IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 July 2022, accepted 22 July 2022, date of publication 1 August 2022, date of current version 8 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195176

RESEARCH ARTICLE

Application of YOLOv5 Based on Attention Mechanism and Receptive Field in Identifying Defects of Thangka Images

YUBO LI[1], YAO FAN[1], SHUAISHUAI WANG[1], JIANXIAN BAI[1], AND KEYING LI[2]
[1]College of Information Engineering, Xizang Minzu University, Xianyang, Shaanxi 712000, China
[2]College of Computer Science, Xi'an Polytechnic University, Xi'an, Shaanxi 710000, China

Corresponding author: Yao Fan (fannyaoo@xzmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62062061.

ABSTRACT Aiming at the problems of target detection network in the defect detection field of thangka images with complex background colors, such as poor small target detection effect, insufficient feature information extraction, prone to error detection and leak detection, and low accuracy of defect detection, this paper proposed the YOLOv5 defect detection algorithm combining attention mechanism and receptive field. First of all, the Backbone network is used for feature extraction, integrating attention mechanism to represent different features, so that the network can fully extract the texture and semantic features of the defect area, and the extracted features are weighted and fused to reduce information loss. Secondly, a weighted fusion of features of different dimensions is transferred by the Neck network, and the combination of FPN and PAN is used to realize the fusion of semantic features and texture features of different layers and to locate the defect target more accurately. Finally, while replacing the GIoU loss function with CIoU, the receptive field is added to the network, so that the algorithm uses a four-channel detection mechanism to expand the detection range of receptive fields, and fuses semantic information between different network layers, so as to achieve fast location and more refined processing of small targets. The experimental results show that compared with the original YOLOv5 network, the detection accuracy of YOLOV5-scSE and YOLOV5-CA networks proposed in this paper has improved by 8.71 percentage points and 10.97 percentage points respectively, and the verification index has been significantly improved. It can quickly and more accurately identify and locate the location of the defect area and has a stronger ability to generalize the defect category, which greatly improves the accuracy of thangka image defect detection.

INDEX TERMS YOLOv5, defective region, tangka dataset, deep learning, scSE, CA.

I. INTRODUCTION

Thangka images emerged in Songtsen Gampo's period, are to use the color satin framed by religious scroll painting, in the development of Tibetan history long river absorbed the advantages of culture, formed its own unique culture and art, keep records of all aspects of politics, religion, economy, history and social life in Tibet's long history, are called the ''encyclopedia of Tibetan culture''. The state has upgraded from the traditional identification, recording and archival protection of cultural heritage to the restoration of damaged cultural heritage. Due to the particularity of the materials made of thangka images, the difficulty of preservation is greatly increased. Moreover, thangka images are also vulnerable to damage from external environmental factors. In the process of detection and repair, the fragile thangka images will be harmed again. Therefore, the use of digital processing has become an important way and trend of conservation and restoration of cultural relics.

A large number of researchers have conducted relevant studies on the detection and research of damaged areas at home and abroad: Wang *et al.* [1] applied the DCNN model to the defect detection of potato surfaces and conducted transfer learning on independently developed data sets to improve the

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

generalization ability of the model. For the defects on the back glass of cell phones, Jiang *et al*. [2] proposed a symmetric neural network with U-net based encoder and decoder to apply defect detection to the detection of defects on the back-glass surface of cell phones. In order to detect aircraft engine defects, Chen *et al*. [3] proposed a fast and accurate feature weighting network (FWNet) to solve the problem of defect scale change. Furthermore, in order to increase the effective feature weight, they proposed a new feature weighting module to improve the accuracy of aircraft engine defect detection. Gu *et al*. [4] proposed an insulator localization algorithm and insulator defect detection algorithm using a deep convolutional neural network in order to solve the problem of high false positives in railroad insulator defect detection, RRPN was used to locate defect targets, which could effectively eliminate unnecessary background in positioning results. Tabernik *et al*. [5] proposed a two-stage deep learning framework based on segmentation, which uses very little data for training and can play a good role in the detection and segmentation of surface anomalies. Zhao *et al.* [6] proposed a novel detection framework with positive sample training, combining GAN and self-encoder to reconstruct defective images and using LBP features for local contrast defect detection in images. Xu *et al*. [7] proposed an efficient defect detection device SEDD based on self-supervised learning strategy and image segmentation, which can detect defect regions without defect samples with labeled data in the pipeline. Defect detection [8]–[13] has been well developed in recent years with the acceleration of the modernization process.

The existing detection algorithms are mainly based on the Two-stage target detection network represented by RCNN [14]–[17] and the One-stage target detection network represented by YOLO [18]–[21], and the fundamental difference between the two detection algorithms is whether there is a region proposal process in the detection, which also creates the advantages and disadvantages of the two algorithms.

Although the Two-stage network has high detection accuracy, it has high training cost and slow detection speed and requires a large number of data sets due to the network depth. The advantage of the One-stage is that the detection speed is fast, and because the depth and number of parameters of the network are less than that of the Two-stage network, the number of data required is also less than that of the Two-stage network, so it is suitable for small sample data sets. The disadvantage is that the detection accuracy is slightly poor, and the detection of small targets is not good. The thangka image data belongs to a small sample dataset, so YOLOv5 is adopted as the basic framework of the network. However, it is found through experiments that: because of the particularity of thangka image production, its background color is rich so that the difference between the foreground and background characteristics of thangka image is not obvious, and the original network cannot learn its defect characteristics well, and the defect areas of small targets are prone to false detection and missed detection. To address the above problems, this paper proposes a defect detection network that integrates the attention mechanism and increases the receptive field, to strengthens the learning of the texture and semantic features of the small target defect area of the thangka image. Thus, when detecting the small target defect area of the thangka image with complex background color, the situation of false detection and omission is reduced. Specific innovation points are as follows:

a) Since the YOLOv5 network is not able to learn the features of its target region well for images with complex background colors, in this paper, the scSE (concurrent spatial and channel squeeze & excitation) mechanism, which is excellent in the field of image segmentation, is combined with the output of Backbone, so that the network can better express the defect characteristics for the defect area.
b) Since the network cannot allocate and process the defective target more finely, detection often occurs in the case of false detection. The CA (Coordinate Attention) module is fused after the output of Backbone and on the output of three branches of FPN respectively, so that the detection accuracy and measurement index of the network for the target area has been further improved.
c) Due to the poor detection effect of One-stage network for small targets, misdetection and omission of small targets often occur. This paper designs four-channel receptive field detection, which adds a small target detection layer on the basis of the network to improve the information fusion of deep semantic features and shallow semantic features, so that the detection network for the small target defect detection effect is improved.
d) In this paper, the algorithm is applied to defect detection of thangka images, which improves the accuracy of defect detection and reduces the chance of loss caused by human factors. It provides a new technical route for defect detection of thangka images and a new idea and provides a new idea and demonstration for digital cultural heritage conservation.

II. OVERALL NETWORK FRAMEWORK

A. IMPROVE NETWORK ARCHITECTURE

According to the characteristics of thangka data set and the unique characteristics of the target defect region combined with the advantages and disadvantages of defect detection, a YOLOv5 algorithm based on attention mechanism and receptive field is proposed to detect the defect region of thangka image. The main work focuses on the design and optimization of the receptive field of Yolov5s network, the design and optimization of the overall architecture of the network, the comparison and data analysis of the experimental data obtained by the application of the same data set with other different networks and the improvement of the loss function and activation function. The improved algorithm can effectively improve the precision rate and recall rate of thangka defect detection, effectively reduce the missed

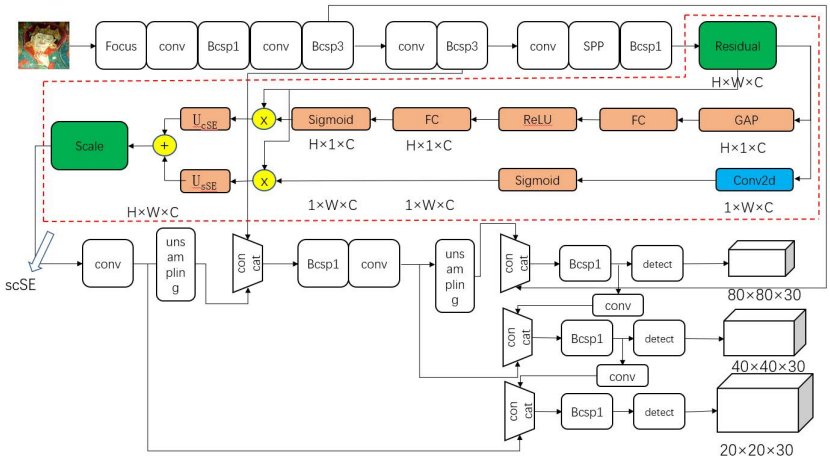


FIGURE 1. Structure diagram of YOLOv5-scSE mechanism.

detection rate and false detection rate of detection, and can detect the defect area more completely and classify its defect categories, so as to complete the defect detection of thangka. From the experimental results, it can be more clearly proved that the network model has been greatly improved in both accuracy and recall.

1) IMPROVE NETWORK ARCHITECTURE

The overall network architecture optimizes the design of the original network from three aspects. First, the scSE network module is imported on the basis of the original network, and the relationship between the space and the channel of the network is optimized, so that the feature map output from Backbone can be further learned in the network, and its features can be further optimized and purified, so that the network can better learn the characteristics of the target area. Second, introducing the CA module into the network, the CA module can not only fuse the cross-channel feature information but also capture the direction-aware and location-aware information, so that the network can have a better feature learning rate. In addition, it also reduces the loss of features, and enables the network to obtain more correlation eigenvalues from position information in different directions, so as to detect and locate more accurately. Third, adding receptive field to the network can make the network better learn its texture and semantic features for the defective areas of small targets, thus reducing the probability of missed detection and false detection in the small target area. The optimized network is compared with its original network to improve the defect detection network so that the network model can better complete the defect detection work.

2) SCSE

The scSE module [22], [23] re-plans and adjusts the characteristics of the network. The idea of scSE is to learn the correlation between channels. By increasing the weight of feature channels, the influence of background features is reduced, which is used to improve the segmentation results. The scSE module applied in any layer of CNN, so as to improve the ability of CNN to encode spatial information and improve the network's feature recognition ability for images. The scSE module is used in the network combination form of encoder-decoder like U-Net, so as to improve the network information extraction of feature map. In this paper, it is applied to defect detection for experimental comparison.

The overall framework is shown in Figure 1, where scSE is in the red box. The cSE module uses global average pooling to change the dimension of the input feature map into $1\times1\times C$, and then uses convolution to process the information to obtain the C-dimensional vector. The processed vector is normalized by Sigmoid activation function, and finally the calibrated feature map is obtained by channel-wise multiplication. The sSE module directly carries out convolution operation on the feature graph, transforms it the $H\times W\times1$ feature, and then obtains the spatial attention diagram through the sigmoid activation function, which is fused with the input feature to complete the calibration of spatial information. Finally, the feature outputs of the cSE and sSE modules are weighted and fused to obtain the feature output of the scSE module.

3) CA

The CA module compared with other lightweight attention method has great advantage on the network. First of all, when acquiring features, it can not only capture cross-channel feature information can also capture the directionality and location-aware information, these advantages can help the CA module more precise positioning and identify the target area; Second, The CA modules are lightweight and flexible, and can be easily deployed in network models to enhance features by strengthening information representation; Finally, The CA can bring huge benefits to downstream feature learning on the basis of lightweight network, especially multi-target detection and positioning tasks.

When acquiring features, the CA module can capture cross-channel feature information as well as direction sensing and location sensing information, and encode the input

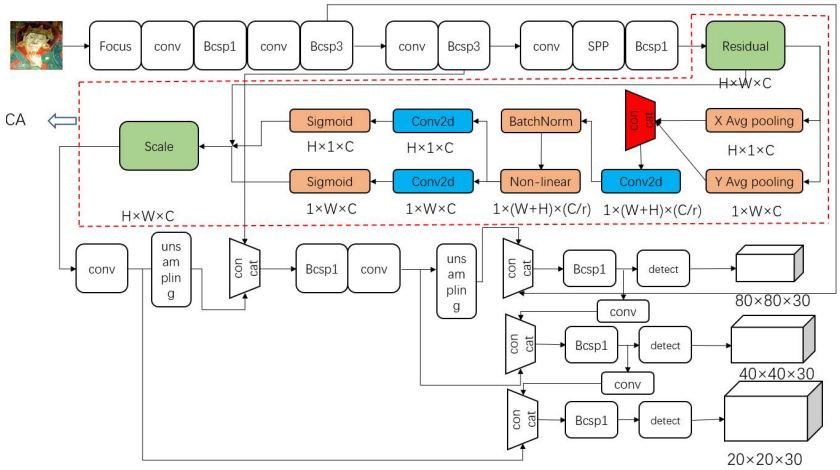


FIGURE 2. Structure diagram of YOLOv5-CA mechanism.

features through coordinate information embedding and coordinate information generation. Its modules are enclosed in the red dashed box in Figure 2. The H × 1 and 1 × W pooling kernels are used to encode each channel along horizontal and vertical coordinate, respectively, and the feature maps of H × 1 × C and 1 × W × C are obtained. Then, the features in different directions were fused to return an attention diagram of direction perception. These two transformations can not only make the attention module capture the long-term dependence along one spatial direction, but also save the precise location information along another spatial direction, which can improve the network positioning of the target to be detected more accurately.

4) RECEPTIVE FIELD

In order to improve the performance of the network and further ensure that the network can finely learn the global characteristics of the defect target area of the thangka image, so that the network can better detect and locate the small target area, this paper adds the receptive field to the original three-channel detection mechanism and extends it to a four-channel receptive field detection mechanism. the network further integrates the semantic information and texture information between different dimensions. As a result, it can achieve more refined processing of small targets, further more accurately identify and locate the defect areas of small targets, and reduce the probability of missed detection and false detection of small targets. The expanded four-channel receptive field structure is shown in Figure 3:

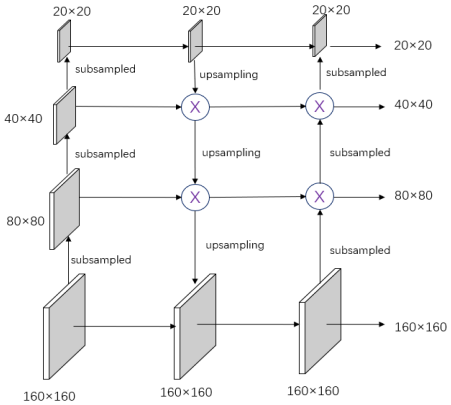


FIGURE 3. Four scale receptive field structure.

B. LOSS FUNCTION

The loss function mainly describes the execution efficiency of the network algorithm for the expected results. The main purpose is to detect the results predicted by the network model to measure the quality of the network model prediction. The loss function of YOLOv5 network is divided into three parts: coordinate loss function (loss_giou), target confidence loss (loss_obj) and target classification loss (loss_cls). The final total loss is the sum of coordinate loss, target confidence loss and target classification loss. Among them, the target confidence error loss and target classification loss use the cross-entropy loss function (BCE), and then use the $GIoU$ [25] loss function at the same time to calculate the score loss of the bbox (bounding box).

Although $GIoU$ increases the penalty for its misdetection compared to IoU [26], $GIoU$ is calculated as the intersection and merging relationship between the target region and the detected region. However, when the target frame and the network detection frame are included, the detection accuracy will be seriously affected. Therefore, the $CIoU$ loss function is used to replace the $GIoU$ of the original network. Moreover, $CIoU$ calculates the Euclidean distance between the real region and the center point of the detection region, so it can effectively solve the inclusion situation, and it can normalize the distance from the center point between the two frames to accelerate the convergence of the network.

$$L = -\sum_{n=1}^{N} y^{(n)} \log x^{(n)} + (1 - y^{(i)}) \log(1 - x^{(n)}) \quad (1)$$

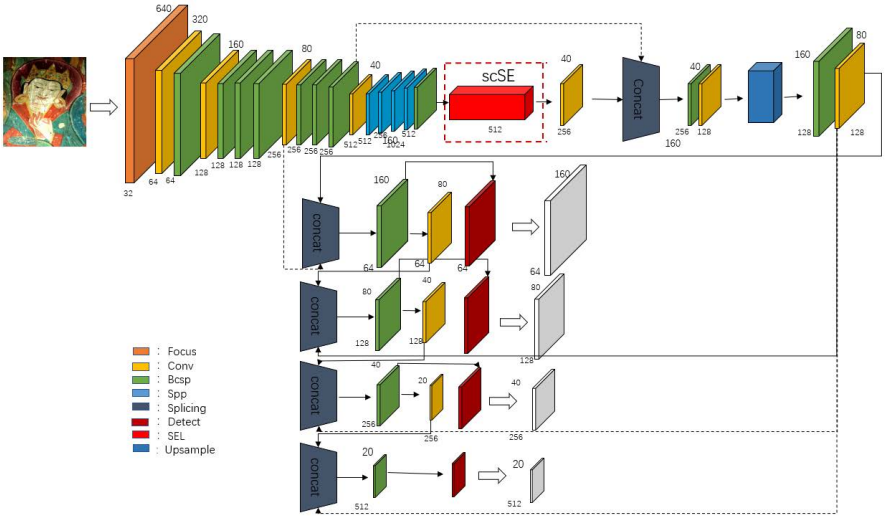


FIGURE 4. Picture size and channel number of each layer of YOLOv5-scSE.

$$GIoU = IoU - \frac{|D - (\mathrm{X} \cup \mathrm{Y})|}{|D|} \tag{2}$$

$$R_{CIoU} = \frac{\rho^2(X, Y)}{b^2} + \alpha\upsilon \tag{3}$$

$$\upsilon = \frac{4}{\pi^2}(\arctan\frac{Y_\omega}{Y_h} - \arctan\frac{X_\omega}{X_h})^2 \tag{4}$$

$$CIoU = 1 - IoU + R_{CIoU} \tag{5}$$

L is the cross-entropy loss, X and Y is the detection frame and the real frame respectively, x is predicting the probability of sample, y is label. R_{CIoU} is penalty term formula, ρ represents the Euclidean distance representing the center point, b is the diagonal distance of the smallest closed region that can contain both X and Y, α is the balance ratio parameter, υ represents the similarity parameter that measures the length and width of the detection frame and the real frame.

III. TRAINING PROCESS

A. YOLOv5-scSE TRAINING PROCESS

Figure 4 shows the training flow diagram of the YOLOv5-scSE network framework as well as the convolution parameters and channel number of each layer network. First of all, the network input the image into Backbone network module. Through continuous convolution, the input image is down-sampled so as to reduce the size of the input image and increase the number of channels. Then, the feature graph after convolution is extracted by shallow feature. Then, the processed feature images are input into the scSE attention module, and the feature images are respectively entered into the cSE and sSE network modules. CSE first conducts a global average pooling to change the feature images from H×W×C to 1×1×C, and then carries out information processing on the convolution of the feature images of 1×1×C/2 and 1×1×C successively, so as to obtain the feature vector of C dimension. Then, the sigmoid activation function was used to normalize the feature vectors to obtain the corresponding masks. Finally, the feature map after information calibration was obtained by channel-Wise multiplication operation. The sSE module is relatively convenient for feature processing. Firstly, the feature map input by the network uses a 1×1×1 convolution operation, and the dimension of the input feature map is changed from H×W×C to H×W×1. Then, the spatial attention map is obtained by sigmoid activation function for the feature graph after convolution operation, and then the calibrated feature graph is obtained by multiplying the feature graph with the feature graph input from the original network. The calibrated feature images of the cSE module and the sSE module are added and operated in corresponding positions to get the final output features of the scSE. Next, the extracted features are input into the Neck network module, and the combination mode of FPN+PAN is used for feature fusion to enhance its feature expression ability, and output the feature maps of four different dimensions. Finally, CIoU is used as the loss function of bbox, and NMS non-maximum suppression is applied to the detection frame, so as to achieve more accurate detection and positioning.

B. YOLOv5-CA TRAINING PROCESS

Figure 5 shows the training flow chart using the YOLOv5-CA network framework and the convolution parameters image size and number of channels for each layer. In this paper, the processed data set is firstly input into the network, the number of input image channels is expanded, and the input image is subjected to shallow feature extraction through a four-layer convolution operation through the Backbone module. The feature map extracted by the Backbone module is input into the attention module, and the CA attention mechanism can enhance the learning ability of the network for the defect area, extract the feature representation that is more discriminative for the target area, and fuse the features of different levels to prevent the loss of feature information. The CA attention

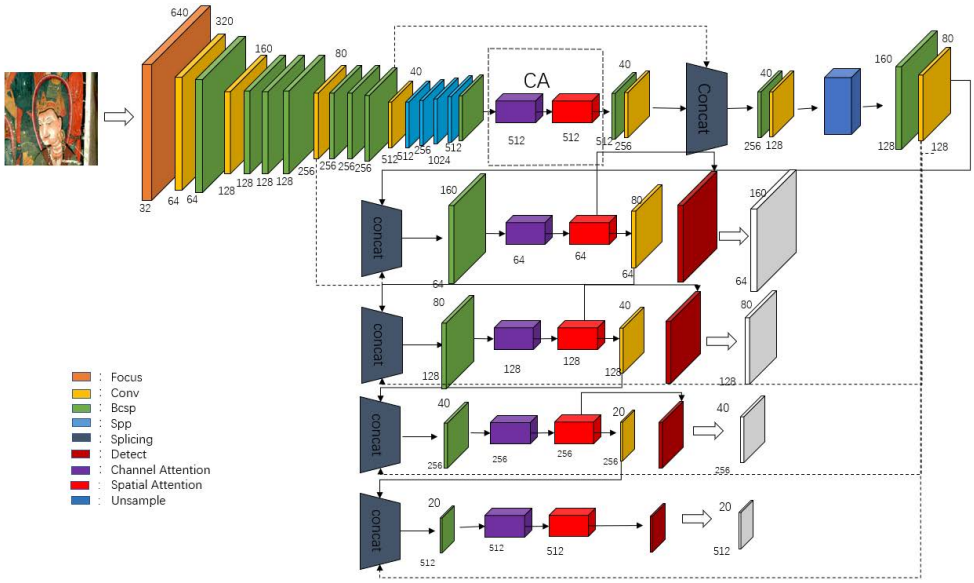


FIGURE 5. Picture size and channel number of each layer of YOLOv5-CA.

module encodes the input features through coordinated information embedding and coordinate information generation. To coordinate information embedding: First, the input feature map of H×W×C uses H×1 and 1×W pooling kernels to encode information for each channel along the horizontal and vertical coordinate directions, respectively, to obtain H×1×C and 1×W×C feature map. Through the above two transformations, feature aggregation is carried out in different directions for the space in horizontal coordinate direction and vertical coordinate direction, and then an attention diagram of direction perception is returned. These two transformations not only enable the attention module to capture the long-time dependence along one spatial direction, but also preserve the accurate position information along the other spatial direction, which is more accurate for the network to locate the target to be detected. These two transformations are relatively simple and efficient for feature map learning, and it can make full use of location information in different spatial directions to locate the target region more accurately. Finally, it can also capture the relationship between channels more effectively. For the generation of coordinate information, a cascade operation should be carried out on the feature graphs processed in horizontal coordinate direction and vertical coordinate direction to generate the intermediate feature graphs of the two transformed spatial information in horizontal and vertical directions, and then the generated feature graphs are divided into two separate tensors along the spatial dimension. The tensor is transformed into the same number of channels as the network input to CA by using two 1 × 1 convolution, and then the two weights obtained are expanded as the final attention weight. Then, the extracted and fused features are input into the Neck network module. The network adopts the up-down sampling combination of FPN+PAN to fuse the input features of different dimensions, so that the output feature maps of four different channel dimensions can carry more semantic features and texture features of target defect regions. Finally, CIoU is used as the loss function of bbox, and the NMS is used for non-maximum suppression of detection frame so that the network can detect and locate the defect area more accurately, so as to better complete defect detection of thangka images.

C. TRAINING TEST PROCESS

In this paper, the whole experimental process is divided into two aspects, namely, using the thangka data set to conduct defect training on the network and testing the images of the test set. In the process of training the thangka data set, the initial learning rate is set to 0.01 and the final learning rate is set to 0.2, the weight attenuation coefficient is set to 5e-4, the momentum is set to 0.937. Batch Size is set to 32. The whole network algorithm updates the algorithm weight through continuous forward propagation and back propagation, and the loss function tends to be stable in the parameter update. When the loss function reaches the expected value or the convergence is stable, the training of the network for the thangka defect data set is completed, and then the trained network model is obtained, and the thangka image with defect target is input into the model for detection. The trained network detects and classifies the defect target area in the picture so as to achieve the effect of detecting the defective area. The specific algorithm flow is shown in Figure 6 below:

IV. EXPERIMENT AND ANALYSIS

A. PREPARATION BEFORE EXPERIMENT

The hardware platform of the data obtained in this experiment is: CPU is Intel (R) Core (TM) i7-11700, the memory is 64.0GB, and the graphics card is NVIDIA GeForce RTX 3090. The software environment for the experimental data is

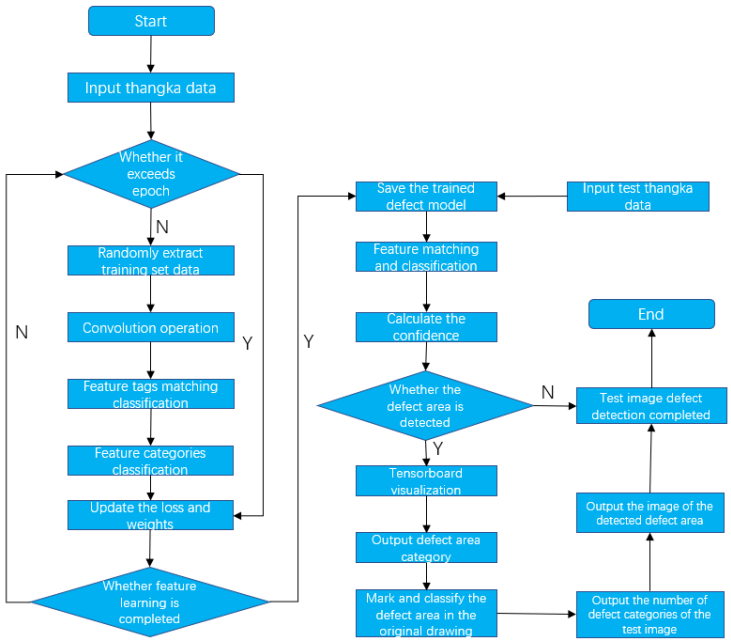


FIGURE 6. Algorithm flow chart.

TABLE 1. Distribution of the number of defect categories in the test set.

Sample	All	Fade	Crack	Dent	Damage	Stain
N	576	210	187	30	107	42

CUDA version 11.0, CUDNN version 8.0, and the operating system is Windows 10 Professional Edition. The data experimental test is carried out by using Python 3.7 programming language and PyTorch1.7.0 framework, and the compiler software is Pycharm2020 _1.2_x64.

The data set used in this experiment is the thangka images unique to Tibet. Due to the particularity and scarcity of its images, there has not been a unified thangka data set so far. What this experiment need is some datasets of thangka images with defective areas. However, because the existing thangka images are not only small in number, but also have low resolution, damage to varying degrees and are difficult to obtain, so the processing and acquisition of datasets has become an important part of this experiment. The thangka images in the data set used in this paper are all taken from Tibet. From the 7000 thangka images acquired and processed, the thangka images with defects are selected to form the data set used in this experiment, then it is divided into the data set according to the proportion of the training set and the test set at 8:2, which is used to train and test the network. And in order to enrich the diversity and robustness of the training samples, this experiment expands the sample data by mirroring, flipping, immediately cutting and other data enhancement methods, so that the network training can achieve better results.

In this experiment, the defect areas of the defect thangka data set are marked by the labeImg marking tool software, and the defect areas are classified into five categories: fade, crack, dent, damaged, stain. The Fade defect means that the pigment on the surface of the thangka image will only fall off slightly, but the bottom plate of the thangka image has not been damaged; the Crack defect indicates that there are serious cracks in the thangka image and the bottom plate is seriously damaged; the Dent defect refers to the phenomenon that the surface of the thangka image is uneven and incomplete under the action of external force; the Damaged defects show that not only the pigment color on the surface of the thangka image has fallen off, but also the base plate has been slightly damaged; the Stain defects indicate that the surface of the thangka image is contaminated by soil, oil, rain stains or other pigments that do not belong to the original image, destroying the image area of the surface. The defect categories are expressed as 0, 1, 2, 3,4 respectively. Table 1 is a defect sample to test the category number distribution of the thangka dataset:

B. ANALYSIS OF COMPARISON EXPERIMENTAL DATA

1) NO INCREASE IN RECEPTIVE FIELD

In this paper, the network without added receptive field was used for training and learning on Thangka images, and the experimental comparative analysis was conducted for 3000 and 6000 iterations respectively.

From the experimental results of Table 2, It can be seen from the experimental results that, in the thangka data set

TABLE 2. No increase in receptive field comparison of network parameters for 3000 iterations.

Network model	P	R	mAP@0.5	mAP@0.5:0.95
Yolov3	87.92	0.7352	0.6829	0.4936
Yolov5s	0.745	0.6093	0.6176	0.3688
Yolov5-scSE	**0.8233**	**0.6395**	**0.6392**	**0.4083**
Yolov5-CA	**0.829**	**0.656**	**0.632**	**0.426**
Yolov5l	0.863	0.699	0.635	0.471

TABLE 3. No increase in receptive field comparison of network parameters for 6000 iterations.

Network model	P	R	mAP@0.5	mAP@0.5:0.95
Yolov3	0.8941	0.7129	0.6829	0.5314
Yolov5s	0.773	0.629	0.6312	0.4114
Yolov5-scSE	**0.849**	**0.66**	**0.6417**	**0.439**
Yolov5-CA	**0.866**	**0.67**	**0.653**	**0.448**
Yolov5l	0.882	0.7034	0.6676	0.5282

TABLE 4. Increase the receptive field comparison of network parameters for 3000 iterations.

Network model	P	R	mAP@0.5	mAP@0.5:0.95
Yolov5s	0.7855	0.6284	0.6312	0.414
Yolov5-scSE	**0.8672**	**0.6524**	**0.6495**	**0.4515**
Yolov5-CA	**0.8731**	**0.666**	**0.6502**	**0.4395**
Yolov5l	0.8656	0.704	0.6696	0.519

used in this experiment, compared with other network models, although the detection accuracy and recall rate of the network model used in this paper did not reach the best, but in 3000 iterations, the detection accuracy of Yolov5s network is improved by 0.0783 and 0.084 respectively, so it can be proved that the network proposed in this paper can effectively improve the detection effect.

When the number of iterations reaches 6000, The experimental results are shown in Table 3, the detection accuracy of the network model YOLOv5-scSE and YOLOv5-CA used in this paper is 0.076 and 0.093 higher than that of the original network YOLOv5s, and the recall rate is 0.031 and 0.041 also higher than it, respectively. Compared with the experimental data with 3000 iterations, the detection accuracy and recall rate are greatly improved. It is fully explained that the network used in this paper has a good effect of feature learning for data sets with complex background colors. but it still lags behind compared with YOLOv3 and YOLOv5l. The reason is that some of the target defect areas in the thangka images belong to small target features, while the network algorithm and the original network algorithm YOLOv5s used in this paper are shallow because of the shallow depth of the network model. The defect features of the small target defect area cannot be well obtained, so the detection accuracy and recall rate of the network algorithm used in this experiment are slightly lower than those of the other two networks.

2) INCREASE THE RECEPTIVE FIELD

In the convolutional neural network, the excessive downsampling rate leads to less small target features in learning, while the excessive receptive field is more unfriendly to the detection of small target features. The receptive field of the feature points on the feature map of the target area is much larger than that in the downsampling rate, which leads to too few effective features occupied by the small targets in the feature points of the feature graph learned by the network algorithm. And it will lose a large number of semantic features and texture features in the area around the small target, resulting in a decline in the detection accuracy, missed detection and false detection in the defect area of the small target. In this experiment, the proportion of small targets in the defect area is too high, and the network cannot learn the texture and semantic features of the small target defects very well, in order to improve the detection efficiency, this paper improves and optimizes the network to increase the receptive field of small targets, so that the network algorithm can accurately learn the characteristics of small target defect regions when detecting small targets, thus improving the accuracy of thangka image defect detection.

From the experimental results in Table 4, we can see that on the thangka image dataset, when the receptive field is added to the network, the accuracy and recall rate of all network models are greatly improved compared with the original receptive field. Table 4 shows that when the receptive field is increased, the detection effect of all network models at epoch 3000 iterations is slightly higher than that of the original network epoch 6000 iterations. And compared with other network models, when the network epoch is iterated 3000 times, the improvement effect of the YOLOv5-scSE used in this paper is the most obvious. compared with the original YOLOv5s network, the detection accuracy of YOLOv5-scSE and YOLOv5-CA is increased by 0.0817 and

TABLE 5. Comparison of network parameters for 6000 iterations.

Network model	P	R	mAP@0.5	mAP@0.5:0.95
Yolov5s	0.8360	0.643	0. 6425	0.4409
Yolov5-scSE	**0.8726**	**0.6761**	**0.6529**	**0.4591**
Yolov5-CA	**0.8952**	**0.6818**	**0.6613**	**0.4786**
Yolov5l	0.8924	0.7034	0.6676	0.5282

0.0876, respectively, and the recall rate is increased by 0.024 and 0.0376, respectively. Although the accuracy and recall of YOLOv5-scSE are slightly less than YOLOv5-CA, in terms of mAP@0.5 and mAP@0.5:0.95 parameters, the former has obvious advantages in average detection accuracy compared with the latter, which further shows that YOLOv5-scSE can converge faster than YOLOv5-CA when epoch iterates for 3000 times.

From the experimental results in Table 5, when epoch iterates 6000 times, the detection effect of the network model and the contrast network model is improved on the thangka image dataset. The detection accuracy of YOLOv5-scSE and YOLOv5-CA is 0.0366 and 0.0592 higher than that of YOLOv5s, and the recall rate is 0.0331 and 0.0388 higher than that of the original network, respectively, and YOLOv5-CA surpasses YOLOv5l in detection accuracy. Compared with Table 4, when the epoch was iterated 6000 times, the detection accuracy of the network increasing receptive field is 0.0871 and 0.1097 higher than that of the original network model YOLOv5s, respectively. By increasing the receptive field, the network can deeply learn the small target region features of complex background colors, so as to improve the detection effect.

3) INCREASE THE RECEPTIVE FIELD COMPARISON OF EXPERIMENTAL EFFECTS

In order to more intuitively illustrate the effectiveness of the improvement of the overall network framework in this paper, the implementation effect charts of adding scSE network module and CA network module to the original network are described in the form of histogram, and two groups of histograms are used to represent the visual analysis of the effect of 6000 network epoch iterations with and without receptive field, respectively.

According to the above analysis, it can be directly observed that the accuracy and recall rate of the network in detecting defective areas and some other quantitative parameters have been steadily improved compared with the original network: As can be seen from Figure 7, the detection accuracy of the original network YOLOv5s is 0.773, while the detection accuracy of the attention mechanism module network added in this paper increased by 0.076 and 0.093 year-on-year respectively, and the detection recall rate increased by 0.031 and 0.041 respectively, and the mAP value also improved correspondingly. It is fully proved that adding the attention mechanism module to the original network can greatly improve the learning of thangka image features of

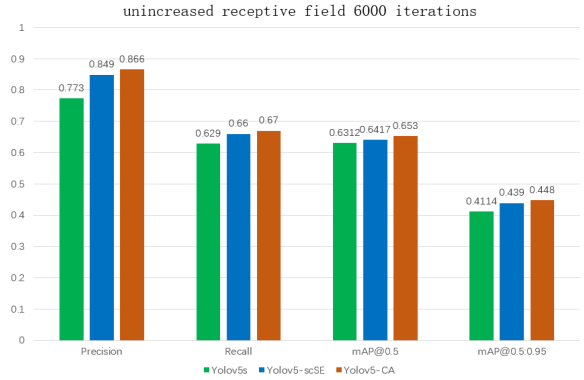


FIGURE 7. Histogram of unincreased receptive field 6000 iterations.

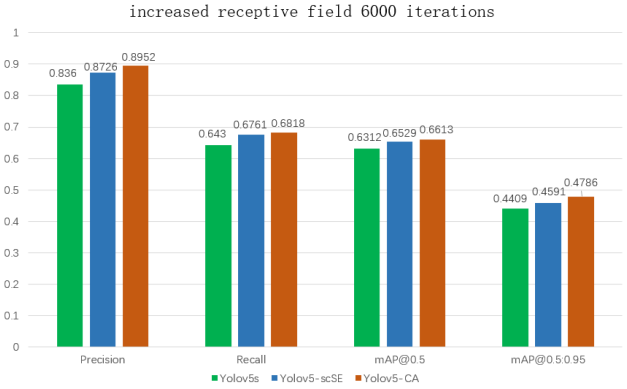


FIGURE 8. Histogram of increased receptive field 6000 iterations.

complex background colors, and can make the network fully learn the defect features of the target area, so as to improve the accuracy of target region detection. Figure 8 shows the comparative bar chart of the detection accuracy of the defect area of the network model after adding the receptive field to the network. It can be clearly seen from the chart that after the attention mechanism is added in this paper, the network data has also been greatly improved. And comparing Figure 7 with Figure 8, after the original network YOLOv5s increases the receptive field, the detection accuracy is improved by 0.063. The recall rate and mAP value have also been correspondingly improved. Compared with figure 7, when the receptive field is not increased, the detection accuracy of the network used in this paper is also increased by 0.0236 and 0.0292 respectively, and the recall rate is also increased by0.0161 and 0.0118, respectively. Other evaluation indicators have also been improved compared with those without increased receptive field.

TABLE 6. Fade defect detection data.

Epoch	Network model	P	R	map@0.5	map@0.5:0.95
6000	Yolov3	0.857	0.895	0.844	0.653
	Yolov5s	0.812	0.861	0.829	0.539
	Yolov5-scSE	**0.832**	**0.862**	**0.83**	**0.574**
	Yolov5-CA	**0.831**	**0.88**	**0.843**	**0.596**
	Yolov5l	0.847	0.9	0.838	0.654

TABLE 7. Crack defect detection data.

Epoch	Network model	P	R	map@0.5	map@0.5:0.95
6000	Yolov3	0.785	0.850	0.720	0.504
	Yolov5s	0.732	0.556	0.547	0.239
	Yolov5-scSE	**0.774**	**0.61**	**0.6**	**0.324**
	Yolov5-CA	**0.792**	**0.631**	**0.61**	**0.321**
	Yolov5l	0.766	0.84	0.696	0.451

TABLE 8. Dent defect detection data.

Epoch	Network model	P	R	map@0.5	map@0.5:0.95
6000	Yolov3	0.930	0.367	0.365	0.286
	Yolov5s	0.693	0.267	0.335	0.167
	Yolov5-scSE	**0.977**	**0.367**	**0.374**	**0.238**
	Yolov5-CA	**0.978**	**0.367**	**0.366**	**0.246**
	Yolov5l	0.975	0.367	0.368	0.313

According to the experimental data in Figure 7 and Figure 8, it can be fully explained that when the receptive field and attention mechanism are added to the original network, the network has better generalization. It can better learn the texture and semantic features of the target region with a complex background. Moreover, the increase of receptive field can effectively solve the problems such as low detection accuracy, missed detection and error detection for small target area. When the network increases the receptive field, the overall learning ability of the network is enhanced, and the network can learn more fine features for small targets, so as to improve the detection effect of the network on the defective areas of small targets. and further improve the generalization and robustness of the whole network model.

C. COMPARISON OF THE DETECTION EFFECT CONCERNING THE FIVE TYPES OF DEFECTS

In order to better analyze the difference between the algorithm network used in this paper and other algorithm networks, this paper makes a further comparative analysis of the experimental data of five different defect types, and sets the number of epoch iterations of all networks to 6000 for quantitative analysis, in which the bold font in the table is the network algorithm used in this experiment.

In this paper, we observe the experimental data in Table 6. For the first type of Fade target defect, the detection accuracy analysis shows that the detection effect of YOLOv3 is relatively good, while the detection accuracy of the original network is the lowest. It is said that the accuracy of the network used in this paper has been slightly improved compared with the original network. From the perspective of recall rate, the highest evaluation index is YOLOv5l, followed by YOLOv3. Because of the large number of network layers and complex network architecture, the two models can learn better for the first type of fade defect areas. But in terms of year-on-year detection speed and training speed, the network used in this paper and the original YOLOv5s network have significant advantages in the comparison of speed and FPS compared with the other two comparison networks.

According to the observation in Table 7, the accuracy of the second type of creak defect is slightly lower than that of the first type of fade defect, including detection accuracy, recall and other evaluation indexes. However, in terms of accuracy, the YOLOv5-CA network used in this paper has the highest detection accuracy, followed by YOLOv3, which shows that after adding the CA module, the network can learn and analyze the features of the second kind of defect area more precisely, so as to improve its detection accuracy. However, compared with the same period of Crack defect recall rate and mAP value, it is not difficult to find that although the network used in this paper has been greatly improved compared with the original network recall rate and mAP, but compared with YOLOv3 and YOLOv5l networks, it lags far behind. Its advantage lies in improving the accuracy of detection while preserving its advantages of detection speed and small proportion.

From the experimental data in Table 8, it can be seen that when the network iterates 6000 epochs, the detection accuracy is greatly improved. Except for the original network YOLOv5s network, the network detection accuracy is more than 0.930. And the two kinds of network detection results used in this paper are better, which are 0.284 and 0.285 higher

TABLE 9. Damaged defect detection data.

Epoch	Network model	P	R	map@0.5	map@0.5:0.95
6000	Yolov3	0.903	0.860	0.892	0.696
	Yolov5s	0.869	0.822	0.829	0.501
	Yolov5-scSE	**0.918**	**0.879**	**0.858**	**0.595**
	Yolov5-CA	**0.909**	**0.879**	**0.872**	**0.6**
	Yolov5l	0.902	0.844	0.866	0.642

TABLE 10. Stain defect detection data.

Epoch	Network model	P	R	map@0.5	map@0.5:0.95
6000	Yolov3	0.926	0.593	0.593	0.535
	Yolov5s	0.760	0.619	0.589	0.425
	Yolov5-scSE	**0.928**	**0.612**	**0.591**	**0.5**
	Yolov5-CA	**0.924**	**0.595**	**0.59**	**0.502**
	Yolov5l	0.918	0.619	0.602	0.538

than the network accuracy used in YOLOv5s. The detection accuracy has been greatly improved. From the analysis of recall rate and mAP, except for the low original network, the recall rates of the other four networks are the same, but the recall rates are too low. For the experimental data of mAP@0.5, YOLOv5-scSE network is the best. The reason why the recall rate and mAP evaluation index of the third kind of Dent defect detection are not high is that the thangka defect dataset used in this paper, the target sample of Dent defect area is too few, and its ratio to the total defect sample data is too low, which leads to the network cannot learn its features more completely, which leads to its evaluation index is too low.

From the experimental data in Table 9, it is clear that for the fourth type of Damaged defect area, the detection accuracy of YOLOv5-scSE is the highest, and the accuracy of YOLOv5-CA is higher than that of YOLOv3 and YOLOv5l. Based on the analysis of the experimental results of the fourth type of defect from the recall rate, it is found that the recall rate of the network used in this paper is also higher than that of other networks, and the year-on-year processing speed is 28.57 frame and 27.78 frame respectively, which is only slightly smaller than that of the original network, but much higher than the 6.7 frame of YOLOv3 and 7.94 frame of YOLOv5l.

Table 10 shows the analysis of the fifth type of stain defect area. From the data state of detection accuracy, both YOLOv5-scSE and YOLOv5-CA can well detect the defect areas, among which the effect of YOLOv5-scSE is the best, followed by YOLOv3, followed by the results of YOLOv5-CA. From the recall rate analysis, the recall rate of YOLOv3 and YOLOv5-CA is relatively low, while the recall rate of the original network is the highest, but its map@0.5:0.95 data is the lowest. It can be analyzed that although the recall rate of its detection is high, the average accuracy of the anchor frame is relatively small.

To sum up, from the analysis of the detection accuracy, the detection accuracy of YOLOv5-scSE network for No. 4 and No. 5 defects is better than that of other comparison networks, while the performance of the YOLOv5-CA network is better than that of No. 2 and No. 3 networks, and the accuracy of No. 1 defect YOLOv3 is the best. The network used in this paper is comparable to YOLOv3 and YOLOv5l in terms of accuracy. However, from the perspective of FPS and training time cost: in the case of improved accuracy, the network used in this paper has greatly improved the detection cost and training time cost compared with YOLOv3 and YOLOv5l networks. In the real sense, it achieves a stable state of the relative balance between speed and accuracy.

D. COMPARISON OF THE VISUAL EFFECT OF DIFFERENT DEFECT DETECTION ALGORITHMS

In order to reflect the visual effect more clearly and intuitively, this paper shows an example of using different algorithm networks to detect the images with the same defect areas in the dataset, and makes a detailed comparative analysis of the detected images. In order to better reflect the difference of the algorithm model, the threshold of the *IoU* anchor box is set to 0.4, that is, when the score of the *IoU* anchor box is greater than 0.4, the defect area detected in this paper is output as a positive sample, on the contrary, the output is a negative sample, so that most of the negative sample areas can be omitted, so as to save time and analyze more clearly. The following Figure 9 is a comparison of visual inspection results.

According to the above different algorithm comparison graphs, it is easy to find that the network optimized by this paper has a great improvement for the detection of unique thangka images: in group A figure, only YOLOV5-CA used in this paper can detect the defect area of No. 0 in the red circle identification area of the original figure, and its *IoU* threshold value is 0.58; and only YOLOV5-SCSE network can detect the defect area of no. 0 in the green circle identification area. However, for the defects in these two areas, none of the three comparison networks can detect the defect areas. in addition, for the green box in the original image, although all network models detect their defects, it is not difficult to

FIGURE 9. Comparison of different algorithms for defective images.

find that the *IoU* thresholds of other networks are relatively high compared with YOLOv5s.

The damage is too serious in the group B figure. In the black box identified in the B figure, only YOLOv5-scSE detects its defect area. Although the detection *IoU* threshold of No. 4 defect is 0.41, the other four networks did not detect the defect, and missed detection occurred. Although the red and green defective areas in the original image have been

TABLE 11. Comparison of network consumption time and FPS for 6000 iterations.

	Yolov3	Yolov5s	**Yolov5s-scSE**	**Yolov5-CA**	Yolov5l	Yolov5x
Time(h)	25.851	2.893	**3.485**	**3.088**	8.324	14.761
FPS	6.7	33.33	**28.57**	**27.78**	7.94	3.76

seriously damaged due to the influence of external factors, the network model in this paper can still detect the defective areas accurately.

The defect No.3 in the yellow box of the group C figure is missed by YOLOv5-scSE. From the green and red circles, it can be found that although the defects in the four areas are detected by all the network models, the *IoU* detected by them is much smaller than that of the two networks used in this paper. For example, in the green area, the network accuracy used in this paper is 0.88 and 0.83, while the detection accuracy of the other three network model algorithms is only 0.58, 0.56 and 0.63. The same is true in the red circle, the accuracy of the network model used in this paper is better than other network models in the detection of No. 0 defect.

In group D, YOLOv5s failed to detect the defect area No. 2 in the red box, No. 1 in the black box and No. 1 in the blue box, and the defect area no. 1 in the black box was also failed to be detected by yOLOV5-SCSE network. For the green box, both YOLOv5l and YOLOv3 comparison networks failed to detect the defect area No. 1 in the box, while both the original network and the network used in this paper were completely detected.

The defective region 0 to the left of the blue box in the group E image was detected only by YOLOV5-CA, but not by other networks. Inside the red, yellow and blue boxes, all defects can be fully detected by all networks.

Only the YOLOv5-CA network model used in this paper detected the defective region 0 in the blue box and red circle in the group F figure. However, in the defect area 1 in the green box, YOLOv5s has missed detection. Defects in other areas can be completely detected and classified by all network models.

A large area of false detection occurred in group G. In the red box of the original image, due to the complex background color of the thangka image and the high complexity of the YOLOv3 network model, the YOLOv3 network mistakenly detected the large color difference as defect No.0 and defect No.1. However, other networks can completely detect the No.2 defect in the original picture with high accuracy.

In the orange area of group H, only the YOLOv5-seSE and YOLOv5-CA networks used in this paper can accurately detect the defect areas, indicating that the network is effective for detecting small target defects. When the attention mechanism module is added to the original network, the network can more finely learn the characteristics of the defect target area, so as to achieve the purpose of effectively identifying the defect area. For the No.3 defect in the black area, only the YOLOv5l network can accurately detect the defect area; the No.1 defect area and the No.3 defect area in the orange area can also be effectively identified by all networks at the same time.

According to the experimental results, it can be seen from the overall visual comparison diagram that the detection effect of YOLOv5-scSE and YOLOv5-CA has been greatly improved, especially the YOLOv5-CA network model, which can detect the defect areas missed by both YOLOv3 and YOLOv5l. Although a certain speed and FPS have been lost after the addition of CA module, the addition of CA makes the network algorithm more fully learn the small features of its defect area. Although the network used in this paper is slightly inferior to YOLOv5s in detection speed and FPS, this paper is comparable to YOLOv3 and YOLOv5l in detection accuracy and recall rate, and has greater advantages in speed and less memory resources. And this paper adds the receptive field to the network, which can make the network pay more attention to the small target area and study it carefully. Compared with other network algorithms, the network used in this paper can learn the advantages of other network algorithms, and can make up for the shortcomings of other network algorithms, so as to carry out defect detection of Thangka image under the condition of both speed and accuracy.

E. COMPARISON OF CONSUMPTION TIME AND FPS FOR 6000 ITERATIONS

From the above experimental results in Table 11, it can be seen that the original network YOLOv5s takes the least time when epoch iterates for 6000 times, followed by YOLOv5-CA, YOLOv5-scSE, YOLOv5l and YOLOv5, respectively. The network model that takes the longest training time is that YOLOv3 takes up to 25.851 hours. The reason for the difference in network training time is that the network architecture and the number of network model layers lead to the difference of training time. Although the training time of YOLOv5-scSE and YOLOv5-CA used in this paper is slightly longer than that of the original network, the detection accuracy and effectiveness of the two algorithms used in this paper are much higher than those of the original network. Compared with YOLOv3, YOLOv5l and YOLOv5x, the detection effect of the algorithm used in this paper is almost the same or even better than that of the original network. However, the size and training time cost of the network model used in this paper is far less than that of the above network model, and FPS has 3 to 7 times of detection advantages over the above network.

F. MODEL DATA SIZE COMPARISON

This paper further analyzes the data on other parameters of the network model with the addition of scSE and CA and compares it with other good network models, including the size of the network model, the depth of the network layers, the

TABLE 12. Comparison of the data size of network model for 6000 iterations.

Network model	Yolov3	Yolov5s	**Yolov5-scSE**	**Yolov5-CA**	Yolov5l
Model size（MB）	117.75	13.76	**19.00**	**17.9**	93.80
layers	333	283	**333**	**330**	499
parameters	61545274	7074330	**8848154**	**8315026**	46652890
GFLOPS	155.1	16.4	**17.8**	**16.9**	114.3

size of the number of network parameters and the complexity of GFLOPS (Giga Floating-point Operations Per Second) to analyze the model in comparison.

Comparative analysis of Table 12 shows that when the network iterates over the thangka dataset for 6000 times, the difference in the size of the network model is very obvious. The maximum model size of YOLOv3 is 117.75MB, while the minimum size of YOLOv5s is 13.76MB. Compared with YOLOv5s, the algorithm model used in this paper increases the size of the network model due to the increase of attention mechanism and receptive field. However, compared with YOLOv3 and YOLOv5l, the model size is much smaller than the above two models, and the network model has great advantages over other network models in storage. Although the algorithm model used in this paper has increased the number of network layers compared with YOLOv5s, the number of parameters has not increased much. On the contrary, although the number of network layers in this paper is similar to that of YOLOv3, the number of parameters is only 12.88% of that of YOLOv3 model and 17.82% of the number of YOLOv5l parameters. GFLOPS is used to measure the complexity of the algorithm model. As can be seen from the above table, the network model algorithm used in this paper has lower complexity than YOLOv3 and YOLOv5l, and does not affect its detection accuracy, so it can get more accurate detection results with very low complexity and number of parameters.

V. CONCLUSION

In the defect detection of thangka images, due to the complex background color of thangka images, it is difficult to extract the features of defective regions, which results in low accuracy of defect detection network in detecting defect areas and is prone to the phenomenon of false detection and missed detection. In this paper, we propose a thangka defect detection network integrating the scSE and CA mechanisms, which can effectively solve the problems of difficult feature extraction and false detection. The experimental results show that the proposed YOLOv5-scSE improves the defect detection accuracy by 0.0871 compared with the original YOLOv5s network, while the YOLOv5-CA network improves the detection accuracy by 0.1097 compared with the original YOLOv5s network. In addition, compared with some classical algorithms (YOLOv3, YOLOv5), the defect detection network used in this paper does not cause a large amount of time loss under the condition of improving accuracy and recall rate, and the detection speed is far higher than YOLOV3 and YOLOv5l networks. The experiment shows that the attention mechanism introduced in the data set with complex background color achieves good results in the detection experiment, and can achieve the purpose of matching the speed and accuracy in real sense.

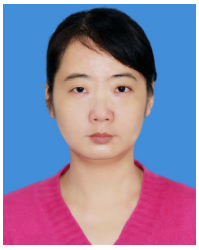
In the future, this paper will continue to explore and research in the following directions: (1) The network model is applied to more data sets to further improve the universality and accuracy of the model. (2) We will try to use more effective methods to extract defective features of thangka images, and further extract more effective features to improve the accuracy and recall rate of defective region detection by using the extracted features. (3) Feature extraction will be carried out on more complex networks so as to achieve the effect of improving detection efficiency.

REFERENCES

[1] C. Wang and Z. Xiao, "Potato surface defect detection based on deep transfer learning," *Agriculture*, vol. 11, no. 9, p. 863, Sep. 2021.
[2] J. Jiang, P. Cao, Z. Lu, W. Lou, and Y. Yang, "Surface defect detection for mobile phone back glass based on symmetric convolutional neural network deep learning," *Appl. Sci.*, vol. 10, no. 10, p. 3621, May 2020.
[3] L. Chen, L. Zou, C. Fan, and Y. Liu, "Feature weighting network for aircraft engine defect detection," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 18, no. 3, May 2020, Art. no. 2050012.
[4] Z. Gu, Y. Wang, X. Xue, S. Wang, Y. Cheng, X. Du, and P. Dai, "Railway insulator defect detection with deep convolutional neural networks," in *Proc. 12th Int. Conf. Digit. Image Process. (ICDIP)*, Jun. 2020, pp. 8–19.
[5] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *J. Intell. Manuf.*, vol. 31, no. 3, pp. 759–776, Mar. 2020.
[6] Z. Zhao, B. Li, R. Dong, and P. Zhao, "A surface defect detection method based on positive samples," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2018, pp. 473–481.
[7] R. Xu, R. Hao, and B. Huang, "Efficient surface defect detection using self-supervised learning strategy and segmentation network," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101566.
[8] Y. Fan, Y. Li, Y. Shi, and S. Wang, "Application of YOLOv5 neural network based on improved attention mechanism in recognition of Thangka image defects," *KSII Trans. Internet Inf. Syst.*, vol. 16, no. 1, pp. 245–265, 2022.
[9] J. Li and H. Wang, "Surface defect detection of vehicle light guide plates based on an improved RetinaNet," *Meas. Sci. Technol.*, vol. 33, no. 4, Apr. 2022, Art. no. 045401.
[10] Y. Chen, G. Wang, and Q. Fu, "Surface defect detection method based on improved attention mechanism and feature fusion model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Mar. 2022.
[11] S. Wang, H. Wu, X. Li, and N. Peng, "Review on surface defect detection methods of solar cells," *Int. Core J. Eng.*, vol. 8, no. 2, pp. 95–102, 2022.
[12] Z. Niu, "Design of defect detection system for semiconductor plastic packaging based on machine vision," *J. Phys., Conf. Ser.*, vol. 2006, no. 1, Aug. 2021, Art. no. 012008.
[13] S. Li and X. Wang, "YOLOv5-based defect detection model for hot rolled strip steel," *J. Phys., Conf. Ser.*, vol. 2171, no. 1, Jan. 2022, Art. no. 012040.
[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 580–587.

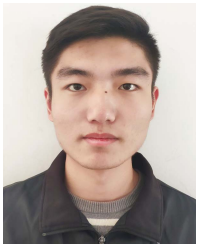
[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[19] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[20] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[22] A. G. Roy, A. N. Nav, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2018.
[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[24] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.
[25] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.
[26] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 784–799.

YUBO LI received the B.S. degree in IoT engineering from Yulin University, in 2020. He is currently pursuing the master's degree with the College of Information Engineering, Xizang Minzu University, Xianyang, China. His research interests include computer vision and defect detection.



YAO FAN received the Ph.D. degree in computer science from Chang'an University. She is currently an Associate Professor with the School of Information Engineering, Xizang Minzu University, Xianyang, China. Her research interests include the digital protection of Tibet culture heritage, and artificial intelligence and image processing. She has published more than ten articles in the above areas.

SHUAISHUAI WANG received the B.S. degree in computer science and technology from Xinxiang University, in 2020. He is currently pursuing the master's degree with the College of Information Engineering, Xizang Minzu University, Xianyang, China. His research interests include computer vision and edge detection.



JIANXIAN BAI received the B.S. degree in IoT engineering from Nantong University, in 2021. He is currently pursuing the master's degree with the College of Information Engineering, Xizang Minzu University, Xianyang, China. His research interests include computer vision and defect detection.

KEYING LI received the B.S. degree in computer science and technology from the Hi-Tech College, Xi'an University of Technology, in 2020. She is currently pursuing the master's degree with the College of Computer Science, Xi'an Polytechnic University, Xi'an, China. Her research interests include cryptography and computer vision.