Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering) 2023 24(3):243-256
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.A2200175

Adaptive cropping shallow attention network for defect detection of bridge girder steel using unmanned aerial vehicle images

Zonghan MU[1,2], Yong QIN[1✉], Chongchong YU[3], Yunpeng WU[4], Zhipeng WANG[1], Huaizhi YANG[5], Yonghui HUANG[5]

[1]*State Key Lab of Rail Traffic Control & Safety, Beijing Jiaotong University, Beijing 100091, China*
[2]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100091, China*
[3]*School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China*
[4]*School of Safety Engineering and Emergency Management, Shijiazhuang Tiedao University, Shijiazhuang 050043, China*
[5]*Beijing-Shanghai High Speed Railway Co., Ltd., Beijing 100038, China*

Abstract: Bridges are an important part of railway infrastructure and need regular inspection and maintenance. Using unmanned aerial vehicle (UAV) technology to inspect railway infrastructure is an active research issue. However, due to the large size of UAV images, flight distance, and height changes, the object scale changes dramatically. At the same time, the elements of interest in railway bridges, such as bolts and corrosion, are small and dense objects, and the sample data set is seriously unbalanced, posing great challenges to the accurate detection of defects. In this paper, an adaptive cropping shallow attention network (ACSANet) is proposed, which includes an adaptive cropping strategy for large UAV images and a shallow attention network for small object detection in limited samples. To enhance the accuracy and generalization of the model, the shallow attention network model integrates a coordinate attention (CA) mechanism module and an alpha intersection over union (α-IOU) loss function, and then carries out defect detection on the bolts, steel surfaces, and railings of railway bridges. The test results show that the ACSANet model outperforms the YOLOv5s model using adaptive cropping strategy in terms of the total mAP (an evaluation index) and missing bolt mAP by 5% and 30%, respectively. Also, compared with the YOLOv5s model that adopts the common cropping strategy, the total mAP and missing bolt mAP are improved by 10% and 60%, respectively. Compared with the YOLOv5s model without any cropping strategy, the total mAP and missing bolt mAP are improved by 40% and 67%, respectively.

Key words: Railway; Bridge; Unmanned aerial vehicle (UAV) image; Small object detection; Defect detection

1 Introduction

As the speed of trains and the density of the railway network increase, safety risks likewise increase. As catastrophic accidents are increasingly caused by the failure of railway facilities, safety problems along railway lines are of particular concern (Arivazhagan et al., 2015). It is therefore increasingly important to carry out timely checking for equipment faults and potential safety hazards along the line (Wu et al., 2022). As important railway infrastructure, bridges are often corroded due to long-term erosion of the steel structure and are missing high-strength bolts (Ramana et al., 2017; Long et al., 2021). Therefore, the inspection of railway bridges is a major task and the corrosion of bolts makes the disassembly of equipment very difficult. These factors have a huge negative impact on the stability of the bridge and on its later safety management (Wang JK et al., 2021; Wang ZQ et al., 2021).

At present, manual inspection has many limitations, such as low detection efficiency, more visual blind areas, high risk, and inconsistent discrimination standards. Unmanned aerial vehicle (UAV) inspection technology has the advantages of flexibility, efficiency, and low cost, and has been widely used in various infrastructure inspections at home and abroad (Duque et al., 2018; Morgenthal et al., 2019). In view of the limitations of manual inspection such as long span along railway lines and scattered inspection sites, UAV and other related equipment are used to take pictures,

✉ Yong QIN, yqin@bjtu.edu.cn
Zonghan MU, https://orcid.org/0000-0003-0198-434X

Received Mar. 29, 2022; Revision accepted July 13, 2022; Crosschecked Feb. 3, 2023

as shown in Fig. 1. Then, manual visual recognition gradually became a supplementary measure for manual inspection. However, due to people's different subjectivity and the negative impact of repetitive work, the efficiency and accuracy of manual reading of pictures are often not guaranteed. Therefore, with the rise of computer vision and image detection, researchers began to pay attention to railway infrastructure defect detection methods based on computer vision (Shao et al., 2020).

Fig. 1 UAV image of steel structure of railway bridge girder

Much research has been done on industrial defect detection based on computer imaging. Cha et al. (2018) proposed a structural visual detection method based on faster region-based convolutional neural network (Faster R-CNN) to detect the damage of concrete cracks, reinforcement corrosion, delamination, and bolt corrosion of bridges. Kang and Cha (2018) proposed a UAV autonomous method using ultrasonic beacons instead of Global Position System (GPS), a deep convolutional neural network (DCNN) for damage detection, and a geotagging method for damage location, with specificity and sensitivity exceeding 90%, to solve the problem of expensive and inefficient artificial visual monitoring and the blind areas in data collected by UAV. Cha et al. (2017) proposed a convolutional neural network (CNN) based on sliding windows, which combined two kinds of redundant paths of sliding windows to accomplish full image coverage, and realized defect location on crack surface without calculating defect features. Ali et al. (2021) proposed an autonomous UAV system integrating an improved Faster R-CNN for identifying various types of structural damage and mapping detected damage to a GPS-denied environment. This method uses a real-time streaming protocol and multi-processing to significantly reduce the number of false positives. Recently, many scholars have used a segmentation network to locate the damage at pixel level in order to obtain more accurate detection results. Kang and Cha (2021) proposed a novel semantic transformer representation network (STRNet) consisting of a squeeze and excitation attention-based encoder, a multi-head attention-based decoder, coarse up-sampling, a focal-Tversky loss function, and a learnable swish activation function, as well as a method for evaluating the complexity of image scenes. The final detection accuracy of that STRNet model is over 90%. Choi and Cha (2020) proposed a CNN consisting of standard convolution, densely connected separable convolution modules, a modified atrous spatial pyramid pooling module, and a decoder module. This CNN effectively eliminated a wide range of complex backgrounds and crack-like features, while having a very small number of parameters and a fast processing speed. Kang et al. (2020) improved the Faster R-CNN algorithm, used a modified tubularity flow field (TuFF) algorithm to locate the crack region, and used a modified distance transform method (DTM) to measure the crack thickness and length from the perspective of pixel measurement; the average accuracy of crack damage detection reached 95%.

In the field of railway infrastructure damage detection, a two-stage network structure model such as Faster R-CNN was introduced to carry out defect detection on railway infrastructure. It can overcome the limitations of manual inspection and has very high accuracy. It can be used as a supplementary means for daily inspection and can even replace manual inspection in some cases. For example, Liu et al. (2019) improved the Faster R-CNN so that it could locate and detect the rotating targets in bird-prevention and fasteners on catenary support devices. Chen P et al. (2019) also used the improved Faster R-CNN algorithm to detect the defects of rail fasteners. Wu et al. (2018) proposed local Weber-like contrast (LWLC) and gray stretch maximum entropy (GSME) threshold segmentation methods. After railway image enhancement, the UAV rail image was processed with gray stretching and denoising, and the optimal segmentation threshold was selected for defect detection. However, although the two-stage algorithm pursues detection accuracy, it ignores detection efficiency and is limited by the requirements of computer equipment, so it cannot meet the requirements of real-time or time-limited detection.

In real-time or time-limited defect object detection, single shot multibox detector (SSD) (Liu et al.,

2016), You Only Look Once (YOLO) (Redmon et al., 2016), and other single-stage network models are gradually becoming widely used in railway infrastructure defect detection and can maintain an accuracy not inferior to part of the two-stage network, but also have extremely fast training and detection speeds, and can be more convenient in deployment and application. For example, Chen et al. (2018) applied DCNN to fastener defect detection on high-speed railway lines and, combined with the classical single-stage network method, constructed a cascade detection network consisting of two detectors and a classifier. Tao et al. (2018) designed a cascade autoencoder (CASAE) to locate the defect region by threshold segmentation based on the different responses of the encoder to the normal region and the defect region. Then, the cropped graph block of the defect area was input to the defect classification network, and finally the defect category was output. Chen Q et al. (2019) proposed a catenary U-bolt missing identification network based on YOLOv3 and squeeze-and-excitation networks (SE). Jia and Luo (2019) proposed a crack image detection and parameter measurement method by combining digital image processing with a CNN. However, the detection object in railway bridge girder steel structure is usually small, and a single-stage network is not friendly to the detection of such a small object because of simultaneous positioning and classification. Moreover, in order to improve the detection accuracy, a single-stage network often adopts a network model with deeper structure, which can obtain stronger semantic information, but also loses higher resolution and reduces the detection accuracy for small objects.

Therefore, experts at home and abroad have done much work on the problem of small object detection. Some people have improved the accuracy of small object detection by improving the network structure. Tang et al. (2018) introduced a new context-assisted network framework to deal with small objects that are difficult to detect, which can make full use of the semantic information around small objects. Noh et al. (2019) designed a high resolution target feature extraction network based on generative adversarial network (GAN) which shared parameters with a low resolution feature extraction network, reduced the number of parameters, and replaced the convolution layer with empty convolution to expand the receptive field. Wei et al. (2022) proposed the Rotated Position Sensitive RoI Align (RPS-RoI-Align) module to improve the quality of candidate regions for objects with intensive detection in aerial images, and extracted rotation-invariant features from them to promote subsequent classification and regression. Yang et al. (2021) proposed Query Det, which used a new query mechanism, cascade sparse query (CSQ), to accelerate a dense object detector based on a feature pyramid.

However, the result of improving the network structure is often an increase in the complexity of the network model which reduces its detection efficiency, and is contrary to the high efficiency required by real-time or time-limited railway defect detection. Therefore, some scholars consider improving the detection effect of small objects by processing images. van Etten (2018) cropped the image for detection of small objects in super-resolution satellite images. Kisantal et al. (2019) oversampled the images containing small objects, and augmented each image containing small objects by copying and pasting small objects many times. Chen et al. (2020) adjusted small object images to smaller sizes, and then spliced them into the same size as conventional images. Loss information was used as feedback to guide the next iteration update. Liu et al. (2021) spliced the images containing more small objects into one image through the feedback of the loss contribution rate of small objects, so as to produce more training data of small objects and thus improve the overall detection accuracy. However, in the data set of a steel structure UAV image of a railway bridge girder, the number of various objects to be detected is extremely unbalanced, and the method of splicing and copying will lead to over-fitting and poor detection effects. Additionally, in the process of UAV shooting, the scale of the same target to be detected in a batch of pictures will also change greatly because of the change of distance, height, focal length, etc., so a common cropping strategy is not suitable.

In view of the difficulty of UAV image processing of the steel structure of a railway bridge girder, the adaptive cropping shallow attention network (ACSANet) is proposed.

The main contributions of this paper are as follows:

1. An adaptive image cropping method is proposed, which can adaptively adjust the image cropping size and cropping overlapping area according to the specific situation of the image, which can eliminate

the negative influences arising from the UAV shooting distance and the unfixed focal length, and so improve the detection effect of small targets.

2. Based on the characteristics of the objects to be detected, a shallow attention network is proposed to make the model pay more attention to the detection objects, so that the corrosion area can be detected more easily.

3. The coordinate attention (CA) mechanism module (Hou et al., 2021) is integrated into the shallow attention network to help the network find the defect areas in a wide range of UAV shooting scenes.

4. The alpha intersection over union (α-IOU) loss function (He et al., 2021) is integrated into the shallow attention network to improve the model detection accuracy on the small sample data set of a railway bridge girder steel structure.

The proposed ASCANet, adaptive cropping strategy, and shallow attention network, including attention mechanism and loss function, are described in Section 2 respectively. Experimental results are shown in Section 3. The conclusions are described in the last section.

2 Methodology

2.1 ACSANet

Due to its advantages in flexibility and speed, the YOLOv5s network model is relatively consistent with the wish for fast and accurate railway bridge girder inspection. The YOLOv5s model is very small, and its weight file is only 1/3 of the YOLOv5m size and 1/12 of the YOLOv5x size. Also, the training time is the shortest among the various models. In addition, considering the model's deployment on a UAV, lightness is very important. The improved model based on YOLOv5s can maximize the advantages of light weight and speed, and so it is selected as the benchmark network for railway bridge girder inspection using UAV images. It is composed of CSPDarknet53 and spatial pyramid pooling (SPP) modules as the backbone network. The structure of feature pyramid network (FPN) and path aggregation network (PANet) is adopted in the neck, followed by three YOLO detection heads (Redmon et al., 2016). At the same time, many data augmentation methods are used in the input of the network, including Zoom, Flip, Mixup (Zhang et al., 2018), Mosaic (Bochkovskiy et al., 2020), and hue-saturation-value (HSV) spatial data augmentation. Mosaic, in particular, is an important method for improving the accuracy of small target detection. The Mosaic schematic diagram of a steel bridge structure is shown in Fig. 2.

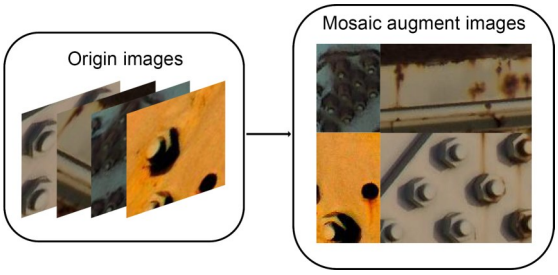

Fig. 2 Schematic diagram of Mosaic of bridge steel structure

However, in the detection process of a steel structure UAV image of railway bridge girder, despite the data augmentation, the detection effect is still not good. In particular, the original YOLOv5s network model has poor detection effect on targets such as a missing bolt or rust in the railings of the steel structure of a railway bridge girder. From the perspective of UAV image processing, the main reasons for the poor effect are as follows:

(1) In the process of UAV shooting, the shooting distance and focal length are not fixed, so that the scale of the same object varies greatly.

(2) The data set is small, and the number of defect samples such as corrosion and missing bridge bolts is very small, so it is difficult to balance the sample set.

(3) The corrosion of the steel structure and railings of the bridge is similar to the complex background around the bridge, making it difficult to distinguish.

(4) The detection objects are small and intensive, and the common one-stage target detection model is often ineffective in detecting small object images directly.

Therefore, we made targeted improvements to the original YOLOv5s network model, including an adaptive cropping strategy for large size images and a shallow attention network for detection of simple features and small samples. We added a CA mechanism module and α-IOU loss function to the shallow attention network. The generalization of the small data set of UAV images for the steel structure of the railway bridge girder was further enhanced and its detection accuracy was improved.

ACSANet is developed based on the above improvements, and its network structure is shown in Fig. 3. During data preprocessing, Mosaic, Mixup, HSV color enhancement, Flip, and other data augmentation methods were applied. Due to the small number of missing bolt samples, Photoshop software was also used for data synthesis, and the synthesized data was confirmed as valid by professionals. Then, adaptive cropping and data enhancement images were input into the shallow attention network to eliminate overfitting and sample imbalanced issues.

2.2 Adaptive cropping

For small target detection of large UAV images, cropping the large UAV images is a simple and effective method, but the traditional cropping strategy often simply crops the image or only considers the cut-off of the target to be detected due to cropping, and thus adopts an overlapping cropping strategy (van Etten, 2018). However, in the process of photographing the steel structures of railway bridges, the scale of the same target to be detected due to the focal length change of the UAV image also changes very significantly, so the traditional cropping strategy is not suitable. When the overlapping cropping strategy is used, the smaller overlap size does not guarantee that the targets of different scale sizes in a batch of images are not cut and can be trained completely, and the larger overlap size reduces the efficiency of detection.

To address the shortcomings of the traditional cropping strategy and the overlapping cropping strategy, an image adaptive cropping strategy is proposed, and the details are as follows:

(1) Uncropped image is defined as p_i (i=1, 2, ⋯, n), and the label corresponding to the uncropped image is q_i (i=1, 2, ⋯, n).

(2) The detection target with the highest proportion in q_i is determined as the main target to be detected, which is represented by τ.

(3) Through preliminary pre-training, the optimal detection ratio α is selected according to the size S_1 of τ and S_2 of cropped image, and α=S_1/S_2.

(4) The average width w, height h, and size S_{ave} of τ in the training data set are determined.

(5) Using w and h of τ, the overlapping width w_c and height h_c of each picture cropping are determined.

(6) The cropping number C_n is determined by using S_{ave} and α.

The adaptive cropping strategy can adapt to the detection targets of different scale sizes in a batch of images and dynamically assign different cropping strategies for each image; these strategies include the number of crops and the size of the cropping overlap region.

The adaptive cropping strategy is determined according to the ratio of the main target to be detected and the cropped image size; for example, the target to be detected in the steel structure of railway bridge girder is mainly nuts, so the nuts are selected as the main target. The actual ratio of the main target to the cropped picture is determined by experimental comparison of the detection accuracy. After the experiment, it is found that this strategy can better improve the detection accuracy. The determination of the size of the cropping overlap area integrates the efficiency of the detection target and model training, and the formulae for calculating the width and height of the overlap are shown in Eqs. (1) and (2), respectively:

$$w_c = \frac{2 \times \sum_{i=1}^{n} (x_{max} - x_{min})}{n}, \tag{1}$$

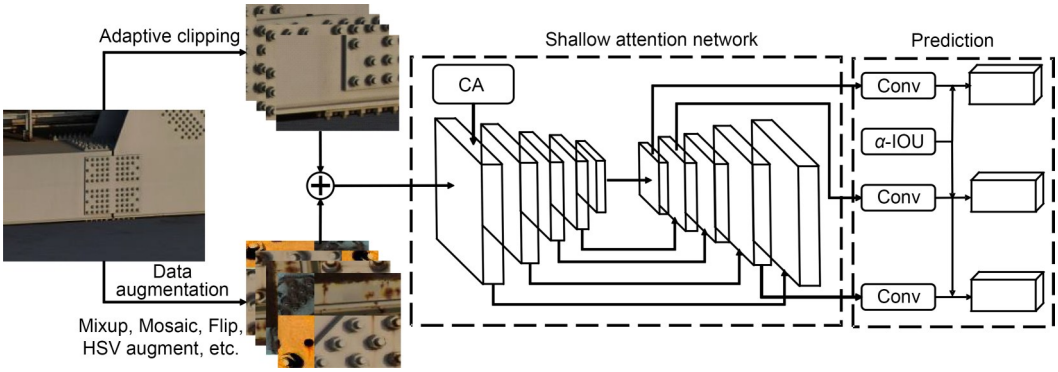

Fig. 3 Architecture of the ACSANet. Conv represents the convolution

$$h_c = \frac{2 \times \sum_{i=1}^{n} (y_{max} - y_{min})}{n}, \quad (2)$$

where x_{max}, x_{min}, y_{max}, and y_{min} are respectively the abscissa and ordinate of the maximum and minimum of the main target tag to be detected, respectively; n is the number of main targets to be detected.

The number of croppings is determined according to the average size of the main target to be detected and the ratio of target to image training, and the formula for calculating the number of croppings is shown in Eq. (3):

$$C_n = \frac{1}{n} \sum_{i=1}^{n} \frac{WH\alpha}{(x_{max} - x_{min})(y_{max} - y_{min})}, \quad (3)$$

where C_n is the number of image croppings; W and H represent the width and height of the original image; α is the size ratio between the main target to be detected and the input image.

The comparison between the image adaptive cropping strategy and the overlapping cropping strategy is shown in Fig. 4. Taking Fig. 4a as an example, the small pictures on the left and right show the cropped image effect, and the small picture on the bottom shows the overlapping part. The adaptive cropping strategy can automatically adjust the size of the cropping frame and the overlap size according to the images, as shown in Figs. 4a and 4c. By contrast, the overlapping cropping strategy can only fix the size of the cropping frame and the overlap size, resulting in some images with too small cropping frame overlap area, so the targets at the segmentation boundary cannot all participate in the training, as shown in Fig. 4b, or for some pictures, cause the cut box overlap area to be too large, as shown in Fig. 4d.

2.3 Shallow attention network

In (Zhu et al., 2021), the method of increasing detection layers was used to improve the detection of small objects, but for the image data of railway bridge girder steel structure, this is not applicable, because some of the features of the target to be detected are relatively simple and too deep a detection network tends to ignore such information. In addition, the deep network fusion of other features may affect the final detection results. Therefore, on the basis of the YOLOv5s detection network, instead of changing the number of detection heads, the size of the detection heads is changed. The features of the backbone network are enhanced from the shallowest point and, at the same time, the detection heads targeting shallow layers are increased. After the 25th layer, the sampling of the feature map continues, and the map continues to expand. At the same time, the feature image of 64×64

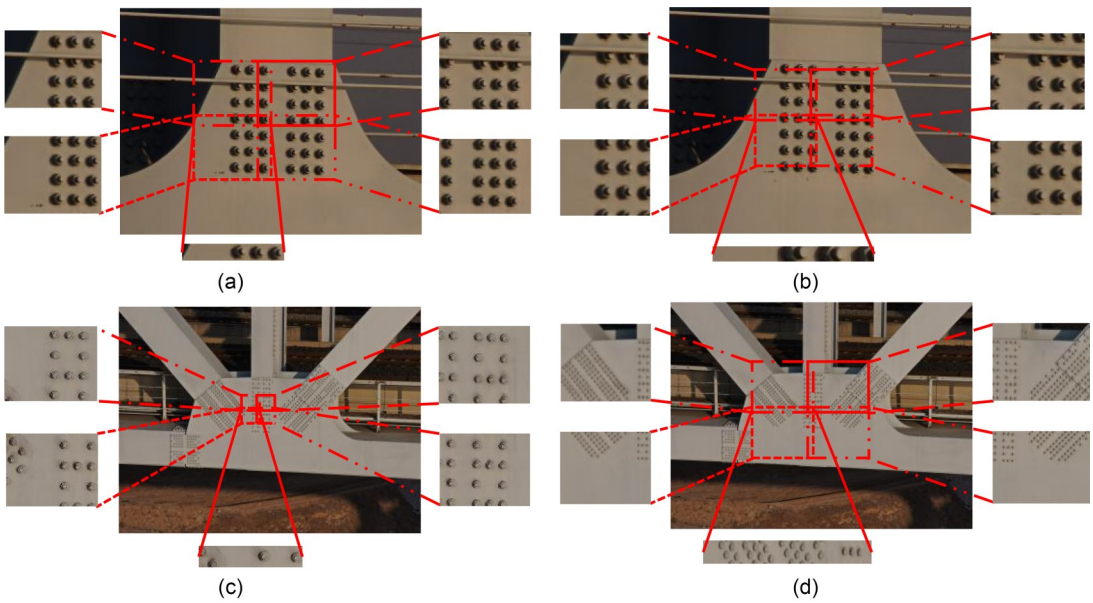

Fig. 4 Comparison diagram of adaptive cropping strategy and overlapping cropping strategy: (a) adaptive cropping of large detection targets; (b) overlapping cropping of large detection targets; (c) adaptive cropping of small detection targets; (d) overlapping cropping of small detection targets

obtained at the 28th layer is spliced and fused with the feature image of the shallowest layer in the backbone network, thus obtaining a large feature map for small target detection.

In addition, for some targets to be detected with simpler features, such as rusting targets like steel structures and railings, the model also adds the CA mechanism module to the shallow network to enhance the detection of rusting targets. The shallow attention network is shown in Fig. 5.

2.4 CA+C3_*N*

In the study (Hou et al., 2021), CA is compared with SE (Hu et al., 2020) and convolutional block attention module (CBAM) (Woo et al., 2018) attention mechanism, and the model performance of the integrated CA mechanism is improved considerably and has better robustness than other attention mechanisms. This verifies that the module is effective in the detection improvement of the model, so it is considered appropriate to add it to the shallow attention network structure. The CA module first averages the horizontal and vertical directions of the input feature map separately to obtain a pair of one-dimensional feature codes, compresses the channels in the spatial dimension by splicing and convolution, and then encodes the spatial information in the vertical and horizontal directions by batch normalization and nonlinear activation functions. Next, it obtains the same number of channels as the input feature map by two convolution and activation operations and, finally, normalizes and weights to obtain the output feature map. The shallow attention network integrates the CA mechanism module and the backbone network C3 module of the original YOLOv5s model. The CA+C3_*N* network structure is shown in Fig. 6.

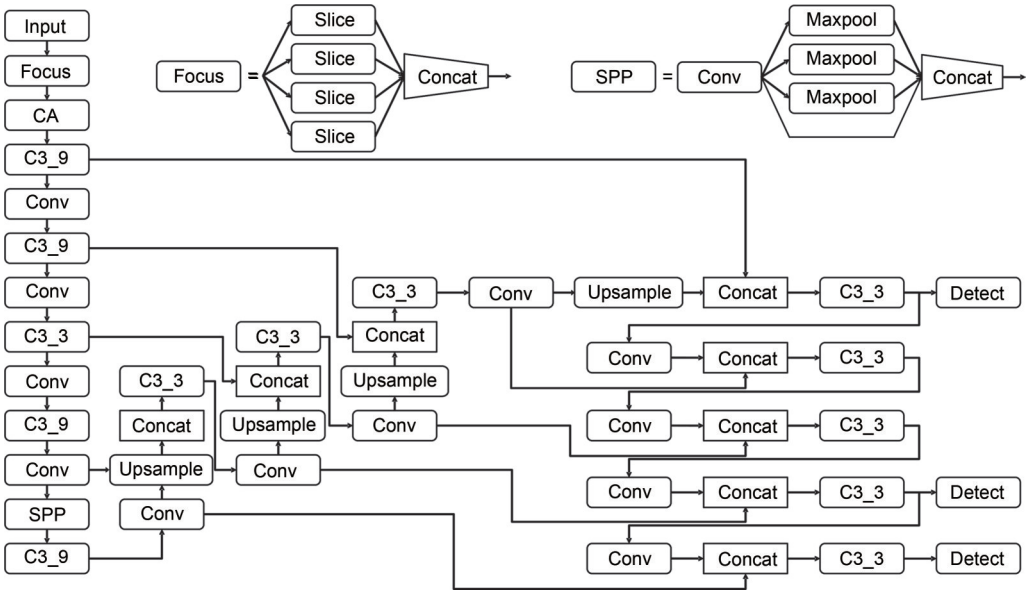

Fig. 5 Architecture of the shallow attention network

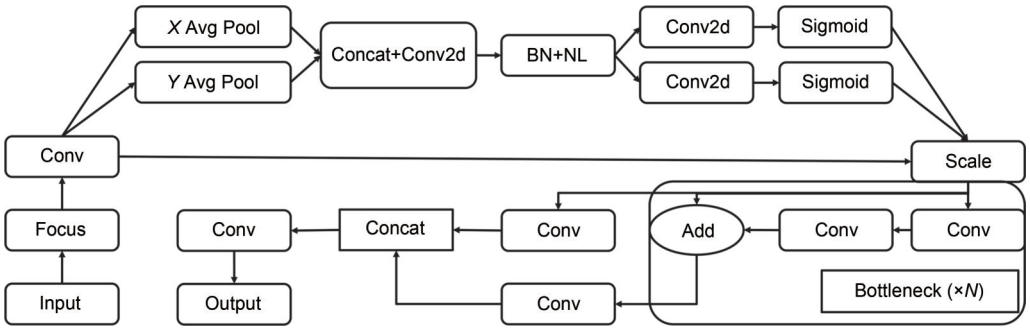

Fig. 6 CA+C3_*N* network structure. BN: batch normal; NL: non-linear

In the images taken by UAV, there is complex background information which affects the accuracy of target detection of the model. Adding a CA mechanism module can enhance the attention area, which can resist complex information in the training process and focus detection on useful target objects.

2.5 α-IOU

The α-IOU loss function is the introduction of the power transform to the existing intersection over union (IOU) loss function (Rahman and Yang, 2016), which does not introduce additional parameters and does not increase the training and inference time. α-IOU applies the Box-Cox transform (Box and Cox, 1964) to IOU loss L_{IOU}=1−IOU, and generalizes it to power IOU loss: $L_{(\alpha\text{-IOU})}=(1-\text{IOU})^{\alpha}/\alpha$, $\alpha>0$, denoted as α-IOU. By adjusting the additional power regularization term α to obtain more accurate boundary box regression and target detection, it can be generalized to a more general form that can generalize existing IOU-based losses, including GIOU (Rezatofighi et al., 2019), DIOU (Zheng et al., 2019), and CIOU (Zhang et al., 2022).

Through multi-target detection benchmarks and model experiments, α-IOU losses can significantly exceed existing IOU based losses (He et al., 2021), and, by adjusting α, the detector has greater flexibility in achieving different levels of box regression accuracy and is more robust to small data sets and noise.

The number of pictures of the data set used for a railway bridge girder steel structure detection is small and the number of samples is unbalanced, which is exactly in line with the conditions for the application of an α-IOU loss function, so the strategy of α-IOU loss is considered in the shallow attention network structure. Later experimental results show that α-IOU can better detect the steel structure of a railway bridge girder.

3 Results and discussion

3.1 Implementation details

The experiment was carried out on Ubuntu20.04 system with NVIDIA GeForce RTX2080 graphics card. The experimental environment was Python3.8, CUDA10.1, CUDNN8.0.2, and Pytorch1.9.1. The stochastic gradient descent (SGD) optimizer was used in training, and the initial learning rate was 0.01 and the momentum value was 0.937. Since a single graphic processing unit (GPU) was used for training, the BatchSize was adjusted to 8 and the Epoch was 300. The evaluation indexes include Precision, Recall, and mAP. mAP is the average value of AP values of all categories. AP is the average precision value of a class of detection targets, and is calculated by:

$$P=\frac{\text{TP}}{\text{TP}+\text{FP}}\times 100\%, \tag{4}$$

$$R=\frac{\text{TP}}{\text{TP}+\text{FN}}\times 100\%, \tag{5}$$

$$\text{AP}=\int_0^1 P(R)\text{d}R, \tag{6}$$

where Precision (P) indicates the proportion of the data set in which the predicted outcome is a positive example, and Recall (R) indicates the proportion of the data set in which the true outcome is a positive example, and the number of predicted outcomes is correct. False positive (FP) means the number of samples judged to be positive but in fact negative, true positive (TP) means the number of samples judged to be positive and in fact positive, and false negative (FN) means the number of samples judged to be negative but in fact positive.

3.2 Data structure

Using a DJI MATRICE 300 RTK UAV equipped with DJI Zenmuse H20 aerial camera, lateral shots of railway bridge girder steel structures were hovered at a safe distance along the railroad line. The steel structure of the bridge was divided into four parts and photographed successively; the acquisition area is shown in Fig. 7.

Fig. 7 UAV data acquisition diagram

In this paper, the samples are labeled as Pascal VOC, which is one of the standard data formats commonly used in the field of object detection. As shown

in Fig. 8, a total of 70 railway bridge girder images collected by UAV are annotated in six detection categories.

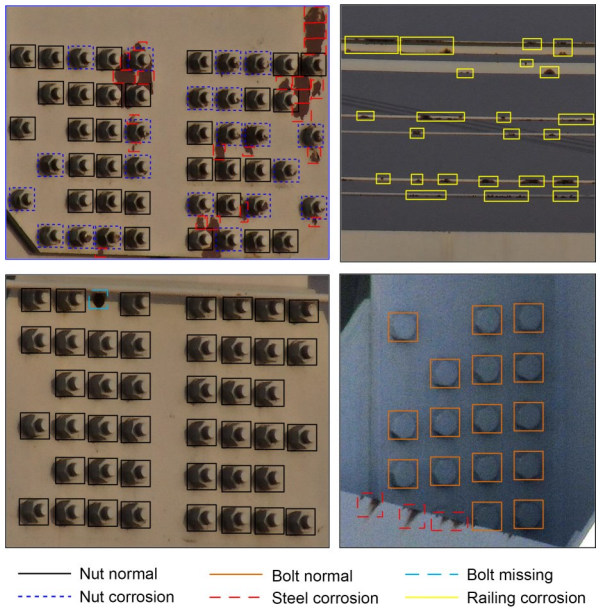

Fig. 8 UAV images-based detection objects of railway bridge girder steel structure

The position information and category information in the images were saved using '.xml' files, while the training and test sets were divided according to the ratio of 6:1. The sample set structures of UAV images of railway bridge girder steel structure are shown in Table 1. Through this annotation method, the sample set of steel structure image of railway bridge girder can be preliminarily constructed.

Table 1 Sample set structures of UAV images of railway bridge girder steel structure

Size of images	Class	Number of images trained	Number of objects trained	Number of images tested	Number of objects tested
5184×3888	Nut normal	60	10323	10	1255
	Bolt normal	32	458	2	36
	Bolt missing	13	17	6	6
	Nut corrosion	52	1025	6	212
	Steel corrosion	42	386	6	104
	Railing corrosion	33	277	5	64

3.3 Implementation details

The scale ratio between the image and the main target (nut) is from 20 to 80, and the adaptive cropping is performed at an interval of 10 for training and testing. Fig. 9 shows the test accuracy (mAP) and training time (time) of different cropping scales, and we can see that the comprehensive training efficiency and test results are most suitable when the ratio of image to main target is 50:1. Therefore, the next experiment will be conducted with a 50:1 ratio of the image to the main target.

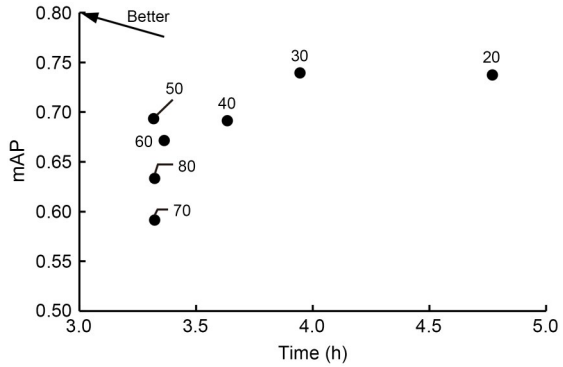

Fig. 9 Adaptive cropping mAP comparison of different cropping scales

3.4 Comparison of attention mechanisms

In the comparison of the effects of attention mechanism, the cropping scale is 50:1. The total prediction time and average prediction time of each image under different attention mechanisms are shown in Fig. 10.

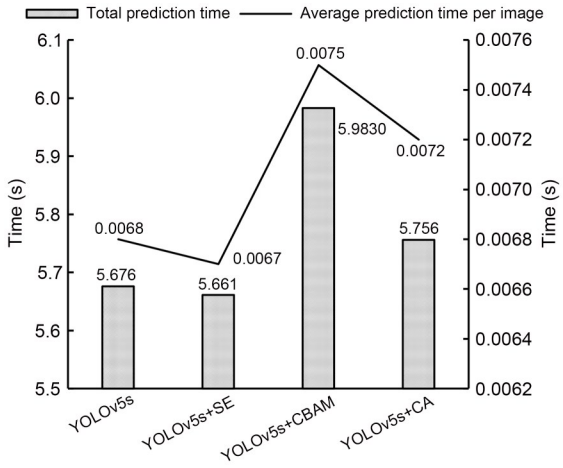

Fig. 10 Total prediction time and average prediction time of each image under different attention mechanisms

The results of mAP, Precision, Recall, and inference speed of the original YOLOv5s model and different attention integration models in the bridge steel structure sample set are shown in Table 2.

Table 2 Detection results of integrated models of different attention mechanisms

Method	Image size	mAP	Precision	Recall	FPS
YOLOv5s	512×512	0.677	0.755	0.597	204
YOLOv5s+SE	512×512	0.623	0.737	0.560	204
YOLOv5s+CBAM	512×512	0.671	0.727	0.633	185
YOLOv5s+CA	512×512	0.712	0.761	0.587	196

FPS: frames per second

The accuracy of each class under different attention mechanisms is shown in Fig. 11. By comparing the detection effects in Fig. 11, it can be found that YOLOv5s+CA is significantly better than the original YOLOv5s model. In particular, the detection accuracy for missing bolts is improved by more than 20%, which is very significant. However, in terms of the average prediction time per image, that for the YOLOv5s+CA model is only 0.0004 s greater than that for the YOLOv5s.

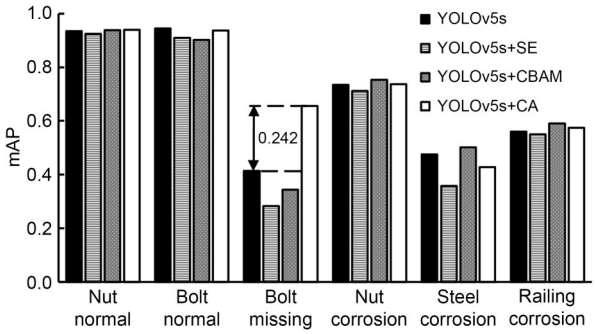

Fig. 11 Accuracy of each class under different attentional mechanisms

3.5 Precision comparison

The cropping scale is 50:1, and Fig. 12 shows the loss curves of YOLOv5s and its different improved models. It can be seen that all models have achieved convergence without overfitting, and ACSANet has the best convergence effect.

Fig. 13 shows the accuracy comparison of different improved models for each class as well as the mAP curve graph.

The test results of the improved model for the steel structure of a railway bridge girder in different

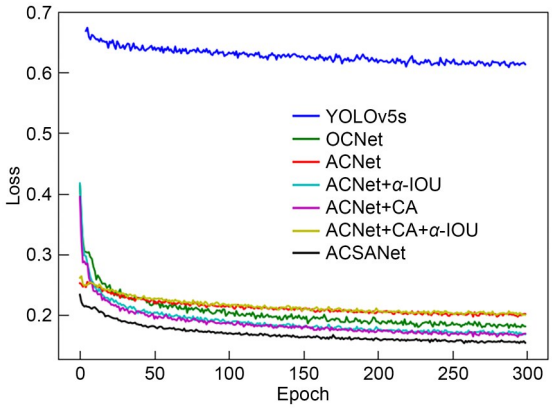

Fig. 12 Loss curves of each model

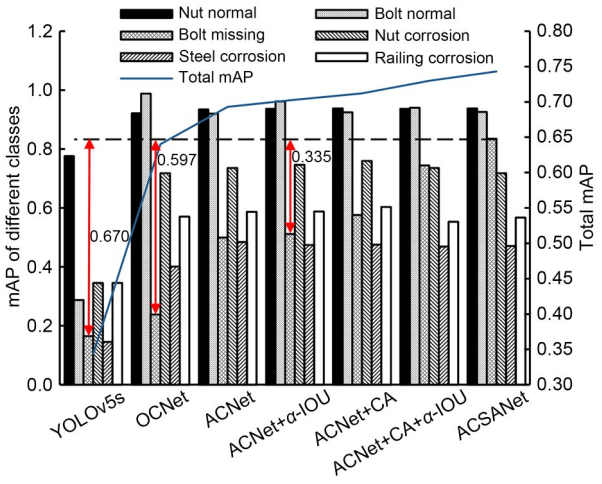

Fig. 13 Accuracy comparison of different improved models and mAP curves

stages are compared in Table 3. OCNet and ACNet denote the network structures with overlapping cropping strategy and adaptive cropping strategy, respectively.

According to Table 3 and Fig. 12, the proposed ACSANet model outperforms the YOLOv5s model using adaptive cropping strategy in terms of the total mAP and missing bolt mAP by 5% and 30%, respectively. Compared with the YOLOv5s model that adopts the common cropping strategy, the total mAP and missing bolt mAP are improved by 10% and 60%, respectively. Compared with the YOLOv5s model without any cropping strategy, the total mAP and missing bolt mAP are improved by 40% and 67%, respectively.

Fig. 14 shows the detection effects of YOLOv5s and ACSANet on the steel structure of a railway bridge girder. Fig. 14a is the detection results for YOLOv5s. The purple circle indicates the detection error (the normal nut is detected as a corroded nut), the green circle

Table 3 Accuracy comparison of YOLOv5s and improved model

Method	Image size	mAP	Precision	Recall	FPS
YOLOv5s	512×512	0.344	0.678	0.353	178
OCNet	512×512	0.640	0.767	0.568	185
ACNet	512×512	0.693	0.829	0.549	208
ACNet+α-IOU	512×512	0.703	0.779	0.600	188
ACNet+CA	512×512	0.712	0.757	0.621	196
ACNet+CA+α-IOU	512×512	0.730	0.719	0.602	196
ACSANet	512×512	0.743	0.761	0.600	66

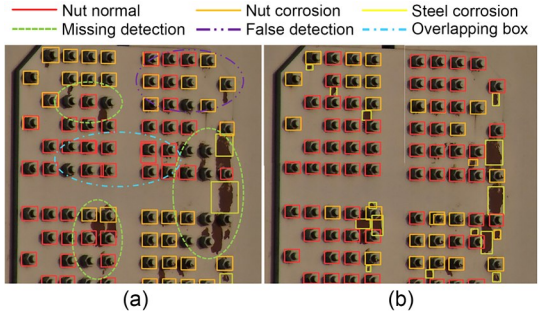

Fig. 14 Detection effect comparison of YOLOv5s (a) and ACSANet (b) on steel structure of railway bridge girder

indicates the detection omission (the normal nut, corroded nut, and corroded steel are missed), and the blue circle represents the duplicate box. Compared with the detection effects of the YOLOv5s model, the ACSANet model in Fig. 14b detects the UAV image of the steel structure of a railway bridge girder better.

The local image visualization and overall image visualization results of ACSANet detection are shown in Figs. 15 and 16, respectively.

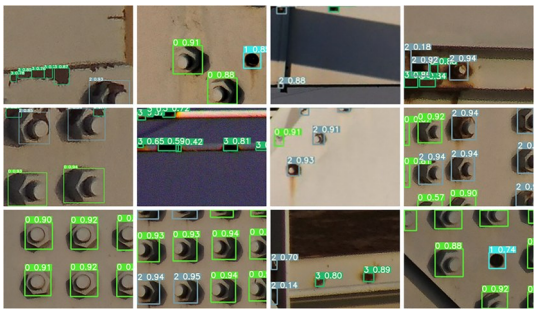

Fig. 15 Local image visualization results of ACSANet detection

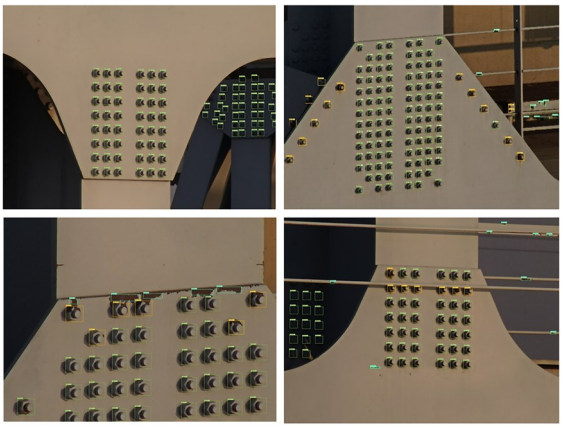

Fig. 16 Overall image visualization results of ACSANet detection

4 Conclusions

This study proposes a network structure ACSANet for UAV image detection of railroad bridge steel structures. The ACSANet network model firstly performs adaptive cropping of UAV images, inputs suitable targets to be detected and image scales, and outputs cropped images with excellent detection effects. Secondly, the YOLOv5s network structure is improved by incorporating the shallowest features of the backbone network, and a shallow attention network is proposed, so that the target to be detected with simple features will not be affected by the deeper network structure or other similar targets and backgrounds. Finally, the CA mechanism module is added to the shallow attention network, which makes the detection process pay more attention to the shallow features, while the α-IOU loss function is used in the IOU loss function in the model to improve the detection effect on small data sets. Through the above improvements and experimental results, the conclusions are summarized as follows:

1. In small data sets, the ratio between the target to be detected and the input image has a significant impact on the final detection result. In the data set used in this study, when the ratio of the image to the main target is between 20:1 and 80:1, the boundary is 50:1. When the ratio is greater than 50:1, the accuracy

changes greatly, but the training time remains basically unchanged. When the ratio is less than 50:1, the accuracy remains basically unchanged, but the training time changes greatly. Therefore, there is a critical point in the training process, when the combined result of training efficiency and test accuracy is the best.

2. Deeper networks will interfere with the detection accuracy of small targets, limited samples, and simple objects. Comparing the detection results under different network structures but keeping other strategies the same, ACSANet improves the accuracy of missing bolts by nearly 10% compared to ACNet+CA+α-IOU.

3. Different attention mechanisms do not necessarily improve detection accuracy due to different attention directions. Appropriate attention mechanism and loss function can better detect the steel structure image of railway bridge taken by UAV, but using an inappropriate attention mechanism will have a negative effect on the detection.

The experimental results show that ACSANet can better detect defects in the steel structure of railroad bridges and, especially, significantly improve the detection accuracy of missing bolts, thus providing greater safety for railway bridges.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61833002).

Author contributions

Zonghan MU designed the research and processed the corresponding data. Zhipeng WANG, Huaizhi YANG, and Yonghui HUANG collected the data. Zonghan MU wrote the first draft of the manuscript. Chongchong YU and Yunpeng WU helped organize the manuscript. Yong QIN revised and edited the final version.

Conflict of interest

Zonghan MU, Yong QIN, Chongchong YU, Yunpeng WU, Zhipeng WANG, Huaizhi YANG, and Yonghui HUANG declare that they have no conflict of interest.

References

Ali R, Kang D, Suh G, et al., 2021. Real-time multiple damage mapping using autonomous UAV and deep faster region-based neural networks for GPS-denied structures. *Automation in Construction*, 130:103831. https://doi.org/10.1016/j.autcon.2021.103831

Arivazhagan S, Shebiah RN, Magdalene JS, et al., 2015. Railway track derailment inspection system using segmentation based fractal texture analysis. *ICTACT Journal on Image and Video Processing*, 6(1):1060-1065. https://doi.org/10.21917/ijivp.2015.0155

Bochkovskiy A, Wang CY, Liao HYM, 2020. YOLOv4: optimal speed and accuracy of object detection. arXiv: 2004.10934. https://doi.org/10.48550/arXiv.2004.10934

Box GEP, Cox DR, 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Cha YJ, Choi W, Büyüköztürk O, 2017. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5):361-378. https://doi.org/10.1111/mice.12263

Cha YJ, Choi W, Suh G, et al., 2018. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):731-747. https://doi.org/10.1111/mice.12334

Chen JW, Liu ZG, Wang HR, et al., 2018. Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network. *IEEE Transactions on Instrumentation and Measurement*, 67(2):257-269. https://doi.org/10.1109/TIM.2017.2775345

Chen P, Wu YP, Qin Y, et al., 2019. Rail fastener defect inspection based on UAV images: a comparative study. Proceedings of the 4th International Conference on Electrical and Information Technologies for Rail Transportation, p.685-694. https://doi.org/10.1007/978-981-15-2914-6_65

Chen Q, Liu L, Han R, et al., 2019. Image identification method on highspeed railway contact network based on YOLO v3 and SENet. Chinese Control Conference, p.8772-8777. https://doi.org/10.23919/ChiCC.2019.8865153

Chen YK, Zhang PZ, Li ZM, et al., 2020. Stitcher: feedback-driven data provider for object detection. arXiv: 2004.12432. https://doi.org/10.48550/arXiv.2004.12432

Choi W, Cha YJ, 2020. SDDNet: real-time crack segmentation. *IEEE Transactions on Industrial Electronics*, 67(9):8016-8025. https://doi.org/10.1109/TIE.2019.2945265

Duque L, Seo J, Wacker J, 2018. Bridge deterioration quantification protocol using UAV. *Journal of Bridge Engineering*, 23(10):04018080. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001289

He JB, Erfani S, Ma XJ, et al., 2021. Alpha-IOU: a family of power intersection over union losses for bounding box regression. arXiv: 2110.13675. https://doi.org/10.48550/arXiv.2110.13675

Hou QB, Zhou DQ, Feng JS, 2021. Coordinate attention for efficient mobile network design. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.13708-13717. https://doi.org/10.1109/CVPR46437.2021.01350

Hu J, Shen L, Albanie S, et al., 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

Jia XY, Luo WG, 2019. Crack damage detection of bridge based on convolutional neural networks. Chinese Control and Decision Conference, p.3995-4000. https://doi.org/10.1109/CCDC.2019.8833336

Kang DH, Cha YJ, 2018. Autonomous UAVs for structural health monitoring using deep learning and an ultrasonic beacon system with geo-tagging. *Computer-Aided Civil and Infrastructure Engineering*, 33(10):885-902. https://doi.org/10.1111/mice.12375

Kang DH, Cha YJ, 2021. Efficient attention-based deep encoder and decoder for automatic crack segmentation. *Structural Health Monitoring*, 21(5):1-16. https://doi.org/10.1177/14759217211053776

Kang DH, Benipal SS, Gopal DL, et al., 2020. Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning. *Automation in Construction*, 118:103291. https://doi.org/10.1016/j.autcon.2020.103291

Kisantal M, Wojna Z, Murawski J, et al., 2019. Augmentation for small object detection. arXiv: 1902.07296. https://doi.org/10.48550/arXiv.1902.07296

Liu G, Han J, Rong WZ, 2021. Feedback-driven loss function for small object detection. *Image and Vision Computing*, 111:104197. https://doi.org/10.1016/j.imavis.2021.104197

Liu JH, Wu YP, Qin Y, et al., 2019. Defect detection for bird-preventing and fasteners on the catenary support device using improved Faster R-CNN. Proceedings of the 4th International Conference on Electrical and Information Technologies for Rail Transportation, p.695-704. https://doi.org/10.1007/978-981-15-2914-6_66

Liu W, Anguelov D, Erhan D, et al., 2016. SSD: single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision, p.21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Long A, Kim CW, Kondo Y, 2021. Detecting loosening bolts of highway bridges by image processing techniques. Proceedings of the 16th East Asian-Pacific Conference on Structural Engineering and Construction, p.119-127. https://doi.org/10.1007/978-981-15-8079-6_11

Morgenthal G, Hallermann N, Kersten J, et al., 2019. Framework for automated UAS-based structural condition assessment of bridges. *Automation in Construction*, 97:77-95. https://doi.org/10.1016/j.autcon.2018.10.006

Noh J, Bae W, Lee W, et al., 2019. Better to follow, follow to be better: towards precise supervision of feature super-resolution for small object detection. IEEE/CVF International Conference on Computer Vision, p.9724-9733. https://doi.org/10.1109/ICCV.2019.00982

Rahman MA, Yang W, 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. The 12th International Symposium on Advances in Visual Computing, p.234-244. https://doi.org/10.1007/978-3-319-50835-1_22

Ramana L, Choi W, Cha YJ, 2017. Automated vision-based loosened bolt detection using the cascade detector. *Sensors and Instrumentation*, 5:23-28. https://doi.org/10.1007/978-3-319-54987-3_4

Redmon J, Divvala S, Girshick R, et al., 2016. You only look once: unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition, p.779-788. https://doi.org/10.1109/CVPR.2016.91

Rezatofighi H, Tsoi N, Gwak J, et al., 2019. Generalized intersection over union: a metric and a loss for bounding box regression. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.658-666. https://doi.org/10.1109/CVPR.2019.00075

Shao ZF, Li CM, Li DR, et al., 2020. An accurate matching method for projecting vector data into surveillance video to monitor and protect cultivated land. *ISPRS International Journal of Geo-Information*, 9(7):448. https://doi.org/10.3390/ijgi9070448

Tang X, Du DK, He ZQ, et al., 2018. PyramidBox: a context-assisted single shot face detector. The 15th European Conference on Computer Vision, p.812-828. https://doi.org/10.1007/978-3-030-01240-3_49

Tao X, Zhang DP, Ma WZ, et al., 2018. Automatic metallic surface defect detection and recognition with convolutional neural networks. *Applied Sciences*, 8(9):1575. https://doi.org/10.3390/app8091575

van Etten A, 2018. You only look twice: rapid multi-scale object detection in satellite imagery. arXiv: 1805.09512. https://doi.org/10.48550/arXiv.1805.09512

Wang JK, He XH, Faming S, et al., 2021. A real-time bridge crack detection method based on an improved inception-resnet-v2 structure. *IEEE Access*, 9:93209-93223. https://doi.org/10.1109/ACCESS.2021.3093210

Wang ZQ, Zhang YS, Yu Y, et al., 2021. Prior-information auxiliary module: an injector to a deep learning bridge detection model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6270-6278. https://doi.org/10.1109/JSTARS.2021.3089519

Wei ZQ, Liang D, Zhang D, et al., 2022. Learning calibrated-guidance for object detection in aerial images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2721-2733. https://doi.org/10.1109/JSTARS.2022.3158903

Woo S, Park J, Lee JY, et al., 2018. CBAM: convolutional block attention module. The 15th European Conference on Computer Vision, p.3-19. https://doi.org/10.1007/978-3-030-01234-2_1

Wu YP, Qin Y, Wang ZP, et al., 2018. A UAV-based visual inspection method for rail surface defects. *Applied Sciences*, 8(7):1028. https://doi.org/10.3390/app8071028

Wu YP, Qin Y, Qian Y, et al., 2022. Hybrid deep learning architecture for rail surface segmentation and surface defect detection. *Computer-Aided Civil and Infrastructure Engineering*, 37(2):227-244. https://doi.org/10.1111/mice.12710

Yang CHY, Huang ZH, Wang NY, 2021. QueryDet: cascaded sparse query for accelerating high-resolution small object detection. arXiv: 2103.09136. https://doi.org/10.48550/arXiv.2103.09136

Zhang HY, Cisse M, Dauphin YN, et al., 2018. Mixup: beyond empirical risk minimization. The 6th International

Conference on Learning Representations.

Zhang YF, Ren WQ, Zhang Z, et al., 2022. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing*, 506:146-157. https://doi.org/10.1016/j.neucom.2022.07.042

Zheng ZH, Wang P, Liu W, et al., 2019. Distance-IOU loss: faster and better learning for bounding box regression. Proceedings of the 34th AAAI Conference on Artificial Intelligence, p.12993-13000. https://doi.org/10.1609/aaai.v34i07.6999

Zhu XK, Lyu SC, Wang X, et al., 2021. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. IEEE/CVF International Conference on Computer Vision Workshops, p.2778-2788. https://doi.org/10.1109/ICCVW54120.2021.00312