

ARTICLE

DSC-RTDETR: An Improved RTDETR Based Crack Detection on Concrete Surface

Yan Zhou and Hengyang Wu*

School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, 201209, China

*Corresponding Author: Hengyang Wu. Email: wuhy@sspu.edu.cn

Received: 09 August 2025; Accepted: 12 September 2025; Published: 20 October 2025

ABSTRACT: Crack Detection is crucial for ensuring the safety and durability of buildings. With the advancement of deep learning, crack detection has increasingly adopted convolutional neural network (CNN)-based approaches, achieving remarkable progress. However, current deep learning methods frequently encounter issues such as high computational complexity, inadequate real-time performance, and low accuracy. This paper proposes a novel model to improve the performance of concrete crack detection. Firstly, the You Only Look Once (YOLOv11) backbone replaces the original Real-Time Detection Transformer (RTDETR) backbone, reducing computational complexity and model size. Additionally, the Dynamic Snake Convolution (DSConv) has been introduced to strengthen the model's ability to extract crack features. Furthermore, integrating Cross-Stage Partial + Parallel Spatial Attention (C2PSA) and BiFormer attention mechanism to enrich the extracted feature information and detection accuracy. Experimental results demonstrate that compared with the RTDETR baseline model, the proposed method achieves a 1.8% improvement in detection accuracy and a 2.1% increase in recall, while reducing computational complexity by 55.3 GFLOPs and parameter count by 14 M. These results highlight the superior performance and efficiency of the proposed model.

KEYWORDS: Crack detection; YOLOv11; RTDETR; DSConv; BiFormer

1 Introduction

Concrete is widely used in engineering and construction due to its high strength, durability, and excellent workability. It plays a critical role in civil infrastructure, including roads, bridges, and tunnels. However, factors such as sustained thermal variations, inferior material quality, improper mixing can contribute to surface cracking. These cracks not only impair the aesthetic appearance of structures but also weaken their structural integrity. In addition, they can serve as initiation points for more severe failures, posing significant safety risks. Timely detection and repair of concrete cracks can help lower maintenance costs and prevent catastrophic structural damage.

In recent years, deep learning methods—particularly CNNs have achieved excellent results in crack detection. Advances in artificial intelligence have further driven progress in this field [1,2]. Laxman et al. [3] proposed a deep learning-based method for detecting cracks in reinforced concrete structures. Furthermore, object detection and instance segmentation techniques from computer vision have increasingly been applied to crack detection [4,5]. Zhang et al. [6] proposed an improved U-net network pavement crack segmentation method, Superior to advanced models such as U-net.



In 2016, Redmon and Farhadi [7] proposed an end-to-end object detection algorithm, YOLOv1, which redefined the object detection task as a regression problem. This approach marked the emergence of YOLO-based object detection methods. Nie and Wang [8] developed a YOLOv3-based method for pavement crack detection, significantly improving real-time detection and accuracy. Yu [9] employed YOLOv5s for concrete crack detection, but the model showed limitations when detecting fine or low-contrast cracks. Ni et al. [10] improved YOLOv7 by using K-means++ clustering and the F-ReLU activation function, achieving improved precision in road crack detection. Li et al. [11] used UAV imagery with YOLOv8 to improve dam crack inspection efficiency.

With the rise of Transformers [12], significant progress has been made in crack detection. Yu and Zhang [13] compared the performance of CNN-based models and Transformer-based models in pavement crack segmentation. Transformer models tend to converge more easily and achieve higher accuracy. Among them, SwinUNet demonstrates the best overall performance, achieving a favorable balance between accuracy and efficiency. Xiao et al. [14] proposed a novel hybrid-window attentive vision transformer framework, named CrackFormer, for pavement crack detection, aiming to provide an effective and automated solution for pavement distress inspection and repair. Sun et al. [15] designed a lightweight detection framework, Pavement Crack DETection Transformer (PCDETR), which fuses convolutional features with sequential features to enable efficient pavement crack detection. However, both U-Net and Transformer-based models generally have large parameter sizes and high computational complexity, resulting in slow inference and limiting their practical deployment in engineering applications.

Facebook AI introduced the DETR model [16] in 2020, eliminating the need for traditional techniques such as anchor boxes and Non-Maximum Suppression [17] (NMS). Liu et al. [18] propose Crack-DETR, an efficient pavement crack detection deep learning network, that can detect complex pavement cracks from complex background noise. Based on DETR, Zhao et al. [19] proposed RTDETR that efficiently processes multi-scale features, achieving superior performance in detecting small and low-contrast objects, particularly in terms of accuracy.

Despite significant progress in crack detection using deep learning methods, several challenges remain. Both CNN- and Transformer-based models require substantial memory and processing resources. The YOLO series provides excellent detection speed but achieves slightly lower accuracy compared to RT-DETR. To overcome these limitations, we integrate the YOLOv11 architecture into RT-DETR by replacing the backbone and the encoder. The proposed model combines YOLOv11's multi-scale feature extraction with the RTDETR decoder as the detection head, unifying the strengths of both frameworks. Based on this design, we further design a C3k2-DSCConv [20] module in the neck to replace standard convolutions, which enables cross-scale feature fusion. This design particularly addresses the elongated and complex crack shapes effectively. Finally, we integrate the BiFormer [21] attention mechanism to enhance the C2PSA module, which further reduces computational cost and improves feature extraction efficiency.

Overall, the proposed DSC-RTDETR model provides a refined architecture that overcomes the limitations of existing methods in terms of accuracy and computational complexity, and provides a more efficient and precise solution for crack detection. By combining the lightweight structure of YOLOv11 with the high-precision detection capability of RTDETR, DSC-RTDETR significantly reduces computational overhead and maintains accuracy. Compared with the YOLO series, DSC-RTDETR performs better in detecting complex cracks, such as fine or low-contrast cracks. Compared with RTDETR, the proposed model reduces computational burden and enhances feature extraction efficiency through architectural optimization and the introduction of the BiFormer attention mechanism, which achieves a better balance between accuracy and speed.

2 Overview of YOLOv11 and RTDETR Algorithms

Transformer-based end-to-end object detection models, such as DETRs, have achieved significant progress, but still face limitations in practical applications due to their high computational cost. To address the drawbacks of NMS in balancing accuracy and speed in traditional object detection, RTDETR was proposed as the first real-time end-to-end object detector.

In Fig. 1, the framework comprises four main components: a Backbone for feature extraction, an Efficient Hybrid Encoder that combines CNN and Transformer strengths, an IoU-aware Query Selection module for initializing object queries with high-quality proposals, and a Decoder & Head that refines features for final prediction.

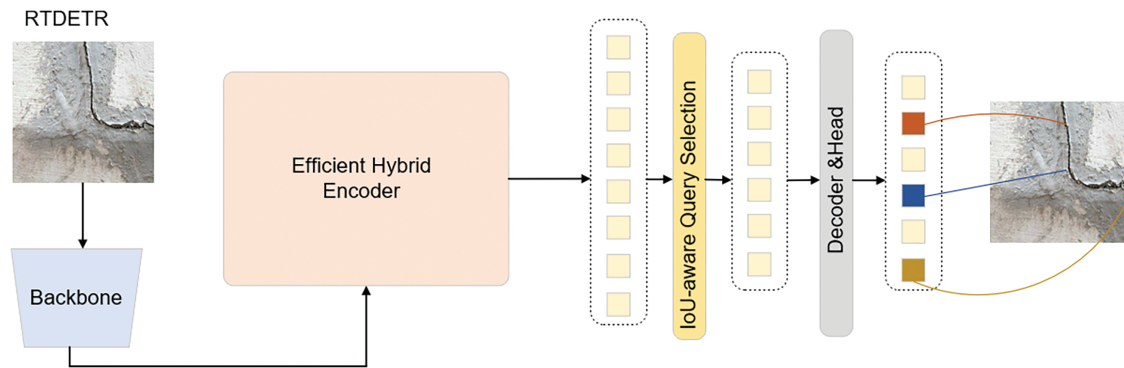


Figure 1: Overall architecture of the RTDETR model

The decoder in RTDETR plays a crucial role in enhancing detection accuracy. It receives object queries and multi-scale features extracted by the encoder. The decoder applies decoupled intra-scale interaction to capture relationships within each feature scale, which ensures that fine-grained and contextual information is preserved. Cross-scale feature fusion allows the decoder to combine information from different feature resolutions, improving the representation of small and low-contrast targets. RTDETR also employs an IOU-aware [22] query selection mechanism in the decoder, which provides high-quality initial object queries. These mechanisms enable the decoder to generate more precise bounding boxes and class predictions, resulting in a significant improvement in overall detection performance.

In Fig. 2, the detailed architecture of the proposed YOLOv11 model. It consists of a Backbone built with novel C2PSA and C3k2 modules for powerful feature extraction, a Neck employing feature pyramid networks with Upsample and Concatenation for multi-scale feature fusion, and a Head for making final predictions at three different scales.

YOLOv11 is a new generation within the YOLO family and demonstrates remarkable performance in real-time object detection. The overall architecture follows the classical design principles of the YOLO series, with its backbone composed of a feature extraction network that combines multiple convolutional blocks and residual Bottleneck structures. Notably, it incorporates the C3k2 module, which merges a Cross Stage Partial (CSP) architecture with dual Bottleneck layers and 3×3 convolutional kernels. This design improves feature reuse and gradient propagation through cross-stage connections. The Neck uses an improved Spatial Pyramid Pooling-Fast (SPPF) module, which allows multi-scale contextual information aggregation. The integration of the C2PSA attention mechanism strengthens the model's perception of small and occluded objects, which improves detection robustness. The design principles of the YOLOv11 architecture are highly adaptable and can inspire improvements in other object detection frameworks to further enhance performance.

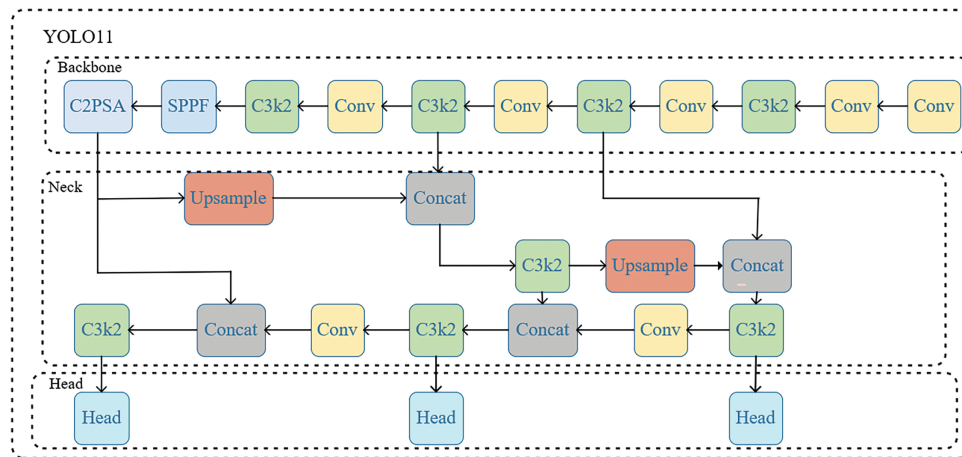


Figure 2: The detailed architecture of the YOLOv11 model

3 DSC-RTDETR

This chapter introduces the implementation details of the enhancements applied to the DSC-RTDETR algorithm.

3.1 RTDETR with YOLO Network Integration

In Fig. 3, the architecture of our proposed DSC-RTDETR model. It enhances the original RTDETR by integrating a deformable convolution-based backbone (featuring C2PSA and Bitformer modules) and a redesigned neck with C3k2_DSConv modules for more efficient and powerful multi-scale feature processing before the hybrid encoder.

Although RTDETR reduces computational demand, it still requires substantially more resources than the YOLO series. To address this issue, this study integrates the lightweight features of the YOLO architecture into RTDETR to create a new network structure. It establishes a more efficient and lightweight detection framework.

In DSC-RTDETR, the YOLOv11 backbone integrated with C3k2_DSConv and the BiFormer attention mechanism is employed as the feature extractor to generate multi-scale feature representations. These features are directly fed into the Transformer decoder of RTDETR, thereby combining the efficient local feature extraction capability of YOLOv11 with the global modeling capacity of the RTDETR decoder. This design unifies lightweight architecture with high-precision detection.

In this study, DSC-RTDETR retains the IOU-aware query selection strategy proposed in RTDETR while adapting it to the multi-scale feature maps output by the YOLOv11 backbone. During training, candidate queries are ranked not only by their classification confidence but also dynamically filtered using the IOU between predicted and ground-truth boxes. This enables the decoder to focus on more representative and informative queries. Such an adaptation ensures that the query selection mechanism of RTDETR remains effective within the lightweight DSC-RTDETR framework, thereby improving model convergence stability and detection accuracy. The enhancement is particularly pronounced in detecting small-scale defect regions.

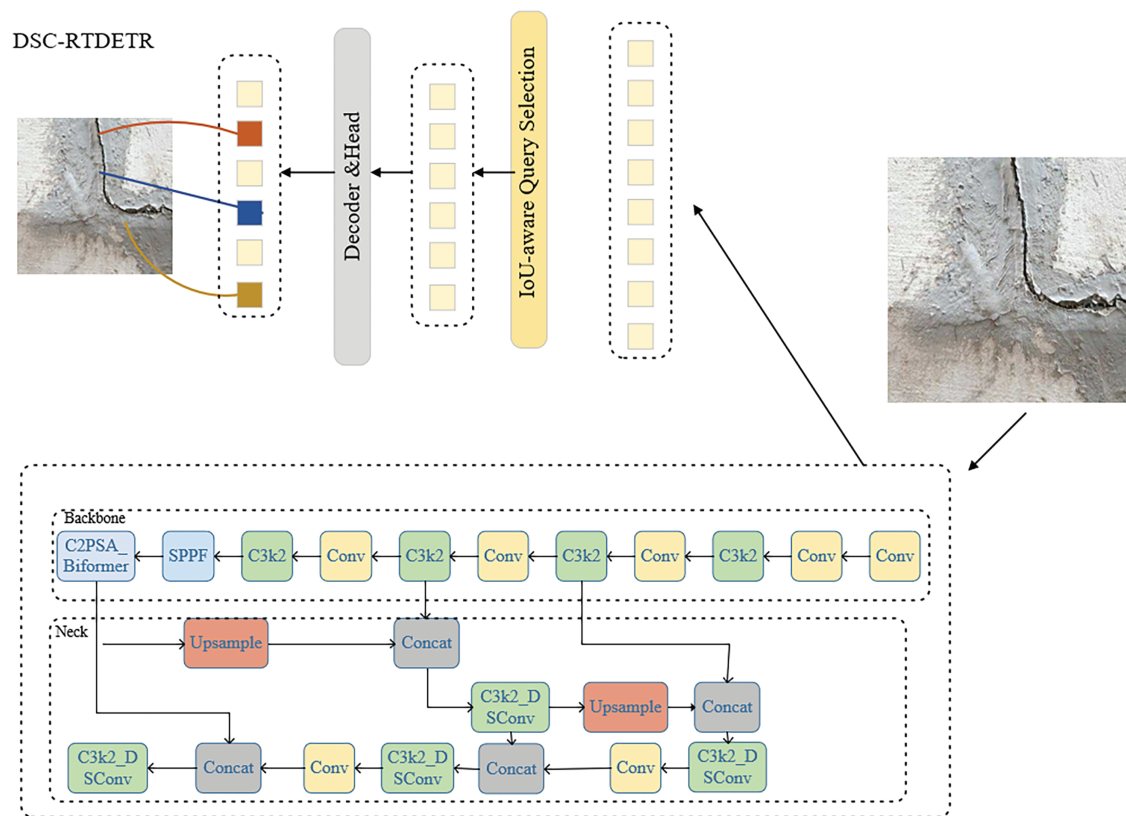


Figure 3: The architecture of DSC-RTDETR model

3.2 C2PSA Optimization Based on the BiFormer Attention Mechanism

BiFormer is a vision transformer model that incorporates Bi-level Routing Attention. It employs a two-level routing mechanism, where each image patch is assigned to a position-based routing unit. The routing units classify patches into upper-level and lower-level routes based on similarity, with upper-level routes capturing global contextual information and lower-level routes focusing on local structural details. This hierarchical design enhances the representation of fine-grained and low-contrast features, which is particularly beneficial for detecting elongated and irregular crack patterns.

Specifically, the routing module is controlled by several key hyperparameters. The top-k parameter determines how many of the most relevant windows are selected for each query window, typically set to 4. The query/key feature dimension defines the size of the projected queries and keys, commonly matching the model hidden dimension. Routing can optionally include learnable parameters (param_routing) to improve adaptability, and soft routing can be enabled to perform weighted aggregation of selected window features. In addition, the differentiable routing flag (diff_routing) controls whether gradients can propagate through the routing operation. These hyperparameters allow the routing module to efficiently capture long-range dependencies while maintaining the computational efficiency of local attention.

In Fig. 4, it replaces the standard full self-attention with a more efficient routing mechanism. The process involves Depthwise Convolution (DWConv), Layer Normalization (LN), the core Bi-level Routing Attention operation that selectively routes tokens, and an MLP for feature transformation.

Compared with conventional convolutional neural networks, Transformers offer superior ability to capture long-range dependencies. This is highly beneficial for concrete crack detection, since cracks are often

elongated and irregular. The two-level routing attention mechanism in BiFormer enables adaptive allocation of attention across different spatial scales and regions, which improves detection accuracy and enhances model robustness.

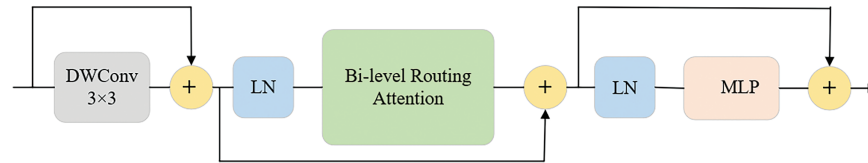


Figure 4: The structure of the Bi-level routing attention block

Although C2PSA has shown strong performance in general feature extraction tasks, it presents limitations when applied to concrete crack detection. Crack morphology varies greatly, from fine hairline fractures to wide and irregular patterns. However, C2PSA relies on convolutional kernels with fixed receptive fields, which restricts its adaptability to the diverse forms of cracks.

In Fig. 5, the detailed design of the C2PSA_Biformer block. It combines PSA blocks with multiple parallel branches of Convolution and BiLevelRoutingAttention, followed by a split and transformation operation (Conv, $s = 1$, $k = 1$) to aggregate information from different representation subspaces.

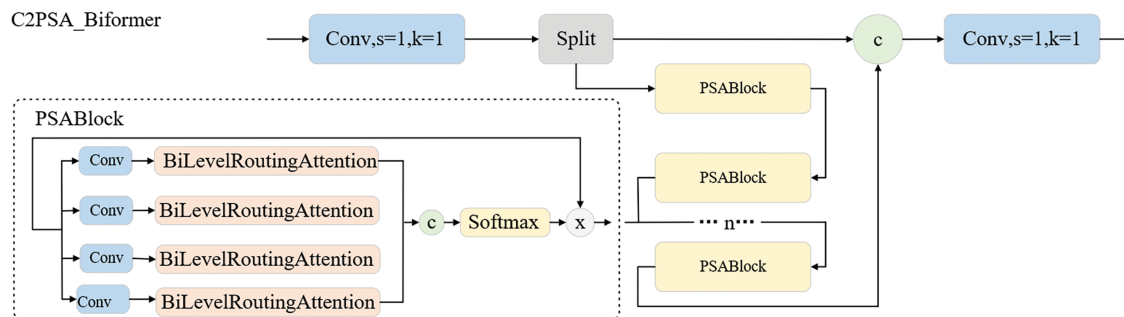


Figure 5: The detailed design of the C2PSA_Biformer block

We propose C2PSA_BiFormer, an enhanced module that integrates C2PSA with BiFormer to address its limitations in concrete crack detection. By balancing local and global feature extraction, it improves detection accuracy, while the dual-path routing attention mechanism enables adaptive selection of key features across scales, enhancing the capture of fine-grained crack details.

In concrete crack detection, the C2PSA_BiFormer module adds important improvements to the DSC-RTDETR framework. The module uses BiFormer's Bi-level Routing Attention to extract better features while keeping inference fast. It works well for complex textures, small cracks, and low-contrast areas. The C2PSA_BiFormer combines global and local attention: upper-level routing collects overall context, and lower-level routing focuses on crack details. This layered attention allows multi-scale feature interaction and strong feature representation. The position-sensitive PSA Block adds query-based sparsity, which lets the module focus on important regions and reduce unnecessary calculations. Overall, the C2PSA_BiFormer improves the detection accuracy and speed of DSC-RTDETR, providing a practical solution for light and accurate crack detection.

3.3 C3k2 Optimization with Dynamic Snake Convolution (DSConv)

DSConv is a new type of convolution inspired by tubular structures. It adapts well to elongated and irregular shapes, making it effective for detecting cracks in concrete. Standard 3×3 convolutions have limited ability to capture overall structural features. To address this issue, Yu [23] introduced dilated convolutions, which expand the receptive field and allow deeper layers to retain higher-resolution feature maps. Deformable convolutions increase flexibility and enable models to handle complex structures. However, these methods may miss small details in thin and irregular cracks. DSConv overcomes this limitation by adding learnable offsets and rules that adjust convolution operations according to the shape and position of the target. It learns convolutional kernel offsets via offset_conv, limits them to $[-1, 1]$ with tanh, and uses the DSConv module to adapt the kernel dynamically to complex shapes such as elongated and curved structures. DSConv provides a more precise and flexible solution for crack detection than conventional convolution methods.

In Fig. 6, it illustrates the process of learning input-dependent offsets (Δ) from the input feature map, which are then used to guide the convolutional sampling locations in the subsequent deformable convolution, allowing it to adaptively focus on non-rigid and irregular regions.

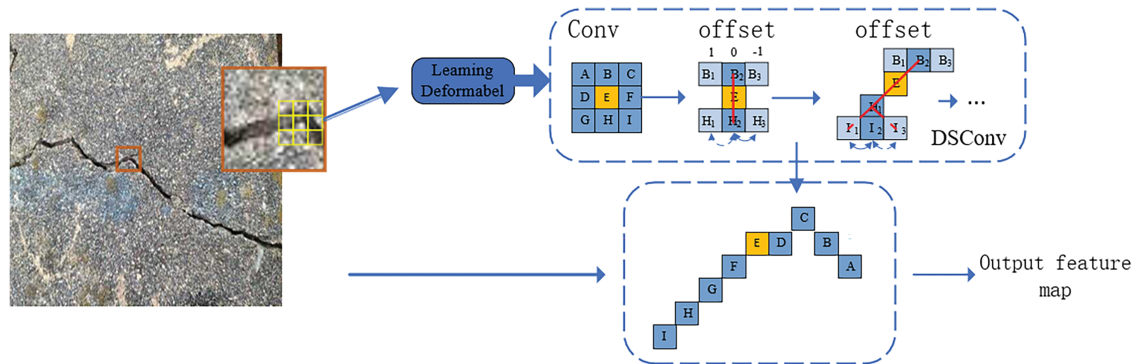


Figure 6: Schematic of the DSConv operation

DSConv is an advanced dynamic convolution method that uses variable kernel weights and a dynamic path selection mechanism. This design lets the model capture crack features of different shapes and greatly increases its flexibility in feature extraction. The dynamic path selection also strengthens DSConv's ability to combine information across a wider spatial context, ensuring high effectiveness in tasks that require both fine and large-scale feature analysis.

In Fig. 7, this enhanced cross-stage partial module integrates DSConv and DySnakeConv operations within its branches, alongside standard convolutions (CBS) and bottleneck blocks, for powerful and adaptive feature learning.

In the DSC-RTDETR architecture, the C3k2_DSConv module improves feature extraction and increases model flexibility. By replacing standard Bottleneck blocks with DSConv-enhanced modules, it introduces dynamically adjustable convolution kernels that trace the snake-like shapes of cracks, enabling the model to focus on fine-grained and spatially varying features. This approach is especially effective in scenarios with high background noise, thin cracks, or varying crack widths. The C3k2_DSConv module maintains the computational efficiency of the original backbone while significantly improving the model's ability to capture complex crack shapes. Compared with traditional parallel feature extraction methods, the combination of C3k2 and DSConv strengthens the model's capacity to capture both local and global crack features, thereby enhancing detection accuracy.

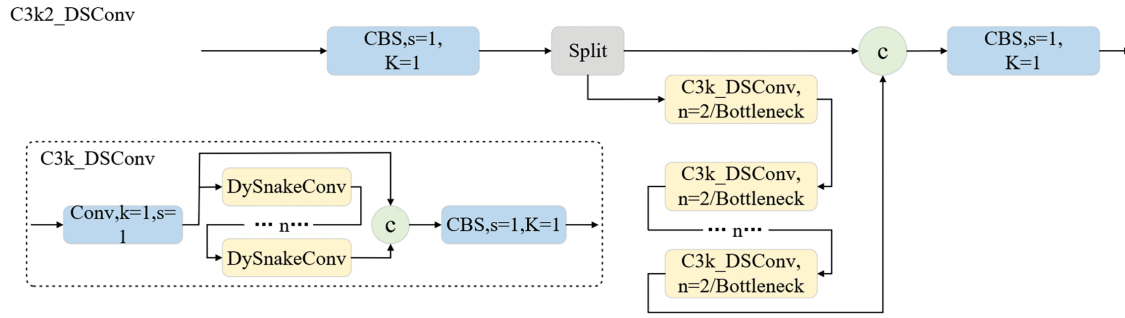


Figure 7: The internal structure of the C3k2_DSConv module

4 Experimental Environment and Results Analysis

4.1 Experimental Environment and Evaluation Metrics

All experiments were performed on an AutoDL virtual server equipped with a Linux-based system, PyTorch 2.0.0 + cull8, and Python 3.8.10. The hardware setup included an RTX 4090D GPU with 24 GB of VRAM. The training hyperparameters are summarized in [Table 1](#).

Table 1: Training parameters

Batch	Epochs	Imgsz	Weight_decay
16	100	416 × 416	0.0001

For the crack-seg dataset, data augmentation was used to increase diversity. The strategies included average blurring (kernel sizes 5×5 to 15×15), random cropping (retaining 40%–80% of the original size), and random rotation (-60° to 60°). All augmented images were resized to 416×416 pixels. After augmentation, 4029 images were generated and split into training, validation, and testing sets with a 7:2:1 ratio.

The AdamW optimizer was chosen for training DSC-RTDETR due to its stability and convergence. By combining adaptive learning rates with decoupled weight decay, AdamW reduces overfitting and improves convergence stability during training. For DSC-RTDETR, which integrates DSConv and a lightweight design, AdamW enables faster convergence and achieves balanced optimization across multiple loss functions, ensuring both accuracy and efficiency.

This model adopts precision, recall, F1 value, parameter number and GFLOPs as evaluation indicators to measure each performance of the model.

Calculation formula of precision P and recall R and mAP :

$$P = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$mAP = \int_0^1 PRdr \quad (3)$$

F_1 values can be combined with precision and recall to reflect overall performance:

$$F_1 = 2 \frac{P \times R}{P + R} \quad (4)$$

The above metrics are essential for evaluating algorithm performance, with higher values reflecting stronger detection capability. In addition, the number of model parameters must be considered, as it directly affects model complexity and requires greater computational power and memory during training.

4.2 Ablation Studies

To evaluate the effectiveness of the proposed DSC-RTDETR algorithm, ablation experiments were conducted to assess the contributions of the BiFormer attention mechanism and DSConv. Table 2 presents the results of experiments where the proposed modules were added or removed. These comparisons highlight the impact of different module configurations on overall model performance.

Table 2: Ablation experiments

Experimental	RTDETR	YOLOV11	DSConv	BiFormer	P (%)	R (%)	FLOPs (G)	Parament (M)
1	✓				90.3	72.7	58.6	20.1
2	✓	✓			91.5	73.5	10.7	5.6
3	✓	✓	✓		92.3	72.7	11.7	6.2
DSC-RTDETR	✓	✓	✓	✓	92.1	74.8	11.3	6.1

Note: Bold entries in the first column represent the experiments or models proposed in this work, while bold values in other columns indicate the best performance.

The study involved four experiments, as summarized in Table 2, where the symbol “✓” denotes the inclusion of a specific module. Experiment 1 corresponds to the baseline RTDETR algorithm, while Experiment 4 represents the proposed DSC-RTDETR model. With the progressive integration of additional modules, model accuracy improves steadily. The DSConv module enhances crack feature extraction, leading to a notable increase in detection accuracy, whereas the BiFormer attention mechanism further optimizes feature representation by mitigating redundancy and strengthening feature fusion.

This combination also reduces model size and parameter count while improving recall. Overall, compared with the original RTDETR, the proposed model achieves a 1.8% gain in precision and a 2.1% gain in recall, while reducing computational cost by 53.3 GFLOPs and parameter count by 14 M. These enhancements yield a more lightweight model that maintains high accuracy and recall, thereby delivering superior overall performance.

4.3 Visualization Comparison of Model Training Process

To evaluate the effectiveness of the proposed DSC-RTDETR algorithm for concrete crack detection, a comparative study was conducted against YOLOv8, YOLOv11, and the original RTDETR over multiple training epochs. The results are presented in Fig. 8.

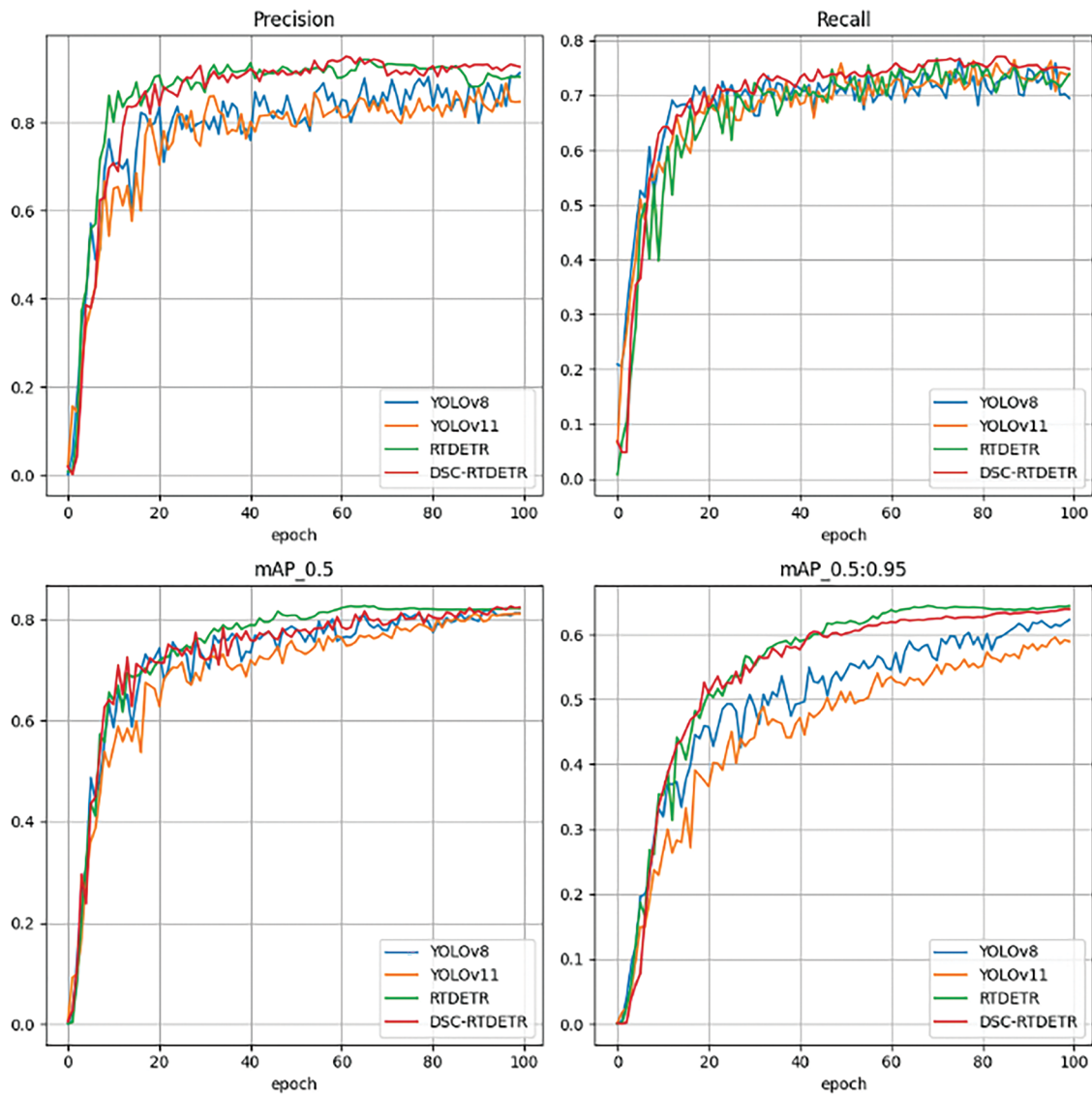


Figure 8: Precision vs. epoch curves comparing the proposed DSC-RTDETR with state-of-the-art detectors including YOLOv8, YOLOv11, and the original RTDETR

As shown in [Fig. 8](#), DSC-RTDETR achieves higher precision and recall than both YOLOv8 and YOLOv11. In the early stages of training, DSC-RTDETR shows slightly lower mAP50 and mAP50-95 than RTDETR due to its reduced number of parameters, which initially limits performance. However, these metrics converge in later epochs, exhibiting minimal differences. Overall, DSC-RTDETR demonstrates slightly superior performance in precision, recall, and F1-score relative to the other models.

4.4 Comparison with Mainstream Algorithms

To objectively assess the performance of the improved DSC-RTDETR model, we conducted training and evaluation experiments in comparison with the original RTDETR and several widely used network models. The performance metrics of these models are summarized in [Table 3](#).

Table 3: Comparative experiment results for different mainstream algorithms

Model	P (%)	R (%)	mAP (%)	F1 (%)
SSD [24]	87.3	75.1	80.6	80.74
Faster RCNN [25]	88.6	74.3	81.5	80.82
RTDETR	90.3	72.7	82.2	80.47
YOLOv8	90.3	69.5	81.3	78.54
YOLOv5	80.7	72.0	76.9	76.1
YOLOv9c	87.0	72.0	80.1	78.7
YOLOv11	88.9	70.8	81.0	77.84
YOLOv10n	70.4	70.3	72.5	69.3
YOLOv12	88.5	68.9	76.4	77.5
RTDETR-x	92.7	72.7	81.1	81.5
DSC-RTDETR	92.1	74.8	82.4	82.55

Note: Bold entries in the first column represent the experiments or models proposed in this work, while bold values in other columns indicate the best performance.

As shown in Table 3, the DSC-RTDETR model outperforms all other compared algorithms. It achieves the highest precision (92.1%) and mAP (82.4%), surpassing the YOLO series in both precision and recall. These results demonstrate the effectiveness and superiority of the proposed algorithm for concrete crack detection.

4.5 Evaluation of Different C3k2 Integration Strategies

To further improve DSC-RTDETR, the C3k2 module was enhanced by integrating DSConv-based dynamic snake convolution. To evaluate the effectiveness of this integration, several recently proposed methods were implemented at the same position in the model under identical experimental settings. The comparative results are shown in Table 4.

Table 4: Comparative experiment results for different strategy implementations

Experiment	P (%)	R (%)	mAP (%)	mAP50-95 (%)
C3k2 + Conv	90.6	73.9	80.6	61.7
C3k2 + DCNv2 [26]	91.8	72.0	79.4	62.2
C3k2 + DWRSeg [27]	88.2	71.9	80.6	62.1
C3k2 + RepVGG [28]	86.1	74.9	80.2	60.7
C3k2 + GhostModel [29]	92.7	71.9	79.7	62.5
C3k2 + FasterBlock [30]	85.2	76.6	82.3	62.2
C3k2 + DSConv	92.1	74.8	82.4	63.9

Note: Bold entries in the first column represent the experiments or models proposed in this work, while bold values in other columns indicate the best performance.

As shown in Table 4, integrating DSConv into the C3k2 module significantly enhances the model's performance in crack detection. The original C3k2 module often misses critical crack-related features due to the complex and elongated nature of cracks. DSConv, with its strong ability to capture tortuous and elongated patterns, effectively overcomes these limitations.

We integrated DSConv, Conv, and DCNv2 into the C3k2 module under identical experimental settings, and the comparative results are presented in Table 4. As shown, C3k2 + DSConv achieves superior performance with higher precision (92.1%) and mAP (82.4%) compared to both C3k2 + Conv (90.6%, 80.6%) and C3k2 + DCNv2 (91.8%, 79.4%). Notably, DSConv improves mAP by 3.0 points over DCNv2, highlighting its stronger ability to capture elongated and irregular crack structures. Overall, experiments demonstrate the domain-specific advantages of DSConv for crack detection.

4.6 Comparative Evaluation of Attention Mechanisms

The BiFormer attention mechanism enhances feature aggregation by dynamically selecting relevant information across multiple spatial scales, thereby effectively improving model performance. Table 5 presents a comparative study of different attention mechanisms, which demonstrates the effectiveness of integrating the BiFormer attention mechanism into the C2PSA module.

Table 5: Comparative experiment results for different attention mechanisms

Experiment	P (%)	R (%)	mAP (%)	mAP50-95 (%)
C2PSA + Acmix [31]	90.6	74.3	81.9	62.4
C2PSA + DLKA [32]	85.0	72.7	79.7	61.4
C2PSA + EMA [33]	83.6	73.9	80.4	60.2
C2PSA + LSKA [34]	85.2	71.7	79.3	59.7
C2PSA + SEAM [35]	86.2	75.1	79.9	59.5
C2PSA + BiFormer	92.1	74.8	82.4	63.9

Note: Bold entries in the first column represent the experiments or models proposed in this work, while bold values in other columns indicate the best performance.

As shown in Table 5, the integration of the BiFormer attention mechanism into the C2PSA module leads to significant improvements in both recall and mAP for crack detection. The BiFormer mechanism enhances the model's feature aggregation capability by dynamically selecting relevant information across multiple spatial scales, enabling the extraction of more comprehensive and accurate feature representations.

4.7 DSC-RTDETR on Crack500

To further validate the generalization ability and practicality of the proposed method, in addition to experiments on the crack-seg dataset, we also conducted training and testing on a road crack dataset derived from Crack500. The Crack500 dataset contains a wider variety of crack types and images from diverse scenes, which facilitates a more comprehensive evaluation of the model's performance under different data distributions. Experimental results demonstrate that the proposed method also achieves outstanding detection performance on the Crack500 dataset, thereby further confirming its applicability in real-world scenarios and the reliability of the conclusions.

In terms of performance comparison, as shown in Table 6, DSC-RTDETR demonstrates a well-balanced performance across multiple evaluation metrics, highlighting clear advantages. The model achieves an mAP of 79.7%, outperforming YOLOv8, RT-DETR, and YOLOv11, indicating superior overall detection accuracy for pavement crack detection tasks. Additionally, DSC-RTDETR attains an F1 score of 77.1%, which is comparable to RTDETR and higher than YOLOv11, further confirming a favorable balance between precision and recall. The results demonstrate that our model maintains competitive performance across different environments, confirming its robustness and adaptability.

Regarding inference speed, DSC-RTDETR reaches 56.09 FPS, representing a substantial improvement over RT-DETR. Although slightly slower than YOLOv8 and YOLOv11, the higher detection accuracy compensates for this, achieving an effective trade-off between lightweight design and real-time performance. From a computational perspective, DSC-RTDETR requires significantly fewer FLOPs than RT-DETR while maintaining higher mAP and faster inference, reflecting the benefits of the architectural optimizations. Compared with YOLOv8 and YOLOv11, the inference speed of DSC-RTDETR is slightly lower, mainly due to the integration of feature enhancement modules and an improved detection head. These additions moderately increase computational complexity but result in higher accuracy and reduced FLOPs, aligning well with practical requirements in industrial applications where both detection performance and computational efficiency are critical.

Table 6: Comparison of DSC-RTDETR and mainstream models

Model	P (%)	R (%)	mAP (%)	F1 (%)	FPS	FLOPs (G)
YOLOv8	86.1	69.0	79.1	76.6	204.08	8.1
RTDETR	89.2	68.2	78.6	77.2	46.56	56.9
YOLOv11	86.6	65.7	77.9	74.7	153.84	6.3
DSC-RTDETR	87.9	68.7	79.7	77.1	56.09	11.3

Note: Bold entries in the first column represent the experiments or models proposed in this work, while bold values in other columns indicate the best performance.

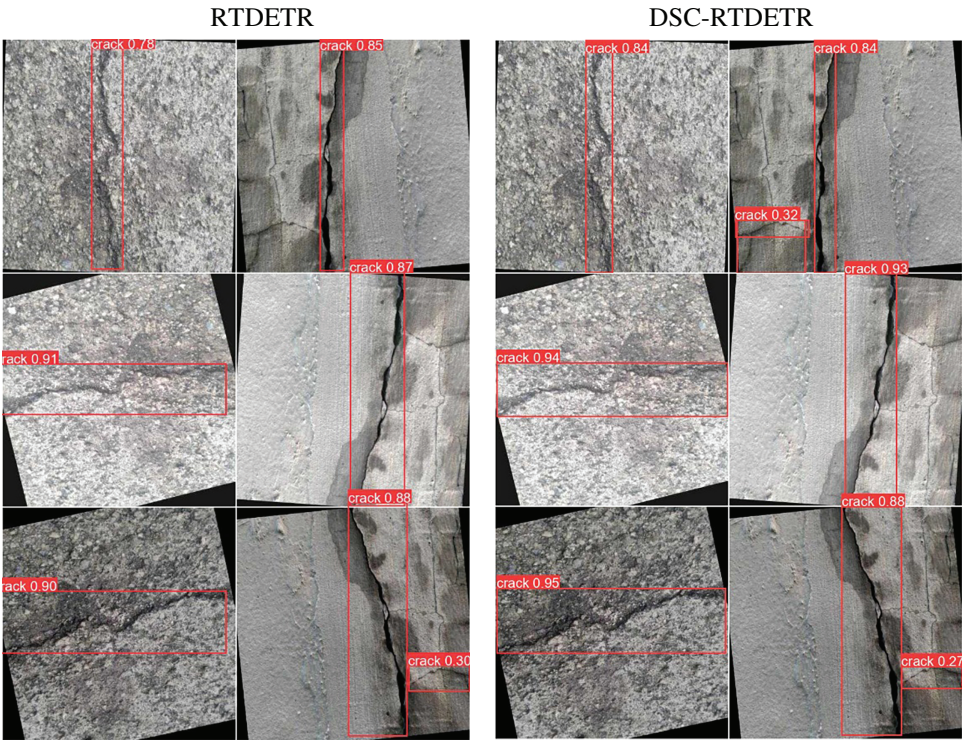


Figure 9: Variation curve of each index with iteration rounds

4.8 Detection Results

To verify the effectiveness of the proposed improvements, we compared the detection results of DSC-RTDETR with those of the original RT-DETR model, demonstrating the superiority of the enhanced approach. As shown in Fig. 9, DSC-RTDETR achieves noticeably higher accuracy in crack detection.

5 Conclusion

In summary, this study proposes DSC-RTDETR, a novel lightweight concrete crack detection model that integrates YOLOv11 with RT-DETR. By incorporating DSConv and BiFormer, the model achieves improved detection accuracy while preserving computational efficiency. The discussion on edge-device deployment is grounded in the reduced computational complexity and parameter count, which make the model suitable for resource-constrained mobile devices. In future work, we will focus on improving recall, validating performance on hardware-limited platforms, and further optimizing the model for practical deployment in diverse real-world environments.

Acknowledgement: The authors acknowledge the contribution and the support of the department of School of Computer and Information Engineering, Shanghai Polytechnic University.

Funding Statement: This work is supported by the National Natural Science Foundation of China (No. 62162014).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yan Zhou, Hengyang Wu; data collection: Yan Zhou; analysis and interpretation of results: Yan Zhou; draft manuscript preparation: Yan Zhou. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset modified and used in this study will be available upon request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Ali R, Chuah JH, Talip MSA, Tan J, Lee K, Lim WL, et al. Structural crack detection using deep convolutional neural networks. *Autom Constr.* 2022;133:103989. doi:10.1016/j.autcon.2021.103989.
2. Hamishebahar Y, Guan H, So S, Yu Y, Smith S, Zhang Y, et al. A comprehensive review of deep learning-based crack detection approaches. *Appl Sci.* 2022;12(3):1374. doi:10.3390/app12031374.
3. Laxman KC, Tabassum N, Ai L, Zhang Y, Wang X, Chen H, et al. Automated crack detection and crack depth prediction for reinforced concrete structures using deep learning. *Constr Build Mater.* 2023;370(2):130709. doi:10.1016/j.conbuildmat.2023.130709.
4. Hafiz AM, Bhat GM. A survey on instance segmentation: state of the art. *Int J Multimed Inf Retr.* 2020;9(3):171–89. doi:10.1007/s13735-020-00195-x.
5. Hacıfendioglu K, Başağa HB. Concrete road crack detection using deep learning-based faster R-CNN method. *Iran J Sci Technol Trans Civ Eng.* 2022;46(2):1621–33. doi:10.1007/s40996-021-00671-2.
6. Zhang Q, Chen S, Wu Y, Ji Z, Yan F, Huang S, et al. Improved U-Net network asphalt pavement crack detection method. *PLoS One.* 2024;19(5):e0300679. doi:10.1371/journal.pone.0300679.
7. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017 Jul 21–26; Honolulu, HI, USA. p. 7263–71. doi:10.1109/CVPR.2017.690.

8. Nie M, Wang C. Pavement crack detection based on YOLO v3. In: Proceedings of the 2019 2nd International Conference on Safety Produce Informatization (IICSPI); 2019 Dec 14–15; Chongqing, China. p. 327–30. doi:10.1109/IICSPI48186.2019.9095956.
9. Yu Z. YOLOv5s-based deep learning approach for concrete cracks detection. In: Proceedings of the SHS Web of Conferences. Les Ulis, France; 2022. (In Chinese). doi:10.1051/shsconf/202214403015.
10. Ni C, Li L, Luo WT, Zhang Y, Chen H, Wang K, et al. Disease detection of asphalt pavement based on improved YOLOv7. *Comput Eng Appl.* 2023;59(13):305–16.
11. Li X, Li L, Liu Z, Peng Z, Liu S, Zhou S, et al. Dam crack detection studies by UAV based on YOLO algorithm. In: Proceedings of the 2023 2nd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC); 2023 Aug 18–20; Qingdao, China. p. 104–8. doi:10.1109/RAIIC59453.2023.10281120.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017;30:5998–6008.
13. Zhang Y, Zhang L. Detection of pavement cracks by deep learning models of Transformer and UNet. *IEEE Trans Intell Transp Syst.* 2024;25(11):15791–808. doi:10.1109/TITS.2024.3420763.
14. Xiao S, Shang K, Lin K, Wu Q, Gu H, Zhang Z. Pavement crack detection with hybrid-window attentive vision transformers. *Int J Appl Earth Obs Geoinf.* 2023;116(18):103172. doi:10.1016/j.jag.2022.103172.
15. Sun Z, Zhai J, Pei L, Li W, Zhao K. Automatic pavement crack detection transformer based on convolutional and sequential feature fusion. *Sensors.* 2023;23(7):3772. doi:10.3390/s23073772.
16. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Proceedings of the 16th European Conference on Computer Vision (ECCV); 2020 Aug 23–28; Glasgow, UK. p. 213–29. doi:10.1007/978-3-030-58452-8_13.
17. Neubeck A, Van Gool L. Efficient non-maximum suppression. In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06); 2006 Aug 20–24; Hong Kong, China. p. 850–5. doi:10.1109/ICPR.2006.479.
18. Liu G, Liu G, Chai X, Li L, Dai F, Huang B. Crack-DETR: complex pavement crack detection by multifrequency feature extraction and fusion. *IEEE Sens J.* 2025;25(9):16349–60. doi:10.1109/JSEN.2025.3549121.
19. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. p. 16965–74.
20. Qi Y, He Y, Qi X, Zhang Y, Yang G. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 6070–9.
21. Zhu L, Wang X, Ke Z, Zhang W, Lau R. BiFormer: vision transformer with bi-level routing attention. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–22; Vancouver, BC, Canada. p. 10323–33. doi:10.1109/CVPR52729.2023.00995.
22. Yu J, Jiang Y, Wang Z, Cao Z, Huang T. Unitbox: an advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia; 2016 Oct 15–19; Amsterdam, The Netherlands. p. 516–20. doi:10.1145/2964284.2967274.
23. Yu F, Koltun V, Funkhouser T. Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 472–80.
24. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. European Conference on Computer Vision; 2016 Oct 8–16; Amsterdam, The Netherlands. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
25. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst.* 2015;28(6):91–9. doi:10.1109/tpami.2016.2577031.
26. Zhu X, Hu H, Lin S, Dai J. Deformable ConvNets V2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 9300–8. doi:10.1109/CVPR.2019.00953.
27. Wei H, Liu X, Xu S, Song G, Zhang L. DWRSeg: rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. *arXiv:2212.01173.* 2022. doi:10.48550/arxiv.2212.01173.

28. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. Repvgg: making VGG-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; Nashville, TN, USA. p. 13733–42.
29. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. Ghostnet: more features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 14–19; Seattle, WA, USA. p. 1580–9.
30. Chen J, Kao S, He H, Zhuo W, Wen S, Lee CH, et al. Run don't walk: chasing higher FLOPS for faster neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–22; Vancouver, BC, Canada. p. 12021–31.
31. Pan X, Ge C, Lu R, Song S, Chen G, Huang Z, et al. On the integration of self-attention and convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19–24; New Orleans, LA, USA. p. 815–25.
32. Azad R, Niggemeier L, Hüttemann M, Kazerouni A, Khodapanah Aghdam E, Velichko Y, et al. Beyond self-attention: deformable large kernel attention for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024 Jan 4–8; Waikoloa, HI, USA. p. 1287–97.
33. Ouyang D, He S, Zhang G, Li X, Shi X, Liu Z. Efficient multi-scale attention module with cross-spatial learning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5. doi:10.1109/ICASSP49357.2023.10096516.
34. Lau KW, Po LM, Rehman YAU. Large separable kernel attention: rethinking the large kernel attention design in CNN. *Expert Syst Appl.* 2024;236(9):121352. doi:10.1016/j.eswa.2023.121352.
35. Wang Y, Zhang J, Kan M, Shan S, Chen X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 14–19; Seattle, WA, USA. p. 12275–84.