RCYOLO: An Efficient Small Target Detector for Crack Detection in Tubular Topological Road Structures Based on Unmanned Aerial Vehicles

Chao Dang and Zai Xing Wang

***Abstract*—Unmanned Aerial Vehicles (UAVs) combined with target detection algorithms can enhance the detection of road cracks. In response to the challenges presented by complex crack shapes and textures, small sizes, and highly integrated backgrounds, this article developed a UAV-based road crack target detection algorithm using road crack you only look once (RCYOLO). RCYOLO was composed of a C2f_DySnakeConv (C2f_DSConv) module located in the ninth layer and a simulated attention mechanism (SimAM) module situated above the Spatial Pyramid Pooling–Fast (SPPF), along with a dyhead attention detection head that integrated three types of attention mechanisms. Initially, the C2f_DSConv was proposed to effectively extract tubular features of cracks. Subsequently, the SimAM addressed the issue of target-background fusion, enhancing feature recognition of the targets while suppressing background interference. Finally, the dyhead strategy incorporated three types of attention mechanisms, effectively resolving the issue of small target omissions. Our results showed that on the custom UAV dataset road crack image, which included close-range and long-range images, RCYOLO outperformed the baseline network YOLOv8 by 5.9% in mAP@0.5, 6.5% in recall, and 9.8% in precision. On the public dataset Detection of Objects in Aerial Images, mAP@0.5 also exceeded YOLOv8 by 5.8%, indicating that RCYOLO performed well in other remote sensing image tasks, making this target detection algorithm more suitable for high-altitude photography of crack targets than other mainstream algorithms.**

***Index Terms*—Dyhead, DySnakeConv, road crack object detection, simulated attention mechanism (SimAM), unmanned aerial vehicle (UAV), YOLOv8.**

I. Introduction

THE construction of highway infrastructure in the northwestern region of China is crucial for economic and regional development. The primary goal of highway construction in this area is to build a comprehensive transportation corridor, while also focusing on enhancing the quality and standardization of highway construction to promote regional economic growth and transportation convenience. Although there has been an increase in the total mileage of highways, the development of highway infrastructure in the western region still lags behind that of the more developed eastern regions. The durability of asphalt pavements is influenced by factors, such as materials, traffic volume, construction quality, and maintenance levels. If these factors are not properly addressed, they could lead to a reduction in the lifespan of highways and increased risks to traffic safety [1].

Due to the continuous rise in road maintenance costs and increasing demands for efficiency, taking measures to prevent severe traffic hazards has become particularly important. However, road cracks, as a common pavement problem, come in various types and are often influenced by multiple factors, such as shadows, climate changes, and data collection noise. They typically present as multiscale, elongated tubular structures, making them difficult to detect accurately using traditional methods. In this context, Li et al. [2] proposed an automatic road crack image (RCI) background processing detection method executed by an automatic road inspection vehicle, which can effectively meet the detection needs of open road areas. However, in crowded urban road environments, this method may lead to traffic congestion due to slow sampling speed and may collect a large amount of invalid data. In contrast, Chen et al. [3] proposed a road surface damage detection method based on the combination of unmanned aerial vehicle (UAV) and machine-learning algorithms. While this method effectively addresses the issue of multiscenario detection, it is only suitable for targets with obvious features and large effective areas, failing to capture small-scale cracks and struggling to achieve accurate detection in complex environmental conditions, lacking generalization capability. Therefore, in remote, hard-to-reach areas or crowded urban areas, there are still many gaps in research on efficient multiscale crack detection methods. Developing a method for timely detection of multiscenario, multiscale targets with high generalization capability are of great significance for road maintenance and preventing safety accidents.

Currently, convolutional neural networks (CNNs), as a representative method in deep learning, have shown immense potential in sectors, such as agriculture [4], industry [5], and military [6], due to their high generalization and feature representation capabilities. Compared to traditional digital image

Manuscript received 22 February 2024; revised 17 May 2024 and 12 June 2024; accepted 24 June 2024. Date of publication 27 June 2024; date of current version 24 July 2024. The work of Gan Caijiao was supported in part by the Integrated Circuit Industry Research Institute under Grant [2023]36, in part by the R&d and Industrialization of Metal Film pressure sensor under Grant 224033, and in part by the Artificial Intelligence chip Development and Application Research under Grant 224024 [2024] 001. *(Corresponding author: Zai Xing Wang.)*

Chao Dang is with the School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu Province 730000, China (e-mail: dc224017@gmail.com).

Zai Xing Wang is with the School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu Province 730000, China, also with the Gansu Integrated Circuit Industry Research Institute, Gansu Province 730000, China, and also with the Microelectronics Industry Research Institute Lanzhou, Gansu Province 730000, China (e-mail: zaixw@mail.lzjtu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3419903

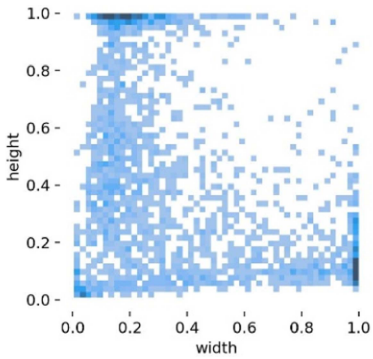

Fig. 1. Width–height distribution of targets in the dataset.

processing methods (like spectral imaging [7] and threshold segmentation [8], [9]), which mainly target individual cracks, such as longitudinal and transverse cracks, CNNs offer superior generalization and detection accuracy for large-scale and diverse tasks. Cracks are classified based on width and edge damage into minor (width <3 mm, fewer branches) and severe (width >3 mm, multiple branches) categories [10]. With the maturity and low cost of drone technology, UAVs are increasingly used in remote sensing and mapping. For road crack detection, high-resolution cameras on UAVs can inspect roads without affecting traffic, quickly acquiring high-definition data [11]. Deep-learning methods combined with UAVs are widely used in specialized target detection tasks. Zhang et al. [12] proposed a drone image detection network with self-attention guidance and multiscale feature fusion, effectively combining local and global information to focus on the target area and reduce background impact. Wang et al. [13] improved YOLOv8 by introducing the global attention mechanism (GAM) into its lower layers to enhance detection performance by retaining image feature information during sampling. Cui et al. [14] introduced SCYNet, a network combining UAVs and YOLO, using a Simi-BiFPN feature structure with skip connections to fully fuse features from various scales without significantly increasing computational overhead. To address crack diversity, Wang et al. [15] proposed a detector combining YOLOv5 with a Vision Transformer to calculate attention weights for longitudinal and transverse crack regions, achieving a mean Average Precision (mAP) of 0.872. He et al. [16] introduced MUENet, which accelerates network training by detecting and classifying crack morphology, color, and type. These studies demonstrate the effectiveness of combining CNNs and UAVs in detection tasks. However, challenges persist in road crack detection research.

1) Environmental Factors and Technical Challenges:

a) The tubular topology and small target characteristics of road cracks, as shown in the width–height distribution map of targets (Fig. 1), indicate that small targets dominate the image, and multiple points are closely aligned or near the axis, reflecting the common tubular structure of cracks. These cracks, extending along the road surface, pose increased detection difficulty due to their complex shapes, sizes, widths, depths, and diverse directions. b) One of the main challenges in road crack detection is the high similarity between cracks and the surrounding background. The color, texture, and lighting conditions of road materials often cause cracks to blend with the road surface, making them difficult to identify. This issue is particularly pronounced under the extreme climatic conditions in northwest China, especially in environments with frequent snowfall. Snow cover not only obscures the cracks on the road surface but also further complicates the detection task.

2) Model and Resource Requirements:

a) Jiang et al. [17] pointed out that while many existing studies use modules to enhance model performance in road surface crack detection, few studies consider how the "method" and "location" of adding modules affect the model's performance. This is an issue that must be addressed. b) Research on road crack detection often relies on ground-based photography samples, leading to poor performance and limited generalization ability of the models in high-altitude operations. Insufficient training samples make it difficult for models to learn complex patterns, especially in extreme environments. Additionally, the increased network size to improve detection accuracy significantly raises the demand for computational resources, posing challenges for real-time high-altitude detection using drones.

In recent years, due to advancements in embedded systems, the integration of UAVs with deep- learning technology has become more mature. However, computational cost and inference time have also emerged as research focuses. Many methods tradeoff detection accuracy for speed. To balance this drawback, one-stage methods, such as YOLO [18] and SSD [1], have been developed. These methods employ a regression-based approach to directly regress the coordinates of bounding boxes and object classes at multiple locations in an input image. The latest YOLOv8 series is widely applied in various real-time detection tasks. To address the issues raised in the previous section, this article proposes a model specifically for road crack detection, named road crack you only look once (RCYOLO). Simultaneously, we constructed a road crack dataset, RCI, using UAVs and environmental simulation methods. Experiments demonstrate that RCYOLO exhibits excellent speed and accuracy in different scenarios. The contributions of this article are as follows.

1) To enhance the robustness of target detection algorithms and broaden their application scope, we address the shortcomings in existing high-altitude road crack detection research and the lack of samples from harsh high-altitude environments by selecting Northwest China as the sampling area to create the RCI dataset (Fig. 2). To ensure sample diversity, we collected images at different times, under various weather conditions, and from different UAV heights and shooting angles. The RCI dataset is the first UAV road crack dataset specifically targeting high-altitude harsh environments and plays a crucial role in multiscale crack detection tasks. Traditional detection algorithms often struggle to effectively identify multiscale and small-target cracks. However, the RCI dataset, with its rich sample diversity and complex environmental conditions,

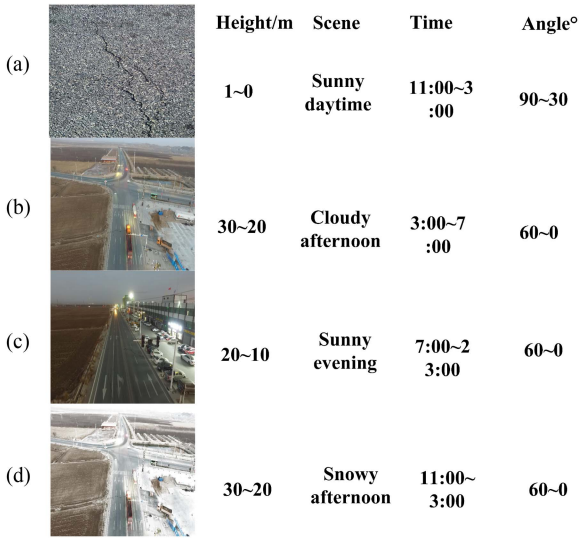

Fig. 2. Sample RCI dataset (a) near-ground captured images, (b) high-altitude UAV captured images, (c) nighttime UAV captured images, and (d) UAV captured images simulating snowy conditions.

provides ample training data, enabling models to more accurately detect and identify road cracks of different scales, especially small-scale cracks. Our dataset not only fills the gaps in existing research but also ensures the efficiency and reliability of models in practical applications.

2) The RCYOLO design is as follows.
First, Dyhead replaces the original detection head, enhancing the model's capability to detect multiscale, multicategory targets, particularly suitable for high-altitude small target detection. Second, the C2f_DySnakeConv (C2f_DSConv) module overcomes the limitations of traditional convolutional networks in processing specific shape features. Through a multiview feature fusion strategy, DSConv effectively captures target structural features from different angles and aggregates key features, significantly improving the model's performance in recognizing complex crack features. Additionally, we introduced the simulated attention mechanism (SimAM) parameter-free attention mechanism, enabling the model to more precisely focus on key areas in the image, filter out complex background interference, and identify fine and irregular road cracks. This enhances detection effectiveness without sacrificing detection speed. These combined technologies allow our model to excel in complex environments and diverse scenarios, effectively addressing the challenges in crack detection.

II. Related Work

The YOLO series networks employ a tripartite structure (backbone, neck, and detection head). Although the YOLOv8 algorithm is already quite efficient, its performance in detecting small, low-pixel, and tubular targets at long distances is not ideal, especially under the influence of the loess soil and extreme weather in Western China. This is due to the relatively low quality of the features extracted from the detection network, which is not specifically designed for small objects and multiscale targets. In this article, the original YOLOv8 network has been upgraded. Fig. 3 shows the proposed network architecture. The RCYOLO network addresses these issues through the following improvement strategies.

1) The Dyhead detection head combines three mechanisms: scale awareness, spatial attention, and channel attention, enhancing the representation ability of target detection. We introduced the Dyhead detection head into YOLOv8 to address the issues of crack size and width through multiscale perception, tackle shape and directional diversity through spatial attention, and enhance feature expression through channel attention, thereby improving detection accuracy to meet the complex challenges of road crack detection.
2) The environment for UAV-based road crack detection is characterized by numerous elongated and curled tubular structures, thin and weak local structures, and complex and variable global patterns. Standard convolution kernels aim to extract local features, while deformable convolutions can adapt to geometric deformations of different objects, enriching their applications. To address these issues, we replaced traditional convolutions in the C2F module of YOLOv8 with dynamic snake convolutions (DSConv). DSConv, with their flexible convolution paths, better capture complex and irregular crack shapes, thus improving detection accuracy and robustness.
3) We introduced the SimAM attention mechanism, which simulates human attention distribution to optimize decision-making. When targets are highly integrated with the background, SimAM strengthens focus on the target and suppresses background interference, thereby improving recognition accuracy. By calculating the importance score of each position on the feature map and adjusting responses accordingly, SimAM effectively filters out irrelevant factors in the complex environments and extreme weather of Northwest China, focusing on detecting road cracks and enhancing the model's accuracy and robustness.
4) Simple improvement strategies can effectively enhance road crack detection capabilities, but the accumulation of too many effective modules can lead to a significant increase in model parameters and a slowdown in detection speed. Moreover, excessive module stacking does not necessarily improve model accuracy. Additionally, whether the combination of improvement strategies can positively impact the network model is another issue we need to consider. We tested the quantity, position, module selection, and combination effects of the three improvement strategies to find the optimal matching model, ultimately deriving the network structure most suitable for this article.

A. Dynamic Head Framework DyHead

Akshatha et al. [18] noted that the three detection heads of YOLOv8 might underperform when dealing with small objects

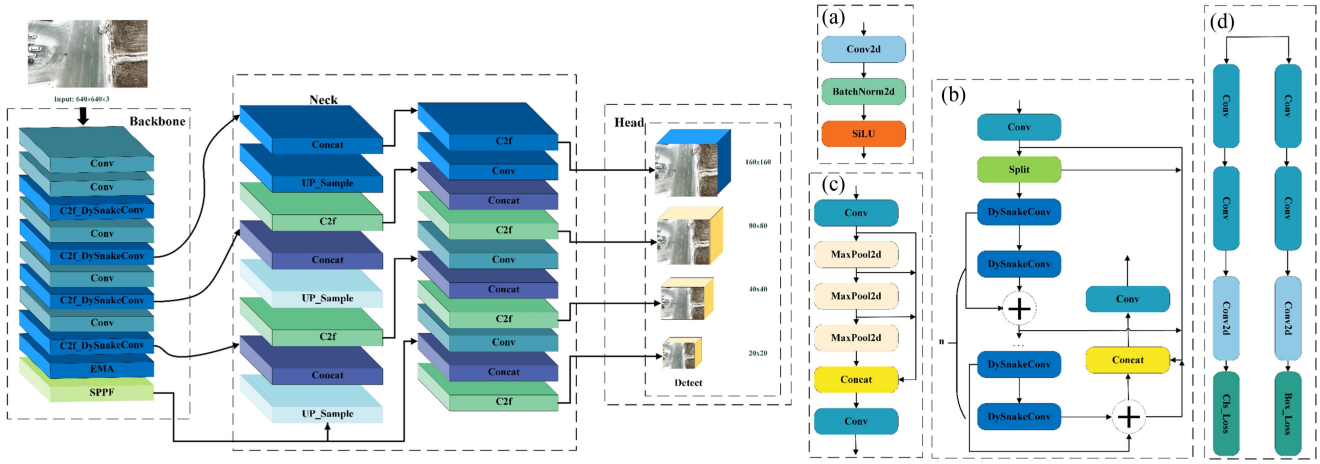

Fig. 3. Network framework of the RCYOLO model.

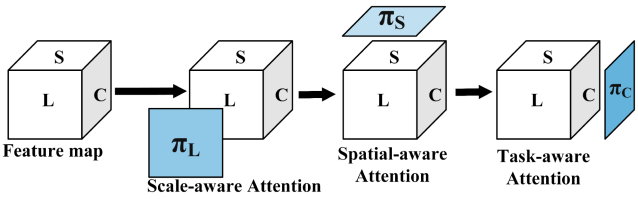

Fig. 4. DyHead block details.

and occluded targets. Most current research efforts are directed toward adding an additional detection head of a different scale. Such direct expansions could lead to significant consumption of computational resources, especially when processing high-resolution images, like those captured by high-altitude UAVs for road crack identification. In contrast, DyHead [19], as a dynamic detection head, can utilize computational resources more efficiently. DyHead offers a more refined attention mechanism. This mechanism combines scale-aware, space-aware, and task-aware capabilities, aiming to achieve more effective information integration on specific feature tensor dimensions. In this way, DyHead can more effectively identify and locate small road cracks in high-resolution images captured by high-altitude UAVs without increasing the computational burden.

First, the four-dimensional (4-D) tensor $L \times H \times W \times C$ is redefined as a 3-D tensor $L \times S \times C$. Here, L represents the number of levels in the feature pyramid, S is derived from the height (H) and width (W) of the feature map, and C denotes the number of channels. DyHead employs a segregated attention mechanism, as shown in Fig. 4, utilizing three types of attention mechanisms in sequence, with each mechanism focusing exclusively on one dimension

$$W\ (F) = \pi_C\ (\pi_S\ (\pi_L\ (F) \cdot F) \cdot F) \cdot F. \tag{1}$$

Among them, $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ are three separate attention functions applied to dimensions L, S, and C, respectively.

To effectively integrate features from different scales based on semantic relevance, we propose a scale-aware attention mechanism. This mechanism, known as the attention function $\pi_L(\cdot)$, is designed to emphasize the significance of scale in feature fusion. The formula for $\pi_L(\cdot)$ is as follows (2):

$$\pi_L\ (F) = \sigma\left(f\left(\frac{1}{SC}\sum_{S,C} F\right)\right) F. \tag{2}$$

The function $f(\cdot)$ is approximated using a 1×1 convolution as a linear function, and $\sigma\ (x) = \max(0, \min(1, (x+1)/2))$.

In the space-aware attention module, deformable convolutions [19] are employed to learn sparsity. Compared to standard convolutional networks, the primary advantage of deformable convolutions lies in their lack of fixed receptive fields and structures. This means they are more efficient in handling spatial transformations in images, such as scaling, rotation, and deformation. Then, features across levels are aggregated at the same spatial location. Its expression is shown as follows:

$$\pi_S\ (F) \cdot F\ = \frac{1}{L}\sum_{K=1}^{K}\sum_{L=1}^{L} \omega_{l,k} \cdot F\ (1; p_k + \Delta p_k; c) \cdot \Delta m_k. \tag{3}$$

K represents the number of sparse sampling locations. $p_k + \Delta p_k$ is the shifted position in different areas, focused around the self-learned spatial offset Δp_k. Δm_k is the self-learned importance scalar at position p_k. All of these are learned from the input features of the intermediate level F.

The task-aware attention module dynamically turns ON or OFF feature channels to select for different tasks. Its expression is detailed as follows:

$$\pi_C\ (F) \cdot (F)) = \max\begin{pmatrix}\alpha^1\ (F) \cdot F_C + \beta^1\ (F) \\ \alpha^2\ (F) \cdot F_C + \beta^2\ (F)\end{pmatrix}. \tag{4}$$

$[\alpha^1, \beta^1, \alpha^2, \beta^2]^T = \theta(\cdot)$ is a hyperfunction used to learn and control activation thresholds. $\theta(\cdot)$ first performs global pooling to reduce dimensions in the $L \times S$ dimension, then employs two fully connected layers and a normalization layer, and finally uses a shifted sigmoid function to normalize the output to the range of $[-1, 1]$. Fig. 5 illustrates this, where π_L, π_S, and $\pi_C(F)$ correspond to the scale-aware, space-aware, and task-aware attention modules, respectively.

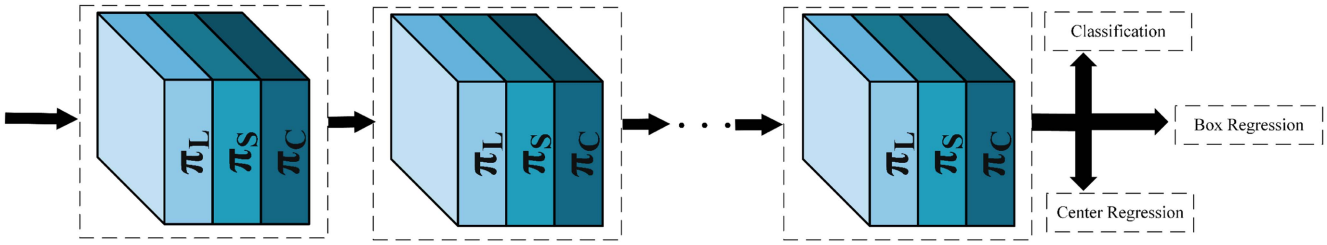

Fig. 5. Connection scheme of DyHead blocks.

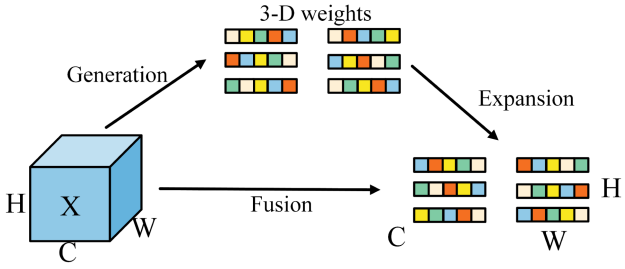

Fig. 6. SimAM module.

B. SimAM Attention Mechanism

The attention mechanism leverages the most important features in the input sequence by weighting and combining the input vectors. Images captured by high-altitude drones, especially those from the RCI dataset in the northwest region, often contain complex backgrounds that cause severe interference. To address this issue in the YOLOv8 architecture, we integrated the SimAM attention mechanism. SimAM is a 3-D attention module that, based on neuroscience theory, proposes an improved energy function and derives its analytical solution to enhance the efficiency of attention weight calculation. SimAM simulates human attention allocation, focusing more on the target while suppressing background interference in cases where the target and background are highly integrated, thus improving recognition accuracy. By calculating the importance score of each position on the feature map and adjusting the feature response accordingly, SimAM enables the model to more effectively highlight key targets and suppress irrelevant background information in complex scenes, thereby enhancing the model's decision-making accuracy.

Its unique reliance on 3-D weights and the energy function significantly speeds up the process. A key feature of SimAM is its lightweight, nonparametric design, which adds almost no additional parameters compared to other attention mechanisms. The SimAM attention module is placed at various levels of the network, and its effects at each position are detailed in the table. Architectural details of SimAM can be seen in Fig. 6.

Overall, SimAM utilizes its 3-D weights and energy function for rapid weight computation. Additionally, as a nonparametric module, the extra advantage is that the energy function for each neuron is as follows:

$$e_t\ (w_t, b_t, y, x_i) = \left(y_t - \hat{t}\right)^2 \frac{1}{M-1} \sum_{i=1}^{M-1} \left(\begin{matrix} y_0 - \\ \hat{x}_i \end{matrix} \right)^2 \tag{5}$$

where $\hat{t} = w_t\, t + b_t$ and $\hat{t} = w_t\, t + b_t$ are linear transformations of t and x_i, respectively, with t being the target neuron and x_i representing other neurons in a single channel of the input feature

$$e_t\ (w_t, b_t, y, x_i) = \frac{1}{M-1} \sum_{i=1}^{M-1} \left(-1 - (w_t x_i + b_t)\right)^2 + \left(1 - (w_t + b_t)\right)^2 + \lambda w_t^2 \tag{6}$$

where w_t and b_t can be easily obtained by

$$w_t = -\frac{2\,(t - \mu_t)}{(t - \mu_t)^2 + 2\sigma_t^2 + 2\lambda} \tag{7}$$

$$b_t = -\frac{1}{2}\,(t + \mu_t)\, w_t \tag{8}$$

where μ_t and σ_t^2

$$\mu_t = \frac{1}{M-1} \sum_{i=1}^{M-1} x_i \tag{9}$$

$$\sigma_t^2 = \frac{1}{M-1} \sum_{i=1}^{M-1} (x_t - \mu_t)^2. \tag{10}$$

Mean and variance are ascertained for all neurons, excluding neuron t. The computation of minimal energy is elucidated as per (11), which illustrates that the lower energy e_t^* signifies the uniqueness of neuron t in contrast to its neighboring neurons

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{11}$$

where $\hat{\mu}$ and $\hat{\sigma}^2$

$$\hat{\mu} = \frac{1}{M} \sum_{i=1}^{M} x_i \tag{12}$$

$$\hat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \hat{\mu})^2. \tag{13}$$

C. Dynamic Snake Convolution Structure

Considering the tubular and sparse nature of road cracks, the authors in [20] introduced a cross-network multiscale feature fusion technique between two different networks to support high-precision vascular segmentation. In [21], a method combining deep and shallow feature fusion with a global transformer and a dual local attention network was explored, capturing

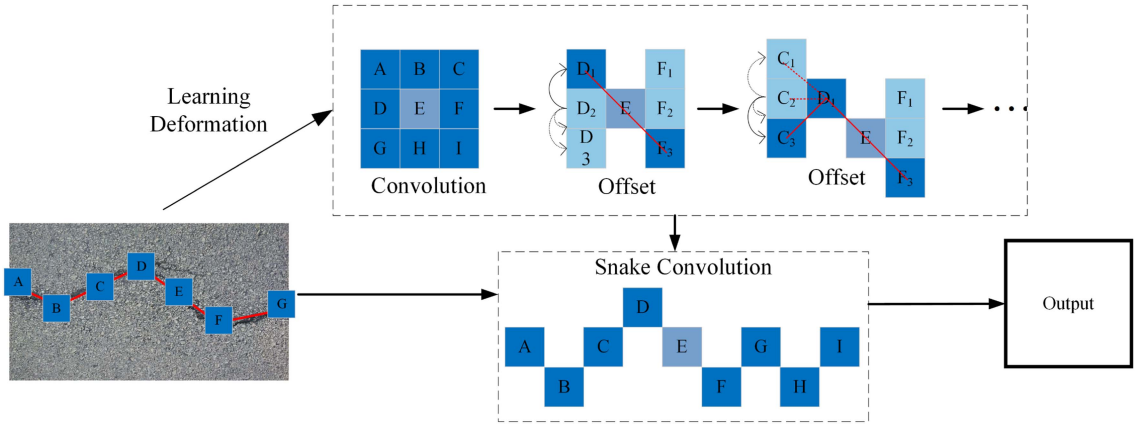

Fig. 7. Principles of dynamic snake convolution (DSConv) for feature extraction.

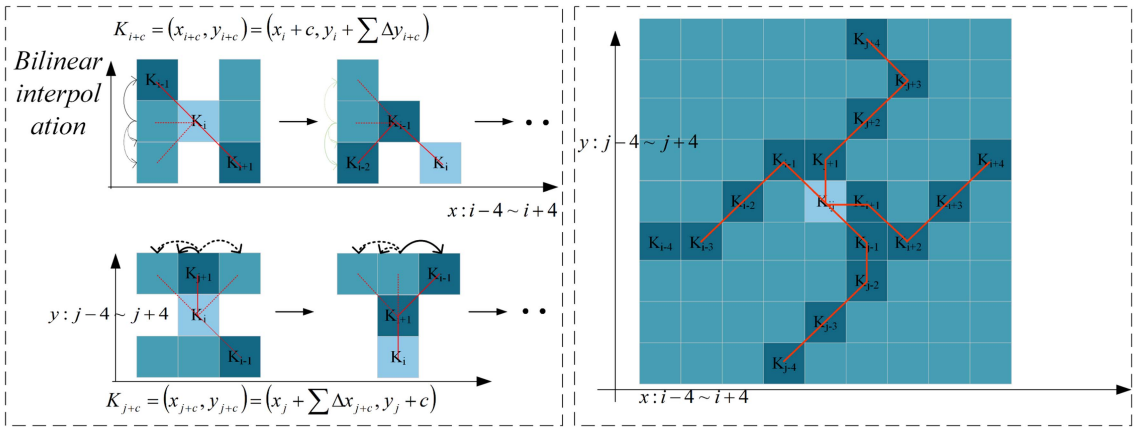

Fig. 8. Coordinate calculation of DSConv (left) and the receptive field of DSConv (right).

both global and local features. The authors in [22] proposed a method integrating contextual anatomical information with the vascular topological structure for precise segmentation of tubular structures. Given the similarity between vascular and road crack features, we propose a multiview feature fusion strategy to enhance attention to significant features from multiple perspectives. Integrating DSConv into YOLOv8 enhances the structural feature performance of road cracks. The structure of DSConv, shown in Fig. 7, introduces continuity constraints in the design of the convolutional kernel, allowing each convolutional position to choose its oscillation direction based on the previous position. This ensures the continuity of the receptive field while maintaining flexibility.

Our objective is to extract the local characteristics of tubular structures. Initially, we assume a coordinate system for the 2-D convolution, denoted as K, with the central coordinate being the most critical, represented as $K_i = (x_i, y_i)$. Within this framework, we employ a 3×3 convolution kernel with a dilation factor of 1. K is represented as

$$K = \{(x-1, y-1), (x-1, y), \cdots (x+1, y+1)\}. \quad (14)$$

The standard convolution kernel is linearized along both the x-axis and y-axis. The kernel size is set to 9, and the position of each grid point in K is expressed as $K\_(i\pm c) = (x\_(i\pm c), y\_(i\pm c))$. Here, $c = \{0, 1, 2, 3, 4\}$ represents the horizontal distance of the grid points from the central grid. The selection of grid positions in K is a cumulative process, starting from the central position Ki. Each adjacent grid position is determined relative to the previous one, with an offset $\Delta = \{\delta | \delta \in [-1, 1]\}$. The total sum of these offsets is denoted by Σ. In summary, these parameters and notations are used to describe and control the grid positions of the convolution kernel and their relative movements. As depicted in Fig. 8, the left side presents a graphical representation of the coordinate calculations in DSConv, whereas the right side shows the receptive field of DSConv

$$K_{i\pm c} = \begin{cases} (x_{i+c}, y_{i+c}) = \left(x_i + c, y_i + \sum_i^{i+c} \Delta y\right) \\ (x_{i-c}, y_{i-c}) = \left(x_i - c, y_i - \sum_{i-c}^{i} \Delta y\right). \end{cases} \quad (15)$$

For A, bilinear interpolation is executed. Here, K denotes fractional positions, while K' represents positions in integer space, and B is the bilinear interpolation kernel. The formula for K is as described as follows:

$$K = \sum_{K'} B(K', K) \cdot K'. \quad (16)$$

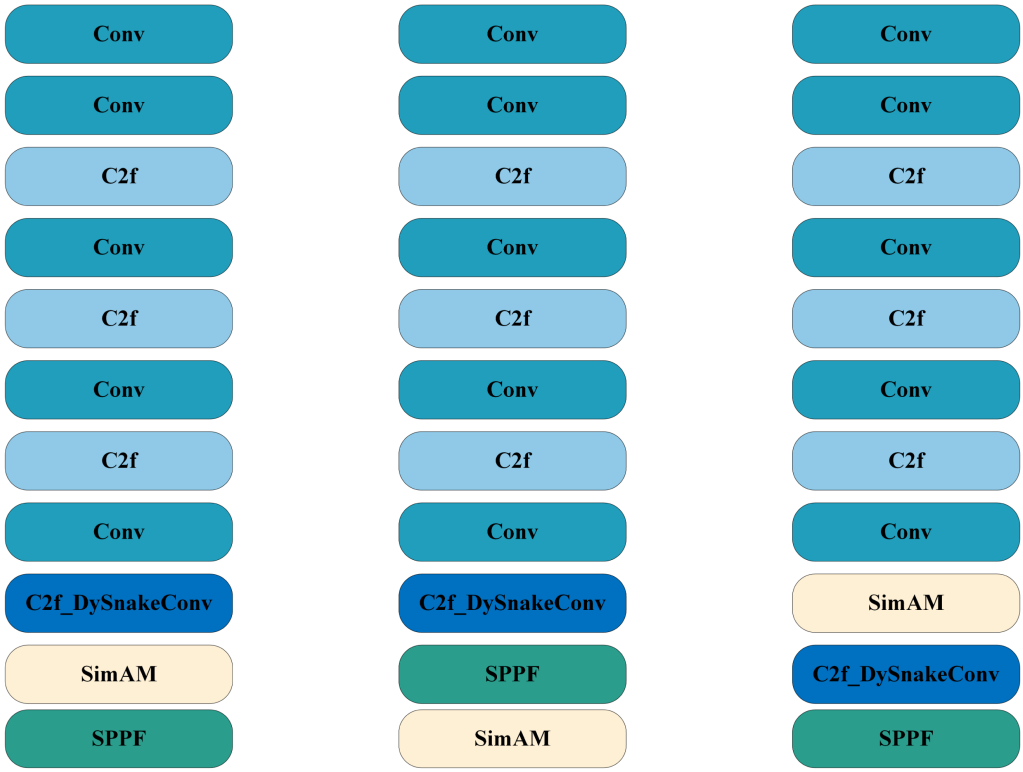

Fig. 9. Position of SimAM, showing (a) above SPPF, (b) below SPPF, and (c) between C2f_DySnakeConv and conv of the final backbone. (a) YOLO-S. (b) YOLO-S1. (C) YOLO-S2.

TABLE I
EXPERIMENTAL ENVIRONMENT SETTINGS

Hardware environment	CPU	Inte(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
	GPU	NVIDIA GeForce RTX 3080
	RAM	64G
Software environment	OS	Windows 10
	CUDA Toolkit	12.2
	Python	3.8.18
Training information	Optimizer	SGD
	Epoch	200
	Batch size	8
	Learning range	0.01

In addition to the research designs mentioned above, we also considered the concerns raised in [17] about the positioning of modules. Currently, there is almost no research on the optimization of the positions of the SimAM module and C2f_DSConv within YOLOv8. In this article, we experimented with three different placements of the SimAM module, as illustrated in Fig. 9.

Regarding the placement of C2f_DSConv, the article is relatively unique. It requires considering the situation of applying C2f_DSConv in a nonglobal backbone from a quantitative perspective, as well as the scenario of applying C2f_DSConv in a global backbone. This is illustrated in Fig. 10.

III. EXPERIMENTS AND RESULTS

A. Experimental Environment and Dataset

1) Environment: All experiments were conducted in the same hardware and software environment to test the performance of the RCYOLO object-detection algorithm. The specific environmental parameters are listed in Table I. We used the stochastic gradient descent (SGD) optimizer, set the training epochs to 200, the batch size to 8, and the learning rate to 0.01. The SGD optimizer was chosen for its stability and efficiency in handling large-scale datasets and deep-learning tasks. The training epochs were set to ensure the model could fully learn

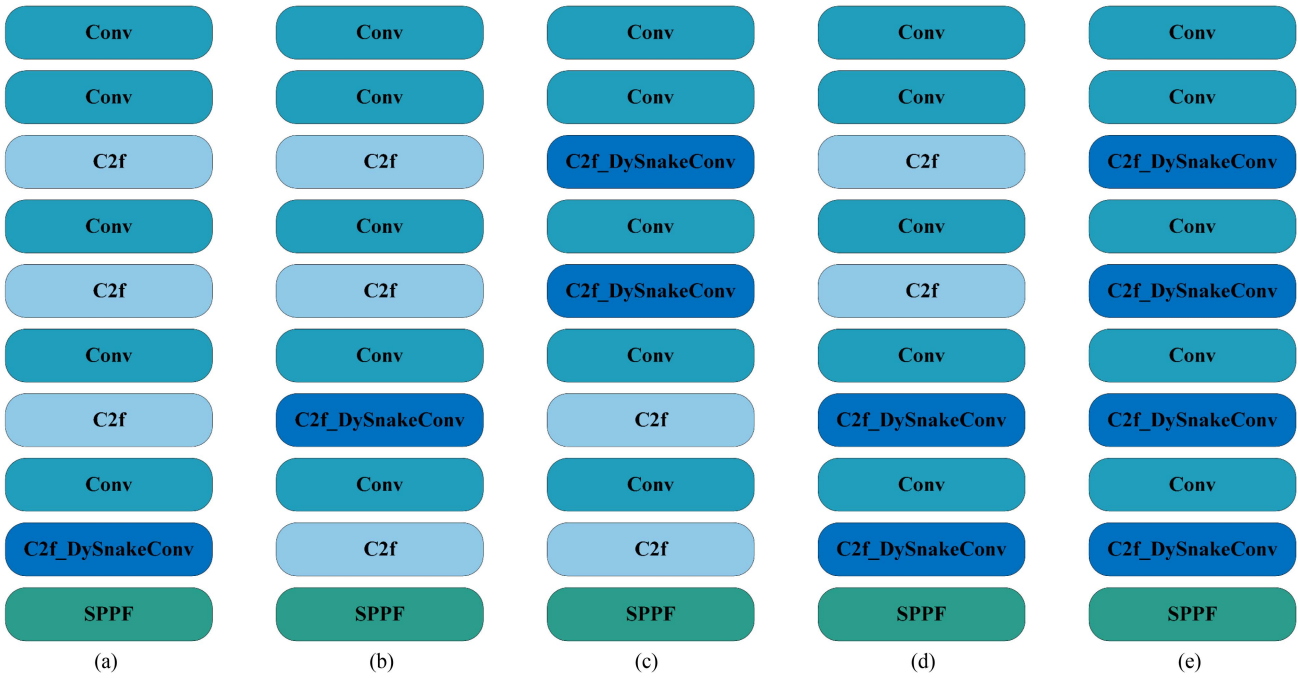

Fig. 10. Application quantity and positions of C2f_DySnakeConv: (a) C2f_DySnakeConv at the position of the fourth c2f, (b) position of the third c2fs, (c) positions of the first and second c2fs, (d) positions of the fourth and third c2f, and (e) C2f_DySnakeConv used globally in the backbone. (a) YOLO-D. (b) YOLO-D1. (c) YOLO-D2. (d) YOLO-D3. (f) YOLO-D4.

the data features, while the smaller batch size provided more stable gradient updates. The initial learning rate of 0.01 ensured a balance between convergence speed and stability. Table I lists the hardware environment, including the Intel Xeon Silver 4210R CPU @ 2.40 GHz, NVIDIA GeForce RTX 3080 GPU, and 64G RAM, as well as the software environment, including the Windows 10 operating system, CUDA Toolkit 12.2, and Python 3.8.18.

2) Aerial Road Crack Dataset—RCI Dataset: The RCI dataset consists of 1880 images focused on the detection of high-altitude road cracks. These images were captured using a DJI Inspire 1 drone in the loess environment of Northwestern China and include 20 images simulating snowy weather conditions. The dataset is divided into 1316 training images, 376 validation images, and 188 test images. A significant advantage of this dataset is its comprehensive coverage of the diverse terrains of Northwestern China, making it an ideal platform for assessing the network's road crack detection capabilities in complex terrains and varying weather conditions. Additionally, the diversity of the RCI dataset makes it an excellent choice for testing the network's robustness in adapting to different environmental conditions, especially under extreme weather conditions. Therefore, this dataset is not only suitable for the identification and classification of road cracks but also for testing the algorithm's performance in terrain recognition and environmental adaptability.

3) Public Dataset—DOTA Dataset: As shown in Fig. 11, the Detection of Objects in Aerial Images (DOTA) dataset [23] is a large-scale aerial image object detection dataset released by Wuhan University on November 28, 2017, specifically designed for object detection in aerial images. It comprises 2806 high-resolution images covering 15 different categories:

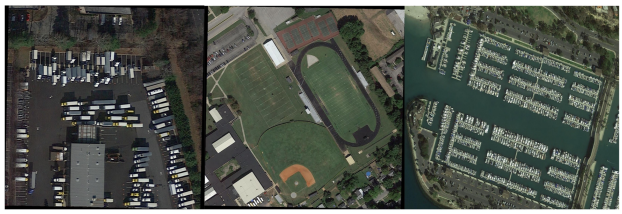

Fig. 11. Public dataset—DOTA sample.

0: small-vehicle, 1: large-vehicle, 2: plane, 3: storage tank, 4: Ship, 5: harbor, 6: ground-track-field, 7: soccer-ball-field, 8: tennis-court, 9: swimming-pool, 10: baseball-diamond, 11: roundabout, 12: basketball-court, 13: bridge, 14: helicopter. The dataset is divided into 1964 training images, 561 validation images, and 281 test images.

The DOTA dataset's advantages lie in its large scale and high-quality images, making it particularly suitable for testing and enhancing the performance of deep-learning models in aerial image analysis. This dataset can be used to assess the network's capabilities in object detection, classification, localization, and semantic segmentation. Due to the diversity and complexity of its images, the DOTA dataset is especially suitable for researching and improving algorithms in handling high background interference and accurately and robustly processing targets of varying scales, angles, shapes, and sizes.

B. Evaluation Criteria

In order to understand the model's performance, we use various metrics including mAP, parameters, GFLOPS, and FPS. Precision: Precision refers to the ratio of the number of samples

TABLE II
RESULTS OF ABLATION OF THE ALGORITHM MODULE IN RCI

Dyhead	C2f_DSConv	SimAM	mAP@0.5	Recall	Precision	FPS
–	–	–	0.8744	0.8183	0.8412	81
√	–	–	0.9018	0.8587	0.8310	80
–	√	–	0.9093	0.8443	0.8895	79
–	–	√	0.9151	0.8452	0.9200	80
√	√	–	0.9235	0.8679	0.9233	77
√	–	√	0.9198	0.8669	0.9192	78
–	√	√	0.9263	0.8795	0.9222	79
√	√	√	0.9336	0.8836	0.9390	79

correctly predicted as positive by the model to the total number of samples predicted as positive. Recall is calculated by the percentage of all positive samples that are correctly predicted. mAP is a performance metric for multiclass classification problems, which considers the precision–recall curve of different categories and calculates their average precision value. mAP@0.5 means that the IoU threshold is set at 0.5 when calculating mAP. GFLOPS represents the number of billion floating-point operations per second. FPS, the number of images processed per second or the time taken to process a single image, is used to assess the detection speed as shown in formulas as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

$$\text{mAP} = \frac{\sum_{j=1}^{M} \text{AP}_j}{M} \tag{19}$$

$$\text{FPS} = \frac{1}{t}. \tag{20}$$

In this context, true positives refer to the number of positive class samples correctly predicted as positive by the model. False positives (FP) are the number of negative class samples that are incorrectly predicted as positive by the model. False negatives (FN) are the number of positive class samples that are incorrectly predicted as negative by the model. True negatives are the number of negative class samples correctly predicted as negative by the model. The term (number) N represents the number of object classes.

C. Ablation Study on RCI Dataset

We employed the RCYOLO model as the baseline in our ablation experiments on the RCI dataset to examine the effectiveness of multiple modules. The experimental results are shown in Table II. It is noteworthy that besides prioritizing the accuracy of the model, we also emphasized its efficiency. In Table II, we compared several key evaluation parameters of concern for practical applications under different module ablations. It can be observed that each of the three modules proposed by us improves the mAP@0.5 without significantly sacrificing the FPS value. Particularly, the C2f_DSConv and SimAM modules increased the mAP@0.5 by 3.5% and 4.1%, respectively. This confirms that our proposed DSConv module is far superior to standard convolution in extracting features of tubular topological small target structures like road cracks, and the nonparametric design of SimAM is powerfully effective in filtering interference factors in the background without overly impacting inference speed. When the Dyhead module is added alone, both mAP@0.5 and Recall increase, but we found that Precision decreases by 1.02%. This is because the DyHead module enhances feature extraction capabilities through dynamic convolution and attention mechanisms, which may enable the model to detect more positive samples (targets), thereby improving Recall. However, this enhancement might also lead to an increase in FP in background or nontarget areas, thus reducing Precision. When Dyhead is combined with other modules, the metrics normalize or even improve due to the complementary nature of the modules. The Dyhead module enhances feature extraction, while the C2f_DSConv and SimAM modules optimize spatial information utilization and attention mechanisms, respectively. This complementarity allows the combination to better balance feature extraction and FP control. Second, multiscale feature fusion plays a role. The Dyhead module handles multiscale features, and the C2f_DSConv and SimAM modules optimize the utilization of these features, improving the model's accuracy in detecting targets of different scales and reducing FP. Overall, our RCYOLO, compared to the original YOLOv8, improved mAP@0.5, Recall, and Precision by 5.9%, 6.5%, and 9.8%, respectively, proving that our optimizations are effective for high-altitude UAV road crack detection applications.

1) Research on Optimal Performance of Attention Mechanisms: In Table III, we tested the performance of YOLOv8 detectors constructed with five different cutting-edge and widely used attention mechanisms on our custom high-altitude UAV road crack dataset. For fairness, the same hyperparameters were used across all tests. The attention mechanisms tested include SE [25], EMA [26], GAM [27], CA [28], and SimAM. It is evident that, due to its nonparametric design, SimAM holds certain advantages in terms of both the number of parameters and computational load. Importantly, SimAM also excels in accuracy performance, showing the most significant improvement. Compared to the detection model without any attention mechanism on the same dataset, SimAM increased the mAP@0.5% by 4.07%, thus proving to be the most optimal in this article.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT ATTENTION MODULES IN RCI

	mAP@0.5	Recall	Precision	Param/M	GFLOPs	FPS
None	0.8744	0.8183	0.8412	3.01	8.2	81
SE[25]	0.9093	0.8551	0.8803	3.10	8.4	80
EMA[26]	0.9018	0.8587	0.8310	3.10	8.4	79
GAM[27]	0.9143	0.8443	0.8695	3.46	9.5	77
SimAM	0.9151	0.8562	0.9200	3.05	8.3	80
CA[28]	0.9112	0.8324	0.8983	3.05	8.3	79

TABLE IV
OPTIMAL PERFORMANCE OF SIMAM IN RCI

	mAP@0.5%	Recall%	Precision%.
YOLO-S	0.9263	0.8795	0.9222
YOLO-S1	0.9142	0.8351	0.9237
YOLO-S2	0.9139	0.8548	0.8924

From Table IV, we can see that the concerns about module placement mentioned in the previous chapters have been addressed. Consequently, we designed experiments to incorporate attention mechanisms at three different positions in our designed backbone structure, as illustrated in Fig. 9. We found that due to the nonparametric nature of SimAM, changes in its position do not affect the number of parameters and computational load, but there are differences in accuracy. We observed that the precision is highest in the YOLO-S placement scheme. Compared to the second-best performing YOLO-S1 placement scheme, YOLO-S achieved higher mAP@0.5 and Recall by 1.2% and 4.4%, respectively, with only a 0.1% loss in Precision. This comprehensive comparison confirms that the other two connection methods are not optimal. Therefore, we chose the YOLO-S structure as our preferred option.

2) Optimal Optimization Study of C2f_DySnake Conv: In the previous phase, we discussed the study of attention mechanism module placement. In this chapter, we address not only the placement issues of the C2f_DSConv module but also its distribution in different quantities. While planning to optimize the network backbone with the C2f_DSConv module, we discovered that using C2f_DSConv across the entire backbone not only yields a minimal increase in detection accuracy but also leads to unnecessary increases in parameters and computational load, making real-time detection more challenging.

As shown in Table V, YOLO-D and YOLO-D4 exhibit the best overall detection performance among several optimization strategies. However, it is evident that YOLO-D4, compared to YOLO-D, which has a better balance of accuracy and speed with a single C2f_DSConv backbone structure, shows a 0.04% decrease in mAP@0.5, but increases in Recall and Precision by 1.68% and 0.28%, respectively. It is important to note that YOLO-D4 brings a substantial computational burden. The Param(M) and GFLOPs of YOLO-D4 are higher than those of YOLO-D by 0.44M and 0.7, respectively, and the FPS decreases by 2. From a comprehensive cost-effectiveness perspective, a single C2f_DSConv backbone structure is superior to a global C2f_DSConv backbone structure. In the two double C2f_DSConv backbone structures YOLO-D2 and YOLO-D3, YOLO-D3 shows slightly higher Param(M) than YOLO-D2 by 0.4, but with better detection performance and nearly the same detection speed as YOLO-D3. However, compared to the single and double C2f_DSConv backbone structures, YOLO-D outperforms YOLO-D3 in mAP@0.5 and Recall by 0.37% and 0.36%, respectively, albeit with a 0.52% lower Precision. Importantly, YOLO-D is significantly better than YOLO-D3 in terms of computational load and detection speed. Considering all these factors, we have chosen the strategy offered by YOLO-D.

3) Research on the Optimal Performance of the Head Section: In Table VI, a performance comparison was conducted between the original YOLOv8 head and three commonly used heads (Dyhead, YOLOX-Head, and Efficient-Head). The results show that Dyhead performs best overall. Its mAP@0.5 is 0.0274 higher than YOLOv8-Head, 0.0387 higher than YOLOX-Head, and 0.0095 higher than Efficient-Head. In terms of Recall, Dyhead is 0.0404 higher than YOLOv8-Head, 0.0140 higher than YOLOX-Head, and 0.0055 higher than Efficient-Head. Although YOLOv8-Head has slightly higher Precision, the difference is only 0.0102 compared to Dyhead, while Dyhead is 0.0009 lower than YOLOX-Head and 0.0056 higher than Efficient-Head. Overall, Dyhead performs the best in terms of mAP@0.5 and Recall, with reasonable parameter and computation requirements, and only slightly lower inference speed than YOLOv8-Head. In multiscale small object detection, Dyhead demonstrates significant advantages in recognizing and locating complex road cracks due to its excellent feature extraction capabilities. Therefore, we chose Dyhead as the head part of RCYOLO.

D. Generalization and Robustness Testing

To verify the generalization ability and robustness of our RCYOLO, we tested our model on the DOTA remote sensing dataset and compared it with the original YOLOv8 model. The DOTA dataset, with its classification containing highly similar objects, 15 detection targets, small objects, and complex background interference, provides an excellent platform to evaluate the generalization and robustness of detection models. As shown in Table VII, RCYOLO's mAP@0.5 overall outperforms YOLOv8 by 5.8%. For smaller sized objects like small-vehicle and helicopter, where the other two models show weaker performance,

TABLE V
SPATIAL DISTRIBUTION AND QUANTITY OF C2F_DYSNAKECONV MODULES IN RCI

	mAP@0.5%	Recall%	Precision%	Param(M)	GFLOPs	FPS
YOLO-D	0.9093	0.8443	0.8895	3.01	8.3	79
YOLO-D1	0.9028	0.8320	0.8863	3.01	8.4	79
YOLO-D2	0.9042	0.8411	0.8940	3.00	8.6	78
YOLO-D3	0.9056	0.8407	0.8947	3.40	8.5	78
YOLO-D4	0.9089	0.8611	0.8923	3.45	9.0	77

TABLE VI
OPTIMAL PERFORMANCE OF SIMAM IN RCI

Head	mAP@0.5	Recall	Precision	Param(M)	GFLOPs	FPS
YOLOv8-Head	0.8744	0.8183	0.8412	3.01	8.2	81
Dyhead	0.9018	0.8587	0.8310	3.02	8.4	80
YOLOX-Head	0.8631	0.8447	0.8319	3.08	8.5	79
Efficient-Head	0.8923	0.8532	0.8254	3.07	8.5	79

TABLE VII
GENERALIZABILITY AND ROBUSTNESS IN DOTA DATASET

Network	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	mAP@0.5%
YOLOv8	0.642	0.879	0.955	0.757	0.791	0.766	0.589	0.617	0.872	0.804	0.565	0.348	0.641	0.675	0.347	0.604
RCYOLO	0.713	0.882	0.944	0.772	0.804	0.756	0.619	0.701	0.867	0.771	0.663	0.533	0.715	0.682	0.424	0.662

RCYOLO improves accuracy over YOLOv8 by 7.1% and 7.7%, respectively. On the other hand, for targets with high similarity and heavy background interference, like basketball-court and baseball-diamond, RCYOLO also demonstrates superior detection effectiveness compared to YOLOv8, with a 7.4% and 9.8% increase in accuracy, respectively. This proves that our optimization effectively addresses the issues of poor generalization and low robustness in existing detectors when dealing with small targets, high similarity, and complex background.

E. Comparative Experiments

We compared our method with other cutting-edge and widely used detectors, such as RetinaNet [29], faster R-CNN [30], YOLOv5 [31], YOLOv7 [32], and YOLOv8, in terms of six parameters: mAP@0.5, Recall, Precision, Param, GFLOPs, and FPS, as shown in Fig. 12. It is evident that RCYOLO outperforms other advanced remote sensing detectors in all these parameters. Compared to the baseline model YOLOv8, RCYOLO has an increase of 0.21M in parameters, 0.3 in GFLOPs, and a decrease of 2 in FPS. However, RCYOLO shows significant improvements in mAP@0.5, Recall, and Precision, with increases of 5.9%, 6.43%, and 3.78%, respectively. From a comprehensive perspective, RCYOLO performs the best among all the compared algorithms.

To visually assess the effectiveness of RCYOLO, the experiment provides images of the detection results. As shown in Fig. 13(a), within the DOTA dataset, YOLOv8 failed to detect small targets. In RCI dataset, in Fig. 13(b), in the near-ground images, it can be observed that the original YOLOv8 produced false detections, mistakenly identifying background areas with high similarity as cracks. In Fig. 13(c), YOLOv8 significantly misidentified land cracks as road cracks in high-altitude images. In Fig. 13(d), due to the interference of the snowy landscape, the YOLOv8 detector misidentified steps on the road as cracks and also exhibited serious omissions in detection. In contrast, RCYOLO demonstrated excellent performance on both the RCI dataset and the DOTA dataset.

IV. DISCUSSION

This article introduces a novel concept of replacing traditional standard convolutions with DSConv. By leveraging the adaptive focus of DSConv on elongated and curved local structures, it accurately captures the tubular structure characteristics of cracks. Compared to methods proposed in recent crack detection studies, such as [24] and [33], there appears to be a scarcity of research dedicated to experimenting with optimizations of standard convolutions. Through the comparison of the first and third groups of ablation experiments in Table II, the incorporation of the DSConv into the new C2f_DSConv module within the RCI dataset for crack detection tasks resulted in improvements of 3.9%, 2.6%, and 4.8% in mAP@0.5, Recall, and Precision, respectively, over the standard convolution-based C2f module. However, in crack detection methods based on UAVs, to address the issue of background interference, attention mechanisms are generally employed. Despite this, existing studies tend to overlook the computational and parameter burden while solving background noise issues. Compared to the SE and CA attention mechanisms mentioned in [34] and [35] for addressing

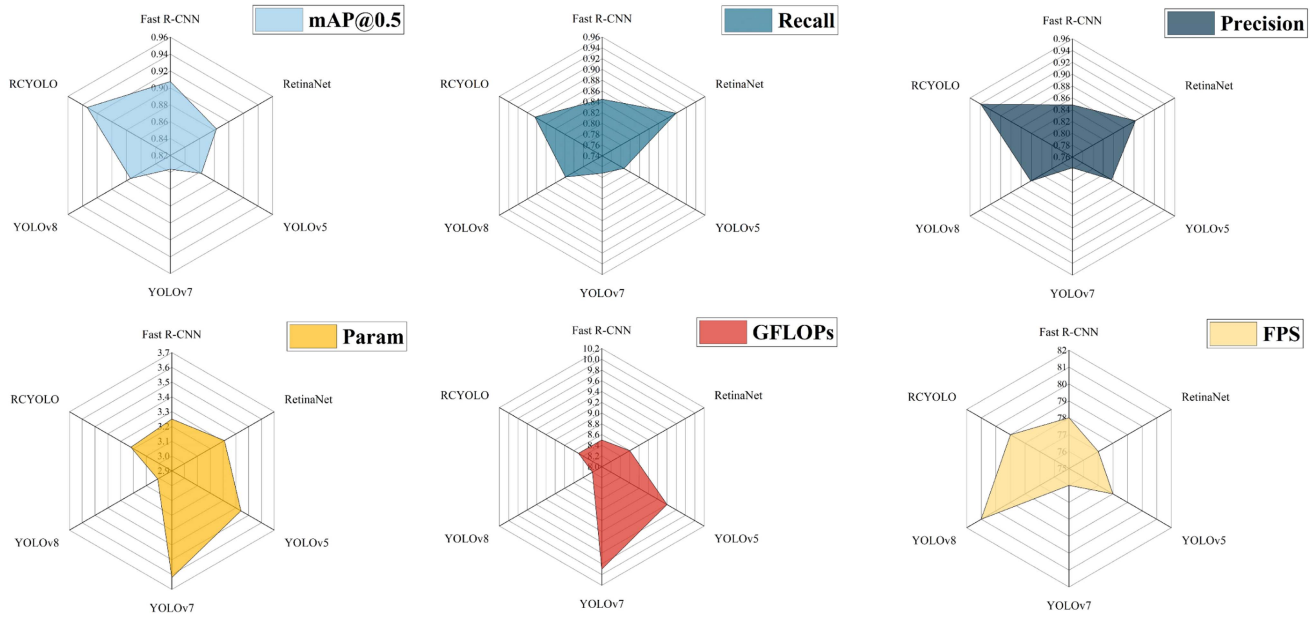

Fig. 12. Public dataset—DOTA sample.

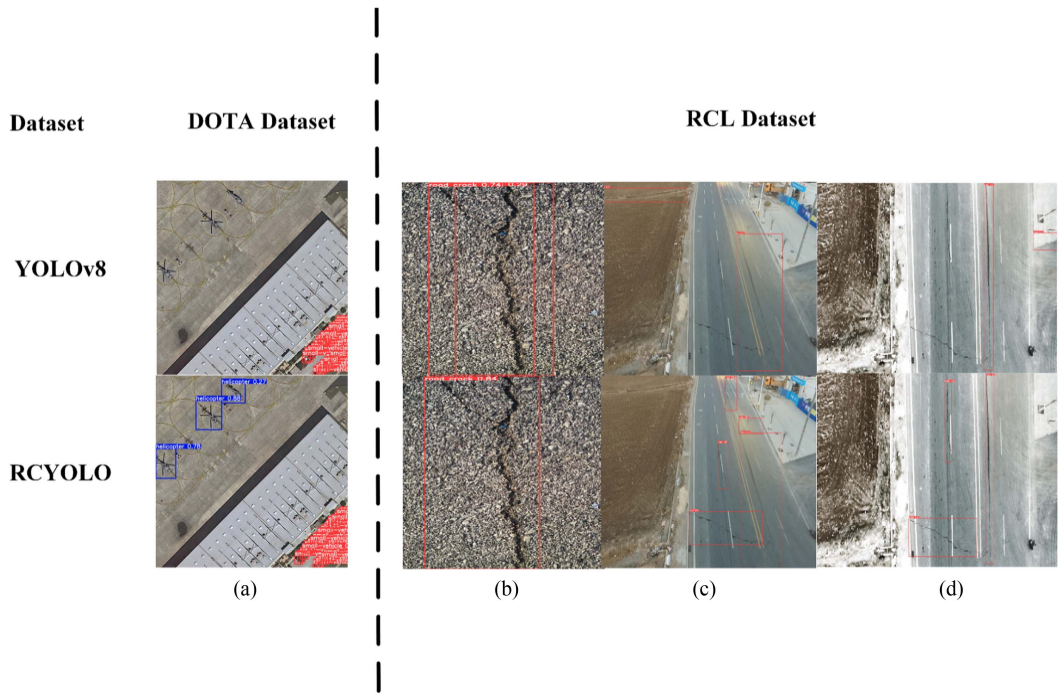

Fig. 13. Compared results with four types of samples. (a) DOTA dataset. (b) Close-up photography of a crack. (c) High-altitude UAV photgraphy of cracks. (d) Simulated snowscape with cracks.

background distractions, our chosen parameterless SimAM attention mechanism showed a reduction of 0.09M in parameters and a decrease of 0.2 GFLOPs in Table III, along with a 0.58% increase in mAP@0.5. In contrast, with CA, where the differences in parameters and GFLOPs were negligible, mAP@0.5 saw an increase of 0.39%. Striving to overcome the prevalent focus on near-ground detection, when employing UAV detection methods, we faced the challenge of detecting small objects. To this end, we adopted a variable detection head featuring three attention mechanisms, which significantly outperformed other algorithms in detecting small objects. In Table VII, testing our RCYOLO model against the advanced YOLOv8 model on the DOTA public remote sensing small object dataset showed an overall improvement of 5.8% in mAP@0.5. Additionally, in Fig. 13(b), the enhanced granularity recognition of RCYOLO in small object detection is clearly observable. Although the

RCYOLO model currently shows promising performance, it still faces several distinct challenges.

A. FN Program

Despite the proposed method showing good robustness in our UAV and close-range images, by observing the experiments in part three, we found that the Recall values are generally lower than Precision, indicating that the issue of FN still exists. In our samples, there are some pixels similar to cracks, leading to the omission of features. We anticipate overcoming this problem by incorporating more pixel regions similar to cracks into the samples.

B. Application Capability Expansion

The current article primarily focuses on tubular-structured road cracks, with the proposal of using drones as carriers. Initially, our tests were conducted solely on PC graphics card devices. However, in practical applications, UAVs should be equipped with portable devices, such as FPGAs and embedded devices. This must undergo experimentation before being deployed in real-world applications. Furthermore, the current scope of article on cracks exhibits a somewhat singular functionality. Future articles could explore areas, such as cracks in buildings and railway tracks. These targeted objectives, characterized by their tubular structures, hold significant value for testing the generalizability of our model.

C. Module Position Study

In the current article on the placement and number of modules, we proposed three placement strategies for SimAM as illustrated in Fig. 9, and for C2f_DSConv, three quantity and five placement strategies as shown in Fig. 10. Experimental results indicate that different quantities and placements have varying impacts on model performance. However, due to the relatively limited experimental samples, we have not conducted an in-depth study on whether the impact of placement and quantity on the model is fixed. In short, it remains to be seen whether the optimal placement and quantity discussed will still hold on different datasets. Further article in this area may still be necessary in the near future.

V. Conclusion

Road defects are a critical aspect of road maintenance management and road safety. With the continuous increase in road mileage, traditional manual road detection methods cannot meet the requirements of large-scale road maintenance in terms of accuracy and efficiency. The detection methods using automated road detection vehicles not only affect normal traffic but also face image occlusion issues.

This article proposes a UAV-based road crack detection system, RCYOLO. The system incorporates the C2f_DSConv backbone module, which effectively extracts complex morphological features of cracks. It also employs a parameter-free SimAM attention mechanism to enhance the network's feature fusion, enabling the algorithm to identify key information for distinguishing cracks. To address the issue of detecting small targets and multiscale detection in high-altitude images, we designed the dyhead, a dynamic detection head that combines various attention mechanisms, improving localization and classification performance. Experimental results show that RCYOLO achieves 93.36% (+5.9%) mAP@0.5 on the RCI dataset and 66.2% (+5.8%) on the DOTA dataset, indicating that this method excels in detecting multi-scale small target cracks and supports intelligent preventive maintenance of roads.

The future article may focus on lightweight crack detection models, such as using quantization to reduce the model complexity for edge device deployment. Additionally, we will explore large-scale road scenarios, such as rainy weather and unstructured roads [8].

Acknowledgment

We sincerely thank the Gan Cai Jiao [2023] No. 36 – Integrated Circuit Industry Research Institute, the Metal Thin Film Pressure Sensor Development and Industrialization Project, and the Artificial Intelligence Chip Development and Application Research Project for their financial support of this research. We also express our deep gratitude to Professor Wang Zaixing for his meticulous guidance during the revision and writing of this paper.

References

[1] H. Yao, Y. Liu, X. Li, Z. You, Y. Feng, and W. Lu, "A detection method for pavement cracks combining object detection and attention mechanism," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22179–22189, Nov. 2022.

[2] F. Li, X. Liu, Y. Yin, and Z. Li, "A novel method for particle instance segmentation and size measurement," *IEEE Trans. Instrum. Meas.*, vol. 73, Jan. 2024, Art. no. 5006817.

[3] J. Chen, N. Zhao, R. Zhang, L. Chen, K. Huang, and Z. Qiu, "Refined crack detection via LECSFormer for autonomous road inspection vehicles," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2049–2061, Mar. 2023.

[4] Y. Xie et al., "Landslide extraction from aerial imagery considering context association characteristics," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 131, 2024, Art. no. 103950.

[5] D.-L. Nguyen, M. D. Putro, and K.-H. Jo, "Lightweight CNN-based driver eye status surveillance for smart vehicles," *IEEE Trans. Ind. Inform.*, vol. 20, no. 3, pp. 3154–3162, Mar. 2024.

[6] J. Zhu, J. Zhang, H. Chen, Y. Xie, H. Gu, and H. Lian, "A cross-view intelligent person search method based on multi-feature constraints," *Int. J. Digit. Earth*, vol. 17, no. 1, 2024, Art. no. 2346259.

[7] H. Jin et al., "Micro-cracks identification and characterization on the sheds of composite insulators by fractal dimension," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1821–1824, Mar. 2021.

[8] W. Xu et al., "Building height extraction from high-resolution single-view remote sensing images using shadow and side information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6514–6528, Mar. 2024.

[9] L. Peng, W. Chao, L. Shuangmiao, and F. Baocai, "Research on crack detection method of airport runway based on twice-threshold segmentation," in *Proc. 5th Int. Conf. Instrum. Meas., Comput., Commun. Control*, 2015, pp. 1716–1720.

[10] X. Dong, D. Li, and J. Fang, "FCCD-SAR: A lightweight SAR ATR algorithm based on FasterNet," *Sensors*, vol. 23, no. 15, 2023, Art. no. 6956.

[11] J. Li, T. Liu, X. Wang, and J. Yu, "Automated asphalt pavement damage rate detection based on optimized GA-CNN," *Automat. Construction*, vol. 136, 2022, Art. no. 104180.

[12] Y. Zhang, C. Wu, T. Zhang, Y. Liu, and Y. Zheng, "Self-attention guidance and multiscale feature fusion-based UAV image object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jan. 2023, Art. no. 6004305.
[13] F. Wang, H. Wang, Z. Qin, and J. Tang, "UAV target detection algorithm based on improved YOLOv8," *IEEE Access*, vol. 11, pp. 116534–116544, 2023.
[14] J. Cui, Y. Qin, Y. Wu, C. Shao, and H. Yang, "Skip connection YOLO architecture for noise barrier defect detection using UAV-based images in high-speed railway," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12180–12195, Nov. 2023.
[15] S. Wang, X. Chen, and Q. Dong, "Detection of asphalt pavement cracks based on vision transformer improved YOLO V5," *J. Transp. Eng.*, vol. 149, no. 2, 2023, Art. no. 04023004.
[16] X. He, Z. Tang, Y. Deng, G. Zhou, Y. Wang, and L. Li, "UAV-based road crack object-detection algorithm," *Automat. Construction*, vol. 154, 2023, Art. no. 105014.
[17] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, 2022.
[18] K. R. Akshatha, A. K. Karunakar, S. B. Shenoy, A. K. Pai, N. H. Nagaraj, and S. S. Rohatgi, "Human detection in aerial thermal images using faster R-CNN and SSD algorithms," *Electronics*, vol. 11, no. 7, 2022, Art. no. 1151.
[19] J.-H. Kim, N. Kim, and C. S. Won, "High-speed drone detection based onYOLO-V8," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.
[20] L. Shen, B. Lang, and Z. Song, "Infrared object detection method based on DBD-YOLOv8," *IEEE Access*, vol. 11, pp. 145853–145868, 2023.
[21] X. Dai et al., "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7369–7378.
[22] G. Zhao et al., "Graph convolution based cross-network multi-scale feature fusion for deep vessel segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 183–195, Jan. 2023.
[23] Yang Li et al., "Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation," *IEEE Trans. Cybern.*, vol. 53, no. 9, pp. 5826–5839, Sep. 2023.
[24] A. Shen, Y. Zhu, P. Angelov, and R. Jiang, "Marine debris detection in satellite surveillance using attention mechanisms," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4320–4330, Jan. 2024.
[25] Y. Liu, F. Liu, W. Liu, and Y. Huang, "Pavement distress detection using street view images captured via action camera," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 738–747, Jan. 2024.
[26] S. Zhu and M. Miao, "SCNet: A lightweight and efficient object detection network for remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, Dec. 2024, Art. no. 6001605.
[27] Y. Luo, J. Xu, C. Feng, and K. Zhang, "An accurate detection algorithm for time backtracked projectile-induced water columns based on the improved YOLO network," *J. Syst. Eng. Electron.*, vol. 34, no. 4, pp. 981–991, 2023.
[28] D. Ouyang et al., "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[29] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.
[30] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.
[31] Y. Zhang and Z. Cai, "CE-RetinaNet: A channel enhancement method for infrared wildlife detection in UAV images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 4104012.
[32] M. Jiang, L. Gu, X. Li, F. Gao, and T. Jiang, "Ship contour extraction from SAR images based on faster R-CNN and Chan–Vese model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5203414.
[33] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, Sep. 2022.
[34] P. Su, H. Han, and M. Liu, "MOD-YOLO: Rethinking the YOLO architecture at the level of feature information and applying it to crack detection," *Expert Syst. Appl.*, vol. 237, 2024, Art. no. 121346.
[35] Q. Qiu and D. Lau, "Real-time detection of cracks in tiled sidewalks using YOLO-based method applied to unmanned aerial vehicle (UAV) images," *Automat. Construction*, vol. 147, 2023, Art. no. 104745.

Chao Dang is currently working toward the master's degree in integrated circuit engineering with the Lanzhou Jiaotong University, Lanzhou, China.

His research interests include computer vision, image processing, FPGA, embedded systems, and machine learning.

Zai Xing Wang received the Ph.D. degree in microelectronics and electronic solids from the Lanzhou University, Lanzhou, China, in 2008.

He is currently an Associate Professor with the Lanzhou Jiaotong University, Lanzhou, China. His research interests include novel device simulation and model, computer software and application, radio electronics, and automation technology.