Journal of Civil Structural Health Monitoring (2025) 15:4107–4131
https://doi.org/10.1007/s13349-025-01031-8

ORIGINAL PAPER

Integrating dual-modal camera systems on unmanned aerial vehicles for bridge tower defect detection

Wang Chen[1,2] · Jian Zhang[1,2]

Received: 19 February 2025 / Revised: 19 August 2025 / Accepted: 1 October 2025 / Published online: 22 October 2025

Abstract
Infrastructure inspection is pivotal in ensuring transportation safety, extending service life, and preventing major accidents. This study addresses the challenges associated with inspecting large-span bridges, particularly high-tower structures, by proposing a novel defect identification method that integrates dual-mode UAV camera information. The proposed method consists of two primary stages. In the global information construction phase, a wide-angle camera captures large-scale scene images, while a synthetic strategy based on deep feature representation focuses on high-tower target areas, generating comprehensive images. The resulting full-scale model of the tower surface provides essential macrostructural information for defect analysis. In the local information analysis phase, a zoom camera is employed to achieve precise focus on smaller regions. A segment-based partitioning strategy is introduced to organically integrate global and local details, establishing a multi-scale information framework that balances macro-level structural assessment with micro-level precision analysis. For localized images, a lightweight detection model, Light-YOLO, based on feature ablation and fusion, is proposed to eliminate redundant information while enhancing multi-dimensional connections between spatial and channel features. By harmonizing the holistic examination of large-scale structures with the meticulous representation of local features, the proposed approach ensures both comprehensive structural assessment and fine-grained defect detection. Validation experiments conducted on real bridges demonstrate that the method provides engineers with efficient and reliable data support, significantly enhancing the accuracy and efficiency of bridge inspections.

Keywords Bridge tower · Unmanned aerial vehicles · Defect detection

1 Introduction

As vital arteries within transportation networks, bridges play a crucial role in public safety and economic development, with their structural integrity and durability directly influencing both domains. However, due to the relentless effects of environmental erosion, fluctuating traffic loads, and material aging, bridges inevitably develop various forms of deterioration, such as surface spalling, reinforcement corrosion, and crack propagation [1]. If left undetected or inadequately addressed, these structural deficiencies can pose significant safety hazards. Consequently, regular inspections to identify potential defects at an early stage have become imperative for ensuring transportation safety and prolonging bridge service life. However, current inspection methods remain largely reliant on manual visual assessments—an approach that is not only inefficient but also highly susceptible to human subjectivity, leading to inconsistencies and inaccuracies in diagnostic results [2]. Thus, there is an urgent need for more scientific and efficient inspection techniques to advance the health management and sustainable development of bridge infrastructure.

With the rapid advancement of unmanned technology, an increasing number of intelligent systems are being integrated into bridge inspection [3]. Devices, such as climbing robots, drones, and automated sensors, are progressively replacing traditional manual inspection methods. Jang et al. [4] designed an annular climbing robot equipped with multiple vision cameras, a propulsion module, and a control computer, enabling high-precision, close-range data acquisition on bridge piers. To facilitate proximity-based data

✉ Jian Zhang
jian@seu.edu.cn

[1] School of Civil Engineering, Southeast University, Nanjing, China

[2] Advanced Ocean Institute of Southeast University, Nantong 22600, China

collection, Yang et al. [5] developed a wall-climbing robot capable of capturing RGB-D images and inertial measurement unit (IMU) data. Field experiments on concrete bridges validated the robot's stability and reliability. Lin et al. [6] introduced the High Mobility Inchworm Climbing Robot (HMICRobot), featuring a hybrid power system, electromagnetically controlled foot pads, and large-wheel core modules, allowing for efficient inspections of both vertical and underside surfaces of steel bridges. Similarly, Ding et al. [7] developed a vacuum-driven climbing robot equipped with a high-resolution camera module and an advanced mobility system, designed for precise image acquisition on sea-crossing bridge towers. While climbing robots excel in high-precision, close-range surface data collection, they face efficiency constraints. Their reliance on prolonged field operations for large-scale structural inspections can hinder overall productivity.

In contrast, drones, with their exceptional mobility, have emerged as indispensable tools in bridge inspection. The widespread adoption of UAV technology has significantly enhanced both inspection efficiency and coverage [8]. Li et al. [9] conducted a series of experiments evaluating the applicability of drones in bridge inspection, focusing on key performance metrics, such as hovering accuracy and data acquisition stability. Liu et al. [10] devised a circular close-range flight path for UAVs, tailored to the geometric characteristics of bridge piers, optimizing inspection accuracy and coverage to enhance the efficiency and precision of surface defect detection. Yoon et al. [11] developed an advanced UAV system integrating high-resolution sensors, GPS, an IMU, and a one-dimensional LiDAR sensor, which was successfully deployed for on-site inspections of the underside and piers of a two-span prestressed concrete box girder bridge. Jiang et al. [12] engineered a vision-based UAV equipped with stereo vision/inertial navigation positioning and sonar-based obstacle avoidance technology to conduct defect inspections on the underside and high towers of the Huai'an Bridge. To address the challenge of aligning the UAV camera plane parallel to the target surface, Peng et al. [13] introduced an innovative UAV system featuring a specially designed top-mounted platform. This system utilizes a three-point laser ranging device to precisely measure the distance and the angle between the target surface and the imaging plane. It was successfully deployed in field inspections of the side and underside of the Xiang-Jiang River Bridge. Ding et al. [14] conducted full-scale calibration of the UAV's gimbal camera, enabling accurate measurement of the true physical dimensions of structural defects in bridge inspections.

The introduction of intelligent inspection devices has significantly streamlined the field data acquisition process. However, as the volume of inspection data continues to grow, engineers spend considerable time making experience-based judgments, which compromises the objectivity and consistency of data analysis. To overcome the limitations of manual assessment, automated detection technologies based on computer vision have been widely adopted [15]. Adhikari et al. [16] designed a digital image processing workflow tailored to the detection of defects in concrete structures. Talab et al. [17] integrated the Otsu method with multi-filtering techniques to propose a three-stage processing scheme specifically for crack detection in concrete. Dias-da-Costa et al. [18] conducted a tracking analysis of defect progression based on image similarity principles. While these digital image processing techniques rely heavily on handcrafted features, their performance often lacks stability under varying conditions, necessitating a degree of human intervention.

In recent years, deep learning algorithms have emerged as a transformative force in image processing. With their self-learning and adaptive capabilities, they have overcome the limitations of traditional methods that depend on manually designed features, significantly enhancing robustness and accuracy [19, 20]. Cha et al. [21] combined a pre-trained convolutional classification network with a sliding window technique to develop an innovative vision-based structural crack detection method, effectively expanding coverage and improving detail resolution. Zhang et al. [22] incorporated convolutional modules with fixed parameters to design an advanced defect classification and recognition model—FF-BLS. Laxman et al. [23] trained a binary classification convolutional neural network to achieve automated identification of concrete surface cracks. Chang et al. [24] introduced a novel convolutional neural network architecture that integrates layer pruning and parameter reduction techniques, aiming to fully automate defect identification.

To extract more comprehensive defect information from images, such as defect locations and quantities, researchers have developed object detection models based on classification frameworks, enabling both identification and precise localization of structural defects. Cha et al. [25] proposed a defect detection method based on the Faster Region-based Convolutional Neural Network (Faster R-CNN), successfully achieving accurate identification of multiple defect types, including delamination, corrosion, and cracks. Jiang et al. [12] integrated MobileNetV2 as the backbone feature extraction network into the YOLOv3 framework, introducing YOLO-MobileNet to enhance the efficiency of multi-type concrete defect detection. Yu et al. [26] optimized the YOLOv4 model using a pruning algorithm, presenting an improved YOLOv4-FPM model specifically designed for bridge crack detection. Moreover, the model's generalization capability was enhanced through multi-scale dataset training.

To further capture finer morphological details of defects, researchers extended the concept of classification models by transforming image-level classification into pixel-wise classification, leading to the development of segmentation models. Ni et al. [27] employed a Generative Adversarial Network (GAN)-based strategy to learn morphological features of defects, significantly improving the segmentation network's performance in defect recognition. Wang et al. [28, 29] developed a series of high-precision segmentation networks tailored for detecting cracks on the inner surfaces of bridge structures. Their effectiveness and feasibility were validated through indoor loading experiments and on-site bridge tests. Song et al. [30] introduced a PSPNet network incorporating an improved self-attention mechanism, enabling highly efficient crack segmentation even in complex backgrounds.

In the field of bridge inspection, a variety of intelligent technologies have been developed to enhance detection precision and operational efficiency. Crawler robot-based methods are capable of acquiring high-resolution visual data, yet they are often constrained by slow deployment, limited mobility, and complex operational requirements, making them less suitable for large-scale or high-elevation structures such as bridge towers. UAV-based approaches, by contrast, offer enhanced flexibility and efficiency in data acquisition. However, existing UAV applications have largely concentrated on localized damage detection, lacking an integrated and scalable framework tailored for the comprehensive inspection of tall bridge towers.

To address this gap, this study proposes a dual-mode camera information fusion strategy designed specifically for UAV-based bridge tower inspection. By simultaneously deploying a wide-angle camera and a zoom camera during flight, the system captures both global structural context and fine-grained local defect details. The wide-angle camera is used to rapidly acquire images of the bridge tower's overall spatial structure, enabling the construction of a global spatial model through targeted perception algorithms. Meanwhile, the zoom camera captures localized regions in high detail, ensuring a seamless linkage between macro- and micro-scale data. Building upon this data acquisition framework, a lightweight detection model, Light-YOLO, is introduced to perform fast and precise defect recognition on the localized images. This enables not only the identification of defect types but also an analysis of their spatial distribution and topological relationships. The proposed method ultimately offers engineers a comprehensive, multi-scale understanding of the bridge tower's operational status, facilitating timely structural health assessment and providing a reliable basis for subsequent maintenance planning.

The structure of this paper is as follows: Sect. 2 outlines the research methodology adopted in this study. Section 3 presents the technical framework for global information construction. Section 4 elaborates on the strategy for integrating global and local information. Section 5 introduces the Light-YOLO algorithm. Section 6 details the execution of field experiments and analyzes the obtained results. Finally, Sect. 7 summarizes the key contributions of this study and offers insights into potential future research directions.

2 Proposed methodology

The inspection of high towers in large-span bridges presents significant technical challenges due to their immense scale and complex inspection environments. Existing detection methods still suffer from limitations in comprehensiveness, efficiency, and fine-detail processing. To address these challenges, this study introduces an innovative defect identification and localization method for bridge towers, leveraging a dual-

mode camera information fusion strategy. A two-stage inspection framework is established, encompassing global information acquisition and in-depth local feature extraction, as illustrated in Fig. 1.

In the global information construction phase, a wide-angle camera, mounted on a UAV platform, rapidly captures the overall structural features of the bridge tower, ensuring full coverage of its macroscopic surface. However, this process often introduces complex backgrounds and redundant information. To mitigate these issues, a targeted region perception strategy based on deep stitching is proposed. This strategy employs a rectangular grid-based prompt engineering technique within the Segment Anything Model (SAM) to precisely filter out extraneous data, preserving only the essential structural features of the bridge tower. Subsequently, a region segmentation technique is used to extract targeted areas T_i^E and retained areas T_i^R. The Local Feature Matching with Transformers(LoFTR) model, built upon a transformer architecture, is applied to detect deep features and achieve high-precision feature matching, thereby generating a comprehensive planar model of the bridge tower $(T_1 + T_2 + \cdots + T_i)$. While the global information construction phase ensures macroscopic integrity, it falls short in capturing intricate defect details. To compensate for this limitation, the local information extraction phase employs a telephoto camera with a narrow field of view to conduct detailed surface inspections, capturing fine-grained defect features. To seamlessly integrate local details with the global planar model, a cross-scale information connection method is introduced, using segments as the fundamental inspection unit. By employing the LoFTR detection tool, this method removes redundant regions from local images and constructs a vertical segmentation framework aligned with the bridge tower. By incorporating the topological structure of segment regions with global-scale information, a deep association between macro- and micro-level data is established. To efficiently process the vast amount of local inspection data,

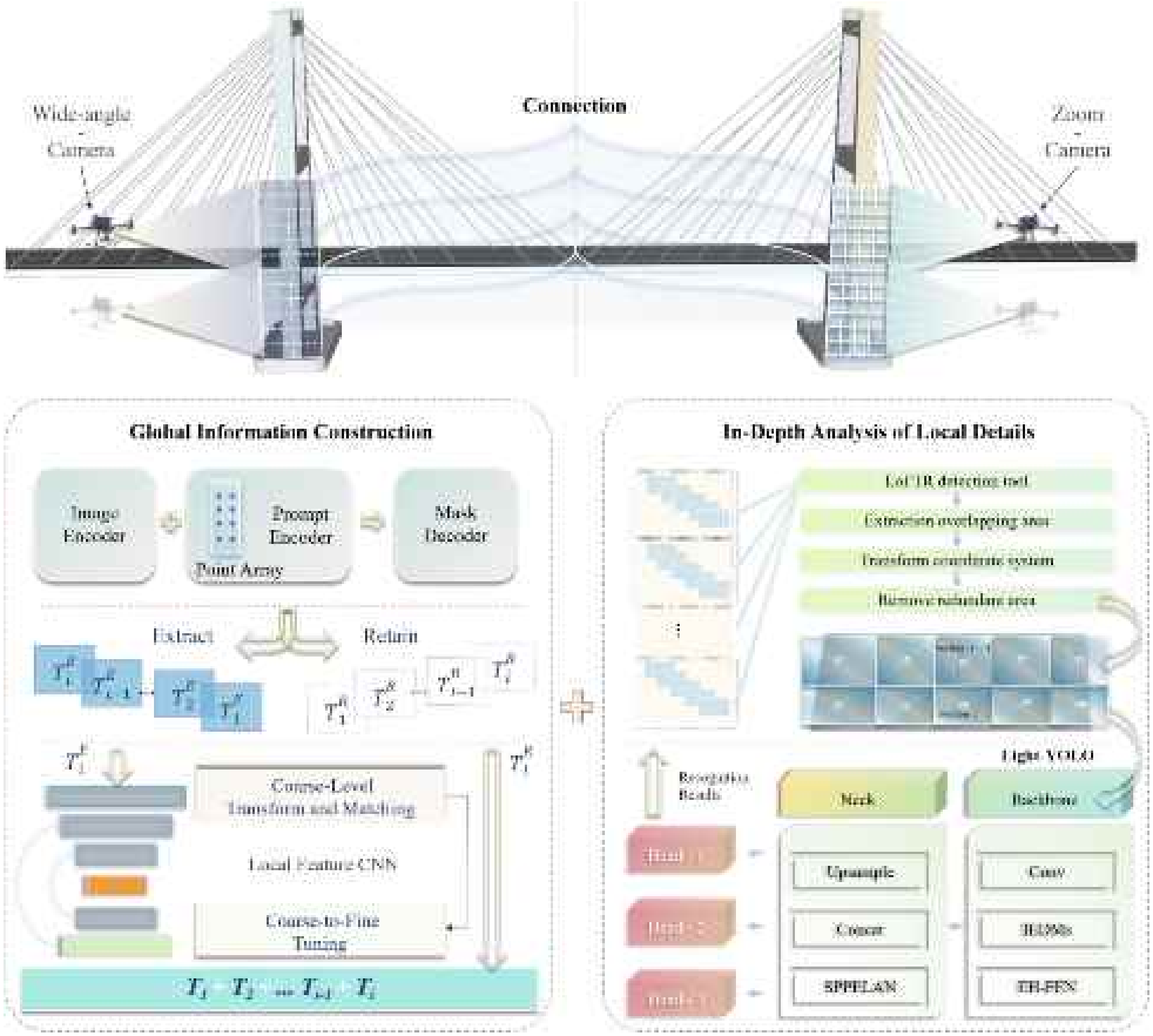


Fig. 1 Framework for defect identification in bridge towers

this study further incorporates an optimized Light-YOLO model, ensuring simultaneous improvements in detection accuracy and computational efficiency. The model refines feature extraction by concentrating on salient defect characteristics while employing a multi-dimensional information fusion strategy to enhance hierarchical feature representation and efficient expression.

The proposed method presents an integrated inspection framework for bridge towers, combining global condition assessment with fine-grained defect identification. By leveraging dual-mode imaging, it enables rapid evaluation of the overall structural state while accurately detecting early-stage defects. This facilitates proactive maintenance by preventing minor issues from escalating, and supports structured data management for efficient tracking and long-term decision-making.

3 Targeted feature extraction for global information establishment

To facilitate effective global information construction, this study introduces a targeted feature extraction-based approach, as depicted in Fig. 2. The method comprises two principal stages: target region awareness and deep feature representation. Although wide-angle cameras effectively capture broad scene coverage, complex backgrounds introduce substantial redundancy, hindering global information

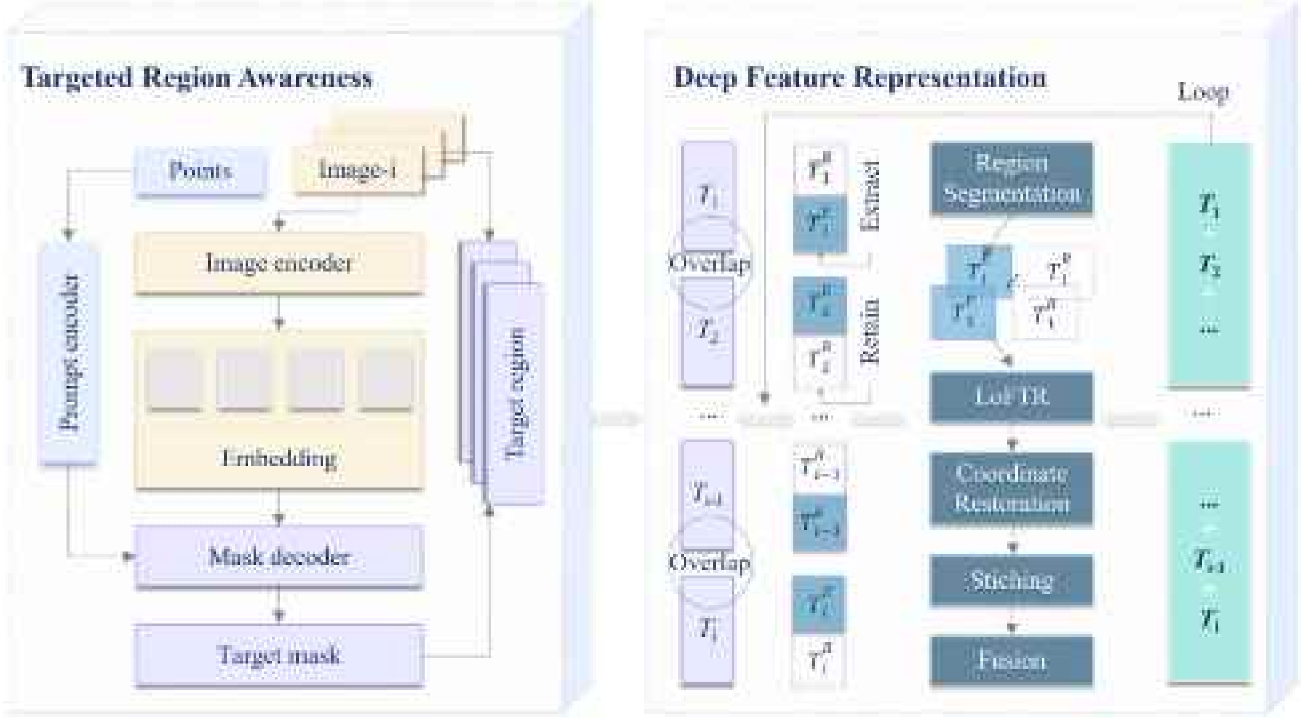


Fig. 2 Process of global information construction

reconstruction. Accurate extraction of target regions from cluttered scenes is thus essential. Manual methods are intuitive but inefficient, while segmentation networks, though effective, demand extensive datasets and high computational costs. Given the limited number of images required for reconstruction, training a dedicated segmentation model often proves cost-inefficient.

To address these challenges, this study integrates the Segment Anything Model (SAM) [31, 32] and incorporates matrix point-based prompt engineering to enable rapid perception and efficient extraction of target regions. SAM consists of three key components: an image encoder, a prompt encoder, and a mask decoder. The image encoder, built on the Vision Transformer (ViT) architecture [33], serves as the core of feature extraction, encoding input images into feature vectors for segmentation. Each input image is divided into fixed-size patches, which are projected linearly to generate embedding vectors. These embeddings are further enriched with position encoding to retain spatial information. ViT employs a multi-layer Transformer encoder structure [34], where each layer comprises a multi-head self-attention mechanism and a feedforward neural network, progressively constructing higher-level feature representations.

The prompt encoder plays a critical role in guiding the SAM segmentation process by encoding external prompt information and integrating it with visual features. In this study, to mitigate potential segmentation errors due to the model's limited understanding of domain-specific structures, matrix point-based prompts are employed. These prompts provide structured and localized guidance, with point coordinates embedded as feature vectors through an embedding layer. This strategy effectively steers the segmentation model toward the bridge tower region, enhancing the accuracy of feature extraction and ensuring reliable global structural representation in complex bridge scenes.

The mask decoder subsequently fuses image features extracted by the image encoder with matrix point features generated by the prompt encoder. Through a cross-attention mechanism, the model facilitates feature interaction, establishing associations between different feature sources. The mask decoder consists of multiple stacked decoding layers, where each layer employs self-attention mechanisms and feedforward neural networks to iteratively refine the image and prompt features. Ultimately, the model generates target masks precisely aligned with the given prompts, enabling accurate extraction of target regions. It is important to note that the performance of this prompt-driven segmentation process relies on the clarity of the input images. In scenarios where image quality is compromised, critical structural features of the bridge tower may become blurred or obscured, thereby reducing the accuracy of prompt-based guidance and potentially leading to incomplete or erroneous segmentation. Therefore, to ensure the robustness of the proposed method, image acquisition must be performed under suitable conditions.

During the construction of a complete planar model, prioritizing feature search and matching within overlapping regions rather than across the entire area significantly enhances efficiency and improves matching accuracy. To achieve this, this study introduces a regional partitioning parameter, ρ, which divides the target region along the height dimension of the overall area $\{W, H\}$ into an extraction region T_i^E and a retention region T_i^R, defined as follows: $[T_i^E:\{w_i^E = W, h_i^E = \rho H\}, T_i^R:\{w_i^R = W, h_i^R = (1-\rho)H\}]$. w_i^E and w_i^R represent the widths, h_i^E and h_i^R denote the heights of the respective subregions. This partitioning method effectively concentrates computational efforts on the most critical portions of the target region. The extraction region T_i^E is subsequently fed into the LoFTR model to facilitate deep feature extraction and high-precision matching.

LoFTR [35] is a high-precision local feature matching approach that integrates convolutional neural networks (CNNs) with Transformers. Its architecture comprises three key components: a feature extraction module, a coarse matching module, and a fine matching module, as illustrated in Fig. 3. The feature extraction module leverages convolutional networks to process input images through multiple hierarchical layers, generating low-resolution feature representations $\widetilde{T_{i-1}^E}, \widetilde{T_i^E}$ at a 1/8 scale and high-resolution feature representations $\widehat{T_{i-1}^E}, \widehat{T_i^E}$ at a 1/2 scale, capturing multi-scale feature information.

In the coarse matching module, the low-resolution features $\widetilde{T_{i-1}^E}, \widetilde{T_i^E}$ are flattened into one-dimensional vectors, enriched with positional encoding, and passed through the LoFTR module. Self-attention layers establish complex contextual relationships among feature points, while a cross-view attention mechanism is introduced to enable deep feature association and fusion between image pairs. The resulting matched features undergo dot product computation and dual Softmax normalization to generate a matching probability matrix. An outlier rejection process using a mutual nearest-neighbor algorithm refines the matches, yielding a preliminary matching prediction $(\tilde{x}, \tilde{y}) \in N_c$, where N_c represents the set of inlier matches from the coarse stage.

In the fine matching module, the coarse matching points $\tilde{x}, \tilde{y}$ are mapped onto the high-resolution feature representations $\widehat{T_{i-1}^E}, \widehat{T_i^E}$, obtaining $\hat{x}, \hat{y}$. The corresponding mapped regions are then cropped and fed into the LoFTR module for fine-grained feature extraction. The center feature of region $\hat{x}$ is compared against all features within region $\hat{y}$, followed by a normalization process. Finally, the refined matching point pairs $\left\{\left(\hat{x}, \hat{y'}\right)\right\}$ are computed using a probabilistic distribution function $\mathrm{E}(\cdot)$, ensuring precise feature alignment.

The high-confidence matching pairs selected by LoFTR are used to derive the transformation matrix between images, establishing the spatial transformation relationship between the two frames. While the transformation of the extraction region has been completed in this process, ensuring global consistency necessitates the effective integration of the

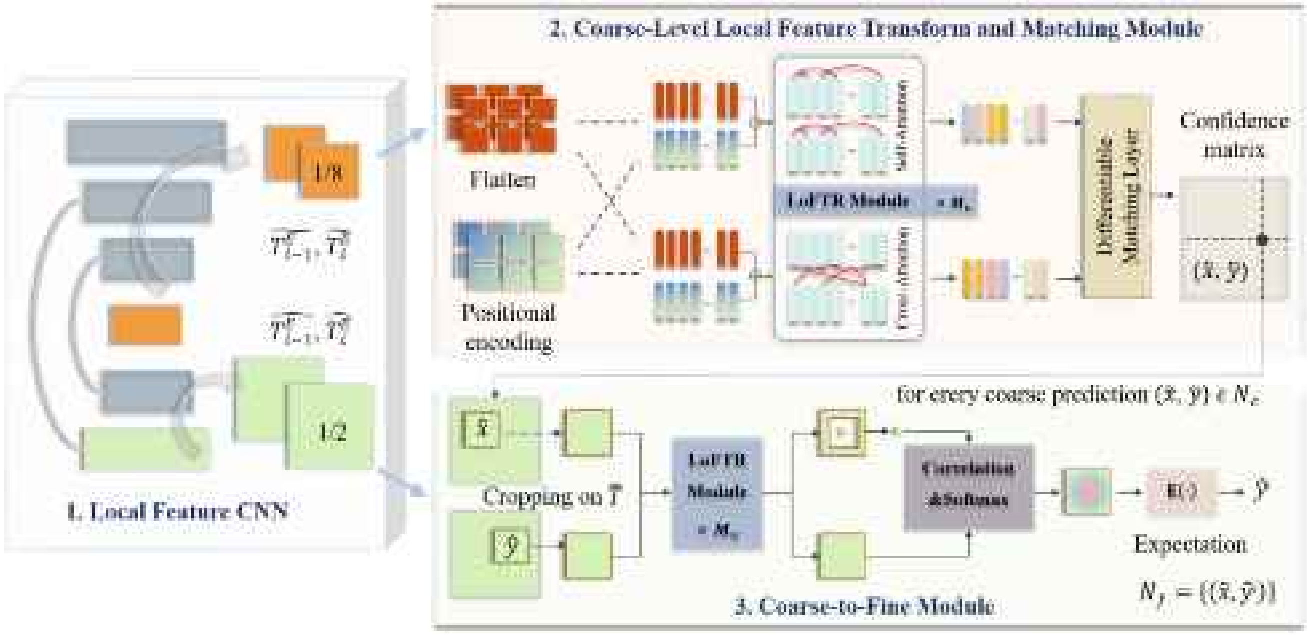


Fig. 3 Procedure of the LoFTR algorithm

retention region. Specifically, the extraction region T_i^E and the retention region T_i^R are first recombined to reconstruct the complete image structure $\{W, H\}$, followed by an update of the spatial relationships of the matched points to compute the overall transformation matrix. Once the complete transformation parameters are obtained, the image T_{i-1}^R is aligned to the coordinate system of T_i^R, thereby completing their seamless stitching. To enhance the quality of the stitching process, a multi-band blending technique is introduced to refine the fused output, ensuring smooth transitions between images. This process serves as the fundamental unit of image stitching. By iteratively executing this operation, multiple images are progressively aligned and merged, ultimately generating a high-precision, structurally complete planar model of the bridge tower.

4 Segment-based global–local information fusion construction method

A comprehensive planar model can swiftly capture global positional information. However, due to the complexity of large-scale images and the vast amount of real-world data represented by each pixel, it becomes challenging to convey fine-grained details. Consequently, relying solely on such images for routine structural surface inspections proves impractical. To compensate for the absence of localized detail, a zoom lens is simultaneously employed to capture high-resolution imagery of structural features. Localized images, with their enhanced focus on fine details, enable precise analysis of defect characteristics and facilitate the extraction of relevant defect information. However, in current inspection practices, while local images effectively identify structural anomalies, their inability to incorporate spatial positioning information prevents the accurate determination of defect locations within the overall structural context.

To address the issue of spatial discreteness in local images, this study introduces the intermediate-level concept of a "segment" serving as a bridge between detailed local imagery and the global planar model. During the local information acquisition phase, a zoom lens with a narrow field of view performs lateral scanning in an upward row-by-row manner. Each complete lateral scan is defined as a "segment," functioning both as a modular unit for integrating local information and as a fundamental building block of the global planar model. By sequentially connecting the local details captured in each row, a segment unit is rapidly constructed. However, to ensure comprehensive coverage during data acquisition, a certain degree of overlap between local images is inevitable, occurring not only in the horizontal direction but also along the vertical axis. Consequently, segments formed purely through sequential connections contain substantial redundant regions. To obtain a refined segment unit, overlapping portions in both directions must be eliminated, as illustrated in Fig. 4. Horizontally, overlap primarily arises between local images within the same segment. In this stage, processing is conducted segment by segment, ensuring that all overlapping regions within a given segment are removed before proceeding to the next. This sequential processing strategy maintains the consistency and orderliness of image registration and redundancy elimination, thereby minimizing potential errors and inconsistencies that could arise from cross-group processing.

Specifically, the local image list is first divided into I groups, each containing N images, based on the number of segments. Within each group, adjacent images are paired sequentially using a sliding window approach to generate image pairs for processing. For instance, the first image in a group is paired with the second to form the first image pair, the second with the third to form the second pair, and so forth, until all image pairs within the group have been processed.

Each image pair undergoes feature point matching using the deep learning model LoFTR, which extracts local features between images. The RANSAC algorithm is then applied to estimate the homography matrix from the matched feature points, further computing the geometric transformation between images. Based on this transformation matrix, the second image is aligned spatially to the coordinate system of the first image, allowing for precise determination of their overlapping region. A mask is then constructed and applied to the second image to remove the redundant overlapping area. This process is repeated iteratively for each group until all groups have been processed.

In the vertical direction, eliminating the overlap between adjacent image groups, $Section_{i-1}$, $Section_i$, is a critical step in redundancy removal. Given two adjacent image groups, the images in $Section_{i-1}$ are sequentially arranged, while those in $Section_i$ are matched correspondingly. Initially, the first image of $Section_{i-1}$, denoted as $image(\mathrm{n1})$, is paired with the first image of $Section_i$, denoted as $image(\mathrm{n2})$, for processing. LoFTR is then employed to identify and match key feature points, from which the homography matrix is estimated to determine the geometric transformation between images. This transformation is applied to extract the overlapping region, which is subsequently removed.

Next, the width of the removed region, $\mathcal{H}_R$, is examined to determine whether it is smaller than the original image width, $\mathcal{H}$, as an initial assessment of redundancy. If excessive overlap remains, the next image in $Section_{i-1}$, $image(\mathrm{n1} = \mathrm{n1} + 1)$, is paired with $image(\mathrm{n2})$ from $Section_i$, and the process repeats until no redundancy remains. The procedure then continues with $image(\mathrm{n1} = \mathrm{n2})$ from $Section_{i-1}$ being matched with $image(\mathrm{n2} = \mathrm{n2} + 1)$ from $Section_i$, iterating until all images in $Section_i$ have been

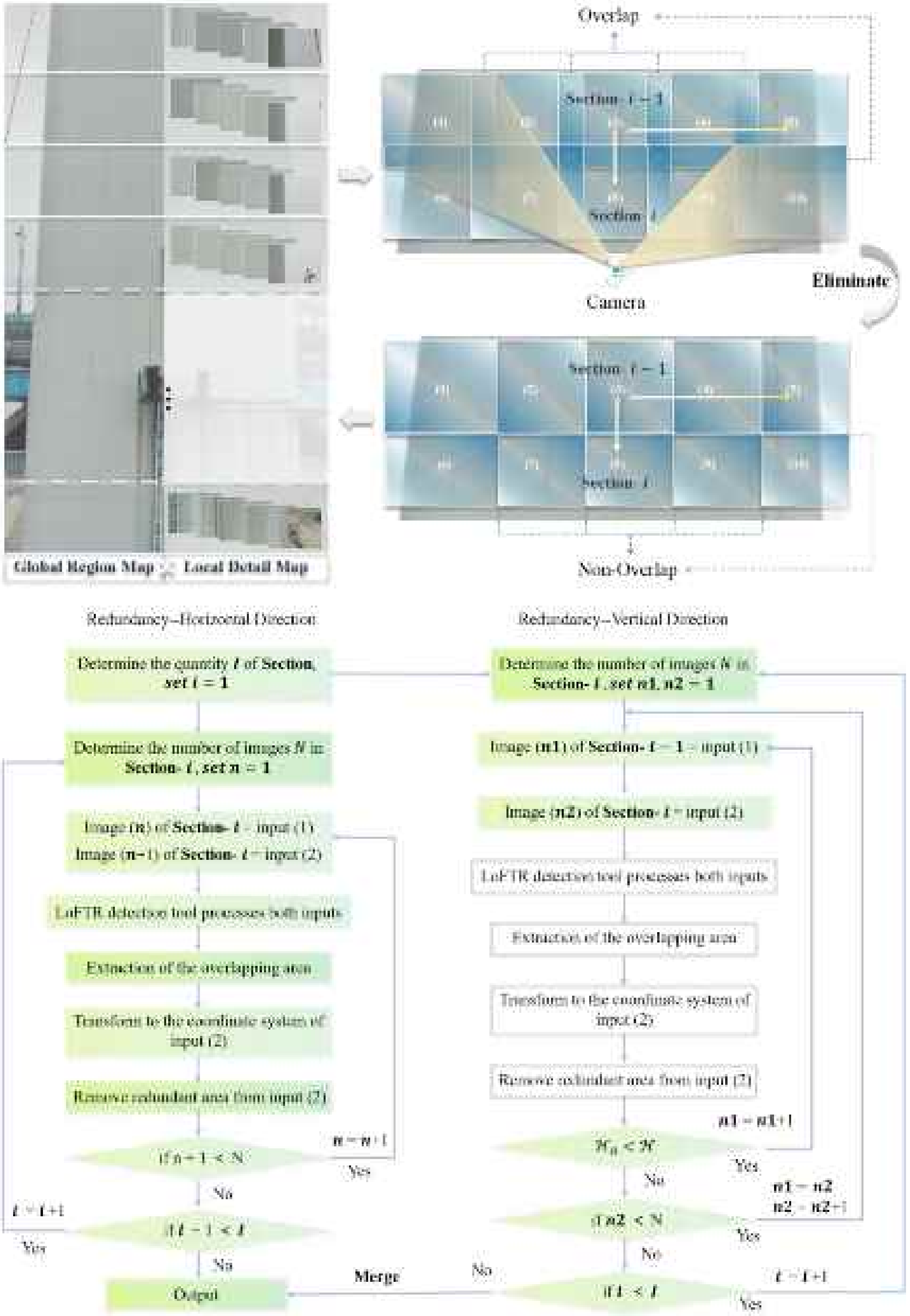

Overlap
Section- $i-1$
Section- i
Camera
Eliminate
Section- $i-1$
Section- i
Non-Overlap
Global Region Map
Local Detail Map
Redundancy–Horizontal Direction
Determine the quantity I of Section, set $i = 1$
Determine the number of images N in Section- i, set $n = 1$
Image (n) of Section- i = input (1)
Image ($n-1$) of Section- i = input (2)
LoFTR detection tool processes both inputs
Extraction of the overlapping area
Transform to the coordinate system of input (2)
Remove redundant area from input (2)
$n = n+1$
if $n+1 < N$
Yes
No
$i = i+1$
if $i-1 < I$
Yes
No
Output
Merge
Redundancy–Vertical Direction
Determine the number of images N in Section- i, set $n1$, $n2 = 1$
Image ($n1$) of Section- $i-1$ = input (1)
Image ($n2$) of Section- i = input (2)
LoFTR detection tool processes both inputs
Extraction of the overlapping area
Transform to the coordinate system of input (2)
Remove redundant area from input (2)
$n1 = n1+1$
$N_{n} < N$
Yes
No
$n1 = n2$
$n2 = n2+1$
if $n2 < N$
Yes
No
No
$i = i+1$
if $i < I$
Yes

Fig. 4 Modeling methods for global–local information fusion

processed. Following this methodology, image groups are paired in a sliding window manner: the first group is matched with the second, the second with the third, and so on, until all groups have been processed. Finally, the composite panoramic view is segmented into equal inspection units based on the number of segments formed by the local images. Each segment is then mapped onto a unified global spatial coordinate system, completing the integration of local information with the global planar model.

This spatial referencing framework enables engineers to swiftly ascertain the exact position of any given segment within the global model. Furthermore, intricate details within each segment, such as cracks, spalling, and corrosion, can be precisely delineated and analyzed. This process not only resolves the spatial discreteness of local information but also ensures that the global model presents a clear and accurate depiction of each segment's condition, providing robust data support for structural defect analysis, maintenance planning, and repair decision-making.

5 Proposed light-YOLO detection network

While the zoom camera captures highly detailed local information, its limited field of view results in an extensive volume of data, imposing a significant computational burden on data analysis. Moreover, to achieve high-precision defect detection, existing detection networks often employ complex model architectures [36, 37], which not only lead to increased computational costs but also restrict their applicability in resource-constrained environments. To address these challenges, researchers have recently introduced a series of lightweight object detection methods [38–40]. The fundamental principle of lightweight detection is to reduce model parameters and computational complexity while preserving detection accuracy, thereby mitigating the high resource demands of traditional approaches. Against this backdrop, YOLO (You Only Look Once) has emerged as a widely adopted object detection framework due to its remarkable efficiency and high detection accuracy [41]. By simultaneously performing object localization and classification within a single network, YOLO eliminates the need for the complex region proposal stages required in traditional methods. However, the model itself still entails a certain level of complexity, necessitating substantial computational resources and storage. To address the dual demands of efficiency and accuracy in analyzing close-range, high-resolution images, this study introduces a lightweight enhancement—Light-YOLO, based on the YOLOv9 [42] framework, as illustrated in Fig. 5. The model is specifically optimized for processing detailed images captured by the zoom camera, which contain rich and precise defect information essential for automatic damage detection. The enhancements primarily target two aspects: sampling strategies and feature extraction mechanisms, aiming to reduce computational cost while preserving high recognition accuracy.

The Light-YOLO framework operates in two stages: training and inference. During training, a main structure comprising Backbone, Neck, and Head is combined with an auxiliary branch to enhance learning and detection performance. The Backbone extracts multi-level visual features, the Neck refines and fuses them across scales, and the Head performs classification and bounding box regression. The auxiliary branch provides additional supervision, helping the model retain fine-grained features and mitigating information loss from deep-layer abstraction. This promotes more balanced feature propagation and improves detection accuracy. In the inference phase, only the main structure is used, ensuring high efficiency and low computational cost, making Light-YOLO suitable for real-time deployment.

To obtain multi-scale feature representations, downsampling is typically employed albeit at the cost of partial information loss. To address the diverse requirements of feature extraction across model scales, Information Enhancement Downsampling Modules (IEDMs) are introduced, as shown in Fig. 6. For the large-scale model Light-YOLO-c, which possesses rich channel information, an IEDM-C module is designed to effectively refine and integrate multidimensional features. This module comprises two complementary branches: a global branch and a local branch. The input feature map $\mathcal{X} \in \mathcal{R}^{C \times H \times W}$ is evenly divided into local $\mathcal{X}_l \in \mathcal{R}^{C_1 \times H \times W}$ and global $\mathcal{X}_g \in \mathcal{R}^{C_2 \times H \times W}$ components, with $C_1 = C_2 = C/2$. The global branch reconstructs the feature space by aggregating spatial and channel-wise information, achieving deep fusion of global features across multiple dimensions. In implementation, $\mathcal{X}_g$ is further partitioned into sub-feature maps corresponding to distinct spatial regions, and the process is formulated as follows:

$$\begin{cases} x_{0,0} = X\left[0{:}C{:}\theta_c, 0{:}H{:}\theta_s, 0{:}W{:}\theta_s\right], x_{1,0} = X\left[1{:}C{:}\theta_c, 1{:}H{:}\theta_s, 0{:}W{:}\theta_s\right] \\ x_{0,1} = X\left[2{:}C{:}\theta_c, 0{:}H{:}\theta_s, 1{:}W{:}\theta_s\right], x_{1,1} = X\left[3{:}C{:}\theta_c, 1{:}H{:}\theta_s, 1{:}W{:}\theta_s\right] \\ \mathcal{X}'_d = SiLU\left(BN\left(Conv_{1*1}\left(x_{0,0} \oplus x_{1,0} \oplus x_{0,1} \oplus x_{1,1}\right)\right)\right) \end{cases} \tag{1}$$

To enable cross-channel information reorganization, a channel-wise sampling coefficient θ_c (default: 4) is introduced. The first sub-feature map $x_{0,0}$ is formed by sampling the input feature map from the first channel with a step of θ_c. Similarly, $x_{1,0}$ denotes the second sub-feature map, obtained from the second channel with the same step. Following this scheme, four sub-feature maps are generated along the channel dimension. In parallel, a spatial sampling coefficient θ_s (default: 2) governs the interval-based selection along spatial dimensions H and W. For instance, $x_{0,0}$ is also used to denote the first spatial sub-feature map derived from the top-left spatial element using a step of θ_s, resulting in a resolution of

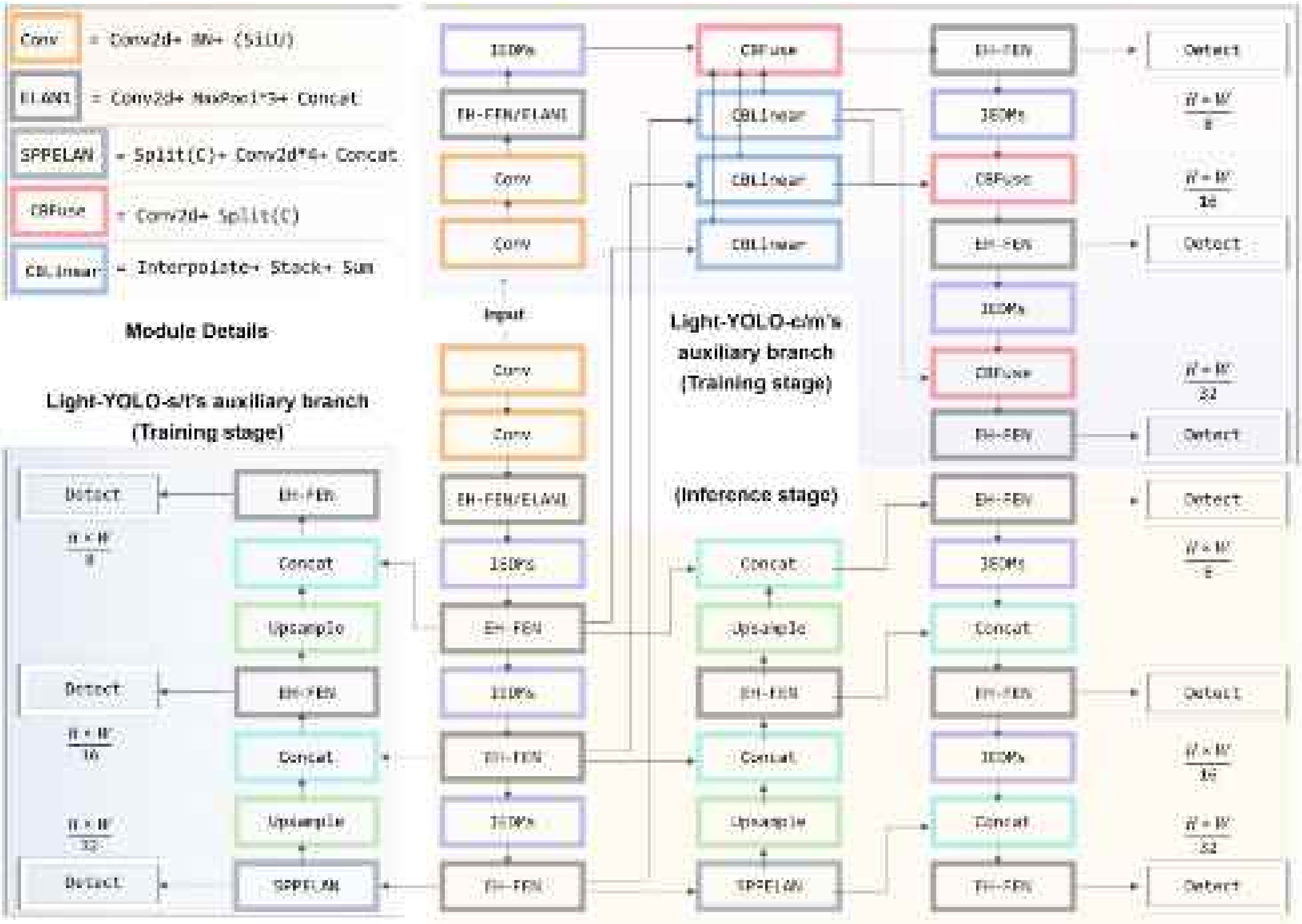

Conv = Conv2d+ BN+ (SiLU)
ELAN1 = Conv2d+ MaxPool*3+ Concat
SPPELAN = Split(C)+ Conv2d*4+ Concat
CBFuse = Conv2d+ Split(C)
CBLinear = Interpolate+ Stack+ Sum
Module Details
Light-YOLO-s/t's auxiliary branch (Training stage)
Light-YOLO-c/m's auxiliary branch (Training stage)
(Inference stage)
Input
Conv
EH-FEN/ELAN1
3EDMs
EH-FEN
CBFuse
CBLinear
Concat
Upsample
SPPELAN
Detect
$\frac{H\times W}{8}$
$\frac{H\times W}{16}$
$\frac{H\times W}{32}$

Fig. 5 The network architecture of Light-YOLO

$[\frac{H}{\theta_s}, \frac{W}{\theta_s}]$. Subsequent maps, such as $x_{1,0}$, are obtained by offsetting the starting point along the H-axis and repeating the sampling. All resulting sub-feature maps are concatenated along the channel dimension ($\bigoplus$) to form a reorganized feature representation, which is then processed by a 1×1 convolutional layer for spatial remapping. Batch Normalization (BN) is applied to stabilize training and address gradient vanishing or explosion.

Finally, the SiLU activation function introduces nonlinearity, enabling the network to capture more complex feature representations, yielding an output $\mathcal{X}'_g \in \mathcal{R}^{C_5 \times \frac{H}{2} \times \frac{W}{2}}$. The local sampling branch, primarily focused on extracting fine-grained details, employs a combination of convolutional operations and max pooling. Specifically, the C_3 and C_4 channel components of $\mathcal{X}_l$ undergo dimensionality reduction, where $C_3 = C_4 = C_1/2$. Ultimately, the three output features from both branches are fused through concatenation ($\oplus$), producing a highly efficient and high-quality output representation $\mathcal{X}'_{out} \in \mathcal{R}^{(C_4+C_5+C_6) \times \frac{H}{2} \times \frac{W}{2}}$.

For the mid-scale model Light-YOLO-m, average pooling is first applied for initial feature fusion to compress local information. Spatial sampling and fusion strategies, inspired by global sampling structures in large-scale models, are then integrated to preserve critical features during downsampling by incorporating global contextual information. In Light-YOLO-s/t, where feature diversity is relatively limited, a simplified downsampling module is adopted: average pooling for preliminary fusion followed by convolution-based downsampling to reduce spatial dimensions while maintaining sufficient representational strength.

Conventional feature extraction methods often incur high computational costs when processing high-dimensional inputs, limiting overall efficiency. To address this issue, an Efficient Hierarchical Feature Extraction Network (EH-FEN) is introduced, featuring a multi-level, multi-dimensional enhancement and fusion framework (Fig. 7). EH-FEN consists of two main components: the Hierarchical Multi-Scale Feature Fusion Module (HMS-FFM) and the Compact Coupled Attention (CCA) module. HMS-FFM captures multi-granular feature representations through depth-varied

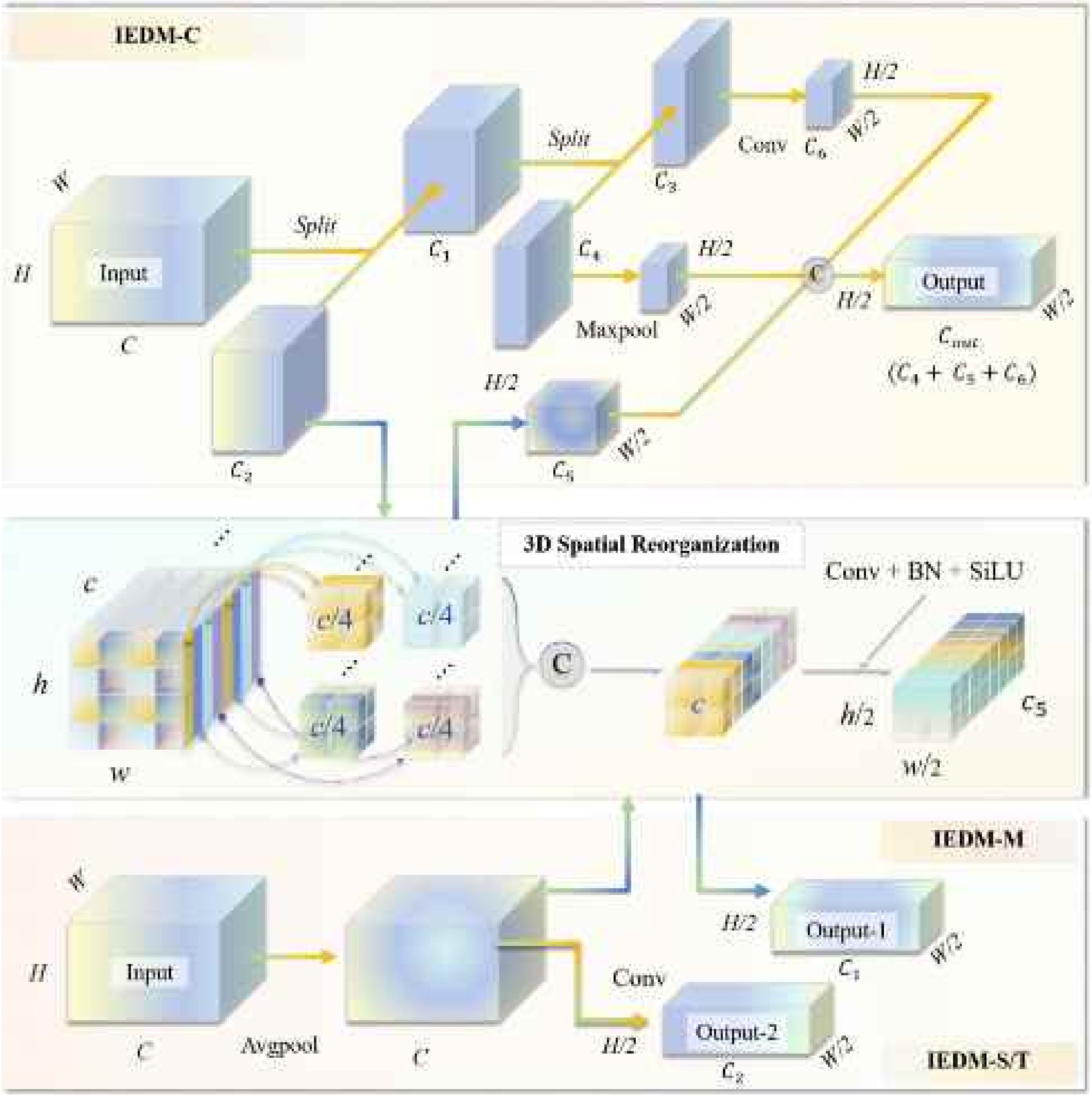


Fig. 6 Illustration of the proposed information enhancement downsampling modules

strategies across hierarchical levels and integrates them via a streamlined fusion mechanism to enable effective cross-level interaction. This hierarchical design enhances representation capacity while maintaining low computational cost. The process begins with shallow feature extraction from the input feature map $\mathcal{X} \in \mathcal{R}^{C\times H\times W}$, where an efficient channel-wise partitioning strategy is applied to optimize computation without compromising informative content.

This module generates two distinct types of feature representations: shallow features $\mathcal{X}_s \in \mathcal{R}^{C_s\times H\times W}$, capturing low-level visual patterns, and deep features $\mathcal{X}_d \in \mathcal{R}^{C_d\times H\times W}$, enabling complex feature learning. Both channels are set as $C_s = C_d = C/2$.

In the deep feature extraction phase, $\mathcal{X}_d \in \mathcal{R}^{C_d\times H\times W}$ is further divided into mid-level features $\mathcal{X}_{mc} \in \mathcal{R}^{C_{mc}\times H\times W}$ and deep-level features $\mathcal{X}_{df} \in \mathcal{R}^{C_{df}\times H\times W}$, where $C_{mc} = C_{df} = C/4$. A 1×1 convolution is first applied to $\mathcal{X}_{mc}$ to obtain coarse mid-level representations $\mathcal{X}'_{mc} \in \mathcal{R}^{C_{mc}\times H\times W}$. For deep-level features $\mathcal{X}_{df}$, initial representation learning is

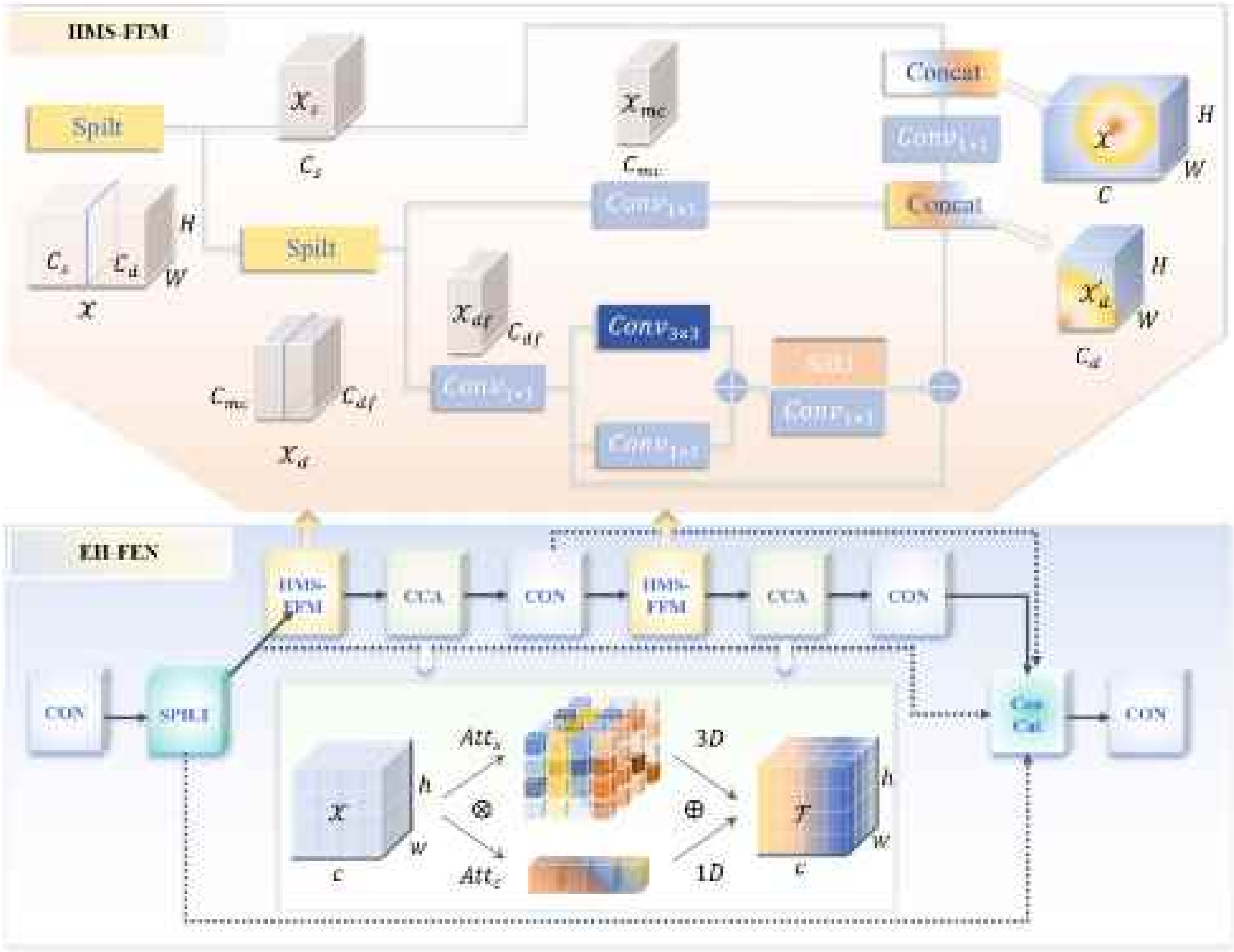


Fig. 7 Architecture diagram of efficient hierarchical feature extraction network

performed via a 1 × 1 convolution. Subsequently, two parallel convolutional branches (1 × 1 and 3 × 3) are used to capture fine-grained local details and wider contextual information. The outputs are element-wise summed, followed by activation and projection through a convolutional layer to produce refined features. These are fused element-wise with the initial convolutional results to form $\mathcal{X}'_{df} \in \mathcal{R}^{C_{df} \times H \times W}$. A progressive fusion strategy is then adopted. Features $\mathcal{X}'_{df} \in \mathcal{R}^{C_{df} \times H \times W}$ and $\mathcal{X}'_{mc} \in \mathcal{R}^{C_{mc} \times H \times W}$ are concatenated along the channel dimension and fused via convolution to obtain $\mathcal{X}'_{d} \in \mathcal{R}^{C_{d} \times H \times W}$. Finally, $\mathcal{X}'_{d}$ is concatenated with $\mathcal{X}_{s}$, forming the unified representation $\mathcal{X}' \in \mathcal{R}^{C \times H \times W}$. This hierarchical integration approach enhances the representational capacity by effectively combining shallow, mid-level, and deep features. The design balances computational efficiency with expressive depth, improving the model's ability to capture intricate patterns and contextual structures.

Attention mechanisms play a critical role in enhancing network performance and feature representation. Traditional designs often rely on additional trainable parameters to learn salient spatial or channel regions, resulting in increased model complexity and computational overhead—particularly problematic in resource-constrained environments such as embedded or mobile platforms. To overcome these limitations, a lightweight attention mechanism termed CCA is introduced. CCA integrates spatial and channel attention within a dual-branch structure to achieve efficient feature enhancement without redundant parameters. The spatial branch captures three-dimensional spatial cues $\mathcal{F}_1 \in \mathcal{R}^{C \times H \times W}$, while the channel branch models inter-channel dependencies to produce $\mathcal{F}_2 \in \mathcal{R}^{C \times H \times W}$. The fused output $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2 \in \mathcal{R}^{C \times H \times W}$ yields a globally enhanced representation with minimal computational burden.

In the spatial attention branch, the input tensor $\mathcal{X} \in \mathcal{R}^{C \times H \times W}$ undergoes channel-wise mean computation across spatial dimensions, serving as the basis for spatial variation analysis. The squared deviation $(\mathcal{X}-\mu)^2$ is then calculated to capture spatial feature dispersion. A global summation over each channel's squared deviations is subsequently performed to obtain the total variation, facilitating stable normalization:

$$S = \sum_{h=1}^{H}\sum_{w=1}^{W}(\mathcal{X}-\mu)^2 = \sum_{h=1}^{H}\sum_{w=1}^{W}\left(\mathcal{X}-\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathcal{X}_{h,w}\right)^2 \tag{2}$$

In this manner, each pixel in the feature map is weighted based on its deviation from the global mean, yielding a three-dimensional spatial attention weight. This weight is further compressed using the Sigmoid function σ, generating the refined three-dimensional spatial attention map $\mathbf{Att}_{\mathrm{s}}$. The final output of this branch is obtained by applying element-wise multiplication between $\mathbf{Att}_{\mathrm{s}}$ and the input feature map $\mathcal{X}$, resulting in the spatially adjusted feature representation $\mathcal{F}_1$:

$$\mathcal{F}_1 = \mathbf{Att}_{\mathrm{s}} \otimes X = \sigma\left(\frac{(\mathcal{X}-\mu)^2}{4\left(\frac{S}{H\times W-1}+\lambda\right)}+0.5\right)\otimes X \tag{3}$$

To prevent division-by-zero errors during computation, a small regularization constant λ is set to 1×10^{-4}. The symbol ⊗ indicates element-wise multiplication. While $\mathcal{F}_1$ enhances the representation of critical regions in 3D space, it lacks modeling of inter-channel dependencies. To overcome this limitation, a channel attention branch is introduced to evaluate the relative importance of feature channels, thereby strengthening the network's focus on informative components.

This branch employs a lightweight and efficient mechanism. A global average pooling (GAP) operation is first applied along spatial dimensions, generating a global descriptor for each channel by averaging all pixel values. This operation discards spatial specificity while summarizing the distribution of feature responses across each channel, effectively capturing channel-wise significance.

Subsequently, a 1D convolution with a small kernel (typically size 3 or 5) is performed along the channel axis to learn inter-channel dependencies and generate weighting coefficients. This convolutional step reduces computational overhead and maintains low parameter complexity. The resulting features are passed through a Sigmoid activation function σ to produce refined channel attention weights $\mathbf{Att}_{\mathrm{c}}$. The final output of the channel attention branch is obtained via element-wise multiplication between $\mathbf{Att}_{\mathrm{c}}$ and the original input feature map $\mathcal{X}$, yielding the channel-enhanced feature representation $\mathcal{F}_2$:

$$\mathcal{F}_2 = \mathbf{Att}_{\mathrm{c}} \otimes X = \sigma\left(Conv1d\left(\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathcal{X}_{h,w}\right)\right)\otimes X \tag{4}$$

*Conv*1*d* refers to a one-dimensional convolution operation with a kernel size of 3. This branch effectively captures inter-channel dependencies while maintaining minimal computational overhead, thereby enhancing overall performance.

5.1 Implementation details

Bridge tower surfaces commonly exhibit defects, such as cracks, corrosion, and spalling. This study leverages the lightweight neural network model, Light-YOLO, to perform automated defect analysis on images captured by drones operating in zoom mode. The model's capacity for accurate defect identification relies on extensive training with a diverse dataset. To address this, data were collected from multiple sources, including smartphones, cameras, and drones, and further enriched with publicly available open-source images [28, 38, 43] to ensure diversity. In total, 3790 defect images of bridge structures were gathered. The defects within these images were manually annotated by experienced engineering technicians, following standardized guidelines that defined defect categories and boundary criteria. The finalized dataset includes 4039 annotated cracks, 3497 instances of spalling, and 1397 corrosion regions. The dataset was partitioned into training and testing subsets in an 8:2 ratio, with 3032 images allocated to the training set for model learning and optimization, and 758 images reserved for evaluating the model's generalization and real-world applicability. All images were resized to 448 × 448 pixels to ensure uniformity in the input data.

The network was trained for 300 epochs with a batch size of 16 per iteration, employing the AdamW optimizer to facilitate convergence. The initial hyperparameter configuration included a weight decay of 5×10^{-4} and a learning rate of 1×10^{-3}. Model performance was rigorously assessed using mean average precision (mAP) at an intersection-over-union (IoU) threshold of 0.5 (mAP-50), a comprehensive metric that amalgamates both precision and recall, thereby providing a robust evaluation of the model's efficacy. The model was trained within the PyTorch framework, running on an Ubuntu 20.04 operating system. The environment specifications included Python version 3.9.7, PyTorch 2.0.0, and CUDA 11.8. For the computational experiments, a GeForce RTX 4090 GPU with 24 GB of memory was utilized.

To conduct a more comprehensive assessment of the performance of the Light-YOLO series models, this study undertakes an extensive comparative analysis with the YOLOv9 series networks. The experimental results, as presented in Table 1, provide a clear depiction of the performance and efficiency across the various models. This

Table 1 Performance comparison of different methods on Dataset

Network	#Params (M)	FLOPs (G)	Pre(%)	Re(%)	mAP-50 (%)
YOLOV9-c	25.2	101.8	89.3	85.1	90.2
YOLOV9-m	19.9	76.0	86.8	81.4	87.8
YOLOV9-s	7.1	26.2	87.8	77.1	84.2
YOLOV9-t	1.9	7.1	80.9	68.3	76.4
Light-YOLO-c	18.8	80.3	89.6	85.0	91.0
Light-YOLO-m	17.5	66.2	87.8	83.4	88.6
Light-YOLO-s	5.5	21.1	88.4	75.3	85.7
Light-YOLO-t	1.5	5.8	73.4	73.4	76.9

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

Table 2 Effectiveness of IEDM and EH-FEN module

IEDM	EH-FEN	#Params (M)	FLOPs (G)	MIoU(%)
		25.2	101.8	90.2
✓		23.3	97.0	90.8
	✓	21.6	88.2	91.2
✓	✓	18.8	80.3	91.0

comparative evaluation reveals that the Light-YOLO series effectively optimizes performance without compromising efficiency, striking an exceptional balance between computational cost and accuracy. Specifically, Light-YOLO-c achieves a 25% reduction in parameter count and a 21% decrease in FLOPs, while simultaneously improving the mAP-50 by 0.8% relative to YOLOv9-c. Among the Light-YOLO variants, the Light-YOLO-s model demonstrates the most substantial improvement in performance. In addition to the efficiency gains, Light-YOLO-s shows a 1.5% increase in mAP-50 over YOLOv9-s, highlighting its superior balance of speed and accuracy. As the lightest model in the Light-YOLO series, Light-YOLO-t enhances performance by 0.5% compared to YOLOv9-t.

To investigate the roles and performance contributions of the key components within the proposed network architecture, designed ablation experiments have been conducted (Light-YOLO-c is employed as a representative case). The IEDM module is introduced during the downsampling stage to enhance multi-scale feature representation while minimizing information loss. The EH-FEN module, on the other hand, is responsible for capturing rich multi-level and multi-dimensional features with minimal computational overhead. The experimental results, presented in Table 2, demonstrate the effectiveness of each module both individually and in combination. The integration of IEDM and EH-FEN leads to a significant reduction in model parameters while maintaining, or in some cases improving, detection precision. This confirms that the proposed structural enhancements not only optimize computational efficiency but also contribute positively to the model's overall detection capability.

The inference results of all models are presented in the detailed visualizations shown in Fig. 8, facilitating a direct comparison of the models' inference performance across various operational scenarios. In particular, Light-YOLO-c exhibits remarkable robustness and precision when confronted with complex backgrounds and large-scale scenes, maintaining high fidelity to the ground truth and even identifying minor cracks that may be overlooked by human inspectors. In contrast, although Light-YOLO-t, the smallest model in the series, offers faster response times, its detection of finer details is less precise. Nevertheless, it remains highly effective for the rapid identification of the most common defects. Overall, the Light-YOLO series achieves an optimal balance between efficiency and performance, demonstrating strong adaptability and stability in a wide range of applications, thereby underscoring its substantial potential for real-world deployment.

6 Filed validation experiments

The test bridge selected for this study is the Jiujiang Yangtze River Highway Bridge in Jiangxi Province, China, as depicted in Fig. 9. The main bridge structure is a double-tower, single-sided hybrid girder cable-stayed bridge with a main span of 1405 m, composed of prestressed concrete girders, steel–concrete composite sections, and steel box girders. The heights of the main bridge towers are 230.854 m for the southern pylon and 242.308 m for the northern pylon. The pylons adopt an H-shaped structural design, consisting of upper, middle, and lower tower columns. This study focuses on the southern pylon, where the upper tower column has a slope of 1/56.43884 and a height of 89.1 m. The middle tower column features an external slope of 1/15.012 and stands 112.31 m tall, while the lower tower column has an external slope of 1/4.545269 and a height of 23.844 m. In the longitudinal direction, the pylon exhibits an I-shaped

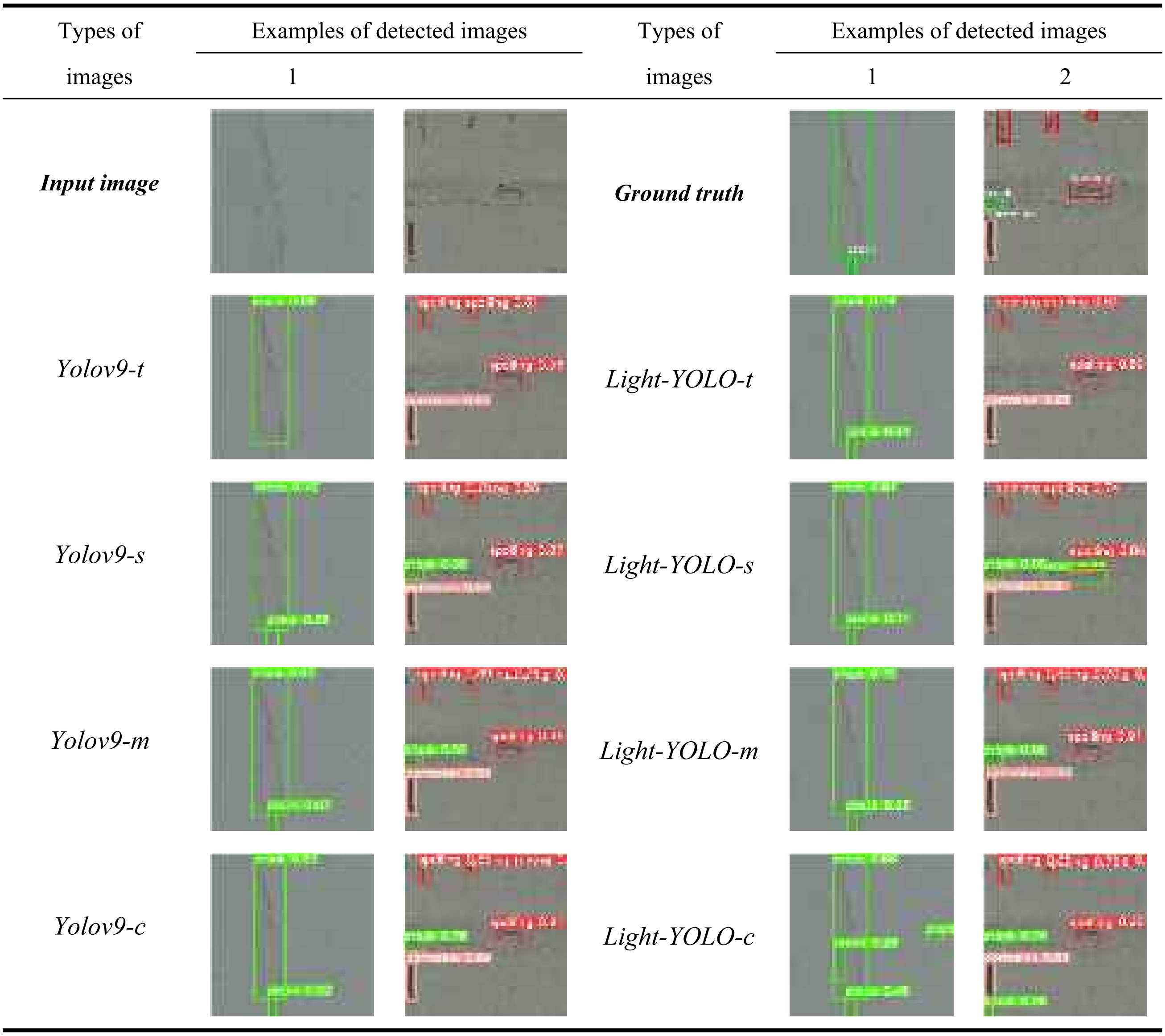


Fig. 8 Visualized results of different networks

configuration, with the upper tower column measuring 8.8 m in width. The middle and lower tower columns have a slope of 1/61.087, with their width gradually increasing from 8.8 to 13.25 m.

For data acquisition, this study employs a DJI M300 RTK UAV, equipped with an H20 camera system. The UAV system supports stable operation in wind conditions up to level 6. It is also rated IP45 for ingress protection, enabling basic resistance against dust and water, which enhances reliability in light rain or mist. Furthermore, for scenarios involving low-light or shaded environments, auxiliary lighting can be integrated into the UAV platform to provide effective illumination compensation. In wide-angle mode, the camera features a 1/2.3-inch CMOS sensor with an equivalent focal length of 24 mm, capable of capturing images at a resolution of 12 megapixels. In zoom mode, the sensor size is 1/1.7-inch CMOS, offering an equivalent focal length range of 31.7–556.2 mm and capturing 20-megapixel high-resolution images. The UAV conducts vertical inspections along the bridge tower, utilizing both wide-angle and zoom imaging modes to detect structural defects, ensuring comprehensive coverage and precise identification of anomalies across different regions and structural details.

The UAV conducted dual-mode data acquisition for different sections of the bridge tower—upper, middle, and lower tower columns—utilizing both wide-angle and zoom imaging modes. The details of the data acquisition process are illustrated in Fig. 10. In wide-angle mode, the

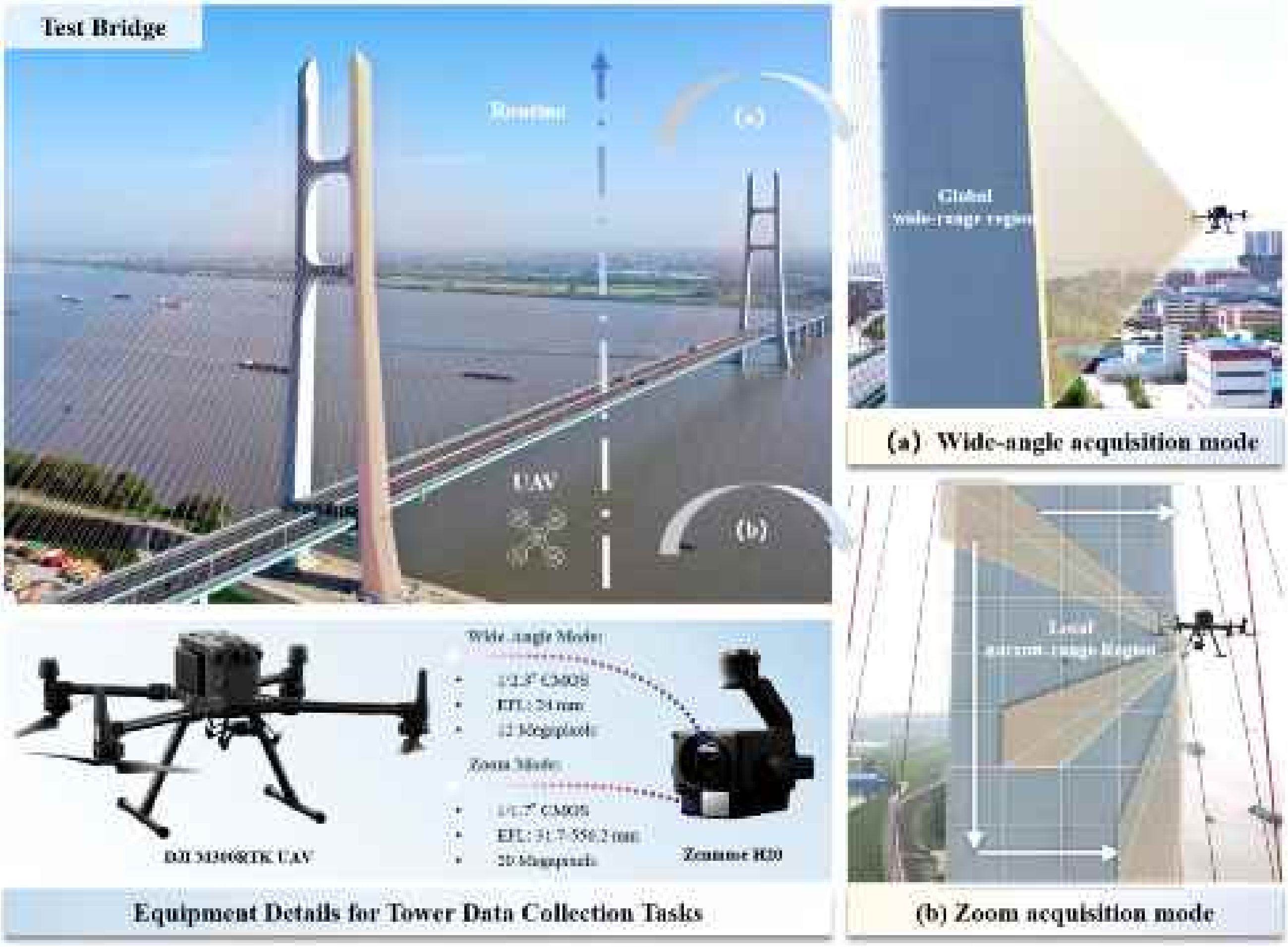


Fig. 9 Schematic diagram of bridge overview

camera, leveraging its large field of view, captured extensive spatial information, producing images with a resolution of 4056 × 3040 pixels. Conversely, in zoom mode, although the camera's field of view was narrower, it provided significantly finer local details, achieving a resolution of 5184 × 3888 pixels. The collected image data are as follows: Lower tower column: 2 wide-angle images, 198 zoom images; Middle tower column: 6 wide-angle images, 488 zoom images; Upper tower column: 4 wide-angle images, 316 zoom images. The limited number of wide-angle images facilitates the rapid construction of a global tower column view. However, due to their lack of fine details, they are not directly suitable for defect detection. Instead, their primary value lies in providing global spatial positioning for the zoomed-in views. The construction of the global view relies on image stitching techniques, where feature detection and matching serve as critical steps. However, the complex background information in wide-angle images introduces significant interference in feature matching, leading to a high rate of erroneous correspondences (see Figs. 11 and 12). To mitigate this issue, this study employs the global view construction method proposed in Sect. 3, aiming to enhance the quality and efficiency of wide-angle image stitching (see Figs. 11 and 12). Specifically, a 12 × 3 rectangular point grid is utilized as a control condition, combined with feature extraction tools based on the SAM model, enabling the rapid elimination of complex background noise in raw data.

Based on empirical observations during UAV data acquisition, the overlapping region between two consecutive wide-angle images generally falls within 30% to 40% of the image height. To ensure sufficient coverage of this overlap while simplifying unified processing, the parameter ρ was empirically set to 0.5. This value consistently yields stable and accurate feature matching results, while reducing the unnecessary computational cost associated with full-image analysis. The operational workflow involves: Dividing images evenly along the height dimension; Cropping non-overlapping areas; Performing feature detection and matching exclusively on overlapping regions. During the global stitching process, the cropped images are subsequently

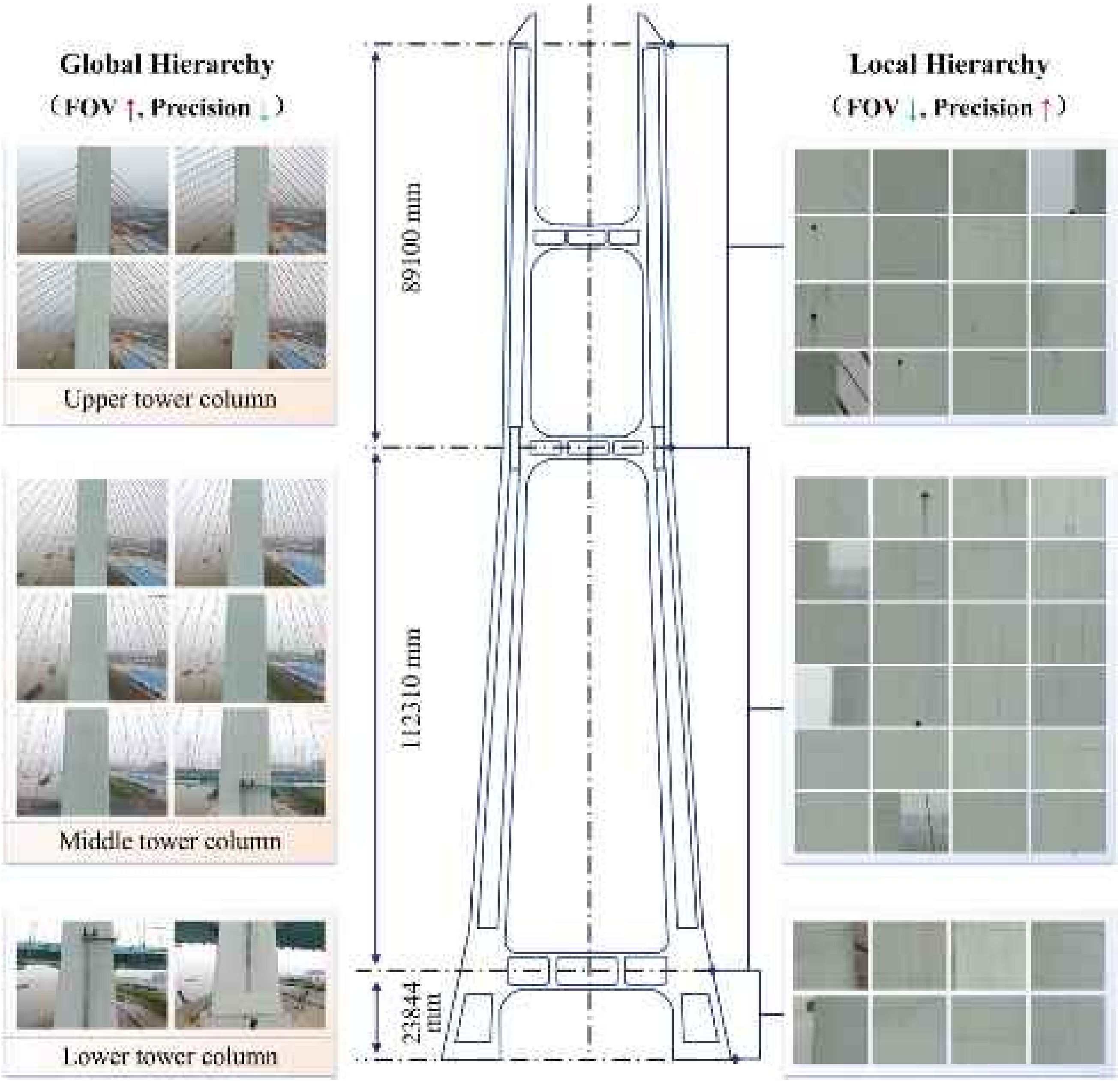


Fig. 10 Dual-mode camera acquisition results

restored, ensuring maximum accuracy and computational efficiency in the final stitched result.

The collaborative operation of wide-angle and zoom cameras is essential for acquiring comprehensive and detailed structural information. While the wide-angle camera excels at capturing the global view of the bridge tower, its limitations in resolving fine local details are effectively compensated by the zoom camera. To ensure completeness in data acquisition, adjacent images captured by the zoom camera are designed with a certain degree of overlap. However, in defect detection analysis, this overlap can lead to redundant computations, causing the same defect to be counted multiple times within overlapping regions. Such redundancy introduces statistical errors, increases computational overhead, and wastes resources. Therefore, eliminating overlapping regions between local images is crucial for improving both data processing efficiency and accuracy. This study proposes an image de-overlapping method based on LoFTR feature detection. The zoom camera scans row by row along the height of the bridge tower, beginning from the uppermost edge of each tower section. The removal process starts with eliminating horizontal overlaps. Subsequently, the method

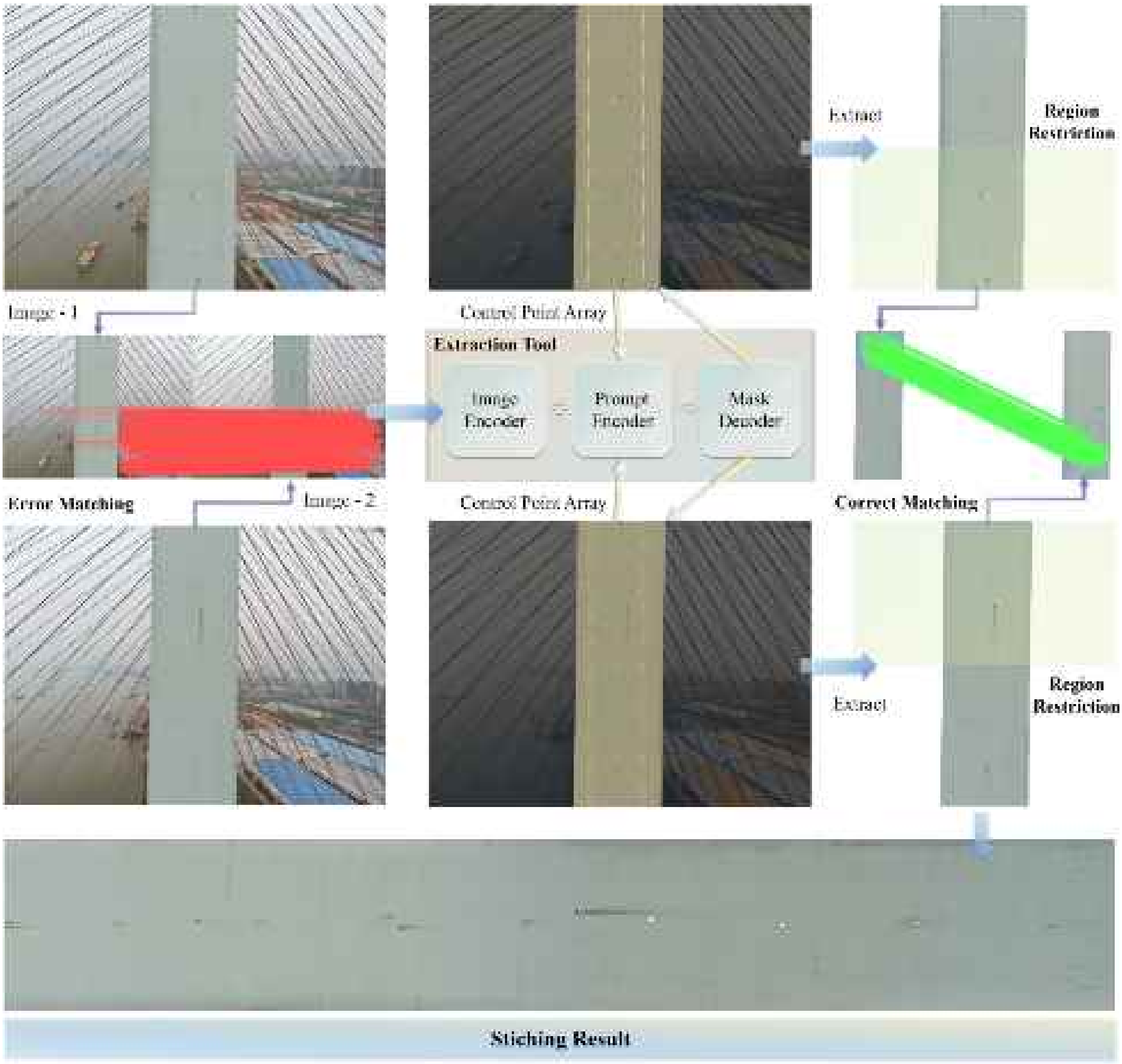


Fig. 11 Global image reconstruction process for upper tower column

progresses downward, detecting and eliminating vertical overlaps by comparing adjacent images to ensure the removal of redundant regions along the vertical axis. This results in a row-wise de-overlapping operation from top to bottom. Figure 13 presents a visual representation of the overlap removal process in both the horizontal and vertical directions for the upper, middle, and lower tower columns. By implementing this approach, a non-overlapping sequence of local images is constructed, providing a precise and accurate local view for subsequent structural analysis.

Given the spatial discreteness of the local image sequence, this study introduces a segmentation-based approach, treating each row as an independent inspection unit. Each row of local images is regarded as a distinct inspection segment, thereby establishing a coherent spatial framework within the sequence. The lower tower column comprises 22 inspection segments, the middle tower column is divided into 75 segments, and the upper tower column consists of 55 segments. Based on this segmentation, the panoramic composite images of the tower columns are evenly divided into corresponding segments, ensuring a one-to-one correspondence between local and global segments. Since the tower columns exhibit a certain inclination in space, their true longitudinal surface length can be derived

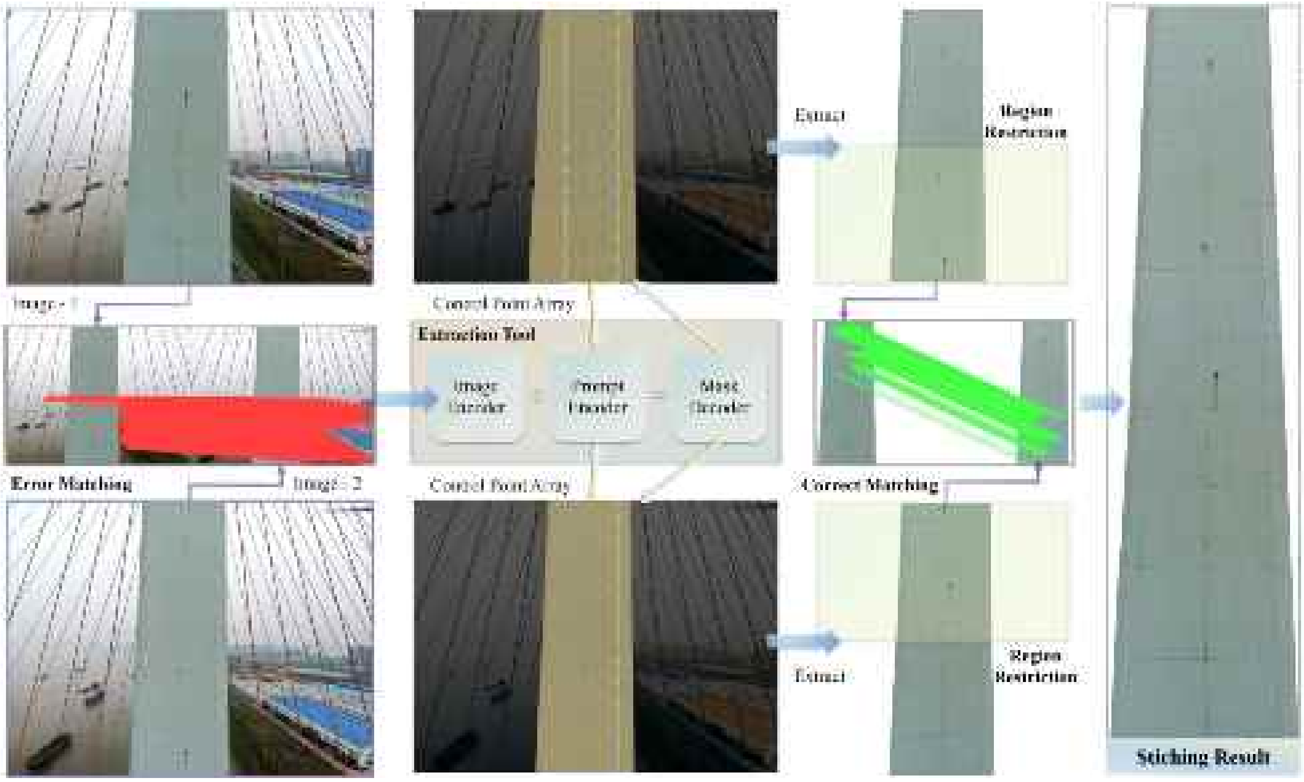


Fig. 12 Global image reconstruction process for middle tower column

based on this slope. However, considering the importance of height coordinates as a navigational reference, using the tower's spatial height rather than the converted longitudinal length as the segment's vertical coordinate proves more intuitive for engineering applications. Thus, this study adopts the center height of each segment as the global positioning reference, providing engineers with a preliminary spatial localization framework for rapid directional guidance. To ensure high alignment precision, a fine-tuning and optimization process is required between local and global segments. The effectiveness of this approach for the upper and lower tower columns is illustrated in Figs. 14 and 15. By comparing the computed position data with actual measurements, the results indicate an average error of 0.477%, achieving an overall positioning accuracy at the centimeter level. Future research will focus on developing automated, high-precision alignment methods to enable precise matching between local and global segments, thereby providing more accurate spatial localization services for engineering applications.

After constructing the global spatial information of the local images, all local image data are fed into the trained Light-YOLO-c model for a detailed analysis of defect types and quantities. The Light-YOLO-c model, a deep learning-based defect detection algorithm, has been specifically trained on tower column defect images, enabling it to efficiently and accurately identify various types of structural anomalies. The detection results include the type, location, and quantity of defects, providing engineering personnel with the ability to quickly pinpoint structural deficiencies while offering precise data support for subsequent maintenance efforts. To ensure the clarity and interpretability of the defect analysis results, selected visual representations and statistical summaries of different defect types are presented in Figs. 14 and 15. These visualizations distinctly mark the defect locations within each segment of the tower column and categorize the various defect types. By comparing defect distributions across different segments, the overall structural health of the tower column can be assessed, facilitating the early identification of potential risk areas. The results indicate that cracks account for a significant proportion of detected defects, and they are typically the most prevalent defect type within each segment. This observation holds critical implications for the long-term maintenance and operational management of the tower column. As cracks are a common form of structural damage, their timely repair is essential to maintaining the stability and safety of the tower. If left unaddressed, even minor cracks could escalate, posing structural risks. Therefore, engineering teams should prioritize the inspection and repair of segments with a high concentration of cracks, effectively mitigating potential hazards and ensuring the structural integrity of the tower column.

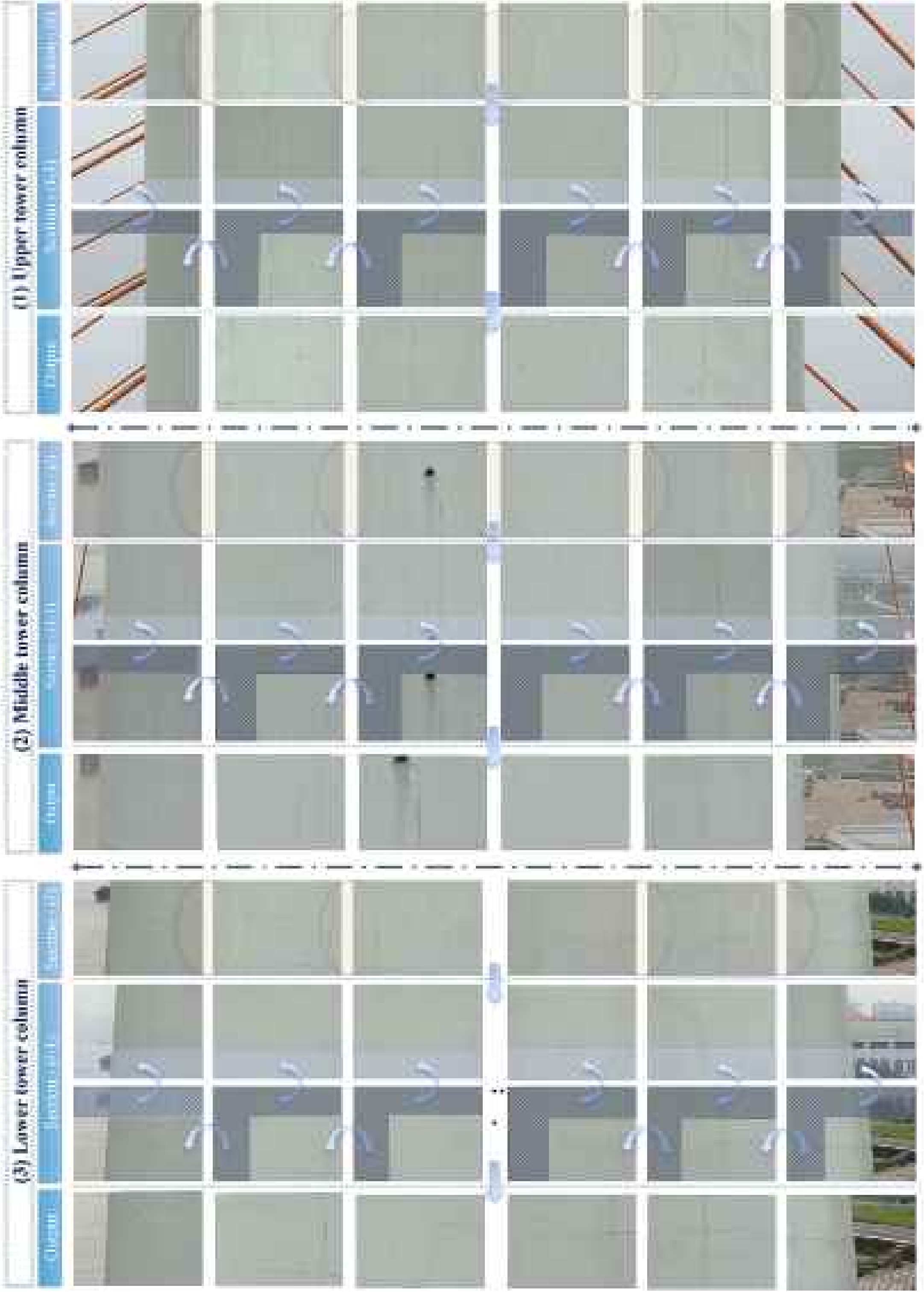
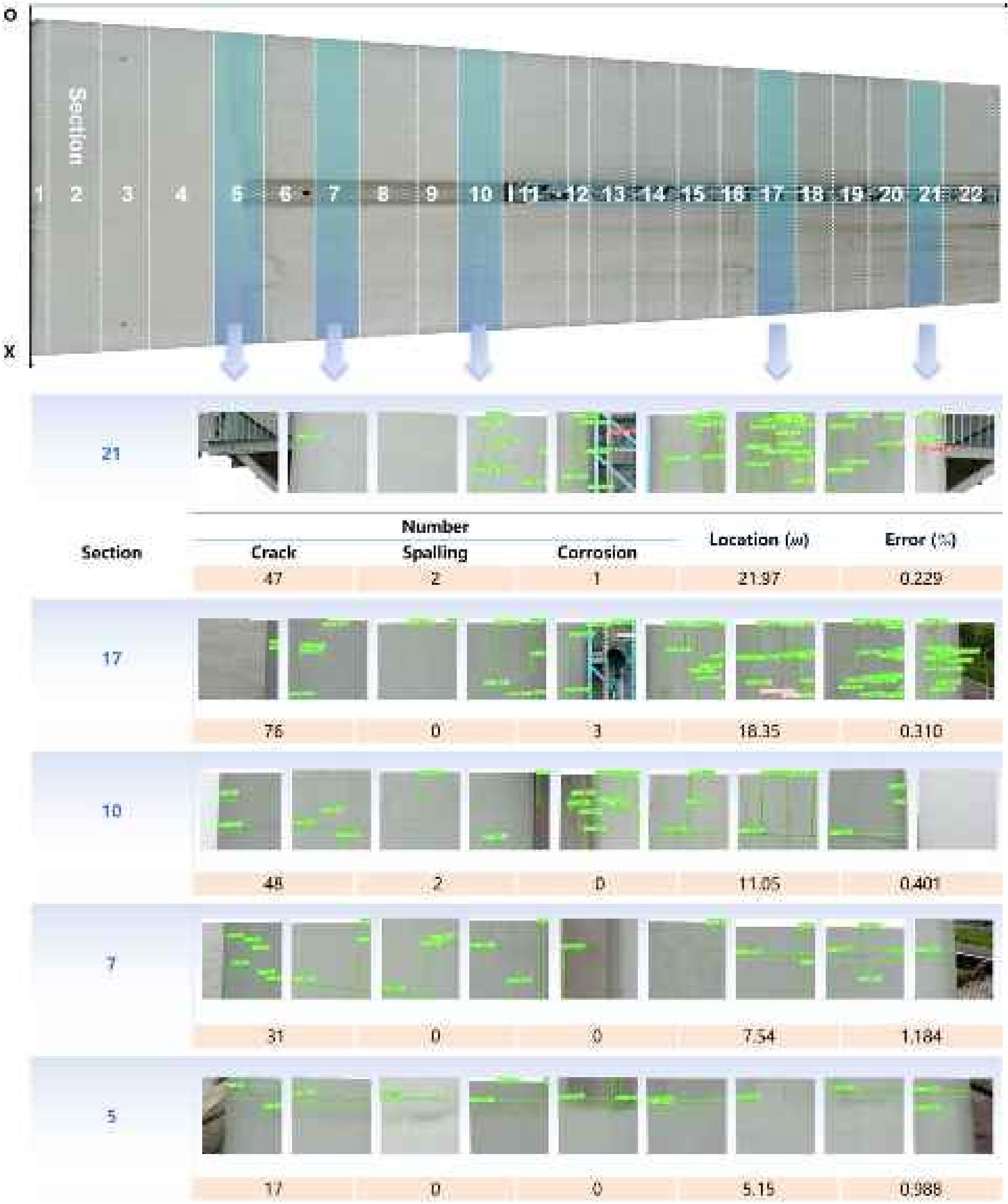


Fig. 13 Process of establishing inspection segments

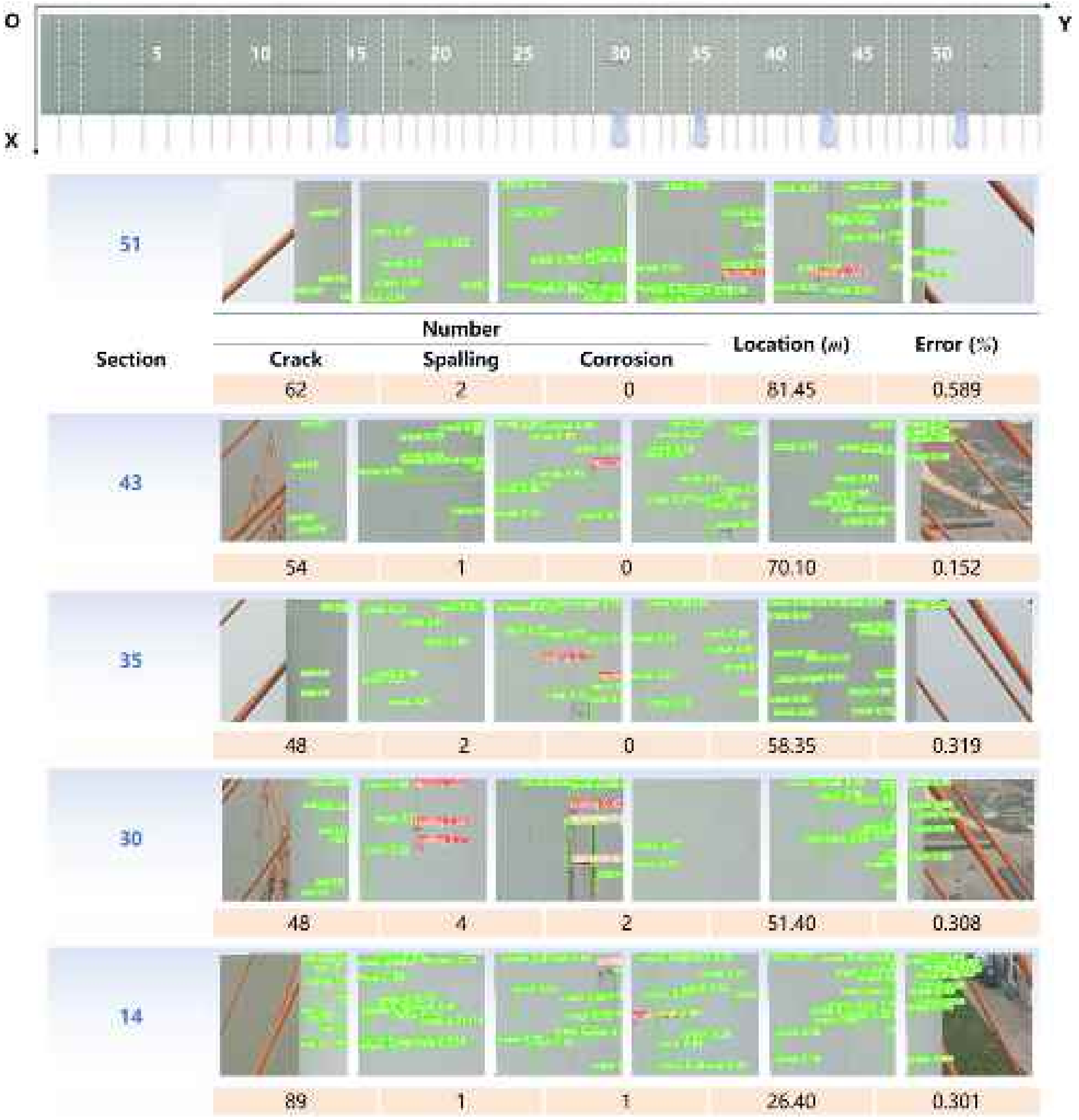
7 Conclusion

This paper addresses the critical challenges in inspecting high-tower bridge structures by introducing an innovative dual-mode drone camera inspection technology. This method employs a wide-angle camera to gather extensive structural data, facilitating the creation of a comprehensive global view, while the zoom camera captures detailed local defect



Section	Number			Location (m)	Error (%)
	Crack	Spalling	Corrosion		
21	47	2	1	21.97	0.229
17	76	0	3	18.35	0.310
10	48	2	0	11.05	0.401
7	31	0	0	7.54	1.184
5	17	0	0	5.15	0.988

Fig. 14 Visualization of the lower tower column analysis results



Section	Number Crack	 Spalling	 Corrosion	Location (*m*)	Error (%)
51	62	2	0	81.45	0.589
43	54	1	0	70.10	0.152
35	48	2	0	58.35	0.319
30	48	4	2	51.40	0.308
14	89	1	1	26.40	0.301

Fig. 15 Visualization of the upper tower column analysis results

information. By merging these two datasets, the proposed approach significantly enhances engineers' understanding of the bridge tower's operational state. The primary contributions of this study are as follows:

(1) Global View Construction: To overcome inefficiencies in image stitching caused by complex backgrounds, this research proposes a feature extraction-based global view construction method. This technique employs a targeted lattice guide for efficient extraction of tower features, coupled with deep feature matching to achieve high-precision global view reconstruction. The method optimizes both the efficiency and accuracy of image stitching, providing a robust solution for constructing global information of high-tower bridge structures.

(2) Zoom Image Data Fusion: This study introduces an innovative approach to handle redundancy and overlap in zoomed images, ensuring seamless integration with the global view. By employing deep feature extraction techniques and introducing segmentation concepts, the method successfully bridges the gap between detailed local images and the global perspective, enhancing the accuracy of structural inspections.

(3) Light-YOLO Detection Model: This study proposes the Light-YOLO model, a lightweight yet highly precise target detection algorithm, specifically designed for close-range, high-resolution image data. This model integrates advanced three-dimensional spatial-enhanced attention mechanisms to strengthen feature extraction, improving detection accuracy without compromising computational efficiency.

The proposed methodology is particularly applicable to high-tower bridge inspections, where it facilitates comprehensive, rapid, and precise defect detection across large-scale structures. The fusion of global and local data offers engineers a reliable foundation for early-stage maintenance planning and proactive structural health management, potentially reducing long-term repair costs and enhancing the safety of critical infrastructure. While the proposed method offers significant improvements, there are inherent limitations that must be addressed. For instance, environmental factors, such as high winds and low visibility, can still affect the UAV's performance, leading to image blurring or noise. The current focus on bridge towers limits the model's applicability to more complex infrastructure, requiring further research to enhance the model's versatility and performance across various types of structures. Future research will focus on enhancing the adaptability of the Light-YOLO model for different structural contexts, such as tunnels and pipelines, where defect types may overlap. Efforts will also be made to integrate multi-scale feature learning and context-aware fusion techniques to further improve detection capabilities across various image resolutions. Additionally, improving the UAV's stability under challenging environmental conditions, such as strong winds and low-light scenarios, remains an area for exploration to ensure consistent and reliable inspection results in diverse settings.

Appendix 1

Abbreviation	Explanation
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once
SAM	Segment Anything Model
LoFTR	Local Feature Matching with Transformers
ViT	Vision Transformer
RANSAC	Random Sample Consensus
IEDMs	Information Enhancement Down-sampling Modules
BN	Batch Normalization
EH-FEN	Efficient Hierarchical Feature Extraction Network
HMS-FFM	Hierarchical Multi-Scale Feature Fusion Module
CCA	Compact Coupled Attention

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

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request. Data will be shared in accordance with ethical guidelines and privacy regulations.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest regarding the publication of this paper. All authors have disclosed any financial or personal relationships that could be perceived as influencing the research presented in this study.

References

1. Deng L, Sun T, Yang L, Cao R (2023) Binocular video-based 3D reconstruction and length quantification of cracks in concrete structures. Autom Constr 148:104743. https://doi.org/10.1016/j.autcon.2023.104743
2. Kim B, Cho S (2019) Image-based concrete crack assessment using mask and region-based convolutional neural network. Struct Control Health Monit. https://doi.org/10.1002/stc.2381
3. Zhou K, Wang Z, Ni Y, Zhang Y, Tang J (2023) Unmanned aerial vehicle-based computer vision for structural vibration measurement and condition assessment: a concise survey. J Infrastruct Intell Resilience 2(2, 2023):100031. https://doi.org/10.1016/j.iintel.2023.100031
4. Jang K, An YK, Kim B, Cho S (2021) Automated crack evaluation of a high-rise bridge pier using a ring-type climbing robot. Comput-Aided Civil Infrastruct Eng 36(1):14–29. https://doi.org/10.1111/mice.12550
5. Yang L, Li B, Feng J, Yang G, Chang Y, Jiang B (2023) Jizhong xiao, Automated wall-climbing robot for concrete construction inspection. J Field Robotics 40:110–129. https://doi.org/10.1002/rob.22119
6. Lin T, Putranto A, Chen P, Teng Y, 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
7. Ding W, Shu J, Debono CJ, Prakash V, Seychell D, Borg RP (2024) Quantitative assessment of cracks in concrete structures using active-learning-integrated transformer and unmanned

robotic platform. Autom Constr 168:105829. https://doi.org/10.1016/j.autcon.2024.105829

8. Zhang C, Zou Y, Wang F, Castillo E, Dimyadi J, Chen L (2022) Towards fully automated unmanned aerial vehicle-enabled bridge inspection: Where are we at? Construct Build Mater 347:128543. https://doi.org/10.1016/j.conbuildmat.2022.128543
9. Li R, Yu J, Li F, Yang R, Wang Y, Peng Z (2023) Automatic bridge crack detection using unmanned aerial vehicle and faster R-CNN. Constr Build Mater 362:129659. https://doi.org/10.1016/j.conbuildmat.2022.129659
10. Liu Y, Nie X, Fan J, Liu X (2020) Image-based crack assessment of bridge piers using unmanned aerial vehicles and three-dimensional scene reconstruction. Comput-Aided Civil Infrastruct Eng 35(5):511–529. https://doi.org/10.1111/mice.12501
11. Yoon S, Jr BFS, Lee S, Jung HJ, Kim IH (2022) A novel approach to assess the seismic performance of deteriorated bridge structures by employing UAV-based damage detection. Struct Control Health Monit 29:7. https://doi.org/10.1002/stc.2964
12. 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
13. 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
14. Ding W, Yang H, Yu K, Shu J (2023) Crack detection and quantification for concrete structures using UAV and transformer. Autom Constr 152:104929. https://doi.org/10.1016/j.autcon.2023.104929
15. Mohan A, Poobal S (2018) Crack detection using image processing: a critical review and analysis. Alexandria Eng J 57(2):787–798. https://doi.org/10.1016/j.aej.2017.01.020
16. Adhikari RS, Moselhi O, Bagchi A (2014) Image-based retrieval of concrete crack properties for bridge inspection. Autom Constr 39:180–194. https://doi.org/10.1016/j.autcon.2013.06.011
17. Talab AMA, 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
18. Dias-da-Costa D, Valença J, Júlio E, Araújo H (2017) Crack propagation monitoring using an image deformation approach. Struct Control Health Monit 24:e1973. https://doi.org/10.1002/stc.1973
19. Alzubaidi L, Zhang J, Humaidi AJ, Dujaili AA, Duan Y, Shamma OA, Santamaría J, Fadhel MA, Amidie MA, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:53. https://doi.org/10.1186/s40537-021-00444-8
20. Xu D, Xu X, Forde MC, Caballero A (2023) Concrete and steel bridge structural health monitoring—insight into choices for machine learning applications. Constr Build Mater 402:132596. https://doi.org/10.1016/j.conbuildmat.2023.132596
21. Cha YJ, Choi W, Büyüköztürk O (2017) Deep learning-based crack damage detection using convolutional neural networks. Comput-Aided Civil Infrastruct Eng 32(2):361–378. https://doi.org/10.1111/mice.12263
22. Zhang Y, Yuen K-V (2021) Crack detection using fusion features-based broad learning system and image processing. Comput-Aided Civil Infrastruct Eng 36(2021):1568–1584. https://doi.org/10.1111/mice.12753
23. Laxman KC, Tabassum N, Ai L, Cole C, Ziehl P (2023) Automated crack detection and crack depth prediction for reinforced concrete structures using deep learning. Constr Build Mater 370:130709. https://doi.org/10.1016/j.conbuildmat.2023.130709
24. Chang S, Zheng B (2024) A lightweight convolutional neural network for automated crack inspection. Constr Build Mater 416:135151. https://doi.org/10.1016/j.conbuildmat.2024.135151
25. Cha YJ, Choi W, Suh G, Mahmoudkhani S, Büyüköztürk O (2018) Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Comput-Aided Civil Infrastruct Eng 33:731–747. https://doi.org/10.1111/mice.12334
26. Yu Z, Shen Y, Shen C (2021) A real-time detection approach for bridge cracks based on YOLOv4-FPM. Autom Constr 122:103514. https://doi.org/10.1016/j.autcon.2020.103514
27. Ni F, He Z, Jiang S, Wang W, Zhang J (2022) A generative adversarial learning strategy for enhanced lightweight crack delineation networks. Adv Eng Inform 52:101575. https://doi.org/10.1016/j.aei.2022.101575
28. 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
29. Chen W, Zhang J (2024) Efficient and lightweight monitoring network for cracks in complex background regions based on adaptive perception. Autom Constr 166:105614. https://doi.org/10.1016/j.autcon.2024.105614
30. Song F, Liu B, Yuan G (2024) Pixel-level crack identification for bridge concrete structures using unmanned aerial vehicle photography and deep learning. Struct Control Health Monit 1299095:14. https://doi.org/10.1155/2024/1299095
31. Ma J, He Y, Li F, Han L, You C, Wang B (2024) Segment anything in medical images. Nat Commun 15:654. https://doi.org/10.1038/s41467-024-44824-z
32. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W, Dollár P, Girshick R (2024) Segment anything, arXiv:2304.02643, https://doi.org/10.48550/arXiv.2304.02643. (accessed 22 July 2024)
33. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D (2023) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):87–110. https://doi.org/10.1109/TPAMI.2022.3152247
34. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2024) An image is worth 16x16 words: transformers for image recognition at scale, arXiv:2010.11929, https://doi.org/10.48550/arXiv.2010.11929. (accessed 15 June 2024)
35. Sun J, Shen Z, Wang Y, Bao H, Zhou X (2021) LoFTR: detector-free local feature matching with transformers. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp 8918–8927. https://doi.org/10.1109/CVPR46437.2021.00881.
36. Dai X, Chen Y, Yang J, Zhang P, Yuan L, Zhang L (2021) Dynamic DETR: end-to-end object detection with dynamic attention. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp 2968–2977. https://doi.org/10.1109/ICCV48922.2021.00298
37. Zhou J, Yang D, Song T, Ye Y, Zhang X, Song Y (2024) Improved YOLOv7 models based on modulated deformable convolution and swin transformer for object detection in fisheye images. Image Vis Comput 144:104966. https://doi.org/10.1016/j.imavis.2024.104966
38. He Z, Jiang S, Zhang J, Wu G (2022) Automatic damage detection using anchor-free method and unmanned surface vessel. Autom Constr 133:104017. https://doi.org/10.1016/j.autcon.2021.104017
39. Cheng S, Song J, Zhou M, Wei X, Pu H, Luo J, Jia W (2024) EF-DETR: A lightweight transformer-based object detector with an encoder-free neck. IEEE Trans Ind Inform 20(11):12994–13002. https://doi.org/10.1109/TII.2024.3431044
40. Hua W, Chen Q, Chen W (2024) A new lightweight network for efficient UAV object detection. Sci Rep 14:13288. https://doi.org/10.1038/s41598-024-64232-z

41. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp 779–788. https://doi.org/10.1109/CVPR.2016.91
42. Wang CY, Yeh IH, Liao HYM (2024) YOLOv9: learning what you want to learn using programmable gradient information, arXiv:2402.13616, https://doi.org/10.48550/arXiv.2402.13616. (accessed 20 July 2024)
43. Jiang S, Wu Y, Zhang J (2023) Bridge coating inspection based on two-stage automatic method and collision-tolerant unmanned aerial system. Autom Constr 146:104685. https://doi.org/10.1016/j.autcon.2022.104685

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.