | | 74.239 | 28.188 | 18.966 | 0.022 |
| | | | | (↓3.38) | (↓70.72) | (↓74.27) | (↓0.020) |
| 6(Ours) | √ | √ | √ | 80.97 | 23.241 | 12.671 | 0.027 |
| | | | | (↑3.36) | (↓75.67) | (↓80.56) | (↓0.015) |

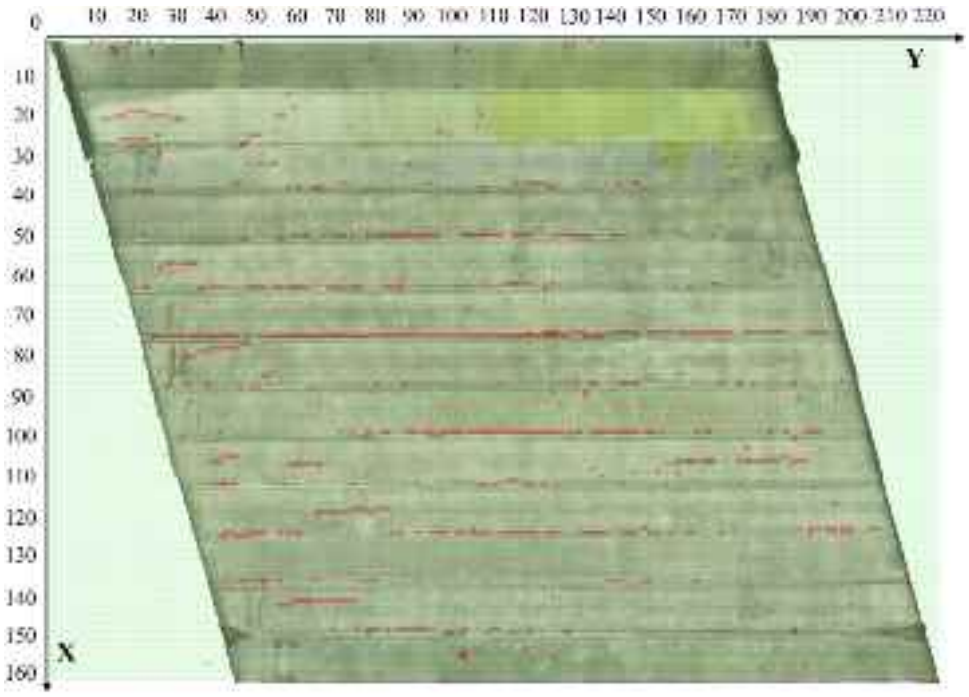


**Fig. 9.** Girder-soffit block-based localisation schematic.

panoramic underside stitching and precise crack detection. First, a UAV equipped with a high-resolution camera acquires gap-free imagery, and an incremental SFM pipeline rapidly reconstructs a sparse point cloud while accurately recovering camera poses. Singular-value decomposition (SVD) is then applied to fit the point cloud to an optimal plane, establishing a unified coordinate frame for the soffit; all views are homographically projected and stitched onto this plane to generate a sub-millimetre-resolution underside panorama. A lightweight, improved Light TransUNet deep-learning model is subsequently employed for fine crack segmentation and precise spatial localisation. Experimental results demonstrate that the system significantly enhances both the efficiency and accuracy of soffit defect detection, providing reliable data support for bridge maintenance and health monitoring. The principal innovations of this work include:

(1) A matrix-pattern UAV imaging and incremental SFM–based rapid pose-estimation technique is proposed. By acquiring soffit images in a grid pattern, inspection coverage is maximised and manual intervention minimised. An incremental Structure-from-Motion (SFM) algorithm then quickly recovers a sparse point cloud and camera poses, and singular-value decomposition (SVD) precisely fits the soffit reference plane. This enables high-precision, low-distortion panoramic stitching of a single girder segment within six minutes.
(2) A seamless stitching and matrix-coding localisation pipeline combining WB-LUTS white-balance correction with Laplacian-pyramid fusion is developed. To address illumination and colour-temperature variations across viewpoints, a deep-learning WB-LUTS algorithm first unifies image colour and brightness. Grid-based image tiling and Laplacian blending then produce a visually continuous, seam-free high-fidelity panorama in an additional two minutes. A matrix-coded spatial indexing scheme assigns each grid cell a unique identifier, facilitating accurate mapping of defect detections and practical deployment.
(3) A lightweight crack-segmentation and spatial-localisation algorithm based on an improved Light TransUNet model is introduced. By replacing the ResNet backbone with the compact StraNet and reducing the number of Transformer layers, the model parameter count is cut by 86.4 %, while inference on 448 × 448 px inputs reaches 37 fps. This enables complete soffit crack segmentation and recognition in under 15 min. Leveraging the matrix-coding index, segmented cracks are precisely reprojected onto the actual soffit, achieving sub-millimetre localisation and providing accurate, reliable data for bridge health monitoring.

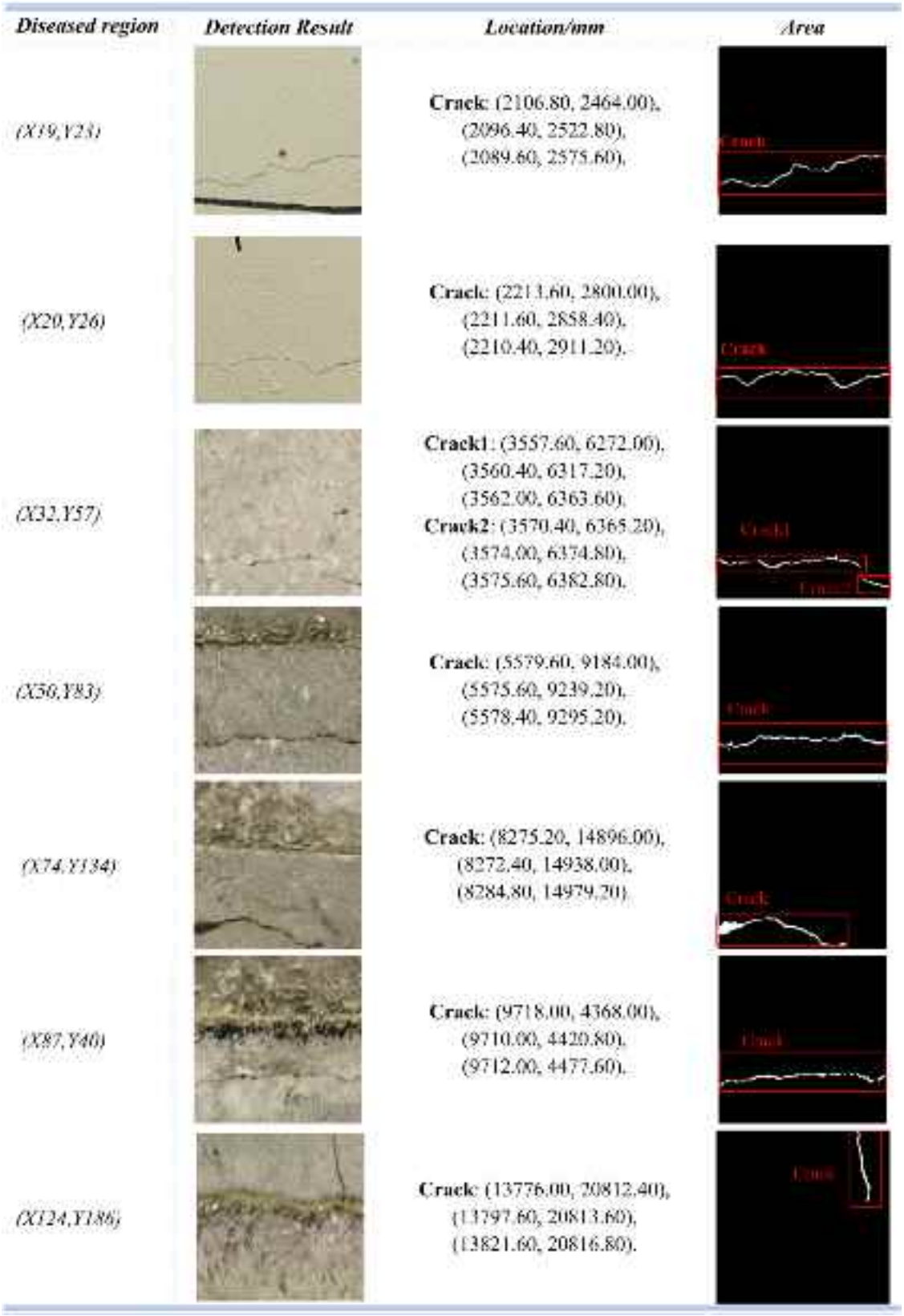


**Fig. 10.** Crack detection and spatial localisation results in representative defect regions.

The proposed panoramic-stitching and crack-identification system effectively reduces manual intervention in soffit-defect inspection, enhances detection efficiency, and improves localisation accuracy, demonstrating strong practical applicability. Nevertheless, several technical challenges remain. First, the system is currently suitable only for planar soffits and is not applicable to girders with significant curvature or complex geometries. Second, SfM-based pose recovery and sparse point-cloud reconstruction for the entire soffit region are heavily affected by the scale of image data, typically requiring over two hours, which necessitates a patch-and-stitch workflow. Third, the generated panoramas are extremely large (4–5 GB each), imposing substantial storage and real-time processing demands and limiting the potential for online deployment. Future research directions include: (1) developing planar-projection techniques capable of handling curved or complex structures by computing the relationships between planes and curved or spatial surfaces to reduce geometric distortion; (2) designing more efficient pose-estimation and point-cloud reconstruction algorithms, such as sparse-dense hybrid feature methods or learning-based rapid pose solvers, to lower computational complexity; (3) incorporating GPU/multi-core parallelisation and streaming-pipeline processing to accelerate end-to-end performance in key stages, including feature matching, incremental bundle adjustment (BA), and image projection stitching; and (4) exploring distributed partition-and-merge frameworks with incremental updates to further compress overall processing time while maintaining stitching quality, thereby enabling efficient, comprehensive, and near-real-time monitoring of soffit defects.

## CRediT authorship contribution statement

**Jiangfan Zhao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jian Zhang:** Writing – review & editing, Validation, Methodology. **WenBin Xu:** Writing – original draft, Software. **Wang Chen:** Writing – review & editing, Validation, Methodology. **Rui Zhao:** Writing – review & editing, Methodology.

## Declaration of Competing Interest

The authors declare that they do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted in this paper.

## Acknowledgements

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).

## References

[1] M.-D. Cui, UAV inspection system and visual damage recognition for hard-to-reach bridge under-deck areas (Ph.D. thesis), Southeast University, 2023, https://doi.org/10.27014/d.cnki.gdnau.2023.003795 (in Chinese).
[2] Gao S-X, Wang X-C. A novel crack-width measurement method and integrated wireless crack detector. Build Technol 2019;3:96–8. ⟨https://www.ndhx.net/tougao/qikan-73875.html⟩ (in Chinese).
[3] Rucka M, Wilde K. Ultrasound monitoring for evaluation of damage in reinforced concrete. Bull Pol Acad Sci Tech Sci 2015;63:65–75. https://doi.org/10.1515/bpasts-2015-0008.
[4] Zhou W. Research on crack depth detection based on impact elastic-wave method. Jiangxi Build Mater 2017;7:13–4. https://doi.org/10.3969/j.issn.1006-2890.2017.07.010 (in Chinese).
[5] Ni F-T, Zhang J, Chen Z-Q. Pixel-level crack delineation in images with convolutional feature fusion. Struct Control Health Monit 2019;26:e2286. https://doi.org/10.1002/stc.2286.
[6] Jiang S, Cheng Y, Zhang J. Vision-guided unmanned aerial system for rapid multiple-type damage detection and localization. Struct Health Monit Int J 2023; 22:319–37. https://doi.org/10.1177/14759217221084878.
[7] Xie R, Yao J, Liu K, Lu X, Liu Y, Xia M, Zeng Q. Automatic multi-image stitching for concrete bridge inspection by combining point and line features. Autom Constr 2018;90:265–80. https://doi.org/10.1016/j.autcon.2018.02.021.
[8] Wang D, Zhang Y, Pan Y, Peng B, Liu H, Ma R. An automated inspection method for the steel box girder bottom of long-span bridges based on deep learning. IEEE Access 2020;8:94010–23. https://doi.org/10.1109/ACCESS.2020.2994275.
[9] Chen W, Yuan B, Chen D, Hu Y, Wang F, Zhang J. Synchronized identification and localization of defect on the bottom of steel box girders based on a dynamic visual perception system. Comput Ind 2025;169:104291. https://doi.org/10.1016/j.compind.2025.104291.
[10] Lin J, Lei S, Huang S, Li M, Xiao Q, Huang Y. Research on intelligent detection technology of bridge surface defects based on UAVs. Proc 4th Int Conf Mach Learn Comput Appl (ICMLCA) 2023:878–87. https://doi.org/10.1145/3650215.3650369.
[11] Jiang S, Cheng Y, Zhang J. Vision-guided unmanned aerial system for rapid multiple-type damage detection and localization. Struct Health Monit Int J 2023; 22:319–37. https://doi.org/10.1177/14759217221084878.
[12] Liu Y-F, Cho S, Spencer BF, Jr, Fan J-S. Concrete crack assessment using digital image processing and 3D scene reconstruction. J Comput Civ Eng 2016;30: 04014124. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000446.
[13] Xu Y, Zhang J. UAV-based bridge geometric shape measurement using automatic bridge component detection and distributed multi-view reconstruction. Autom Constr 2022;140:104376. https://doi.org/10.1016/j.autcon.2022.104376.
[14] Feng C-Q, Li B-L, Liu Y-F, Zhang F, Yue Y, Fan J-S. Crack assessment using multi-sensor fusion simultaneous localization and mapping (SLAM) and image super-resolution for bridge inspection. Autom Constr 2023;155:105047. https://doi.org/10.1016/j.autcon.2023.105047.
[15] DeTone D, Malisiewicz T, Rabinovich A. SuperPoint: Self-supervised interest point detection and description. Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops. 2018. p. 224–36. https://doi.org/10.1109/CVPRW.2018.00060.
[16] Zhao X, Wu X, Miao J, Chen W, Chen PCY, Li Z. ALiKE: accurate and lightweight keypoint detection and descriptor extraction. IEEE Trans Multimed 2022;25: 3101–12. https://doi.org/10.1109/TMM.2022.3155927.
[17] Potje G, Cadar F, Araujo A, Martins R, Nascimento ER. XFeat: accelerated features for lightweight image matching. Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2024:2682–91. ⟨https://arxiv.org/abs/2404.19174⟩.
[18] Lindenberger P, Sarlin PE, Pollefeys M. LightGlue: local feature matching at light speed. Proc IEEE/CVF Int Conf Comput Vis 2023:17581–92. ⟨https://arxiv.org/abs/2306.13643⟩.
[19] Zhang L, Yang F, Zhang YD, Zhu YJ. Road crack detection using deep convolutional neural network. Proc IEEE Int Conf Image Process (ICIP) 2016:3708–12. https://doi.org/10.1109/ICIP.2016.7533052.
[20] Cha Y-J, Choi W, Buyukozturk O. Deep learning-based crack damage detection using convolutional neural networks. ComputAided Civ Infrastruct Eng 2017;32: 361–78. https://doi.org/10.1111/mice.12263.
[21] Cha Y-J, Choi W, Suh G, Mahmoudkhani S, Buyukozturk O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. ComputAided Civ Infrastruct Eng 2018;33:731–47. https://doi.org/10.1111/mice.12334.
[22] Yang C, Chen J, Li Z, Huang Y. Structural crack detection and recognition based on deep learning. Appl SciBasel 2021;11:2868. https://doi.org/10.3390/app11062868.
[23] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. Proc. IEEE Conf. Computer Vision and Pattern Recognition. 2016. p. 779–88. https://doi.org/10.1109/CVPR.2016.91.
[24] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC. SSD: single shot MultiBox detector. Lecture Notes Computer Science (ECCV) 2016;9905:21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
[25] Peng YN, Liu M, Wan Z, Jiang WB, He WX, Wang YN. A dual deep network based on the improved YOLO for fast bridge surface defect detection. Acta Autom Sin 2022;48:1018–32. https://doi.org/10.16383/j.aas.c210807.
[26] Zhang J, Qian S, Tan C. Automated bridge crack detection method based on lightweight vision models. Complex Intell Syst 2023;9:1639–52. https://doi.org/10.1007/s40747-022-00876-6.
[27] Luo Y, Ling J, Wang J, Zhang H, Chen F, Xiao X, Lu N. SFW-YOLO: a lightweight multi-scale dynamic attention network for weld defect detection in steel bridge inspection. Measurement 2025:117608. https://doi.org/10.1016/j.measurement.2025.117608.
[28] Zhang A, Wang KCP, Fei Y, Liu Y, Tao S, Chen C, Li JQ, Li B. Deep learning-based fully automated pavement crack detection on 3D asphalt surfaces with an improved CrackNet. J Comput Civ Eng 2018;32:04018041. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000775.
[29] Ni F, Zhang J, Chen Z. Zernike-moment measurement of thin-crack width in images enabled by dual-scale deep learning. ComputAided Civ Infrastruct Eng 2019;34: 367–84. https://doi.org/10.1111/mice.12421.
[30] Zhu S-Y, Du J-C, Li Y-S, Wang X-P. Method for bridge crack detection based on the U-Net convolutional networks. J Xidian Univ 2019;46(4):35–42. https://doi.org/10.19665/j.issn1001-2400.2019.04.006 (in Chinese).
[31] Zhang Y, Lin Z. Arxiv. Detect Pavement cracks Deep Learn Models Transform UNet 2023. ⟨https://arxiv.org/abs/2304.12596⟩.

[32] Schönberger JL, Frahm J-M. Structure-from-Motion revisited, proceedings of. IEEE Conf Comput Vis Pattern Recognit (CVPR) 2016:4104–13. https://doi.org/10.1109/CVPR.2016.445.
[33] S.K.R. Manne, M. Wan, W.B. LUTs: Contrastive Learning for White Balancing Lookup Tables, arXiv preprint arXiv:2404.10133. (2024) 1–16, https://doi.org/10.48550/arXiv.2404.10133.
[34] Lowe DG. Distinctive image features from Scale-Invariant keypoints. Int J Comput Vis 2004;60:91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
[35] Ma X, Zhang J. Arxiv preprint. arXiv:2403.19967 ( Rewrite Stars Reveal Power Star Oper Effic Netw 2024. https://doi.org/10.48550/arXiv.2403.19967. arXiv:2403.19967 (.
[36] L. Chen, D. Merhof, MixNet: Multi-modality Mix Network for Brain Segmentation, arXiv preprint. arXiv:2004.09832 (2020), https://doi.org/10.48550/arXiv.2004.09832.
[37] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Arxiv preprint. arXiv:2102.04306 ( TransUNet Transform Make Strong Encoders Med Image Segm 2021. https://doi.org/10.48550/arXiv.2102.04306. arXiv:2102.04306 (.
[38] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (MICCAI) 2015;9351:234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
[39] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-UNet: U-Net-like pure transformer for medical image segmentation. Lect Notes Comput Sci (ECCV Workshops) 2022;13669:205–18. https://doi.org/10.1007/978-3-031-25066-8_9.
[40] Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U$^2$-Net: going deeper with nested U-structure for salient object detection. Pattern Recognit 2020;106:107404. https://doi.org/10.1016/j.patcog.2020.107404.
[41] Chen D, Qin F, Ge R, Peng Y, Wang C. ID-UNet: a densely connected UNet architecture for infrared small target segmentation. Alex Eng J 2025;110:234–44. https://doi.org/10.1016/j.aej.2024.09.108.
[42] Liu Y, Yao J, Lu X, Xie R, Li L. DeepCrack: a deep hierarchical feature learning architecture for crack segmentation. Neurocomputing 2019;338:139–53. https://doi.org/10.1016/j.neucom.2019.01.036.