DS-YOLO-RC: Geometric-Aware Rotated Crack Detection via Over-parameterized Dilated Convolution and Topology-Preserving Alignment

Zhimin Yue[1,2] and Li Li[1,2,*]
[1]Southwest University of Science and Technology, Mianyang 621010, China
[2]Sichuan Engineering Technology Research Center of Industrial Self-Supporting and Artificial Intelligence, Mianyang 621010, China
*Corresponding author: llki2000@163.com

***Abstract*—Standard attention-centric detectors, while powerful, often suffer from inductive bias mismatch when processing slender, high-frequency manifolds like hydraulic cracks. To address this, we propose DS-YOLO-RC, a lightweight rotated object detection framework tailored for geometric-aware feature modeling. By re-introducing morphological priors via Depth-wise Over-parameterized Dilated Convolution (DoDConv) and ensuring spatial fidelity via Dynamic Upsampling (DySample), we fundamentally reconstruct the feature fusion network. Specifically, DoDConv effectively elongates the receptive field along the crack trajectory to capture global topology, while DySample utilizes content-aware inverse reprojection to rectify feature misalignment caused by static interpolation. Extensive experiments on a self-constructed hydraulic concrete dataset and the CRACK500 benchmark demonstrate that our method consistently outperforms state-of-the-art detectors. Compared with the YOLOv12 baseline, DS-YOLO-RC achieves a 2.5% improvement in mAP@50-95 (reaching 45.1%) while reducing computational costs to 5.9 GFLOPs, striking an optimal balance between high-precision localization and real-time inference (96 FPS).**

***Index Terms*—Crack detection, rotated object detection, feature alignment, multi-scale perception, YOLO.**

I. Introduction

Hydraulic concrete structures, such as dams and ship locks, constitute the critical backbone of water conservancy infrastructure. However, under the influence of complex environmental stresses, surface cracks are inevitable. As highlighted by Cai *et al.* [1] and Fan *et al.* [2], cracks are not merely visual indicators of deterioration but also early warnings of potential structural failure. Therefore, utilizing computer vision and Unmanned Aerial Vehicle (UAV) technology to achieve automated, high-precision crack detection has become an urgent industrial imperative [3]. Recently, Feng *et al.* [3] and Han *et al.* [4] explored UAV-based intelligent inspection technologies for earth-fill dams and hydraulic surfaces, respectively, providing crucial data acquisition means for non-contact monitoring.

Early crack characterization mainly relied on image segmentation. Lu *et al.* [5], Wang *et al.* [6], and Harb *et al.* [7] achieved notable progress in feature quantization and complex background segmentation by improving deep learning models. However, despite their precision, pixel-level segmentation models typically incur substantial computational costs, making it difficult to meet the rigorous real-time requirements for large-scale inspection of hydraulic structures.

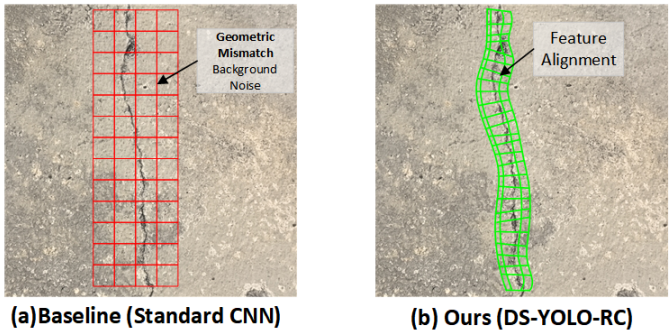

Fig. 1. Comparison of feature extraction mechanisms. **(a) Baseline:** Standard CNNs utilize fixed square receptive fields, leading to geometric mismatch and background noise interference. **(b) Ours (DS-YOLO-RC):** The proposed method employs DoDConv and DySample to achieve pixel-level feature alignment, effectively capturing the meandering geometry of cracks.

To balance accuracy and efficiency, YOLO-based object detection has gradually taken the lead. Li *et al.* [8] and Zhang *et al.* [9] proposed real-time monitoring methods for concrete and pavement cracks based on the YOLO architecture, respectively. However, directly transferring these generic detectors faces a severe "Geometric Mismatch" challenge. As noted by Wang *et al.* [10], for extremely slender objects distributed in arbitrary directions, traditional Horizontal Bounding Boxes (HBB) introduce a large amount of background redundancy, leading to a significantly reduced signal-to-noise ratio. Although Oriented Bounding Box (OBB) technologies (such as the work of Ma *et al.* [11] and Yang *et al.* [12]) have alleviated the orientation issue to some extent, and some studies have begun to focus on multi-scale features and feature alignment [13], [14], existing state-of-the-art detectors—**even the powerful YOLOv12**—are fundamentally limited by their **inductive bias towards generic objects**. They rely heavily on isotropic square convolution kernels and static grid sampling at the underlying level, which are designed for compact targets. This "generalist" design leads to two core defects in crack detection: first, the fixed square receptive field struggles to decouple slender cracks from background noise; second, static

upsampling (such as nearest neighbor interpolation) ignores the geometric pose of the object, causing severe spatial feature misalignment between deep semantic features and shallow localization features.

To address these challenges, we propose **DS-YOLO-RC**, a rotated crack detection framework based on geometric-aware feature representation and alignment. Unlike prior works that focus on stacking layers, we fundamentally reconstruct the neck network of YOLOv12 to mitigate the architectural mismatch in standard Convolutional Neural Networks (CNNs). **The main contributions of this paper are summarized as follows:**

- **Framework Design:** We propose **DS-YOLO-RC**, a tailored rotated object detection framework for hydraulic cracks. We fundamentally reconstruct the feature fusion network by introducing a geometric-semantic dual alignment mechanism, which effectively mitigates the inductive bias mismatch inherent in standard CNNs when processing slender, high-frequency manifolds.
- **Perception Enhancement:** We design a novel **Depth-wise Over-parameterized Dilated Convolution (DoD-Conv)**. By constructing an elastic receptive field via structural re-parameterization, this module enables the network to capture the long-range global topology of cracks along their growth direction, addressing the "Geometric Mismatch" issue.
- **Feature Alignment:** We incorporate the content-aware **Dynamic Upsampling (DySample)** [15] mechanism into the Feature Pyramid Network (FPN) architecture. We strategically deploy it to form a synergy with DoD-Conv: while DoDConv captures global morphology during downsampling, DySample strictly preserves these spatial details during upsampling via inverse reprojection, effectively solving the feature misalignment problem.
- **Performance & Efficiency:** Extensive experiments demonstrate that DS-YOLO-RC achieves state-of-the-art performance (45.1% mAP@50-95) on the self-built hydraulic dataset. Crucially, our geometric-aware design achieves this high precision with reduced computational cost (5.9 GFLOPs), validating its potential for real-time industrial deployment.

Open Science: In accordance with the conference's open science initiative and to facilitate reproducibility, the source code and the self-built hydraulic dataset are publicly available at https://github.com/yue7254-dotcom/DS-YOLO-RC.

II. Related Work

A. Concrete Crack Detection: From Segmentation to Detection

Defect detection in concrete has undergone a paradigm shift from pixel-level segmentation to object detection. Although early segmentation methods (such as CNN CCT Net [16]) achieved high accuracy, high annotation costs and inference latency limited their engineering implementation. Consequently, YOLO-based lightweight detection solutions have gained attention. Zhang *et al.* [9] proposed an enhanced YOLOv8 that achieved high-precision performance in pavement crack detection, while Bai *et al.* [14] further optimized detection performance through a multi-scale information gain algorithm. Furthermore, addressing extreme environments such as underwater scenarios, Ma *et al.* [11] developed a hybrid neural network, demonstrating the generalization potential of deep learning in multi-defect detection. As current state-of-the-art real-time detectors, YOLOv8 [17] and the latest YOLOv12 [18] have refreshed the accuracy-speed boundary by optimizing the Cross Stage Partial (CSP) backbone. However, when directly transferring them to crack detection, the fixed square receptive fields inherent in standard CNNs struggle to adapt to the slender manifold structure of cracks. This "Geometric Mismatch" causes tiny crack features to be easily overwhelmed by background noise.

B. Evolution of Oriented Object Detection (OBB)

Oriented Object Detection (OBB) aims to solve the background redundancy problem of horizontal boxes when processing non-axisymmetric objects. RoI Transformer [19] and Oriented R-CNN [20] established high-precision benchmarks by introducing Rotated Regions of Interest (RoI). Regarding single-stage algorithms, Yang *et al.* [21] proposed R3Det to address the feature misalignment problem, while Xu *et al.* [22] proposed the Gliding Vertex method, providing a new regression perspective for multi-oriented objects. Addressing the uncertainty of angle regression, Llerena *et al.* [23] proposed the Probabilistic IoU (PIoU) loss function, which effectively optimized the fitting quality of bounding boxes. Although the aforementioned methods perform excellently in aerial images (e.g., vehicles, ships), hydraulic cracks possess extreme aspect ratios and arbitrary curvature, making their sensitivity to bounding box regression far higher than that of ordinary rigid objects. A survey by Wang *et al.* [10] indicates that maintaining rotation invariance during the feature extraction stage and overcoming the feature sparsity of slender objects remain core challenges facing OBB technology in the field of crack detection.

C. Feature Alignment and Multi-Scale Perception

Feature alignment and multi-scale perception are key to solving the aforementioned geometric challenges. Regarding feature alignment, Han *et al.* [24] pointed out that the static upsampling of standard FPN leads to spatial misalignment between deep semantics and rotated instances. Although CARAFE proposed by Wang *et al.* [25] alleviated this issue, its computational cost is high. In contrast, DySample proposed by Liu *et al.* [15] abandons the heavy computation of dynamic convolutions, achieving low-cost pixel-level alignment through a content-based reverse re-projection flow. Regarding multi-scale perception, Yu *et al.* [26], [27] established the theoretical foundation for Dilated Convolution to expand the receptive field. Meanwhile, DO-Conv proposed by Cao *et al.* [28] demonstrated the effectiveness of the Over-parameterization strategy in enhancing feature extraction capabilities, achieving a perfect balance between accuracy and efficiency by

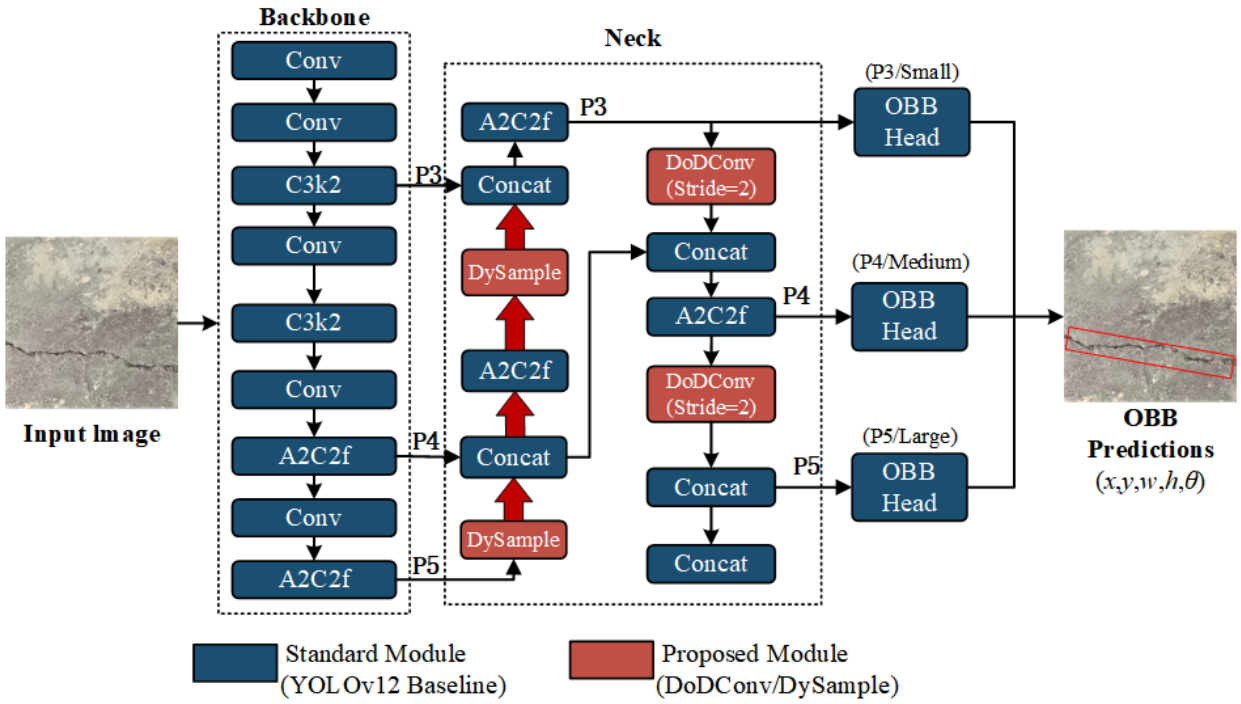

Fig. 2. The overall architecture of the proposed DS-YOLO-RC network. The framework comprises a CSP Backbone, a Geometric Alignment Neck, and a Decoupled OBB Head. Specifically, **DySample** is integrated for content-aware upsampling to preserve crack topology, while **DoDConv** (with a multi-dilation configuration) enhances the receptive field for slender objects. The output OBB is represented by (x, y, w, h, θ) following the *le90* angle convention to ensure stable training on high-aspect-ratio cracks.

constructing a multi-branch structure during training and performing equivalent fusion during inference. Inspired by this, we construct the DoDConv operator, creatively introducing "dynamic dilation rates" into the over-parameterization mechanism. This design aims to combine long-range perception with strong feature extraction capabilities, fundamentally correcting the inductive bias limitations of fixed convolution kernels when processing extreme-scale cracks.

III. Methodology

A. Overall Architecture

To address the challenge of "Geometric Mismatch" in hydraulic concrete crack detection—specifically the contradiction between the extreme aspect ratios of cracks and the fixed square receptive fields of standard CNNs—as well as the inherent Feature Misalignment challenge in rotated detection, we construct a rotated object detection framework based on geometric-semantic dual alignment, named **DS-YOLO-RC**, utilizing the state-of-the-art YOLOv12 as the baseline.

As shown in Fig. 2, while retaining the efficient CSP backbone architecture of YOLOv12, this network performs a deep reconstruction of the Neck network from two dimensions: manifold perception and topological alignment:

Morphological Receptive Field Adaptation In the downsampling path of the Neck, we introduce Depth-wise Over-parameterized Dilated Convolution (**DoDConv**) to replace standard strided convolutions. Unlike the static aggregation of conventional convolutions, DoDConv combines the strong feature extraction capability of Over-parameterization with the long-range perception advantage of Dynamic Dilation. This mechanism empowers the network with a "zoom lens"-like capability, enabling it to adaptively capture long-range dependencies along the crack direction. Consequently, it significantly enhances the simultaneous capture capability for tiny cracks and penetrating long cracks without increasing computational redundancy.

Topology-Preserving Feature Alignment During the upsampling stage of FPN feature fusion, we utilize the Dynamic Upsampling operator (**DySample**) to discard traditional Nearest Neighbor Interpolation. Traditional interpolation ignores the geometric pose of the object, easily leading to spatial decoupling between deep semantic features and shallow localization features. Conversely, DySample generates point sampling offsets with pixel-level precision by learning a content-based Inverse Reprojection Flow. This mechanism essentially performs dynamic rectification on the feature map, ensuring strict spatial alignment of rotated crack features during the multi-scale fusion process, thereby substantially improving the angle regression precision of OBB.

B. Morphological Multi-Scale Feature Fusion based on DoDConv

Although YOLOv12 achieves remarkable efficiency, relying solely on standard strided convolutions limits its ability to model the extreme aspect ratios of cracks. While standard

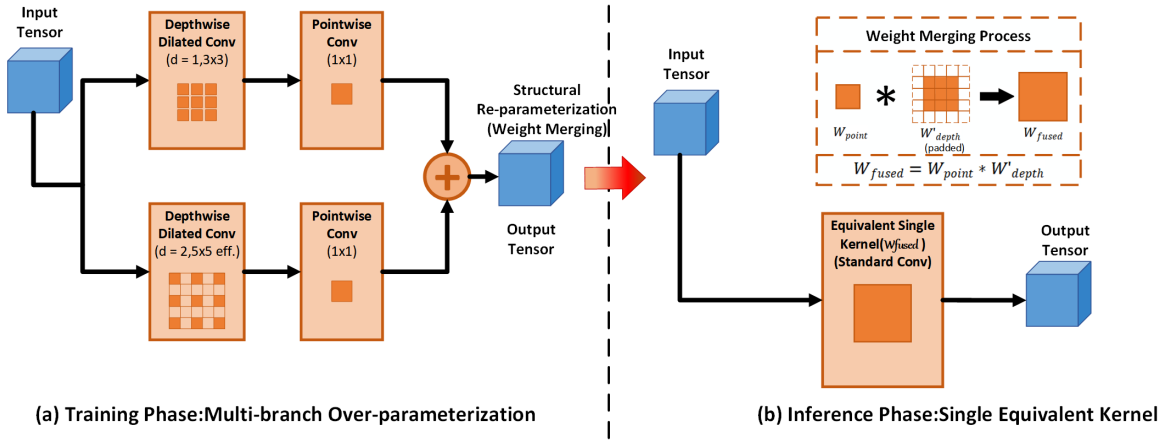

Fig. 3. Schematic diagram of the Structural Re-parameterization mechanism of the DoDConv module. (a) **Training Phase** (Multi-branch Over-parameterization): A multi-branch topology is adopted, where each branch consists of a cascade of Depth-wise Dilated Conv and Point-wise Conv. Branches with different dilation rates (e.g., $d = 1, 2, 4$) work collaboratively to capture multi-scale morphological features of cracks. (b) **Inference Phase** (Structural Re-parameterization): Utilizing the linear property of convolution, multi-branch weights are mathematically equivalently collapsed into a standard single convolution kernel. The top-right inset visually demonstrates the weight merging process, i.e., the Point-wise kernel convolving the padded Depth-wise kernel ($W_{point} * W'_{depth}$). This design achieves the unification of "high-dimensional feature search during training" and "low-dimensional efficient computation during inference."

Dilated Convolutions can theoretically enlarge the receptive field, they suffer from two inherent issues: the *Gridding Effect* (loss of local continuity) and *Optimization Instability* (sparse gradients make it hard to converge).

To resolve this dilemma, we propose **DoDConv** (Depth-wise Over-parameterized Dilated Convolution). Instead of relying on dynamic inference computation, DoDConv introduces a **Multi-branch Re-parameterization** mechanism to decouple the training-time structural complexity from the inference-time computational efficiency.

Training Phase: Multi-Scale Elasticity We construct a multi-branch topology where each branch possesses a distinct fixed dilation rate (e.g., $d = 1, 2, 4$). This design is not intended to dynamically select a branch per input, but to create an **Elastic Receptive Field** during optimization. The formulation in Eq. (1) is redefined as an ensemble of multi-scale morphological priors:

$$Y_{train} = \sum_i (\mathcal{D}_{d_i}(X) * W^{(i)}_{point}) + B \tag{1}$$

where $\mathcal{D}_{d_i}$ represents depth-wise convolution with dilation rate d_i. By combining the dense local features ($d = 1$) with sparse global contexts ($d > 1$), the over-parameterized branches act as a **Learnability Enhancer**. They provide dense gradient flow to stabilize the training of large-dilation branches, effectively filling the semantic gaps caused by the gridding effect.

Inference Phase: Equivalent Fusion Since the multi-branch structure is linear, we apply Structural Re-parameterization to mathematically collapse all branches into a single equivalent kernel W_{fused} before inference:

$$W_{fused} = \sum_i \text{Pad}(W^{(i)}_{depth}, d_i) * W^{(i)}_{point} \tag{2}$$

This ensures that DS-YOLO-RC enjoys the rich feature representation capability of a multi-scale network while maintaining the high inference speed of a single-branch structure. DoDConv is thus a static but highly expressive operator tailored for slender crack manifolds.

C. Dynamic Feature Alignment based on DySample

In the FPN structure of the YOLO series, upsampling operations typically employ Nearest Neighbor Interpolation or Bilinear Interpolation. These static interpolation strategies rely solely on geometric distance, ignoring the semantic content of feature maps. For OBB detection tasks, such coarse upsampling leads to blurred object edges, subsequently causing angular deviations in prediction boxes.

To address this, we **adopt** the content-aware **DySample** module [15] into our framework to achieve dynamic feature alignment. Unlike standard usage, we specifically leverage its point-sampling mechanism to preserve the high-frequency spatial details of rotated cracks during the upsampling process.

Offset Generation DySample does not use a fixed sampling grid; instead, it dynamically predicts sampling positions based on the content of the input feature map. Given an input feature map X, we first generate offsets O and scope modulation factors S via a Linear Projection layer:

$$O = \mathcal{W}_{offset}(X), \quad S = \sigma(\mathcal{W}_{scope}(X)) \tag{3}$$

where $\mathcal{W}_{offset}$ and $\mathcal{W}_{scope}$ denote convolution operations for generating offsets and scope coefficients, respectively, and σ is the Sigmoid activation function. The introduction of S constrains the offset magnitude, preventing sampling points from drifting to irrelevant regions.

Dynamic Grid Resampling To achieve pixel-level feature alignment, we superimpose the generated offsets onto the original static grid $\mathcal{G}$ to construct a dynamic sampling grid $\mathcal{G}'$:

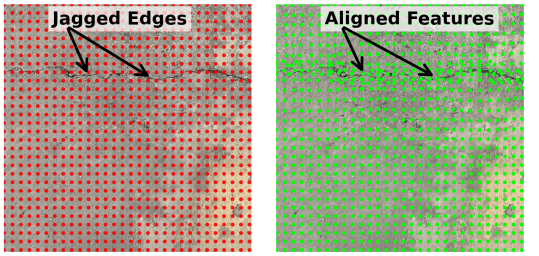

Fig. 4. Microscopic visualization comparison of feature alignment. (a) **Standard Nearest Neighbor Upsampling**: Uses a fixed grid (red dots), causing sampling points to fall into the background regions, leading to feature misalignment. (b) **DySample (Ours)**: Drives sampling points (green dots) to shift towards the semantic boundaries of cracks via inverse reprojection flow. This dynamic alignment ensures feature purity for oriented bounding box regression.

$$\mathcal{G}' = \mathcal{G} + 0.5 \cdot O \odot S \tag{4}$$

where $\odot$ denotes element-wise multiplication (Hadamard Product), and the coefficient 0.5 is used to normalize the sampling step size. Through this step, the regular grid points undergo non-rigid deformation, shifting towards the geometric boundaries of the cracks. The final upsampling output is obtained via the bilinear resampling function:

$$Y = \text{GridSample}(X, \mathcal{G}') \tag{5}$$

DySample utilizes this learnable deformation mechanism to enable the upsampled feature maps to adaptively fit the geometric boundaries of rotated cracks (as shown in Fig. 4), thereby effectively resolving the feature misalignment issue and improving the regression precision of the detection head for angular parameters.

D. Optimization Objective

To ensure precise convergence during training, specifically addressing the sensitivity of high-aspect-ratio cracks to angular deviations, we formulate the total loss function L_{total} as a weighted sum of three components: classification loss, bounding box regression loss, and distribution focal loss. The total objective is defined as:

$$L_{total} = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{dfl} L_{dfl} \tag{6}$$

where λ_{box}, λ_{cls}, and λ_{dfl} are hyperparameters balancing the tasks. In our experiments, we empirically set $\lambda_{box} = 7.5$, $\lambda_{cls} = 0.5$, and $\lambda_{dfl} = 1.5$ to ensure balanced gradient flow.

Classification and Distribution Loss For the classification task (L_{cls}), we employ the Binary Cross-Entropy (BCE) loss to handle class probabilities. For the distribution focal loss (L_{dfl}), we utilize the vector formulation to refine the regression of the bounding box boundaries, enabling the network to focus on the probability distribution near the ground truth labels.

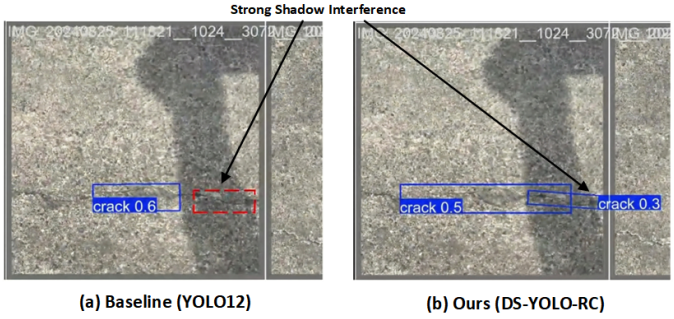

Fig. 5. **Qualitative comparison under strong shadow interference.** (a) The baseline YOLOv12 suffers from *Missed Detection* within the shadowed region (indicated by the red dashed box). (b) Our DS-YOLO-RC successfully localizes the crack target under the same illumination conditions, demonstrating enhanced environmental robustness.

Rotated Bounding Box Regression (L_{box}) Standard Intersection over Union (IoU) based losses (e.g., CIoU) exhibit significant limitations when dealing with extremely slender cracks. A slight angular deviation in a high-aspect-ratio object causes a drastic drop in IoU, leading to gradient instability. To address this, instead of treating angle regression as a simple independent variable ($\lambda_{angle} L_{angle}$), we adopt the **ProbIoU (Probabilistic IoU)** loss strategy. We model the rotated bounding box (x, y, w, h, θ) as a 2-D Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. The similarity between the predicted box P and the ground truth G is measured using the Bhattacharyya Distance (BD):

$$BD = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln \left(\frac{|\Sigma|}{\sqrt{|\Sigma_1||\Sigma_2|}} \right) \tag{7}$$

where $\Sigma = (\Sigma_1 + \Sigma_2)/2$. The final regression loss is calculated as:

$$L_{box} = 1 - \exp(-C \cdot BD) \tag{8}$$

where C is a scaling constant. By converting the geometric rotation problem into a probability distribution matching problem, ProbIoU effectively overcomes the discontinuity in angle regression and provides smoother gradients for slender crack targets.

IV. Experiments

A. Experimental Settings

To fully verify the effectiveness and robustness of the proposed method, we conducted extensive experiments on a self-constructed hydraulic concrete crack dataset containing 4,748 high-resolution images (partitioned at a ratio of 7:2:1) and the public CRACK500 benchmark containing 500 complex pavement images. For CRACK500, we converted its original mask annotations into rotated bounding boxes using the Minimum Area Rectangle algorithm to verify cross-domain generalization.

All experiments were implemented on the PyTorch framework and executed in a resource-constrained hardware environment (NVIDIA GeForce GTX 1660 SUPER, 6GB RAM;

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS. (ALL MODELS ARE TRAINED FROM SCRATCH WITHOUT PRE-TRAINED WEIGHTS.)

Method	Backbone	Strategy	Params (M)	FLOPs (G)	mAP@50 (%)	mAP@50-95 (%)	FPS
YOLOv5n-OBB [29]	CSPDarknet	Scratch	2.25	6.0	74.2	36.9	148
YOLOv5s-OBB [29]	CSPDarknet	Scratch	8.10	19.7	74.5	37.0	64
YOLOv8n-OBB [30]	CSPDarknet	Scratch	2.76	7.1	77.2	39.3	115
YOLOv11n-OBB [31]	YOLO11	Scratch	2.66	6.5	69.4	30.3	138
YOLOv12n-OBB [32]	YOLO12	Scratch	2.58	6.1	79.0	42.6	99
DS-YOLO-RC (Ours)	**YOLO12**	**Scratch**	**2.97**	**5.9**	**80.5**	**45.1**	**96**

TABLE II
GENERALIZATION PERFORMANCE EVALUATION ON DEEPCRACK AND CRACK500 DATASETS

Method	Backbone	DeepCrack mAP	CRACK500 mAP	FPS	FLOPs (G)
YOLOv8n-OBB [30]	CSPDarknet	72.4	63.4	115	7.1
YOLOv12n-OBB [32]	Attention	75.1	64.0	99	6.1
DS-YOLO-RC (Ours)	**Attention**	**78.2**	**65.1**	96	**5.9**

TABLE III
ABLATION STUDY ON THE EFFECTIVENESS OF EACH MODULE

Exp.	Baseline (YOLOv12)	DoDConv	DySample	mAP@50 (%)	mAP@50-95 (%)
1	✓			79.0	42.6
2	✓	✓		79.3	43.2
3	✓		✓	79.8	44.5
4	✓	✓	✓	**80.5**	**45.1**

Intel Core i5-9400F CPU). This setup was intentionally selected to simulate real-world engineering deployment on cost-effective edge devices. Although the baseline YOLOv12-n is capable of achieving over 600 FPS on high-end data-center GPUs (e.g., Tesla T4), our priority is accessibility in industrial scenarios. Despite the additional structural complexity introduced by DoDConv and DySample, the proposed DS-YOLO-RC maintains a real-time inference speed of 96 FPS on this entry-level hardware, demonstrating its high practicality for edge computing. To eliminate inductive bias from pre-trained weights, all models adopted the "Train from Scratch" strategy. The input image size was set to 640×640, batch size to 8, and total training epochs to 300. The SGD optimizer was employed (momentum 0.937, weight decay 0.0005) with an initial learning rate of 0.01 under Cosine Annealing decay.

B. Comprehensive Performance Comparison

To rigorously validate the performance boundaries of DS-YOLO-RC, we compared it with mainstream rotated detectors (Table I). Based on the quantitative results, we draw the following conclusions:

Efficiency-Accuracy Trade-off. A pivotal advantage of DS-YOLO-RC is its ability to boost precision without compromising computational efficiency. As evidenced in Table I, compared to the baseline YOLOv12n-OBB, our model improves mAP@50-95 by 2.5% (42.6% $\rightarrow$ 45.1%) while achieving a superior mAP@50 of 80.5%. Crucially, despite a marginal increase in parameters (2.58 M $\rightarrow$ 2.97 M) due to the multi-branch training topology, the floating-point operations (FLOPs) actually decrease from 6.1 G to 5.9 G. This indicates that the proposed geometric-aware design utilizes computational resources more efficiently. Specifically, the **DoDConv** module utilizes structural re-parameterization to collapse complex training branches into a single equivalent kernel during inference. Thus, it incurs **no additional inference latency compared to standard convolutional layers**, effectively decoupling training representational power from inference speed. Meanwhile, the **DySample** module employs a lightweight point-sampling mechanism, avoiding the heavy computational burden often associated with learnable upsamplers. Consequently, our model maintains a high inference speed of 96 FPS on the GTX 1660 SUPER. This is comparable to the baseline (99 FPS) and far exceeds the real-time threshold (30 FPS) required for industrial inspection, validating its practicality for edge deployment.

Comparison with SOTA Models. To further position our work, we analyze the dominance over previous generations. When compared with widely adopted lightweight detectors, DS-YOLO-RC demonstrates a significant performance advantage. For instance, YOLOv8n-OBB requires higher FLOPs (7.1 G) but trails our method by 5.8% in mAP@50-95. Furthermore, compared to YOLOv11n-OBB (30.3% mAP@50-95), our method achieves a remarkable improvement of 14.8% in high-precision localization.

C. Generalization Analysis

To verify robustness under domain shift, we evaluated the models on two public benchmarks: **DeepCrack** and **CRACK500** (asphalt pavement), which differ significantly from the training domain (hydraulic concrete). As shown in Table II, despite utilizing the identical backbone, DS-YOLO-RC demonstrates superior generalization capabilities. Specifically, our method outperforms the YOLOv12 baseline by **3.1%** on the DeepCrack dataset ($75.1\% \rightarrow 78.2\%$) and **1.1%** on the CRACK500 dataset ($64.0\% \rightarrow 65.1\%$). This consistent improvement indicates that the morphological features extracted by the DoDConv module are not overfitting to the specific background texture of hydraulic structures, but rather capture the intrinsic geometric topology of cracks across different material surfaces.

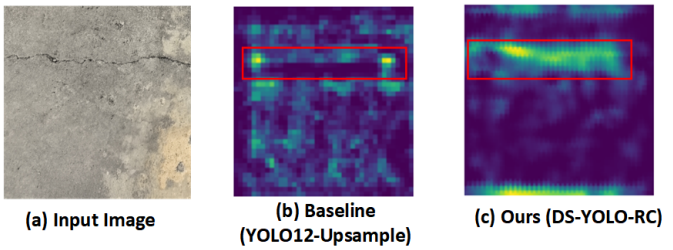

Fig. 6. **Visualization comparison of feature response integrity.** (a) Original input image containing tiny cracks. (b) Feature map of the baseline YOLOv12-OBB. Due to static interpolation, the response exhibits severe fragmentation, activating only the crack endpoints. (c) Feature map of our DS-YOLO-RC. With the proposed DySample module, the model successfully reconstructs the continuous topological structure of the crack and suppresses background noise.

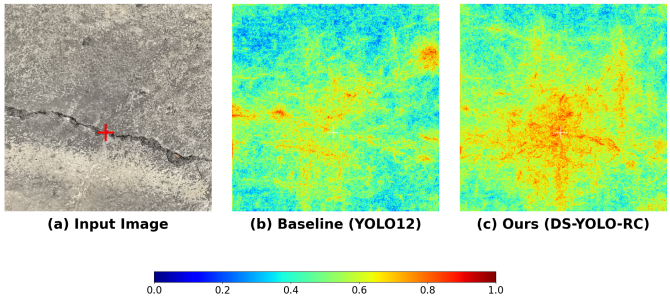

Fig. 7. **Visualization of Effective Receptive Field (ERF).** The red cross in (a) indicates the target pixel on the crack. (b) The baseline model (YOLOv12) exhibits a scattered and isotropic response, failing to focus on the defect structure. (c) Our DS-YOLO-RC demonstrates a highly concentrated receptive field that adaptively aligns with the crack's topology, verifying the morphological perception capability of the proposed DoDConv module.

D. Ablation Study Analysis

To quantify the specific contributions of each module, we designed stepwise ablation experiments under identical conditions (Table III).

Multi-scale Perception Role of DoDConv. Comparing Exp. 1 and Exp. 2, replacing the neck downsampling layers with DoDConv resulted in a steady improvement in mAP@50 (79.0% $\rightarrow$ 79.3%) and mAP@50-95 (42.6% $\rightarrow$ 43.2%). This confirms that the dynamic receptive field mechanism effectively captures the scale variations of cracks, reducing the False Negative rate.

Feature Alignment Role of DySample. Moreover, integrating DySample individually (Exp. 3) yielded a notable gain, boosting mAP@50-95 to 44.5%. Finally, combining both modules (Exp. 4) achieved the SOTA performance with an mAP@50 of 80.5% and mAP@50-95 of 45.1%. Compared to the baseline (Exp. 1), our method achieves a comprehensive improvement of 2.5% in high-precision localization (42.6% $\rightarrow$ 45.1%).

E. Qualitative Visualization Analysis

To intuitively verify the effectiveness of feature reconstruction, we visualized the intermediate feature maps. As shown in Fig. 6(b), the baseline model exhibits *Feature Fragmentation*, where activation exists only at crack endpoints due to the limited receptive field. In contrast, DS-YOLO-RC (Fig. 6(c)) successfully achieves *Topological Reconstruction*. Benefiting from the content-aware sampling of DySample, the model reconnects broken segments into a continuous structure and suppresses background noise, visually explaining the quantitative gains in localization accuracy.

Mechanism Verification via ERF. To further investigate the theoretical cause of this improvement, we visualized the Effective Receptive Field (ERF), as shown in Fig. 7. In standard CNNs (Fig. 7(b)), the receptive field presents a Gaussian-like isotropic distribution that is agnostic to object geometry, leading to significant background noise interference. In contrast, thanks to the dynamic dilation mechanism of DoDConv, our model (Fig. 7(c)) exhibits a distinct *morphological awareness*: the high-response region is adaptively elongated along the crack's trajectory. This alignment ensures that the network concentrates its attention on the slender defect rather than the surrounding pavement, providing a solid theoretical explanation for the mAP gains reported in Table I.

V. Conclusion

To address the issues of feature fragmentation and geometric mismatch in hydraulic crack detection, we propose DS-YOLO-RC. By introducing DoDConv to realize morphological adaptation of receptive fields and utilizing DySample to reconstruct the spatial topology of cracks, the model effectively enhances the feature capture capability for slender rotated objects. Experimental results demonstrate that while achieving a superior **80.5% mAP@50** (surpassing the strong baseline), the proposed method elevates the high-precision localization metric mAP@50-95 to **45.1% (+2.5%)**. Furthermore, it achieves a real-time inference speed of **96 FPS** with only 2.97 M parameters, achieving a perfect balance between detection accuracy and computational efficiency.

Future work will focus on enhancing the model's environmental adaptability and edge deployment capabilities. On the one hand, we aim to explore multi-modal fusion of visible light and infrared/sonar data to tackle detection challenges in extreme low-light or turbid underwater environments. On the other hand, we will integrate model pruning and quantization techniques to further compress the model size, making it adaptable to compute-constrained micro-embedded platforms.

Acknowledgment

This research was partially supported by the Key Program of the National Natural Science Foundation of China under Grant No. U21A20157.

References

[1] M. Cai, H. Wang, X. Lv, Z. Liu, and Y. Chen, "Deep learning-based defect identification in hydraulic structures: A comprehensive review," *KSCE Journal of Civil Engineering*, p. 100410, 2025.

[2] C. Fan, Y. Ding, X. Liu, and K. Yang, "A review of crack research in concrete structures based on data-driven and intelligent algorithms," *Structures*, vol. 75, p. 108800, 2025.

[3] W. Feng, S. Cao, L. Fang, W. Du, and S. Ma, "Crack detection and displacement measurement of earth-fill dams based on computer vision and deep learning," *Sustainability*, vol. 17, no. 22, p. 10186, 2025.

[4] F. Han and C. Gu, "Surface damage detection in hydraulic structures from uav images using lightweight neural networks," *Remote Sensing*, vol. 17, no. 15, p. 2668, 2025.

[5] X. Lu, Q. Li, J. Li, and L. Zhang, "Deep learning-based method for detection and feature quantification of microscopic cracks on the surface of concrete dams," *Measurement*, vol. 240, p. 115587, 2025.

[6] L. Wang and J. Li, "Concrete dam apparent crack detection algorithm based on improved deeplabv3+ model," in *2024 International Academic Conference on Edge Computing, Parallel and Distributed Computing*, 2024, pp. 272–276.

[7] S. Harb, P. Achanccaray, M. Maboudi, and M. Gerke, "Multi-temporal crack segmentation in concrete structure using deep learning approaches," *arXiv preprint arXiv:2411.04620*, 2024.

[8] G. Li, Y. Yang, Y. Wen, and J. Li, "Cgsw-yolo enhanced yolo architecture for automated crack detection in concrete structures," *Symmetry*, vol. 17, no. 6, p. 890, 2025.

[9] Z. Zhang, H. Zhang, and T. Zhang, "Enhanced yolov8-based pavement crack detection: A high-precision approach," *PLoS One*, vol. 20, no. 5, p. e0324512, 2025.

[10] K. Wang *et al.*, "Oriented object detection in optical remote sensing images using deep learning: a survey," *Artificial Intelligence Review*, vol. 58, no. 11, p. 350, 2025.

[11] C. Ma, Z. Chen, H. Wang, and G. Shen, "Underwater structural multi-defects automatic detection via hybrid neural network," *Journal of Marine Science and Engineering*, vol. 13, no. 11, p. 2029, 2025.

[12] X. Yang, E. del Rey Castillo, Y. Zou, L. Wotherspoon, J. Yang, and H. Li, "Automated concrete bridge damage detection using an efficient vision transformer-enhanced anchor-free yolo," *Engineering*, 2025.

[13] Z.-H. Zhu, W. Lu, S.-B. Chen, C. H. Ding, J. Tang, and B. Luo, "Real-world remote sensing image dehazing: Benchmark and baseline," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[14] Y. Bai, Z. Li, R. Liu, J. Feng, and B. Li, "Crack-detection algorithm integrating multi-scale information gain with global–local tight–loose coupling," *Entropy*, vol. 27, no. 2, p. 165, 2025.

[15] W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to upsample by learning to sample," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6027–6037.

[16] H. Ling and F. Sun, "Cct net: A dam surface crack segmentation model based on cnn and transformer," *Infrastructures*, vol. 10, no. 9, p. 240, 2025.

[17] Q. Yao, D. Zhuang, Y. Feng, Y. Wang, and J. Liu, "Accurate detection of brain tumor lesions from medical images based on improved yolov8 algorithm," *IEEE Access*, 2024.

[18] U. Aymon, N. S. Kamarudin, and A. F. A. Nasir, "Facial expression recognition with yolov11 and yolov12: A comparative study," *arXiv preprint arXiv:2511.10940*, 2025.

[19] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[20] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3520–3529.

[21] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3163–3171.

[22] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1452–1459, 2020.

[23] J. M. Llerena, L. F. Zeni, L. N. Kristen, and C. Jung, "Gaussian bounding boxes and probabilistic intersection-over-union for object detection," *arXiv preprint arXiv:2106.06072*, 2021.

[24] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[25] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "Carafe: Content-aware reassembly of features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3007–3016.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[27] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 472–480.

[28] J. Cao *et al.*, "Do-conv: Depthwise over-parameterized convolutional layer," *IEEE Transactions on Image Processing*, vol. 31, pp. 3726–3736, 2022.

[29] G. Jocher *et al.*, "YOLOv5 by Ultralytics," 2020, accessed: 2025-01-21. [Online]. Available: https://github.com/ultralytics/yolov5

[30] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023, accessed: 2025-01-21. [Online]. Available: https://github.com/ultralytics/ultralytics

[31] ——, "Ultralytics YOLOv11," 2024, accessed: 2025-01-21. [Online]. Available: https://github.com/ultralytics/ultralytics

[32] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025. [Online]. Available: https://arxiv.org/abs/2502.12524