UDK 625.821.5:004.414.23
Original scientific article/Izvirni znanstveni članek

ISSN 1580-2949
Mater. Tehnol.

A STUDY OF THE INTELLIGENT RECOGNITION OF CONCRETE-STRUCTURE CRACKS BASED ON YOLOv7 AND C2E-Net

ŠTUDIJ INTELIGENTNEGA RAZPOZNAVANJA RAZPOK BETONSKIH ZGRADB NA OSNOVI MODELOV YOLOv7 IN C2E-Net

Linbin Li[1,2,#] Jianfeng Li[3,4,5,#], Yong Luo[6,#,*]

[1]Key Laboratory of Data Science and Intelligent Computing, Fuzhou University of International Studies and Trades, 350202, China
[2]Fuzhou Softimage Information Technology Co., Ltd, Fuzhou 350001, China
[3]Hainan Cloud Spacetime Information Technology Co., Ltd, Sanya572025, Hainan, China
[4]Xing Yun Chen (Hong Kong) Technology Limited, Hong Kong999077, China
[5]School of Civil Engineering, Fuzhou University, Fuzhou Fujian, 350108, China
[6]Wuxi Zhuyun Biotechnology Co., Ltd., Xinan Sub-district, Xinwu District, Wuxi 214000, China
[#]These authors contributed equally to this work.

Prejem rokopisa – received: 2025-09-21; sprejem za objavo – accepted for publication: 2025-12-10

doi:10.17222/mit.2025.1565

Cracks, a common problem in concrete structures, severely compromise their safety and reliability. Accurate and efficient crack identification across diverse environmental conditions is pivotal for ensuring the safe operation and health monitoring of buildings. Accordingly, this paper innovatively proposes a novel intelligent crack-recognition model for concrete structures by integrating YOLOv7 and C2E-Net. First, crack features are extracted via the CA, ELG, and CCF modules of C2E-Net. These features are then inputted into an enhanced YOLOv7 model for fusion, which improves the recognition capability of tiny cracks. Additionally, the Inner-CIOU loss is introduced to optimize small-target detection, addressing the issue that traditional loss functions struggle to accurately identify small targets such as tiny cracks. Following the model's refinement, we carry out ablation experiments. These experiments aim to assess the contribution of each functional module in the model's structure and to uncover how the internal components work together. Finally, to validate the model's performance, this paper selects SSD, Faster-RCNN, and Ghost-YOLO as contrast models and employs multiple evaluation metrics, including the Dice coefficient, IoU, detection accuracy, recall rate, and F1 score, for comprehensive analysis. The experimental results demonstrate that the C2E-Net-YOLOv7 model for concrete-structure-crack intelligent recognition performs exceptionally well during both the training and testing phases, with a Dice coefficient of 0.83, IoU of 0.81, detection accuracy of 91 %, recall rate of 89 %, and F1 score of 0.91. Compared to the traditional SSD model, the Dice coefficient increases by 50.91 %, IoU by 88.37 %, and detection accuracy by 59.65 %; relative to the Ghost-YOLO model, the Dice coefficient improves by 9.21 %, IoU by 22.73 %, and detection accuracy by 5.81 %. The experiments indicate that the YOLOv7 model with the C2E-Net module introduced achieves performance improvements in concrete-structure crack-detection tasks, breaking through the bottlenecks of traditional models and demonstrating remarkable superiority and effectiveness. This model offers an efficient and accurate method for intelligent crack recognition, helping to detect safety hazards promptly and ensuring the long-term stability and safety of buildings.

Keywords: YOLOv7 (object detection); C2E-Net; Inner-CIOU loss; Concrete cracks; Deep learning; Computer vision

Razpoke so pogoste poškodbe na betonskih strukturah, ki lahko resno ogožajo njihovo varnost, stabilnost in zanesljivost. Natančna in učinkovita identifikacija vrste razpok nastalih v različnih pogojih je ključna za varno obratovanje in spremljanje »zdravja« zgradb. Avtorji v tem članku opisujejo inovativni pristop oziroma model za inteligentno razpoznavanje vrste razpok betonskih struktur z integracijo programskih orodij YOLOv7 in C2E-Net. Najprej se značilnosti razpok vgradijo v module CA, ELG in CCF programskega orodja C2E-Net. Te karakteristike oziroma značilnosti razpok se nato vgradijo in zlijejo skupaj v naprednem modelu YOLOv7, kar izboljša sposobnost prepoznavanja tankih razpok. Dodatno so avtorji uvedli tudi orodje Inner-CIOU loss, ki optimizira zaznavanje manjših ciljanih področij. To omogoča reševanje problemov, ki se nanašajo na tradicionalne funkcije izgub. Le-te težko natančno prepoznajo majhna področja z drobnimi razpokami. Po izboljšanju modela so avtorjii izvedli posamezne ablacijske (korekcijske) poskuse. Namen le-teh je bil oceniti prispevek vsakega funkcionalnega modula v strukturi modela in odkriti, kako notranje komponente delujejo skupaj. Nazadnje so avtorji za potrditev učinkovitosti modela za primerjavo izbrali modele SSD, Faster-RCNN v Ghost-YOLO. Za celovito analizo pa so uporabili številne metrike ocenjevanja, vključno s koeficientom Dice, IoU, natančnostjo zaznavanja, stopnjo priklica in oceno F1. Eksperimentalni rezultati so pokazali, da izdelani model za razpoke betonskih struktur imenovan C2E-Net-YOLOv7 izjemno inteligentno deluje, tako med treningom kot tudi med testnimi fazami z Dice koeficientom 0,83, vrednostjo 0,81 za IoU, 91 %-no natančnostjo zaznavanja, 89 % stopnjo odpoklica in 0,91 za oceno F1. V primerjavi s tradicionalnim SSD modelom je Dice koeficient narastel za 50,91 %, IoU za 88,37 %, in natančnost zaznavanja za 9,65 %. Glede na Ghost-YOLO model se je Dice koeficient izboljšal za 9,21 %, IoU za 22,73 % in natančnost zaznavanja za 5,81 %. V zaključkih avtorji članka povdarjajo, da YOLOv7 model z vgraditvijo C2E-Net modula pomembno izboljša naloge povezane z zaznavanjem razpok betonskih zgradb in premaguje ozka grla tradicionalnih modelov. S tem je dokazana izjemna superiornost in učinkovitos tega modela. Ta model ponuja učinkovito in natančno metodo za inteligentno prepoznavanje razpok in pomaga pri pravočasnem odkrivanju nevarnosti in zagotavljanju dolgoročne stabilnosti ter varnosti betonskih zgradb.

Ključne besede: YOLOv7 (model za zaznavanje objektov); C2E-Net; Inner-CIOU loss; razpoke na betonu; globoko učenje; računalniška vizija, umetna inteligenca

*Corresponding author's e-mail:
yongluo1990@outlook.com (Yong Luo)

1 INTRODUCTION

In contemporary construction, concrete serves as the most ubiquitous structural material. Its integrity and stability are profoundly influential in ensuring the secure functioning of constructed facilities and in safeguarding human life and property.[1] Traditional manual visual inspection methods dominate concrete-crack detection but are limited by their dependence on personnel quality and experience, lack of objective evaluation standards, and difficulty in comprehensively and effectively inspecting large and complex concrete structures or special environment projects, thus failing to meet modern society's demands for efficient and precise building-safety monitoring.[2]

As scientific and technological progress continues unabated, particularly with the swift evolution of computer vision and deep-learning techniques, automated approaches for detecting concrete cracks through image recognition and analysis are steadily emerging as a prominent research area and prevailing trend.[3] Park *et al*.[4] merged deep learning with structured light technology, successfully employing this fusion for the identification and measurement of surface cracks in concrete structures. Kim *et al*.[5] introduced a framework for detecting concrete surface cracks, which is built upon a shallow convolutional neural network, attaining remarkable detection precision while maintaining a minimal computational demand. Lei *et al*.[6] used convolutional neural networks (CNNs) for crack classification, obtaining better classification results than support-vector-machine classifiers. Bae *et al*.[7] proposed a super-resolution crack-detection network that improves the resolution of raw data through deep-learning methods, thereby enhancing crack-detection capabilities. Sarhadi *et al*.[8] proposed a new methodology for image segmentation, employing an enhanced UNet++ architecture, to efficiently detect the cracks present in concrete structures.

Crack-extraction and analysis methods based on machine learning and traditional image-processing technologies face challenges such as complex crack backgrounds and fine, unclear crack shapes in practical applications, which can undermine the precision of detection outcomes.[9] In response to these challenges, scholars have put forward various network models such as RCNN, U-Net, YOLO, and SegNet, and improved model performance by optimizing network architectures and introducing different feature-extraction modules. Zhang *et al*.[10] employed the MobileNet network to diminish the parameter count in object-detection models and incorporated the CBAM attention mechanism to enhance the model's accuracy. M. M Brigante and Sumbatyan[11] proposed a stochastic global search algorithm combined with genetic algorithms to identify linear cracks and improve the accuracy of recognition algorithms. Deng *et al*.[12] embedded deformable convolutions into detection networks to enhance the detector's accuracy in detecting out-of-plane cracks. Ji *et al*.[13] put forward the Feature Enhanced and Differential Pyramid Network (FBDPN) to boost crack detection across various scales while preserving the efficacy of features. Li *et al*.[14] presented the CGSW-YOLOv5 algorithm, which incorporates modules such as a concrete-crack feature-enhancement module and a special adaptive multi-scale feature-aggregation attention mechanism, which jointly work to improve the efficiency and accuracy of detection.

However, in practical applications, automated concrete-structure crack-detection technologies face challenges such as difficulties in accurate feature extraction and the recognition of cracks,[15] susceptibility to complex background interference,[16] and high computational resource consumption.[17] In response, this paper proposes the YOLOv7-C2E-Net model for intelligent concrete-structure crack recognition, using YOLOv7 as the baseline. By optimizing the model architecture and computational process, this model improves feature extraction and complex background adaptability while maintaining detection accuracy. It also reduces overfitting risks and exhibits excellent generalization performance. In practical concrete-crack detection tasks, it balances accuracy and applicability (e.g., detecting tiny cracks in complex scenarios of concrete structures such as tunnels and bridge piers), offering an efficient and reliable solution for building-safety monitoring.

2 METHODOLOGY

2.1 Intelligent Recognition Model of Concrete Cracks Based on YOLOv7-C2E-Net

The YOLO algorithm serves as a real-time object-detection method underpinned by fully convolutional neural networks.[18] It works on the principle of processing the entire image in a single forward pass to simultaneously predict bounding boxes and class information. The input image is divided into multiple grid cells, each tasked with predicting a certain number of bounding boxes, confidence scores, and class probabilities. Specifically, bounding boxes define the position and size of objects; confidence scores indicate the likelihood of an object within the box and the accuracy of the prediction; class probabilities represent the probability of the object belonging to a specific category.[19] YOLOv7 inherits the fast and efficient detection features of its predecessors. Equipped with new architectural designs and optimization techniques, it has improved accuracy, generalization ability, and processing speed, making it an important achievement in object detection. As shown in **Figure 1**, its network structure consists of three main components: the Input module, Backbone, and Head.

The YOLOv7-C2E-Net model presented in this paper uses YOLOv7 as the baseline and integrates the core modules of C2E-Net, combining their advantages. The YOLOv7 component employs its unique input module to standardize concrete images and apply Mosaic data augmentation, simulating complex real-world conditions and

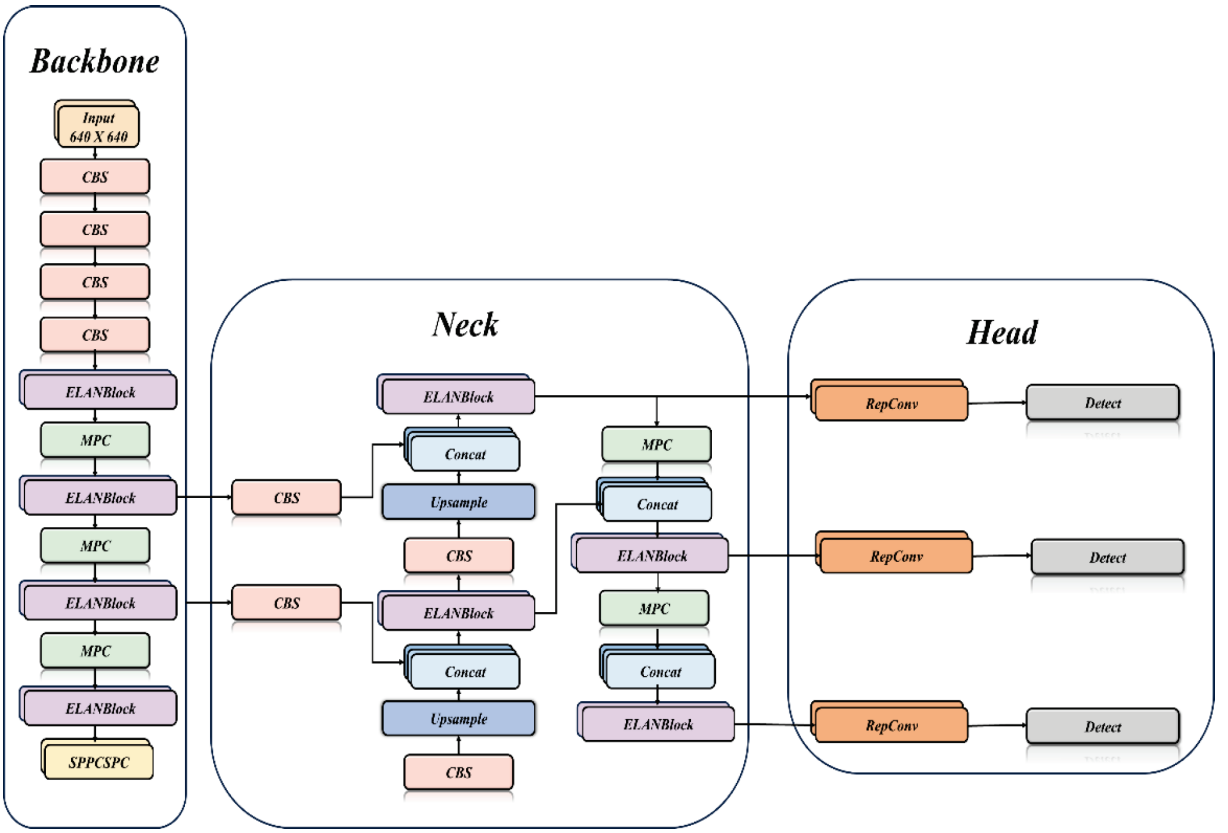

Figure 1: YOLOv7 Network Architecture

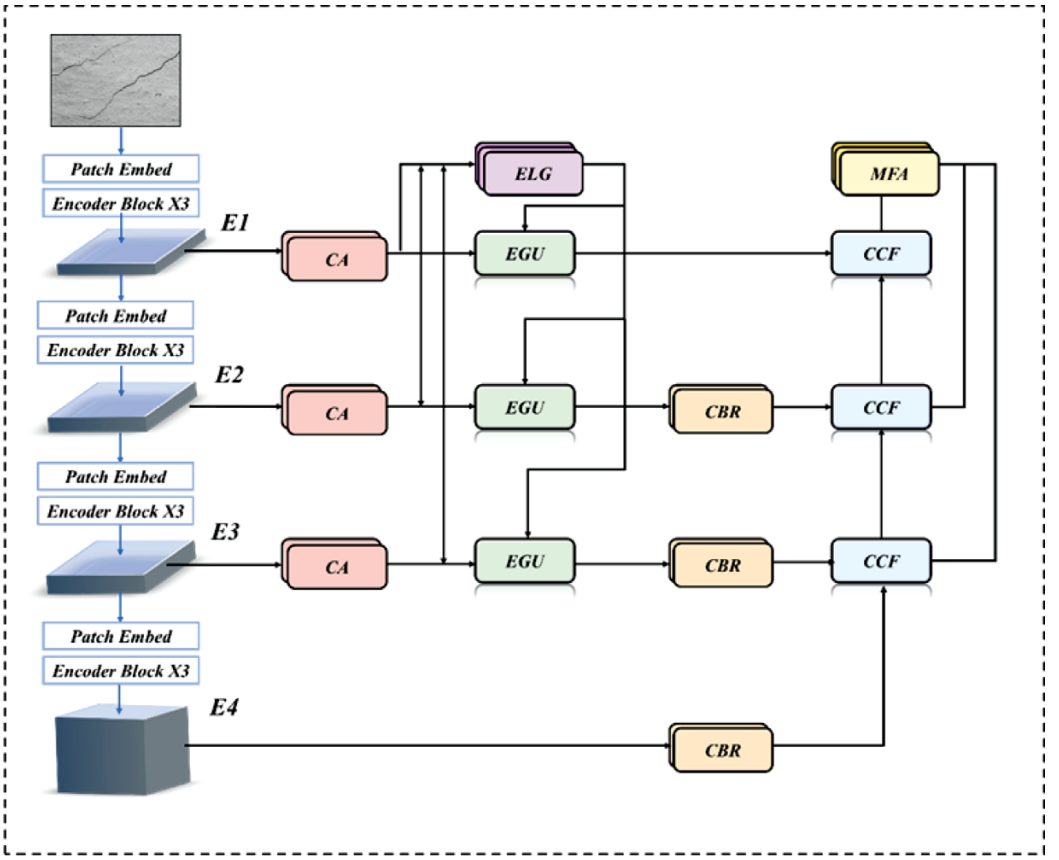

Figure 2: C2E-Net Architecture

enhancing crack recognition under varying lighting and angles. Its backbone feature-extraction network, based on a modified EfficientNet architecture, uses CBS modules to capture the local texture features, ELAN modules to fuse multi-branch features, and MP modules to reduce feature-map dimensions while preserving key crack features, achieving efficient feature extraction. In the detection head, the SPPCSPC module fuses multi-scale crack features, the PAN module strengthens feature semantics, and the Rep Conv module balances feature representation during training and computational efficiency during inference.

Serving as a high-precision module for feature enhancement, C2E-Net represents an advanced network that leverages edge-learning guidance. It employs a cascaded attention mechanism and incorporates context-aware cross-layer feature fusion, all aimed at achieving efficient and accurate extraction of image features.[20] As shown in **Figure 2**, C2E-Net uses a PVT-V2-based encoder-decoder structure and innovatively integrates three modules: cascaded attention (CA), edge learning guidance (ELG), and context-aware cross-layer fusion (CCF). These modules enable comprehensive feature capture from local details to global context, strongly supporting subsequent object detection.

(1) The Cascaded Attention (CA) Module, as depicted in **Figure 3**, is formulated to address the semantic discrepancy between the encoder and decoder components. It employs a multi-scale feature-extraction strategy through dilated convolutions with progressive dilation rates, complemented by cross-attention mechanisms to bridge the semantic divide. The module operates by receiving encoder features, integrating cross-dimensional attention, and distributing them across six parallel processing pathways. Subsequent decomposition occurs through five distinct perspectives, extracting hierarchical low-level features from the encoder stream. Finally, a multi-branch fusion strategy is implemented, combining spatial semantic information from the five pathways using Batch Normalization (BN) and Squeeze-and-Excitation (SE) modules. This architectural design empowers the CA module to perform multi-scale feature capture while harmonizing low-level spatial granularity with high-level contextual semantics through cross-attention interactions, establishing a robust feature representation framework for subsequent fusion and recognition tasks.

(2) Edge Learning Guidance (ELG) Module is developed to tackle the limitations of conventional edge-detection methods, which often fail to capture subtle edge details while introducing irrelevant noise. Leveraging three parallel processing pathways that receive inputs from the CA module's outputs, it employs a sophisticated deep convolutional neural architecture to identify edge characteristics. This configuration enables the extraction of detailed spatial information and edge-specific cues, which subsequently guide feature-integration processes and enhance the precision of object-boundary feature extraction. By augmenting both semantic understanding and edge-representation capabilities, the ELG module strengthens the overall edge-characterization framework, facilitating more nuanced and comprehensive information mining.

(3) Context-Aware Cross-Layer Fusion (CCF) Module As illustrated in **Figure 4**, the CCF module tackles challenges arising from object variations caused by scale fluctuations, viewing angle differences, and illumination disparities. Its architecture incorporates a dual-branch configuration featuring a Multi-Scale Channel Attention

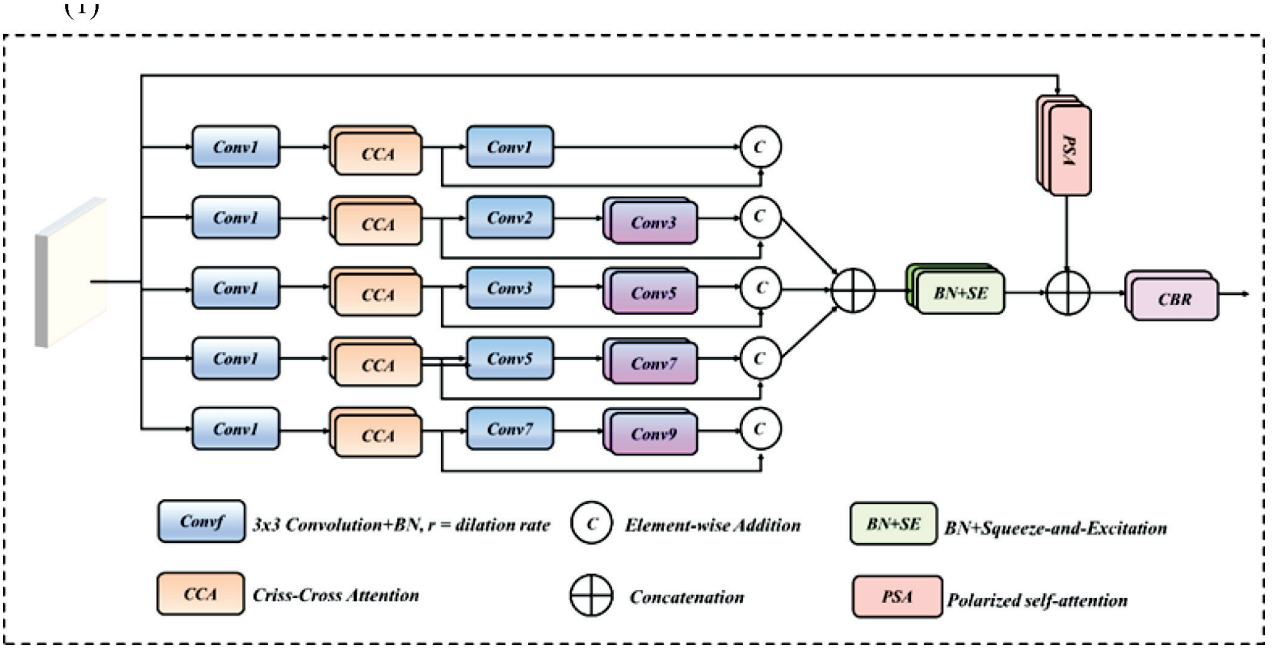

Figure 3: CA Module Architecture

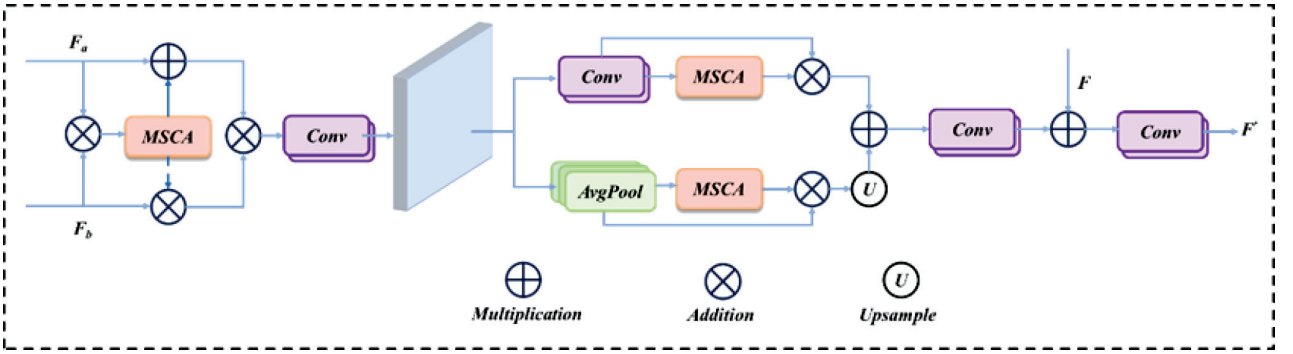

Figure 4: CCF Structure

(MSCA) mechanism, enabling the simultaneous capture of localized and holistic contextual data. Through cross-layer feature aggregation, the module synthesizes multi-level information, distilling comprehensive global context from hierarchical feature representations. This fusion strategy effectively disambiguates foreground objects from complex backgrounds in ambiguous regions while attenuating the adverse effects of geometric transformations and environmental variations on detection performance through contextual reasoning. The resultant enhanced feature representation establishes a robust foundation for accurate object localization under challenging conditions.

The YOLOv7-C2E-Net model excels in concrete crack identification. Its hybrid design combines YOLOv7's swift detection with C2E-Net's fine-grained feature extraction, enabling real-time, precise detection of tiny, complex-shaped cracks. The CA and ELG modules work together to sharpen the model's perception of crack details, reducing false and missed detections. The CCF module integrates features across layers, enhancing adaptability to various crack types and sizes, and ensuring stable detection in complex concrete structures. The C2E-Net module improves the model's accuracy and generalization ability in complicated environments, effectively balancing detection efficiency and accuracy.

2.2 Improved Algorithm Based on Inner-CIOU Loss Function

The YOLOv7 algorithm's objective function is mainly composed of three parts: classification loss, localization loss, and bounding-box regression loss. The classification loss uses the cross-entropy loss function to measure the difference between predicted and actual categories, helping the model to distinguish features of different classes. For localization loss, the Complete Intersection over Union (CIOU) loss function is employed. It considers factors like bounding-box overlap area, central point distance, and aspect-ratio consistency. In bounding-box regression loss, a combination of mean-squared error loss and CIOU loss is used to enhance the coordinate prediction accuracy. The CIOU loss function is an extension of the conventional IOU metric. It considers center-point separation and aspect-ratio balance to better assess the alignment between predicted and actual bounding boxes. However, despite its advantages over traditional IOU-based methods, CIOU still has limitations. These include task-specific performance instability, lack of adaptive tuning mechanisms, and suboptimal generalization across diverse detection scenarios.

$$L_{\mathrm{CIOU}} = 1 - L_{\mathrm{IoU}} + \frac{\rho^2(b, b^{\mathrm{gt}})}{c^2} + av \tag{1}$$

$$L_{\mathrm{IOU}} = \frac{\left| b \cap b^{\mathrm{gt}} \right|}{\left| b \cup b^{\mathrm{gt}} \right|} \tag{2}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{3}$$

$$a = \frac{v}{(v + (1 - L_{\mathrm{IoU}})} \tag{4}$$

In the formula, w and h denote the width and height of the predicted bounding box, respectively, whereas w^{gt} and h^{gt} correspond to the dimensions of the ground-truth bounding box. V captures the correlation in aspect ratio between the predicted and ground-truth bounding boxes. A serves as the weight coefficient, and c represents the diagonal distance of the enclosing box that contains both the predicted and ground-truth bounding boxes. $\rho^2(b, b^{\mathrm{gt}})$ is the squared Euclidean distance between the centers of the predicted bounding box (b) and the ground-truth box (b^{gt}). LCIOU refers to the Complete-IOU loss, and LIOU is the standard Intersection-over-Union loss, which is the overlapping ratio of the predicted bounding box and the ground-truth box.

To enhance YOLOv7's small object-detection accuracy, we propose the Inner-CIOU improvement algorithm. Building upon the original CIOU formulation, this method incorporates a dynamic scaling factor (S_{ratio}) to adjust the auxiliary bounding-box dimensions. By optimizing the correlation between the auxiliary box scaling and the IOU values, the algorithm enables a more precise overlap calculation between the predicted and ground-truth boxes. This results in improved localization precision, accelerated model convergence, and enhanced re-

gression efficiency during training. The Inner-CIOU loss architecture is illustrated in **Figure 5**.

$$B_{\mathrm{l}} = x_{\mathrm{c}} - \frac{w \times S_{\mathrm{ratio}}}{2} \tag{5}$$

$$B_{\mathrm{r}} = x_{\mathrm{c}} + \frac{w \times S_{\mathrm{ratio}}}{2} \tag{6}$$

$$B_{\mathrm{t}} = y_{\mathrm{c}} - \frac{w \times S_{\mathrm{ratio}}}{2} \tag{7}$$

$$B_{\mathrm{b}} = y_{\mathrm{c}} + \frac{w \times S_{\mathrm{ratio}}}{2} \tag{8}$$

$$I_{\mathrm{inter}} = \left(\min(B_{\mathrm{r}}^{\mathrm{gt}}, B_{\mathrm{r}}) - \max(B_{\mathrm{l}}^{\mathrm{gt}}, B_{\mathrm{l}})\right) \times \times \left(\min(B_{\mathrm{b}}^{\mathrm{gt}}, B_{\mathrm{b}}) - \max(B_{\mathrm{t}}^{\mathrm{gt}}, B_{\mathrm{t}})\right) \tag{9}$$

$$U_{\mathrm{nion}} = (w^{\mathrm{gt}} \times h^{\mathrm{gt}}) \times (S_{\mathrm{ratio}})^2 + + (w \times h) \times (S_{\mathrm{ratio}})^2 - I_{\mathrm{inter}} \tag{10}$$

$$L_{\mathrm{IOU}^{\mathrm{Inner}}} = \frac{I_{\mathrm{inter}}}{U_{\mathrm{union}}} \tag{11}$$

$$L_{\mathrm{Inner\text{-}CIOU}} = L_{\mathrm{CIOU}} + L_{\mathrm{IOU}} - L_{\mathrm{IOU}^{\mathrm{Inner}}} \tag{12}$$

In the formula, (x_c, y_c) represents the x and y coordinates specifying the center location of the predicted bounding box; B_l, B_r, B_t, B_b denotes the boundary coordinates (left, right, top, bottom) encompassing the predicted region, while B_l^{gt}, B_r^{gt}, B_t^{gt}, B_b^{gt} correspond to those of the ground-truth box; S_{ratio}, the scale factor ratio controlling the auxiliary bounding box's size, ranges from 0.5 to 1.5; Union is the combined area of the predicted and ground-truth boxes; L_{IOU}^{Inner} and $L_{Inner\text{-}CIOU}$ represent the Inner-IOU loss and Inner-CIOU loss values, respectively.

3 MODEL INSTANTIATION ANALYSIS

3.1 Dataset Construction

To support the rigorous training and testing requirements of this model's deep neural networks for extensive image datasets, the study compiled a corpus exceeding 1400 high-resolution photographs of concrete cracks through various means, such as reviewing major concrete monitoring reports, taking on-site photographs at construction sites, and collecting data from the internet. After acquiring the initial image data, to ensure it meets the stringent requirements for model training, a rigorous evaluation and meticulous screening of the selected photographs were carried out. The assessment mainly focused on the image's pixels, clarity, and the integrity of crack features. Ultimately, more than 1500 high-quality images of concrete cracks, with a resolution of 640×640 pixels and captured under various scenarios and lighting conditions, were confirmed as the basic dataset for model training. As shown in **Figure 6**, to further enhance the model's robustness and generalization ability, data augmentation was performed by applying methods such as rotation, mirroring, cropping, and adjusting brightness, expanding the number of images to over 1500. Given the deep neural network's high dependence on sample labels during the feature-extraction phase, the labelling data annotation tool was employed to accurately annotate the images. The tool precisely delineated the boundaries of the cracks and generated standardized XML label files. This meticulously constructed high-quality dataset has laid a solid foundation for model training and testing, effectively supporting the model in achieving efficient and accurate image recognition and analysis in subsequent processes.

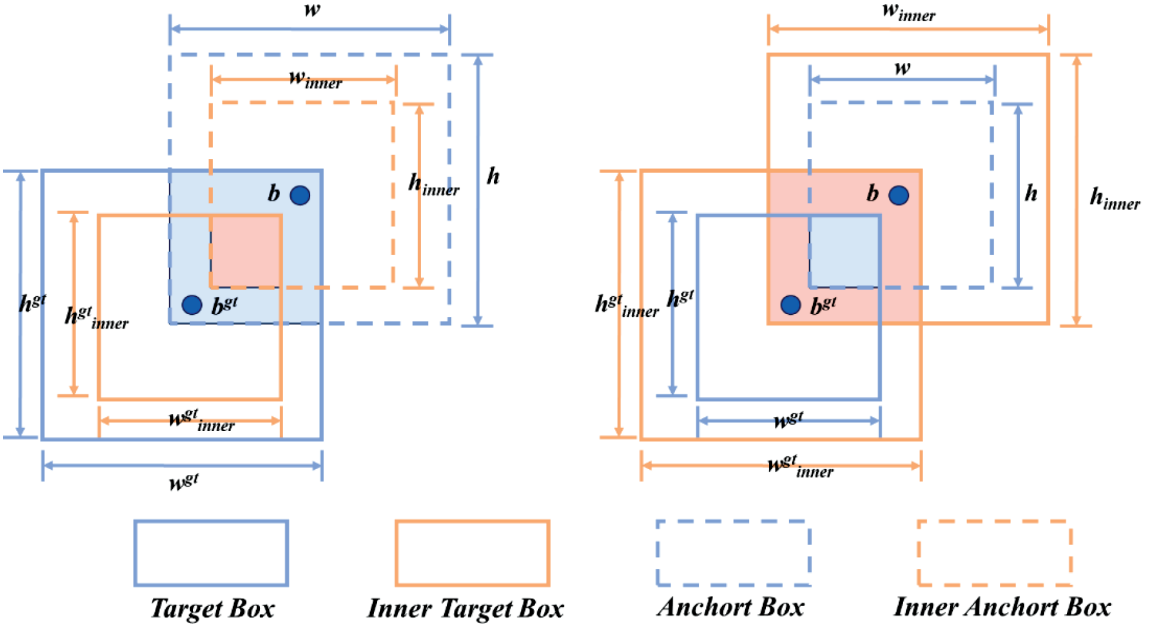

Figure 5: Inner-CIoU Mathematical Principle

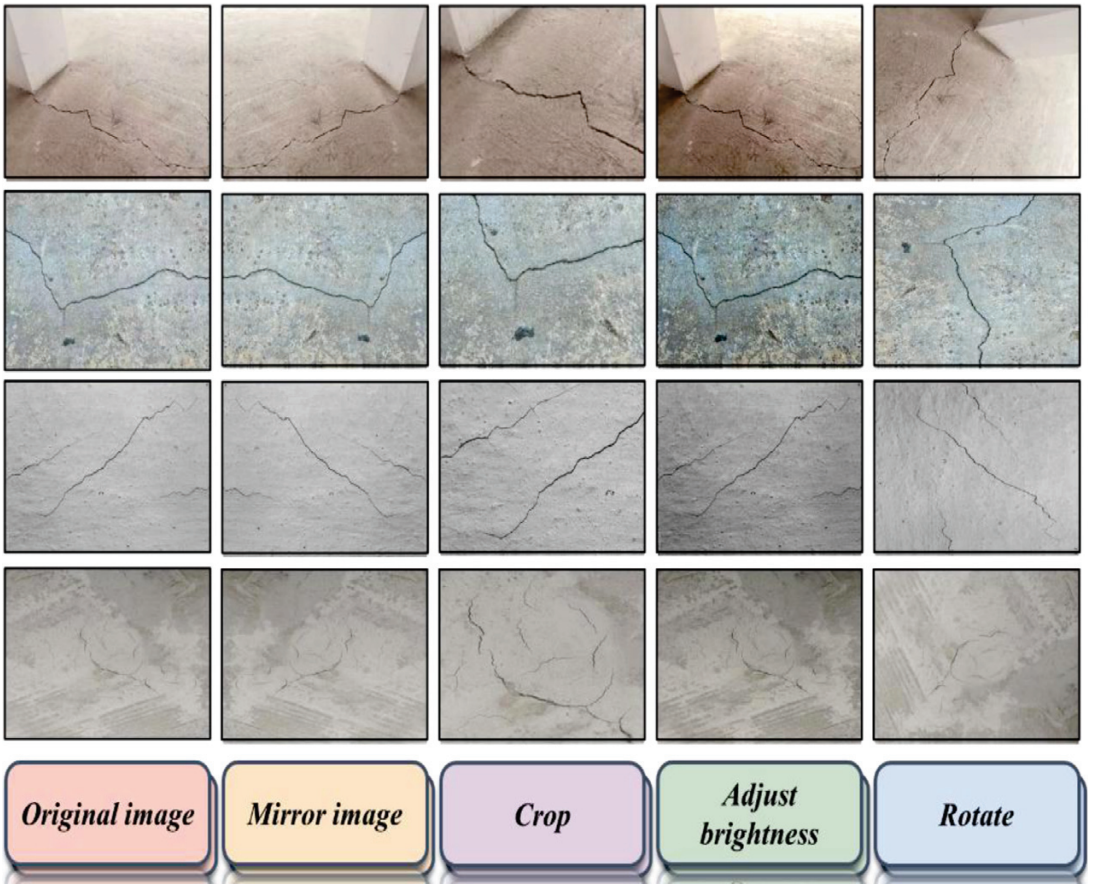

Figure 6: Image Data Augmentation Example

3.2 Experimental Environment and Parameter Settings

The experimental setup comprises: Windows 11 operating system, NVIDIA Quadro RTX 3090 GPU acceleration, Pytorch-based deep-learning architecture, and Python implementation for model execution. A thorough environmental validation was conducted to ensure experimental stability, with key configurations summarized in **Table 1**.

Table 1: Environmental Configuration

Name	Model
GPU	NVIDIA Quadro RTX 3090
GB	24
Operating System	Windows 11
Pytorch	2.4
Python	3.10.5
CUDA	11.8
Torchvision	0.19.1

The model training configuration includes the following: input resolution of 640×640 pixels, batch size of 32, initial learning rate set at 0.001, and stochastic gradient descent (SGD) optimization with momentum. The learning rate is adjusted using a cosine annealing schedule. The dataset is split into training and testing sets following a 7:3 ratio, and the total training duration is 2000 epochs. **Table 2** provides detailed hyperparameter specifications.

Table 2: Training Parameters

Parameter	Parameter value
Input image size	640×640
Epochs	2000
Batch size	32
Warmup	3
Optimizer	SGD
Initial learning rate	0.01
Cosine annealing algorithm	0.2

3.3 Evaluation Metrics

On object detection tasks, a multi-metric evaluation framework is adopted for comprehensive performance assessment, incorporating the Dice coefficient, IoU (Intersection over Union), detection precision, recall rate, and F1-score. The Dice coefficient measures spatial overlap consistency between predictions and ground truth annotations, while IoU serves as the primary localization accuracy indicator. Both metrics operate within the [0,1] interval, with higher values denoting superior spatial alignment and detection precision. Detection precision reflects the ratio of correctly identified objects to

total predicted instances, whereas recall rate quantifies the proportion of accurately detected objects relative to all ground-truth annotations. The F1-score, computed as the harmonic mean of precision and recall, provides a balanced assessment of detection completeness and correctness. This multi-dimensional evaluation paradigm ensures thorough performance characterization, capturing both localization fidelity and classification performance while mitigating the limitations of single-metric assessments. The calculation formulas for each evaluation metric are as follows:

$$Dice = \frac{|A| + |B|}{2|A \cap B|} \tag{13}$$

$$IOU = \frac{A \cup B}{A \cap B} \tag{14}$$

$$Precision = \frac{TP}{FP + TP} \times 100\% \tag{15}$$

$$Recall = TP\frac{TP}{FP + FN} \times 100\% \tag{16}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \tag{17}$$

In these formulas, A denotes the predicted segmentation results, while B represents the ground-truth segmentation labels. The term A∩B signifies the overlap between the predicted and true regions, whereas A∪B indicates the combined extent of both regions. The notation |A| corresponds to the cardinality of set A, quantifying the number of elements in the predicted results, and đBđ denotes the cardinality of set B, reflecting the count of elements in the ground-truth labels. In addition, TP (True Positive) represents scenarios in which the model correctly identifies positive samples, whereas FP (False Positive) indicates instances where the model mistakenly classifies negative samples as positive.

3.4 Ablation Experiments

To verify the effectiveness of the C2E-Net and Inner-CIOU loss function improvements in our model, we conducted four sets of ablation experiments on each module without altering the dataset division. These experiments used YOLOv7 as the baseline model. In the table, "√" indicates the improvements made to the baseline algorithm. The results of the ablation experiments are shown in **Table 3**.

Table 3: Ablation Study Model Evaluation Metrics

Model	C2E-Net	Inner-CIOU	Dice	IoU	Precision (%)	Recall (%)	F1
1			0.65	0.53	78	75	0.76
2	√		0.71	0.61	82	80	0.81
3		√	0.72	0.63	84	83	0.83
4	√	√	0.83	0..81	91	89	0.91

As can be seen from **Table 3**, after introducing the C2E-Net and Inner-CIOU loss function optimization algorithms, the model's performance was enhanced. The Dice coefficient increases from 0.65 to 0.71 (a 9.23 % improvement) and the IoU from 0.53 to 0.61 (a 15.09 % boost) with C2E-Net alone. When only the Inner-CIOU loss function is applied, the Dice coefficient rises from 0.65 to 0.72 (up by 10.77 %) and the IoU from 0.53 to 0.63 (an 18.87 % increase). This indicates C2E-Net effectively extracts image features, and the Inner-CIOU loss function improves small-target identification by more accurately measuring the bounding-box overlap, with the latter showing a slightly better enhancement effect than the former. When combined, the Dice coefficient reaches approximately 0.83 and the IoU approximately 0.81, representing improvements over the original model of 27.69 % and 52.83 %, respectively. This shows that C2E-Net enables efficient and accurate image-fea-

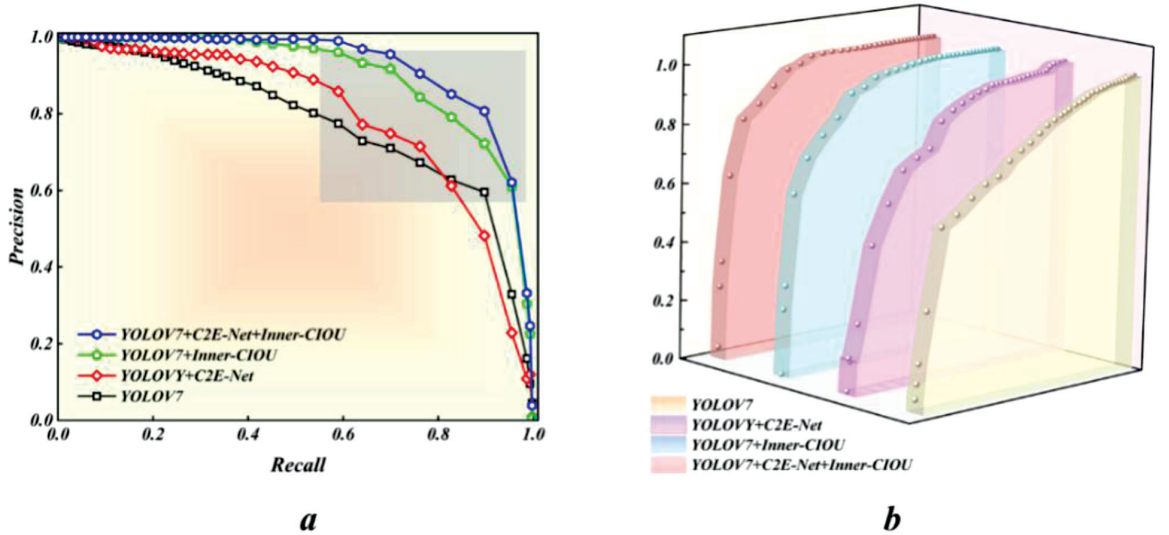

Figure 7: P-R Curves of Baseline and Improved Models (a: Scatter Plot; b: Wall Plot)

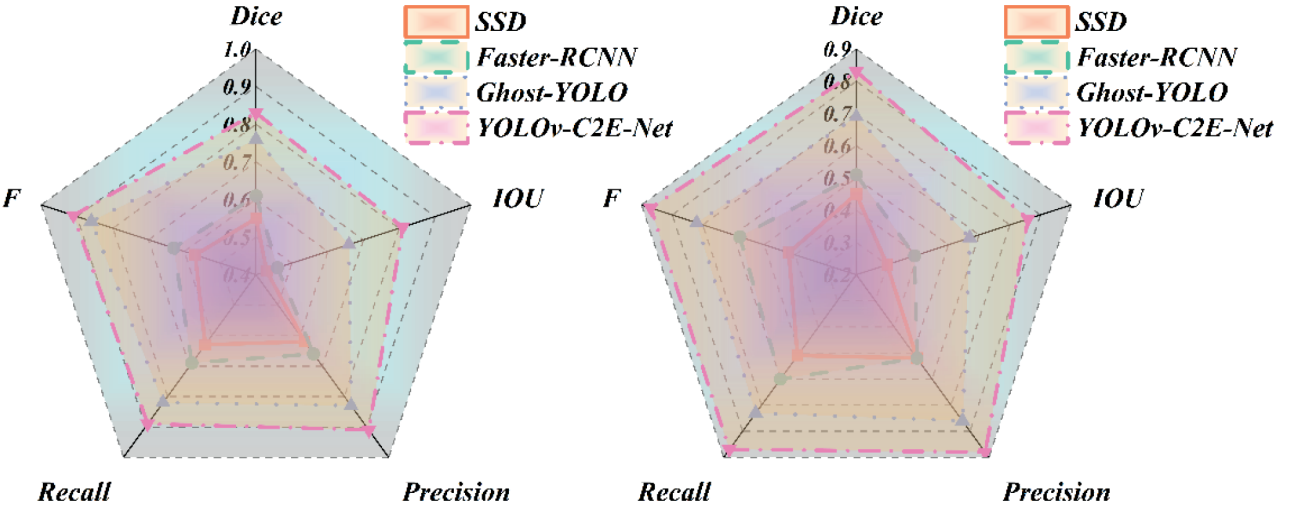

Figure 8: Model Evaluation Metrics (a: Training Set; b: Testing Set)

ture extraction by enhancing key-feature channel-weight allocation and cross-layer information fusion. The Inner-CIOU loss function, which adds target center distance and aspect-ratio penalties when calculating the bounding-box overlap, improves small-target recognition accuracy and refines overlap calculation.

As shown in **Figure 7**, after introducing the C2E-Net and the improved Inner-CIOU loss function algorithm, the area under the Precision-Recall (P-R) curve of the model is significantly larger than that of the baseline model. Specifically, the area under the P-R curve of YOLOv7-C2E-Net is 0.781, which is 42 %, 38 % and 11 % higher than that of SSD, Faster-RCNN, and Ghost-YOLO, respectively. Moreover, the model maintains higher precision across different recall rates. The two improvements work together to make the model more robust and adaptable in practical use, enabling efficient and accurate identification of the concrete cracks even against complex backgrounds.

3.5 Comparative Experiments on Various Object Detection Algorithms

To assess the YOLOv7-C2E-Net model's performance and efficiency, SSD, Faster-RCNN, and Ghost – YOLO were chosen as baseline models for comparison. Multiple common object detection metrics, including the Dice coefficient, IoU, detection accuracy, recall rate, and F1 score, were used for the evaluation. **Figure 8** presents the experimental results. Analysis shows that traditional SSD and Faster-RCNN models exhibit weak performance on both training and test sets, underperforming compared to YOLO-based models. This suggests YOLO-based improved models are more suited for complex concrete crack detection tasks. On the training set, Ghost-YOLO performs well with a Dice coefficient of 0.76, IoU of 0.66, and F1 score of 0.86. But on the test set, its performance drops significantly. The Dice coefficient, IoU, and F1 score decrease by 10.15 %, 15.79 %, and 19.44 % respectively. While it demonstrates high accuracy in identifying concrete cracks within the training dataset, it struggles to generalize by capturing broader contextual features, revealing susceptibility to overfitting and limited adaptability to practical detection environments. Conversely, YOLOv7-C2E-Net exhibits consistent performance across both training and validation datasets. Specifically, on the validation set, it achieves a Dice coefficient of 0.83, an IoU of 0.81, and an F1 score of 0.91, with marginal performance declines of only 2.5 %, 6.58 %, and 4.59 % respectively, compared to training set results. This suggests that the C2E-Net and Inner-CIoU loss function improvements adopted in this paper can significantly enhance the YOLO base model's performance. They enable better feature extraction of the concrete cracks, demonstrating strong feature-extraction and generalization abilities, and can effectively handle the complex task of concrete crack detection.

To further validate the model's applicability, we conducted a 3D visual analysis of the P-R curves for different models, with the results presented in **Figure 9**. As

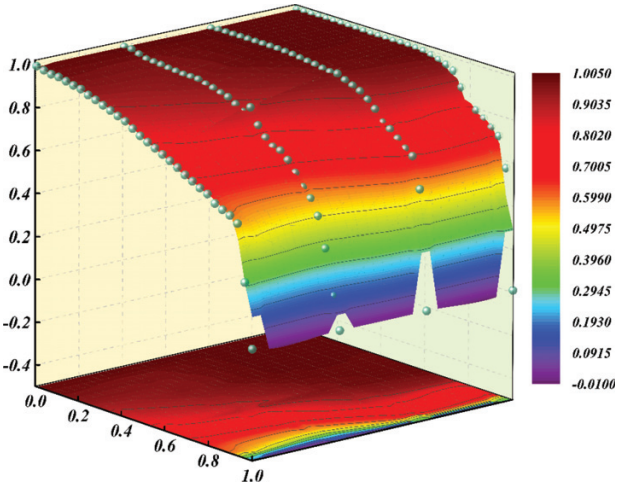

Figure 9: 3D Visualisation of P-R Curves for Different Models

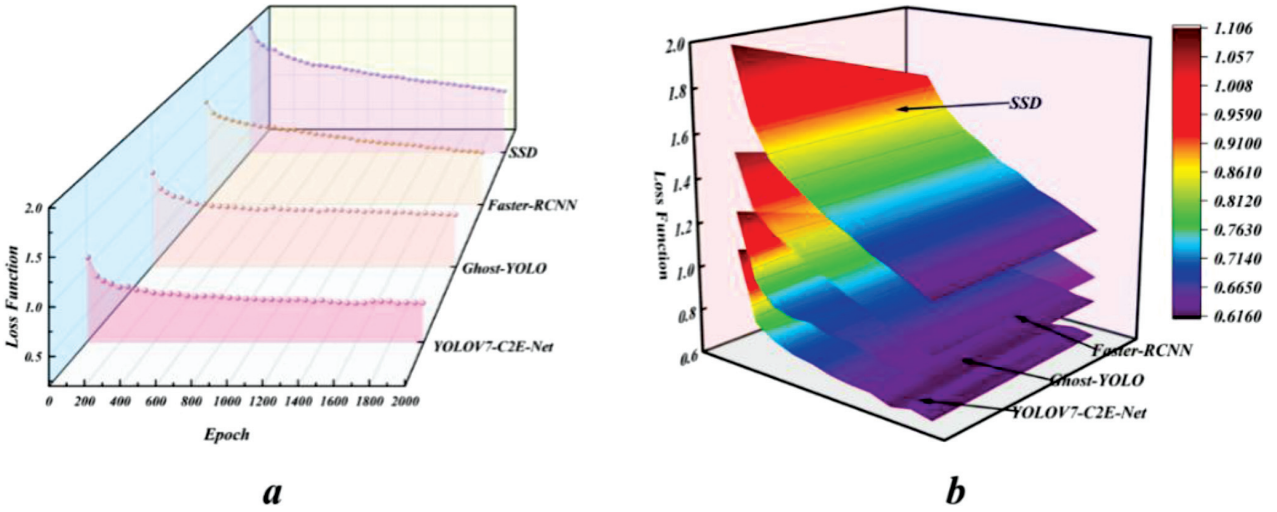

Figure 10: Loss Functions of Different Models

shown in the next figure, the area under the P-R curve of the YOLOv7 – C2E – Net model is significantly larger than that of the other three comparison models. This indicates that the YOLOv7 – C2E – Net model demonstrates superior feature extraction capability, offering distinct advantages in both detection accuracy and precision. In the context of concrete-crack detection against complex backgrounds, this model reveals remarkable applicability.

As shown in **Figure 10**, the loss function values of different models indicate that after 2000 training iterations, all four models converge. Notably, YOLOv7 – C2E – Net converges faster and has a lower loss value. This confirms that the improved algorithm significantly boosts the model's regression efficiency and accelerates its convergence speed. These three comparative experiments fully validate the superior performance of our proposed model, which is characterized by robust feature extraction capability, excellent generalization, rapid convergence, and high detection accuracy. In summary, our model is highly applicable to concrete crack detection in actual construction scenarios and provides a reliable, effective technical solution for related engineering tasks.

4 CONCLUSIONS

This paper presents an innovative YOLOv7-C2E-Net intelligent identification model for concrete building cracks. Based on YOLOv7, it integrates the C2E-Net module and the Inner-CIOU loss function optimization algorithm, bringing a new breakthrough to the field of concrete building crack detection. The proposed architecture integrates the CA, ELG, and CCF modules from C2E-Net to robustly capture crack characteristics with high precision. Furthermore, the implementation of the Inner-CIOU loss function enables accurate quantification of bounding box disparities, substantially enhancing the model's proficiency in detecting small-scale targets.

(1) To verify the optimized algorithms' performance improvements over the baseline, four ablation experiments were conducted with YOLOv7 as the baseline. Results showed that using only C2E-Net increased the Dice coefficient by approximately 9.23 % and IoU by about 15.09 %; using only the Inner-CIOU loss function increased the Dice coefficient by approximately 10.77 % and IoU by about 18.87 %; and using both together resulted in a Dice coefficient of 0.83 and an IoU of 0.81, representing improvements of about 27.69 % and 52.83 % over the original model, respectively. This demonstrates that C2E-Net effectively captures image features precisely, while the Inner-CIOU loss function optimizes small-target detection; their combined integration substantially enhances the model's ability to identify concrete cracks efficiently and accurately, even in challenging environments.

(2) To assess the efficacy and computational efficiency of the proposed YOLOv7-C2E-Net model, benchmarking experiments were conducted against alternative object detection frameworks. The evaluations revealed consistent superiority across critical performance metrics, with a Dice coefficient as high as 0.83, IoU reaching 0.81, detection accuracy of 91 %, recall rate of 89 %, and an F1 score of 0.91. Compared to the traditional SSD model, the Dice coefficient increased by 50.91 %, IoU increased by 88.37 %, and detection accuracy improved by 59.65 %. Compared to the Ghost-YOLO model, the Dice coefficient increased by 9.21 %, IoU increased by 22.73 %, and detection accuracy improved by 5.81 %.

(3) In practical applications, the YOLOv7-C2E-Net model demonstrated strong feature extraction and generalization capabilities. It showed good performance on both the training and test sets. On the test set, the Dice coefficient, IoU, and F1 score only decreased by 2.5 %, 6.58 %, and 4.59 %, respectively. In contrast, the Ghost-YOLO model exhibited a significant decline in

performance on the test set, indicating overfitting and insufficient generalization ability. Compared to other models, the YOLOv7-C2E-Net model can accurately identify concrete cracks of different scales and shapes, providing an efficient and reliable intelligent identification tool for the safety monitoring and maintenance of concrete buildings, and showing great potential for widespread application.

In the future the model's application scenarios can be further expanded: combining it with UAV aerial photography for comprehensive crack inspection of large bridges and tunnels, integrating real-time environmental data (e.g., temperature, humidity) to build a multi-dimensional monitoring model and improve crack risk early warning accuracy, or optimizing it via lightweight design to adapt to edge computing devices—meeting on-site rapid detection needs and further advancing the intelligent upgrading of concrete structure safety monitoring.

5 REFERENCES

[1] BA RAGAA A, AL-NESHAWY F, NOURELDIN M. AI-based framework for concrete durability assessment using generative adversarial networks and bayesian neural networks[J/OL]. Construction and Building Materials, 2025, 471: 140722. DOI:10.1016/j.conbuildmat.2025.140722

[2] ALKANNAD A A, AL SMADI A, YANG S, *et al*. Crack Vision: Effective Concrete Crack Detection With Deep Learning and Transfer Learning[J/OL]. IEEE Access, 2025, 13: 29554-29576. DOI:10.1109/ACCESS.2025.3540841

[3] WU Y, PIERALISI R, B. SANDOVAL F G, *et al*. Optimizing pervious concrete with machine learning: Predicting permeability and compressive strength using artificial neural networks[J/OL]. Construction and Building Materials, 2024, 443: 137619. DOI:10.1016/j.conbuildmat.2024.137619

[4] Park E, Eem S H, Jeon H. Concrete crack detection and quantification using deep learning and structured light [J]. Construction and Building Materials, 2020, 252(5): 119096

[5] Kim B, Yuvaraj N, Preethaa K R S, *et al*. Surface crack detection using deep learning with shallow CNN architecture for enhanced computation[J]. Neural Computing and Applications, 2021, 33(15): 9289–9305

[6] Lei Z, Yang F, Zhang D, *et al*. Road crack detection using deep convolutional neural network [C]// IEEE International Conference on Image Processing. IEEE, 2016

[7] Bae H, Jang K Y, An Y K. Deep super resolution crack network (SrcNet) for improving computer vision?based auto? mated crack detectability in in situ bridges [J]. Structural Health Monitoring, 2021, 20(4): 1428-1442

[8] Sarhadi A, Ravanshadnia M, Monirabbasi A, *et al*. Using an improved U-Net++ with a T-Max-Avg-Pooling layer as a rapid approach for concrete crack detection[J]. Frontiers in Built Environment, 2024, 101485774-1485774

[9] RYUZONO K, YASHIRO S, ONODERA S, *et al*. Performance evaluation of crack identification using density-based topology optimization for experimentally visualized ultrasonic wave propagation[J/OL]. Mechanics of Materials, 2022, 172: 104406. DOI:10.1016/j.mechmat.2022.104406

[10] ZHANG Y X, HUANG J, CAI F H. On Bridge Surface Crack Detection Based on an Improved YOLOv3 Algorithm [J]. IFAC-Papers OnLine, 2020, 53(2): 8205-8210

[11] M Brigante, M A Sumbatyan. On Multiple Crack Identification by Ultrasonic Scanning [J].Journal of Physics: Conference Series, 2018, 991(1)

[12] DENG Li, CHU Han-han, SHI Peng, *et al*. Region-based CNN method with deformable modules for visually classifying concrete cracks[J]. Applied Sciences, 2020, 10(7): 2528

[13] JI X, CHEN S, HAO L Y, *et al*. FBDPN: CNN-Transformer hybrid feature boosting and differential pyramid network for underwater object detection[J/OL]. Expert Systems with Applications, 2024, 256: 124978. DOI:10.1016/j.eswa.2024.124978

[14] LI G, YANG Y, WEN Y, *et al*. CGSW-YOLO Enhanced YOLO Architecture for Automated Crack Detection in Concrete Structures[J/OL]. Symmetry, 2025, 17(6): 890. DOI:10.3390/sym17060890

[15] HACIEFENDIOĞLU K, BAŞAĞA H B. Concrete Road Crack Detection Using Deep Learning-Based Faster R-CNN Method[J/OL]. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 2022, 46(2): 1621–1633. DOI:10.1007/s40996-021-00671-2

[16] PAL M, PALEVIČIUS P, LANDAUSKAS M, *et al*. An Overview of Challenges Associated with Automatic Detection of Concrete Cracks in the Presence of Shadows[J/OL]. Applied Sciences, 2021, 11(23): 11396. DOI:10.3390/app112311396

[17] LIN Y, AHMADI M, ALNOWIBET K A, *et al*. Concrete crack detection using ridgelet neural network optimized by advanced human evolutionary optimization[J/OL]. Scientific Reports, 2025, 15(1): 4858. DOI:10.1038/s41598-025-89250-3

[18] HUSSAIN M. YOLOGv1 to YOLOGv8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection[J]. Machines, 2023, 11(7): 677

[19] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C]// Proceedings of the IEEE / CVF conference on computer vision and pattern recognition. Vancouver, Canada: IEEE, 2023: 7464-7475

[20] MAO X, LI H, LI X, *et al*. C2E-Net: Cascade attention and context-aware cross-level fusion network via edge learning guidance for polyp segmentation[J/OL]. Computers in Biology and Medicine, 2025, 185: 108770. DOI:10.1016/j.