A novel MO-YOLOv4 for segmentation of multi-class bridge damages

Zengsheng He\textsuperscript{a}, Cheng Su\textsuperscript{a,b,c,*}, Yichuan Deng\textsuperscript{a,b,c}

\textsuperscript{a} School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China.

\textsuperscript{b} State Key Laboratory of Subtropical Building Science, South China University of Technology, Guangzhou, China.

\textsuperscript{c} Guangdong Artificial Intelligence and Digital Economy Laboratory, Guangzhou, China

* Corresponding author

Abstract: Convolutional Neural Network (CNN)-based semantic segmentation methods have demonstrated remarkable achievements in the field of concrete crack identification. However, when it comes to the more crucial task of multi-class damage segmentation in bridge inspection, most semantic segmentation methods fail to achieve the desired performance due to major challenges posed by the significant variations in the scale of different damages and the susceptibility of network training to background interferences. To tackle these challenges, a novel multi-class damage segmentation method, MO-YOLOv4, is proposed based on the widely used YOLOv4 structure. MO-YOLOv4 incorporates the Multi-Scale Attention (MSA) module to greatly enhance the feature extraction ability for multi-class damages, and employs the Oriented Bounding Box (OBB)-based segmentation strategy for effectively reducing the influence of background interferences during the network training phase. The effectiveness of the MSA module and OBB-based segmentation strategy is validated through experiments, and a comparative study is conducted to compare the proposed MO-YOLOv4 with the existing semantic segmentation methods for multi-class damages. The experimental
results effectively demonstrate the superior performance of MO-YOLOv4 in multi-
class damage segmentation.

**Keywords:** Multi-class damage segmentation; Convolutional neural network;
YOLOv4; Multi-scale attention; Oriented bounding box

1. Introduction

Bridges play a significant role in enhancing transportation efficiency and fostering
economic development within urban areas. However, as time passes, bridge structures
are subject to the cumulative effects of vehicle load, wind load and environmental
factors such as temperature and humidity. Consequently, these bridge structures
inevitably exhibit various forms of visible damage, including concrete cracks, rebar
exposure, spalling, pits and holes, which serve as visual indicators of structural
degradation. Therefore, conducting regular inspections to assess the visual condition of
in-service bridges becomes imperative for effectively evaluating their long-term
degradation and providing substantial support for the bridge health monitoring.
Nevertheless, the traditional method of bridge inspection predominantly relies on
manual labor, resulting in low efficiency and unreliable evaluation.

To address the aforementioned challenges, the civil engineering sector has been
actively engaged in investigating an automated and accurate method for bridge
inspection. Extensive research has revealed that the Convolutional Neural Network
(CNN) [1], a fundamental model of deep learning techniques, possesses the capability
of autonomously extracting highly relevant feature information from images through
its multi-layer structure. By utilizing this remarkable feature learning ability, the CNN-
based method has surpassed the traditional image processing techniques and swiftly
emerged as the mainstream method in the current domain of bridge damage recognition.
Cha et al. [2] successfully integrated CNN with the sliding window strategy to achieve
highly accurate concrete crack detection, even under varying lighting and shadow
method based on CNN, demonstrating its robustness when applied to on-site
environments. However, the above methods primarily concentrate on identifying
damages at the image level, essentially functioning as the CNN-based image
classification method.

To achieve damage identification at the image sub-region level, researchers have
devoted significant attention to CNN-based object detection methods, such as Faster R-
CNN [4] and YOLO [5], which utilize bounding boxes to locate damages. Cha et al. [6]
employed Faster R-CNN to successfully detect concrete cracks, steel corrosion, bolt
corrosion and steel delamination, achieving an average accuracy of 87%. Deng et al.
[7] demonstrated the effectiveness of Faster R-CNN in concrete crack detection, even
in the presence of handwriting script interferences. Li et al. [8] further enhanced the
performance of Faster R-CNN in concrete crack detection by designing a multi-layer
fusion method. Zhang et al. [9] proved that the third-generation YOLO series algorithm,
YOLOv3, can attain a comparable level of accuracy in damage detection to that of
Faster R-CNN, while maintaining computational speed. Furthermore, Yu et al. [10]
designed a novel method YOLOv4-FPM for concrete crack detection, which combines
the focal loss, pruning algorithm and multi-scale dataset. Roy and Bhaduri [11]
proposed a novel DenseSPH-YOLOv5 model for detecting road damages, such as
cracks, potholes and ruts, achieving high detection accuracy as well as real-time
detection speed. Xu et al. [12] further improved the performance of YOLOv7 in crack
detection, which is capable of real-time monitoring of crack dynamic expansion. In
contrast to the aforementioned methods that employ horizontal bounding boxes for
damage detection, He et al. [13] have shown that the utilization of Oriented Bounding
Boxes (OBB) can effectively encapsulate concrete cracks. This approach greatly
reduces the impact of background interferences during the network training phase,
thereby substantially enhancing the efficacy of network training.

In comparison to the previously mentioned CNN-based object detection methods
that primarily focus on identifying damages at the image sub-region level, the CNN-
based semantic segmentation method enables damage identification at a more granular pixel level. The results of semantic segmentation can be directly utilized for subsequent calculations of damage characteristic parameters, such as length, width and area. Therefore, the semantic segmentation method proves to be more suitable for the bridge damage inspection compared with the other CNN-based methods. Concrete cracks, as the most common and significant form of damage on bridge surface, have prompted extensive research in the development of concrete crack segmentation methods. Yang et al. [14] employed the Full Convolutional Network (FCN), a widely recognized semantic segmentation method, to effectively segment concrete cracks and subsequently quantify their lengths and widths. Alipour et al. [15] developed CrackPix, an FCN-based method that incorporates additional shallow feature information, which results in a high level of accuracy in concrete crack segmentation. Ren et al. [16] employed the dilated convolution to effectively enlarge the receptive field size in FCN, thereby enhancing the network capability for feature extraction. Building upon this, they proposed a novel network called CrackSegNet, which demonstrates precise segmentation of concrete cracks. Additionally, Ni et al. [17] devised a lightweight network with high performance for concrete crack segmentation, incorporating a novel strategy based on Generative Adversarial Network (GAN). In order to alleviate the requirement for a large number of labeled data during network training, Jian and Liu [18] introduced a novel semi-supervised method for concrete crack segmentation, capable of utilizing both labeled and unlabeled data.

However, owing to the small proportion of concrete crack pixels in the image, semantic segmentation networks are prone to interference from a large amount of background pixels during the training phase. As a consequence, the efficiency of extracting concrete crack features is significantly diminished. To tackle this issue, the attention mechanism [19] has emerged as a focal point of research in recent years. Its objective is to enhance the feature extraction ability of segmentation networks by assigning weights to prioritize the target object while suppressing irrelevant
background information. Following that, Chen and He [20] introduced the attention gate module to enhance Unet, another well-established semantic segmentation method. Experimental results demonstrate that the enhanced Unet achieved a high accuracy of over 90% in concrete crack segmentation. Similarly, Kang and Cha [21] put forth STRNet, an efficient and lightweight network for concrete crack segmentation, which harnesses the power of attention mechanism.

The task of multi-class damage segmentation in bridge inspection is considerably more intricate and demanding compared to the single-class concrete crack segmentation. Consequently, the advancement of research in the domain of multi-class damage segmentation has been relatively sluggish. In 2019, Li et al. [22] made a pioneering contribution by developing an FCN-based segmentation method for multi-class damages, including concrete cracks, spalling, efflorescence and holes. Subsequently, in 2022, Dong et al. [23] proposed a novel Road-Seg-CapsNet, which leverages feature fusion technology to achieve more accurate segmentation results for concrete cracks and potholes. In the same year, Li and Zhao [24] successfully improved the segmentation accuracy of concrete cracks and spalling by employing the ensemble learning strategy that combines multiple semantic segmentation networks. Despite this, there still exist segmentation challenges posed by multi-class damages, primarily due to the substantial variations in damage scale and the susceptibility of network training to background interferences.

To address the aforementioned challenges, this paper proposes a novel segmentation method for multi-class damages, which builds upon the widely utilized YOLOv4 framework [25]. Specifically, the proposed method introduces the Multi-Scale Attention (MSA) module, which enhances the performance of traditional YOLOv4 by improving its ability to extract features that are relevant to multi-class damages with substantial variations in their scales. Subsequently, the proposed method further integrates the OBB-based segmentation strategy into the enhanced YOLOv4 mentioned above, thereby reducing the segmentation area of damages and effectively
mitigating the impact of background interferences during network training. To
distinguish it from the traditional YOLOv4, the proposed method is termed MO-
YOLOv4, owing to the utilization of the MSA mechanism and the OBB-based
segmentation strategy. The effectiveness of MSA module and OBB-based
segmentation strategy is validated through experiments, and a comparative study is
conducted to compare the proposed MO-YOLOv4 with the existing semantic
segmentation methods, indicating the superior performance of MO-YOLOv4 in multi-
class damage segmentation.

2. Methodology

2.1 Architecture of MO-YOLOv4

In the traditional YOLOv4, the Cross Stage Partial Darknet53 (CSPDarknet53),
Spatial Pyramid Pooling (SPP) module are employed to extract initial features of the
target object. Subsequently, the Path Aggregation Network (PAN) module is further
utilized to capture the key information from the initial features. Finally, the extracted
key features are utilized by the YOLO head module to generate horizontal bounding
boxes that locate the target object. To propose the novel MO-YOLOv4 for multi-class
damage segmentation, two enhancements have been implemented in the traditional
YOLOv4 framework: (1) The MSA mechanism is introduced into the original PAN,
resulting in the MSA-PAN module. This enhancement improves its capability to extract
multi-scale damage features, which is significant for achieving high performance in
multi-class segmentation. (2) The YOLO head module is modified to generate OBBs
that tightly enclose the damages, referred to as the detection branch, and the
segmentation branch is incorporated to perform damage segmentation within the OBBs,
leading to the OBB-based segmentation for multi-class damages.

The architecture of the proposed MO-YOLOv4 is illustrated in Fig. 1. Initially,
the CSPDarkent53 and SPP module are utilized to extract Feature 1, Feature 2 and
Feature 3 from the input damage image with a default resolution of 416×416.
Subsequently, these initial features undergo optimization through three MSA modules embedded in the MSA-PAN module, resulting in the enriched Feature 1*, Feature 2* and Feature 3* with multi-scale characteristic information. Finally, in the OBB-based segmentation module, these multi-scale features are utilized to accurately locate the damage regions using OBBs through the detection branch, and both the multi-scale features and OBBs are employed to generate the segmentation mask within these located damage regions through the segmentation branch.

Fig. 1. The architecture of MO-YOLOv4.

2.2 MSA module

The attention mechanism has emerged as a prominent research topic in the field of CNN in recent years. Its effectiveness has been demonstrated through extensive studies across various CNN-based visual tasks, including image classification, object detection and image segmentation. The core principle of the attention mechanism involves extracting distinctive information from the input feature map, enabling the generation of an attention map. This attention map is then employed to adaptively adjust the weighting of the input feature map. Through this adaptive weighting process, the attention mechanism effectively suppresses irrelevant information while simultaneously reinforcing the representation of the target object, thereby facilitating the extraction of target features. However, traditional attention mechanism modules typically possess a single-scale receptive field [26], which limits their capability to extract features of multi-class damages with significant scale variations. To overcome this limitation, this study draws inspiration from current attention mechanism studies [27]-[30] and designs an attention mechanism module with multi-scale receptive fields, namely the MSA module.
The attention mechanism module can be viewed as a miniature CNN, where the shallow convolutional layers correspond to small-scale receptive fields, facilitating the feature extraction from small-scale damages. Conversely, the deep convolutional layers correspond to large-scale receptive fields, enabling the capture of features relevant to large-scale damages. For the input feature map with a size of $W \times H \times C$ shown in Fig. 2, the MSA module introduces four convolutional kernels: one $1 \times 1$ kernel, one $3 \times 3$ kernel and two $3 \times 3$ kernels with dilation rate (DR) of 2 and 3 [31], [32], respectively. This process generates four convolutional layers with varying receptive field scales, resulting in four feature maps that contain feature information at different scales. Subsequently, through channel-wise concatenation and sigmoid activation, these feature maps, which have the same size of $W \times H \times C / 4$, are fused into an attention map with a size of $W \times H \times C$, enabling the encompassment of multi-scale feature information. Following the principle of the attention mechanism, the attention map is further utilized to adaptively assign weights to each element of the input feature map through the element-wise multiplication. This process will effectively generate the output feature map with multi-scale feature information.

In this study, the three MSA modules designed above are integrated into the MSA-PAN module, as illustrated in Fig. 1. This integration leads to a notable enhancement in the feature extraction process for multi-class damages.

**2.3 OBB-based segmentation strategy**
Due to the constrained shooting conditions in the on-site environment, the bridge damage images obtained in practice typically contain a significant amount of background information. This greatly increases the difficulty of network training for traditional semantic segmentation methods. In response to this challenge, the MO-YOLOv4 introduces the OBB-based segmentation strategy that combines the functionalities of object detection and semantic segmentation. As illustrated in Fig. 1, the detection branch of MO-YOLOv4 initially generates OBBs to obtain the precise damage regions, and then the segmentation branch further generates the segmentation mask within these obtained damage regions. Compared with the traditional Horizontal Bounding Box (HBB), the adopted OBB has the capability to tightly enclose the damage. This enables the detection branch to obtain the damage regions with less background noise and other irrelevant information. Consequently, under the OBB-based segmentation strategy, the segmentation branch can solely employ these regions with prominent damage characteristics for segmentation training. This leads to higher training efficiency and segmentation accuracy of MO-YOLOv4.

To achieve the aforementioned objective, the detection branch undergoes the modification by augmenting the channel dimension of the convolutional layer in the YOLO head of the traditional YOLOv4. This augmentation introduces the rotation angle parameter, thereby enabling detection branch to generate OBBs for the detection of damage regions. Specifically, each OBB can be identified by the parameters \((c, x, y, w, h, \theta)\), in which \(c\) denotes the category of the damage within the box, \(x\) and \(y\) denote the coordinates of the center point of the box; \(w\) and \(h\) denote the width and height of the box, respectively; and \(\theta\) denotes the rotation angle of the box.

Subsequently, the segmentation branch adopts a post-processing operation to select the segmentation mask within the damage region from the full region segmentation mask. The input feature maps of the segmentation branch include Feature 1\(^*\), Feature 2\(^*\) and Feature 3\(^*\) with the corresponding sizes of \(52 \times 52 \times 128\), \(26 \times 26 \times 256\) and \(13 \times 13 \times 512\), respectively, as shown in Fig. 3. To achieve full region
segmentation, Feature 2* and Feature 3* are initially modified by a 1×1 convolutional kernel and then magnified by two upsampling operations with the magnifications of 2 and 4, respectively. This results in two new feature maps with the same size as Feature 1*, visually represented by the green and blue box within the dashed boxes in Fig. 3. Then, these three feature maps are fused into a new feature map with a size of 52×52×384 through the process of channel-wise concatenation. Following common techniques employed in semantic segmentation methods, the fused feature map undergoes the processing utilizing multiple convolutional kernels, upsampling operations and a sigmoid activation function. This process generates the full region segmentation mask with a size of 416×416×C, which represents the identification results of C damage categories for each pixel in the input image with the resolution of 416×416. To extract the segmentation mask within the damage region from the full region segmentation mask, the post-processing operation in segmentation branch selectively incorporates the information within the OBB characterized by the parameters (c, x, y, w, h, θ) obtained by the detection branch.

Fig. 3. The network structure of segmentation branch.
(Note: Each convolution layer is followed by batch normalization and Leaky Relu activation)

3. Experimental validation

3.1 Dataset construction

In the present study, a total of 1203 damage images, including images of concrete cracks, spalling, rebar exposure and potholes, were carefully selected from a public bridge crack dataset [33] and bridge inspection reports provided by a bridge inspection company in Guangzhou, China. Subsequently, in order to construct the multi-class
damage dataset, the professional annotation software RoLabelImg [34] was utilized to obtain the labels of OBBs for the damage images, which serve as the ground-truth boxes. Then, another professional annotation software Labelme [35] was further employed to generate ground-truth masks, which represent the labels of segmentation masks for the damage images. The visual representation of the four types of damages, along with the corresponding ground-truth boxes and ground-truth masks, is illustrated in Fig. 4.

It is worth noting that, to facilitate the detection branch of MO-YOLOv4 in generating OBBs that can tightly enclose various types of damages, this study employed different numbers of ground truth boxes based on the distribution characteristics of damages during the labeling process. For the concrete crack displaying multiple distribution directions, as shown in Fig. 4 (a), the objective of tightly enclosing the damage is achieved by employing multiple ground-truth boxes. Conversely, for damages exhibiting a relatively uniform distribution direction, as depicted in Fig. 4 (b) to (d), a single ground-truth box is utilized to tightly enclose the damage.

![Fig. 4. The visual representation of the four types of damages with the corresponding ground-truth boxes and ground-truth masks: (a) concrete crack, (b) spalling, (c) rebar exposure, (d) pothole.](image-url)
The above multi-class damage dataset was divided into the training set and the testing set with an approximate ratio of 8:2. The total samples of training set and testing set, as well as the sample distribution of different damages, are presented in Table 1. In the following experiments, the damage samples from the training set are used for network training, and subsequently, the segmentation performance of multi-class damages is evaluated on the testing set.

**Table 1.** Details of the training and testing set.

<table>
<thead>
<tr>
<th>Type</th>
<th>Total number of samples</th>
<th>Number of samples for each type of damages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Concrete crack</td>
</tr>
<tr>
<td>Training set</td>
<td>975</td>
<td>386</td>
</tr>
<tr>
<td>Testing set</td>
<td>228</td>
<td>87</td>
</tr>
</tbody>
</table>

(Notes: Each damage sample contains the damage image with the corresponding ground-truth box and ground-truth mask)

### 3.2 Training details

Due to the utilization of the OBB-based segmentation strategy, the proposed MO-YOLOv4 incorporates the loss functions from both object detection and semantic segmentation methods into its own loss function, denoted as \( L_{\text{total}} = L_{\theta} + L_S \). During the training phase, the object detection loss function is defined as

\[
L_D = L_f + L_{ce} + L_{\text{CloU}} + L_{\text{smooth L1}}.
\]

Similar to the traditional YOLOv4, the focal loss function \( L_f \) [36] and cross entropy loss function \( L_{ce} \) [1] serve as the confidence loss function and the classification loss function for the OBB, respectively, and the CIoU loss function \( L_{\text{CloU}} \) [37] is employed to optimize the position parameters \((x, y, w, h)\) of the OBB. Furthermore, an additional smooth L1 loss function \( L_{\text{smooth L1}} \) [4] is utilized in \( L_D \) for the optimization of the rotation angle parameter \( \theta \) of the OBB. As for the semantic segmentation loss function \( L_S \), it is defined as a specific cross entropy loss function \( L_S \text{ (target, prediction)} \), where the target is the ground truth mask within the OBB provided by the detection branch, and the prediction is the segmentation mask generated by the segmentation branch within the OBB.
To improve the training efficiency of MO-YOLOv4, its feature extraction module, CSPDarknet53, employed pre-trained weight parameters obtained from the widely used public visual dataset Pascal VOC [38]. During the training phase, MO-YOLOv4 adopted an epoch size of 100 and a batch size of 16, along with the Adam optimizer [39]. Additionally, considering the favorable network performance of CSPDarknet53, which utilized the pre-trained weight parameters, the network training process exclusively updated the weights of the other modules in MO-YOLOv4 for the first 25 epochs, aiming to swiftly improve the network performance of these modules. Subsequently, for the last 75 epochs, all weights of MO-YOLOv4 were optimized to comprehensively enhance its overall performance. Throughout this training process, the initial learning rates for the first 25 epochs and the last 75 epochs were set as 0.001 and 0.0001, respectively. The corresponding learning rate decay strategy was as follows: after every 5 epochs, the current learning rate was adjusted to 90% of the learning rate from the previous epoch.

In the present study, the open-source deep learning library Pytorch [40] was utilized to establish MO-YOLOv4 network. All the experiments involved were conducted on the hardware configuration of the Intel ® Core ™ i5-279 10400 CPU @ 64-bit 2.90 GHz, 16 GB memory and NVIDIA GeForce RTX 3090 GPU.

### 3.3 Effectiveness of MSA module

In the present study, the MSA module with multi-scale receptive fields is designed to improve the segmentation performance of multi-class damages with varying scales. To validate the effectiveness of the MSA module in multi-class damage segmentation, comparative studies were conducted to assess its multi-class feature extraction ability in comparison to traditional attention modules with single-scale receptive fields, namely Convolutional Block Attention (CBA) [41] and Efficient Channel Attention (ECA) [42] module. For the sake of convenience in subsequent discussions, the methods that incorporate the CBA and ECA module, which are also built upon the...
OBB-based segmentation strategy, are denoted as CO-YOLOv4 and EO-YOLOv4, respectively, following the naming convention of the proposed MO-YOLOv4.

To facilitate an intuitive comparison of the multi-class damage feature extraction abilities among the aforementioned attention mechanism modules, Fig. 5 visually presents the damage feature maps obtained by these modules in the form of heat maps [43]. These heat maps highlight the regions of focus identified by the respective attention mechanism modules. Notably, Fig. 5 (b) reveals that CO-YOLOv4 exhibits an excessively larger focused region compared to the actual region of spalling in the original image. Similarly, the focused regions of CO-YOLOv4 and EO-YOLOv4, as depicted in Fig. 5 (d), fail to distribute near the actual region of pothole. Consequently, both EO-YOLOv4 and CO-YOLOv4 demonstrate limited effectiveness in extracting the features of spalling and pothole, respectively. In contrast, the heat maps generated by MO-YOLOv4 provide a more distinct outline of the appearances of all four types of damages. This observation highlights the stronger feature extraction capability of the MSA module for multi-class damages, surpassing the capabilities of the traditional CBA and ECA module.
Fig. 5. The damage heat maps obtained by different attention mechanism modules: (a) concrete crack, (b) spalling, (c) rebar exposure, (d) pothole.

Subsequently, a comparison of Mean Intersection over Union (MIoU) metric [44] on the testing set is conducted to further demonstrate the enhancement brought by the MSA module in multi-class damage segmentation. In essence, the MIoU metric represents the average of the Intersection over Union (IoU) values across multi-class objects (including the background). As shown in Table 2, the IoU metric for spalling achieved by CO-YOLOv4 is significantly lower compared to that attained by MO-YOLOv4. Furthermore, the IoU metrics for pothole obtained by CO-YOLOv4 and EO-YOLOv4 exhibit notably inferior performance in contrast to those achieved by MO-YOLOv4. Overall, the integration of the MSA module with multi-scale receptive fields in MO-YOLOv4 yields a higher MIoU metric of 64.28%, surpassing the metrics achieved by CO-YOLOv4 and EO-YOLOv4, both of which employ single-scale receptive fields.
Table 2. Comparison of the attention mechanism modules with different scales of receptive fields.

<table>
<thead>
<tr>
<th>Method</th>
<th>Background</th>
<th>Concrete crack</th>
<th>Spalling</th>
<th>Rebar exposure</th>
<th>Pothole</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO-YOLOv4</td>
<td>96.80</td>
<td>49.27</td>
<td>55.21</td>
<td>46.91</td>
<td>51.51</td>
</tr>
<tr>
<td>EO-YOLOv4</td>
<td>96.89</td>
<td>49.49</td>
<td>57.34</td>
<td>48.97</td>
<td>59.05</td>
</tr>
<tr>
<td>MO-YOLOv4</td>
<td>96.95</td>
<td>52.69</td>
<td>60.27</td>
<td>48.32</td>
<td>63.16</td>
</tr>
</tbody>
</table>

3.4 Effectiveness of OBB-based segmentation strategy

By employing the OBB-based segmentation strategy for multi-class damage segmentation, the proposed MO-YOLOv4 is capable of generating OBBs to precisely locate the damage regions, which are subsequently utilized for segmentation training. In comparison to the utilization of traditional HBBs, this approach effectively reduces the interference caused by non-damage regions during the training phase, thereby improving the performance of MO-YOLOv4 in multi-class damage segmentation. To validate this perspective, a counterpart utilizing the HBB-based segmentation strategy, namely MH-YOLOv4, was established while incorporating the MSA module. During the training phase, the ground-truth masks of MO-YOLOv4 and MH-YOLOv4 were provided within the OBB and HBB, respectively. Additionally, to ensure the objectivity of this comparative experiment, both MO-YOLOv4 and MH-YOLOv4 followed the same training process stated in Section 3.2.

The average proportion of damage pixels within the bounding box during the training process is presented in Table 3. It can be observed that the OBB exhibits a higher proportion of all four types of damages compared to the proportion within the HBB. This observation further highlights the effectiveness of the OBB in tightly enclosing damages and obtaining the damage region with less background information. Consequently, when utilizing this damage region for training, MO-YOLOv4 achieves a higher performance in multi-class damage segmentation. As shown in Table 4, both the IoU and MIoU metrics of MO-YOLOv4 surpass those of MH-YOLOv4,
demonstrating the effectiveness of the OBB-based segmentation strategy for enhancing the segmentation performance of multi-class damages.

Table 3. The damage proportion within different bounding boxes during training process

<table>
<thead>
<tr>
<th>Box type</th>
<th>Concrete crack</th>
<th>Spalling</th>
<th>Rebar exposure</th>
<th>Pothole</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBB</td>
<td>10</td>
<td>45</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>OBB</td>
<td>25</td>
<td>60</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 4. Comparison of MH-YOLOv4 and MO-YOLOv4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Background (%)</th>
<th>Concrete crack (%)</th>
<th>Spalling (%)</th>
<th>Rebar exposure (%)</th>
<th>Pothole (%)</th>
<th>MIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH-YOLOv4</td>
<td>96.76</td>
<td>41.23</td>
<td>58.35</td>
<td>47.49</td>
<td>58.51</td>
<td>60.46</td>
</tr>
<tr>
<td>MO-YOLOv4</td>
<td>96.95</td>
<td>52.69</td>
<td>60.27</td>
<td>48.32</td>
<td>63.16</td>
<td>64.28</td>
</tr>
</tbody>
</table>

3.5 Comparison with the existing semantic segmentation methods

In order to comprehensively evaluate the promising performance of the proposed MO-YOLOv4, a comparison study was conducted with existing semantic segmentation methods for multi-class damages, namely FCN [22], improved FCN [45], Deeplabv3+ [46]. To ensure the objectivity of this comparative experiment, all aforementioned methods adopted an identical training process, as detailed in Section 3.2, where their respective feature extraction modules utilized pre-trained weight parameters acquired from the Pascal VOC dataset.

Upon examining the multi-class damage segmentation results, it becomes evident that both FCN and improved FCN tend to misclassify the pixels of crack-like interferences and exposed rebars as concrete crack pixels, as depicted in Fig. 6 (b) and (c), respectively. Similarly, Deeplabv3+ also exhibits erroneous segmentation by misidentifying the pixels of exposed rebars and potholes as concrete crack and spalling pixels, as shown in Fig. 6 (c) and (d), respectively. This observation indicates the inadequacy of FCN, improved FCN and Deeplabv3+ in effectively distinguishing the
differences among these types of damages, owing to their limited feature extraction ability for multi-class damages. Furthermore, all these selected semantic segmentation methods generate incomplete segmentation masks of concrete cracks compared to their ground-truth masks, as shown in Fig. 6 (a). This reflects the influence of a large amount of background information in network training of tradition semantic segmentation methods, which increases the training difficulty involved in the segmentation of concrete cracks with a relatively small scale among the four types of damages.

Conversely, by incorporating the MSA module to facilitate feature extraction for multi-class damages and implementing the OBB-based segmentation strategy to mitigate the influence of background interferences during network training, the present MO-YOLOv4 achieves the most accurate segmentation results for the four types of damages when compared to the aforementioned semantic segmentation methods, as illustrated in Fig. 6. The corresponding MIoU metric of MO-YOLOv4 reaches up to 64.28%, which represents a significant improvement of 19.62%, 13.8% and 10.49% in comparison to the metrics of FCN, improved FCN and Deeplabv3+, respectively, as presented in Table 5. These results clearly demonstrate the superior performance of MO-YOLOv4 in multi-class damage segmentation.
422 Fig. 6. Multi-class damage segmentation results of different semantic segmentation methods: (a) concrete crack, (b) spalling, (c) rebar exposure, (d) pothole.

424 Table 5. Comparison of different semantic segmentation methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Background</th>
<th>Concrete crack</th>
<th>Spalling</th>
<th>Rebar exposure</th>
<th>Pothole</th>
<th>MIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>96.05</td>
<td>33.65</td>
<td>35.42</td>
<td>14.17</td>
<td>44.01</td>
<td>44.66</td>
</tr>
<tr>
<td>Improved FCN</td>
<td>96.33</td>
<td>34.50</td>
<td>55.28</td>
<td>17.34</td>
<td>48.97</td>
<td>50.48</td>
</tr>
<tr>
<td>Deeplabv3+</td>
<td>96.23</td>
<td>46.36</td>
<td>43.57</td>
<td>33.44</td>
<td>49.36</td>
<td>53.79</td>
</tr>
<tr>
<td>MO-YOLOv4</td>
<td>96.95</td>
<td>52.69</td>
<td>60.27</td>
<td>48.32</td>
<td>63.16</td>
<td>64.28</td>
</tr>
</tbody>
</table>

4. Conclusions

The segmentation task of multi-class bridge damages poses two major challenges, i.e., the substantial variations in the scale of different damages and the susceptibility of network training to background interferences, which greatly impede the segmentation performance of existing semantic segmentation methods for multi-class damages. To tackle these challenges, this paper proposes a novel method for multi-class damage segmentation based on the widely recognized YOLOv4 architecture, which is referred to as MO-YOLOv4. Specifically, the proposed method integrates the MSA module to augment its feature extraction capability in effectively addressing multi-class damages characterized by significant variations in their scales. Subsequently, the proposed method further introduces the OBB-based segmentation strategy to reduce the segmentation area of damages, effectively mitigating the influence of background interferences during network training.

The experimental results indicate that MO-YOLOv4 achieves an MIoU metric of 64.28% for four types of damages, including concrete cracks, spalling, rebar exposure and potholes. This MIoU is 3.71% and 1.93% higher than the MIoUs achieved by CO-YOLOv4 and EO-YOLOv4, respectively. These results effectively showcase the promising feature extraction ability of the MSA module in multi-class damage segmentation, in contrast to the limited capabilities of the CBA and ECA module with
single-scale receptive fields. Furthermore, this MIoU realized by MO-YOLOv4 surpasses that of MH-YOLOv4 by 3.82%, where the latter employs traditional HBBs for the identification of damage regions. This distinction underscores the ability of OBBs to tightly enclose damages compared to that of HBBs, resulting in segmentation regions with clearer damage characteristics and fewer background interferences during training. When compared to existing semantic segmentation methods, namely, FCN, improved FCN and Deeplabv3+, the present MO-YOLOv4, incorporating the MSA module and the OBB-based segmentation strategy, achieves the most accurate segmentation results for the four types of damages. In particular, the MIoU of MO-YOLOv4 exceeds the MIoUs of FCN, improved FCN and Deeplabv3+ by up to 19.62%, 13.8% and 10.49%, respectively. All of the aforementioned experimental results effectively demonstrate the superior performance exhibited by MO-YOLOv4 regarding multi-class damage segmentation.

However, due to the adoption of the large-scale architecture and complex segmentation strategy, MO-YOLOv4 requires approximately two to three times the training time compared to traditional semantic segmentation methods. Considering that efficiency plays a crucial role in the development of automatic bridge inspection methods, future research endeavors will explore promising techniques to develop lightweight networks for MO-YOLOv4, such as knowledge distillation [47] and depthwise separable convolution [48]. These techniques aim to further enhance the efficiency and effectiveness of MO-YOLOv4. Furthermore, comprehensive experiments will be conducted in more complex on-site bridge environments to evaluate the robustness and applicability of MO-YOLOv4.

**Acknowledgements**

The research is funded by the National Natural Science Foundation of China (52308314), Guangdong Basic and Applied Basic Research Foundation (2022A1515010174, 2023A1515030169), Guangzhou Science and Technology...
Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References


CapsNet of feature fusion. Construction and Building Materials, 324 (2022) 126719,

stacking ensemble learning of multiple semantic segmentation networks. Sensor, 22 (2022)


enhancement for detection of targets with small image sizes. Pattern Recognition Letters, 166

multi-scale residual attention. 2021 International Joint Conference on Neural Networks
(IJCNN), 2021, https://doi.org/10.1109/IJCNN52387.2021.9534200.


module for object detection in remote sensing images. IEEE Access, 10 (2022) 87266-87291,
https://doi.org/10.1109/ACCESS.2022.3199368.

[31] F. Yu, V. Koltun. Multi-scale context aggregation by dilated convolutions,


[34] RoLabelImg, 2020, available at: https://github.com/roLabelImg-master.


[40] Pytorch, available at: https://pytorch.org/.


