Research on Road Defect Detection Based on Deep Learning

Jinchi Zhao, Chunna Zhang*, Shengqiang Cong, Yang Yu, Xiaoping Yue, Yuming Shen, and Yinglun Hui

Abstract—**It aims to address the problems of low accuracy of existing road defect detection algorithms and false detection and missed detection of small cracks. It proposes a road detection algorithm, YOLOv10n-DSLD, that improves the YOLOv10n model. The algorithm uses Large Separable Kernet Attention (LSKA) in the backbone network to improve Spatial Pyramid Pooling Fast (SPPF) to enhance the ability of multi-scale feature extraction. Since the size and aspect ratio of cracks vary greatly in complex backgrounds, Dynamic Snake Convolution (DSC) is used to improve the sensitivity of identifying the shapes and boundaries of small road cracks by adaptively adjusting the shape and size of the convolution kernel. Finally, Dynamic Sample (DySample) is introduced to improve UpSample in YOLOv10 to enhance the feature fusion capability of the neck network and reduce the computational complexity of the model without reducing the detection accuracy. Experiments show that the YOLOv10n-DSLD model has achieved remarkable results. The improved model has improved mAP50 and mAP50-95 by 2.5% and 1.1%. In addition, the improved model has higher detection accuracy than other advanced object detection models, including YOLOv5, YOLOv7, YOLOv8, YOLOv10 series, YOLOv11, and Faster R-CNN. It has demonstrated a high detection capability, and the ablation experiment further shows that the improved model is effective in improving the overall performance.**

Index Terms—**Deep Learning; Road Detection; YOLOv10; DSC**

I. Introduction

WITH the rapid advancement of urbanization and the continuous increase in vehicle numbers, urban road infrastructures are under mounting pressure. Traditional maintenance methods often fail to detect defects such as cracks and potholes in a timely manner, resulting in frequent traffic accidents and escalating maintenance costs. In this context, achieving efficient, accurate, and real-time road defect detection has become a pressing need for modern traffic management. Deep learning-based object detection techniques have increasingly supplanted conventional approaches, delivering significant gains in detection accuracy [1]. Nevertheless, their heavy reliance on computational resources poses challenges to their deployment in certain real-time and resource-constrained scenarios [2].

Current deep learning target detection methods are mainly divided into two categories: two-stage (such as Faster R-CNN) and single-stage (such as YOLO series) [3]. The former has high accuracy but slow speed, while the latter has stronger real-time performance.The YOLOv10 version still has the problem of insufficient feature extraction in the detection of small road defects.To improve performance, Zhang et al. introduced the Swin Transformer module to enhance the detection of small targets [4]; Ao et al. proposed the dynamic acceleration network DyFasterNet, which improves feature extraction capabilities through multi-core adaptive convolution and improves average accuracy by 1.9%. However, DyFasterNet has additional computational overhead due to its complex structure.Although the abovement ioned improvement methods have improved the detection accuracy, they have also brought about a significant increase in parameters, which is not conducive to the lightweight [5]. In order to balance accuracy and efficiency, Lu et al. introduced lightweight modules GSConv and VoV-GSCSP to reduce computational costs while maintaining good feature extraction capabilities [6]. Jin et al. combined AKConv and VanillaNest to propose AKVanillaNet, which effectively captures target shape features and significantly reduces the number of parameters [7]. Wu et al. designed a lightweight multi-scale detection model GAS-YOLO, which integrated the GSF-ST architecture, improved BiFPN and Swin Transformer, improved the small target detection capability, introduced Wiou loss to optimize the sample imbalance problem, and improved the detection accuracy by 10.8% [8].

In the road detection model,Liu et al. proposed a fast improved road detection model MMS-YOLOv10 based on YOLOv10, adding the multi-co-attention (MCA) mechanism to the C2f module to enhance the adaptability of objects of different scales [9]. Secondly, a multi-level feature fusion (MFF) module is designed to enhance semantic and detail information and improve the feature expression ability with different levels of features. Finally, a sample-related weighted loss function is introduced in the network training process to solve the sample imbalance problem.Although existing road detection methods have achieved good results in terms of accuracy, they are still prone to missed detection or false detection in complex environments such as occlusion, lighting changes, and bad weather, affecting the stability and practicality of the model. To address these challenges, an enhanced YOLOv10-DSLD model is proposed, which introduces advanced network structure and optimization strategies to improve detection accuracy and generalization ability,

Manuscript received April 29, 2025; revised July 10, 2025.

Jinchi Zhao is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, Liaoning China (e-mail: 232085400142@stu.ustl.edu.cn).

Chunna Zhang* is an Associate Professor at University of Science and Technology Liaoning, Anshan 114051, Liaoning China (Corresponding author to provide phone: +86-130-1961-7131; e-mail: lkdzcn@ustl.edu.cn).

Shengqiang Cong is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoning China (e-mail: 222085400551@stu.ustl.edu.cn).

Yang Yu is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoning China (e-mail: 222085400542@stu.ustl.edu.cn).

Xiaoping Yue is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoning China (e-mail: 222085400571@stu.ustl.edu.cn).

Yuming Shen is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoning China (e-mail: 232085400119@stu.ustl.edu.cn).

Yinglun Hui is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoning China (e-mail: 243208540506@stu.ustl.edu.cn).

and enhance its robustness in complex scenarios. This study mainly made three improvements:

1) To solve the problem that the standard Large Separable Attention(LSA) relies too much on texture features and ignores the global shape of objects, LSKA is introduced.

2) In order to enhance the sensitivity of feature extraction to small cracks in slender and tortuous local structures, DSC is introduced in the convolutional layer of the backbone network.

3) To enhance the detection capability of small targets and complex scenes and achieve a better balance between accuracy and speed, the DySample upsampler is used to reduce computational and resource overhead while maintaining performance.

The road defect detection technology based on YOLOv10n-DSLD has significantly advanced infrastructure modernization. This method offers the advantages of high accuracy and lightweight deployment, enabling effective identification of defects such as cracks and potholes. Consequently, it enhances road management efficiency, reduces maintenance costs, and prolongs the service life of road infrastructure. This technology has achieved the transformation from manual inspection to data-driven intelligent management, promoted precise road maintenance and sustainable development, and become a key supporting tool in the intelligent transportation system, helping the automation and modernization of the transportation system [10].

II. Related works

A. YOLOv10 Model

YOLO is one of the most popular real-time target detection algorithms. YOLOv10 has been lightly optimized in the backbone network, using Depthwise Separable Convolutions and Bottleneck Layers, which significantly reduces the amount of computation and memory consumption.YOLOv10 introduced Path Aggregation Network (PAN) and Efficient Channel Attention (ECA), which improved the small object detection accuracy and feature fusion capabilities.In terms of the backbone network, YOLOv10 balances the accuracy and computational efficiency of the model through lightweight design and the introduction of a new feature extraction module, and performs particularly well in small object detection and low-resource environments [11]. YOLOv10 has five different models, each with different parameter sizes to meet various application requirements. For the actual needs of road defect detection, the YOLOv10n model achieves a good balance between detection speed and accuracy. On this basis, this paper proposes an improved model YOLOv10n-DSLD as shown in Fig. 2.

B. LSKA

Traditional LKA performs well in visual tasks, but its computational and memory overhead caused by the increase of convolution kernel size limits its application.To solve this problem, LSKA is proposed, which effectively reduces computational complexity and memory consumption by decomposing large kernel convolution into horizontal and vertical one-dimensional convolutions. Experiments show that LSKA maintains comparable performance to LKA in Visual Attention Network (VAN), while significantly optimizing resource usage and enhancing the model's ability to focus on object shapes.

The working principle of LSKA is shown in Fig. 2 LSKA decomposes a large kernel into two depth-wise convolutions to obtain a wide receptive field with different features, which helps in the subsequent selection of kernels of different sizes. Moreover, decomposing a large kernel can effectively reduce the number of parameters in the model compared to using a single kernel.In addition, LSKA can dynamically select the appropriate kernel according to the characteristics of the input target, thereby adapting to the contextual information of different target objects.In order to dynamically select the appropriate kernel [12], LSKA divides the input features into different sub-feature maps, and then applies different depth convolutions to these sub-feature maps to obtain different output feature maps. Then, the features obtained from different kernels are cascaded with different receptive field ranges:

$$\tilde{U} = \left[\tilde{U}_1; \ldots; \tilde{U}_i\right] \tag{1}$$

Under this operation,denotes the features extracted by different kernels, capturing spatial relationships through a combination of channel-wise average pooling and max pooling operations [13]:

$$SA_{\text{avg}} = P_{\text{avg}}(\tilde{U}), \quad SA_{\text{max}} = P_{\text{max}}(\tilde{U}) \tag{2}$$

Then, after SA connection, the features are converted into N spatial attention maps through convolution kernels and sigmoid processing to ensure that they have the same number of deep convolutions [14]:

$$\widetilde{SA} = F^{2\to N}([SA_{\text{cvg}}; SA_{\text{max}}]) \tag{3}$$

The final N spatial attention maps are passed through a sigmoid activation function to obtain different spatial selection weights, and the final output is the element-wise product between the input features X and S.

$$\widetilde{SA}_i = \sigma(\widetilde{SA}_i) \tag{4}$$

$$S = F\left(\sum_{i=1}^{N}(\widetilde{SA}_i \cdot \tilde{U}_i)\right) \tag{5}$$

$$Y = X \cdot S \tag{6}$$

C. DSC

Tubular structures are typically elongated and tortuous, posing challenges for conventional convolutional methods in capturing their intricate local features. Deep learning-based image processing approaches like U-Net often fall short in detecting fine-grained structures, and their loss functions do not explicitly enforce topological constraints, leading to inadequate preservation of topological continuity.A method called DSC is proposed, which can adaptively focus on elongated and tortuous local structures. DSCN effectively addresses the challenges of capturing local features, preserving global morphological integrity, and maintaining topological continuity in tubular structure segmentation by innovatively integrating dynamic convolution, feature fusion, and topological constraint techniques. Extensive evaluations demonstrate that

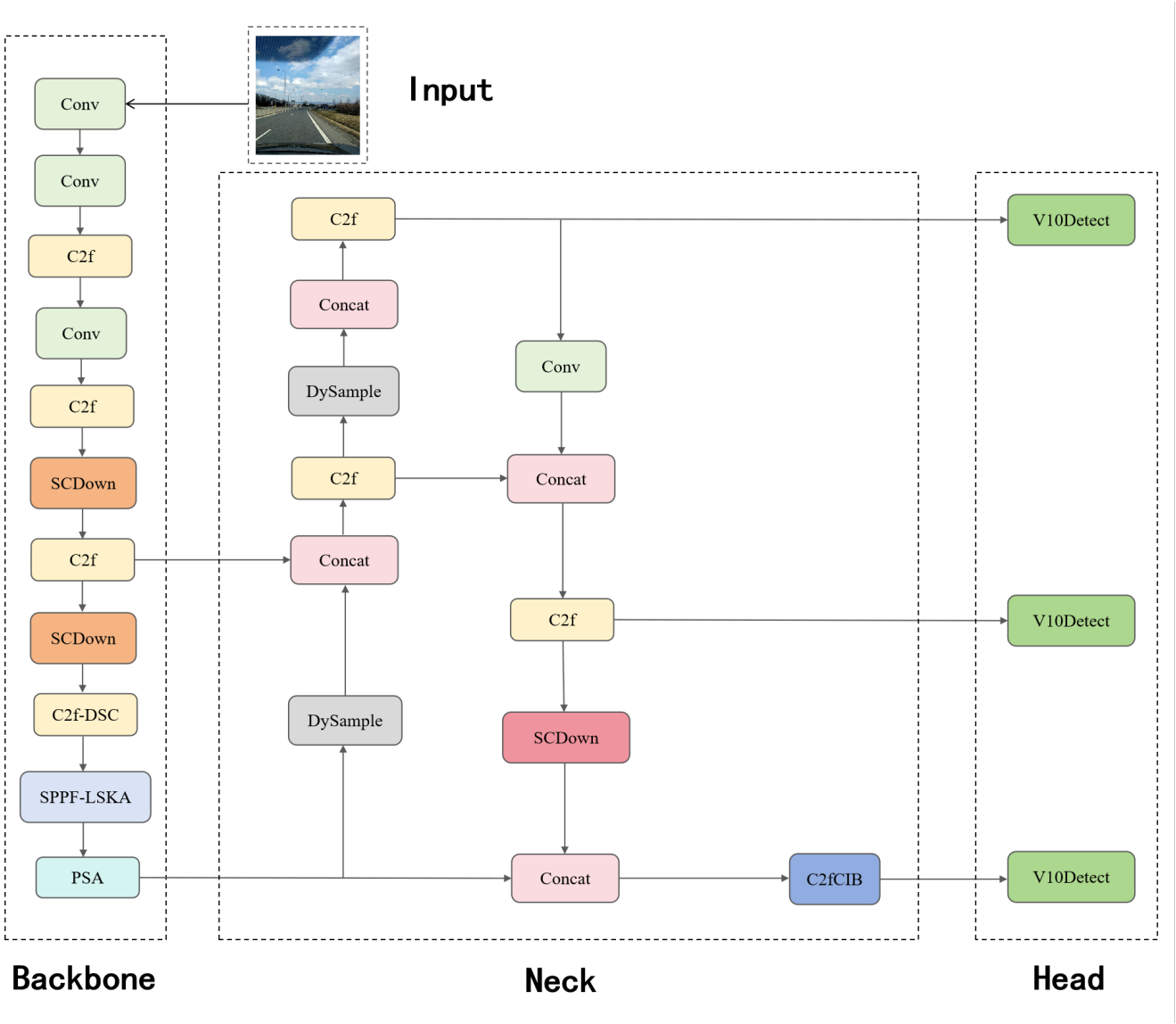


Fig. 1. YOLOv10n-DSLD model

this method achieves outstanding performance on both 2D and 3D datasets, providing a more accurate and continuous segmentation framework for tubular structures, significantly improving segmentation accuracy and reliability [15].

D. DySample

Conventional dynamic upsamplers (such as CARAFE, FADE, and SAPA) typically rely on dynamic convolutions and additional sub-networks to generate dynamic kernels, which incurs significant computational overhead in terms of parameters, FLOPs, and GPU memory consumption.These methods typically rely on high-resolution input features, which restrict their applicability in various scenarios. DySample overcomes the complexity of dynamic convolutions by employing a point-sampling approach for upsampling, effectively avoiding the high computational cost associated with dynamic convolutions and significantly reducing resource consumption.DySample does not require custom CUDA packages, significantly reducing the number of parameters, FLOPs, GPU memory usage, and latency. Compared to traditional kernel-based dynamic upsamplers, it offers clear advantages. Therefore, we integrate DySample into our YOLOv10n-DSLD network, focusing on upsampling low-resolution images to higher resolutions with minimal overhead [16].

III. An Improved Road Defect Detection Model Based on YOLOv10

To address the issues of low accuracy, false positives, and missed detections of fine cracks in existing YOLOv10n-based road defect detection algorithms, this paper proposes the YOLOv10n-DSLD model. The proposed approach introduces several technical enhancements to improve algorithm performance, effectively overcoming limitations in feature representation and extraction, and enhancing the accuracy of small-object detection in complex scenarios.First, we integrate the LSKA module into the SPPF pooling layer of the YOLOv10n backbone, which facilitates the aggregation of both local and global features after multi-scale pooling operations. Second, a dynamic deformable convolution module is incorporated into the C2f block of the backbone, significantly enhancing the efficiency and accuracy of feature extraction, particularly for fine details and multi-scale features in complex images. Finally, we adopt a lightweight DySample

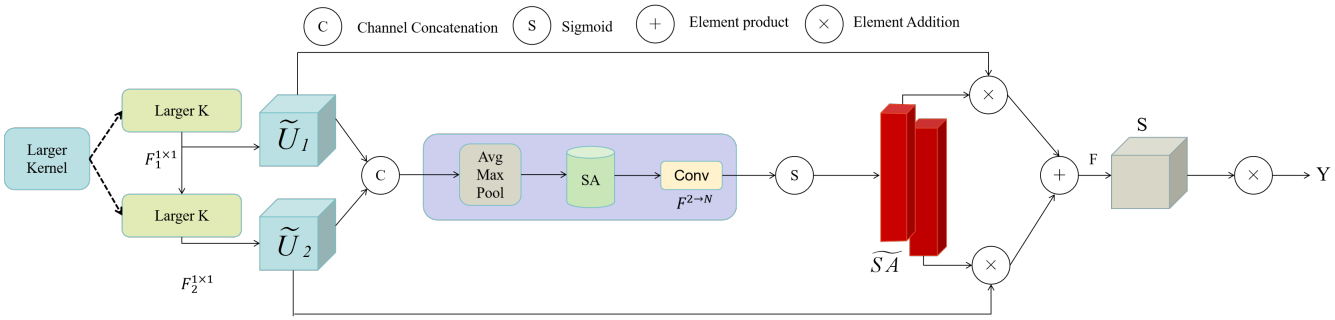


Fig. 2. LSKA working principle diagram

upsampler, which replaces traditional interpolation with a point sampling approach, improving object detection accuracy while reducing computational overhead.

A. SPPF-LSKA

In YOLOv10, the introduced SPPF module significantly enhances the network's capability in detecting small and multi-scale objects while maintaining high computational efficiency. By employing three consecutive 5×5 max pooling operations followed by feature concatenation, SPPF expands the receptive field and accelerates feature extraction. Compared to the traditional SPP, SPPF reduces both computational cost and memory usage, making it particularly suitable for efficient object detection. The LSKA module, on the other hand, reduces computational complexity and memory overhead by decomposing 2D convolution kernels into 1D horizontal and vertical kernels. It further enhances feature extraction through an integrated attention mechanism [17]. The combination of LSKA and SPPF enables YOLOv10 to perform more efficient and accurate multi-scale feature extraction, especially when dealing with large-scale datasets and complex visual tasks.

The SPPF-LSKA module is constructed by augmenting the original SPPF with the LSKA.The architecture of the module is illustrated in Fig. 3: A schematic diagram of the LSKA module is shown in Fig. 4 For a given feature map F, the operation can be formulated as follows:

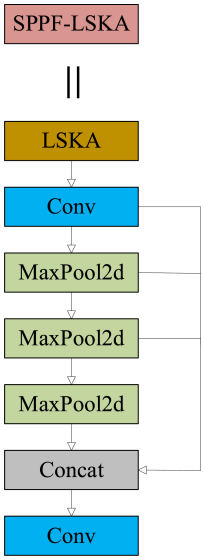


Fig. 3. SPPF-LSKA

$$F \in \mathbb{R}^{C \times H \times W} \tag{7}$$

where C denotes the number of input channels, and H

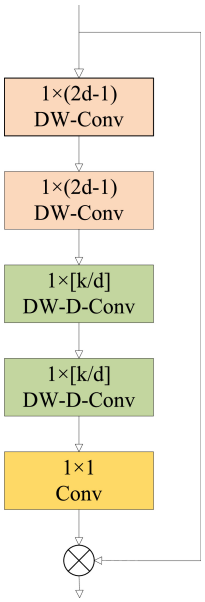


Fig. 4. LSKA

and W represent the height and width of the feature map, respectively. The output of the LSKA module is defined as follows:

$$\overline{Z}^C = \sum_{H,W} W^C_{(2d-1)\times 1} * \left(\sum_{H,W} W^C_{1\times(2d-1)} * F^C \right) \tag{8}$$

$$Z^C = \sum_{H,W} W^C_{[\frac{k}{d}]\times 1} * \left(\sum_{H,W} W^C_{1\times[\frac{k}{d}]} \times \overline{Z}^C \right) \tag{9}$$

$$A_C = W_{1\times 1} * Z^C \tag{10}$$

$$\overline{F}^C = A^C \otimes F^C \tag{11}$$

where, the symbols $*$ and $\otimes$ represent the convolution operation and the Hadamard product, respectively.Represents the depthwise convolution output obtained by convolving the input feature map F with a convolution kernel W of size $k \times k$. The left side of equation (8) shows the output of the depthwise convolution with a kernel size of $1 \times (2d-1)$, which helps capture local spatial information and, as shown in equation (9), is used to compensate for the lattice effect

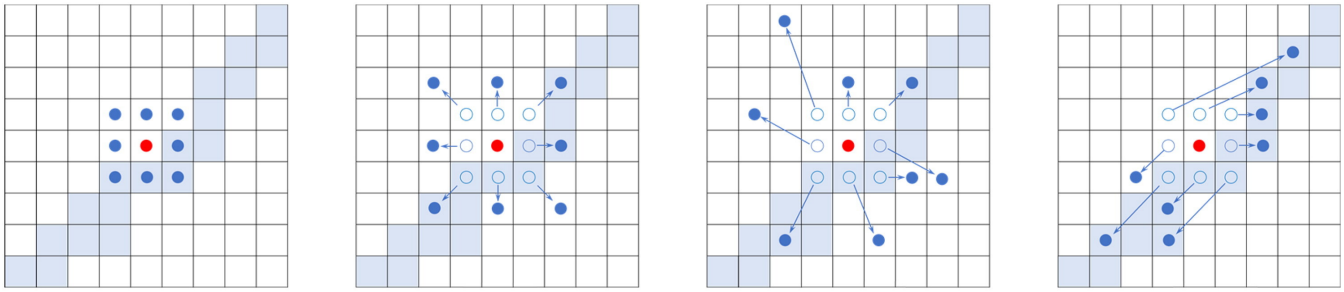


Fig. 5. Different Convolution Types: (a) Standard Convolution; (b) Dilated Convolution; (c) Deformable Convolution; (d) DSC.

that may occur in the subsequent depthwise convolution.The kernel size of the deep convolution is $\lfloor \frac{k}{d} \rfloor \times 1$, where k denotes the receptive field of the kernel W, and d is the dilation rate. A 1×1 convolutional kernel A_C is applied to produce the attention map. The Hadamard product between the attention map A^C and the input feature map F^C is represented by the left-hand side and the middle term of Equation (11) [18]. In summary, integrating the SPPF-LSKA module into the YOLOv10n model for road defect detection significantly improves the model's performance in complex road environments. Road defects such as fine cracks and potholes often appear at varying scales and locations. By incorporating the SPPF and LSKA modules, the model gains a more comprehensive understanding of road surface features, thereby enhancing detection accuracy. Specifically, the SPPF module reduces the dimensionality of feature maps through pooling operations, effectively lowering computational cost. Meanwhile, the LSKA module decomposes large convolutional kernels into 1D kernels, reducing both computational complexity and memory usage. The combination of these two modules enhances the model's ability to capture shape-related information, making it particularly effective for analyzing the geometric characteristics of road defects, and thus achieving superior detection performance in challenging scenarios.

B. C2f-DSC

Standard convolution performs poorly when handling datasets with diverse features (Fig. 5a), particularly for targets with tubular or fine curved structures, such as road cracks. To address this limitation, several alternative convolutional approaches have been proposed, including dilated convolution and deformable convolution [19]. However, dilated convolution lacks the ability to adaptively adjust its receptive field to focus on tubular structures (Fig. 5b). Although deformable convolution can learn regions of interest adaptively based on feature characteristics, it fails to preserve the connectivity of the focus regions, especially for fine, curved, tubular targets like road cracks (Fig. 5c).To overcome these issues, [20] proposed a dynamic snake convolution, which introduces convolutional kernels capable of dynamically adjusting their shape to enhance feature perception. During feature learning, this method allows the kernel to adapt its shape to better focus on elongated and tortuous local features within tubular structures (Fig. 5), thereby enabling more accurate representation of such structures and significantly improving detection and segmentation performance.

DSC introduces a deformable offset mechanism to adaptively adjust the shape of convolutional kernels, enabling more precise capture of geometric features in curved or highly continuous road defects. By iteratively learning optimal offsets, this method enhances the flexibility of convolution operations in the 2D spatial domain, improving the model's ability to perceive complex object shapes. In road defect detection tasks, DSC effectively addresses scenarios involving multi-scale and morphologically diverse defects, significantly improving both detection accuracy and robustness. The underlying principle of Dynamic Snake Convolution is as follows [21]:

First, for a standard 2D convolution with a given coordinate K (size $N \times N$), the center coordinate is denoted as $K_i = (x_i, y_i)$, where i is an integer. For a 3×3 convolutional kernel with a dilation rate of 1, K is represented as:

$$K = \{(x-l, y-l), (x-l, y), \cdots, (x+l, y+l)\} \quad (12)$$

where, K denotes the standard 2D convolution coordinates, with x and y representing the horizontal and vertical grid positions.

To better adapt convolutional kernels to the complex geometric characteristics of road defects, a deformable offset Δ is introduced in this study. However, if the model learns these offsets in a purely random manner, the receptive field may deviate from the actual defect areas, particularly in cases involving irregular patterns such as cracks or potholes. To address this issue, an iterative strategy is adopted, as illustrated in Fig. 6, where each defect region is progressively aligned with its optimal observation point. This approach ensures that feature attention remains consistently focused, preventing excessive expansion or displacement of the receptive field caused by large offset values.

In DSC, the standard convolutional kernel is linearized along the x-axis and y-axis. For a convolution kernel of size 9, for the x-axis, each specific position $K_i \pm c$ is calculated as: $(x_i \pm c, y_i \pm c)$, where $c = \{0, 1, 2, 3, 4\}$ represents the horizontal distance from the center grid. The selection of each grid position $K_i \pm c$ within the convolution kernel follows a progressive strategy: starting from the center, each subsequent position $K_i + 1$ is determined based on its predecessor with an incremental offset $\Delta = \{\delta \mid \delta \in [-1, 1]\}$. This accumulated offset Σ ensures that the convolution kernel is aligned with the morphological structure of road defect features along the x-axis. The modification along the y-axis

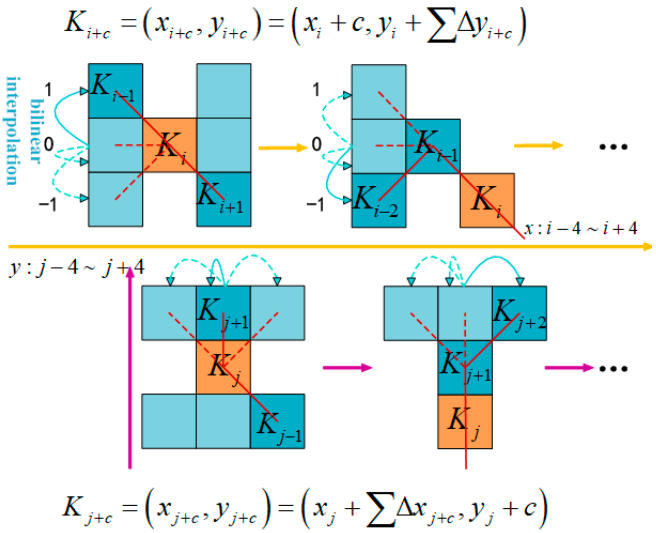


Fig. 6. DSC Coordinate Computation Diagram

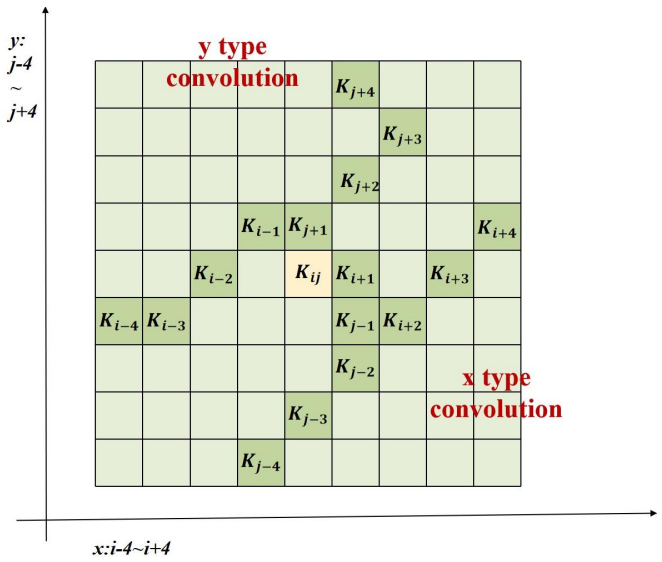


Fig. 7. Dynamic Structural Diagram of the DSC Module

is expressed in Equation (13) as follows:

$$K_{i\pm c} = \begin{cases} (x_{i+c}, y_{i+c}) = (x_i + c, y_i + \sum_i^{i+c} \Delta y), \\ (x_{i-c}, y_{i-c}) = (x_i - c, y_i + \sum_{i-c}^{i} \Delta y), \end{cases} \tag{13}$$

Equation (14) is modified along the y-axis as follows:

$$K_{j\pm c} = \begin{cases} (x_{j+c}, y_{j+c}) = (x_j + \sum_j^{j+c} \Delta x, y_i + c), \\ (x_{j-c}, y_{j-c}) = (x_j + \sum_{j-c}^{j} \Delta x, y_i - c), \end{cases} \tag{14}$$

Given that the offset Δ is generally a non-integer bilinear interpolation is employed as follows:

$$K = \sum_{K'} B(K', K) \cdot K' \tag{15}$$

where, K denotes the fractional positions in Equations (13) and (14), while K' enumerates all integer grid locations. B represents the bilinear interpolation kernel, which can be decomposed into two one-dimensional kernels as follows:

$$B(K, K') = b(K_x, K'_x) \cdot b(K_y, K'_y) \tag{16}$$

As illustrated in Fig. 7, due to the two-dimensional deformation, DSC covers a 9×9 region during its transformation, effectively expanding the receptive field. This adaptability to dynamic structures better captures the morphological characteristics of elongated road cracks, thereby enhancing the perception of critical features. Integrating DSC with the backbone network C2f significantly enhances segmentation accuracy, robustness, and real-time performance in road defect detection tasks. DSConv improves the perception of elongated and curved defects by employing adaptive convolutional kernels, while C2f strengthens the model's adaptability to complex environments through multi-scale feature fusion and contextual awareness. The synergy of these two components enables precise and stable detection of various road defects, making the approach particularly effective for complex road surfaces and dynamic environmental conditions.

C. Dysample

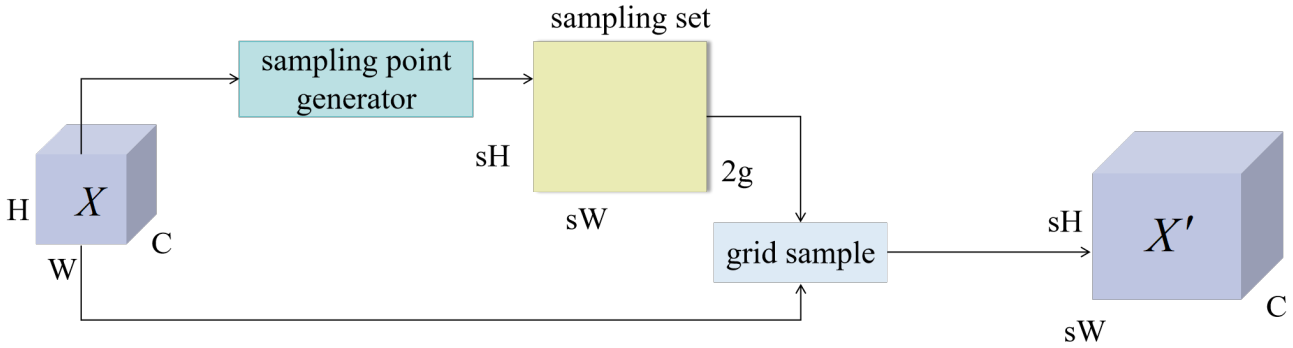
Due to variations in shooting angles and distances, the size of defects and the texture of the road surface in defect images may change, leading to pixel distortion and resulting in blurred defect edges. This makes it difficult for the model to accurately learn their features [22]. YOLOv10 employs the UpSample module in the neck network to restore high-resolution feature maps, thereby enhancing the detection capability for small defects (such as fine cracks) [23]. At the same time, it strengthens multi-scale feature fusion, reducing missed detections and false positives.To further optimize the upsampling performance, DySample is introduced to replace the original UpSample module. DySample enhances the representation of low-resolution or distant defects through a dynamic point sampling approach, enabling clearer extraction of defect region features. Moreover, it does not rely on additional CUDA libraries, which improves the model's upsampling efficiency.Fig. 8 illustrates the dynamic sampling mechanism and structural design of DySample.

Fig. 8 demonstrates the feasibility of the upsampling method based on dynamic point sampling. The execution process of DySample can be summarized as follows: first, the input feature map is passed through a sampling point generator to produce a sampling set, which contains a group of sampling locations. Next, a grid sampling operator resamples the input feature map according to these locations, typically using bilinear interpolation to obtain the feature value of each sampling point. Finally, the resampled feature map is output [3], providing new spatial structural information to better capture and represent target features. Specifically, this process can be formalized as:

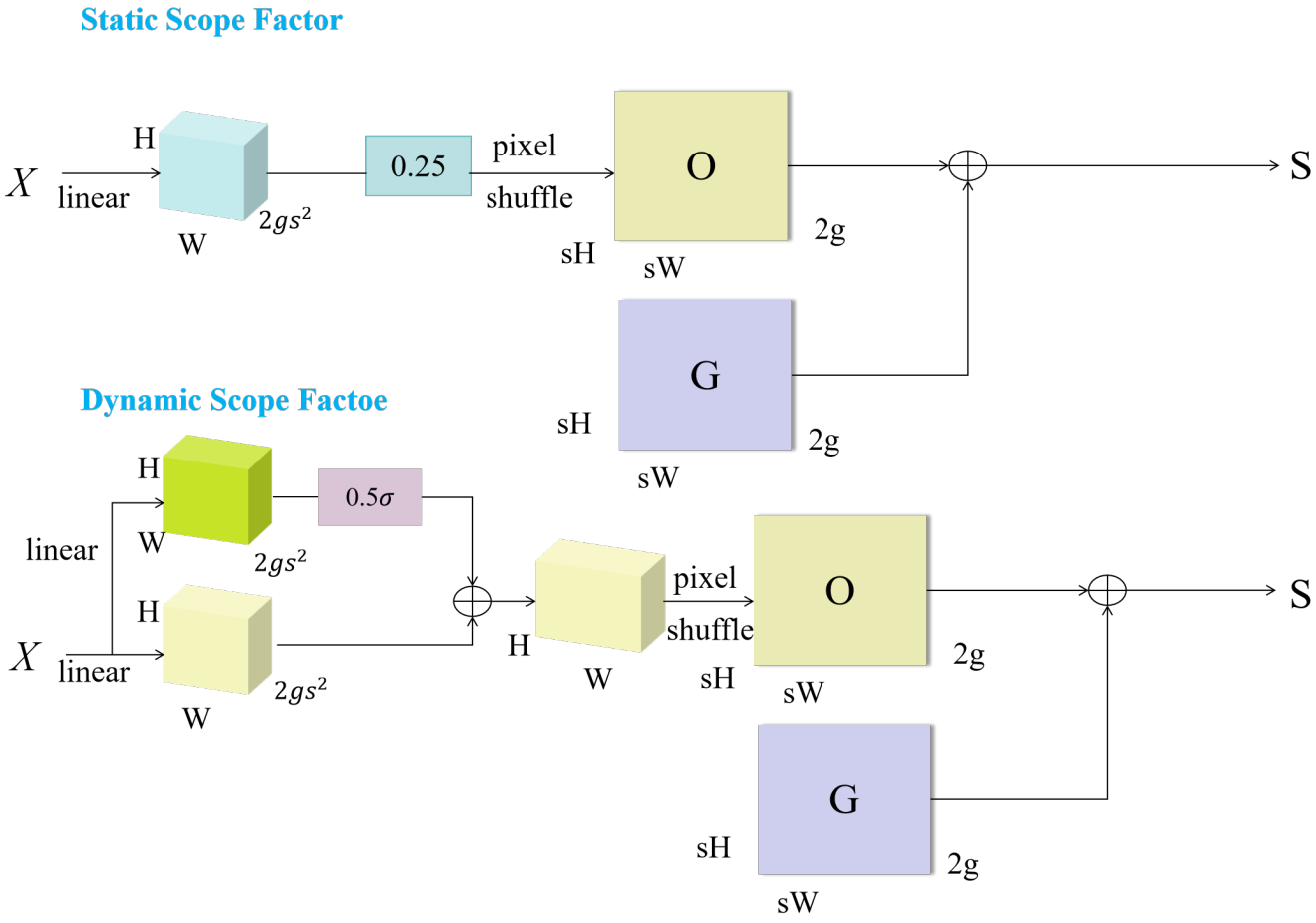
$$X' = \text{grid\_sample}(X, \delta) \tag{17}$$

where, X denotes the input feature map, and δ represents the generated sampling set. The `grid_sample()` function operates on X using the locations in δ to perform resampling, resulting in a new feature map X'.

Let the upsampling scale factor be s, and the dimensions of the feature map X be $C \times H \times W$.An output offset O of size $2s^2 \times H \times W$ is generated by a linear layer with C input channels and $2s^2$ output channels.Then, the offset O is rearranged into a shape of $2 \times sH \times sW$ using the pixel shuffle algorithm described in [24]. Finally, the sampling set



(a)Sampling based dynamic upsampling



(b)Sampling point generator in DySample

Fig. 8. DySample module

δ is obtained by adding the offset O to the original sampling grid G. This process is defined as follows:

$$O = \text{linear}(X) \tag{18}$$

$$\delta = G + O \tag{19}$$

The reshaping operation is omitted in this section. Finally, the upsampled feature map X', with dimensions $C \times sH \times sW$, is generated using the sampling set and the `grid_sample()` function, as shown in Equation (17). Fig. 8 illustrates the sampling-based dynamic upsampling technique and the module design of DySample.

IV. EXPERIMENT

A. Datasets and Experimental Setup

We conduct model training using the RDD2020 dataset from the CrowdAI Road Damage Detection Challenge (CRDDC2020). The dataset consists of 21,040 road images collected from Japan, India, and the Czech Republic, with 2,829, 7,705, and 10,506 samples respectively. It includes nine types of road damage: longitudinal cracks (D00), longitudinal construction joints (D01), transverse cracks (D10), transverse construction joints (D11), alligator cracks (D20), potholes (D40), blurred intersections (D43), blurred white lines (D44), and manholes (D50).This section presents a comprehensive overview of the model settings, training pro-

TABLE II
ABLATION EXPERIMENT

Method	SPPF-LSKA	C2f-DSC	DySample	mAP@0.5
YOLOv10n	-	-	-	52.60
Proposed method1	✓	-	-	54.80
Proposed method2	-	✓	-	54.30
Proposed method3	-	-	✓	54.30
Proposed method4	✓	✓	-	54.40
Proposed method5	✓	-	✓	55.50
Proposed method6	-	✓	✓	54.80
YOLOv10n-DSLD	✓	✓	✓	55.10

cess, evaluation metrics, ablation studies, and comparative experiments. All experiments are conducted on a machine equipped with a 10GB NVIDIA GeForce RTX 3080 GPU. The implementation is based on PyTorch 2.0.0 with Python 3.8, CUDA 11.8, and Ubuntu 20.04.

B. Network Training

The images are divided into training and validation sets with an 80:20 ratio. The maximum number of training epochs is set to 300. For practical reasons, the input images are standardized to a size of 640×640, representing the maximum size that the model can accommodate. The initial learning rate is set to 0.01, and the SGD optimizer is used for adjustments. To ensure fairness and accuracy, the same set of hyperparameters is used for both training and ablation experiments. The parameter configurations are detailed in Table I.

TABLE I
PARAMETER TABLE

Parameters	Setup
Epochs	300
Batch Size	16
Imgsize	640
Learning Rate	0.01
Patience	50
Optimizer	SGD
Workers	8
Weight-Decay	0.0005

C. Ablation Study Results Analysis

This paper improves the model by adding modules to both the backbone and neck networks, which reduces the model's parameter count and improves detection accuracy. To evaluate the impact of each module, comparative experiments are conducted to validate the effect of different combinations of the SPPF-LSKA, C2f-DSC, and DySample modules on road defect detection performance. Table II shows that introducing the combination of the SPPF-LSKA, C2f-DSC, and DS modules into YOLOv10n results in an increase of 2.2%, 1.7%, and 1.7% in mAP50, respectively. Notably, the performance improves significantly when the SPPF-LSKA module is added. The enhanced model exhibits higher accuracy, with each module contributing positively to the overall performance. Incorporating the LSKA large-kernel attention mechanism into SPPF achieves a better balance between expanding the receptive field and enhancing feature representation, thereby making the road defect detection model more accurate. The introduction of the DSC module enhances the model's ability to detect small road defects, improving accuracy. When all three modules are integrated, a significant increase in mAP50 accuracy is achieved.

Fig. 9 presents the normalized confusion matrix, illustrating the model's performance across different categories. The diagonal elements represent recall rates, with higher diagonal values indicating better detection performance. The off-diagonal elements reflect the confusion between categories, such as when a category is misclassified as background.

The improved road defect detection model demonstrates excellent accuracy, as shown in Fig. 10 The figure clearly illustrates the model's successful identification of various defect types, significantly enhancing its practicality and supporting road safety improvements. By accurately detecting different defects, the model aids relevant authorities in taking timely measures, thereby reducing the risk of traffic accidents.

D. Comparison Experiment

1) Comparison Experiment with Classic Models: YOLOv10 offers multiple versions to meet various object detection requirements, with differences in model size, computational complexity, architecture design, weights, and performance, including YOLOv10n, YOLOv10s, YOLOv10m, YOLOv10l, and YOLOv10x. Considering that road defect detection requires both accuracy and a certain detection speed, this paper selects YOLOv10n as the baseline network model. The YOLOv10n-DSLD model is then compared with YOLOv8n, YOLOv10s, YOLOv11n, and other road defect detection models to validate its effectiveness.

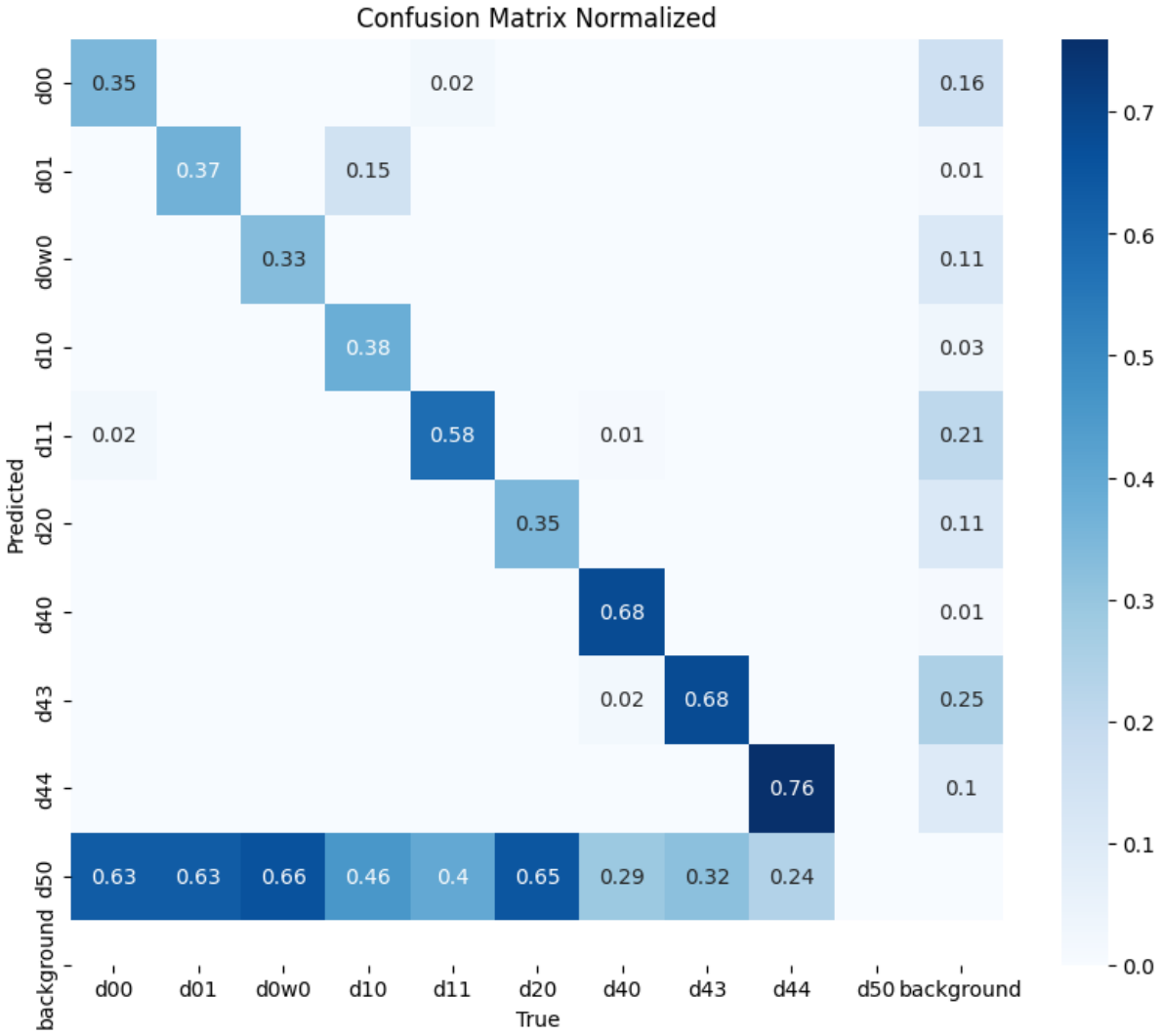


Fig. 9. Normalized Confusion Matrix of the Model

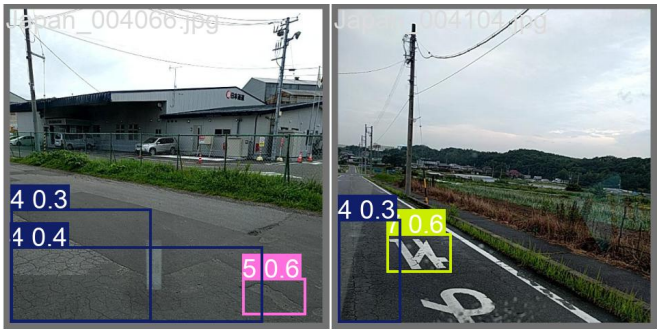


Fig. 10. Road Defect Detection Results

As shown in Fig. 12, the experimental results indicate that the performance of YOLOv10n-DSLD and other improved models stabilizes after 200 training epochs. Based on this, the training period (Epoch) is set to 300 epochs, with an early stopping mechanism (Patience=100) applied to effectively prevent overfitting while ensuring sufficient training. Regarding hardware configuration, empirical testing shows that with the NVIDIA RTX 3080 GPU, a batch size of 16 maximizes the utilization of GPU memory. The input image size is set to 640×640 pixels, which meets the real-time requirements of embedded devices while maintaining good detection accuracy.

In terms of optimizer parameter settings, the learning rate is set to 0.01 based on the following consideration: this value aligns with the characteristics of the Stochastic Gradient Descent (SGD) optimizer, ensuring stable parameter updates. Additionally, the number of data loading workers is set to 8, effectively preventing the GPU from being idle due to data I/O bottlenecks, thus improving training efficiency. Furthermore, the weight-decay parameter is set to 0.0005, which has been validated to strike a good balance between model capacity and regularization, significantly reducing the risk of overfitting. This parameter combination has been systematically optimized to maintain model performance while considering both training efficiency and generalization ability.

YOLOv10n-DSLD performs best among lightweight models, with a mAP50 of 55.1 and a mAP50-95 of 28.7, surpassing both YOLOv8n and YOLOv11n. Although its computational complexity is slightly higher than YOLOv11n, it achieves better detection accuracy, with a parameter count of only 2.8M, significantly lower than YOLOv10s. Compared to YOLOv8-DSCFEM and LeYOLOv8, this model outperforms in terms of accuracy, parameter size, and computational complexity, making it highly suitable for practical road defect

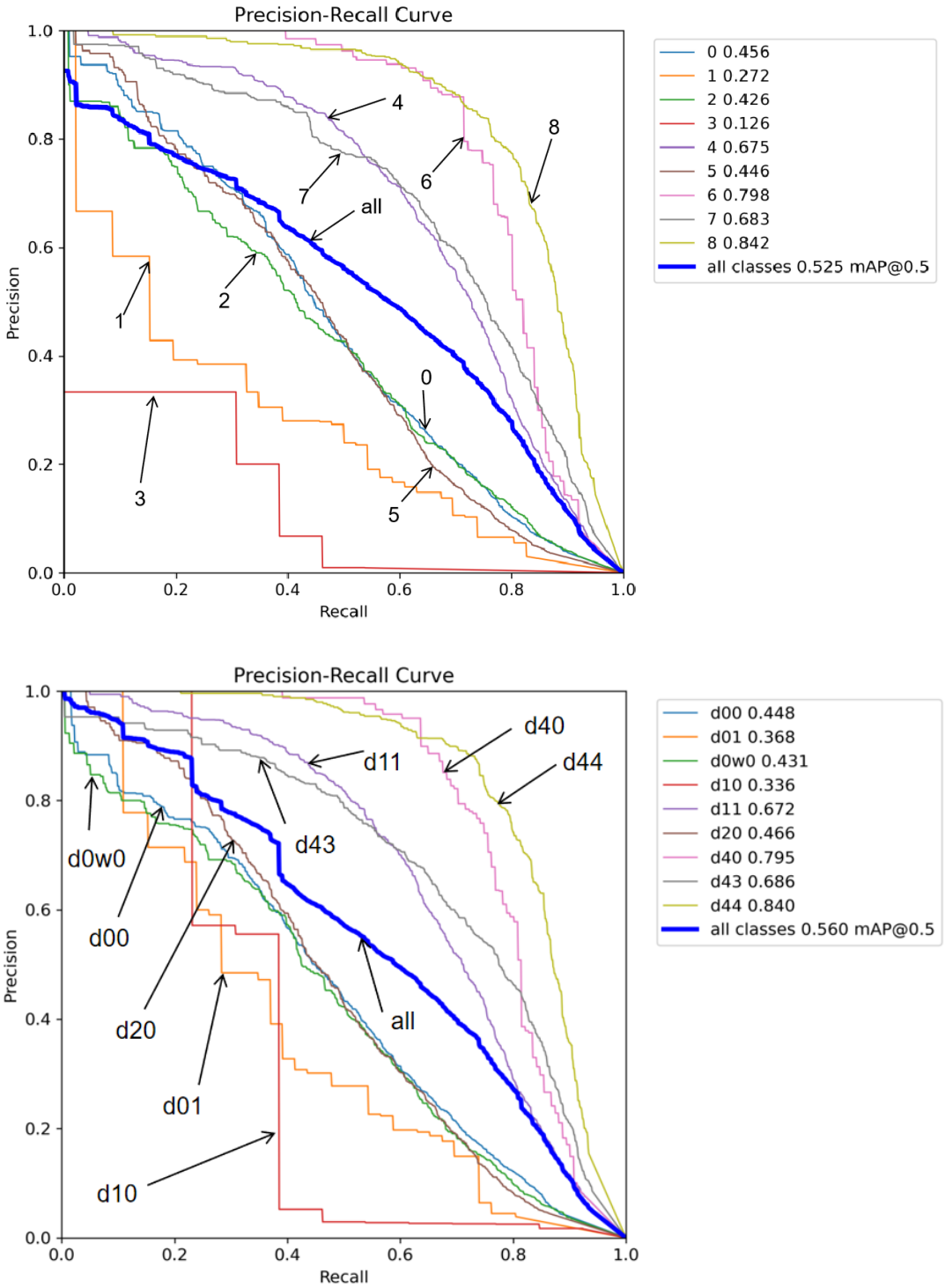


Fig. 11. Normalized Confusion Matrix of the Model

detection applications.

2) Impact of Different Attention Mechanisms on Network Performance: This section focuses on evaluating the role of LSKA in road detection models and compares it with other common attention mechanisms. iAFF enhances detection accuracy by integrating multi-scale channel attention modules, which fuse features of different scales and semantic inconsistencies. CBAM combines channel and spatial attention, improving feature representation while maintaining low computational cost. CGA improves computational efficiency and reduces parameter count by segmenting the input features into distinct parts and independently computing self-attention maps. Compared to these mechanisms, LSKA achieves a significant improvement in detection accuracy, as shown in Table IV.

V. CONCLUSIONS

This study presents a high-precision optimization of the YOLOv10n model, resulting in the improved YOLOv10n-DSLD model. The model incorporates the LSKA attention mechanism in the SPPF module to enhance contextual information extraction, integrates DSC in the C2f module

TABLE III
COMPARISON OF DETECTION PERFORMANCE OF DIFFERENT MODELS

Object Detection Algorithm	Params(M)	GFLOPS	mAP0.5	mAP0.5-0.95
YOLOv10n-EMSCP	3.2	8.2	54.1	27.2
YOLOv10n	2.7	8.7	52.6	27.6
YOLOv10s	8.0	24.8	55.2	28.5
YOLOv11n	2.6	6.5	52.3	27.3
YOLOv8-DSCFEM	3.0	8.2	54.7	27.8
LeYOLOv8s	3.1	8.0	54.1	27.8
YOLOv10n-DSLD	2.8	6.7	55.1	28.7

TABLE IV
COMPARISON OF DETECTION PERFORMANCE OF DIFFERENT MODELS

Object Detection Algorithm	Params(M)	GFLOPS	mAP0.5	mAP0.5-0.95
YOLOv10n	2.7	6.7	52.6	27.6
YOLOv10n+iAFF	3.0	8.2	52.5	27.2
YOLOv10s+CGA	2.9	6.7	52.9	27.3
YOLOv10n+CBAM	8.0	24.8	53.6	27.8
YOLOv10n+LSKA	3.2	6.5	54.8	28.3

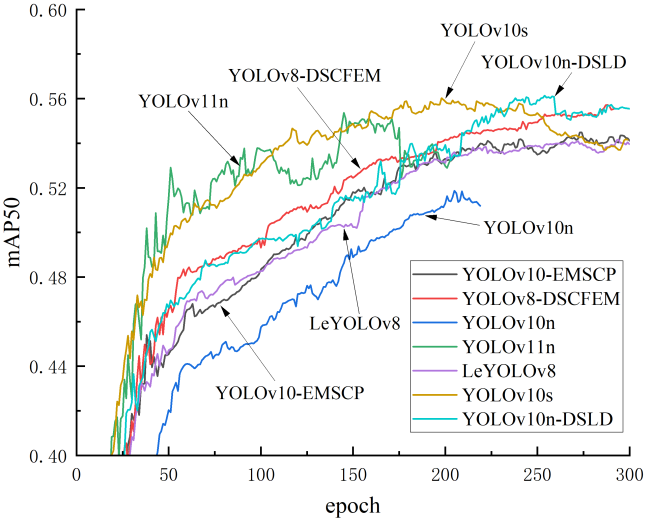


Fig. 12. Comparison curve of mAP0.5-0.95 and epoch of the improved YOLOv10n-DSLD models

to improve the perception of complex defects, and employs DySample to replace traditional upsamplers, thereby clarifying defect region features. Experimental results show that YOLOv10n-DSLD demonstrates stronger robustness in complex environments and small-object detection, improving detection accuracy, reducing false positive rates, and balancing both precision and efficiency, making it suitable for road defect detection. Its application in autonomous driving and intelligent transportation is significant, contributing to enhanced road safety and maintenance efficiency. Ablation experiments further validate the contribution of each module to performance improvement. Future research can focus on optimizing model speed while maintaining accuracy to better meet real-time detection requirements.

REFERENCES

[1] Jiaxu Leng, Yongming Ye, Mengjingcheng Mo, Chenqiang Gao, Ji Gan, Bin Xiao, and Xinbo Gao. Recent advances for aerial object detection: A survey. *ACM Computing Surveys*, 56(12):1–36, 2024.

[2] Shengqiang Cong, Chunna Zhang, Yang Yu, Xiaoping Yue, Jinchi Zhao, and Yuming Shen. Road defect detection model based on yolov8. IAENG International Journal of Computer Science, vol. 52, no. 4, pp986–994, 2025.

[3] Hu Haoyan, Tong Jinwu, Wang Haibin, and Lu Xinyun. Ead-yolov10: Lightweight steel surface defect detection algorithm research based on yolov10 improvement. *IEEE Access*, 2025.

[4] Pengchun Zhang, Haoran Chen, Jiahui Gao, Liqiang Ma, and Rong He. Improved yolov10 for high-precision road defect detection. In *2024 4th International Conference on Computer Science and Blockchain (CCSB)*, pages 79–83. IEEE, 2024.

[5] Ao Li, Chunrui Wang, Tongtong Ji, Qiyang Wang, and Tianxue Zhang. D3-yolov10: Improved yolov10-based lightweight tomato detection algorithm under facility scenario. *Agriculture*, 14(12):2268, 2024.

[6] Yunxuan Wang, Yang Yong, and Chuan Li. Rotating object detection method of insulator defect base on improved yolov5. In *Frontier Academic Forum of Electrical Engineering*, pages 739–748. Springer, 2024.

[7] JIN Xuemeng, LIANG Xiyin, and DENG Pengfei. Lightweight daylily grading and detection model based on improved yolov10. *Smart Agriculture*, 6(5):108, 2024.

[8] Jianxi OU, Jianqin Zhang, Haoyu Li, and Bin Duan. An improved yolov10-based lightweight multi-scale feature fusion model for road defect detection and its applications. *Available at SSRN 4970753*.

[9] Jin Liu, Guorui Zhan, Diandian Wang, Yanqin Kang, Kun Wang, Jun Qiang, and Yikun Zhang. Mms-yolov10: A fast and improved pavement surface defect detection model based on yolov10. 2024.

[10] Di Wu, Ao Zheng, Wenshuai Yu, Hongbin Cao, Qiuyuan Ling, Jiawen Liu, and Dandan Zhou. Digital twin technology in transportation infrastructure: A comprehensive survey of current applications, challenges, and future directions. *Applied Sciences*, 15(4):1911, 2025.

[11] Muhammad Hussain. Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision. *arXiv preprint arXiv:2407.02988*, 2024.

[12] Kin Wai Lau, Lai-Man Po, and Yasar Abbas Ur Rehman. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Systems with Applications*, 236:121352, 2024.

[13] Jun Tie, Chengao Zhu, Lu Zheng, HaiJiao Wang, ChongWei Ruan, Mian Wu, Ke Xu, and JiaQing Liu. Lska-yolov8: A lightweight steel surface defect detection algorithm based on yolov8 improvement. *Alexandria Engineering Journal*, 109:201–212, 2024.

[14] Li Deng, Siqi Wu, Jin Zhou, Shuang Zou, and Quanyi Liu. Lska-yolov8n-wiou: An enhanced yolov8n method for early fire detection in airplane hangars. *Fire*, 8(2):67, 2025.

[15] Kaiwei Yu, I Chen, Jing Wu, et al. Dscformer: A dual-branch network integrating enhanced dynamic snake convolution and segformer for crack segmentation. *arXiv preprint arXiv:2411.09371*, 2024.

[16] Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6027–6037, 2023.

[17] Jing Rui, Chi Kin Lam, Tao Tan, and Yue Sun. Dlw-yolo: Improved yolo for student behaviour recognition. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 332–337. IEEE, 2024.

[18] Xinrong Zhang, Yanlong Wang, and Huaisong Fang. Steel surface defect detection algorithm based on esi-yolov8. *Materials Research Express*, 11(5):056509, 2024.

[19] Ming Lu, Wangqi Sheng, Ying Zou, Yating Chen, and Zuguo Chen. Wss-yolo: An improved industrial defect detection network for steel surface defects. *Measurement*, 236:115060, 2024.

[20] Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6070–6079, 2023.

[21] Guandong Li and Mengxia Ye. Spatial-geometry enhanced 3d dynamic snake convolutional neural network for hyperspectral image classification. *arXiv preprint arXiv:2504.04463*, 2025.

[22] Bensheng Yun, Xiaohan Xu, Jie Zeng, Zhenyu Lin, Jing He, and Qiaoling Dai. An improved unmanned aerial vehicle forest fire detection model based on yolov8. *Fire*, 8(4):138, 2025.

[23] WU Liuai and XU Xueke. Lightweight tomato leaf disease and pest detection method based on improved yolov10n. *Smart Agriculture*, 7(1):146, 2025.

[24] Yin Zhang, Mu Ye, Guiyi Zhu, Yong Liu, Pengyu Guo, and Junhua Yan. Ffca-yolo for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.