Research on Object Detection in Animal Images Based on Convolutional Neural Networks

Yuxin Niu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: niuyx07l@163.com

Zhongsheng Wang
State and Provincial Joint Engineering Lab. of Advanced Network, Monitoring and Control
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: wzhsh1681@163.com

***Abstract*—Object detection is the use of computer to find out all the objects of interest in the image, determine their categories and locations, is one of the core problems in the field of computer vision. Traditional animal image target detection usually adopts the sliding window method, but due to the different sizes of the input images, this method has some problems such as insufficient training samples, low detection accuracy and slow speed. In order to solve such problems, based on the development of deep learning in recent years, this paper proposes an object detection algorithm based on convolutional neural network. YOLOv5 was used to effectively distinguish, identify and mark animal categories, which accelerated the training of the model and greatly improved the accuracy of target detection. Through the analysis of experimental data, it was concluded that the algorithm studied in this paper had good performance and good target detection results. Finally, the key problems of object detection research are summarized, and the future development trend of this field is prospected. When the number of training rounds is 30, the accuracy rate has reached about 70%, and after 50 rounds of training, some accuracy can reach 90%, which is excellent and better than other traditional algorithms.**

Keywords-Deep Learning; Object Detection; YOLO Algorithm; Animal Recognition

I. INTRODUCTION

Nowadays, computer information technology has become an indispensable part of human life, such as mobile phone payment, takeaway food that can be eaten without leaving home, convenient and practical smart furniture at home and powerful Visualize system, etc. With the continuous development of network technology and the wide application of monitoring system, a large number of visual information data spread rapidly on the network. These huge data make a variety of imaging equipment gradually replace human vision, and use the powerful computing power of computers to analyze the data [1] , so as to more accurately recognize people and things in the real world. In recent years, convolutional neural network-based methods have been widely used in the field of image recognition. The visual information data required for processing and analyzing information is very complex, which also poses a huge challenge to the progress of computer vision technology.

Animal resources are as valuable as ecological resources and have high scientific, economic and medical value. For human beings, protecting animal resources is to protect human beings themselves. Animal image target detection can be used to estimate the biological richness of an area, help biologists and conservationists understand animal resources, and can help people assess the richness and diversity of wildlife, recognize endangered and non-endangered species, and establish better conservation plans and management mechanisms. Animal image target detection can be used to detect the disease situation of wild animals, real-time monitoring and prevention of wild animal outbreaks. For example, in the endemic areas of African rinderpest and foot-and-mouth disease, the use of animal image target detection technology can realize the rapid identification and isolation of sick animals in order to control and contain the spread of the epidemic. Animal image target detection can also be used to study animal behavior, such as the interspecies interaction of birds, the habitat, migration and

DOI: 10.2478/ijanmc-2023-0076

predation of animals, and the interaction between animals and humans. Through the analysis of animal image target detection results, we can better understand the ecological habits of animals, behavior patterns and the interaction with the environment. Animal image target detection has important practical value and research significance in the fields of ecology, zoology and conservation biology, and is expected to be widely used in practical ecological protection and animal protection.

The research on object detection of animal images based on convolutional neural networks helps to strengthen the understanding of object detection in deep learning and computer vision, deepen the understanding of convolutional neural networks, help to promote the research progress of deep learning and animal detection, and pave the way for further combining deep learning with all aspects of life in the future.

II. Related work

In the rapid development of computer vision, object detection technology plays a key role, providing basic support for image recognition and understanding. This paper will discuss the overall status of object detection research and the remarkable progress made in this field. It also studies the latest progress of animal target detection, and pays attention to the efforts of domestic and foreign scholars in using deep learning technology to solve the problem of animal recognition. This research direction not only has practical significance for protecting ecological environment and realizing intelligent agriculture, but also provides new challenges and opportunities for the future development of computer vision.

A. Research status of target detection

Computer vision, which was born around 1960, refers to the image recognition technology based on computer algorithms and is the result of the combination of artificial intelligence and cognitive neuroscience. Visual processing also helps us better understand the workings of the human visual system, especially the brain. In the past 20 years, computer vision has experienced two major stages of development, with 2012 as the dividing point. Before 2012, it was mainly the traditional object detection stage, and after 2012, it was the deep learning-based object detection stage [2]. However, due to the limitation of image processing technology and the lack of effective image representation technology at that time, scholars in the traditional object detection stage can only design some complex or non-obvious images to check the computing resources at that time. In 2012, R. Gershick et al. made use of the characteristics of CNN [3] to break through the barriers in the field of target detection, bringing the target detection technology into a new period. In the detection process, algorithms that use deep learning for target detection can be roughly divided into two main lines of development. One is the two-stage algorithm, that is, the detection process is divided into two successive stages. The former uses a network to generate the proposed region, while the latter uses another network to send the proposed region to the classifier for classification [4]. The second is the one-stage algorithm [4] that is, the detection process is only divided into One Stage, and bounding box and classification label are output directly through a network. As a fundamental part of the field of computer vision, object detection has made great progress in the past few years.

Object detection algorithms have made great progress under the deep learning framework, integrating more algorithms and methods to solve real-world problems, and have gradually achieved a balance in terms of speed, accuracy and effectiveness. There are still many difficulties and challenges to overcome in the future, such as the application of small target detection, edge computing and other fields, higher precision and need to face higher complexity of the network structure, but in general, target detection will still be the core field of computer vision.

B. Current status of animal recognition research

At home and abroad, researchers have adopted various methods to explore and improve the animal target detection technology. At present, there are few researches in the field of animal image object detection. In terms of research status at home and abroad, in China, some researchers use deep convolutional neural networks (CNN) to identify wild animals, and some researchers use

transfer learning to solve the problem of livestock target detection in agricultural images. However, these methods have some problems in practice, such as insufficient training samples, low detection accuracy and slow speed. Foreign researchers have proposed a variety of animal object detection methods based on deep learning. For example, methods such as Faster R-CNN, SSD, and YOLO are widely used in the field of animal identification. For example, the European ZooScan project is dedicated to the development of automated aquatic organism classifiers [5] , and uses machine learning algorithms and computer vision technology to achieve fast and accurate classification of biological samples. In addition, National Geographic magazine has published an article introducing a deep learning-based animal object detection technology that can identify dogs, cats, rabbits and other animals.

With the continuous development of convolutional neural network technology, foreign scholars have gradually carried out in-depth research on cats and dogs, and applied it to more complex biometric identification field. In 2017, Tibor Tmovszky [6] used shallow convolutional neural networks to classify and identify animals, such as deer, Wolf, pig, fox, etc., with an accuracy rate of 98%. In 2018, Schneider et al. [7]applied Faster R-CNN to the Reconyx Camera Trap and Snapshot Serengeti groups, and achieved good results of 93.0% and 76.7%. In 2019, Li Anqi [8] proposed an image feature extraction method combining ROI and CNN, which was superior to FastR-CNN to a certain extent. Cheng Zhe'an [9] used the existing wildlife database of the Inner Mongolia Horse Racing Reserve to build a deep residual network containing 6 species, including white stork, wild boar, quagga, deer, mink and spores, and modified the balance loss function of Faster R-CNN, with the mAP value reaching 92.2%.

Animal target detection is an important research direction both at home and abroad, and methods to study this problem are constantly emerging. Scholars are insisting on exploring more efficient and accurate animal target detection methods to cope with more complex and changeable practical application scenarios.

III. TECHNICAL MODEL

A. Convolutional neural network

Convolutional neural network (CNN) is a kind of deep learning neural network for processing high dimensional information such as speech, image, video, and natural language.

CNN network generally uses convolution and pooling operations to preprocess the input data to extract the spatial structure information and timing information in the data. The core idea of CNN is that through the neuronal connections between layers, the network can automatically learn the features of the input data, so as to provide more accurate prediction results when the model makes predictions. As shown in Figure 1.

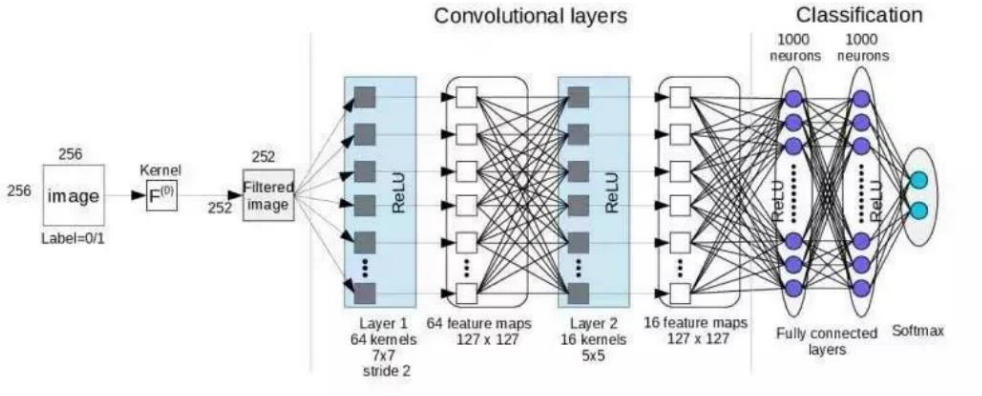

Figure 1. CNN network architecture

B. Object detection algorithm

Target detection is mainly through the input of images or videos, and then use the algorithm to identify the target inside, and determine its position and size, target detection algorithms can usually be divided into the following categories, based on traditional methods; Methods based on deep learning, representative algorithms include Faster R-CNN, YOLO, SSD, etc. Two-stage detector, Faster R-CNN is the representative algorithm; The first stage detector, YOLO and SSD are well-known representative algorithms; Candidate region estimation algorithms, such as RCNN and Faster R-CNN.

C. YOLOv5 algorithm

At present, the YOLOv5 algorithm model is divided into four basic modules: input, Backbone, Neck network and output. Figure 2 shows the structure.

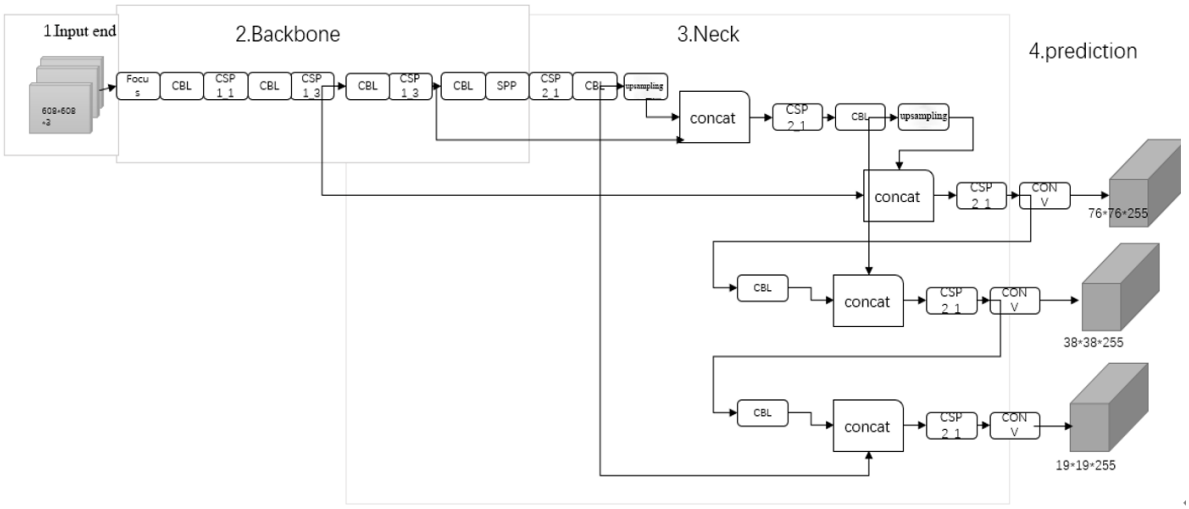

Figure 2. Structure of YOLOv5s

1) Input end

The input image size of YOLOv5 is 608×608, so it generally needs to be preprocessed accordingly. Usually, the original image scale is scaled first and normalized to the range of [0,1]. In the network training part, YOLOv5 not only introduced adaptive anchor frame calculation and adaptive image scaling, but also added Mosaic data enhancement algorithm.

Mosaic data enhancement algorithm. By means of random scaling, random clipping and random arrangement of the four pictures, the data set is enriched, and the training speed and accuracy are improved.

Adaptive anchor frame calculation. This method dynamically adjusts the default bounding box size and shape based on the size and distribution of the object. In the network training stage, the model will output prediction boxes based on these anchor boxes, calculate the difference between them and the real boxes, and then perform reverse updates to update the network parameters [10] . A proper initial anchor box is critical to the accuracy of the target detection algorithm, and for YOLO versions 3 and 4, separate programs need to be set up for different data sets. In YOLOv5, this function has been embedded in the code, which can automatically calculate the best anchor box according to the name of the data set [11] , and the user can turn this function on or off according to the needs during each model training.

Adaptive picture scaling. In order to match different target detection algorithms, we generally scale the image to be detected to a certain size, and then send it to the detection network. However, in the real scene, the aspect ratio of the image is inconsistent, so the commonly used scaling will increase the number of black edges due to the inconsistency of the image, resulting in large data duplication and reducing the reasoning efficiency

of the algorithm. In order to further accelerate the inference efficiency of YOLOv5, this algorithm adopts an adaptive algorithm which can minimize the number of black edges added in the reduced image. Figure 3 shows the implementation effect.

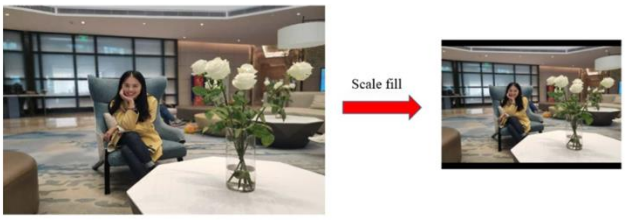

Figure 3. Adaptive picture scaling

2) Backbone network

Generally composed of multi-layer convolution and pooling layers, this module is used to extract some common feature representations. The base network of YOLOv5 uses the CSPDarknet53 and Focus structures. The Focus structure is used to slice the images before they enter the backbone network, and convolution joins the sliced images, thus avoiding the multi-convolution kernel structure and retaining important feature information, as shown in Figure 4. For example, in the network structure of YOLOv5s, the original image with input of 608×608×3 is processed by slicing the Focus structure, and then convolution is processed once by 32 convolution kernels to obtain a feature graph of 304×304×32 [12]. However, different versions of YOLOv5 use different numbers of convolution kernels.

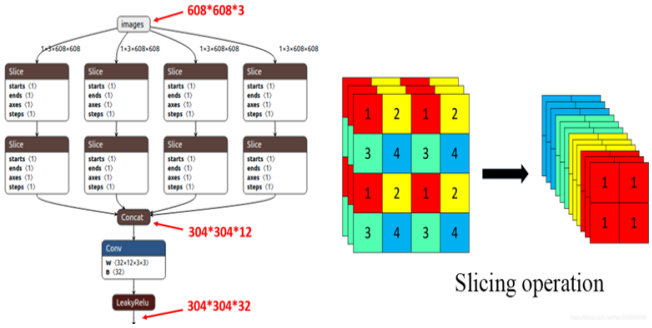

Figure 4. Focus structure

3) Neck network

Neck network is between the baseline network and the header network, which can enhance the diversity and robustness of features [13] . Although YOLOv5 also uses FPN+PAN module, as shown in Figure 3-5, the implementation details are slightly different. Specifically, the FPN structure forms a feature pyramid with multi-scale characteristics by gradually fusing feature maps from low to high levels, which enables the model to process objects of multiple scales at the same time, and improves the detection accuracy of small objects. The Path-aggregation part uses a novel Path-aggregation network for depth feature fusion. The feature maps of different resolutions are connected in series. Meanwhile, the PAN structure weights the feature maps of all scales to optimize the scale invariance. The detection of objects of different sizes and scales is realized, and the detection performance of YOLOv5 is effectively improved. FPN solves the problem of heterogeneity in size, PAN solves the problem of scale, FPN layers convey strong semantic features from the top down, and PAN layers convey positioning features from the bottom up.

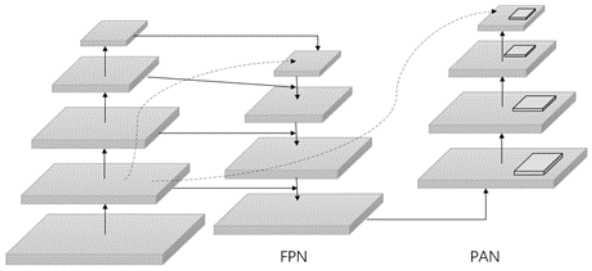

Figure 5. Structure of FPN+PAN

The structure of the whole Neck network is very simple and has very high computational efficiency. In addition, both SPP-Pyramid and Path-aggregation in Neck network can solve some problems existing in YOLOv4, such as false positives caused by unclear counting near edge anchor points and slow network convergence. Therefore, compared with YOLOv4, the Neck network of YOLOv5 not only has better detection performance, but also high computational efficiency and accuracy.

4) Head end

The Head end of YOLOv5 refers to the last layer of the network, which is mainly used to predict the location, category and confidence of the target. Compared with YOLOv4, the Head end

of YOLOv5 has been slightly improved to improve detection performance and speed.

Specifically, the Head end of YOLOv5 adopts GIoU Loss, as shown in equation (1). First, the minimum closure area of the two frames is calculated (popular understanding: The smallest area of the predicted box and the real box), and then calculate the IoU, and then calculate the proportion of the area that does not belong to the two boxes in the closure area, and finally subtract this proportion from IoU to get GIoU. This loss function is dedicated to bringing the predicted box closer and closer to the real box, it is to put the calculated value of the IoU in a norm term that contains the position and shape information of the bounding box, and redefine the IoU so that it can quantify the "proximity" between the two boxes. In this way, the optimizer will backpropagate on the basis of the GIoU objective function, gradually adjusting the output of the network, balancing the algorithm between classification and reconstruction, and achieving better performance.

$$GIoU = IoU - \frac{|A_c - U|}{|A_c|} \quad (1)$$

The Head end of YOLOv5 also uses a new predictive method called Dynamic Convolution, which can effectively adapt to the different shapes and sizes of each target, thus improving the detection performance. In addition, the Head end of YOLOv5 also uses a multi-scale prediction mechanism, that is, by using different sizes of anchor points to detect different sizes of targets. In the prediction, YOLOv5 uses the up-sampling method to fuse the feature maps with different resolutions, which can ensure the detection performance is not lost and improve the calculation speed. The non-maximum suppression (NMS) algorithm used in YOLOv5 is a technique used to remove duplicate boxes in detection results. The main idea is to merge boxes whose overlapping area is larger than a certain threshold, and eventually only one box is retained to represent the target.

The NMS is mainly implemented through the following steps.

Step1. Sort all prediction boxes from highest confidence to lowest

Step2. Traverse each prediction box from top to bottom to check whether the IOU value between this prediction box and all subsequent prediction boxes is greater than a certain threshold (generally 0.5 or 0.6). If the value is greater than the threshold, the prediction box is deleted. Otherwise, the prediction box is retained

Step3. Continue through the next prediction box and repeat the above steps until all prediction boxes have been traversed

Step4. The remaining prediction box is the result after the NMS processing, that is, each object corresponds to only one prediction box [14] .

The YOLOv5 algorithm improves the YOLOv4 algorithm. For example, YOLOv5 uses a new computer vision module, namely SPP module, which can make the difference in the size of the feature map irrelevant, which is conducive to enhancing the feature extraction capability. In addition, YOLOv5 also adopts a new Network design, namely CSP Net (Cross-Stage-Partial Network), which can effectively increase the capacity and feature expression capability of the network, so as to improve the detection accuracy. At the same time, YOLOv5 also applies Mix Up and Cut Mix and other technologies in the training process, which can enhance the generalization ability of neural network. Whether it is a new module design or a new training strategy, these improvements have enabled YOLOv5 to make significant progress in the field of object detection.

The YOLOv5 algorithm has the advantages of higher efficiency, faster speed, higher accuracy, and can be trained on a smaller GPU memory, so it is suitable for target detection scenarios of lightweight devices. At the same time, due to its innovative design, YOLOv5 achieved first place in several target detection competition tasks.

IV. EXPERIMENT AND ANALYSIS

A. Experimental content

Animal image datasets are used to evaluate the proposed target detection method. This dataset

contains images of cats, dogs, birds, horses and sheep. We conducted experiments on this dataset to evaluate the performance of our method under different rounds of training. Experimental results show that the proposed YOLO algorithm has a good effect in the task of animal target detection. Compared with traditional image-based methods and some previously proposed convolutional neural network methods, the proposed YOLO algorithm shows better results in terms of detection accuracy and processing speed.

The environment used in this experiment was Windows 11 system, PyCharm Community Edition 2020.2.1, Python 3.8 and LabelImg.

B. Experimental process

The general flow chart of the experiment is shown in Figure 6.

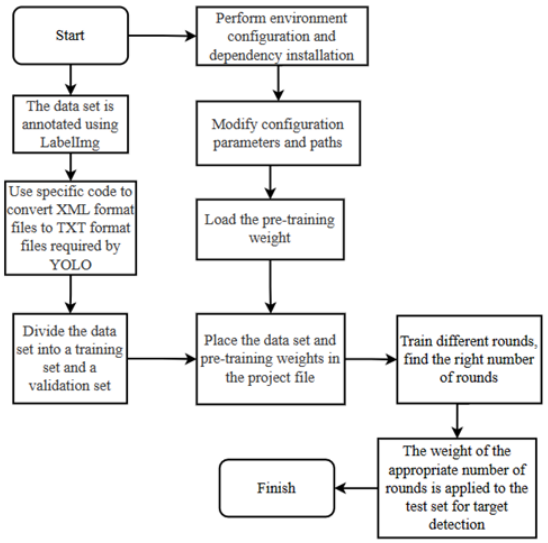

Figure 6. Flowchart of the experiment

First, the target detection data set is annotated using the LabelImg tool. Set VOCdevkit and VOC2007 folders in the project, create JPEG Images in VOC2007 to store unannotated images, store annotated file for Annotations, and create txt file to store the name of the target class. The data set contains five categories and is saved to the Annotations folder after labeling images using LabelImg. Next, convert the XML format to the TXT format required by YOLO. The data set is divided into a training set and a validation set, and the 8:2 ratio is used in this project. Configure the environment, install dependencies, modify parameter paths, and load yolov5s.pt pre-training weights. Download the weights to the project folder, train several times, select the best weights, and train further. Finally, the optimal weight is used to detect the target and the detection result is obtained.

In this experiment, we selected some images from COCO dataset (including horses, sheep and birds) and Cats vs Dogs dataset. COCO dataset is a large and rich object detection, segmentation and captioning dataset, and its images include 91 types of objects, 328,000 images and 2,500,000 labels. There are 80 categories provided, with more than 330,000 images, of which 200,000 are annotated, and the number of individuals in the entire dataset exceeds 1.5 million [15], which are commonly used to train and evaluate advanced computer vision models such as object detection, instance segmentation, and key point detection. The Cats vs Dogs dataset is mainly used for image classification tasks, in which the goal is to distinguish the images of cats and dogs. It is relatively small, including the training set and the test set. The number of pictures of cats and dogs in the training set is 12,500 and sorted in order, and the mixed pictures of cats and dogs in the test set are 12,500 [16]. Used primarily for entry-level computer vision projects to demonstrate how to build and train basic image classification models.

The experiment ran 3, 10, 30 and 50 rounds respectively. After the model is trained, a runs folder will appear, which contains the folders with weights, best is the best weight, last is the weight of the last training, it will automatically do validation, parameters and so on. Next, for inference, first paste the copied and trained best weights into the weights folder, then modify the paths and parameters in the file where detect.py is deduced, and finally put the image path to be detected into the folder in the inference path and run detect.py.

C. Experimental result

The specific data obtained from the experiment are shown in the following figures, including the accurate information of a series of curves during the whole training process and the values of

different loss functions, which are visually displayed in the form of matrix and discount.

The experimental confusion matrix is shown in Figure 7, where each row represents the predicted category and each column represents the true belonging category of the data. The following figure shows that there is little difference between the predicted results of cats, dogs, birds and horses and the true category, but the predicted results of sheep category are not good.

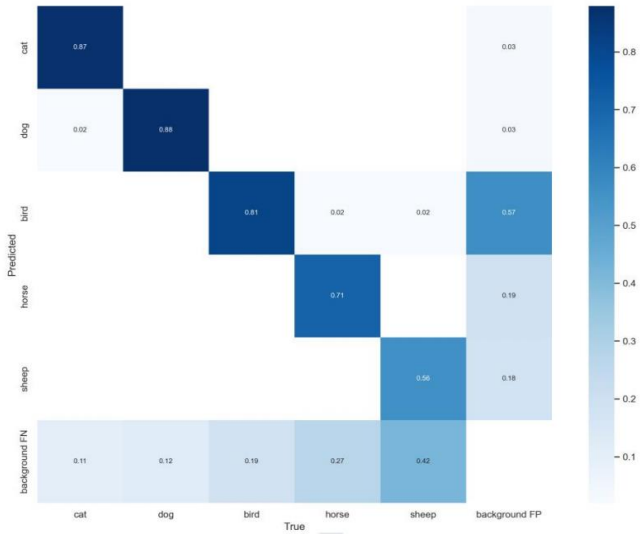

Figure 7. Confusion matrix

As shown in Figure 8, the experimental PR diagram draws a curve between the accuracy rate and the recall rate, showing the performance of the model under different thresholds. The recall rate, that is, the probability that the correct category in the sample is predicted correctly, is shown in formula (2), where TP indicates that the correct category is predicted as the correct category number, and FN indicates that the correct category is predicted as the negative category number.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

The experimental RCC curve is shown in Figure 9, which is a visual representation of the change of recall rate under different confidence thresholds. It is shown in the figure that except for sheep category, other categories can still maintain a high recall rate after filtering out the prediction box with low confidence, indicating that the model has a good performance.

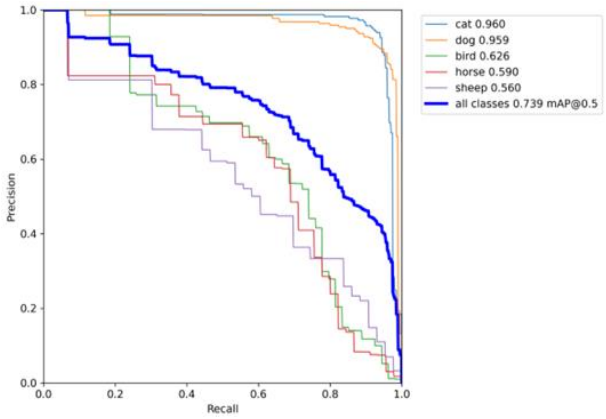

Figure 8. PR curve

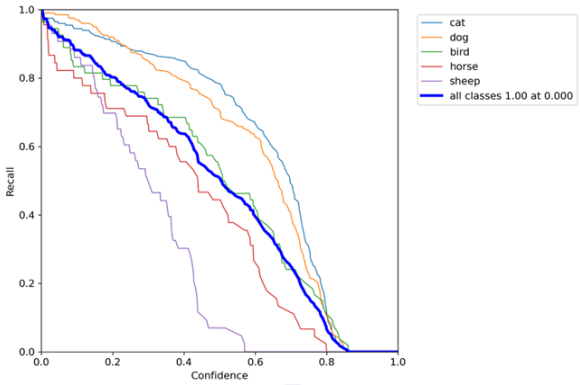

Figure 9. RCC curve

The general figure of the training results of the algorithm on the data set is shown in Figure 10, which comprehensively evaluates the performance of the classification model through different indicators.

The experimental results are shown in Figure 11 and Figure 12, showing the test results of different kinds of animals.

Different rounds of training were conducted in the experiment, and the accuracy, mAP50 and MAP50:95 values were different, as shown in Table 1.

TABLE I. DIFFERENT ROUNDS OF TRAINING

Epoch	Recall	mAP50	mAP50:95
3	0.062	0.009	0.015
10	0.259	0.220	0.119
30	0.507	0.520	0.248
50	0.753	0.739	0.350

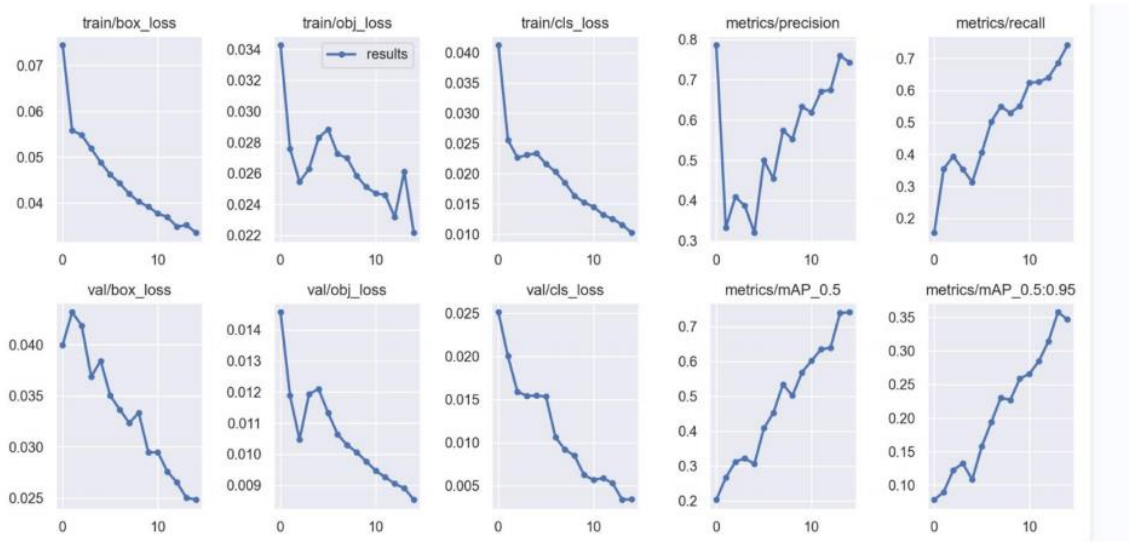

Figure 10. Training results of the algorithm on the data set

Figure 11. Test results of different types of animals(1)

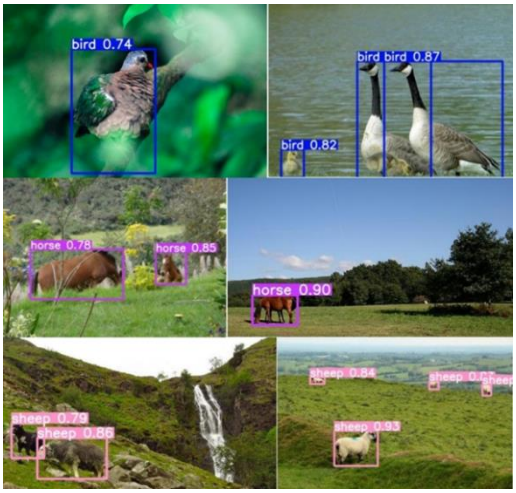

Figure 12. Test results of different types of animals(2)

The above are the results of animal image detection based on YOLOv5 algorithm. This experiment requires different rounds of training, but the increasing number of training rounds will lead to the increase of network pressure, and the training is relatively time-consuming, so this experiment only trains 3 rounds, 10 rounds, 30 rounds and 50 rounds. The accuracy rate, recall rate and loss rate of each round of training are observed. It can be observed that when the number of training rounds is 50, the detection effect is the best, the average accuracy is more than 70%, and the accuracy of some categories can reach 90%. In the later stage, the category of data set and the number of training rounds can be gradually increased to find out whether there is a better detection result.

V. CONCLUSIONS

The traditional target detection algorithm has some disadvantages, such as tedious process, fast convergence and difficult learning, which lead to poor target detection results. The algorithm in this paper is based on a specific model for detection. In order to solve the problems caused by the traditional algorithm, the algorithm will be trained and learned in advance. Through multiple training with different rounds, the training results are compared, and the appropriate number of rounds is

selected for image detection. The algorithm in this paper still has some problems, such as insufficient sample class richness, variable target scale and Angle, etc. In order to further improve the detection accuracy and practical application, future research can consider the introduction of multi-scale training, image enhancement and other technologies, and combine more features and information to improve the detection effect of animal targets in complex scenes in terms of model performance, application expansion, interpretability research and other aspects.

REFERENCES

[1] Gao Hui. Research on Video Object Detection Algorithm Based on Deep Learning [D]. University of Electronic Science and Technology of China, 2021.

[2] Zhang Xin. Image Recognition System of Engineering Vehicle Equipment Based on Deep Learning [D]. Xidian University, 2021.

[3] Zhao Yongqiang, Rao Yuan, Dong Shipeng et al. Review of Deep learning object detection methods [J]. Journal of Image and Graphics, 2019, 25(04):629-654.

[4] Fan Lili, Zhao Hongwei, Zhao Haoyu, et al. Optics and Precision Engineering, 2020, 28(05):1152-1164.

[5] Wang Yuzheng, Cheng Yuan, Bi Hai et al. Marine single-cell algae recognition algorithm based on deep learning VGG network model [J]. Journal of Dalian Ocean University, 2021, 36(02):334-339.

[6] Trnovszky T, Kamencay P, Orjesek R, et al. Animal recognition system based on convolutional neural network [J]. Advances in Electrical and Electronic Engineering, 2017, 15(3):517.

[7] Schneider S, Taylor G W, Kremer S. Deep learning object detection methods for ecological camera trap data [C]. IEEE Conference on Computer and Robot Vision, 2018: 321-328.

[8] Li Anqi. Research on automatic recognition of wildlife monitoring images based on Convolutional neural networks [D]. Beijing Forestry University, 2020.

[9] Cheng Z A. Automatic recognition of terrestrial wildlife in Inner Mongolia based on deep convolutional neural networks [D]. Beijing Forestry University, 2019.

[10] Ma Linlin, Ma Jianxin, Han Jiafang et al. Research on Object Detection Algorithm based on YOLOv5s [J]. Computer Knowledge and Technology, 2021, 17(23):100-103.

[11] Zhou Wenhui, JIA Yonghong, Jiao Yang. Research on detection method of Chinese sturgeon in underwater video [J]. Computer Science and Applications, 2022, 12(8): 1998-2005.

[12] Fan Youchen, Ma Xu, Ma Shuli et al. Evaluation method of laser interference effect based on deep learning [J]. Infrared and Laser Engineering, 2021, 50(S2): 39-45.

[13] Lin Sike, Chen Jinwei, Huang Sihua. Research on Student Behavior Detection based on Deep Learning [J]. China Journal of Multimedia and Network Teaching (Mid-day), 2022(06):237-240.

[14] Jin Y. Research on pig quantity monitoring method based on machine vision [D]. Jiangxi Agricultural University, 2021.

[15] He Yuzhe, He Ning, Zhang Ren et al. Research on Training Unbalance of Deep Learning-oriented object detection Model [J]. Computer Engineering and Applications, 2022, 58(05):172-178.

[16] Xu Bo. Research on Object Detection and Semantic Segmentation Algorithms based on Deep Learning [D]. Northeastern University, 2019.