Computers in Industry 169 (2025) 104291

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Synchronized identification and localization of defect on the bottom of steel box girders based on a dynamic visual perception system

Wang Chen [a], Binhong Yuan [b], Dongliang Chen [b], Yong Hu [b], Feiyu Wang [a], Jian Zhang [a,c,*]

[a] School of Civil Engineering, Southeast University, Nanjing, China
[b] Guangdong Jiaoke Testing Co. Ltd., Guangzhou, China
[c] Jiangsu Key Laboratory of Mechanical Analysis of Infrastructure and Advanced Equipment, Nanjing, China

ARTICLE INFO

Keywords:
Bridges
Deep learning
Panoramic image
Identification and localization

ABSTRACT

Inspecting the underside of large-span bridges is a major challenge due to the extensive area and inaccessibility. This study developed a system that integrates advanced equipment with intelligent algorithms, designed to achieve precise identification and rapid localization of defects on the underside of bridges. The key components of the system are summarized as follows: (1) The dynamic visual perception system is composed of a perception module, a control and transmission module, and a motion module. It enables automated data collection at any position beneath the bridge structure. (2) A block-based panoramic generation strategy is employed, which uses a spatially ordered block concept to simplify the panorama stitching process and enhance accuracy. (3) Deep learning-driven two-phase synchronous identification and localization method. In the first phase, MobileNetV4 serves as the primary feature representation tool, facilitating the lightweight reconstruction of panoramic images. In the second phase, the YOLOv9 detection framework is employed to perform a precise analysis of the identified defect regions, providing detailed defect information on a localized level. The design of this system significantly enhances the efficiency and accuracy of inspections of large-span bridge undersides, offering robust technical support for bridge health maintenance. Experimental results indicate that the proposed method achieves over 90 % accuracy in defect recognition tasks, alongside millimeter-level precision in localization.

1. Introduction

Bridges are vital components of modern infrastructure, serving as critical nodes in transportation networks and fundamental support structures for economic development and resource management. During their service life, bridges are subjected to various challenges such as structural fatigue, material degradation, environmental corrosion, and overload stress, all of which can significantly impact their lifespan and safety performance. To effectively mitigate these risks, regular health inspections have become an indispensable practice. However, as technology advances and inspection standards become more stringent, traditional manual methods are increasingly inadequate for meeting the efficiency and accuracy requirements of contemporary bridge assessments (Rubio et al., 2019). This issue is particularly pronounced in inspecting the undersides of bridges, where hidden locations and complex environments greatly complicate the process (Liu and Chou, 2023). Consequently, there is an urgent need to implement advanced technologies for these inspections.

In recent years, researchers have developed specialized equipment to address the bottlenecks in the aforementioned scenarios. These technologies can be broadly categorized into three main types. First, Unmanned Aerial Vehicles (UAVs) have become a popular tool for inspections due to their flexibility and efficiency (Yoon et al., 2022; Yasuda et al., 2022). Sanchez-Cuevas et al. (2017) designed a novel multi-rotor UAV capable of adhering to the underside of bridge girders for precise contact-based inspections. This system leverages the ceiling effect to enhance both endurance and stability. Conversely, Ikeda et al. (2019) addressed the ceiling effect by integrating a multi-degree-of-freedom robotic arm into UAVs, enabling high-precision, non-contact inspections. Further advancing the field, Wang et al. (2020) created a system that crawls along the underside of girders to collect real-time data, enhancing the efficiency of defect image acquisition. Humpe (2020) aimed to improve data acquisition efficiency by combining panoramic cameras with UAVs to expand the field of view. However, this method encounters issues such as severe image distortion from the camera type and acquisition mode, which

* Correspondence to: School of Civil Engineering, Jiangsu Key Laboratory of Engineering Mechanics Southeast University, Nanjing, China.
E-mail address: jian@seu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.compind.2025.104291
Received 10 November 2024; Received in revised form 4 March 2025; Accepted 30 March 2025
Available online 15 April 2025

affects the accurate representation of real-world conditions. To address this, Peng et al. (2021) constructed a system integrated with a three-point laser rangefinder to accurately restore defect dimensions in images. Additionally, weak GPS signals pose a significant challenge for UAV operations under bridges. To overcome this obstacle, Jiang et al. (2023) introduced a vision-guided UAV that combines stereo vision with Inertial Measurement Unit (IMU) fusion technology for accurate positioning. Wang et al. (2023) proposed a Fiducial Marker Corrected Stereo Visual-Inertial Localization (FMC-SVIL) method that achieves high-precision positioning on the underside of bridges. Bolourian et al Bolourian and Hammad, (2020) investigated LiDAR technology for 3D spatial awareness in these inspections. Apart from UAVs, robotic systems offer unique benefits for bridge inspections. Evan et al. (2020) used a mobile robotic platform that efficiently collects multi-source defect data. This platform combines RGB cameras, infrared cameras, LiDAR, Real-Time Kinematic (RTK) GPS, and Inertial Measurement Units (IMUs) for inspections where the girder underside is accessible from the ground. However, this system is not applicable in scenarios where the underside is over water. Meng et al. (2023) created a robotic platform with an extending arm that inspects the girder underside from above. Lyu et al. (2024) presented a vacuum-suction climbing robot for traversing vertical surfaces of concrete bridge towers or piers. Building on these advancements, Lin et al. (2023) developed a mobility inchworm climbing robot with effective vertical and horizontal climbing abilities, making it well-suited for complex inspection tasks. In engineering applications, these intelligent inspection devices are valued for their portability and flexibility. However, a key challenge remains: in the intricate environments beneath bridge girders, most devices still rely heavily on manual operation. This reliance limits automation and results are greatly influenced by operator skill and experience. Moreover, the limited endurance of these devices restricts inspection efficiency.

Vehicle-based detection systems represent a third category of solutions that address the challenges mentioned earlier. Xie et al. (2018) designed a mobile platform with a large mechanical arm fitted with high-resolution industrial CCD cameras, 3D cameras, lasers, and ultrasonic rangefinders, enabling detailed inspection of bridge undersides. Sutter et al. (2018) enhanced this system by adding a mobile collection module to the arm, improving the efficiency of data capture from both the sides and undersides of bridge structures. Jiang et al. (2021) employed the BIR-X vehicle system, which includes multiple cameras on an extended mechanical arm for comprehensive data collection. Lorenzo et al Capponi et al. (2021) advanced this idea by integrating the vehicle system with bridge structures, resulting in an innovative inspection platform with a mechanical arm extending up to 17 m, allowing for extensive area coverage of bridge sections. These platforms minimize the need for manual control and facilitate stable, thorough inspections of bridge defects. However, practical challenges include limited portability and flexibility, as well as specific environmental constraints. For example, these systems frequently occupy emergency lanes. Moreover, in large-span bridge inspections, coverage blind spots may occur, posing significant limitations to their effectiveness.

Variations in individual experience can lead to technicians drawing different conclusions from the same inspection data. This subjectivity not only introduces uncertainty but also affects the efficiency of the analysis. The emergence of deep learning technology, a transformative force across numerous fields, has revolutionized traditional data processing methods. Deep learning models, which mimic the structure of the human brain, use multilayered neural networks to automatically extract and learn complex features from data (Jha and Babiceanu, 2023). Cha et al. (2017) introduced a convolutional neural network for automatic defect identification, which, when combined with a sliding window technique, enables accurate analysis of large-scale concrete crack images. Rao et al. (2021) utilized a non-overlapping window--based method to segment images into regions for automated detection. Kung et al. (2021) developed a detector using a pre-trained VGG-16 network and employed Class Activation Mapping (CAM) to localize defect regions. While these methods show promise for automated analysis, they still require post-processing to manage detailed information. To advance automation, Lin et al. (2021) introduced an architecture that integrates Feature Pyramid Networks (FPN) with Faster R-CNN. This design effectively fuses multi-scale features, enabling end-to-end defect detection with bounding boxes. Wan et al Kung et al. (2021) applied the Single Shot Multibox Detector (SSD) to bridge engineering, enhancing the automation of detection and using the eight-neighborhood algorithm to improve output accuracy. In an innovative study, Zhang et al. (2023) adapted the YOLOv4 model by implementing a lightweight structural design, significantly boosting processing speed without compromising accuracy. To better mimic human visual processing, Wang et al. (2021) employed DeepLabv3 + for pixel-level image analysis, effectively isolating cracks from the background. Kang and Cha, (2022) introduced the multi-head self-attention mechanism from the Transformer architecture to address the limitations of traditional convolutional neural networks in capturing global features. Their method, tested on 545 images, achieved a notable mean Intersection over Union (mIoU) of 92.6 %. Chen et al. (2023) introduced a lightweight attention mechanism-based segmentation network, CrackSeU, which enhances significantly crack detection by integrating multi-level and multi-scale features. To address the challenges of accurately identifying crack boundaries, He et al He et al. (2024a) developed the Boundary Guidance Crack Segmentation Model (BGCrack), which excels in precise boundary identification.

Current research on bridge underside inspection technologies has primarily focused on small to medium-sized bridges, with equipment that heavily relies on manual operation, leaving automation underdeveloped. Inspecting large-span bridges, especially in difficult-to-access underside areas, presents significant challenges due to the complexity of comprehensive and efficient data acquisition. To overcome these challenges, this study developed a dynamic visual perception system that integrates sensing, control, and transmission functions. This system enables engineers to remotely control the equipment via a software interface, facilitating automated and comprehensive inspections. Additionally, deep learning tools are employed to automate the analysis of large datasets, enhancing data processing efficiency. Building on this, the study introduces an innovative two-phase analysis method that provides intuitive guidance for defect regions on a macro level and precisely captures detailed defect progression on a micro level. This approach seamlessly integrates global and local perspectives, offering a novel solution for comprehensive bridge underside inspections.

This paper is organized into seven distinct sections. It begins with an overview of the methodological framework and the dynamic visual perception system. The third section details the block template-based parallel panoramic image reconstruction strategy. In the fourth section, the diffusion model-guided simulation of real-world defect scenarios is explored. The two-phase characterization method is introduced in the fifth section. The sixth section validates the proposed framework through field experiments. Lastly, the seventh section provides insights into potential avenues for future research.

2. Proposed methodology

Inspecting the underside of large-span bridges is crucial due to the vast area and difficult accessibility of these regions. Although intelligent inspection technologies have advanced, they frequently fall short in these complex environments. Currently, visual inspection via under-bridge vehicles is the standard practice. However, this approach is labor-intensive, time-consuming, and susceptible to inaccuracies, comprehensiveness issues, and lack of traceability due to human involvement. To address these challenges, this study introduces an innovative solution—a dynamic visual perception system. The system overcomes existing limitations by efficiently coordinating the perception, control, and motion modules. The perception module works seamlessly with the control module's core edge computing component,

utilizing various wireless sensing techniques. Supported by the motion module, this integration allows experts to remotely operate the system via an interactive interface, automating the thorough inspection of the bridge underside without requiring on-site human presence, as depicted in Fig. 1. The extensive data collected presents significant processing challenges. To tackle this, a panoramic image generation strategy based on a block concept is proposed. This method simplifies and accelerates the stitching process by dividing images into blocks for independent parallel processing. The staple theory is utilized during feature extraction and matching to streamline image alignment and reduce computational complexity by focusing on overlapping feature points. Panoramic images, while providing a comprehensive view, can be overly complex, complicating interpretation. A two-phase deep learning-based recognition and localization method is proposed. In the first phase, MobileNetV4 is employed to rapidly identify key features. A diffusion model

is used during training to simulate realistic defect scenarios, improving the model's generalization. Panoramic images are then reconstructed with lightweight techniques, incorporating feature adjustment and positional encoding to enhance global information representation and accuracy in defect recognition and localization. The second phase concentrates on the detailed analysis of defect areas identified earlier. Positional encoding facilitates quick tracing back to the original block images. The YOLOv9-based detector is employed to further extract detailed information from the defect regions. This method not only offers a global perspective for inspecting bridge undersides but also provides detailed information within defect regions, aiding engineers in implementing targeted maintenance strategies.

2.1. Dynamic visual perception system

The device consists of three main components: the Perception Module (PM), the Control and Transmission Module (CTM), and the Mobility module (MM), as shown in Fig. 2. Positioned at the front, the PM captures environmental data and forwards it to the CTM. Utilizing the distributed cameras, the PM replicates human vision to accurately depict and characterize the bridge's condition. It comprises imaging devices arranged in a spatially parallel setup, interconnected through Bluetooth and WLAN, creating a novel data acquisition network. This design enhances flexibility and scalability by eliminating complex wiring, allowing for adaptable placement and movement of the camera units, ideal for inspecting wide bridge undersides. Each unit can function independently or in synchronized coordination, based on time synchronization, thus improving data acquisition and processing efficiency.

The CTM plays a crucial role within the device, tasked with both controlling and transmitting data. Edge computing is utilized within this module to facilitate these functions. The CTM links to the PM via Bluetooth or WLAN, enabling remote control over operations like initiation, termination, and data transfer. A status feedback mechanism ensures the PM functions optimally. Given the substantial volume of data collected, high-speed wireless communication technology is employed to establish the data transmission channel between the PM and the CTM. A user-friendly control interface has been designed to enhance interaction, allowing technicians to easily configure and adjust data acquisition parameters like timing, mode, focal length, and collection start/stop. The interface also provides control over the data transmission process, allowing for efficient management of the data flow. This design enhances convenience and flexibility for technicians, improving overall system performance and efficiency.

The static PM configuration limits its ability to capture visual data from fixed viewpoints. The MM is integrated into the device, enabling it to overcome the limitations of static sensing. This module introduces a controllable motion mechanism, enabling dynamic visual sensing and providing richer spatial data. The under-bridge inspection vehicle, the core of the MM, allows the PM to capture data from any position beneath the bridge. This design greatly improves bridge underside inspection coverage, ensuring high-quality image data. Its specialized structure and functional setup also simplify module installation and routine maintenance.

3. Panorama generation strategy based on block concept

The dynamic visual perception system, with its integrated design, provides substantial technical benefits for comprehensive data collection from bridge undersides. The vast amount of data captured contains crucial insights into the bridge's condition. However, efficiently extracting and analyzing this data is challenging. Traditional methods for processing large-scale images are resource-heavy, time-consuming, and largely dependent on manual input. Moreover, these methods often focus on localized analysis of individual images, overlooking the connections between images, which limits a thorough assessment of the bridge's overall condition. This study presents a panorama-based method to effectively connect discrete image data, addressing these challenges. By reconstructing the bridge underside panorama with block-based units, the method offers a structured and comprehensive view. This strategy minimizes redundant analysis of overlapping image regions, enhancing data analysis efficiency and providing better feedback on the bridge's overall structural condition.

Traditional panoramic construction methods (Brown and Lowe, 2007; Wang and Yang, 2020), which involve stitching discrete images into a complete image containing all content, have significant limitations when applied to the complex structures of large-span bridges. Consequently, this study proposes a horizontal division of the bridge's underside, based on the coverage area of each unit within the dynamic perception system. Each sub-region is assigned a unique position code $B_n (n \leq N,$ where N is the total number of perception units). Furthermore, the underside is divided in the vertical direction. Based on structural design characteristics, each sub-region's segments are treated as independent blocks, allowing for an orderly discretization of the entire underside space. To build on the original sub-region position codes, segment information is incorporated to generate block position codes $B_n - S_i (i \leq I,$ where I represents the number of segments). Using the block concept, correlations between images within each block of the underside are established. These block images are then integrated based on their spatial position codes to form a complete panoramic image of the underside. The block concept allows for independent and simultaneous processing of each block, as depicted in Fig. 3. This strategy greatly enhances the efficiency of panoramic image generation while also improving the final image quality. Each block is given a unique code, which overcomes the limitations of relying solely on coordinate information for spatial positioning. This block coding mechanism enables the rapid localization of any area within the overall space.

The generation of block images, a crucial part of constructing panoramic images, heavily relies on the effectiveness of feature extraction and matching processes. Feature extraction identifies key details in images and converts them into a format suitable for processing, while feature matching compares and aligns these extracted features. However, images of the underside of steel box girders often lack distinct features, hindering the efficiency and quality of feature extraction. The primary issue is the ineffective control over searching for feature points across the entire image, leading to significant computational resources being wasted on extracting points from non-informative areas. This study draws an analogy to stapling, where two sheets of paper are easily secured at specific points on the left and right. Similarly, in adjacent images, alignment can be efficiently achieved by focusing on feature points within the left and right overlapping sub-regions. This direct approach effectively resolves image alignment challenges, reducing both computational complexity and resource consumption. Specifically, the image is initially divided vertically into feature extraction and reserved regions. The reserved regions are excluded from the processes of feature detection and matching. Taking into account the motion

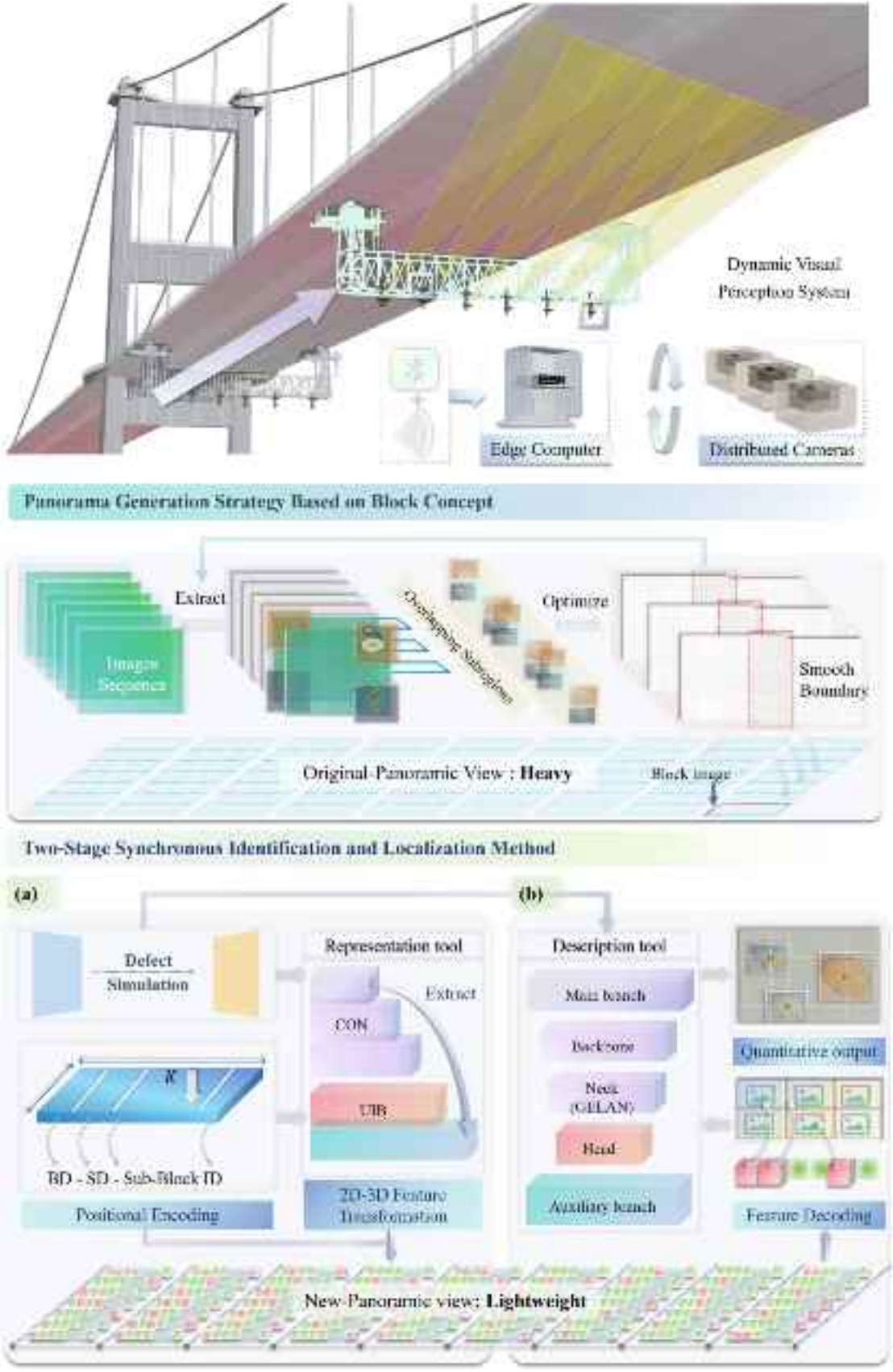


Fig. 1. Framework of defects dynamic perception method.

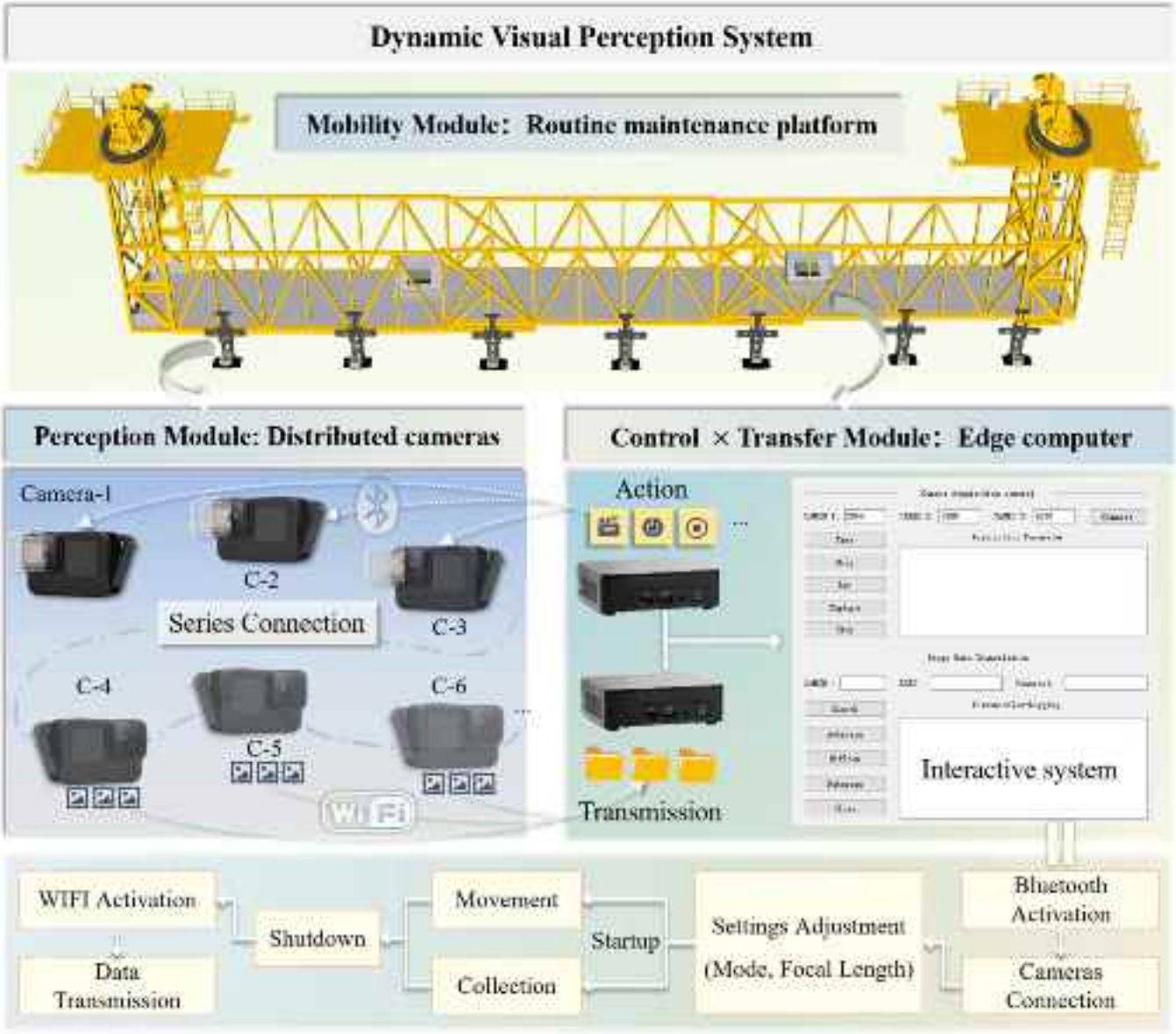


Fig. 2. Schematic diagram of dynamic visual perception system components.

module's maximum speed and the overlap between images captured at adjacent moments, the division ratio between these regions is set at 0.5. The feature extraction region is then horizontally subdivided into left and right staple regions (EL and ER, as depicted in Fig. 3). This subdivision concentrates on feature information within these regions, enhancing efficiency through parallel processing. During matching, similar to stapling, feature points in the left and right staple regions of images from adjacent moments are identified. Finally, homography matrix calculations produce a stable transformation matrix, enabling successful image stitching. With adequate computational resources, image sequences within block images can be grouped, allowing simultaneous stitching of adjacent image pairs. This iterative strategy continues until the block image generation process is complete.

Initial block images may show visible boundaries or transitions in overlapping areas captured at adjacent time intervals. To resolve this, the study applies a pyramid decomposition method (Zhao and Zheng, 2017), utilizing a Gaussian filter and sampling techniques to break down the image into multiple frequency levels. The high- and low-frequency components at each level are then smoothly weighted and combined. Finally, these frequency band components are recombined, effectively eliminating edge discontinuities.

4. Realistic defect scenario simulation guided by diffusion models

Acquiring defect data from the underside of large-span bridges poses significant challenges, which can be effectively addressed through data synthesis technology. Deep learning-based synthesis methods like GAN (Krichen, 2023), VAE (Kingma and Welling, 2024), and Diffusion Models (He et al., 2024b) enable the simulation of scenarios that are challenging to capture directly. These techniques offer reproducible resources that are convenient for developing and testing computer vision algorithms, thereby significantly reducing costs and time. This study utilizes diffusion models, valued for their high-quality image generation and ease of training, to simulate key defect characteristics of the steel box girder underside, such as spalling and rust. While diffusion models produce quality image synthesis results, they have limitations in generation speed. Consequently, the focus was not on generating large-scale defect scenes but rather on synthesizing localized defect areas. The synthesized defect information was then mapped and

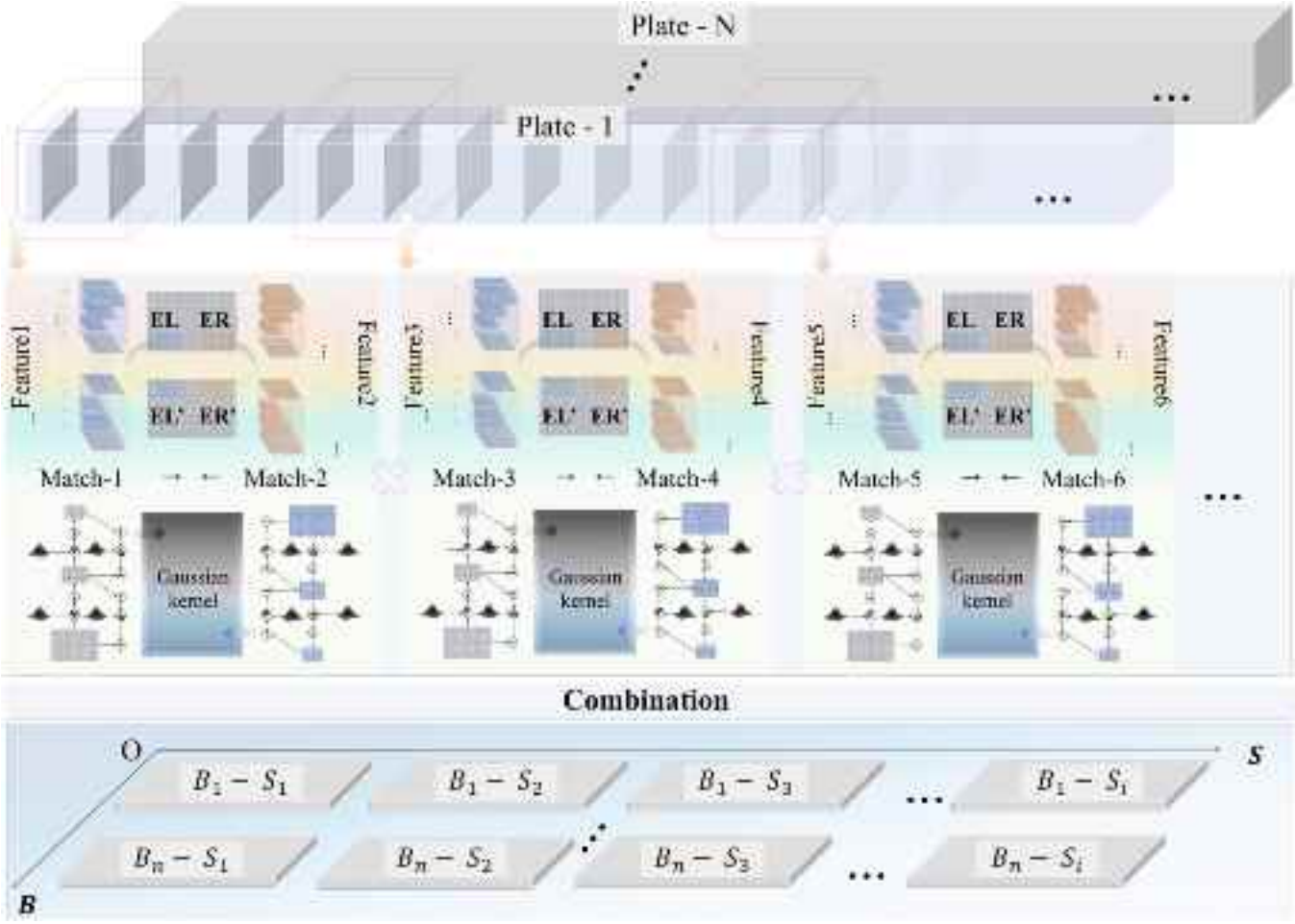


Fig. 3. Panoramic generation process based on the block concept.

integrated into real inspection scenarios using techniques like poisson blending (Pérez et al., 2003), resulting in realistic simulations, as depicted in Fig. 4.

The diffusion model is comprised of two main stages: forward diffusion and reverse denoising (Ho et al., 2020), as shown in Fig. 4. The initial phase transforms a sampled image $Z_0 \sim \varphi(Z)$ (where $\varphi(Z)$ denotes the empirical distribution of real images) into progressively noisier representations. At each timestep t, the image state Z_t is generated by adding controlled Gaussian noise to the preceding state Z_{t-1}. Through $\mathcal{T}$ successive transformations, the original image structure is gradually degraded, ultimately converging to a Gaussian noise distribution at $Z_{\mathcal{T}}$.

The reconstruction phase aims to learn the inverse mapping from noise $Z_{\mathcal{T}}$ back to the data manifold. A parameterized neural network θ is trained to approximate the conditional probability $p_\theta(Z_{t-1}|Z_t)$ by estimating the mean and variance of the reverse transition. Through iterative refinement, the trained model can generate photorealistic defect

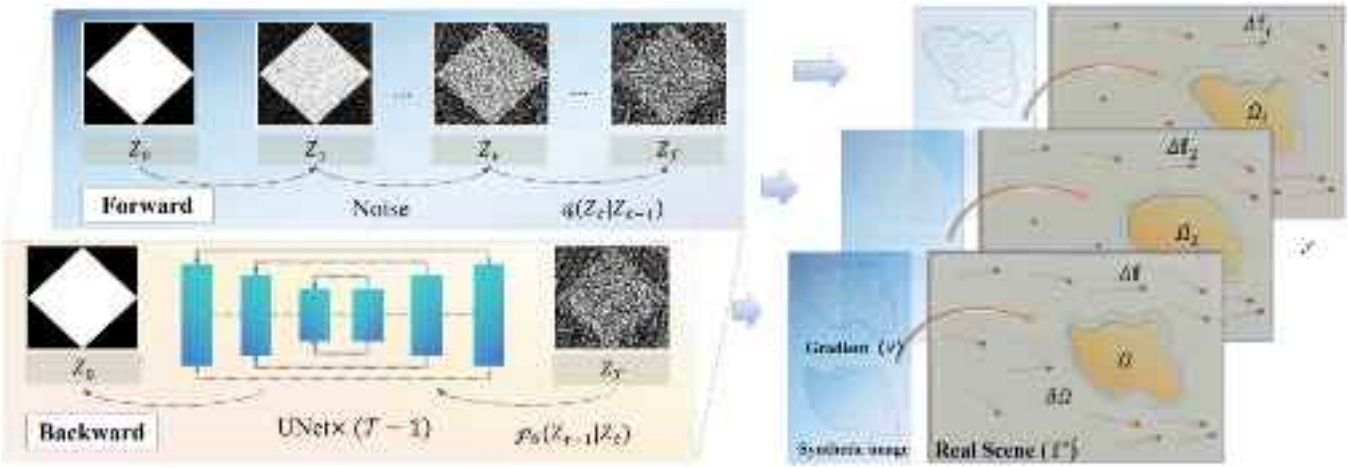


Fig. 4. Realistic defect scenario generation using diffusion modeling and fusion techniques.

patterns, providing high-fidelity synthetic data for industrial inspection systems.

In further research, this study employs poisson blending techniques. The localized defect images generated by the diffusion model are seamlessly integrated into standard steel box girder background images, effectively simulating real defects on the girder undersides. Before blending, the generated defect images are rotated, scaled, and adjusted for color and brightness. These augmented samples, which vary in perspective, size, and lighting, enrich the data foundation for the simulation process. A blending region is then arbitrarily selected in the target image, and the gradient information of this region is calculated to capture the trends in image detail changes. Finally, a poisson equation-based mathematical model is constructed, using the boundary characteristics of the source image and the gradient information of the target image as constraints. Solving the poisson equation yields the pixel value distribution in the blended region:

$$\min_{\mathbb{f}} \iint_{\Omega} |\Delta \mathbb{f} - \mathbb{V}|^2 with \mathbb{f}|\partial\Omega = \mathbb{f}^*|\partial\Omega \tag{1}$$

Where $\mathbb{f}$ is the resulting blended image, $\mathbb{f}^*$ denotes the real-world scene, and $\Delta\mathbb{f}$ indicates the gradient information of $\mathbb{f}$. $\mathbb{V}$ represents the gradient of the synthesized defect image, Ω is the blending region, and $\partial\Omega$ refers to its boundary.

5. Two-phase synchronous identification and localization method

Large-scale block images of the bridge underside offer preliminary insights into localized defects. However, when these images are merged into a panoramic view, a nonlinear decay in information occurs, meaning the combined result is less than the sum of its parts (i.e., $1+1<2$). This can lead to critical defect features being overlooked or diminished in the broader context, compromising the accuracy of the analysis and diminishing the overall value of the panoramic image. To address this, an innovative two-phase method for synchronized defect identification and localization is proposed, as illustrated in Fig. 5. The first phase involves reconstructing the panoramic image using a deep feature representation method, creating a synergistic effect where $1+1>2$. This approach not only enhances defect visibility but also enables quick localization of defect areas while maintaining a lightweight global information representation. In the second phase, the features learned in the first stage are further refined through deep learning, providing a comprehensive understanding of the bridge's condition, including the number, type, and precise location of defects.

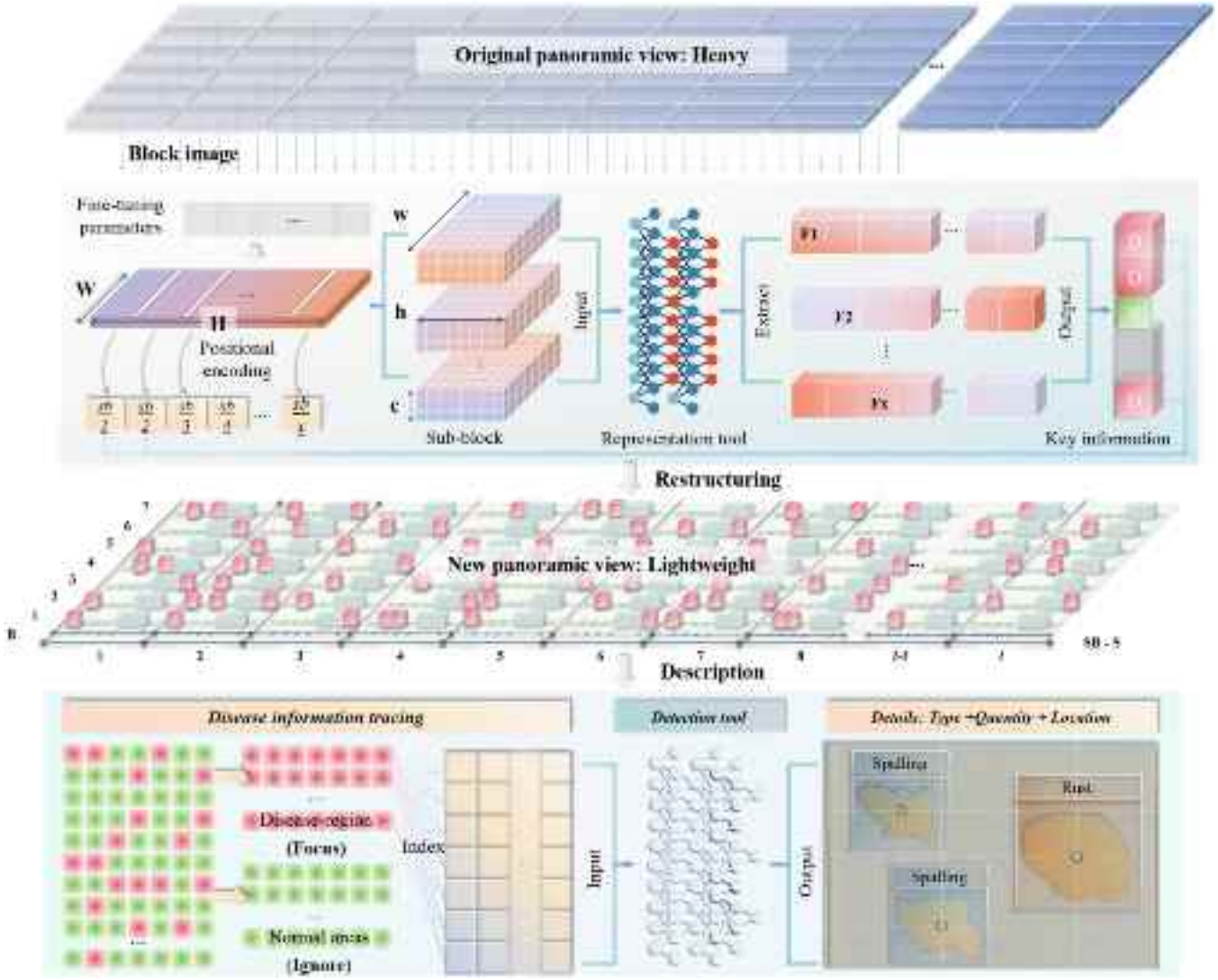


Fig. 5. Two-phase approach for synchronized identification and localization of steel box girder bottom defect.

5.1. Deep feature representation-based lightweight panoramic view reconstruction

Block images contain non-essential background information that does not significantly aid operational analysis. An innovative image processing strategy is proposed to remove irrelevant background details and focus on defect-related features. This study introduces intelligent analysis techniques based on deep learning to efficiently extract these critical features. Deep learning, which mimics the multi-layered structure of human brain neural networks, can autonomously learn and identify complex patterns from data (LeCun et al., 2015). Compared to traditional handcrafted feature extraction methods (Cord and Chambon, 2012; Talab et al., 2016), it offers greater robustness and practicality, especially in processing large-scale image data. In computer vision, two primary architectures dominate: Convolutional Neural Networks (CNNs) (Li et al., 2022) and Transformers (Talab et al., 2016). CNNs are effective at capturing local image features and progressively building more complex, abstract representations. Meanwhile, Transformer architectures, initially successful in natural language processing (Vaswani et al., 2017), have recently gained attention in computer vision due to the self-attention mechanism (Dosovitskiy et al., 2020; Liu et al., 2021), which handles long-range dependencies. Although Transformers offer better prediction accuracy, their computational complexity can lower efficiency. Through structural design and algorithm optimization, CNNs can match Transformer architectures in accuracy (Liu et al., 2022). Given the practical engineering focus of this study, the highly efficient and high-performing MobilenetV4s (Liu et al., 2021) was selected as the feature representation tool.

The exceptional performance of MobilenetV4s stems from its innovative Universal Inverted Bottleneck (UIB) design, as illustrated in Fig. 6. The MobileNet series is distinguished by its integration of depthwise convolution (DW) and pointwise convolution (PW), optimizing parameter efficiency and pioneering research into lightweight convolutional modules. UIB employs a standardized inverted bottleneck architecture, building on DW and PW to create a universal framework. The design features an expansion layer and a projection layer, with DW modules embedded before and after the expansion layer. This modular strategy results in three variants: Extra Depthwise Inverted Bottleneck (ED), ConvNext (CN), and Inverted Bottleneck (IB). The IB module follows the classic inverted bottleneck design, enhancing information encoding by expanding feature channels in the DW layer. It refines intermediate feature representations through DW, promoting smooth gradient propagation. By decoupling the input/output feature expressiveness from the transition layer, it offers an efficient framework for deeper network analysis. The ED module adds a DW before the IB module's expansion layer, expanding the network's receptive field and capturing richer contextual information. By separating spatial convolution from channel mixing, DW reduces parameters and multiply-accumulate operations, allowing greater network depth without a significant computational burden. The CN module simplifies the network by omitting the DW between the expansion and projection layers, following the ED module's design philosophy. This approach reduces the

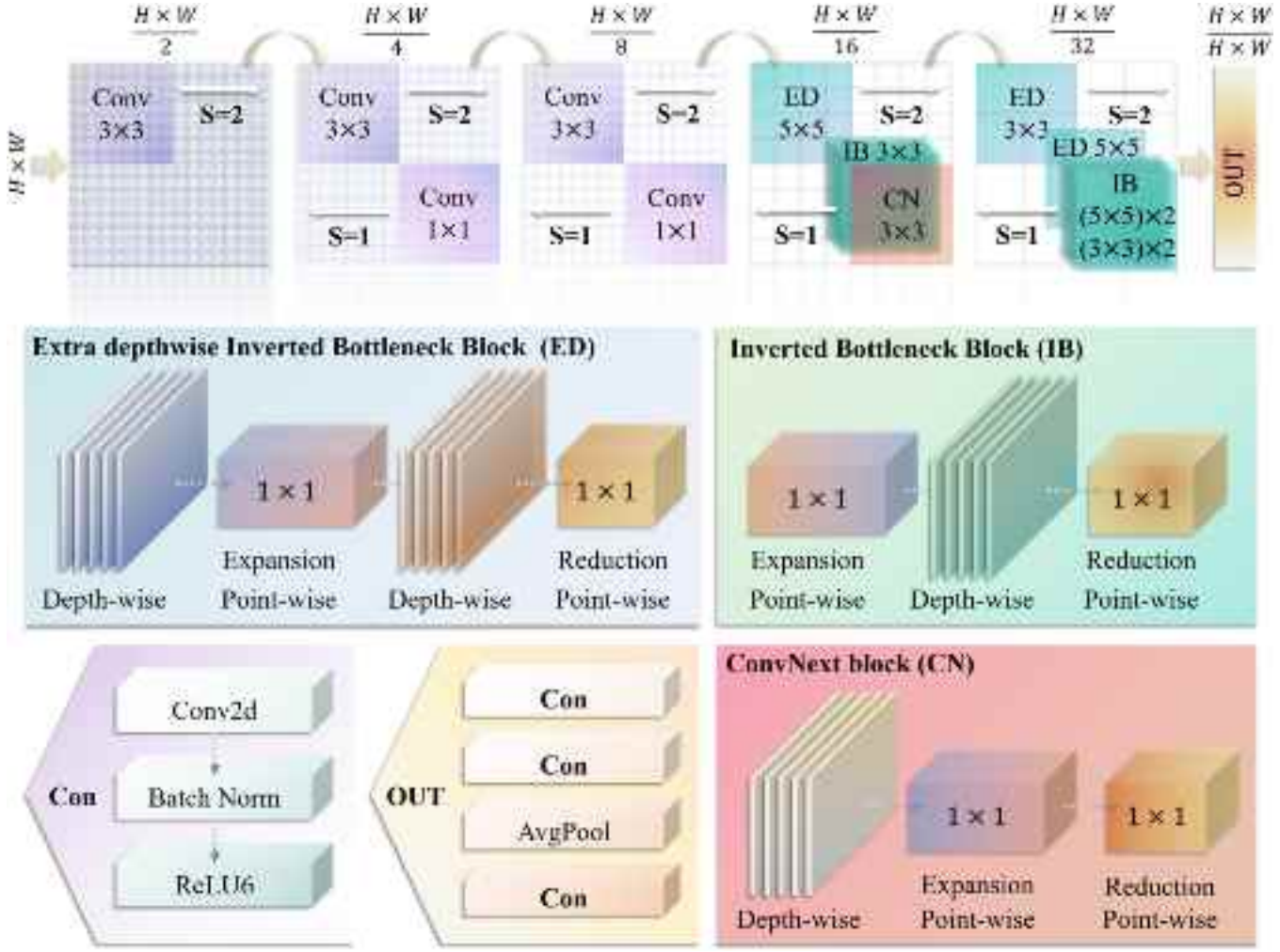


Fig. 6. Architecture of the feature representation tool.

model's computational burden while ensuring efficient and expressive integration of spatial features.

This study introduces the MobilenetV4s network for the initial analysis of block images. MobilenetV4s, known for its lightweight design and efficient feature extraction, facilitates end-to-end automated processing from input to deep feature representation. Then, these features are characterized using a straightforward digital encoding system. In this system, '0' indicates no detected defect, signifying normal service, while '1' indicates defects, suggesting an abnormal service state. It simplifies data structure, preserves critical spatial information via positional encoding, and offers an effective data compression method for lightweight panoramic image reconstruction. However, simple 0–1 encoding may not fully capture the complex attributes of defect in machine vision. color labels are introduced to enhance the visualization. '0' is assigned a red label and '1' a green label, offering intuitive cues for quick identification of damaged areas. To further highlight defects and enrich spatial information, a multi-dimensional feature representation method is proposed. In this method, the feature representation of damaged areas is extended into three-dimensional space, while normal areas remain in two-dimensional space. Building on the original 0–1 encoding, key information is replicated in three-dimensional space by assigning each block an additional attribute value along the third dimension, which is perpendicular to the two-dimensional plane. In this space, damaged areas have a third dimension value of 1, while normal areas retain a value of 0. The reconstructed panoramic image retains informational integrity while remaining significantly lightweight. It is crucial for large-scale bridge projects, as it reduces data storage and processing burdens, facilitates cross-platform sharing and analysis, and serves as a convenient decision-support tool for maintenance teams.

In the reconstruction phase, precise feature extraction is essential for characterizing defects. When processing large-scale block images, relying on a single feature can lead to insufficient defect capture and excessive background noise, wasting computational resources and reducing information effectiveness. A feature granularity adjustment parameter $K(K \leq h,$ where h is the block image's vertical length) is introduced to optimize defect feature representation during global reconstruction. Focusing on high aspect ratio scenarios common in bridge engineering, K is applied to the vertical axis of the block image, enabling uniform subdivision into multiple sub-blocks. This subdivision facilitates more refined feature extraction from each sub-block, accurately reflecting the girder's underside condition during panoramic reconstruction. To enhance defect region localization accuracy, the study improves the positional encoding by incorporating parameter K, resulting in the new encoding system $B_n - S_i - SB_k (k \leq K)$. During block image refinement, each sub-block is explicitly encoded with positional information, allowing engineers to precisely identify defect locations.

5.2. Fine-grained defect analysis approach utilizing regional information reshaping

The lightweight reconstruction of panoramic images serves as a qualitative analysis tool at a macro level, allowing for quick preliminary screening of structural defect areas. However, deep feature extraction and compression in the neural network's representation of block images can lead to the loss of crucial details. This loss restricts the model's capacity to deliver precise quantitative data on defects, including their number, type, and location. Retaining detailed features to enrich quantitative information, can result in higher resource consumption and diminished information efficiency. To overcome the challenge of integrating global and local information effectively, this study introduces an innovative strategy. The essence of this strategy lies in utilizing macro-level global information to guide the tracing and localization of defect areas within lightweight panoramic images. A preliminary screening of the reconstructed lightweight panoramic images is performed based on the defect representation outlined in Section 5.1. An efficient approach is employed to individually analyze all areas within the panoramic images, specifically targeting the characteristic responses of potential defect regions (where the digital code is 1 or dimension > 2). To enhance processing speed and minimize data redundancy at this stage, the system records only the position codes of defect areas, excluding related information from normal areas. The next step involves accurately mapping the position codes of defects from the lightweight panoramic image back to their corresponding block image representation in the original panoramic image. The mapped areas are then finely segmented according to predefined fine-tuning parameters, extracting the corresponding sub-block images. This process ensures that each defect area is displayed in high resolution, enabling detailed subsequent analysis and processing. Finally, the sub-block images undergo detailed analysis using the detection tools.

The fine detection stage demands stringent accuracy, requiring the model to precisely determine the number, type, and bounding boxes of defects. Deep learning models use multi-layer neural networks to efficiently extract feature information. However, during forward propagation, some original information may be lost, leading to discrepancies between the model's predictions and actual observations (Tishby and Zaslavsky, 2015). To address this challenge, this study utilizes YOLOv9, an advanced object detection framework known for its precision and efficiency (Wang et al., 2024). YOLOv9 integrates two key components: Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN). PGI improves data retention through a reversible architecture that maintains input data integrity, optimizing model performance. During training, the PGI incorporates an auxiliary branch to support gradient flow, enhancing feature extraction capabilities without adding extra inference computational burden. GELAN, a lightweight architecture that blends elements from CSPNet and ELAN, enhances accuracy while reducing resource consumption. This combination of advanced gradient protection and efficient feature aggregation makes YOLOv9 an ideal choice for detecting defects with high accuracy in this study.

5.3. Implementation details

The extensive underside of steel box girders in large-span bridges presents significant challenges in collecting defect data, limiting the effectiveness of automated analysis models due to data scarcity. The study employs the diffusion model from Section 4 to simulate realistic defect patterns in a virtual environment. Following on-site investigations, spalling and rust were identified as the primary defect types for detection. A targeted dataset was established to enhance recognition accuracy for this specific bridge scenario. The database will be a foundation for further expansion and application in other infrastructure studies. Through manual selection of defect areas and the application of image enhancement techniques, 3000 training images of steel structure surfaces were generated. These images vary in size, ranging from 15×21 pixels to 2290×1678 pixels. To optimize convergence speed during training, the AdamW optimizer was utilized with an initial learning rate of 1×10^{-4}. The network was trained over 3000 epochs with a batch size of 8. The model was trained using the PyTorch framework on an Ubuntu 20.04 OS. The environment requirements included Python 3.9.7, PyTorch 2.0.0, and CUDA 11.8. A GeForce RTX 4090 GPU with 24 GB of memory was used for this study. Post-training, the image generation tool can flexibly produce defect pattern images. By incorporating fusion techniques, the digitally generated damage is seamlessly integrated into real-world scenarios, supplying essential training data for defect detection models, as illustrated in Fig. 7. A total of 1000 synthetic images were subjected to both subjective and objective evaluations. First, 10 field inspection technicians were invited to visually assess the images, unanimously finding them indistinguishable from real data. Additionally, the Frechet Inception Distance (FID) metric (Heusel et al., 2017) was employed for quantitative analysis, yielding a score of 20.5, indicating a high degree

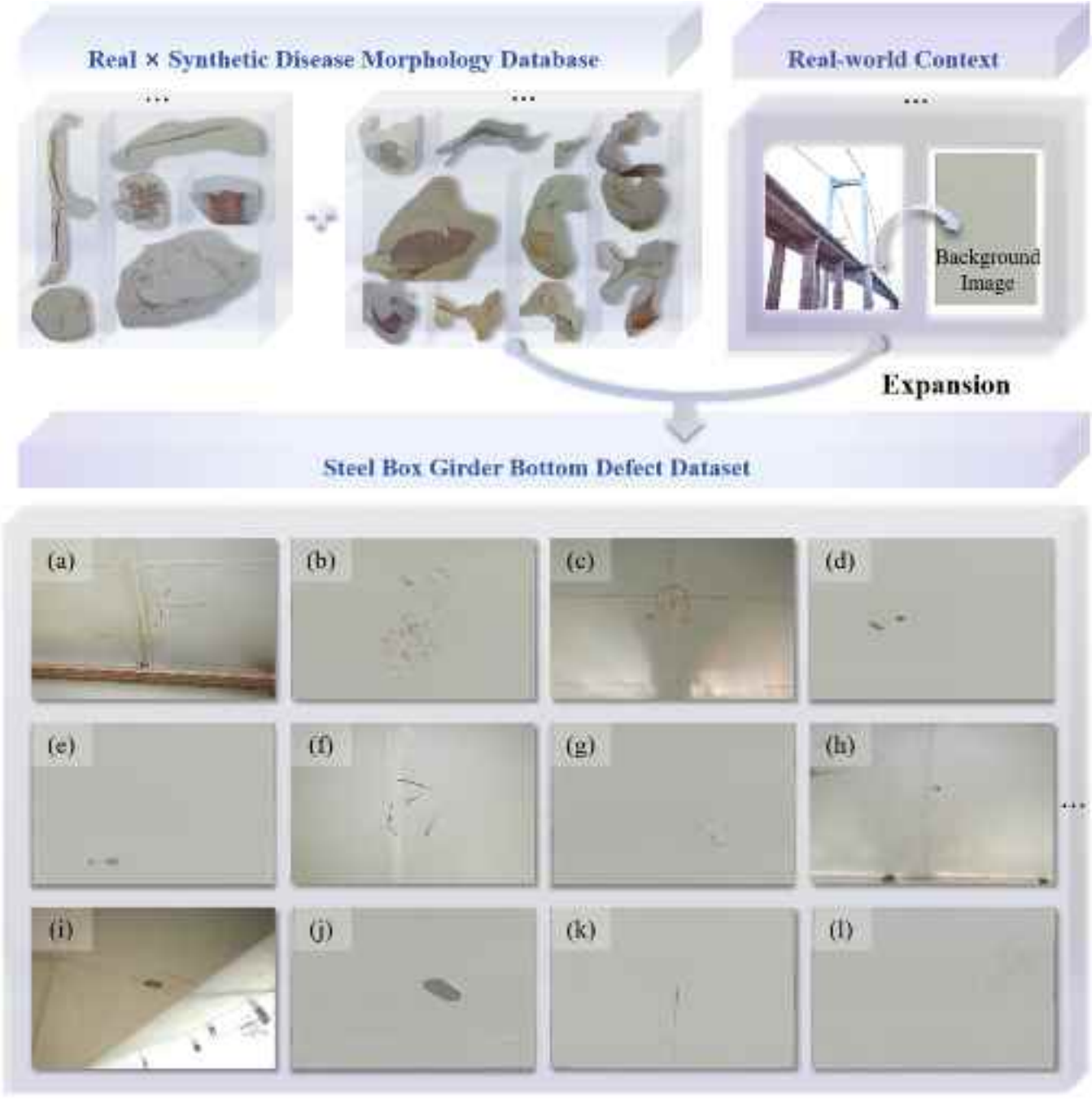


Fig. 7. Process of model data creation (a, c, f, h, i, k represent the collected real defect data, while the rest are synthetic data).

of similarity between the synthetic and real datasets. Utilizing this method, a training dataset comprising 4449 damage-type images was constructed for the Stage One feature representation model, including 2500 generated images. Additionally, 1722 images were used for normal background training data. The test set includes 1000 damage-type images, with 500 generated images, and 670 images of normal backgrounds. The feature representation model's dataset was meticulously labeled using Labelme software, identifying spalling and rust as the two primary defect types. Ultimately, a dataset of 5649 images was constructed for the Stage Two defect detection model, with a training-to-test set ratio of 8:2.

The dataset from Phase One was used to train the feature representation model, employing the same hardware parameters as the defect simulation tool to ensure consistent and comparable training processes.

The neural network was trained using the AdamW optimizer, with initial parameters set to a weight decay of 5×10^{-5} and a learning rate of 1×10^{-3}. The training schedule consisted of 100 epochs, with a batch size of 8 for each iteration. Top-1 accuracy was used as the primary metric to evaluate the performance of the proposed classification model. This metric measures the percentage of instances where the model's top prediction matches the true class label, offering a clear and intuitive assessment of the model's accuracy.

To evaluate the performance of the feature representation model, comparative experiments were conducted with five deep learning models, as shown in Table 1. Considering the parameter count, FLOPs, and Top-1 accuracy in Table 1, MobileNetV4s demonstrates an excellent balance between computational efficiency and classification accuracy. MobileNetV4s achieves a Top-1 accuracy of 90.2 %, outperforming

Table 1
Performance comparison of different models in two phases.

Feature Representation			
Network	#Params (M)	FLOPs (G)	Top1-accuracy(%)
Resnet−18	11.8	30.3	85.1
Yolov9-c-backbone	8.99	42.8	84.3
Yolov9-m-backbone	7.98	32.5	85.7
Yolov9-s-backbone	2.99	11.1	87.0
Yolov9-t-backbone	0.84	2.9	84.9
Mobilenetv4s	**2.50**	**4**	**90.2**
Defect Detecion			
Network	#Params (M)	FLOPs (G)	mAP−50(%)
Yolov9-c	25.23	101.8	92.0
YolovMobile-c	18.04	61.3	86.5
Yolov9-m	19.92	76.0	88.9
YolovMobile-m	13.87	46.9	83.2
Yolov9-s	7.08	26.2	87.4
YolovMobile-s	5.80	18.6	81.6
Yolov9-t	1.88	7.1	79.5
YolovMobile-t	2.54	7.5	76.9

Notes: (1) #Param represents the number of model parameters, measured in **M**, where $\mathbf{1M} = \boldsymbol{10^6}$; FLOPs indicate the computational complexity of the model, measured in G, $\mathbf{1G} = \boldsymbol{10^9}$.

ResNet-18's 85.1 % accuracy while using only 13.2 % of the FLOPs and 21.2 % of the parameter count. When compared to the YOLOv9 series, MobileNetV4s offers a compelling case, with YOLOv9-c-backbone and YOLOv9-m-backbone showing higher computational costs but lower accuracy. Even YOLOv9-s-backbone, the most accurate model in the YOLOv9 series, has a 3.2 % lower Top-1 accuracy compared to MobileNetV4s, which also provides superior efficiency metrics.

The defect detection model was trained using the Phase Two dataset, with hardware parameters consistent with those used in training the feature representation model. The network was trained using the AdamW optimizer, with initial parameters set to a weight decay of 5×10^{-4} and a learning rate of 1×10^{-3}. The model was trained for 200 epochs, with a batch size of 8 for each iteration to optimize the learning process. The model's performance was evaluated using mean average precision at an IoU threshold of 0.5 (mAP-50), which provides a comprehensive measure by integrating both precision and recall.

The YOLOv9 series models were comprehensively trained and tested to assess their performance in real-world detection scenarios. Earlier in this research, MobileNetV4 showed an optimal balance between computational efficiency and detection performance, leading to its integration as the backbone of the YOLOv9 series. During implementation, MobileNetV4 was integrated as the backbone of the YOLOv9 network, replacing the prior architecture. Comparative experiments in Table 1 show that the YOLOv9-c model achieves superior recognition performance. However, as model complexity decreases, so does performance. Although the modified models optimize efficiency metrics, this comes at the cost of reduced overall performance. The high efficiency of the YOLOv9-t backbone likely explains the observed reduction in performance metrics after modifications. The modified models show about a 4 % drop in performance compared to their original versions. Since this phase of the study prioritizes precise and fine-grained defect analysis, high-precision recognition is valued over efficiency. Therefore, the YOLOv9-c model will be used for detailed defect detection in future projects.

The model inference results, shown in Fig. 8, clearly demonstrate the recognition capabilities of these models. The larger YOLOv9 models effectively identify and classify complex patterns, achieving accuracy that closely matches ground truth values. In contrast, the smaller YOLOv9 models struggle with misrecognition in some complex scenarios. In intricate background settings and varied object features, the YOLOv9-c model's predictions are precise, offering detail that rivals or exceeds manual annotations. This underscores the model's robustness and practical applicability in real-world tasks.

6. Filed validation experiments

The Nansha Bridge project includes seven approach bridges, three interchanges, and two major-span suspension bridges, each over a kilometer long. This study concentrates on the Dasha Channel Bridge, a double-tower, single-span suspension bridge with a span of 360 + 1200 + 480 m. The main span includes a steel box girder with a total width of 49.7 m. Longitudinally, segments 3 through 84, each 12.8 m long, were selected for testing, covering a total length of 1049.6 m. Transversely, plate 5–11 were chosen for inspection. The general layout of the bridge is shown in Fig. 9.

The layout of the PM in the dynamic visual perception system is carefully designed to ensure functionality and long-term stability on the inspection platform, as shown in Fig. 9. The PM is mounted externally on the platform using custom peripheral brackets. This setup preserves the workspace of the maintenance platform, facilitating work on the underside of the steel box girders. To account for extreme weather like typhoons and long-term vibrations, the peripheral brackets are designed with specific stabilizing features. High-strength bolts secure the brackets to the inspection vehicle. The PM is precisely positioned directly below the center of the plate it monitors. With a field of view covering approximately 3.5 m, each module fully meets the detection requirements of its assigned plates, as shown in Fig. 10. A dual-module configuration was adopted for the computing control module: one controls three cameras, and the other controls four. This configuration evaluates the stability of the CTM in multitasking scenarios. Before activating the MM, collection instructions are synchronously sent to each perception unit through the CTM's software interface, ensuring activation of the PM. After data collection, the control software securely transmits the data from the PM to the central control system via wireless transmission. The data is then transmitted over the intranet to a remote server, where it undergoes in-depth analysis and processing using the server's computing resources.

Following the block-based panoramic generation strategy from Section 3, the discrete data collected from the girder bottom were meticulously integrated. Each plate within the standard girder segment measures 3 *m* in width and 12.8 *m* in length. Through advanced stitching techniques and homography matrix correction, the output image resolution was achieved at 3000 × 12,800 pixels, with each pixel representing a physical size of 1 *mm*. To validate the accuracy of the block images, the research team designed a series of precise experimental steps before data collection. Red rectangular labels of various sizes were manually created, with dimensions carefully measured and recorded to serve as reference standards for subsequent verification. These labels were accurately placed at pre-determined inspection points, with their positions precisely measured beforehand. Upon completing image collection, the research team analyzed the block images, comparing the red rectangular labels within the images to the previously recorded size and location data. The results indicated that the system's dimensional measurement error was within 2*mm*, with a standard deviation of 0.93. The positioning calculation error was within 3*mm*, with a standard deviation of 1.57. These findings confirm the reliability and effectiveness of the adopted image collection and processing methods in practical applications. The detailed comparison results are presented in Fig. 12.

In the panoramic generation phase, 574 independent block images were produced, each corresponding to a specific section of the beam and collectively covering a length of 1049.6 *m*. Traditional localization methods, which depend on precise coordinate information, often fail to enable engineers to navigate to the areas indicated by the coordinates in real-world applications. To overcome this limitation, a block-based localization system was implemented. This system shifts from relying on coordinate data to using position encoding for defect representation. Block images can be easily combined according to established encoding rules to reconstruct a complete panoramic view of the beam bottom, as

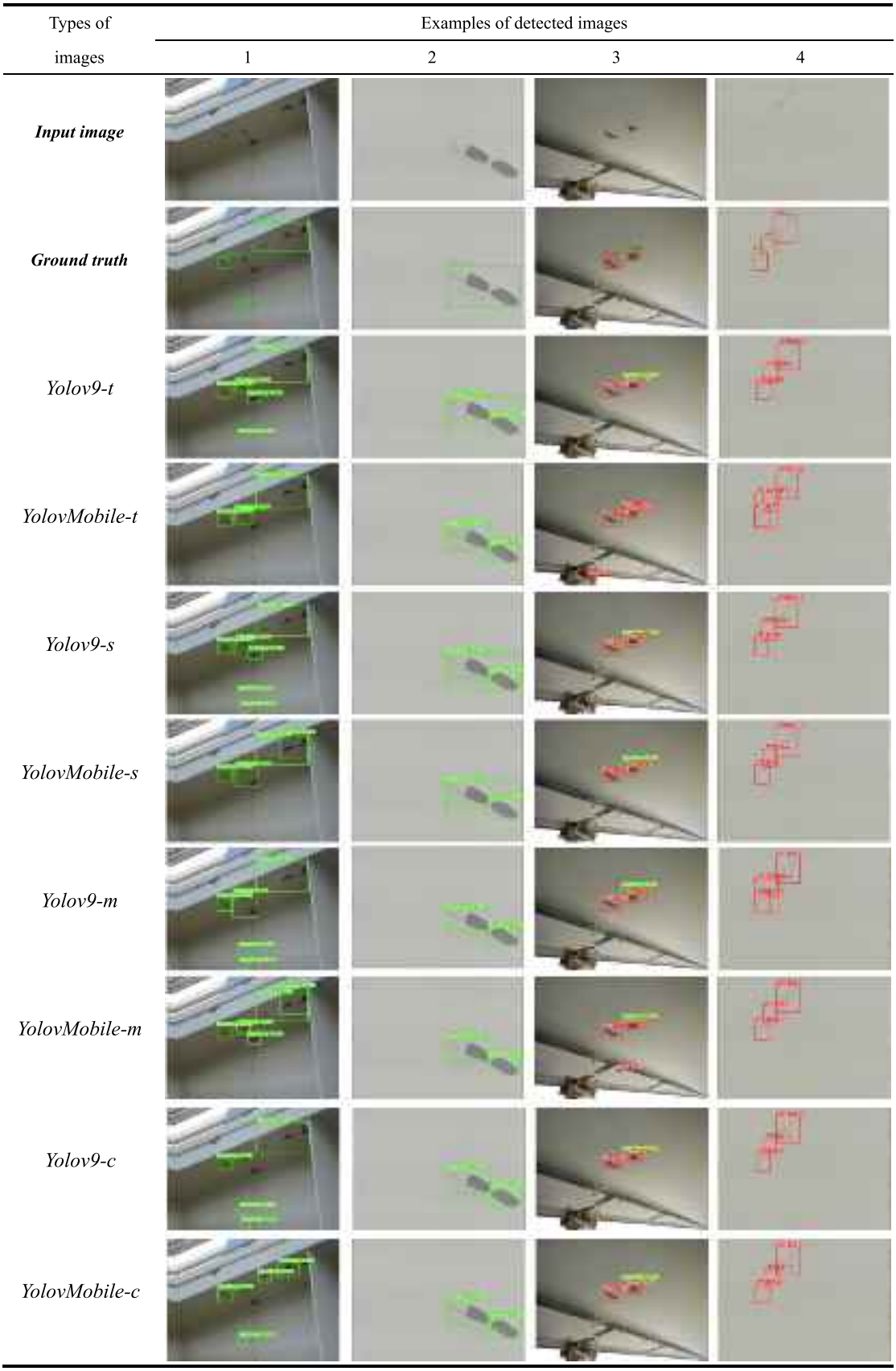


Fig. 8. Visualized results of different networks.

Fig. 9. Overview of bridge specifications and equipment setup.

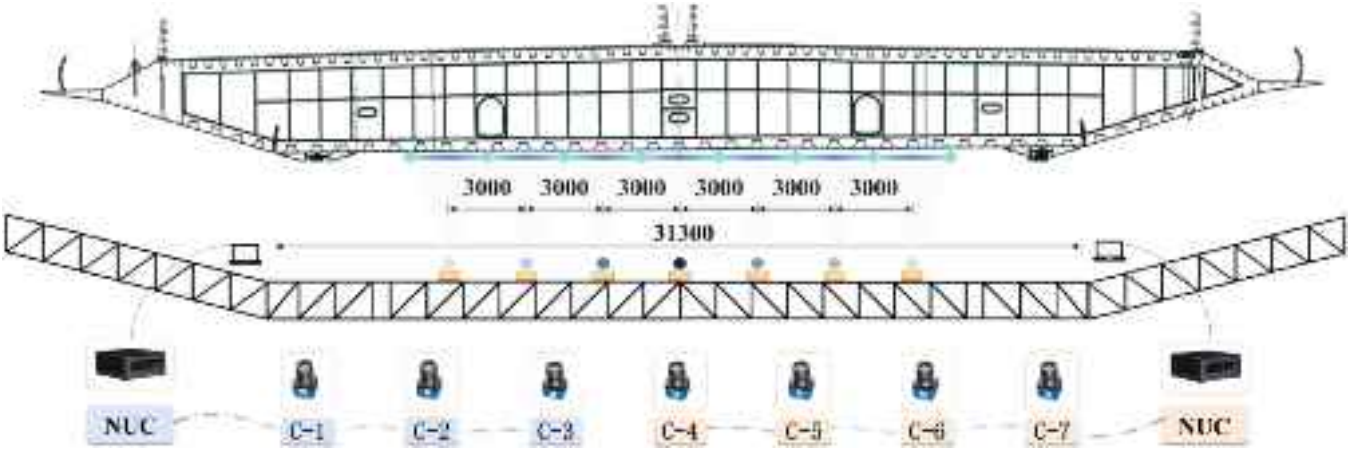


Fig. 10. Detailed layout of the dynamic visual perception system components.

shown in Fig. 11. Fig. 13 provides detailed views of specific segments within this panoramic image. However, when the panoramic image undergoes pixel analysis, it consists of $21000 \times 12800 \times 82 = 2.2 \times 10^{10}$ pixels, creating an image at the scale of tens of billions of pixels. Large-scale panoramic images present challenges related to information overload in practical applications, rendering the straightforward

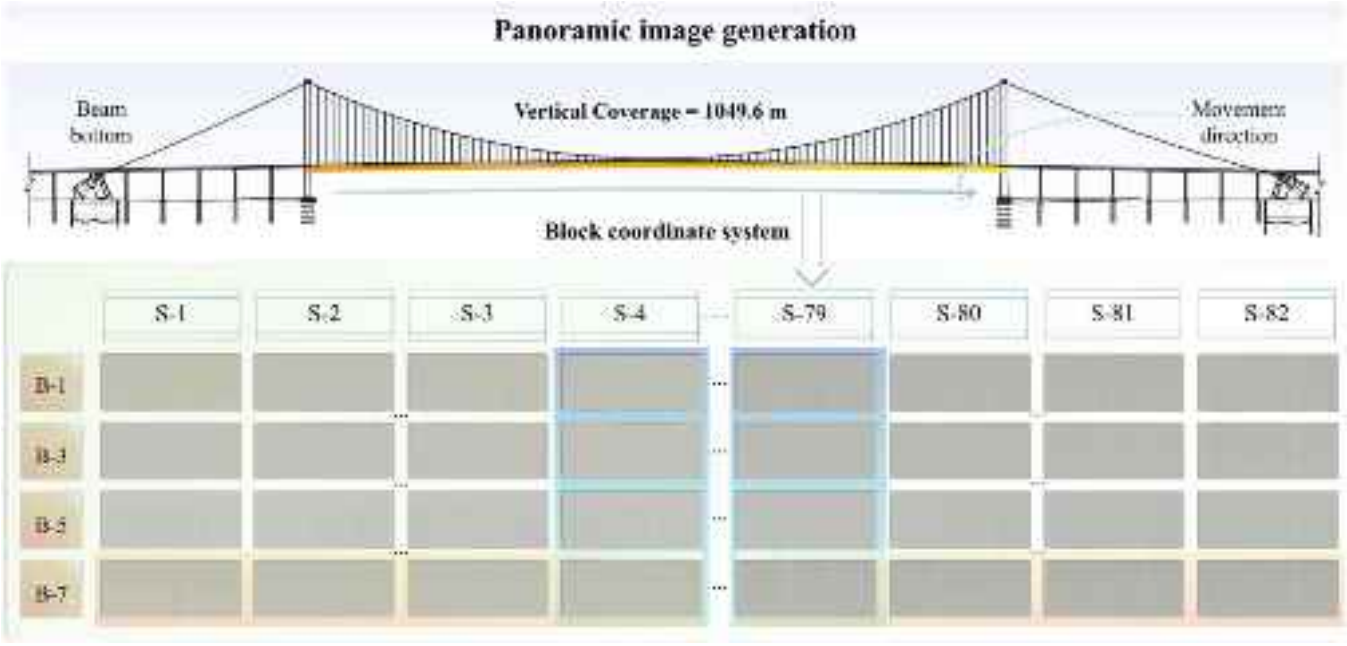


Fig. 11. Application of dynamic visual perception system for panoramic imaging of beam bottoms.

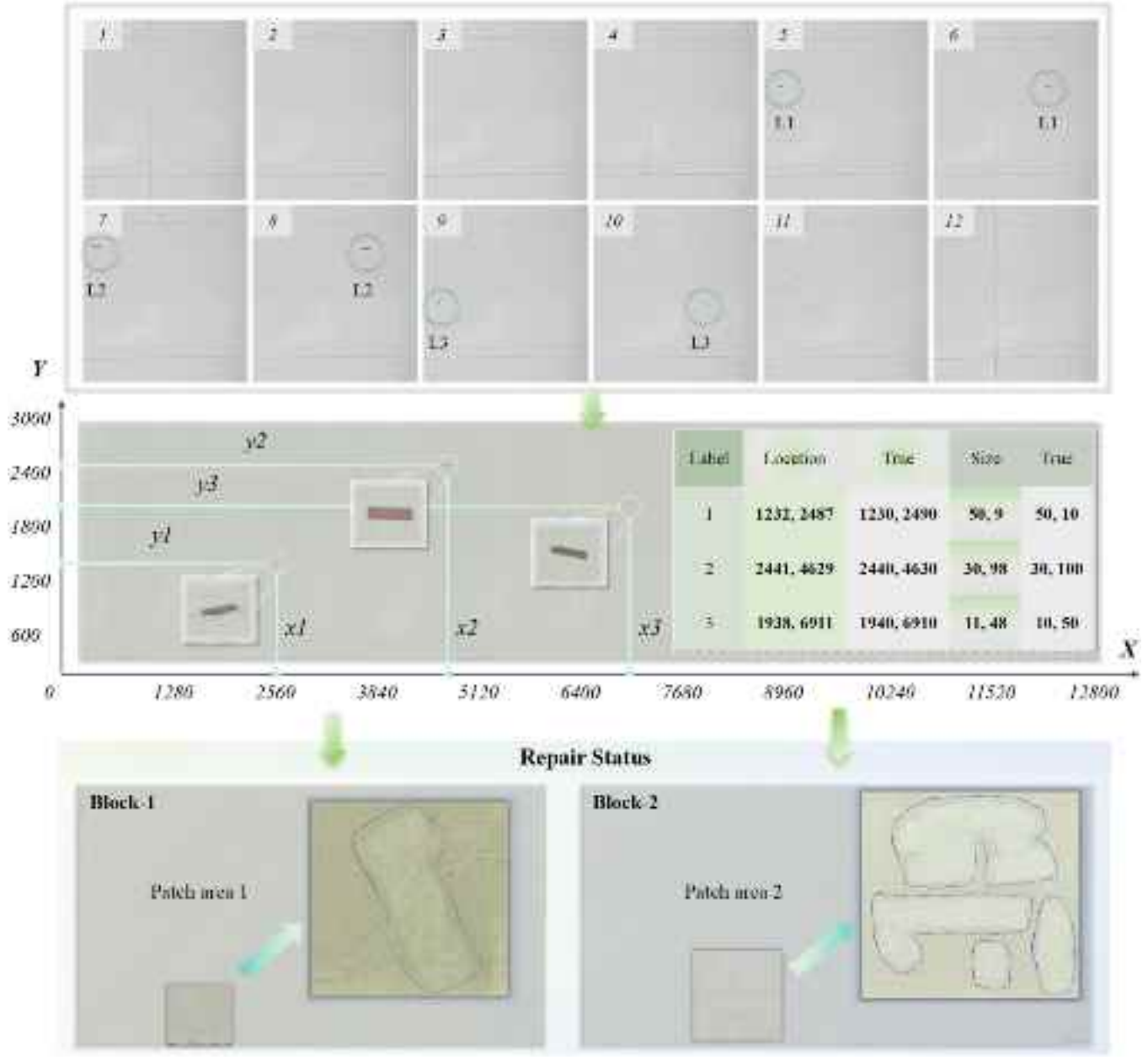
combination of block images into a panoramic view impractical. This study employs the two-phase synchronized defect identification and localization method outlined in Section 5. It divides the problem into global and local levels. On the global level, the focus is on capturing primary features, thus avoiding the complexity of displaying every detail across the entire image. The subsequent local level zooms in on critical defect areas, utilizing high-precision processing techniques to analyze these details. This approach allows for the efficient presentation of both global and local information within the panoramic view, enhancing the overall utility of the system.

The objective of the initial stage is to describe the beam bottom's service condition, focusing on accurately identifying key defects on a global scale. The feature representation method detailed in Section 5.1 was utilized to conduct an in-depth analysis of 574 block images. Due to the significant longitudinal dimensions of the panels, which increase processing complexity, a fine-grained parameter $K = 6$ was introduced. This parameter not only simplifies the processing but also ensures a thorough capture of defect distribution characteristics. Following lightweight reconstruction, the pixel count of the panoramic image was reduced to $82 \times 7 \times 6 = 3444$, representing a 99.99 % reduction in size compared to the image-based representation, thereby significantly improving processing efficiency. However, factors such as the bridge's age and the timeliness of steel box girder maintenance pose limitations when using field-collected data to

validate the proposed method. Fig. 12 highlights the maintenance conditions of defects at the bottom of the steel box girder. Therefore, simulated defect rehearsal experiments were designed using a synthetic database described in Section 4, which includes both real and synthetic defect morphologies. A complete beam bottom dataset was first created from finely processed real scene data. The dataset consists of 7 transverse target areas, each divided into 492 sub-blocks, yielding a total of 3444 sub-block images. To simulate the random distribution of real defects, numerical codes $\mathcal{N}(\mathcal{N} \leq 492)$ for 40 defect regions were randomly generated within each sub-region, corresponding to position codes in the block coordinates. Defect morphology data were then randomly extracted from the synthetic database 280 times, ensuring no duplication in subsequent operations. These extracted defect data were integrated into the corresponding sub-block images. The entire process employed an end-to-end model, with 3444 sub-block images input and 3444 sub-block images containing defect regions output, maintaining a "black-box" approach.

The sub-block images were then analyzed using the first stage's feature representation tool, resulting in the identification of 275 defect regions, while 5 regions remained undetected. Further analysis revealed that these unrecognized regions involved defects that were small in scale and lacked distinct morphology. In the future, additional data on small targets will be incorporated into the existing dataset. Simultaneously, the network architecture will be enhanced by integrating global self-attention mechanisms and shallow feature information fusion, aimed at improving the defect feature representation tool's ability to recognize small targets with greater precision. A lightweight panoramic image was generated based on the output defect feature representation (including the 5 manually added unrecognized regions) and position codes, as shown in Fig. 14. In the block coordinate system, each sub-block was added along the block coordinate axis rather than being expanded along the longitudinal segment axis, forming a sub-block coordinate axis. The transformation effectively addressed the issue of information diffusion caused by an excessive aspect ratio and demonstrated the lightweight panoramic image's effectiveness in quickly identifying and localizing defect regions.

The second stage focuses on conducting a detailed analysis of defect areas to accurately assess their progression. Initially, the position codes of the defect areas are automatically extracted, and feature representation is mapped back to the original sub-block images. Detailed identification of defect types, quantities, and locations was then performed on all sub-block images using the detection tools outlined in Section 5.2, with specific results presented in Fig. 15. The first column displays the position codes of the sub-block images, while the second column shows the defect detection results for the corresponding areas, confirming the accurate identification of spalling and corrosion defects. The third column provides detailed coordinate information of the defects within the block images. To enhance the precision of defect localization, a coordinate system was established with the upper-left corner of each block image as the origin. By integrating the local positioning coordinates in the sub-blocks with the S encoding information, these coordinates were converted into global block coordinates. The study aimed to approximate the impact area of defects by calculating the area covered by the target detection boxes. However, this approach may lead to significant errors when addressing defects with elongated or irregular shapes. In computer vision tasks related to object detection, heatmap techniques



Label	Location	True	Size	True
1	1232, 2487	1230, 2490	50, 9	50, 10
2	2441, 4629	2440, 4630	30, 98	30, 100
3	1938, 6911	1940, 6910	11, 48	10, 50

Fig. 12. Detailed information display of block images.

are often employed to focus on image regions where targets are likely to be found. Drawing on this approach, heatmap techniques were applied to estimate the impact area of defects, yielding results that more accurately reflect the defect morphology. The final column of the figure visualizes the results of defect impact area estimation using heatmaps.

7. Conclusion

The challenges of inspecting the underside of large-span bridges are significant. An integrated hardware-software system that combines advanced equipment with intelligent algorithms has been developed and implemented. By leveraging automated data acquisition and a two-phase defect analysis process, the system streamlines the inspection process, achieving over 90 % accuracy in defect identification while providing millimeter-level localization precision. This, in turn, enables engineers to develop precise, targeted maintenance strategies, thereby enhancing overall bridge health management. The primary achievements of this paper are as follows:

(1) Dynamic visual perception system: It integrates perception, control-transmission, and mobility modules to facilitate flexible bridge inspections. Distributed cameras and wireless communication technologies create a novel data acquisition network, while edge computing enhances data transmission and processing efficiency. The inclusion of a controllable mobility module overcomes the limitations of traditional static inspection methods, broadening detection coverage and improving image data quality. It is successfully applied to the routine inspection of kilometer-scale bridges.
(2) Panoramic image generation and optimization: Introducing the block concept, simplifies the process of stitching panoramic images and enhances image processing accuracy. The method's spatially ordered segmentation and extraction of feature points

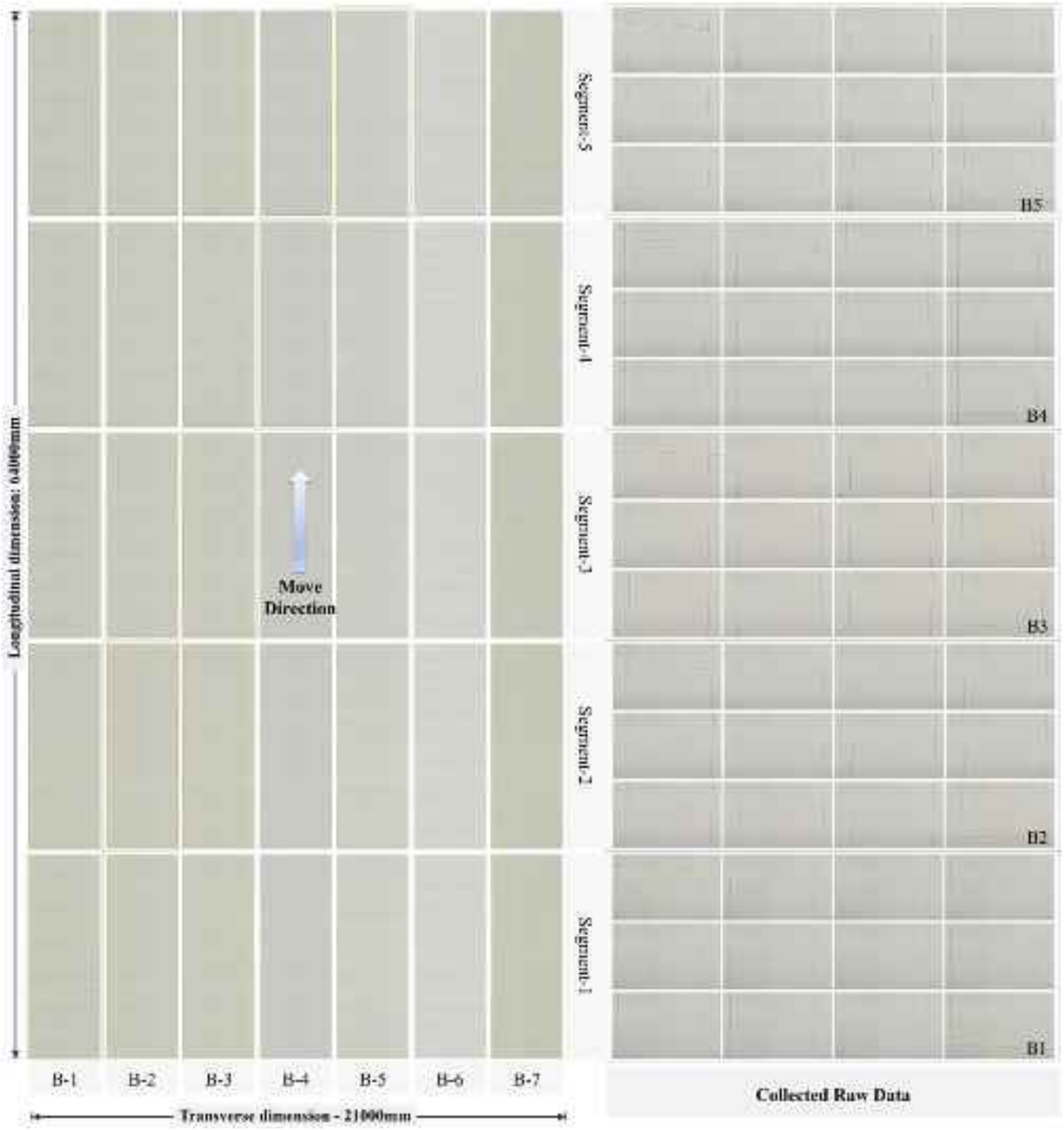


Fig. 13. Expanded view of partial panoramic images.

from overlapping regions reduce processing complexity and optimize image alignment.

(3) Deep learning-driven two-phase identification and localization: In the first phase, the method rapidly filters key features using the MobilenetV4 architecture, complemented by diffusion models that simulate and augment defect scenarios. Its recognition accuracy reaches 90.2 %. Lightweight reconstruction and fine-grained feature adjustments optimize the identification and localization of defect in panoramic images. In the second phase, the YOLOv9 detection framework provides detailed analysis of identified defect areas, achieving a mAP-50 of 92 %.

The proposed system demonstrates exceptional transferability across diverse infrastructure components, owing to its modular architecture and adaptable design framework. In tunnel or pipeline applications, the system's hardware configuration can be optimized through the integration of dimensionally compatible mobile units, addressing data acquisition challenges. Furthermore, the system's capability can be further improved by integrating diverse sensor technologies, including thermal imaging, radar, temperature, humidity, and atmospheric sensors. The system's data processing algorithms should be tailored to analyze infrastructure-specific patterns. To enhance the method's applicability, future research will focus on: 1. Large-Scale Defect

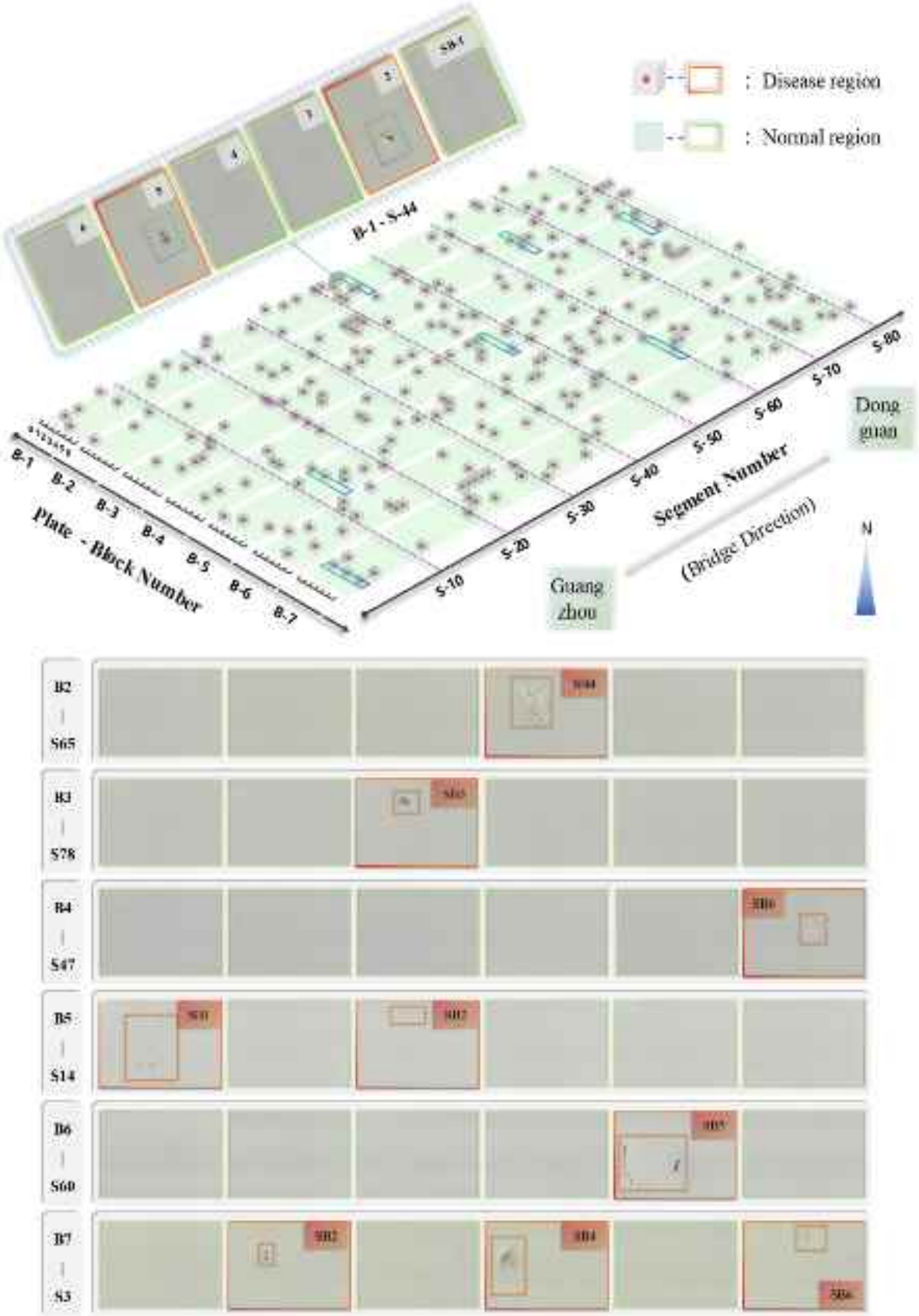


Fig. 14. Lightweight reconstructed panoramic image.

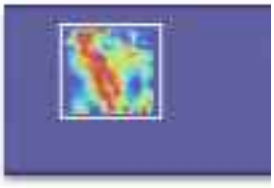
Diseased region	*Detection Result*	*Location/mm*	*Area*
B_1 - SB_2 - S_{44}		**Spalling** (2101, 2910) **Spalling** (2124, 2805) **Rust** (1888, 3281) **Rust** (2268, 2771)	
B_1 - SB_5 - S_{44}		**Spalling** (883, 10107)	
B_2 - SB_4 - S_{65}		**Rust** (1145, 7220)	
B_3 - SB_3 - S_{78}		**Spalling** (1182, 4818)	
B_4 - SB_6 - S_{47}		**Spalling** (1644, 11902) **Spalling** (1622, 11662) **Spalling** (1877, 11582) **Spalling** (1754, 11428)	
B_5 - SB_1 - S_{14}		**Spalling** (1004, 1610) **Spalling** (1426, 885) **Spalling** (1422, 1609)	
B_6 - SB_5 - S_{60}		**Rust** (252, 9758) **Rust** (410, 10256) **Rust** (1578, 9798)	
B_7 - SB_4 - S_3		**Spalling** (778, 7627) **Spalling** (695, 7546) **Spalling** (811, 7912) **Spalling** (823, 7773) **Spalling** (596, 7510) **Spalling** (511, 7298)	

Fig. 15. Presentation of defect analysis results.

Database: Collaborating with infrastructure maintenance companies to acquire diverse defect data and build a comprehensive database using generative models for training and research. 2. Real-Time Data Processing: Designing a lightweight detection model with reduced complexity through techniques like pruning and distillation to enable fast anomaly detection and timely maintenance. 3. Predictive Maintenance System: Integrating multi-source data to develop predictive models that assess the likelihood of damage, combining historical and real-time data.

CRediT authorship contribution statement

Chen Wang: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Yuan Binhong:** Writing – review & editing, Validation, Conceptualization. **Chen Dongliang:** Writing – review & editing, Validation. **Hu Yong:** Writing – review & editing, Validation. **Wang Feiyu:** Writing – review & editing, Software. **zhang jian:** Methodology, Funding acquisition, Conceptualization.

Funding

The research presented was financially supported by the Key R&D Program of Jiangsu (No.: BE2020094) and National Key Research and Development Program of China (No.: 2022YFC3801700).

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.

Data availability

Data will be made available on request.

References

Bolourian, N., Hammad, A., 2020. LiDAR-equipped UAV path planning considering potential locations of defects for bridge inspection. Autom. Constr. 117, 103250. https://doi.org/10.1016/j.autcon.2020.103250.

Brown, M., Lowe, D.G., 2007. Automatic panoramic image stitching using invariant features. Int. J. Comput. Vis. 74, 59–73. https://doi.org/10.1007/s11263-006-0002-3.

Capponi, L., Tocci, T., D'Imperio, M., Abidi, S.H.J., Scaccia, M., Cannella, F., Marsili, R., Rossi, G., 2021. Thermoelasticity and ArUco marker-based model validation of polymer structure: application to the San Giorgio's bridge inspection robot. Acta IMEKO 10 (4). https://doi.org/10.21014/acta_imeko.v10i4.1148.

Cha, Y.J., Choi, W., Büyüköztürk, O., 2017. Deep learning-based crack damage detection using convolutional neural networks. Comput. Aided Civ. Infrastruct. Eng. 32, 361–378. https://doi.org/10.1111/mice.12263.

Chen, W., He, Z., Zhang, J., 2023. Online monitoring of crack dynamic development using attention-based deep networks. Autom. Constr. 154, 105022. https://doi.org/10.1016/j.autcon.2023.105022.

Cord, A., Chambon, S., 2012. Automatic road defect detection by textural pattern recognition based on adaBoost. Comput. Aided Civ. Infrastruct. Eng. 27, 244–259. https://doi.org/10.1111/j.1467-8667.2011.00736.x.

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv: 2010.11929, https://doi.org/10.48550/arXiv.2010.11929. 2020 (accessed 15 June 2042).

Evan, M., Nicholas, C., Sriram, N., 2020. Automated defect quantification in concrete bridges using robotics and deep learning. J. Comput. Civ. Eng. 34 (5). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000915.

He, Z., Chen, W., Zhang, J., Wang, Y.H., 2024a. Crack segmentation on steel structures using boundary guidance model. Autom. Constr. 162, 105354. https://doi.org/10.1016/j.autcon.2024.105354.

He Z., Wang Y.H., Zhang J., Generative Structural Design Integrating BIM and Diffusion Model, arXiv:2311.04052, https://doi.org/10.48550/arXiv.2311.04052. 2023 (accessed 20 May 2024b).

Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S., GANs trained by a two time-scale update rule converge to a local nash equilibrium, 2017 International Conference on Neural Information Processing Systems (NIPS), ACM, Long Beach, CA, USA (2017), pp. 6629-6640, https://dl.acm.org/doi/10.5555/3295222.3295408..

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851. https://doi.org/10.5555/3495724.3496298.

Humpe, A., 2020. Bridge inspection with an off-the-shelf 360° camera drone. Drones 4 (4), 67. https://doi.org/10.3390/drones4040067.

Ikeda, T., Minamiyama, S., Yasui, S., Ohara, K., Ichikawa, A., Ashizawa, S., Okino, A., Oomichi, T., Fukuda, T., 2019. Stable camera position control of unmanned aerial vehicle with three-degree-of-freedom manipulator for visual test of bridge inspection. J. Field Robot. 36 (7), 1212–1221. https://doi.org/10.1002/rob.21899.

Jha, S.B., Babiceanu, R.F., 2023. Deep CNN-based visual defect detection: survey of current literature. Comput. Ind. 148, 103911. https://doi.org/10.1016/j.compind.2023.103911.

Jiang, S., Cheng, Y., Zhang, J., 2023. Vision-guided unmanned aerial system for rapid multiple-type damage detection and localization. Struct. Health Monit. 22 (1), 319–337. https://doi.org/10.1177/14759217221084878.

Jiang, W., Liu, M., Peng, Y., Wu, L., Wang, Y., 2021. HDCB-net: a neural network with the hybrid dilated convolution for pixel-level crack detection on concrete bridges. IEEE Trans. Ind. Inform. 17 (8), 5485–5494. https://doi.org/10.1109/TII.2020.3033170.

Kang, D.H., Cha, Y.J., 2022. Efficient attention-based deep encoder and decoder for automatic crack segmentation. Struct. Health Monit. 21 (5), 2190–2205. https://doi.org/10.1177/14759217211053776.

Kingma D.P., Welling M., Auto-encoding variational bayes, arXiv:1312.6114, https://doi.org/10.48550/arXiv.1312.6114. 2022 (accessed 15 May 2024).

Krichen, M., 2023. Generative adversarial networks (Delhi, India). 2023 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT) 1–7. https://doi.org/10.1109/ICCCNT56998.2023.10306417.

Kung, R.Y., Pan, N.H., Wang, C.C.N., Lee, P.C., 2021. Application of deep learning and unmanned aerial vehicle on building maintenance. Adv. Civ. Eng. 1, 5598690. https://doi.org/10.1155/2021/5598690.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., 2022. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. Neural Netw. Learn. Syst. 33 (12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827.

Lin, J.J., Ibrahim, A., Sarwade, S., Golparvar-Fard, M., 2021. Bridge inspection with aerial robots: automating the entire pipeline of visual data capture, 3D mapping, defect detection, analysis, and reporting. J. Comput. Civ. Eng. 35 (2). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000954.

Lin, T.H., Putranto, A., Chen, P.H., Teng, Y.Z., Chen, L., 2023. High-mobility inchworm climbing robot for steel bridge inspection. Autom. Constr. 152, 104905. https://doi.org/10.1016/j.autcon.2023.104905.

Liu, C.Y., Chou, J.S., 2023. Bayesian-optimized deep learning model to segment deterioration patterns underneath bridge decks photographed by unmanned aerial vehicle. Autom. Constr. 146, 104666. https://doi.org/10.1016/j.autcon.2022.104666.

Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 9992-10002, https://doi.org/10.1109/ICCV48922.2021.00986..

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11966-11976, https://doi.org/10.1109/CVPR52688.2022.01167..

Lyu, G., Wang, P., Li, G., Lu, F., Dai, S., 2024. A heavy-load wall-climbing robot for bridge concrete structures inspection. Ind. Robot 51 (3), 465–478. https://doi.org/10.1108/IR-11-2023-0273.

Meng, Q., Yang, J., Zhang, Y., Yang, Y., Song, J., Wang, J., 2023. A robot system for rapid and intelligent bridge damage inspection based on deep-learning algorithms. J. Perform. Constr. Facil. 37 (6). https://doi.org/10.1061/JPCFEV.CFENG-4433.

Peng, X., Zhong, X., Zhao, C., Chen, A., Zhang, T., 2021. A UAV-based machine vision method for bridge crack recognition and width quantification through hybrid feature learning. Constr. Build. Mater. 299, 123896. https://doi.org/10.1016/j.conbuildmat.2021.123896.

Pérez, P., Gangnet, M., Blake, A., 2003. Poisson image editing. ACM Trans. Graph. 22 (3), 313–318. https://doi.org/10.1145/882262.882269.

Rao, A.S., Nguyen, T., Palaniswami, M., Ngo, T., 2021. Vision-based automated crack detection using convolutional neural networks for condition assessment of infrastructure. Struct. Health Monit. 20 (4), 2124–2142. https://doi.org/10.1177/1475921720965445.

Rubio, J.J., Kashiwa, T., Laiteerapong, T., Deng, W., Nagai, K., Escalera, S., Nakayama, K., Matsuo, Y., Prendinger, H., 2019. Multi-class structural damage segmentation using fully convolutional networks. Comput. Ind. 112, 103121. https://doi.org/10.1016/j.compind.2019.08.002.

Sanchez-Cuevas P.J., Heredia G., Ollero A., Multirotor UAS for bridge inspection by contact using the ceiling effect, In 2017 International Conference on Unmanned Aircraft Systems (ICUAS), Miami, FL, USA, 2017, pp. 767-774, https://doi.org/10.1109/ICUAS.2017.7991412..

Sutter, B., Lelevé, A., Pham, M.T., Gouin, O., Jupille, N., Kuhn, M., Lulé, P., Michaud, P., Rémy, P., 2018. A semi-autonomous mobile robot for bridge inspection. Autom. Constr. 91, 111–119. https://doi.org/10.1016/j.autcon.2018.02.013.

Talab, A.M.A., Huang, Z., Xi, F., HaiMing, L., 2016. Detection crack in image using Otsu method and multiple filtering in image processing techniques. Optik 127 (3), 1030–1033. https://doi.org/10.1016/j.ijleo.2015.09.147.

Tishby N., Zaslavsky N., Deep learning and the information bottleneck principle, in: 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 2015, pp. 1-5, https://doi.org/10.1109/ITW.2015.7133169..

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I., Attention is all you need, in 31st International Conference on Neural Information Processing Systems (NIPS), ACM, Red Hook, NY, USA, 2017, pp. 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349..

Wang, H., Li, Y., Dang, L.M., Lee, S., Moon, H., 2021. Pixel-level tunnel crack segmentation using a weakly supervised annotation approach. Comput. Ind. 133, 103545. https://doi.org/10.1016/j.compind.2021.103545.

Wang, Z., Yang, Z., 2020. Review on image-stitching techniques. Multimed. Syst. 26, 413–430. https://doi.org/10.1007/s00530-020-00651-y.

Wang C.Y., Yeh I.H., Liao H.Y.M., YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, arXiv:2402.13616, https://doi.org/10.48550/arXiv.2402.13616. 2024 (accessed 20 July 2024).

Wang, H., Zhai, L., Huang, H., Guan, L., Mu, K., Wang, G., 2020. Measurement for cracks at the bottom of bridges based on tethered creeping unmanned aerial vehicle. Autom. Constr. 119, 103330. https://doi.org/10.1016/j.autcon.2020.103330.

Wang, F., Zou, Y., Zhang, C., Buzzatto, J., Liarokapis, M., Castillo, E.R., Lim, J.B.P., 2023. UAV navigation in large-scale GPS-denied bridge environments using fiducial marker-corrected stereo visual-inertial localisation. Autom. Constr. 156, 105139. https://doi.org/10.1016/j.autcon.2023.105139.

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

Yasuda, Y.D.V., Cappabianco, F.A.M., Martins, L.E.G., Gripp, J.A.B., 2022. Aircraft visual inspection: a systematic literature review. Comput. Ind. 141, 103695. https://doi.org/10.1016/j.compind.2022.103695.

Yoon, S., Lee, S., Kye, S., Kim, I.H., Jung, H.J., B. F. S. Jr, 2022. Seismic fragility analysis of deteriorated bridge structures employing a UAV inspection-based updated digital twin. Struct. Multidiscip. Optim. 65, 346. https://doi.org/10.1007/s00158-022-03445-0.

Zhang, J., Qian, S., Tan, C., 2023. Automated bridge crack detection method based on lightweight vision models. Complex Intell. Syst. 9, 1639–1652. https://doi.org/10.1007/s40747-022-00876-6.

Zhao, N., Zheng, X., 2017. Multi-band blending of aerial images using GPU acceleration. 2017 10th Int. Congr. Image Signal Process Biomed. Eng. Inform. (CISP-BMEI), Shanghai, China 1–5. https://doi.org/10.1109/CISP-BMEI.2017.8302068.