npj Heritage Science

https://doi.org/10.1038/s40494-026-02681-7

Article in Press

CWADE-Net: a deep learning framework for vegetation invasion and brick spalling defect detection on Nanjing Ming City Wall

Received: 14 February 2026

Accepted: 14 May 2026

Cite this article as: Yuan, X., Wang, N., Wang, Y. *et al.* CWADE-Net: a deep learning framework for vegetation invasion and brick spalling defect detection on Nanjing Ming City Wall. *npj Herit. Sci.* (2026). https://doi.org/10.1038/s40494-026-02681-7

Xianglong Yuan, Nannan Wang, Yuliang Wang, Shenglan Du, Mingming Sui, Yueqian Shen, Shihuan Li, Ziyou Wang, Dong Chen, Jiju Poovvancheri & Liqiang Zhang

CWADE-Net: A Deep Learning Framework for Vegetation Invasion and Brick Spalling Defect Detection on Nanjing Ming City Wall

Xianglong Yuan[a], Nannan Wang[b], Yuliang Wang[c,*], Shenglan Du[d], Mingming Sui[a], Yueqian Shen[e], Shihuan Li[a], Ziyou Wang[a], Dong Chen[a,h], Jiju Poovvancheri[f] and Liqiang Zhang[g]

[a] *College of Civil Engineering, Nanjing Forestry University, Nanjing, 210037, China*
[b] *Nanjing City Wall Protection and Management Center, Nanjing, 210001, China*
[c] *School of Computer and Information Engineering, Chuzhou University, Chuzhou, 239000, China*
[d] *Faculty of Architecture and the Built Environment, Delft University of Technology, Delft, 2628 BL, the Netherlands*
[e] *School of Earth Sciences and Engineering, Hohai University, Nanjing, 211100, China*
[f] *Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS B3H 3C3, Canada*
[g] *Department of Geography, State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, 100875, China*
[h] *Jiangsu Highway Intelligent Detection and Low-Carbon Maintenance Engineering Research Center, Nanjing Forestry University, Nanjing, 210037, China*

ARTICLE INFO

Keywords:
Nanjing Ming Dynasty City Wall
Cultural heritage preservation
Deep neural network
Disease detection
Vegetation invasion
Brick spalling

ABSTRACT

Focusing on the cultural heritage of the Nanjing Ming Dynasty city wall, this study presents CWADE-Net, a deep learning framework specially designed to detect surface defects arising from herbaceous/woody and vine-type vegetation invasion, as well as brick spalling. Its novelty lies in the capability to address challenging conditions such as uneven illumination, complex backgrounds, and large defect-scale variations. CWADE-Net jointly integrates illumination enhancement, edge information encoding, and spatial–frequency feature extraction in its backbone to improve feature representation. Its neck employs bidirectional feature fusion to enhance multi-scale semantic interaction. Moreover, we adopt a lightweight detection head that enables real-time model deployment. Experiments on images acquired using a Nikon D300, iPhone 15 Pro Max, and DJI Matrice 4E demonstrate mAP50 scores of 82.4%, 87.9%, and 54.8% for three defect types, outperforming mainstream methods by 5–12%, thus effectively supporting intelligent monitoring, conservation, and World Cultural Heritage nomination efforts.

Introduction

The Nanjing City Wall of the Ming Dynasty, one of the most outstanding representatives of the ancient Chinese city defense system, is the longest, largest, and best-preserved ancient city wall complex remaining in the world [1]. It was built in the first year of the Hongwu reign of Emperor Zhu Yuanzhang of the Ming Dynasty (AD 1368), with its construction continuing for over 20 years. The wall was built in accordance with the hills and rivers, stretching around 35 kilometers, which has served as a defensive structure and a symbolic landmark of the Ming's capital, Nanjing. As the core component of the urban system, "the palace city, the imperial city, the capital city", of Ming's early period, it carries significant historical memory and urban cultural information. Furthermore, it provides invaluable materials and inspiration for follow-up research on the premodern urban planning, defense system evolution, and construction technologies of ancient China. The Nanjing Ming Dynasty city wall adopts a composite structure, with rammed earth as the core and brick and stone as the outer layer. Therefore, it yields strong durability and high adaptability to the climate. Its brick surfaces often bear inscriptions indicating the production kilns, supervising officials, and kiln craftsmen. Under the strict accountability system, also known as "mandatory attribution of workmanship", implemented during the Hongwu reign of the Ming Dynasty, each brick was required to document the identities of responsible personnel at multiple administrative levels, including the titles and names of officials at the prefectural, subprefectural, and county

*Corresponding author
xianglongyuan@njfu.edu.cn (X. Yuan); wangn1211@126.com (N. Wang); ylw@chzu.edu.cn (Y. Wang); shenglan.du@tudelft.nl (S. Du); mingmingsui@njfu.edu.cn (M. Sui); y.shen_lidar@hhu.edu.cn (Y. Shen); shihuan.li@njfu.edu.cn (S. Li); ziyouwang@njfu.edu.cn (Z. Wang); chendong@njfu.edu.cn (D. Chen); jiju.poovvancheri@smu.ca (J. Poovvancheri); zhanglq@bnu.edu.cn (L. Zhang)

levels, as well as grassroots organizations and the actual producers. These inscriptions provide direct evidence of the Ming Dynasty's official kiln system, labor regime, and local social organization, which are of significant value for archaeological and art-historical research. However, after more than 600 years of natural weathering, foundation settlement, rainwater erosion, and human activities, the wall has suffered from varying defects, including brick spalling, cracking, salt precipitation, weathering, peeling, and vegetation invasion. Among these complicated factors, the natural environmental change (*e.g.*, precipitation, humidity, and weathering) can severely weaken the strength of bricks and bonding materials, while foundation deformation, groundwater seepage, and plant root invasion may further intensify the structural damage. Currently, brick detachment and collapse have already emerged in certain sections of the city wall, seriously threatening its structural stability and historical integrity. How to achieve rapid, accurate, and non-contact automated detection of the city wall defect (without damaging its original structure) has become a key scientific challenge in the protection and restoration work. This enables structural safety assessment, digital documentation, and long-term health monitoring of the heritage. Furthermore, such technical diagnosis can also be extended to other large-scale brick-and-stone heritages, such as the Ming Dynasty city walls of Xi'an and Jingzhou.

Traditional defect detection of built heritage structures primarily relies on human inspection, which is the earliest method and still remains widely used. Professionals identify surface defects such as cracks, peeling, and weathering based on visual inspection, photogrammetry, and domain expertise. Although straightforward and low-cost, this method suffers from low efficiency and high subjectivity, limiting its capacity for large-scale, continuous, and quantitative city wall monitoring. With the development of non-destructive testing (NDT) technology, imagery acquired from infrared thermography (IRT) and ground-penetrating radar (GPR) has been gradually adopted for the detection of structural defects in cultural heritage. IRT monitors surface temperature distributions to indirectly infer the internal cavities, collapse, and humidity accumulations. It can identify both superficial and internal damages induced by physical or chemical processes that take place within the material, and has been widely applied in the diagnosis and assessment of historical monuments [2, 3]. Nevertheless, the performance of IRT is highly sensitive to environmental factors such as temperature, wind speed, and solar radiation. On the other hand, GPR uses high-frequency electromagnetic signals with strong penetration to identify internal cracks and cavities, and thus can be utilized to monitor interior structural issues of cultural heritage buildings [4, 5]. However, GPR's resolution is greatly influenced by material properties and antenna frequency, limiting its scalability to large-scale heritage scenes. Moreover, the accurate interpretation of GPR signals relies on human expert experience. Recently, light detection and ranging (LiDAR) has emerged as a new NDT technique that has been increasingly used in the research of heritage sites. LiDAR acquires dense three-dimensional (3D) point clouds under non-contact conditions, facilitating surface reconstruction, geometric deformation analysis, and weathering damage detection with high spatial resolution and strong visual effects. 3D LiDAR point clouds have been extensively investigated for the automated detection of cultural heritage surface defects, such as the bulging of the Beijing Forbidden City wall [6] and the stone weathering deterioration of the medieval Zsámbék Church in Hungary [7]. However, due to the massive point volume, complex data processing, and insufficient automation in defect detection, the use of LiDAR technology is still restricted when extended to large-scale city wall monitoring applications. Additionally, a number of studies also seek the technique of Interferometric Synthetic Aperture Radar (InSAR) to achieve high spatiotemporal-resolution monitoring of surface subsidence [8], abrupt displacements [9], as well as structural changes of the built heritage.

From the methodological point of view, the rapid and widespread development of Artificial Intelligence (AI) in recent years has greatly revolutionized the digital interpretation, analysis, and conservation of cultural heritage sites. How to apply AI, deep learning, and computer vision technologies for the automated detection of city wall defects has become an increasingly important research direction [10]. Deep learning is capable of learning powerful feature representations and capturing various types of wall defects from raw input data. Compared to traditional manual or physical inspection methods, deep learning networks achieve automatic feature extraction through end-to-end training, allowing them to automatically recognize objects from imagery [11], classify remote sensing scenes [12], assess image quality [13], detect anomalies in videos [14], and analyze diseases in vegetation [15] and buildings [16]. Specifically, in the field of heritage conservation, many recent researchers have explored deep learning tools, including tasks such as classification, segmentation, and object detection networks, for automated defect detection of cultural heritage structures such as weathering, surface cracks, vegetation, and joint damage, leading to remarkable progress in this direction [17].

Deep classification networks assign an entire input image to one of the predefined classes, making them suitable for quick and accurate defect recognition of cultural heritage sites on the image level. Hatir et al. [18] trained a classification network to recognize eight types of rock weathering (*e.g.*, flaking, contour scaling, cracking, differential

erosion, black crust, efflorescence, higher plants, and graffiti) of historical stone monuments located in Konya. A similar approach was adopted by Karimi et al. [19] to classify various brick defects, such as erosion, flaking, and salt efflorescence, focusing on Isfahan historical bridges. D'Orazio et al. [20] trained a convolutional neural network (CNN) classifier to recognize the microalgae and cyanobacteria growth process on historical brick facades. Amin et al. [15] implemented a multi-head external attention mechanism and integrated synthetic data for training to improve the accuracy of pine wilt disease classification. Focusing on Portuguese cultural heritage buildings, Karimi et al. [21] integrated the deep classifier MobileNet into an object detection network to achieve binary classification of damaged and intact tiles on heritage wall surfaces. Besides, classification networks can be extended to other data modalities. Seo et al. [22] developed a CNN classifier to identify brick cracking of masonry heritage buildings from IRT imagery. Using voxel map representations of 3D built heritage structures as input, Muñoz-Silva et al. [23] identified severe structural damage of historical buildings through a 3D CNN classification network. Nevertheless, classification networks can only generate predictions on the image level. They often require additional data processing, such as slide-window-based techniques, to achieve accurate localization of heritage structure defects [24].

More recently, many research works have leveraged object detection networks, including CNN, Faster R-CNN [25], and YOLO-family networks, to precisely detect and localize surface defects of cultural heritage sites. For example, Hacıefendioğlu et al. [26] employed ResNet50 [27] in combination with Grad-CAM [28] to detect cracks in historical brick-and-stone structures, verifying the effectiveness of CNN architectures in crack detection. Ali [29] adopted Faster R-CNN to detect defects of brick-and-stone structures, automatically recognizing cracks and erosion on the ancient temple walls of Ayutthaya, Thailand, from unmanned aerial vehicle (UAV) imagery. Pathak et al. [30] also utilized Faster R-CNN to specifically detect cracking and spalling on the surfaces of this temple heritage. They further localized the detected surface damage on the corresponding 3D models. Wang et al. [31] integrated Faster R-CNN with a smartphone-based webcam system to achieve real-time damage detection for historic masonry structures. Faster R-CNN was also used in stone-structured cultural heritage for the automated detection of damages such as cracks, detachment, and brick loss [32]. Besides Faster R-CNN, YOLO-based networks have also been widely used for the digital monitoring and conservation of cultural heritage. For instance, YOLOv4 was applied to identify surface defects from imagery data in the Shanhaiguan section of the Great Wall [33] and the buffer zone of the world heritage in Macau [34]. Later, the YOLOv5 model was leveraged to detect wall damage such as brick loss, exposure, and spalling on historic structures of the Dadi-Poti tombs in New Delhi [35] and limestone castles in the Loire Valley [36]. Focusing on the vegetation invasion disease, Guo [37] proposed Re-YOLO, which improves the disease detection accuracy of the YOLO baseline through the attention mechanism and multi-scale feature fusion. Based on YOLOv8 [38], Zhang et al. [39] performed non-destructive inspection of surface defects, such as weathering and plant invasion, on the gray-brick walls of ancient houses in Fuzhou. Singh et al. [40] precisely detected superficial damages in the form of cracks, spalling, and vegetation on the Darbhanga Fort, the historic center of Darbhanga, Bihar. Adopting the more recent YOLOv12 model, Zhang et al. [41] achieved accurate detection of multiple types of defects of the ancient Jingzhou city wall, including missing bricks, cracks, weathering, and vegetation invasion. Long et al. [42] constructed MSD-Det, a comprehensive dataset of heritage masonry structure images. It benchmarked 17 object detection networks, demonstrating that YOLO-based algorithms consistently outperformed Faster R-CNN in surface damage detection.

Despite their promising performance, object detection networks are generally limited to coarse localization of heritage defects. In contrast, instance segmentation networks, represented by Mask R-CNN [43], enable more precise, pixel-level localization and delineation of structural damage. Thus, they have been widely used for the detection and localization of defects in cultural heritage sites worldwide. Hatır et al. [44] used Mask R-CNN to map and segment deteriorations of Yazılıkaya monuments in the Hattusa archaeological site in Turkey, covering defects such as biological colonization, cracks, higher plants, and missing parts. Similarly, Saravanan and Bhaskar [45] explored Mask R-CNN for segmenting vegetation defects at the instance level within historical stone structures of World Heritage sites across India. Vandenabeele et al. [46] combined CNN with the traditional watershed segmentation technique to precisely segment individual bricks on the Basilica of St Anthony in Padua, Italy, which enables automated large-scale surveying of such historic masonry structures. Chao et al. [47] integrated the DeepLabV3+ backbone and the self-attention mechanism to accurately segment vegetation invasion of two species: *Erigeron annuus* (L.) Pers. and *Erigeron canadensis* L. at the pixel level. Wang et al. [48] adopted a transformer-based architecture, incorporating a lightweight parameter-free attention mechanism for precise damage segmentation of historical building components such as doors and windows. Similarly, Liu et al. [49] proposed an MEP network, which integrated the Efficient Channel Attention (ECA) for lightweight disease detection of the Great Wall brick surfaces.

Unlike the rapid development and significant advancements in automated defect detection and quantitative inspection of cultural heritage assets worldwide, currently, there is relatively limited research for systematic identification and detection of surface defects in the Nanjing Ming Dynasty city wall. For example, Jin et al. [8] applied the permanent scatterers InSAR technology to monitor the continuous surface subsidence and deformation of Nanjing Ming Dynasty city wall. Li et al. [1] quantitatively evaluated the influence of vegetation on the weathering deterioration of the city wall through field measurements and numerical modeling. With a focus on city wall crack defects, Wu et al. [50] performed visual inspection on the 3D city wall model reconstructed from the UAV oblique photogrammetry to manually identify cracks with lengths greater than 30 cm. Wang et al. [51] proposed an automated approach for crack extraction from point cloud data of the Nanjing city wall using Euclidean clustering. While prior studies have verified the effectiveness of deep learning techniques, especially segmentation networks and object detection networks, in detecting surface damage of other ancient city wall structures, such as the Jinzhou city wall [52] and the Great Wall [33, 49], the application of advanced deep learning methodologies for the inspection, analysis, and long-term monitoring of surface defects in the Nanjing Ming Dynasty city wall remains insufficiently explored.

Furthermore, even though all the aforementioned studies have demonstrated the promising performance of deep learning techniques for the rapid detection and recognition of cultural heritage defects, most approaches rely on general-purpose architectures (*e.g.*, YOLO-based frameworks) without task-specific adaptation or optimization for heritage defect detection. Only a limited number of studies have introduced parameter-free or lightweight attention modules [48, 49] to enhance network efficacy. However, these improvements are insufficient to ensure robust defect detection under challenging conditions, such as poor illumination, complex backgrounds, and significant variations in defect scales, hindering their reliable applications in real-world heritage conservation scenarios. Therefore, there is still a pressing need for developing advanced deep learning methodologies specifically tailored for the robust and accurate detection of surface defects on the Nanjing Ming city wall under challenging environments, to better support its long-term monitoring, preservation, and conservation as a critical cultural heritage asset.

To this end, we have introduced the city wall anomaly detection network (CWADE-Net), a specialized deep learning framework for automated defect detection on the Nanjing Ming Dynasty city wall. In particular, we focus on detecting three representative types of surface defects: herbaceous/woody vegetation invasion, brick spalling, and vine-type vegetation invasion. Considering that the Nanjing Ming Dynasty city wall retains around 25.091km (the longest ancient defensive wall worldwide), we take cross-platform 2D imagery acquired from the digital camera and the mobile device to enable rapid, large-scale capture of the wall's structural features.

Specifically, the novelty of our proposed CWADE-Net lies in its capability to address common challenges in image processing, such as uneven illumination, complex backgrounds, and variations in object scale. To achieve this, we have designed and optimized a series of network modules, significantly enhancing CWADE-Net's robustness against uneven illumination on the wall surface, blurred defect edges along brick-spalling regions, complex textures of brick surfaces, and varying scales of brick defect areas. In the backbone of CWADE-Net, a self-calibrated illumination network, SCI-Net [53], is introduced to enhance feature representations under low-light or non-uniform illumination conditions, which adjusts image brightness and contrast through iterative brightness estimation and adaptive calibration. We also construct an edge information encoding (EIE) module to extract multi-scale edge cues, fusing them with semantic features at the corresponding scales to improve feature learning near vegetation-invasion contours and brick-spalling regions. Besides, a spatial–frequency feature extraction module, named C3k2-FSM, is designed to jointly extract spatial and frequency features. Complementary incorporation of these two features enables fine-grained texture extraction, which improves CWADE-Net's capability of learning discriminative feature representations in complex scenes. Furthermore, we introduce a bidirectional feature fusion module in the network neck to facilitate the interaction and fusion between high-level semantic feature maps and low-level local feature maps. Combined with a lightweight detection head, CWADE-Net effectively integrates multi-scale features while maintaining a balance between accuracy and computational efficiency. Experimental results on the Nanjing Ming Dynasty city wall dataset demonstrated that CWADE-Net outperforms mainstream approaches, including YOLOv8 [38], YOLO11 [54], RetinaNet [55], DINO [56], and Faster R-CNN [25], in detecting vegetation invasion and brick-spalling defects. Specifically, CWADE-Net exhibited notably stronger performance and robustness in low-light and complex-texture conditions. It also generalized well to unseen UAV imagery, demonstrating its robustness and scalability in large-scale city wall heritage scenes. Overall, our research aims to practically recognize and assess large-scale defects in historic city walls using deep learning techniques. The proposed CWADE-Net provides an efficient and reliable technical solution for the intelligent inspection and detection of the Nanjing Ming Dynasty city wall. It directly supports downstream conservation,

restoration, documentation, and long-term monitoring of culturally significant built heritage such as ancient city walls, assisting researchers, structural engineers, and decision-makers in quantitative heritage preservation efforts.

The remainder of this paper is organized as follows. The Methods section introduces CWADE-Net's methodological details, including the backbone for feature extraction, the neck for feature aggregation and fusion, the detection head for defect detection, and the supervision loss terms. The Results section presents the implementation details, qualitative results achieved by CWADE-Net, and its quantitative comparison with other mainstream object detection approaches. A detailed ablation study is further provided to validate the effectiveness of CWADE-Net's architectural design. The Discussions section comprehensively analyzes CWADE-Net's robustness to image sizes, scales, varying image qualities, and cross-platform imagery. Last, we conclude the paper with reflections on the proposed CWADE-Net, limitations, and directions for future work.

Methods

Due to its long construction history, the Nanjing Ming Dynasty city wall exhibits varying types of defects, such as vegetative invasion and brick spalling, which not only affect the wall's appearance but may also threaten its structural safety. In this study, we propose a deep learning-based methodology for detecting these defects solely from RGB images acquired by various devices. More specifically, we acquired on-site images along the northern section of the Nanjing Ming city wall, which starts from Xuanwu Gate and extends through Jiefang Gate and Taiping Gate to Fugui Mountain. Image data were collected using both a digital camera (Nikon D300) and a mobile device (iPhone 15 Pro Max) for training and prediction, with a focus on vegetation-invasion damage and missing-brick defects. The DJI Matrice 4E UAV was also used to acquire wide-angle images of the city wall for defect prediction. The image acquisition covered various illumination conditions (including early mornings, middays, and cloudy days) as well as multiple viewpoints (including front, side, and top-down perspectives) to enhance sample diversity and scene representativeness. We annotated all images with GT bounding boxes using the LabelImg tool [57]. Annotation quality assurance was performed through manual inspection and cross-checking. The wall surface defects involved three categories: herbaceous/woody vegetation invasion, brick spalling, and vine-type vegetation invasion. Woody invasion is dominated by paper mulberry (*Broussonetia papyrifera*) and cutleaf chaste tree (*Vitex negundo var. cannabifolia*), while the main invasion herbaceous invaders are carpetgrass (*Arthraxon prionodes*), spider brake fern (*Pteris multifida*) and holly fern (*Cyrtomium fortunei*). Vine-type invasion mainly involves creeping fig (*Ficus pumila*) and trumpet vine (*Campsis radicans*). Invasion by these plants commonly occurs through seed dissemination, vegetative root expansion, or human-mediated introduction, thereby causing diverse effects across physical, chemical, and biological domains. Brick palling, as one of the main weathering defects, can compromise the structural security of the wall and undoubtedly further impair the integrity of the ancient city wall and shorten its lifespan. These three defects threaten the structural health of the Nanjing Ming Dynasty city wall, which has driven us to prioritize their detection in the proposed methodology.

We acquired 3206 valid images, splitting them into training, validation, and test sets following an 8:1:1 ratio. Since vine-type vegetation samples are relatively sparse in the dataset, this can lead to class imbalance and biased network training, ultimately harming the detection performance. We augmented the images that contain vine-type vegetation by using horizontal and vertical flipping, scaling transformations, and adjustments to color saturation. After data augmentation, the training set contained 3473 herbaceous/woody vegetation instances, 11,356 brick-spalling instances, and 3518 vine-type vegetation instances.

Accurate detection and localization of such defects currently face several challenges, including but not limited to uneven illumination, complex scene backgrounds, large-scale variations in defect regions, and blurred edge details. To address these issues, this paper proposes CWADE-Net, a defect detection network targeting three common types of wall surface defects: herbaceous/woody vegetation invasion, brick spalling, and vine-type vegetation invasion. CWADE-Net consists of three components: a backbone, a neck, and a detection head, which are described as follows:

- **Backbone**. CWADE-Net incorporates the modules of SCI-Net, C3k2-FSM, and EIE, focusing on image-quality enhancement under low-light conditions, spatial-frequency feature extraction, and edge detail augmentation, respectively. Together with these modules, CWADE-Net is able to learn robust feature representations for the precise defect detection in brick-spalling and vegetation-invasion areas.

- **Neck**. We design a bidirectional feature fusion module, combining the top-down and bottom-up network information flows to effectively fuse the high-level, abstract semantic features with low-level, detailed local

features. We also integrate a lightweight attention module, the convolutional block attention module (CBAM) [58], that guides the network to focus on the most discriminative feature channels and spatial regions for enhancing surface defect detection of the Nanjing Ming Dynasty city wall.

- **Detection head**. To further reduce the network's complexity, we propose to perform cross-scale convolution with shared parameters and independent batch normalization. This design effectively restricts the number of network parameters, improving the overall inference efficiency while maintaining the detection accuracy.

The architectural design of CWADE-Net highly aligns with the feature characteristics of vegetation attachment and brick spalling on the Nanjing Ming Dynasty city wall. With the proposed modules, the network can better capture textures and edge details under low-light conditions, learn more discriminative features by focusing on specific defect regions, and achieve efficient inference through the lightweight detection head. Overall, our method provides robust technical support for defect identification and intelligent conservation of the Nanjing Ming Dynasty city wall. Fig. 1 provides an overview of the CWADE-Net architecture, where the blue region at the top presents input images of various wall-surface defect scenes, the grey region illustrates the CWADE-Net backbone embedded with SCI-Net, C3k2-FSM, and EIE modules, the pink region denotes the neck performing top-down and bottom-up multi-scale feature fusion, the yellow region represents the detection head at different scales for target prediction, and the green region shows detection results at each scale as well as the final detection outputs through non-maximum suppression (NMS). In the following subsections, we present technical details of CWADE-Net's backbone, neck, detection head, and supervision losses, respectively.

Backbone

Images of the Nanjing Ming Dynasty city wall commonly suffer from lighting issues, such as uneven illumination, backlighting, shadow occlusion, and low light in certain regions. All these issues can significantly impair the network's capability to correctly detect brick detachment, vegetation invasion, and other types of surface wall defects. In particular, images captured under natural lighting conditions often contain defect areas that appear blurred or indistinct due to insufficient illumination, which can lead to reduced detection accuracy. To address this problem, we introduce the SCI-Net module [53] at the initial stage of the CWADE-Net backbone for image enhancement. SCI-Net models illumination components using multi-scale Gaussian functions, which effectively smooths the brightness and enhances local details in the frequency domain. While preserving the overall visualization of the original image, SCI-Net improves the visibility of low-light defect regions. In the meantime, it suppresses overexposure in high-reflection areas and enhances local structural details that are often obscured, such as crack edges, detachment textures, and weathered surface areas. By introducing SCI-Net, the subsequent network layers receive more discriminative, higher contrast visual cues that are robust against low-light conditions and illumination variations.

SCI-Net hypothesizes that the relationship between the low-light image y and its desired clear image z can be expressed using the following equation:

$$y = z \otimes x \tag{1}$$

where x denotes the illumination component, and $\otimes$ denotes element-wise multiplication. Insufficient illumination during image acquisition often leads to blurred edges near brick-detachment regions and texture loss in vegetation-invasion areas. The adopted SCI-Net enhances these obscured details by iteratively estimating the illumination component. It consists of an illumination estimation module F and a self-calibration module G. The illumination estimation module F is iteratively updated using the following equation:

$$F(x^t) = \begin{cases} u^t = H_\theta\left(x^t\right) \\ x^0 = y \\ x^{t+1} = x^t + u^t \end{cases} \tag{2}$$

where y represents the original low-light image, $H_\theta(\cdot)$ is a mapping operator used to estimate the illumination residual. This mapping operator is parametrized by θ, which is shared across all stages. u^t and x^t denote the residual and the illumination estimate at the t^{th} stage, correspondingly.

A self-calibration module G is applied in each iteration to encourage alignment across multi-stage outputs. By comparing the current illumination estimate x^t and the original image y, G generates a calibration signal s^t and adds it back to the original image. In this way, it provides an input image that better approximates the true illumination for the next stage of illumination residual estimation. Since both G and F share parameters across stages, the training process continuously guides the outputs from all stages toward a unified target, ultimately ensuring consistency among these output feature maps. In network testing, only a single F module is retained for fast inference. The self-calibration module G is iteratively updated as follows:

$$G(x^t) = \begin{cases} z^t = y \oslash x^t \\ s^t = K_\theta(z^t) \\ v^t = y + s^t \end{cases} \tag{3}$$

where y is the original low-light image, x^t is the illumination at the t^{th} stage, and the symbol $\oslash$ denotes element-wise devision. $K_\theta(\cdot)$ represents a self-calibration operator with its parameters shared across all stages. s^t converts the illumination residual at the t^{th} stage into a corrective term added to the original image, and v^t represents the input for the next stage.

During training, SCI-Net progressively refines and optimizes the illumination map through multiple iterations of the module F and the module G. In each iteration, F predicts the illumination residual from the current illumination estimate and updates the illumination map. Then, G generates a calibration signal based on the ratio between the original low-light image and the updated illumination map, adding it back to the original image to provide a more accurate input for the next iteration of F. This iterative calibration and residual refinement process progressively recovers local edge details of brick-spalling regions and fine-grained textures in vegetation-occupied areas, which enables the output of each iteration to approximate the optimal enhancement result. We have collected the images of the Nanjing Ming Dynasty city wall as input for the network training. Supervision loss is applied to the illumination map at every iteration to improve the robustness and generalizability of the network across diverse low-light scenes. Due to that multiple iterations of the modules F and G have encouraged outputs from all stages to converge toward a unified target, in the network testing, only a single F module is needed for image enhancement. Therefore, when deployed for real-world solutions, CWADE-Net enables fast restoration of image edges and texture details under minimal computational cost, providing high-quality visual input for subsequent detection and localization of surface defects on the Nanjing Ming Dynasty city wall.

Besides, accurate detection of defects on the Nanjing Ming Dynasty city wall also heavily relies on rich edge information. However, brick-spalling regions are often characterized by discontinuities or even breaks in edges, while vegetation-invasion areas are characterized by dense contours and rapid texture changes on the wall surface. Therefore, it is important to enhance the network's capability in perceiving edge information for improved wall defect detection. Meanwhile, due to the complexity of the background scenes, which may exhibit varying textures, dust deposits, weathering degradation, or illumination changes, directly extracting edges from the raw images often introduces substantial feature noise to the network. As a result, it could harm the training stability and robustness of the network. To address this issue, we incorporate an EIE module [59] into the CWADE-Net backbone to extract salient edge features, as shown in Fig. 2. EIE consists of two parts: the edge-information generation component and the edge-information fusion component. The edge information generation component is applied after the shallow convolutional layers, leveraging shallow feature maps that preserve fine-grained local structures while suppressing irrelevant background noise to extract multi-scale edge information. Multi-scale edge information is extracted to enhance edge feature representation while mitigating background interference. These edge features are then injected into the mid- and high-level feature maps of the network and fused with semantic features across multiple scales, achieving global edge feature enhancement and maintaining the feature consistency across various scales.

As shown in Fig. 2, the edge-information generation component contains a dual-channel Sobel operator [60], which performs edge detection along two spatial directions in the X and Y axes of the input image. The Sobel-X operator and the Sobel-Y operator are applied to extract spatial gradient features along the horizontal direction and the vertical direction, respectively. We combine the outputs from the two operators through element-wise addition to obtain the edge-response map. This dual mechanism can effectively extract edge features enriched with discriminative geometric structural information. To best preserve the edge sharpness and prevent its diminishment during network downsampling, we adopt the max pooling strategy. Max pooling retains the highest activation in a local image patch,

which can effectively emphasize responses of edges while filtering out background noise. Therefore, compared to average pooling or other downsampling strategies, max pooling is more suitable for edge information preservation. Subsequently, a 1×1 convolution is applied to reshape the dimension of the multi-scale edge features, aligning with the backbone. The edge information generation process enhances the edge feature representation along vegetation-invasion contours and brick-spalling surface boundaries, while significantly reducing irrelevant background noise.

To enhance the feature representation of the wall surface defects, the EIE module further performs the edge information fusion operation. A 1×1 convolution is applied to fuse the edge features extracted from the Sobel operator with the convolutional features from the backbone network at the same scale, enabling interconnection between edge information and texture information. Besides, to strengthen the perception of structural changes on the wall surface, a 3×3 convolution is applied to capture the spatial context of the fused features. Another 1×1 convolution is then employed to reshape the fused features to match the channel dimensions required by subsequent network layers, ensuring that edge information can be effectively merged and utilized by the network. As shown in Fig. 1, three EIE modules are embedded in the backbone, responsible for extracting and fusing edge features across multiple resolutions, thus enhancing the overall edge-awareness capability of CWADE-Net.

In YOLO11, the C3k2 block is adopted to achieve faster and more efficient feature aggregation through channel compression and local convolution. However, the representational capacity of C3k2 highly depends on the local receptive field, resulting in limited selective responses to different frequency components. Thus, it may fail to identify weak edges and repetitive fine texture patterns in complex scenes, especially when confronting uneven illumination, weathered textures, or dense vegetation. Prior studies have shown that joint feature extraction from the spatial and frequency domains can significantly enhance the recognition of geometric and texture details [61]. Driven by this insight, we design the FreqSpatial module for spatial-frequency feature extraction. As shown in the zoomed-in view of Fig. 3(a), the module of FreqSpatial consists of two branches corresponding to space-domain feature extraction and frequency-domain feature extraction. The spatial branch employs Scharr-X and Scharr-Y operators [62] to extract horizontal and vertical gradients of the input features to obtain edge-response maps. Besides, a 3×3 convolution is applied to capture spatial texture features. These two outputs are then merged to enhance the structural details of edges and contours. The frequency branch transforms the input features into the frequency domain through two-dimensional fast Fourier transform (FFT2D) [63]. Then, it performs convolution in the frequency domain to extract high-frequency textures and low-frequency structural information. The extracted features are projected back to the original domain using a two-dimensional inverse fast Fourier transform (IFFT2D) [63], followed by an additional convolution operator to refine local details. The outputs of the two branches are merged through element-wise addition, leading to high-quality feature embeddings that well preserve both edge sharpness and fine-grained textures. As a result, the network is enhanced for more accurate detection and localization of wall surface defects under complex backgrounds.

By integrating the concept of joint spatial–frequency domain feature extraction into the C3k2 block, we design the C3k2-Freqspatial (C3k2-FSM) module. The Bottleneck unit in the original C3k2 block is replaced by the Freqspatial unit, which combines the efficient spatial feature aggregation with the frequency-domain feature extraction. Through the dual-branch fusion of the features at the unit level, information from both domains is aligned and complemented within a single computational module.

The C3k2 block uses a Boolean parameter *c3k* to control its complexity. If *c3k* is *false*, n Bottleneck sub-units participated in local feature extraction with a relatively low number of parameters and computational cost. If it is *true*, n C3k sub-units are applied instead, each consisting of several Bottleneck micro-units fused with bypass features to further enhance the spatial and contextual feature learning. However, this incurs more computational cost. In our network, the original Bottleneck units are replaced with the FreqSpatial units when *c3k* is *false* (Fig. 3(a)), and they are replaced with the C3k-FreqSpatial units when *c3k* is *true* (Fig. 3(b)). Given that the shallow feature maps in the backbone (see Fig. 1) have relatively high spatial resolution, enabling C3k-FreqSpatial (*i.e.*, *c3k* is *true*) at these stages can significantly increase computational cost and memory usage. Moreover, as shallow layers mainly focus on extracting low-level visual cues such as edges and textures, the lightweight FreqSpatial module (*i.e.*, *c3k* is *false*) is sufficient for stable feature extraction. Adopting overly complex modules may amplify pseudo-textures or lead to overfitting, due to the fact that shallow layers are more vulnerable to background noise from shadows, weathered patterns, or other artifacts. On the contrary, feature maps in deeper layers have lower spatial resolutions but more enriched semantics. We apply the multi-branch C3k-FreqSpatial modules at these stages to achieve effective global awareness with only a slight increase in computation, providing more precise high-level representations for the subsequent network feature learning. Based on the above analysis, we set *c3k* to *false* for the first three C3k2-FSM modules embedded in the backbone, while *c3k* is set to *true* only for the final C3k2-FSM module in the backbone.

Neck

To achieve high-quality feature integration in the complex scenes of the Nanjing Ming Dynasty city wall, we employ a bidirectional feature fusion strategy at the neck of CWADE-Net, which combines the network information flows in both the top-down and bottom-up directions. A CBAM attention block is inserted after each fusion module to further suppress irrelevant textures and highlight defect-related regions. On the one hand, our proposed fusion strategy progressively injects high-level semantics into low-level local features. On the other hand, it propagates shallow edge and texture cues back to high-level feature embeddings, enhancing the overall network representation capability for improved detection of both small-scale vegetation-invasion regions and large-scale brick-spalling regions.

Specifically, starting from the high-level feature map with the lowest resolution and most enriched semantics, we upsample it to $40\times40\times1024$ to align with the feature map at the mid-level stage. It is then passed through a C3k2-FSM block to reduce its channel dimension to 512, yielding a $40\times40\times512$ mid-level feature map. We further upsample this feature map to 80×80 to spatially match the high-resolution feature representation at the low level, followed by another C3k2-FSM block to adjust the channel dimension to 256, resulting in an $80 \times 80 \times 256$ low-level feature map. This bottom-up strategy progressively injects deep semantic information into lower-level feature maps while leveraging C3k2-FSM blocks to extract spatial-frequency features, which effectively suppresses background texture noise and enhances structural edges.

In the top-down direction, we take the feature map obtained after the second convolution of the backbone and downsample it using a 3×3 convolution with a stride of 2, obtaining a $80\times80\times256$ feature map. Another downsampling is performed to further reduce its dimensionality to $40 \times 40 \times 512$, and finally to $40 \times 40 \times 1024$. This top-down pathway propagates local-level edge and texture information with high spatial resolution to high-level feature maps, allowing deep features at the latent space to maintain strong semantics with improved structural continuity and edge discriminability.

Feature fusion is performed through channel concatenation at three different network stages. At the low level with high spatial resolution (80×80), we concatenate the top–down features, the bottom-up features, and the backbone features of the same level to obtain an $80 \times 80 \times 1024$ fused representation. Analogously, at the middle level of 40×40, and the high level of 20×20 feature maps, we fuse the three streams of features to form a $40 \times 40 \times 1536$ representation and a $20 \times 20 \times 2048$ representation. The fused features at these three stages are processed by CBAM modules [58] to recalibrate channel responses and enhance spatial saliency. They are further reshaped through C3k2-FSM blocks to the dimensions of $80 \times 80 \times 256$, $40 \times 40 \times 512$, and $20 \times 20 \times 1024$, respectively, for the follow-up defect detection. Through top-down and bottom-up feature fusion at the neck, CWADE-Net enables deep semantic information and shallow structural, textural details to mutually complement each other. Furthermore, the introduction of the attention mechanism (*i.e.*, CBAM module) significantly enhances the channel selectivity and spatial saliency, ultimately improving the discriminative capacity of the obtained feature representations.

Detection head

The detection head is a key component in object detection networks, as it directly bridges high-level semantic features with the final prediction outputs. Therefore, its design highly influences the computational efficiency and the detection performance. A common practice in the YOLO family of networks is to construct a single convolutional branch for each detection scale, independently performing feature refinement, bounding-box regression, and category prediction. Though this strategy effectively leverages the spatial resolution and semantics at different scales for object detection, it introduces overly complex computation and redundant parameters. Especially, in multi-scale detection tasks, these replicated branches significantly increase the model size and inference latency, which prevents the efficient deployment of the network on resource-restricted devices.

To address this limitation, CWADE-Net incorporates a lightweight detection head, the lightweight shared convolutional with separate BN detection (LSCSBD) [64]. The core idea of LSCSBD is to perform shared convolution and independent batch normalization. Specifically, a single set of convolution kernels is reused across feature maps at multiple scales to achieve cross-scale parameter sharing, and independent batch normalization is performed for each scale to ensure that different feature maps undergo separate normalization processes. Through this design, LSCSBD preserves detection accuracy while significantly reducing the number of parameters and the model complexity, thereby saving computational cost and memory consumption to achieve efficient inference.

As illustrated in Fig. 1, we pass the obtained feature maps at different scales to the corresponding LSCSBD heads P3 through P5. Among them, P3 receives the feature map with the highest spatial resolution and is mainly used to detect small targets, such as fine vegetation fragments and minor surface defects. P4 achieves a balance between spatial

detail and semantic abstraction, which allows it to recognize mid-scale defects, such as local brick damage and partial structural loss. P5 receives the feature map with the strongest semantics and the lowest spatial resolution. Therefore, it is suitable for detecting large-scale brick spalling and severe wall surface defects.

Given that feature maps at different scales vary in their channel dimensions and statistical distributions, LSCSBD first uses a 1×1 convolutional module, Conv_BN, to align the channel dimensions for the input features. These aligned features are then fed into a 3×3 convolutional layer, Conv_GN, wherein a single set of convolutional kernels is shared across scales. In this way, LSCSBD enables cross-scale parameter sharing, reduces redundancy, and greatly improves computational efficiency in multi-scale detection tasks. To prevent statistical interference among feature maps of different scales during normalization, LSCSBD preserves branch-specific batch normalization and activation modules (BNAct). By keeping independent batch normalization in each branch, it encourages better numerical stability and more reliable training convergence. Moreover, an additional 3×3 convolutional layer, Conv_GN, followed by branch-specific BNAct layers, is applied to further enlarge the effective receptive field, enhance feature representations, and maintain training stability. At the output stage, two convolutional branches are used at each scale: Conv_Reg performs bounding-box regression by employing a scale-specific parameter to adaptively rescale and re-calibrate the shared convolutional features; Conv_Cls is dedicated to category prediction, performing NMS [65] (Fig. 1) to rank the candidate boxes by descending confidence. It iteratively retains the highest-scoring box and removes those with an intersection over union (IoU) greater than 0.6 until all candidates have been evaluated. The remaining boxes are the final detection results.

Overall, the LSCSBD head not only keeps the multi-scale object perception capability of the original YOLO networks, but also reduces a significant amount of parameters through shared convolution and independent batch normalization. Its design effectively simplifies the network architecture and lowers the computational cost, facilitating the efficient inference, deployment, and practical usability of the detection framework. We incorporate the LSCSBD detection head into our proposed CWADE-Net for its lightweight structure and efficiency. In particular, it is well-suited for real-time wall surface defect detection tasks on embedded devices or edge-computing platforms, where it enables a highly compact and efficient detection pipeline without compromising the detection performance.

Loss functions

Similar to the YOLO11 object detection network, we propose to supervise CWADE-Net using the classification loss, distribution focal loss, and box loss. We provide a detailed explanation of the adopted loss terms as follows:

First, we adopt the classification loss (CLS) to supervise the network to correctly assign each detected bounding box to one of the predefined object categories. Focal loss [55] is used to alleviate the class imbalance issue in this classification task. In the training set, sample instances of brick-spalling defect greatly outweigh the samples of herbaceous/woody vegetation defect and vine-type vegetation defect. Under such an imbalance, a standard cross-entropy loss often misleads the model to be biased toward the majority class. Contrarily, focal loss uses an adjustment factor to down-weight major-class samples and assign higher relative weights to minor-class samples, which enhances the network's capability and robustness in recognizing underrepresented defect types. The classification loss is formulated using the following equation:

$$L_{cls} = -\Sigma_{i=1}^{S \times S} \Sigma_{j=1}^{B} \Sigma_{c=1}^{C} \mathrm{II}_{ij}^{obj} \omega_c (1 - \hat{y}_{ijc})^{\gamma} y_{ijc} \log(\hat{y}_{ijc}) \tag{4}$$

where L_{cls} denotes the obtained classification loss, S is the number of divided grids of the input image along the height/width direction, B is the pre-defined number of anchor boxes, C is the number of categories of defect types, II_{ij}^{obj} represents an exponential function (if the j^{th} anchor box in the i^{th} grid contains a defect object, then $\mathrm{II}_{ij}^{obj} = 1$), ω_c is the weight of the corresponding class c, γ is the focus parameter for hard samples, y_{ijc} is the ground truth (GT) label, and $\hat{y}_{ijc}$ denotes the network's prediction probability.

Then, for bounding-box regression, we adopt the distribution focal loss (DFL) to achieve precise boundary localization of bounding boxes in the defect detection task. The wall surface defect often exhibits complex-shaped boundaries in brick-spalling regions and fragmented patterns in vegetation-invasion areas. By utilizing DFL, we discretize continuous boundary distances into a group of bins and supervise the distribution across adjacent bins with soft labels. The final continuous distance is recovered by computing the expectation of the predicted distribution. This loss achieves more precise regression of surface defect boundaries, effectively reducing discrepancies between predicted boxes and the actual defect regions. The formulation of DFL is given as follows:

$$L_{dfl} = -\Sigma_{i=1}^{N} \Sigma_{c=1}^{4} y_{ic} (1 - p_{ic})^{\gamma} \log(p_{ic}) \tag{5}$$

where L_{dfl} represents the distribution focal loss term, N is the total number of samples, c corresponds to the four dimensions (*i.e.*, x, y, w, and h) of the bounding box, y_{ic} is the GT box label, p_{ic} is the predicted probability of the box, and γ denotes the hyperparameter of the adjustment factor in the focal loss term.

Furthermore, to obtain the optimal geometric correspondence between predicted bounding boxes and GT boxes, CWADE-Net formulates the box loss by calculating complete intersection over union (CIoU). We integrate the discrepancies in the overlap area, center-point distance, and width-height ratio between box predictions and GT boxes into the loss term. This refined box loss encourages the predicted boxes to align more closely with the true targets in both position and geometric shape. The box loss is formulated using the following equation:

$$L_{box} = \lambda_{scale}(1 - IoU + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha \mathrm{v}) \tag{6}$$

where L_{box} denotes the box loss term, IoU represents the intersection over union between the predicted box and the GT box, $\rho^2(b, b_{gt})$ is the Euclidean distance between the predicted box and the GT box, c denotes the diagonal length of the smallest enclosing rectangle that fully covers both the predicted box and the GT box, α is the adaptive weight, and υ describes the width-height ratio consistency between the predicted box and the GT box. λ_{scale} is a dynamic weight defined as follows:

$$\lambda_{scale} = 1 - \frac{\omega^g h^g}{2WH} \tag{7}$$

where $\omega^g h^g$ represents the GT defect box area. W and H are the width and the height of the input image.

CWADE-Net combines the three loss terms, L_{cls}, L_{dfl}, and L_{box} to simultaneously perform bounding box regression and box semantic classification. The final loss term is defined as the weighted sum of the three losses:

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{box} L_{box} + \lambda_{dfl} L_{dfl} \tag{8}$$

where L_{total} is the total loss. λ_{cls}, λ_{box}, and λ_{dfl} are the weights to balance the corresponding loss terms, which are set to 0.5, 7.5, and 1.5, respectively.

Results

Evaluation metrics

We adopted five evaluation metrics to assess the quantitative performance and computational efficiency of CWADE-Net, including precision, recall, mean Average precision (mAP), number of parameters, and floating point operations (FLOPs).

Precision refers to the proportion of predicted positive box detections that are truly positive, and is often used to indicate the reliability of the model's predictions. Its computation is given in Eq. 9:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{9}$$

where TP denotes the number of positive samples that are correctly detected by the network, also referred to as true positives. FP represents the number of negative samples that are incorrectly classified as positive, known as false positives. In the Nanjing Ming Dynasty city wall surface defect detection task, TP refers to instances where the network successfully detects and localizes actual defects, and FP refers to instances where the network mistakenly labels healthy regions as defects.

Recall measures the proportion of actual object instances that are correctly predicted by the network, serving as an indicator of the network's completeness. We compute the recall metric using the following equation:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{10}$$

where FN is the number of positive samples that are incorrectly classified as negative, referred to as false negatives. In our task, FN represents the defect instances that the network fails to detect.

mAP is one of the most commonly used evaluation metrics in object detection, which assesses the model's detection accuracy for different categories under various IoU thresholds. A higher mAP indicates that the model achieves better accuracy and stability when handling multiple target classes. In practice, mAP is often computed at an IoU threshold of 0.5, denoted as mAP50. Its calculation is given as follows:

$$mAP50 = \frac{1}{N}\Sigma_{i=1}^{N} AP_i \tag{11}$$

where N is the total number of categories, and AP_i is the average precision of the i^{th} category.

The number of parameters refers to the total number of trainable weights and biases in a deep learning network, commonly used to measure the model's computational complexity. For CWADE-Net, the parameter number corresponds to the sum of all trainable parameters across the entire network.

FLOPs represents the number of floating-point computations involved in a single forward inference pass, often used to indicate the network's computational cost. It is mainly determined by the network architecture (*e.g.*, the use of convolutional layers and fully connected layers). The computation of FLOPs is given in Eq. 12:

$$FLOPs = 2 \times \Sigma(I \times O \times G \times T) \tag{12}$$

where I represents the spatial dimension of the input feature map to the convolutional layer, O is the spatial dimension of the output feature map, G refers to the size of the convolutional kernel, and T is the number of channels of the input feature map.

The experiments were performed on an Ubuntu 20.04 operating system, using an NVIDIA GeForce RTX 3090 GPU. We used Python 3.8 and PyTorch 1.1 to build the network, leveraging CUDA 11.3 for GPU-accelerated computation. Regarding hyperparameters, we trained the network for 300 epochs with a batch size of 32. Stochastic Gradient Descent (SGD) was chosen as the optimizer. The learning rate was initialized as 0.01 and gradually converged to 0.0001. Besides, we set the momentum as 0.937 and the weight decay as 0.0005. The size of input image patches was set to 640×640.

Qualitative evaluation

To evaluate the effectiveness of CWADE-Net in complex city-wall scenes, we performed qualitative inspection using real-world data collected from the Nanjing Ming Dynasty city wall, with results shown in Fig. 4. The experiments covered a wide range of challenging detection conditions, such as low-illumination shadowed regions, large contiguous brick spalling, co-occurrence of vegetation and missing bricks, wall surfaces intruded by vine-type vegetation, and small-scale defects. By combining the detection results with the output heatmaps, we can visually observe the network's focus regions and prediction behaviour under different scenarios.

It can be observed from Figs. 4(a)–(c) that the network successfully identifies brick-spalling edges and vine-type vegetation, even under low-light conditions. The heatmaps show concentrated high responses in defect regions and low responses in the rest regions, indicating that the network has robust discriminative capability against insufficient illumination. This performance robustness is mainly attributed to the proposed SCI-Net, which optimizes the brightness and contrast to better preserve structural and textural details of the input image, thus allowing accurate defect detection in weak-light environments. Figs. 4(c)-(e) present the network's detection results of brick-spalling defect, which demonstrate that the network can effectively detect relatively large missing-brick regions formed by multiple connected bricks. The corresponding heatmaps also show continuous high responses that spatially well align with the defect areas. This effectiveness results from that the EIE module significantly enhances long edge details of the large brick-spalling regions at the shallow layers, benefiting high-level feature extraction at deeper layers. Meanwhile, the C3k2-FSM blocks further strengthen the long edges of defect regions, which facilitates more reliable detection of large, connected, and continuous brick-spalling defects. Figs. 4(f) and (g) visualize some scenes where vegetation invasion and brick spalling coexist. The network can successfully detect both types of defect. Also, the heatmaps exhibit strong, distinct responses in both the vegetation-covered and brick-spalling regions, demonstrating robust inter-class discrimination. This performance benefits from the adopted CBAM module in feature fusion, which enhances feature channels highly relevant to defects in the channel dimension while emphasizing the spatial positions of vegetation and missing bricks.

Table 1
Quantitative comparison of disease detection models for the Ming city wall. FPS (frames per second) indicates the number of images processed per second, and latency indicates the time required to process each image.

Methods	P (%)	R (%)	mAP50 Invasion of herbaceous/ woody vegetation(%)	mAP50 Brick spalling(%)	mAP50 Invasion of vine-type vegetation(%)	mAP50 (%)	GFLOPs	Parameters (M)	Latency (ms/img)	FPS (img/s)
Retinanet	62.6	63.0	78.9	78.0	31.4	62.8	210.0	36.5	29.2	34.3
Dino	69.0	62.4	76.0	77.6	45.1	66.2	274.0	47.6	76.3	13.1
Faster R-CNN	72.4	64.5	80.8	80.0	45.8	68.9	208.0	41.3	30.6	32.7
YOLOv8n	67.0	65.0	78.4	81.6	48.4	69.5	8.1	3.2	4.7	214.4
YOLO11n	73.6	63.3	80.6	79.6	50.9	70.4	6.3	2.6	3.9	256.9
CWADE-Net	73.0	68.1	82.4	87.9	54.8	75.0	11.4	5.6	7.6	131.2

Thus, the network can attend to and differentiate between the two defect categories within complex scenes. Figs. 4(a)–(b) and (g)–(k) show representative examples of the city wall surface covered by climbing vegetation. The network remains capable of detecting multiple defect types in such complex environments, including herbaceous/woody vegetation invasion, brick-spalling defect, and vine-type vegetation invasion. The corresponding heatmaps exhibit precise, concentrated responses along defect boundaries. Notably, in these regions covered by climbing vegetation, the network can distinguish between vegetation-induced defects and brick-spalling defects. Besides, the high-response areas in the heatmaps are not restricted to the vines but precisely align with the edges of the missing bricks. This suggests the effectiveness of the proposed CWADE-Net in correctly identifying multiple defect types when confronting complex background interference. Figs. 4(k)-(n) include small-scale surface defects, such as fine vegetation and minor brick spalling. The network can successfully detect those small targets, with the heatmaps showing clear and concentrated high-response regions, which indicates the network's strong sensitivity to fine-grained surface defects. This capability primarily arises from the EIE module that enhances small edge details in shallow layers and propagates them to deeper layers for high-level feature abstraction. Moreover, the bidirectional feature fusion strategy effectively integrates low-level edge and texture cues with high-level semantic information, facilitating accurate detection and localization of small-scale defects within complex scenes.

Based on the qualitative analysis, our proposed CWADE-Net has demonstrated strong detection performance under various challenging scenes, including low-light conditions, contiguous brick-spalling defects, coexisting defect types, complexly textured surfaces, and small-scale defect areas. The detection results, together with the corresponding heatmaps, mutually validate that the network maintains high robustness and reliability in complex environments.

Quantitative comparison

To quantitatively evaluate the effectiveness of the proposed CWADE-Net on the city wall surface defect detection task, we compared it with multiple mainstream object detection networks, including YOLOv8 [38], YOLO11, RetinaNet [55], Faster R-CNN [25], and DINO [56]. Experiments were conducted on the same dataset under the same training configuration to achieve a fair comparison. Table 1 presents the quantitative comparison, while Fig. 5 provides the visual comparison of the detection performance of these networks under several representative scenes.

From the quantitative comparison results in Table 1, CWADE-Net achieved a precision of 73.0% and a recall of 68.1%, surpassing all competing networks in both metrics. For the three defect types, herbaceous/woody vegetation invasion, brick spalling, and vine-type vegetation invasion, the network achieved the mAP50 of 82.4%, 87.9%, and 54.8%, respectively. It also gained the highest overall mAP50, demonstrating a significant performance improvement in the task of ancient city-wall surface defect detection. These quantitative findings further aligned with the visual detection results shown in Fig. 5. Specifically, Fig. 5(a) presents a scene where vegetation defect overlaps on top of the brick spalling regions. Compared to other methods, CWADE-Net could accurately detect and localize the occluded brick-spalling defect despite vegetation coverage. In Figs. 5(b) and (c), Faster R-CNN and RetinaNet were likely to misclassify irrelevant twigs or small root structures on the brick surface as vine-type vegetation defects. On the contrary, CWADE-Net greatly reduced such false detections. Besides, in Fig. 5(d), both YOLO11 and DINO failed to completely detect large-scale, contiguous brick-spalling defects, while the detection results of CWADE-Net closely matched the GT missing-brick regions. This is mainly attributed to the fact that the EIE module better preserves the

long-edge features in the shallow layers and passes them to deeper layers. At the same time, the use of C3k2-FSM blocks also enhances the long edges of defect regions, which allows the network to robustly recognize large, contiguous areas of brick spalling and achieve complete detection of such defects. In Fig. 5(e), RetinaNet and DINO mistakenly classified an area of surface-whitened bricks as brick-spalling defects, while CWADE-Net correctly identified it as a non-defect region. This performance gain arises from C3k2-FSM, which enhances surface features in the frequency domain to effectively distinguish between pure color variations and true brick structural loss. In Figs. 5(f) and (g), YOLOv8, DINO, RetinaNet, and YOLO11 all showed missed detections of brick-spalling defect. Compared to them, CWADE-Net achieved complete detection. This is because of the use of the EIE and C3k2-FSM modules for edge information enhancement, which effectively captures the edge details of brick-spalling regions. The collective design of these modules enables the network to robustly detect surface defect targets in both small-scale vegetation invasion and large-scale brick-spalling scenes, thereby achieving higher overall performance in metrics such as recall and mAP50.

Regarding the number of parameters and computational efficiency, CWADE-Net requires a slightly higher computational cost compared to YOLOv8n and YOLO11n. However, it remains significantly lighter than DINO, RetinaNet, and Faster R-CNN. This suggests that CWADE-Net maintains a good trade-off between lightweight design and high detection performance. Although its computational cost is marginally higher than that of extremely lightweight networks (*e.g.*, YOLOv8n and YOLO11n), the improvement gains in recall and mAP50 sufficiently outweigh the extra computational cost. When compared to large networks, CWADE-Net remains relatively low network complexity, thus enabling better deployability for practical applications. Regarding inference efficiency, CWADE-Net achieves a latency of 7.6 ms per image and a throughput of 131.2 FPS. Although the computational complexity and number of parameters are higher than those of YOLO11n, they remain at a relatively low level in general and do not impose a significant extra computational burden on model deployment. Although YOLOv8n and YOLO11n have a slight advantage in inference speed, CWADE-Net maintains relatively high inference efficiency with improved detection performance, demonstrating its strong real-time detection capability. Compared with traditional two-stage detectors and Transformer-based detection models, CWADE-Net clearly exhibits faster per-image inference and higher FPS, indicating that it achieves a favorable balance between detection accuracy and inference efficiency. Overall, considering precision, recall, mAP50, and the qualitative visual results, CWADE-Net achieves a strong balance among accuracy, robustness, and efficiency.

Ablation studies

To quantitatively analyze the effectiveness and contribution of each component in the network, we conducted eight ablation experiments, with the results summarized in Table 2.

Case1 refers to the baseline network, YOLO11n. It achieved a precision of 73.6%, a recall of 63.3%, and an mAP50 of 70.4%, with a computational cost of 6.3G FLOPs and 2.6M parameters. Meanwhile, the model achieved an inference latency of 3.9 ms per image and a throughput of 256.9 FPS, indicating its high inference efficiency. In subsequent ablation experiments (Case2 - Case5), we independently analyzed the standalone contribution of the SCI, EIE, C3k2-FSM, and LSCSBD modules. Incorporating the SCI module in Case2 improved recall from 63.3% to 66.1% and mAP50 from 70.4% to 71.1%. Also, precision showed a marginal decline to 72.9%. These results indicated that SCI enhances visibility in low-illumination shadowed regions to reduce missed detections, although its contribution to overall precision remains limited. Case3 integrated the EIE module, achieving a precision of 71.1%, a recall of 66.0%, and an mAP50 of 70.5%. These minor changes suggested that edge information encoding alone offers limited improvement in detection performance. However, it enriches edge representations that benefit subsequent network feature fusion. In Case4, adopting the C3k2-FSM module yielded a precision of 68.8%, a recall of 63.7%, and an increased mAP50 of 71.1%. It can be observed that the module enhances structural information by modeling spatial-frequency features. Nevertheless, its standalone use only provided limited performance gains due to insufficient global semantic guidance. In Case5, replacing the original detection head with the lightweight LSCSBD head gave a precision of 73.8%, a recall of 61.8%, and an mAP50 of 71.2%. Meanwhile, the computational cost and parameter count were reduced to 6.2G FLOPs and 2.5M, achieving the lowest model complexity among all configurations. According to these results, the LSCSBD head can effectively reduce computational cost while maintaining comparable detection performance, which offers a more advantageous option for lightweight deployment.

Having analyzed the standalone effect of each independent network component, we jointly incorporated the SCI, EIE, and C3k2-FSM modules in Case6. Under this combined configuration, precision reached 73.0%, recall significantly increased to 67.4%, and mAP50 improved to 73.1%. These results suggested that illumination enhancement, edge information encoding, and spatial–frequency feature extraction complement each other at the input and feature

Table 2
CWADE-Net ablation experiment.

Config.	YOLO11n	SCI	EIE	C3k2-FSM (Backbone)	Neck	LSCSBD	P (%)	R (%)	mAP50 (%)	GFLOPs	Parameters (M)	Latency (ms/img)	FPS (img/s)
Case1	✓						73.6	63.3	70.4	6.3	2.6	3.9	256.9
Case2	✓	✓					72.9	66.1	71.1	7.4	2.6	7.4	135.7
Case3	✓		✓				71.1	66.0	70.5	7.6	2.9	5.8	172.4
Case4	✓			✓			68.8	63.7	71.1	6.8	3.3	7.0	143.1
Case5	✓					✓	73.8	61.8	71.2	6.2	2.5	7.0	143.3
Case6	✓	✓	✓	✓			73.0	67.4	73.1	9.2	3.7	8.2	122.1
Case7	✓	✓	✓	✓	✓		73.6	66.5	74.0	13.8	5.2	9.2	108.7
Case8	✓	✓	✓	✓	✓	✓	73.0	68.1	75.0	11.4	5.6	7.6	131.2

aggregation stages. SCI improves visibility in low-light scenes, EIE refines edge details, and C3k2-FSM enhances discrimination under complex textures. Case7 further introduced an improved neck structure, achieving 73.6% in precision, 66.5% in recall, and 74.0% in mAP50. This is attributed to that the bidirectional feature fusion strategy and the CBAM module enable effective multi-scale information propagation, thus improving the network's capacity for detecting small object targets in complex scenes.

Case8 refers to the final, fully configured CWADE-Net by integrating the SCI, EIE, C3k2-FSM, the improved neck, and LSCSBD modules. Under this complete configuration, it achieved a precision of 73.0%, a recall of 68.1%, and the highest mAP50 of 75.0%, with a computational cost of 11.4G FLOPs and 5.6M parameters. It also achieved an inference latency of 7.6 ms per image and a throughput of 131.2 FPS, outperforming Case6 and Case7. This suggests that the incorporation of LSCSBD further improves inference efficiency while maintaining high detection performance. The performance gains demonstrated that the incorporated modules complement each other across illumination refinement, edge encoding, spatial–frequency feature extraction, and multi-scale fusion. This mutual enhancement leads to an effective accuracy balance in precision and recall, as well as an overall optimal detection performance.

Meanwhile, to justify our choice of the CBAM attention mechanism in the neck, we further performed ablation studies by comparing the effectiveness of different attention mechanisms while keeping all other network components and training settings unchanged. The attention module in CWADE-Net's neck was replaced with alternative mechanisms, including plain neck without attention (w/o), SE [66], CoordAttention [67], EMA [68], LSKA [69], TripletAttention [70], CAA [71], and CBAM (adopted by us) [58]. The plain neck structure without the attention mechanism achieved a mAP50 of 68.0%. Incorporating the mechanisms of SE, CoordAttention, EMA, LSKA, TripletAttention, CAA, and CBAM resulted in mAP50 scores of 69.1%, 71.1%, 72.9%, 68.7%, 71.7%, 67.7%, and 75.0%. Among them, CBAM achieved the highest mAP50 score, leading to a 7.0% performance gain over the plain neck structure. Compared to the second-best method, EMA, CBAM showed a better performance with 2.1% improvement in mAP50, demonstrating its superiority among all comparative attention mechanisms.

The neck of CWADE-Net is a key component responsible for multi-scale feature aggregation and information filtering. In the channel aspect, it must select discriminative sub-channels of the feature maps produced by the spatial-frequency encoding. In the spatial aspect, it further needs to filter out irrelevant wall surface textures and vegetation clutter. The adopted CBAM attention mechanism well aligns with this requirement by jointly performing channel recalibration and spatial attention. Specifically, channel attention adaptively reweights feature channels according to their importance, which highlights structural and textural cues mostly relevant to defect targets while suppressing irrelevant wall textures and background noise. In addition, spatial attention enables the network to focus on critical regions such as brick-spalling edges and vegetation-invasion textures, thus effectively reducing both false positives and false negatives in complex scenes. Unlike CBAM, CoordAttention performs coordinate-aware directional attention, which tends to discard fine-grained textural details during feature fusion. Meanwhile, LSKA focuses on enlarged receptive fields, and EMA performs local self-attention. Nevertheless, at the neck where multi-scale high- and low-frequency information are jointly integrated, these mechanisms may over-smooth features or introduce redundant feature correlations. TripletAttention introduces cross-dimensional attention, yet it brings little additional performance gains in this task. CAA's mechanism is limited in small-target detection tasks with category-imbalanced input data, which tends to degrade the detection performance with the reduced mAP50 score.

To further analyze the effectiveness of joint spatial-frequency feature extraction of the proposed C3k2-FSM block, we conducted three ablative experiments by keeping only the frequency branch (Case1), keeping only the spatial branch

Table 3
C3k2-FSM module ablation results.

Config.	Branch (freq)	Branch (spacial)	P (%)	R (%)	mAP50 Invasion of herbaceous/ woody vegetation(%)	mAP50 Brick spalling(%)	mAP50 Invasion of vine-type vegetation(%)	mAP50 (%)	GFLOPs	Parameters (M)	Latency (ms/img)	FPS (img/s)
Case1	✓		73.3	63.7	82.3	84.0	50.5	72.3	11.4	5.6	5.5	181.9
Case2		✓	72.6	65.2	81.6	83.7	53.0	72.8	11.4	5.6	5.9	170.0
Case3	✓	✓	73.0	68.1	82.4	87.9	54.8	75.0	11.4	5.6	7.6	131.2

(Case2), and keeping both branches (Case3). All three experiments were conducted under the same training settings for a fair evaluation.

Fig. 6 presents the visual comparison. Results of Case1 showed high sensitivity to color variations and coarse textures on the brick surface, with the corresponding heatmaps exhibiting dispersed responses. These results suggested that by only performing frequency-domain feature encoding, the network tends to misinterpret variations in surface color and texture as defects, thereby increasing false positives. On the other hand, for small-scale or edge-blurred brick-spalling defects, the relatively weak texture contrast is likely to result in false negatives. In Case2, the heatmap responses mainly concentrated along defect edges, which effectively suppressed background noise. Nevertheless, it still failed to detect defect targets in certain regions where the wall surface is covered by fine vegetation or climbing plants. Case3 showed the optimal qualitative detection performance. The heatmaps formed continuous responses encompassing the boundaries of defect areas, and maintained stably high responses within fine-grained vegetative regions. Moreover, high-response artifacts in irrelevant backgrounds were significantly reduced. Regarding defect target detection, the predicted bounding boxes in Case3 more closely matched the shapes and sizes of the GT boxes. Adjacent small targets were also accurately distinguished without merging or omission.

The quantitative results of the three experiments are provided in Table 3, which are consistent with the visual observations. Case1, keeping only the frequency-domain branch, achieved a mAP50 of 72.3% and a lowest recall rate of 63.7%. The mAP50 scores of herbaceous/woody vegetation invasion and brick-spalling defect were 82.3% and 84.0%, whereas the score of vine-type vegetation invasion dropped to only 50.5%. In Case2, where only the spatial-domain branch is kept, the overall mAP50 slightly increased to 72.8%, and the recall improved to 65.2%. The mAP50 increased from 50.5% to 53.0% for vine-type vegetation-invasion defect. However, the scores slightly decreased to 81.6% and 83.7% for herbaceous/woody vegetation invasion and brick-spalling defect. These results suggested that incorporating spatial-domain gradient cues can enhance edge responses and improve detection for certain small-scale targets, but are less effective than frequency-domain feature extraction in filtering out irrelevant background textures. Case3 improved the overall mAP50 to 75.0% by integrating both frequency- and spatial-domain branches. The mAP50 scores of herbaceous/woody vegetation invasion, brick-spalling defect, and vine-type vegetation invasion reached 82.4%, 87.9%, and 54.8%, respectively, while the overall precision and recall increased to 73.0% and 68.1%. Compared to Case1 and Case2, Case3 achieved higher detection accuracy and more stable recall across all three defect categories, which validated the effectiveness of jointly spatial- and frequency-domain feature extraction.

Both the qualitative and quantitative results indicated that the spatial-domain branch provides more salient edge and shape cues, while the frequency-domain branch shows better robustness against texture and illumination variations. Combining both branches yields complementary feature representations at the channel level, which not only reduces false positives arising from irrelevant background textures but also reduces false negatives and edge shifts under weak-texture and low-contrast conditions. As demonstrated in Table 3 and Fig. 6, the network with both branches achieved the optimal detection performance, especially in complex city-wall scenes.

Model generalization assessment on crack defects

To evaluate the generalization capability of CWADE-Net to other types of defects, we chose surface cracks, another type of commonly observed defect on the Nanjing Ming Dynasty city wall. We collected a total of 597 crack samples and applied data augmentation techniques, including horizontal flipping, vertical flipping, image rotation, affine transformations, random cropping and scaling, as well as color enhancement, same as previous experiments. The train-test-split ratio was set as 8:2. Then, we trained CWADE-Net independently on this dataset to evaluate its generalization performance. The quantitative results were reported in Table 4. CWADE-Net achieved a precision of 79.3%, a recall of 60.0%, and an mAP50 of 71.0% in crack detection, while maintaining 11.4 GFLOPs, 5.6 million parameters, 145.4 FPS,

Table 4
results of CWADE-Net on crack detection.

Index	P (%)	R (%)	mAP50 (%)	GFLOPs	Parameters (M)	Latency (ms/img)	FPS (img/s)
CWADE-Net	79.3	60.0	71.0	11.4	5.6	6.9	145.4

and a per-image latency of 6.9 ms. According to the qualitative results visualized in Fig. 7, our method demonstrated strong localization capability under varying wall texture backgrounds and crack shapes, suggesting that CWADE-Net is not only effective for vegetation intrusion and brick spalling detection, but also generalizes well to crack detection.

Discussion

In this section, we provide a systematic analysis to verify CWADE-Net's robustness to image sizes, scales, varying image qualities, and cross-platform imagery. For all analyses, we directly adopted the CWADE-Net model trained in the previous Results section, and performed inference on these newly unseen image data to achieve a fair evaluation. We also analyze the methodological limitations of our method. Then, we end this section with the conclusions of the proposed CWADE-Net and future recommendations.

To analyze CWADE-Net's robustness against image sizes and scales, we performed experiments on images acquired from a DJI Matrice 4E UAV, with the single frame resolution of 5280×3956. We rescaled and fed the full UAV images into CWADE-Net for surface defect detection without any cropping. Figs. 8(a)-(b) visualized the results. The detection results showed that the network could only detect some large-scale vegetation-invasion defects. Brick-spalling defects were often misdetected. The underlying reason is that when the pixels of the full image are downsampled through rescaling, the effective pixels representing individual bricks are greatly reduced. Considering that the network also performs a sequence of downsampling during feature aggregation, the key discriminative features, such as brick-spalling edges and textures, can be easily diminished. On the other hand, the network is trained with small-scale images captured by ground-based cameras and mobile phones, where the relative scale of brick-spalling defects is remarkably larger than their relative scale in full-frame UAV imagery. This scale mismatch leads to significant discrepancies between the brick-spalling features learned from the training set and the actual brick-spalling features present in the test set, thus resulting in a severe domain gap that degrades the network's detection performance in small-scale brick-spalling defects.

Directly feeding the full high-resolution UAV image into the network leads to the spatial compression and loss of details of small-scale targets, such as missing-brick defects. To mitigate this issue, we divided the original 5280×3956 image by a 3×3 grid, producing nine equal-sized subimages. Each subimage was then rescaled to the network's required input dimensions and individually processed by CWADE-Net. By seamlessly merging these sub-outputs, we obtained the detection result of the original image, as visualized in Figs. 8(c)-(d). Compared with the results obtained from the full-frame input, its detection performance improved remarkably. Specifically, in large-scale continuous vegetation-covered regions, the network could completely delineate the major vegetation zones and reliably detect the small-scale vegetation invasion. For wall surfaces densely distributed with missing-brick defects, the detection results obtained after image cropping could more comprehensively identify and localize the previously overlooked brick-spalling defect. However, it also introduced certain false positives. For example, embrasures (red-circled regions) and rock-layer peeling at the base of the wall (green-circled regions) were mistakenly detected as brick-spalling defects. This is mainly due to that the embrasure, being a functional rectangular opening on the battlement, shares geometrical similarities with the regions of single or multi brick-spalling defect. Meanwhile, rock-layer peeling and brick-spalling defects both exhibit irregular depressions arising from weathering, thus they are observed to have highly similar textural characteristics. Moreover, the diversity of labeled training samples is limited. Due to that images in the training set were captured by ground-based cameras or mobile phones, embrasures and rock-layer peeling samples are significantly underrepresented. Embrasures typically occur only along the top of the city wall, and rock-layer peeling appears only at the base of certain wall sections. Their data sparsity prohibits the network from learning discriminative features of these two categories. Despite this limitation, overall, the *cropping-detection–recomposition* strategy effectively enhances small-scale defect detection without modifying the network's architecture, which significantly improves the generalizability of CWADE-Net when extended to large-scale UAV-captured city wall scenes.

To evaluate the robustness of CWADE-Net on multi-scale imagery, we conducted the experiments by adjusting the shooting distance and focal length. Three imaging scales (far, medium, and near) were obtained. The detection results were visualized in Fig. 9. For the far-scale images, captured with a 6.7mm focal length, the defect targets occupy a relatively small proportion of the image. Therefore, we applied a 3×3 patch-based cropping strategy. Results indicated that the detection performance on far-scale images was comparable to that of medium-scale images after patching, suggesting that this strategy can effectively enhance small target detection in long-distance imaging. Medium-scale images were acquired using a 19.2mm focal length. As the target scale increased, the model exhibited more stable responses to brick loss and vegetation invasion defects. However, missed detections still occurred for certain small-scale defects. Near-scale images, captured with a 40mm focal length, preserved more detailed target information and exhibited scales closer to those of the training samples. They achieve the best detection performance, showing more accurate and complete localization of both brick loss and vegetation invasion, and significantly reducing missed detections of small-scale defects. Across all three imaging scales, the current CWADE-Net model has demonstrated consistently strong detection performance, demonstrating robust scale adaptability. Moreover, incorporating more diverse multi-scale image data in training will further enhance its cross-scale generalization capability.

In addition, to evaluate CWADE-Net's robustness against variations in image quality, we collected images under complex environmental conditions such as motion blur, rainfall, and nighttime scenarios. The defect detection results were presented in Fig. 10. Under normal conditions (Fig. 10(a)), the model achieved accurate localization of both herbaceous/woody vegetation intrusion and brick loss defects. Under motion blur conditions (Fig. 10(b)), although most prominent defect regions could still be detected, certain small-scale vegetation targets (*e.g.*, the orange dashed box in Fig. 10 (b)) were omitted. This is because image blur degrades fine texture and edge features, thus decreasing the network's sensitivity to small vegetation defects. In rainy conditions (Fig. 10(c)), CWADE-Net achieved a generally good performance. However, variations in illumination and the darkening moisture wall surfaces due to moisture could lead to misclassification (e.g., the yellow dashed box in Fig. 10(c) was wrongly detected as brick loss). For nighttime imagery (Fig. 10(d)), CWADE-Net could detect prominent defect regions, suggesting a certain level of adaptability to low-light environments. Nevertheless, due to insufficient illumination, small vegetation defects tended to be mis-detected (e.g., the orange dashed box in Fig. 10(d)). Overall, our proposed method demonstrated good robustness under challenging environmental conditions. However, its performance may degrade when detecting small targets under extremely low illumination and reduced image clarity conditions.

To further assess the generalizability of CWADE-Net to different types of input imagery, we selected several representative scenes of the Nanjing Ming Dynasty city wall, capturing the same wall sections using a UAV and a mobile phone. We then performed qualitative comparison analysis of images captured from these two devices. Experiment results (Fig. 11) demonstrated that, although the two types of images differ in imaging distances, viewing angles, and resolution, their overall detection performances are highly consistent. This suggests that the CWADE-Net trained on smartphone and camera data can effectively transfer its knowledge to UAV imagery, demonstrating good adaptability to multi-device data. However, due to imaging perspective and illumination conditions, there still exist some local variations in the detection performance. UAV images were acquired from greater distances and with wider viewing angles, which exhibited little geometric distortion but tended to compress 3D structural details of city wall defects. They were more sensitive to complex illumination effects such as highlights and shadows. Contrarily, smartphone images, captured at closer range with more focused viewpoints, preserved finer defect details and thus showed superior detection performance of small-scale defects. In the red box of Fig. 11, small vegetation defects located in the lower region of the smartphone image (highlighted by orange circles in the right) were accurately detected, while the same defects in the UAV imagery were missed. This discrepancy indicated that CWADE-Net is more sensitive to small vegetation defects on smartphone images. On the other hand, in the purple box (Fig. 11), four instances of brick loss defects in the smartphone image were not detected, while the corresponding defects in the UAV imagery (highlighted by yellow circles) were successfully identified, suggesting better performance in detecting brick loss defects in UAV images. All these analyses have shown that the network trained on mobile-phone and camera imagery in general performs reliable inference on unseen UAV data and can achieve effective cross-device adaptation.

Despite the superior defect detection performance of the proposed CWADE-Net on most Nanjing Ming Dynasty city wall scenes, it still suffers from three limitations when extended to practical applications, which are summarized as follows:

First, CWADE-Net struggles in handling defect objects that belong to minority categories or are under complex textural backgrounds. As illustrated in Fig. 12(a), CWADE-Net fails to detect some withered vine-type defects due to their low color contrast, irregular spatial distribution, and appearance similarity to brick defects such as cracks

and weathering. As shown in Fig. 12(b), a few green vine-type defects tend to be wrongly detected as herbaceous or woody vegetation defects, mainly because they are highly similar to the surrounding vegetation in color, texture, and local appearance. In addition, vine-type vegetation belongs to a minority defect category in the dataset, which further increases the difficulty of accurate discrimination. Besides, CWADE-Net occasionally misidentifies darker-colored bricks as brick loss defects (Fig. 12(c)). This can be attributed to the color and texture similarity between dark-colored bricks and actual brick loss areas. In Fig. 12(d), CWADE-Net shows limitations in detecting vertically connected brick-loss regions. Although the main defective area can still be identified, the predicted boxes have low confidence and fail to fully cover the actual defect region, mainly because the model is primarily trained on single-brick-scale or locally discrete brick-spalling patterns, whereas vertically connected missing regions cover a larger area and are therefore difficult to identify as an integral whole.

Second, although CWADE-Net has effectively met the requirements for rapid localization of multiple defect categories on the Nanjing Ming city wall, it essentially belongs to object detection–based approaches, and thus is by nature limited in pixel-level identification of defects compared to segmentation methods. Particularly, as detection outputs are often expressed as rectangle bounding boxes, they cannot precisely delineate the true boundaries of defects, which may pose challenges to subsequent contour extraction, area estimation, and fine-grained damage assessment.

Last, CWADE-Net adopts a lightweight approach design, showing high potential for deployment on UAV platforms such as DJI for real-time detection. Nevertheless, its practical application in large-scale real-world cultural heritage monitoring is still subject to certain data-source limitations. Specifically, CWADE-Net's training data purely consists of close-range images captured by smartphones and cameras. When transferred to UAV platforms, the domain discrepancies in imaging perspective, scale, and image quality may lead to a suboptimal detection performance.

To sum up, this paper aims to address the automated detection of surface defects on the Nanjing Ming Dynasty city wall, including the invasion of herbaceous/woody and vine-type vegetation, as well as brick spalling and material loss. To this end, we have proposed a deep learning-based detection network, CWADE-Net. It integrates the modules of the self-calibrated illumination enhancement, edge information encoding, and spatial-frequency feature extraction. Specifically, SCI-Net is incorporated in CWADE-Net to effectively improve the visibility of shadowed, reflective, and low-light regions through multiple stages of illumination estimation and adaptive calibration. This allows the network to learn robust and discriminative wall surface defect features under insufficient or uneven illumination. Furthermore, CWADE-Net employs the EIE modules and the C3k2-FSM blocks to enhance the feature representation along the edges of brick-spalling regions and vegetation-invasion contours, while significantly suppressing background noise. As a result, we obtain high-quality city wall defect features with both sharp-edge information and rich textural details. Benefiting from these deep feature representations, CWADE-Net achieves accurate and robust city wall defect detection under various challenging conditions, such as low illumination, occlusion, and large-scale weathering. Besides, bidirectional feature fusion is performed at the neck of CWADE-Net to facilitate comprehensive interaction between high-level semantics and low-level fine-grained structural information, enabling the network to simultaneously capture global large-scale defect patterns and local small-scale defect details. We utilize a lightweight detection head, LSCSBD, which significantly reduces computational cost while maintaining detection accuracy, thereby improving the real-time performance and deployability of the network. Compared to mainstream deep learning–based object detection frameworks, CWADE-Net achieved mAP50 scores of 82.4%, 87.9%, and 54.8% for three representative defect categories, herbaceous/woody vegetation invasion, brick spalling, and vine-type vegetation invasion, respectively. The overall mAP50 reached 75.0%, demonstrating a notably superior performance over both classical single-stage detectors (*e.g.*, YOLO11, RetinaNet, and DINO) and multi-stage detection networks (*e.g.*, Faster R-CNN). These results indicate that CWADE-Net achieves a favorable balance among detection accuracy, robustness, and inference efficiency, and provides an effective technical solution for intelligent defect inspection of the Nanjing Ming Dynasty city wall.

In future work, we will apply the proposed CWADE-Net to perform comprehensive, full-coverage surface defect detection along the existing Nanjing Ming Dynasty city wall, with a total remaining length of approximately 25km. To mitigate the dependence on fully annotated datasets, we consider incorporating active learning strategies that allow the network to incrementally learn from a small number of annotated training samples [72]. To strengthen the methodological framework of CWADE-Net, we plan to integrate more powerful backbones, such as SegmentAnything and graph neural networks, to achieve more effective feature learning and contextual understanding [11]. In terms of defect categories, we will include more fine-grained defect types in ancient wall structures, such as brick cracking, bulging, displacement, and settlement. Furthermore, we plan to perform precise semantic segmentation and geometric delineation of the detected surface defects to quantitatively analyze their severity level. Thus, CWADE-Net can be used to support the effective monitoring, early-stage warning, conservation, and restoration of the Nanjing Ming Dynasty city

wall. At the meantime, by achieving pixel-level segmentation, the proposed CWADE-Net will also demonstrate strong potential for application in remote sensing tasks. In particular, our pre-trained model can serve as an effective foundation for general-purpose remote sensing scene segmentation through knowledge adaptation and transfer learning techniques [12, 73]. Last, the inscriptions engraved on the ancient bricks of the city wall in the Ming Dynasty represent one of the most unique and culturally valuable characteristics of the monument. Among the hundreds of millions of bricks made in the Ming Dynasty, more than 90% bear inscriptions. These inscriptions record detailed information such as the production region, names of responsible officials, kiln craftsmen, place of origin, and firing dates, vividly reflecting the strict accountability system, also known as "mandatory attribution of workmanship", in the Ming Dynasty. In the future, we aim to develop a deep learning–based network specifically designed for brick inscription recognition, further uncovering the historical and cultural value of the Nanjing Ming Dynasty city wall and contributing to the nomination of the "Ming and Qing Dynasty City Walls of China" for World Cultural Heritage.

Data availability

The datasets generated and analyzed during the current study are publicly available at https://github.com/Yuanxlhh/CWADENet.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 42571513, and Grant 42271450, in part by the Major Program of the National Natural Science Foundation of China under Grant 42293272, and in part by the Project of Anhui Provincial Department of Education Scientific Research under Grant 2025AHGXZK31210.

Author contributions

X.Y. performed the study, designed and implemented the algorithms, and drafted the manuscript. N.W., Y.W., and D.C. provided supervision and data support. S.D. assisted in manuscript writing. S.L. and Z.W. assisted in figure and table preparation. M.S., Y.S., J.P., and L.Z . provided constructive comments. All authors reviewed the manuscript.

Competing interests

The authors declare no competing financial or non-financial interests.

References

[1] Li, Y. *et al.* Role of the urban plant environment in the sustainable protection of an ancient city wall. *Building and Environment* **187**, 107405 (2021). URL https://www.sciencedirect.com/science/article/pii/S0360132320307733.

[2] Alexakis, E. *et al.* A novel application of deep learning approach over irt images for the automated detection of rising damp on historical masonries. *Case Studies in Construction Materials* **20**, e02889 (2024).

[3] Resende, M. M. *et al.* Infrared thermal imaging to inspect pathologies on façades of historical buildings: A case study on the municipal market of são paulo, brazil. *Case Studies in Construction Materials* **16**, e01122 (2022).

[4] Ludeno, G., Cavalagli, N., Ubertini, F., Soldovieri, F. & Catapano, I. On the combined use of ground penetrating radar and crack meter sensors for structural monitoring: Application to the historical consoli palace in gubbio, italy. *Surveys in Geophysics* **41**, 647–667 (2019).

[5] Qian, W., Wu, R., Tian, W., Zhang, T. & Li, N. Non-destructive detection and three-dimensional imaging of internal defects in beijing ming great wall. *npj Heritage Science* **14** (2026).

[6] Hu, Y., Feng, B. & Hou, M. A study on the detection of bulging disease in ancient city walls based on fitted initial outer planes from 3d point cloud data. *Heritage Science* **11** (2023). URL http://dx.doi.org/10.1186/s40494-022-00856-6.

[7] Fehér, K. & Ákos Török. Detecting short-term weathering of stone monuments by 3d laser scanning: lithology, wall orientation, material loss. *Journal of Cultural Heritage* **58**, 245–255 (2022).

[8] Jin, L., Zhang, H., Sun, Y., Li, M. & Zhao, Y. Research on the application of ps-insar technology in the deformation monitoring of nanjing city wall. *Geomatics & Spatial Information Technology* **47**, 5–8, 13 (2024).

[9] Tapete, D. *et al.* Integrating radar and laser-based remote sensing techniques for monitoring structural deformation of archaeological monuments. *Journal of Archaeological Science* **40**, 176–189 (2013). URL https://www.sciencedirect.com/science/article/pii/S0305440312003512.

[10] Kwon, D. & Yu, J. Automatic damage detection of stone cultural property based on deep learning algorithm. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-2/W15**, 639–643 (2019).

[11] Scarrica, V. M. & Staiano, A. Unveiling graph power: Segmentanything and gcn synergy for instance segmentation and classification. In *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 1–7 (IEEE, 2024).

[12] Ngo, B. H., Chae, Y. J., Park, J. H., Kim, J. H. & Cho, S. I. Easy-to-hard structure for remote sensing scene classification in multitarget domain adaptation. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 1–15 (2023).

[13] Haider, Z. A. *et al.* A comprehensive approach for image quality assessment using quality-centric embedding and ranking networks. *Pattern Recognition* **173**, 112890 (2026).

[14] Ul Amin, S. *et al.* Eadn: An efficient deep learning model for anomaly detection in videos. *Mathematics* **10**, 1555 (2022).

[15] Amin, S. U., Jung, Y., Fayaz, M., Kim, B. & Seo, S. Enhancing pine wilt disease detection with synthetic data and external attention-based transformers. *Engineering Applications of Artificial Intelligence* **159**, 111655 (2025).

[16] Marín-García, D., Bienvenido-Huertas, D., Carretero-Ayuso, M. J. & Torre, S. D. Deep learning model for automated detection of efflorescence and its possible treatment in images of brick facades. *Automation in Construction* **145**, 104658 (2023).

[17] Mishra, M. & Lourenço, P. B. Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review. *Journal of Cultural Heritage* **66**, 536–550 (2024). URL https://www.sciencedirect.com/science/article/pii/S1296207424000050.

[18] Hatir, M. E., Barstuğan, M. & İsmail İnce. Deep learning-based weathering type recognition in historical stone monuments. *Journal of Cultural Heritage* **45**, 193–203 (2020).

[19] Karimi, N., Valibeig, N. & Rabiee, H. R. Deterioration detection in historical buildings with different materials based on novel deep learning methods with focusing on isfahan historical bridges. *International Journal of Architectural Heritage* **18**, 981–993 (2024).

[20] D'Orazio, M., Gianangeli, A., Monni, F. & Quagliarini, E. Automatic monitoring of the bio colonisation of historical building's facades through convolutional neural networks (cnn). *Journal of Cultural Heritage* **70**, 80–89 (2024).

[21] Karimi, N., Mishra, M. & Lourenço, P. B. Deep learning-based automated tile defect detection system for portuguese cultural heritage buildings. *Journal of Cultural Heritage* **68**, 86–98 (2024).

[22] Seo, H., Raut, A. D., Chen, C. & Zhang, C. Multi-label classification and automatic damage detection of masonry heritage building through cnn analysis of infrared thermal imaging. *Remote Sensing* **15**, 2517 (2023).

[23] Muñoz-Silva, E. M., Vasquez-Gomez, J. I., Merlo-Zapata, C. A. & Antonio-Cruz, M. Binary damage classification of built heritage with a 3d neural network. *npj Heritage Science* **13** (2025).

[24] Wang, N., Zhao, Q., Li, S., Zhao, X. & Zhao, P. Damage classification for masonry historic structures using convolutional neural networks based on still images. *Computer-Aided Civil and Infrastructure Engineering* **33**, 1073–1089 (2018).

[25] Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**, 1137–1149 (2017).

[26] Hacıefendioğlu, K., Altunışık, A. C. & Abdioğlu, T. Deep learning-based automated detection of cracks in historical masonry structures. *Buildings* **13**, 3113 (2023).

[27] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (IEEE, 2016).

[28] Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626 (IEEE, 2017).

[29] Ali, L. Damage detection and localization in masonry structure using faster region convolutional networks. *International Journal of GEOMATE* **17**, 98–105 (2019).

[30] Pathak, R., Saini, A., Wadhwa, A., Sharma, H. & Sangwan, D. An object detection approach for detecting damages in heritage sites using 3-d point clouds and 2-d visual data. *Journal of Cultural Heritage* **48**, 74–82 (2021).

[31] Wang, N. *et al.* Automatic damage detection of historic masonry buildings based on mobile deep learning. *Automation in Construction* **103**, 53–66 (2019).

[32] Kwon, D. & Yu, J. Automatic damage detection of stone cultural property based on deep learning algorithm. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-2/W15**, 639–643 (2019).

[33] Li, Q. *et al.* Non-destructive testing research on the surface damage faced by the shanhaiguan great wall based on machine learning. *Frontiers in Earth Science* **11** (2023).

[34] Yang, X., Zheng, L., Chen, Y., Feng, J. & Zheng, J. Recognition of damage types of chinese gray-brick ancient buildings based on machine learning—taking the macau world heritage buffer zone as an example. *Atmosphere* **14**, 346 (2023).

[35] Mishra, M., Barman, T. & Ramana, G. V. Artificial intelligence-based visual inspection system for structural health monitoring of cultural heritage. *Journal of Civil Structural Health Monitoring* **14**, 103–120 (2022).

[36] Idjaton, K. *et al.* Detection of limestone spalling in 3d survey images using deep learning. *Automation in Construction* **152**, 104919 (2023).

[37] Guo, X. Research on automatic detection algorithm of plant invasion based on computer vision. In *2025 IEEE 7th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 689–692 (IEEE, 2025).

[38] Varghese, R. & M., S. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6 (IEEE, 2024).

[39] Zhang, L. *et al.* Investigating the surface damage to fuzhou's ancient houses (gu-cuo) using a non-destructive testing method constructed via machine learning. *Coatings* **14**, 1466 (2024).

[40] Singh, S. K., Maity, D. & Kumawat, P. K. Deep learning-based damage detection and segmentation in the battledore of darbhanga fort. *Journal of Cultural Heritage* **73**, 510–523 (2025).

[41] Zhang, G., Dou, X. & Li, L. Intelligent defect detection of ancient city walls based on computer vision. *Sensors* **25**, 5042 (2025).

[42] Long, L., Gan, Z., Liu, Z., Zhao, B. & Li, Q. Msd-det: Masonry structures damage detection dataset for preventive conservation of heritage. *Journal of Cultural Heritage* **73**, 358–370 (2025).

[43] He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988 (2017).

[44] Hatır, E., Korkanç, M., Schachner, A. & İsmail İnce. The deep learning method applied to the detection and mapping of stone deterioration in open-air sanctuaries of the hittite period in anatolia. *Journal of Cultural Heritage* **51**, 37–49 (2021).

[45] Saravanan, T. J. & Bhaskar, B. Automated evaluation of degradation in stone heritage structures utilizing deep vision in synthetic and real-time environments. *Journal of Building Engineering* **98**, 111117 (2024).

[46] Vandenabeele, L., Loverdos, D., Pfister, M. & Sarhosis, V. Deep learning for the segmentation of large-scale surveys of historic masonry: A new tool for building archaeology applied at the basilica of st anthony in padua. *International Journal of Architectural Heritage* **18**, 1749–1761 (2024).

[47] Chao, J. *et al.* Invasive plants detection and distribution patterns analysis through self-attention enhanced semantic segmentation in uav imagery and moran's index. *Computers and Electronics in Agriculture* **229**, 109811 (2025).

[48] Wang, G. *et al.* Damage detection and safety assessment for historic-district buildings using a semantic segmentation model. *npj Heritage Science* **13**, 636 (2025).

[49] Liu, F. *et al.* Identification methods and evaluation metrics for the condition of the beijing masonry great wall. *npj Heritage Science* **14**, 122 (2026).

[50] Wu, J., Shi, Y., Wang, H., Wen, Y. & Du, Y. Surface defect detection of nanjing city wall based on uav oblique photogrammetry and tls. *Remote Sensing* **15**, 2089 (2023).

[51] Wang, H., Shi, Y., Yuan, Q. & Li, M. Crack detection and feature extraction of heritage buildings via point clouds: A case study of zhonghua gate castle in nanjing. *Buildings* **14**, 2278 (2024).

[52] Li, M., Wang, H., Wang, K., Wang, Z. & Li, Y. City wall multispectral imaging disease detection method based on convolutional neural networks. *Laser & Optoelectronics Progress* **61**, 0437006 (2024).

[53] Ma, L., Ma, T., Liu, R., Fan, X. & Luo, Z. Toward fast, flexible, and robust low-light image enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5627–5636 (IEEE, 2022).

[54] ultralytics. Yolov11. `https://github.com/ultralytics/ultralytics` (2023). Accessed: 2025-12-01.

[55] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007 (2017).

[56] Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640 (IEEE, 2021).

[57] Human Signal. Labelling. `https://github.com/HumanSignal/labelImg` (2017). Accessed: 2025-12-01.

[58] Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, 3–19 (Springer-Verlag, Berlin, Heidelberg, 2018). URL `https://doi.org/10.1007/978-3-030-01234-2_1`.

[59] Shao, J., Mao, Y. & Zhang, J. Learning task-oriented communication for edge inference: An information bottleneck approach. *IEEE Journal on Selected Areas in Communications* **40**, 197–211 (2022).

[60] Engel, K., Hadwiger, M., Kniss, J., Rezk-Salama, C. & Weiskopf, D. *Real-Time Volume Graphics* (AK Peters/CRC Press, 2006).

[61] Liu, J. *et al.* Unified spatial-frequency modeling and alignment for multi-scale small object detection. *Symmetry* **17**, 242 (2025).

[62] Jähne, B., Scharr, H. & Körkel, S. Principles of filter design. *Handbook of Computer Vision and Applications* **2**, 125–151 (1999). URL `https://api.semanticscholar.org/CorpusID:62350295`.

[63] Cooley, J. W. & Tukey, J. W. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation* **19**, 297–301 (1965).

[64] Huang, Y., Ouyang, H. & Miao, X. Lsod-yolov8: Enhancing yolov8n with new detection head and lightweight module for efficient cigarette detection. *Applied Sciences* **15**, 3961 (2025).

[65] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 1627–1645 (2010).

[66] Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141 (IEEE, 2018).

[67] Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13708–13717 (IEEE, 2021).

[68] Li, X., Li, N., Wu, D., Yu, X. & Guo, Y. Infrared dim small target detection method based on yolov8. *LASER & INFRARED* **55**, 789–797 (2025).

[69] Lau, K. W., Po, L.-M. & Rehman, Y. A. U. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Systems with Applications* **236**, 121352 (2024).

[70] Misra, D., Nalamada, T., Arasanipalai, A. U. & Hou, Q. Rotate to attend: Convolutional triplet attention module. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3138–3147 (IEEE, 2021).

[71] Huang, Y., Kang, D., Jia, W., Liu, L. & He, X. Channelized axial attention - considering channel relation within spatial attention for semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* **36**, 1016–1025 (2022).

[72] Amin, S. U., Hussain, A., Kim, B. & Seo, S. Deep learning based active learning technique for data annotation and improve the overall performance of classification models. *Expert Systems with Applications* **228**, 120391 (2023).

[73] Ngo, B. H., Kim, J. H., Park, S. J. & Cho, S. I. Collaboration between multiple experts for knowledge adaptation on multiple remote sensing sources. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–15 (2022).

Figure legends

Figure 1. The overall architecture of CWADE-Net. It consists of a backbone, neck, and detection head for detecting defects on the Nanjing Ming city wall. The backbone includes SCI-Net, EIE, and C3k2-FSM modules. The neck performs multi-scale feature fusion. The detection head outputs defect predictions at different scales.

Figure 2. Edge information encoding module, redrawn from Shao et al. [59]. In this module, Sobel-X and Sobel-Y extract horizontal and vertical gradients, respectively, while max pooling and convolution operations preserve and fuse multi-scale edge information.

Figure 3. Structure of the C3k2-FSM block. The C3k2-FSM block integrates spatial and frequency-domain feature extraction for effective surface defect detection on the Nanjing Ming city wall. a) C3k2-FSM's schematic diagram under the condition of $c3k = false$, where lightweight FreqSpatial modules are applied for low-level feature extraction. b) C3k2-FSM's schematic diagram under the condition of $c3k = true$, where C3k-FreqSpatial modules are employed for high-level contextual understanding. FFT2D and IFFT2D denote two-dimensional fast Fourier transform and inverse fast Fourier transform, respectively.

Figure 4. Detection results of CWADE-Net under different defect scenarios. We visualize 14 representative city wall scenes with surface defects and CWADE-Net's detection results in panels a)-n). In the predicted heatmaps, red color indicates strong network responses while blue color indicates weaker responses.

Figure 5. Comparison of detection results among different models. We compare the detection performance of six methods, including YOLOv8, YOLO11, RetinaNet, Faster R-CNN, DINO, and CWADE-Net in panels a)–g), where representative scenes are selected and visualized including overlapping vegetation and brick spalling defects, vine-type interference, large-area brick spalling, and surface discoloration.

Figure 6. Ablation results of the C3k2-FSM module. The influences of adopting different branches in the C3k2-FSM block on the defect detection results are visualized. a) Original image. b) Detection heatmap with only the frequency-domain branch retained. c) Detection heatmap with only the spatial-domain branch retained. d) Detection heatmap with both branches retained. Red color indicates strong network responses while blue color indicates weaker responses.

Figure 7. Crack detection results of CWADE-Net. The detection performance of CWADE-Net when generalizing to crack defects is visualized. a) Original image. b) Ground-truth crack annotations. c) Detection results on crack defects.

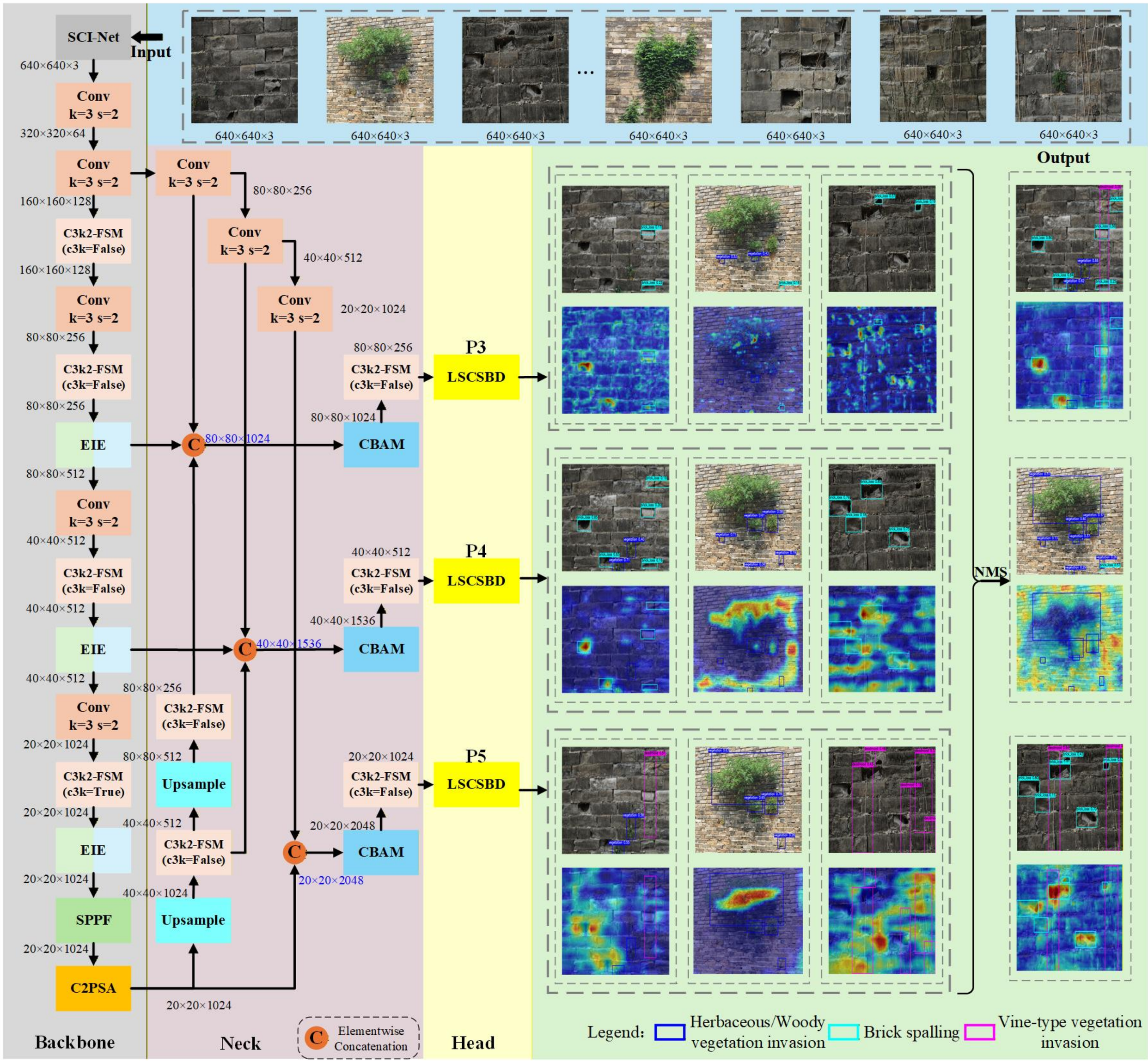
Figure 8. Defect detection results using drone imagery. The full-frame and patch-based city wall defect detection on drone images of the Qingliang Gate section are visually compared. a) and b) Detection results by directly using full-frame images as input. c) and d) Detection results obtained by dividing the same input images into three-by-three patches, detecting defects in each patch separately, and merging the final outputs.

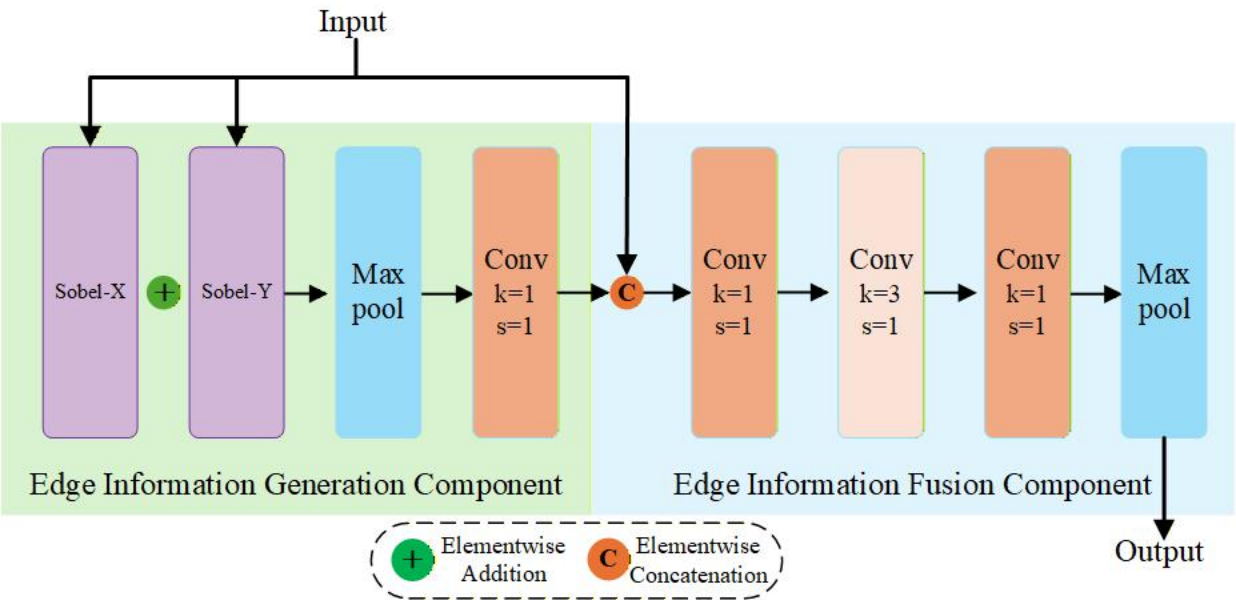
Figure 9. Detection results of varying input image scales. CWADE-Net's defect detection results on far-, medium-, and near-scale images are visually compared. a) Far-scale detection results. b) and d) Medium-scale detection results of the corresponding region highlighted in the red box. c) and e) Near-scale detection results of the corresponding region highlighted in the red box.

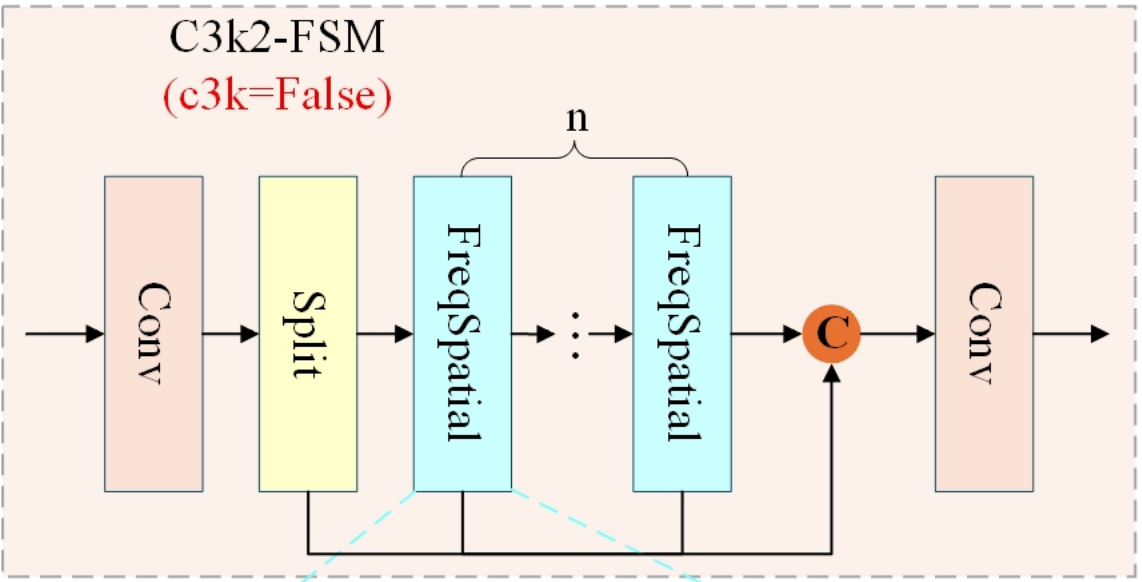
Figure 10. Detection results under varying image-quality conditions. CWADE-Net's defect detection results under four different weather conditions in image qualities are visualized. a) Normal weather. b) Motion blur. c) Rainfall. d) Nighttime. Yellow dashed boxes indicate incorrectly detected defect regions, and orange dashed boxes indicate missed defect regions.

Figure 11. Comparison of drone and mobile-phone detection results. CWADE-Net's defect detection results using drone and mobile-phone images of the same wall section are visually compared. a) and c) Drone-image detection results. b) and d) Mobile-phone image detection results. Yellow circles highlight brick-spalling examples, and orange circles highlight small vegetation examples.

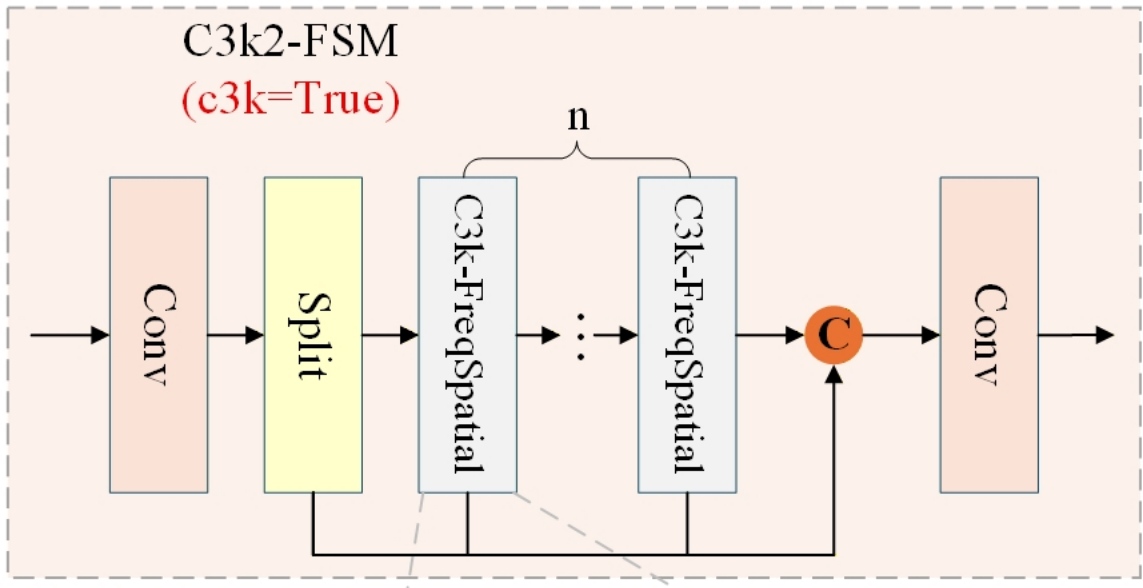
Figure 12. Typical limitation cases of CWADE-Net. Several representative failure cases of CWADE-Net in Nanjing Ming city wall defect detection are selected and visualized. a) Missing detections of withered vine-type vegetation. b) Vine-type vegetation misclassified as herbaceous/woody vegetation. c) Dark-colored bricks incorrectly detected as brick spalling. d) Incomplete detection of vertically connected brick-spalling regions. Yellow dashed boxes highlight the detection error regions.


SCI-Net
Input
640×640×3
Conv k=3 s=2
320×320×64
Conv k=3 s=2
160×160×128
C3k2-FSM (c3k=False)
160×160×128
Conv k=3 s=2
80×80×256
C3k2-FSM (c3k=False)
80×80×256
EIE
80×80×512
Conv k=3 s=2
40×40×512
C3k2-FSM (c3k=False)
40×40×512
EIE
40×40×512
Conv k=3 s=2
20×20×1024
C3k2-FSM (c3k=True)
20×20×1024
EIE
20×20×1024
SPPF
20×20×1024
C2PSA
20×20×1024
Conv k=3 s=2
80×80×256
Conv k=3 s=2
40×40×512
Conv k=3 s=2
20×20×1024
C
80×80×1024
80×80×1024
CBAM
80×80×256
C3k2-FSM (c3k=False)
C
40×40×1536
40×40×1536
CBAM
40×40×512
C3k2-FSM (c3k=False)
80×80×256
C3k2-FSM (c3k=False)
80×80×512
Upsample
40×40×512
C3k2-FSM (c3k=False)
40×40×1024
Upsample
C
20×20×2048
20×20×2048
CBAM
20×20×1024
C3k2-FSM (c3k=False)
P3
LSCSBD
40×40×512
C3k2-FSM (c3k=False)
P4
LSCSBD
20×20×1024
C3k2-FSM (c3k=False)
P5
LSCSBD
640×640×3
640×640×3
640×640×3
640×640×3
640×640×3
640×640×3
640×640×3
Output
NMS
Backbone
Neck
Head
C Elementwise Concatenation
Legend:
Herbaceous/Woody vegetation invasion
Brick spalling
Vine-type vegetation invasion

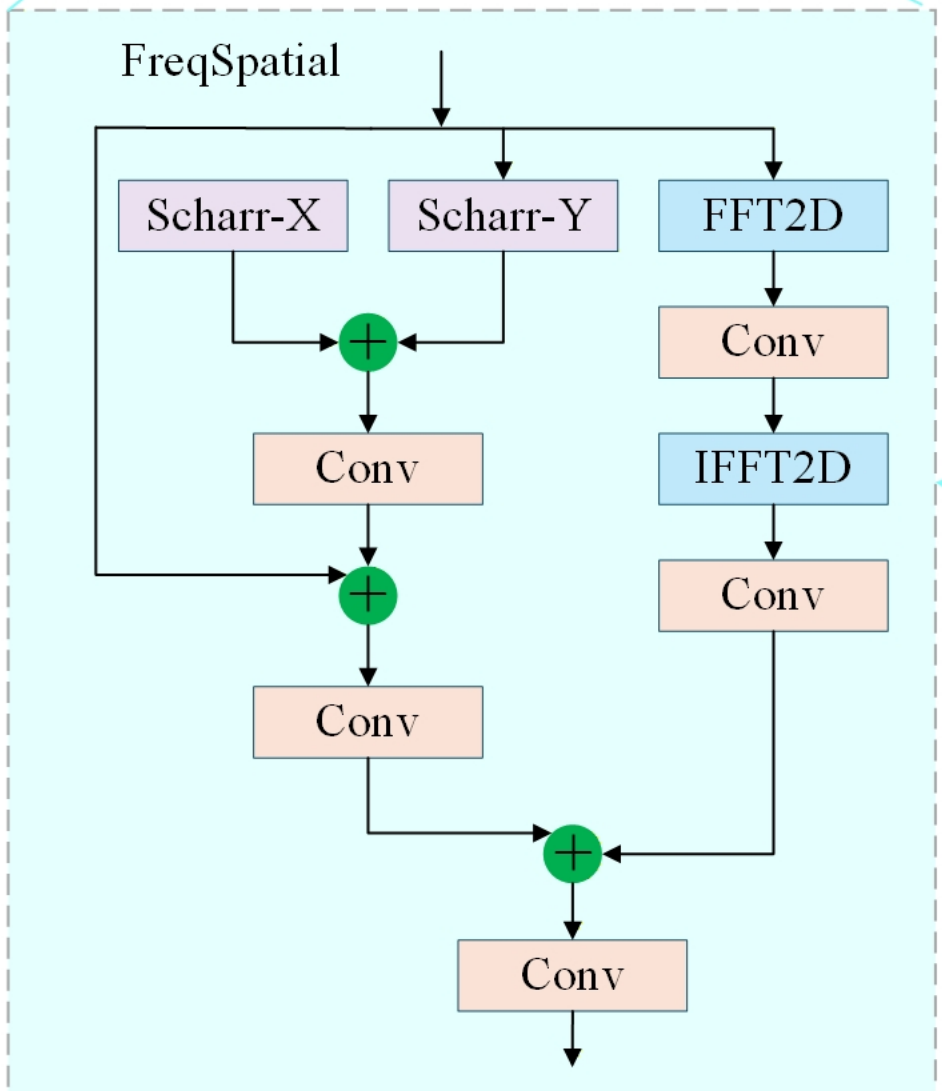

Input
Sobel-X
Sobel-Y
Max pool
Conv k=1 s=1
C
Conv k=1 s=1
Conv k=3 s=1
Conv k=1 s=1
Max pool
Edge Information Generation Component
Edge Information Fusion Component
Output
Elementwise Addition
Elementwise Concatenation

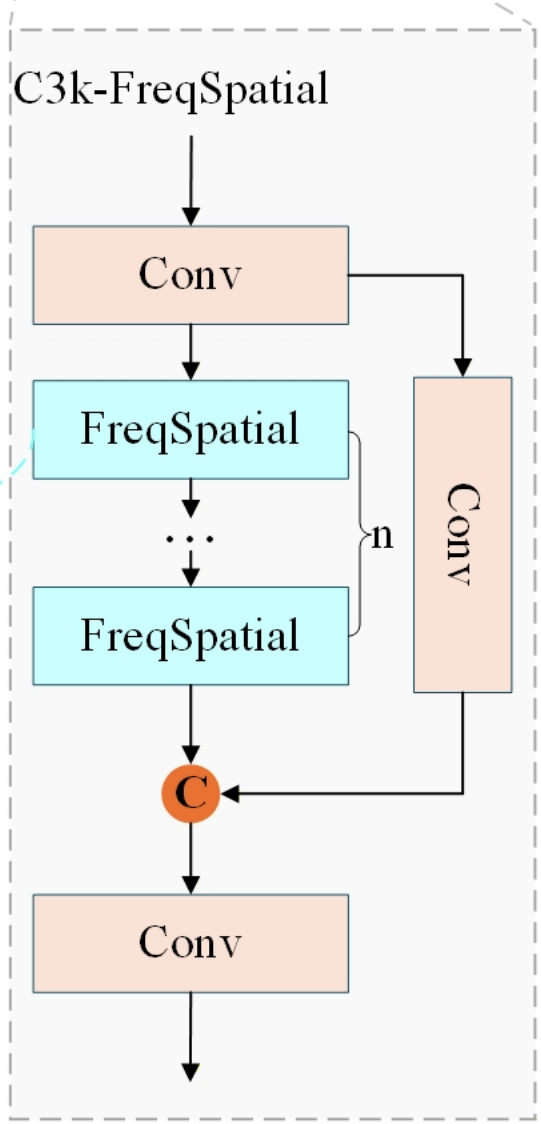

C3k2-FSM
(c3k=False)
n
Conv
Split
FreqSpatial
FreqSpatial
C
Conv

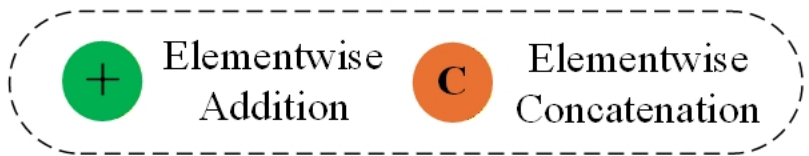
a)

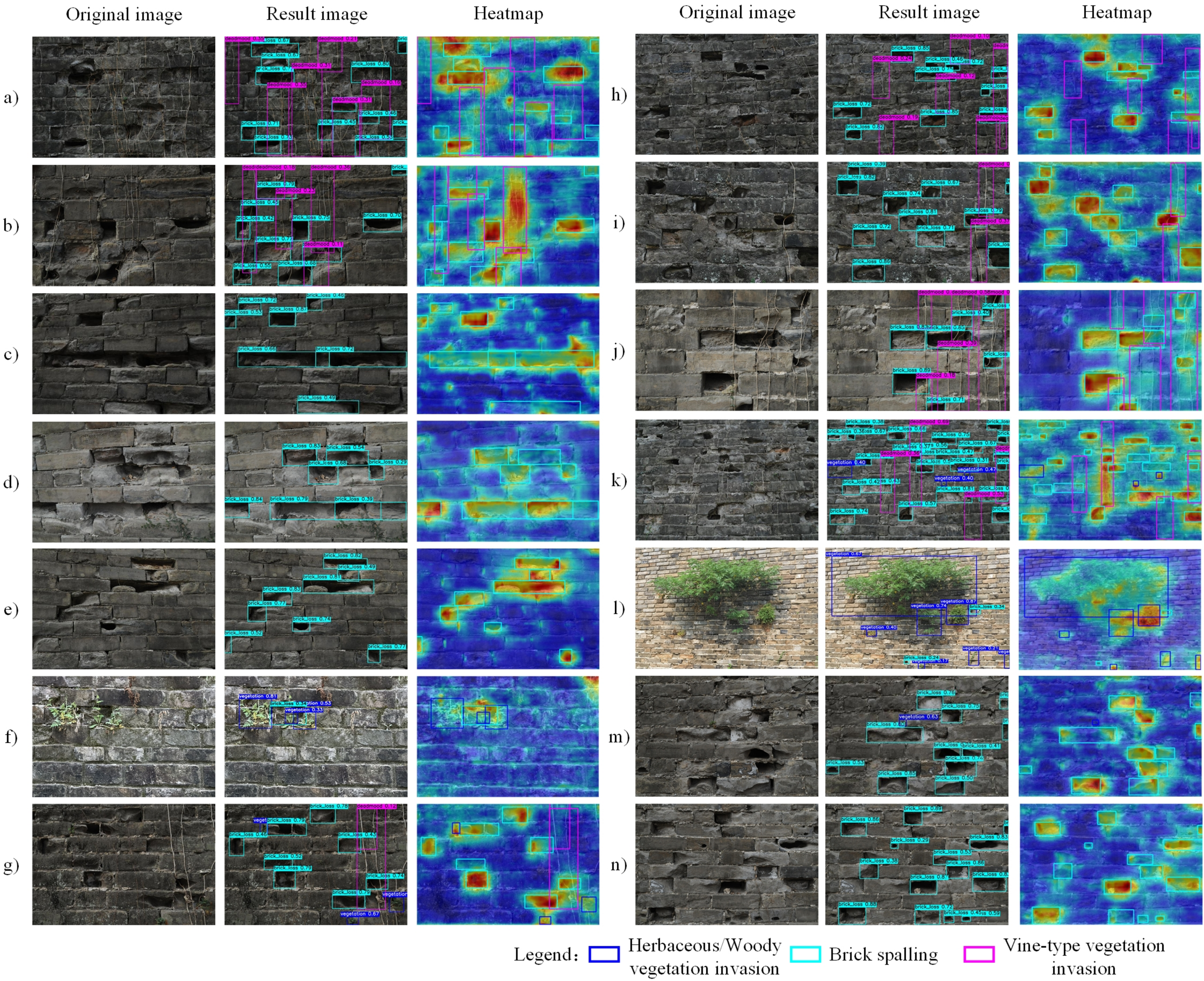

C3k2-FSM
(c3k=True)
n
Conv
Split
C3k-FreqSpatial
C3k-FreqSpatial
C
Conv

b)


FreqSpatial
Scharr-X
Scharr-Y
FFT2D
Conv
Conv
IFFT2D
Conv
Conv
Conv
Conv


C3k-FreqSpatial
Conv
FreqSpatial
FreqSpatial
n
Conv
C
Conv


Elementwise
Addition
C
Elementwise
Concatenation


Original image
Result image
Heatmap
Original image
Result image
Heatmap
a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
k)
l)
m)
n)
Legend：
Herbaceous/Woody vegetation invasion
Brick spalling
Vine-type vegetation invasion

a)

b)

c)

d)

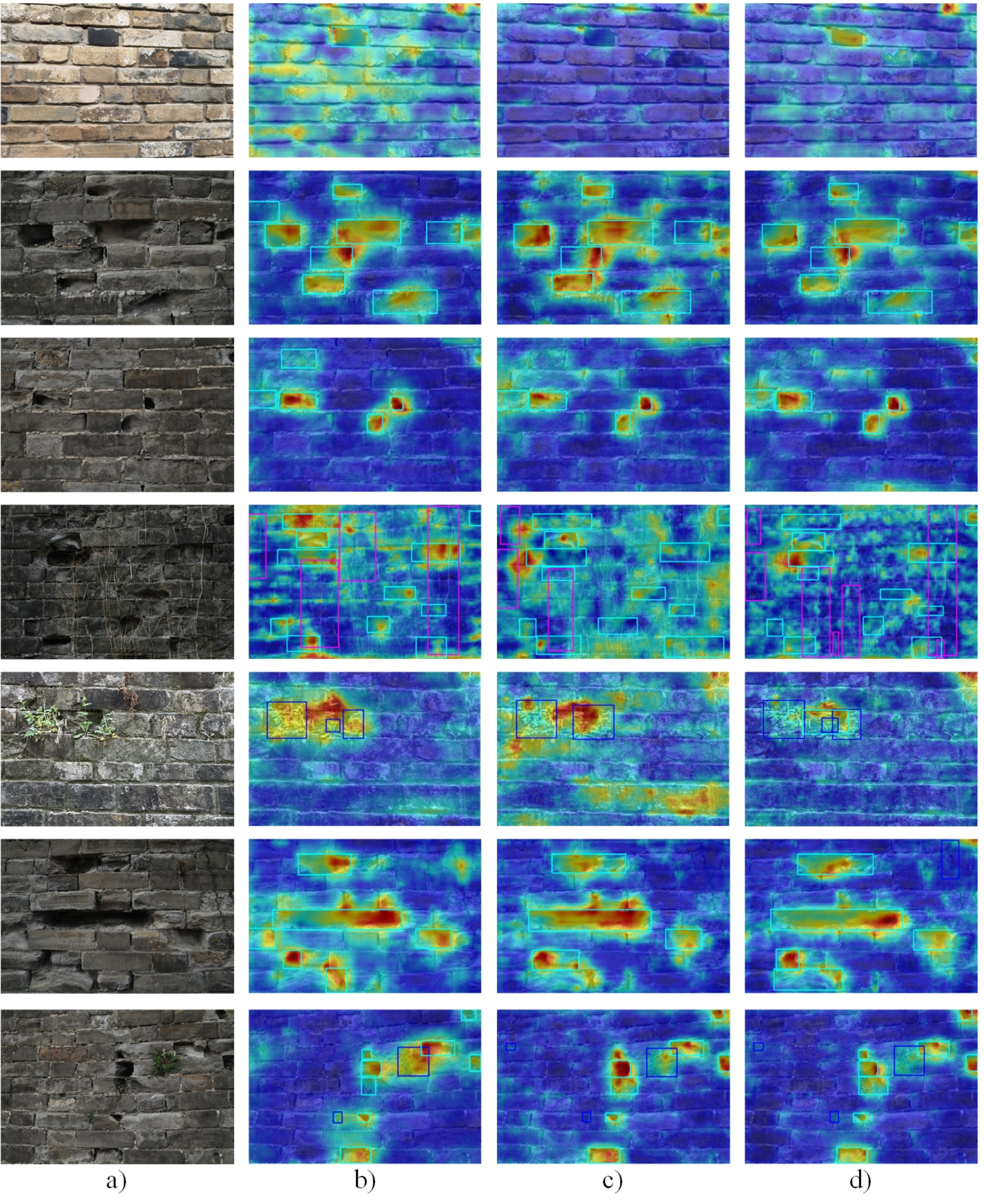
e)

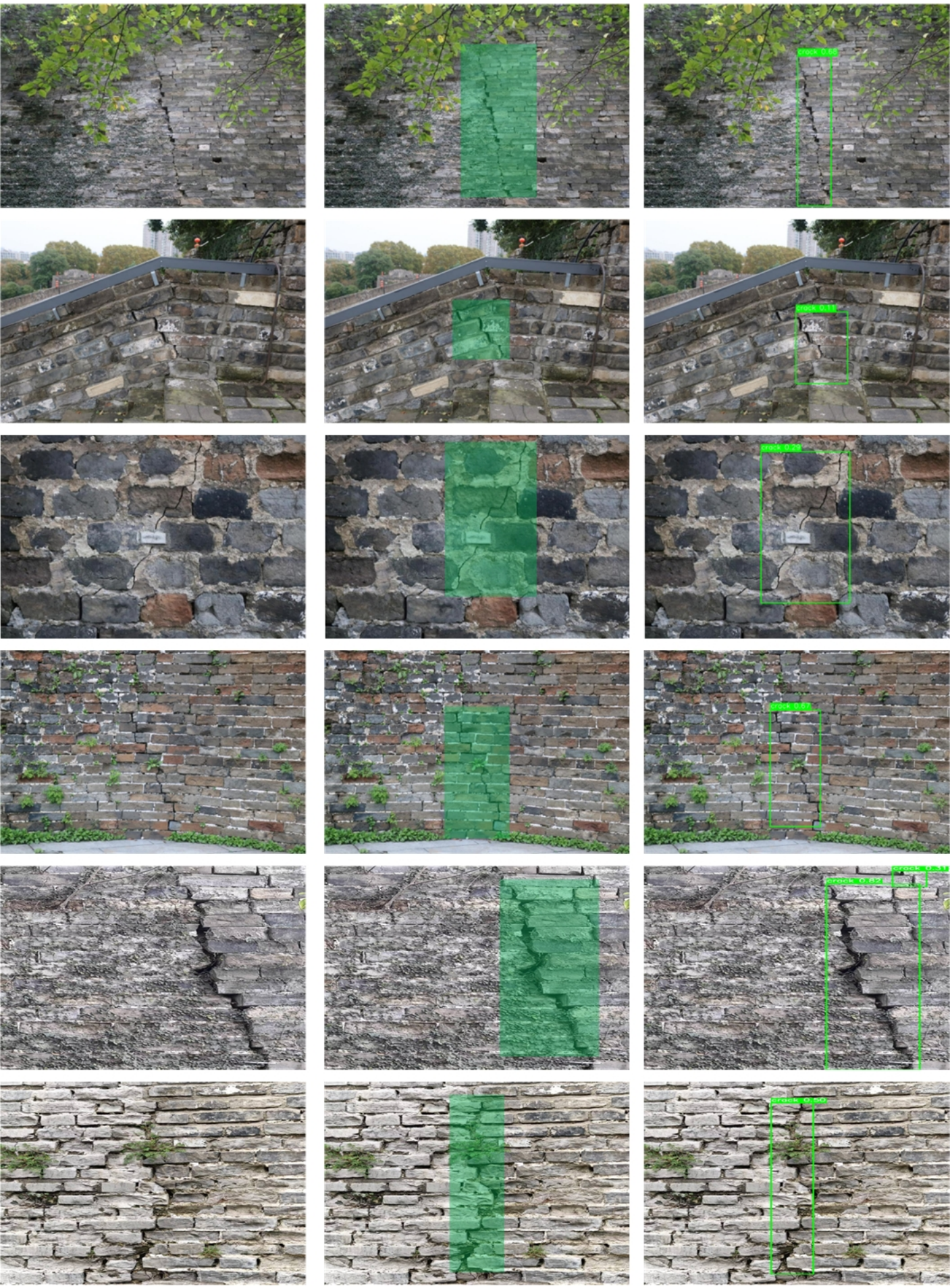
f)

g)

Original image | Ground truth | YOLOv8 | YOLO11 | RetinaNet | Faster R-CNN | Dino | CWADE-Net

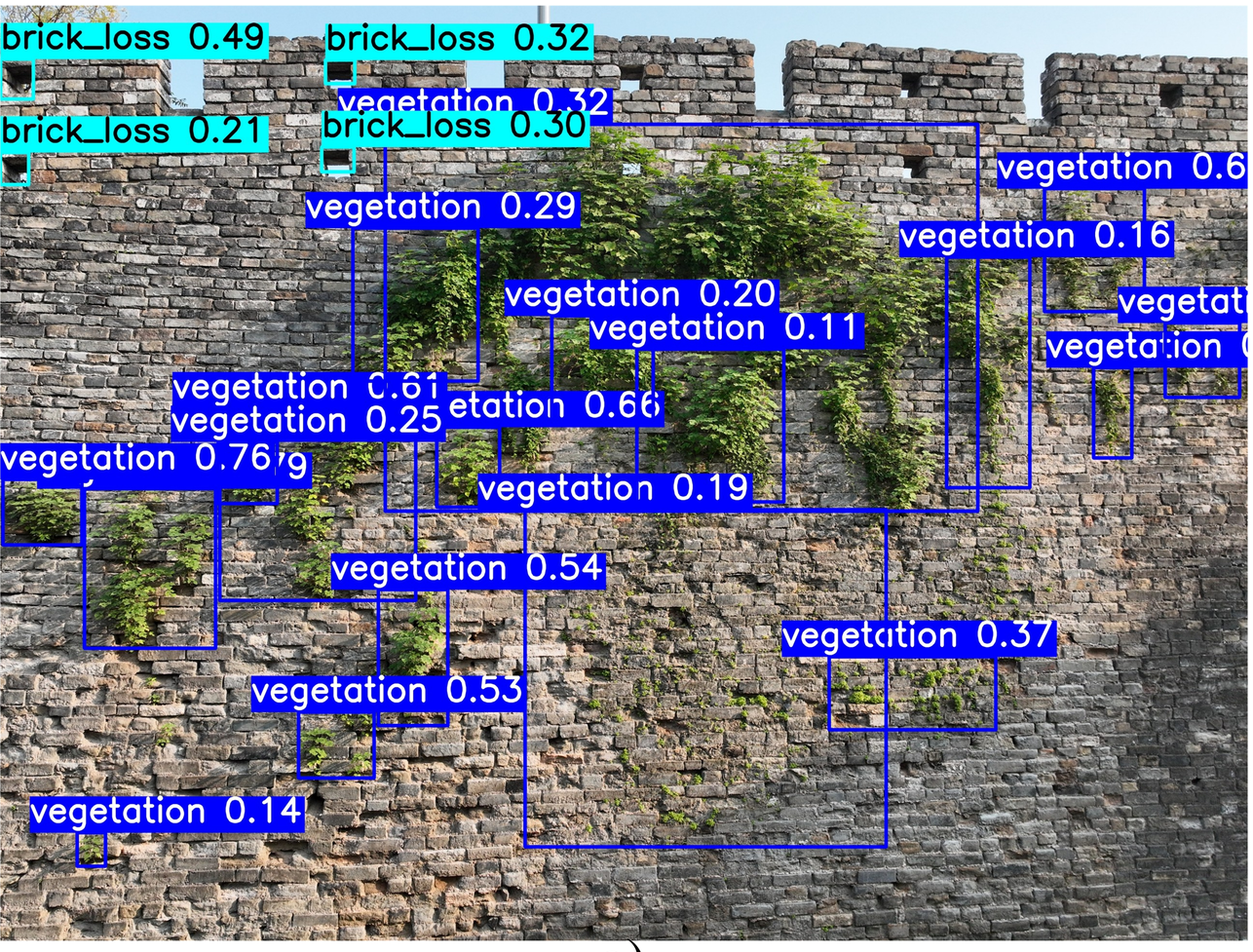
Legend: Herbaceous/Woody vegetation invasion | Brick spalling | Vine-type vegetation invasion


a)
b)
c)
d)
Legend:
Herbaceous/Woody vegetation invasion
Brick spalling
Vine-type vegetation invasion

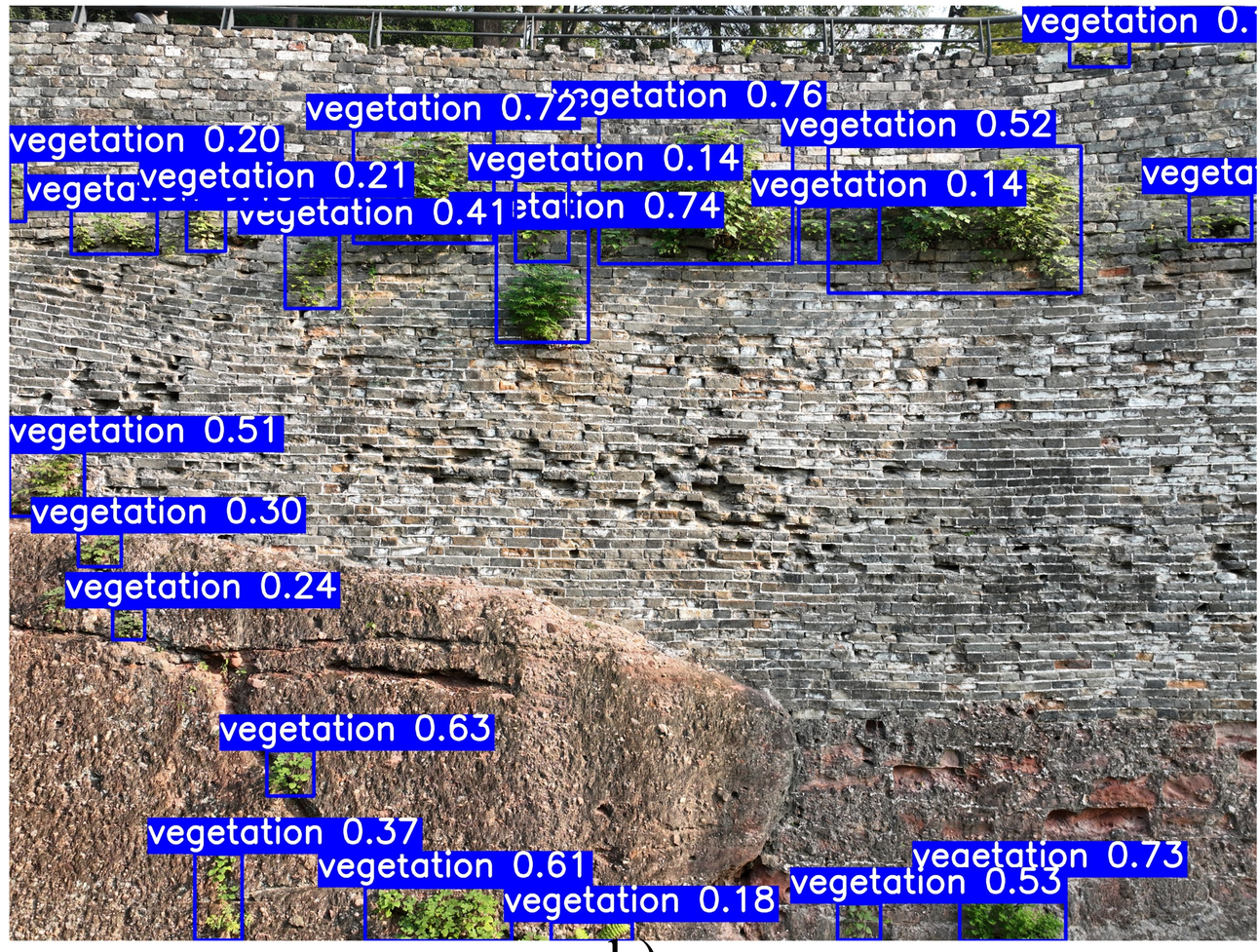


a) b) c)

Legend: Ground truth Crack

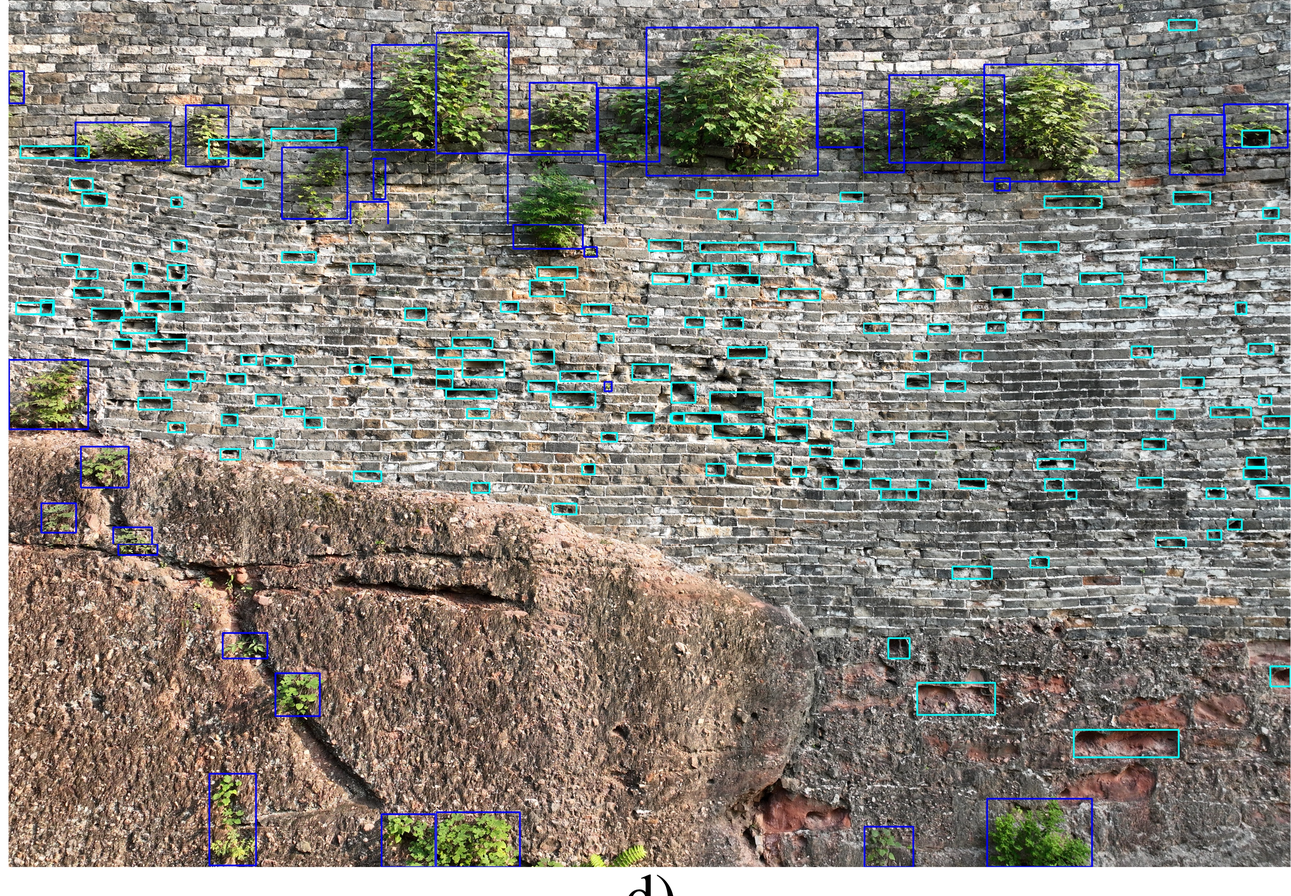


a)



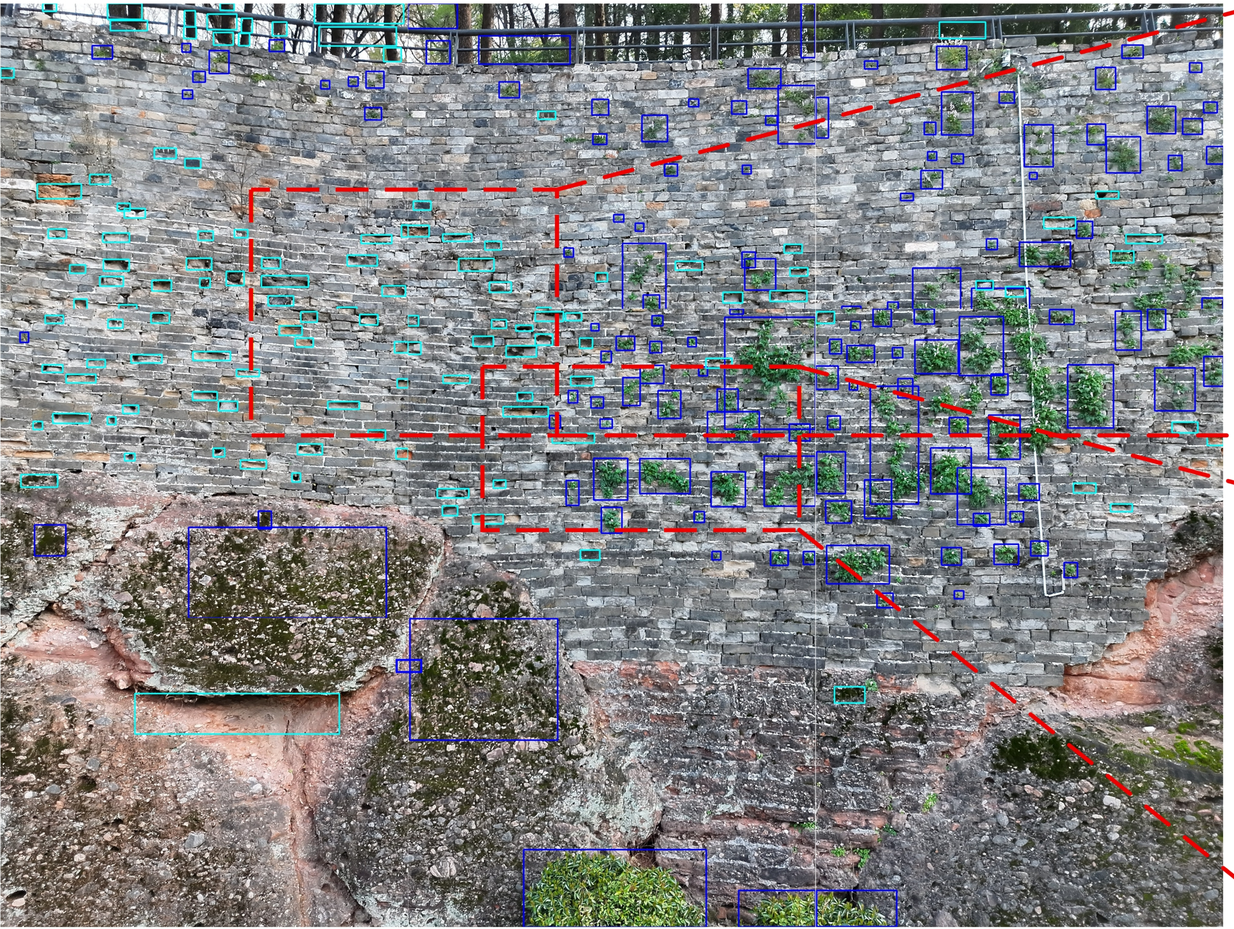
b)

c)

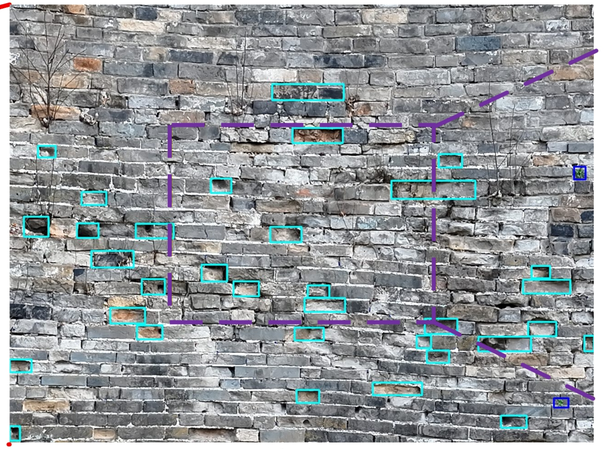


d)

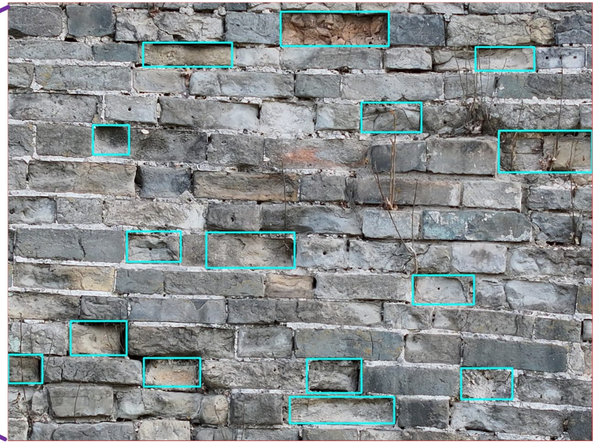
Legend: Herbaceous/Woody vegetation invasion; Brick spalling



a)

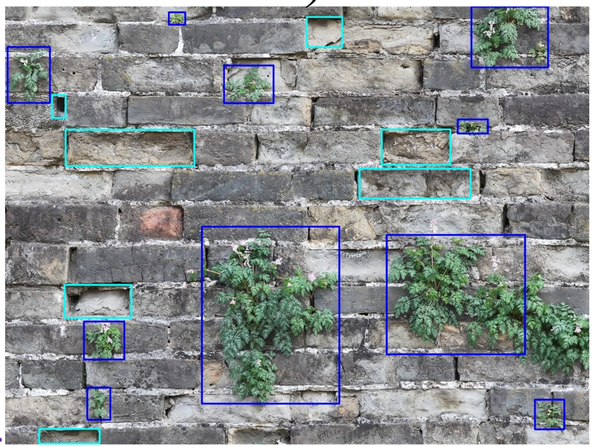


b)



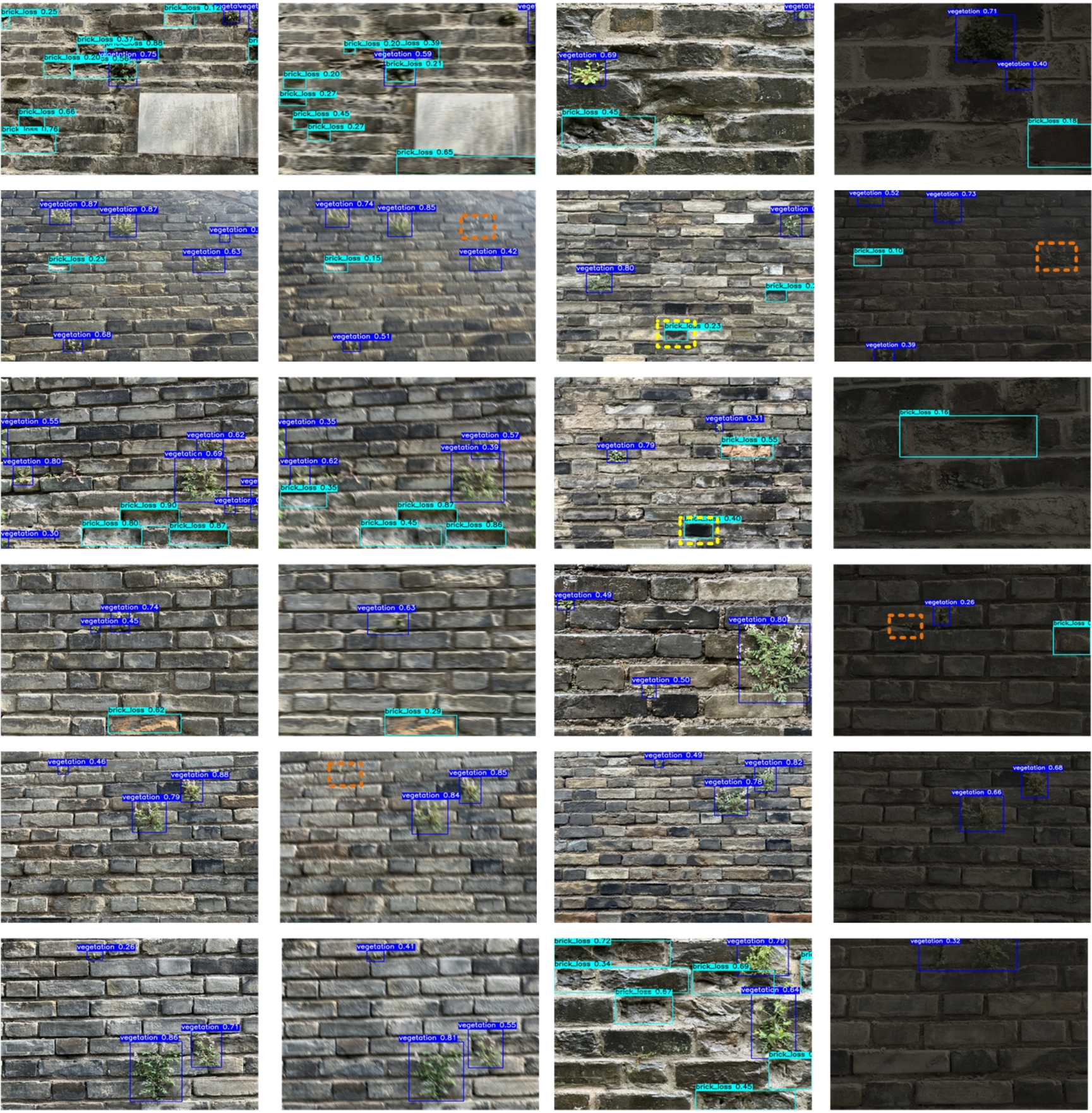
c)

d)



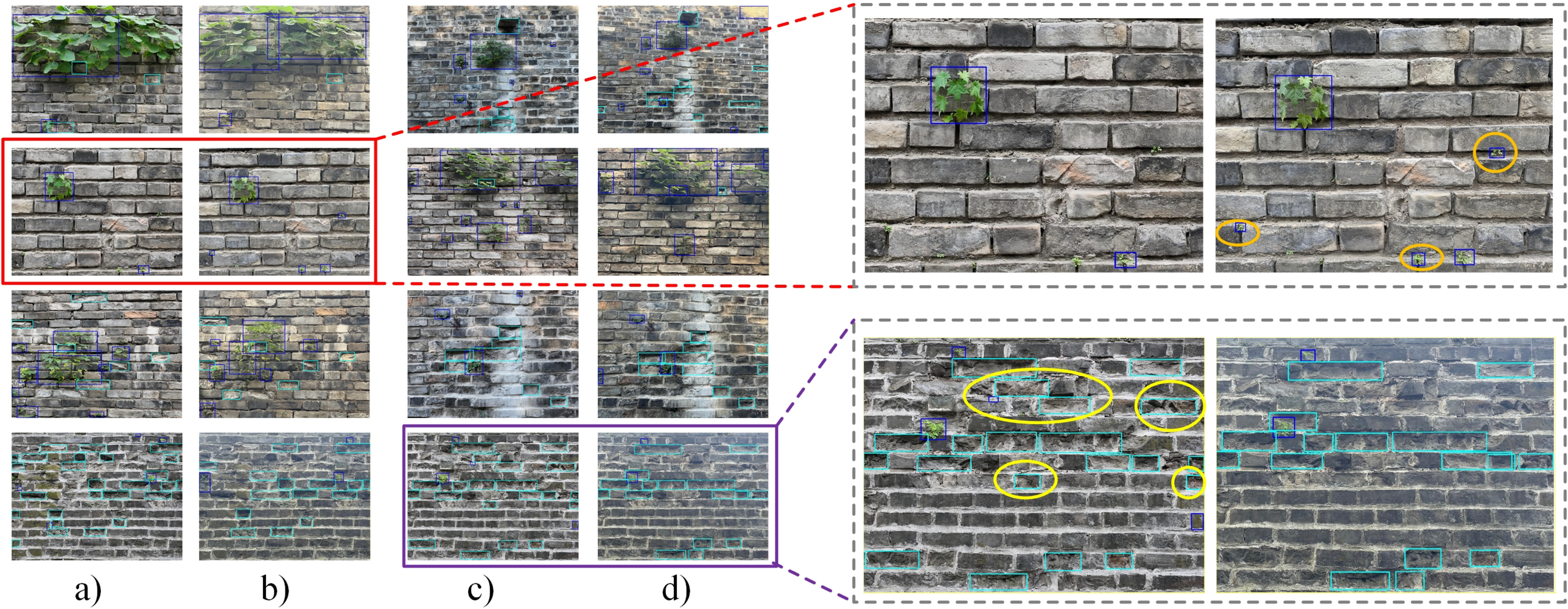
e)

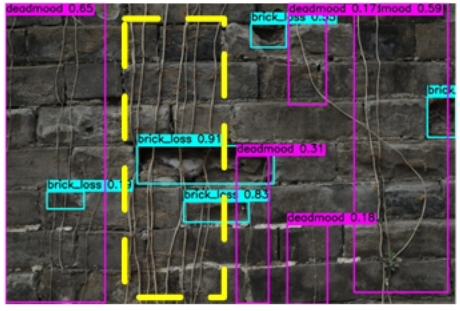
Legend: Herbaceous/Woody vegetation invasion Brick spalling



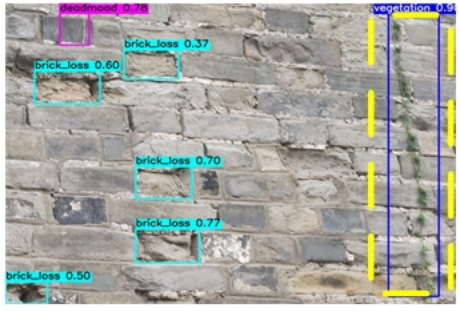
a) b) c) d)

Legend: Herbaceous/Woody vegetation invasion; Brick spalling

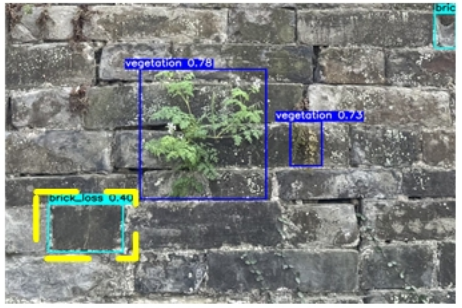

a)
b)
c)
d)
Legend:
Herbaceous/Woody vegetation invasion
Brick spalling



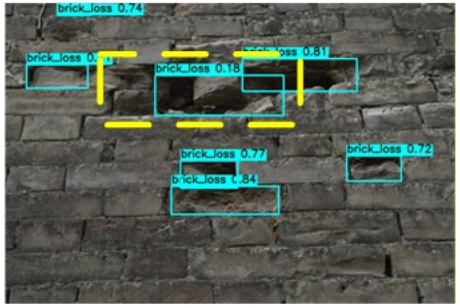
a)



b)

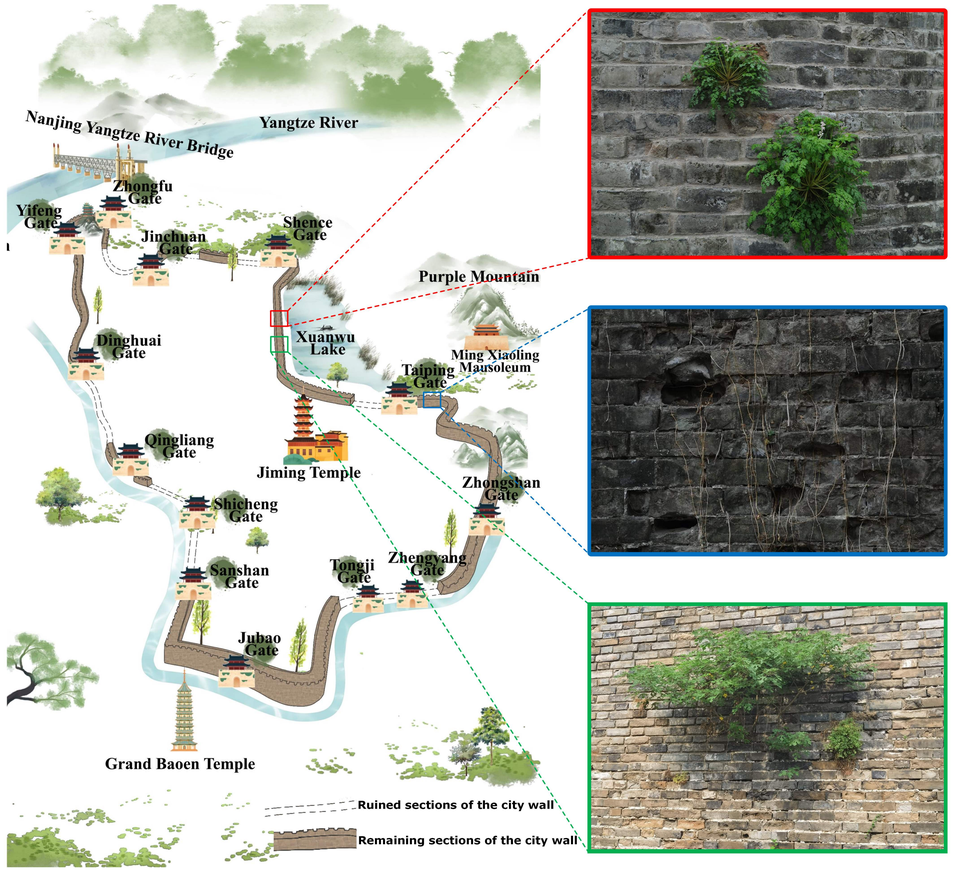


c)

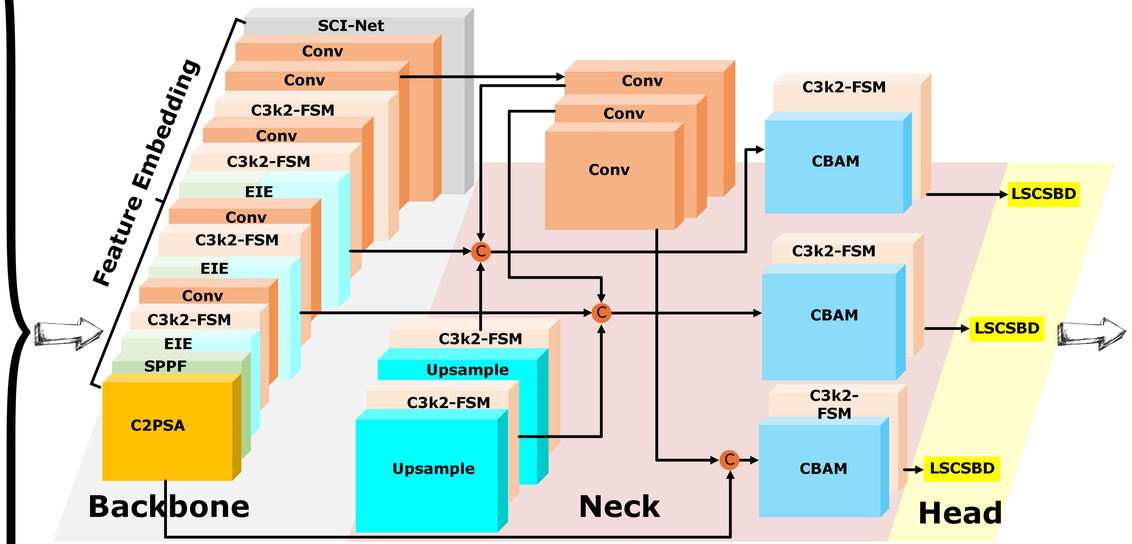


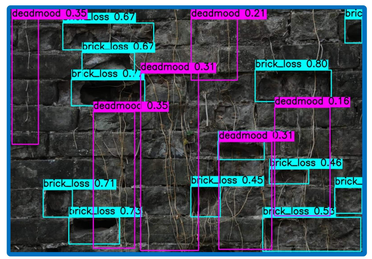
d)

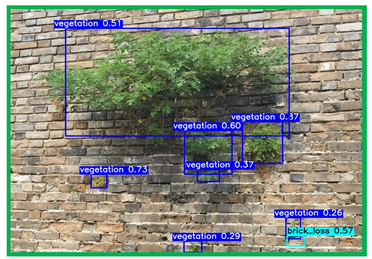
Legend: Herbaceous/Woody vegetation invasion; Brick spalling; Vine-type vegetation invasion



CWADE-Net (City Wall Anomaly Detection Network)







The Nanjing Ming Dynasty Capital City Wall, with a current existing length of 25.091 km and a history of over 600 years

A specialized deep learning framework for vegetation invasion and brick spalling defect detection on the Nanjing Ming Dynasty Capital City Wall

Defect detection results