2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) | 978-1-6654-6468-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICML57342.2022.10009855

YOLO-Animal: An efficient wildlife detection network based on improved YOLOv5

Ding Ma
Ningxia University
School of Information
Engineering
Yinchuan,China
official_martin@163.com

Jun Yang(Corresponding author)
Ningxia University
School of Information
Engineering
Yinchuan,China
dragon@nxu.edu.cn

Abstract—**With the continuous development of modern society, the rapid expansion of human civilization squeezes the living space of other organisms, and the extinction of more and more biological species has sounded the alarm for us. Therefore, in order to timely understand the changes of wild animals and other resources in a specific area, and facilitate relevant personnel to formulate effective restoration and protection measures, this paper proposes a wild Animal species recognition network based on deep learning and improved YOLOv5: YOLO-Animal. The application of artificial intelligence and computer vision has covered all aspects of human's daily life and work. The benchmark YOLOv5 algorithm can quickly and accurately deal with problems associated with images. Through the fusion of weighted Bidirectional Feature Pyramid Network (BiFPN) and Effective Channel Attention (ECA) module, the original YOLOv5s network structure is enhanced, and the detection accuracy of small targets and occluded and fuzzy targets is effectively improved. The YOLO-Animal model outperforms the benchmark YOLOv5s model by 3.2% on mAP and achieves 95.5% accuracy on the test set.**

Keywords—animal detection; YOLO - Animal; BiFPN; ECA

I. Introduction

In recent years, with the rapid development of deep learning and computer vision, object detection, as one of its important tasks, has also been widely used in real life, such as medical diagnosis, intrusion detection, etc. Object detection techniques aim to identify and detect objects of certain object classes in a given image and assign corresponding class labels to each object class. These techniques are carried out in different forms in different network architectures, such as extracting candidate boxes, formulating training strategies and activating and optimizing functions[1].

Animal resources are a valuable wealth of ecological resources. Protecting animal resources is equivalent to protecting human beings themselves. However, with the increasing production activities such as open-pit sand mining and coal mining, the development of human society also affects the living space and conditions of wild animals. While the ecological environment is being destroyed, the wild animals living there are also facing severe challenges for survival. Therefore, wild animals need regular detection and identification by human beings to prevent the reduction of animal species diversity. In order to better protect animal resources and detect animal species in real-time, object detection technology has been applied to this field, which will locate and classify wild animals in the target image. In this paper, the object detection algorithm based on deep learning is combined with intelligent surveillance cameras, infrared cameras, Beidou satellites, etc., to solve the related problems that will occur in traditional detection, such as fuzzy animal targets, unclear discrimination between near and far scenic spots, etc., which has good practical value.

At present, the target detection algorithm based on deep learning has developed into two technical routes: Anchor-based method (one-stage, two-stage) and Anchor-free method. The two-stage algorithm generates a candidate region (a pre-selected box that may contain the object to be detected), and then classifies the samples by convolutional neural network. Its recognition accuracy is high, but its real-time performance is poor. Instead of generating candidate regions, the one-stage algorithm extracts features directly from the network to locate and classify objects. The final detection result can be obtained directly after a stage, so it has a faster detection speed while ensuring accuracy. However, traditional object detection algorithms cannot meet the needs of wildlife species recognition due to the complex field ecological landform and environment.

In this paper, a lightweight, efficient and fast wildlife detection network model is designed based on the benchmark YOLOv5s algorithm. The model integrates weighted Bidirectional Feature Pyramid Network (BiFPN)[2] and Effective Channel Attention (ECA) module to improve the network's attention to local important information. Compared with the previous algorithms, it has a better performance in solving the problems of different degrees such as blurred target, indistinct distinction between near and far scenic spots, and occluded target for the wild animal target in the image.

II. Related work

In recent years, many researchers have been trying to assign tags to images to identify animals. After analyzing 25 types of animals, Atri Saxena et al.[3] built a dataset containing 31,774 images and designed a target detection model based on the one-stage algorithm SSD and the two-stage algorithm Faster R-CNN[4]. The mAP obtained by this method is 80.5%. Mai Ibraheam et al.[5] used four different R-CNN models[6] and a deformable convolutional neural network to improve the accuracy and speed of animal species detection in order to reduce the negative impact of human-wildlife encounters on highways, and applied them to three wildlife datasets, and achieved good results. Ramakant Chandrakar et al.[7] proposed an animal detection and recognition system based on genetic segmentation and convolutional neural network, which uses genetic algorithm in the segmentation process and three-layer neural network in the classification process. The mAP obtained by this method is 99.02%. Mengyu Tan et al.[8] first constructed the wildlife image dataset of Northeast Tiger and Leopard National Park, and selected the YOLOv5 series model, the cascade R-CNN under the feature extractor HRNet32, etc. The experimental results show that the target detection model of day and night joint training has good performance, and the mAP obtained is 98%. Chiagoziem et al.[9] used a deep feature pyramid design with lateral connections to make semantic features of a small project more sensitive. Due to the densely connected convolutional network, function transmission and reuse are enhanced throughout the classification phase, leading to more accurate classification with fewer parameters. Arshita Verma et al.[10] developed a self-trained animal detection model, which uses machine learning and artificial intelligence algorithms to classify vector machines, k-mean nearest neighbors and group trees. The experimental results demonstrate that the classification accuracy of the system is up to 91%.

Li Anqi[11] proposed an automatic wildlife monitoring image recognition algorithm based on ROI and convolutional neural network. The object detection method based on regression algorithm is used to detect and segment the wildlife area in the monitoring image, and generate ROI(Region Of Interest) image, so as to reduce the interference of complex background information on species recognition. Jiang Fuhao et al.[12] applied Swin-Transformer technology[13] to the wildlife target detection model and compared its performance with other excellent detection models. The experimental results show that compared with other excellent detectors, the average detection accuracy of Swin-Transformer is 0.928, which is at least 5% ahead of other detection models.

According to the preliminary investigation, due to the complex landscape and ecological environment in the wild, the size of animal targets in the camera and image is affected by the angle of near and far, and the same kind of animals may have different sizes, large near and small far, which also leads to the small scale of common large animals in the picture. In addition, the animal targets in the image may also have different degrees of problems, such as target ambiguity, unclear distinction between near and far scenic spots, target occlusion, light intensity and so on, which will increase the difficulty of real-time detection to a certain extent, reduce the accuracy of wildlife species identification technology, resulting in missed detection, false detection and other problems.

Because the accuracy of these proposed algorithms is not high, in order to solve the problem of low detection accuracy of the existing algorithms, this paper summarizes the existing related work, and finally designs an improved wildlife species recognition algorithm based on YOLOv5 -- YOLO-Animal.

III. YOLO-Animal Model

A. Lightweight channel attention mechanism

This paper introduces the attention mechanism based on baseline. Its core idea is to find the special correlation between features based on the basic data, and then focus on some important features, mainly including channel attention, pixel attention, multi-order attention, etc.

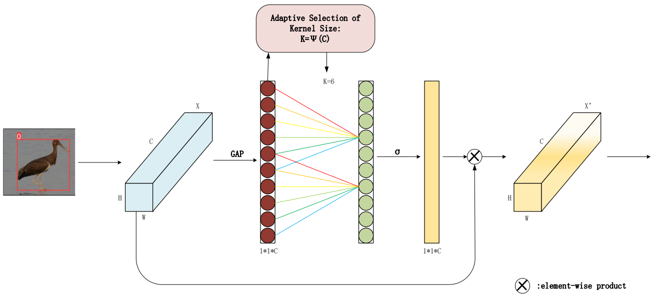

Fig. 1. ECA module architecture.Feature fusion is carried out by global average pooling, and channel weights are further generated by a one-bit convolution of size k.

The channel attention mechanism has been proven to significantly improve the performance of convolutional neural networks. SE-Net[14] improves the performance of the network by modeling the channel relationship, first obtaining a weight matrix through operation, and then reconstructing the corresponding features. The two most important operations of SE-Net are Squeeze and Excitation. In this process, there are two questions: Is the corresponding spatial information lost in the process of global average pooling the spatial information of the features into the channels, which gives each channel a large receptive field? So in the case of Excitation, why go through two fully connected layers when there are a lot of parameters in there? To solve the above problems, Qilong Wang et al.[15] proposed an improved lightweight channel attention mechanism -- ECA-Net in 2020. ECA-Net uses a local cross-channel interaction strategy without dimensionality reduction, which can be effectively implemented by one-dimensional convolution. Its module architecture is shown in Figure 1. ECA-Net focus on all the input channel spatial information, and is worth to pay close attention to choose the part, and gives different weights to different channels, rather than before the convolutional neural network gives the same weight to all channel, so as to make the network more important to focus to extract information, wildlife detection accuracy can be improved. The introduction of this mechanism will help the network to deal with the detection of more than one type of organism in an image, and improve the distinction between near and far spots. The fine-grained ECA module reduces the error detection rate of targets with similar features and makes the model more robust.

B. Multi-channel feature extraction and fusion

In order to better perform multi-channel feature extraction and feature fusion, a new neck structure is introduced: BiFPN (weighted Bidirectional Feature Pyramid Network). The neck structure adopted by the baseline is PANet, which establishes a bottom-up pathway on the basis of fpn. The high-level feature map has stronger semantic information, which is conducive to object classification. The underlying feature map has stronger location information, which is conducive to object positioning. Although the fpn structure improves the semantic information of the predicted feature map, it theoretically loses a lot of location information. In the detection of wild animals, most of them are images with insufficient resolution and clarity. PANet will not be sufficient in feature extraction, resulting in inaccurate target positioning. Unlike PANet, which has only one top-down and bottom-up path, BiFPN processes each bidirectional path as a feature network layer and repeats the same layer multiple times to achieve a higher level of feature fusion. Its module architecture is shown in Figure 2.

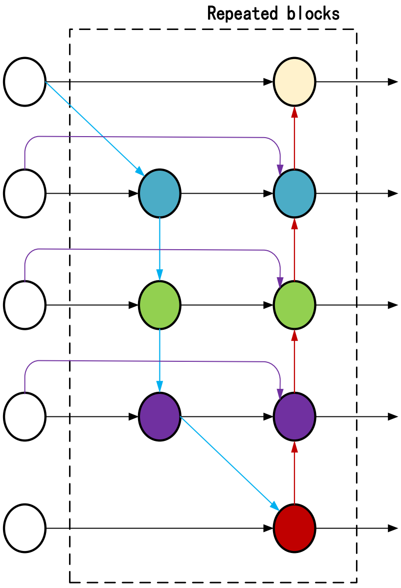

Fig. 2. BiFPN module architecture

C. depth-separable convolution

In order to reduce the number of model parameters, quantify parameters, reduce the memory occupied and further improve the detection speed, this paper introduces depth-separable convolution[16] to meet the real-time detection requirements. Aiming at the problem of robustness of the model, the data enhancement technology[17] is used to further improve its reliability and robustness.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Dataset preparation and training

In order to achieve ideal results, it is necessary to establish a certain number and high quality wildlife data sets. The data source of this wildlife dataset is provided by a national nature reserve, and the corresponding quality is guaranteed. Data will be screened again after data is obtained. On this basis, this paper also obtains high-quality wildlife image and video data from the Internet to expand the dataset. In addition, in the process of data set construction, data cleaning[18], data normalization, data enhancement[19] and other operations will be carried out, and finally 3500 labeled data sets will be obtained. This dataset mainly contains eight types of wildlife commonly found in a national nature reserve, as shown in Table I.

TABLE I. DATASET DESIGN

Data content	Training set	Validation set	Test set
Black Stork	400	100	100
Red-tailed Plover	200	100	100
Golden Eagle	300	100	100
Blue Horse Chicken	200	100	100
Yak	200	100	100
Lynx	150	100	100
Pika	150	100	100
Snow Leopard	300	100	100

The experiment environment is Ubuntu18.04 system, GTX2080Ti graphics card server, CUDA10.1, Python3.7. Set the maximum number of iteration rounds to 1000 and adjust the different hyperparameters for it to expect the best results. The initial learning rate was 0.1. After three rounds of warm-up training, the learning rate was dynamically adjusted, the batch size was set to 64, the optimizer was set to SGD, and the adaptive anchor frame was set to optimally match the wildlife dataset used in this experiment. The final results show that the model converges at round 619 and achieves the best effect of 95.5%. The final network training and detection effect diagram and each curve are shown in Figure 3 and Figure 4, respectively.

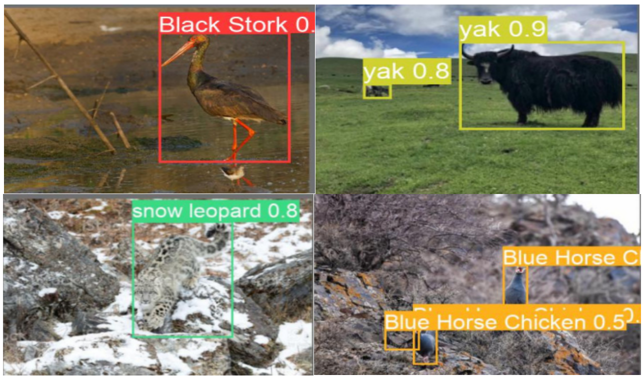

Fig. 3. Test results 1, 2, 3, 4

B. Data analysis and comparison

We calculated the detection accuracy, recall and mAP of the model through a test set containing 800 samples. The calculation formula is as follows:

$$P = \frac{TP}{TP+FP} \qquad (1)$$

$$R = \frac{TP}{TP+FN} \qquad (2)$$

TP is true positive, FP is false positive, FN is false negative, TN is true negative.

$$AP = \sum_{i=1}^{n-1}(r_{i+1} - r_i)P_{inter}(r_i + 1) \qquad (3)$$

The r1, r2... Rn is the Recall value corresponding to the first digit of the Precision interpolation segment in ascending order.

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \qquad (4)$$

This model takes YOLOv5s as baseline. Ablation experiments have been conducted on ECA and BiFPN optimized modules of this model, and the specific results are shown in Table II:

TABLE II. ABLATION EXPERIMENT RESULTS

Model	Precision	Recall	mAP
yolov5s	0.900	0.923	0.933
yolov5s_BiFPN	0.917	0.896	0.939
yolov5s_BiFPN_ECA(YOLO -Animal)	**0.926**	0.918	**0.955**

The ablation experiment results show that the proposed YOLO-Animal can effectively improve the detection effect of wild animals on the same dataset. Compared with YOLOv5s model, our model can better extract important features in complex situations and pay attention to the spatial and semantic information we need to pay attention to more quickly. In addition, the addition of BiFPN module further enables our model to carry out feature extraction and feature fusion at a deeper level, which can be faster and more efficient when dealing with complex image backgrounds and many natural environmental factors.

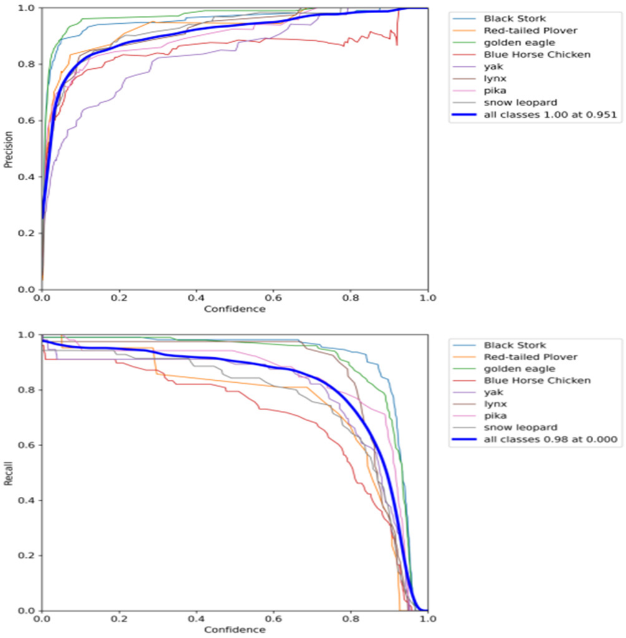

Fig. 4. Precision curve and Recall curve

Furthermore, we selected the classical algorithm in the field of object detection at the present stage and conducted a comparative experiment with our YOLO-Animal model. These algorithms cover classical one - stage and two - stage algorithms. The experimental results are shown in Table III:

TABLE III. COMPARATIVE EXPERIMENTAL RESULTS

Model	Precision	Recall	mAP
SSD	0.893	0.826	0.912
Faster-RCNN	0.742	0.961	0.947
yolov5s	0.900	0.923	0.933
YOLO-Animal	**0.926**	0.918	**0.955**

Target detection algorithms based on deep learning mainly include one-stage algorithm, two-stage algorithm and Anchor-Free algorithm. The two-stage algorithm generates candidate regions, and then classifies samples by convolutional neural network. The representative algorithms include R-CNN, Fast R-CNN[20], Faster R-CNN, etc. Instead of generating candidate regions, the first-stage algorithm directly uses convolutional neural network to extract features and further classify and locate objects. Representative algorithms include: YOLO series[21], SSD series[22], etc. Through the comparative experiment, it is not difficult to find that the detection accuracy and accuracy of our algorithm model have been optimized and improved to a certain extent, no matter comparing with the one-stage SSD algorithm with a high detection speed or the two-stage Faster-RCNN algorithm with a high detection accuracy. Due to a phase algorithm has the disadvantage of objects for small target detection effect is bad, the recall rate is not high, YOLO - Animal model in order to solve this problem to join the ECA attention mechanism, gives different weights to different channel or area, let the model network focus on value, worthy of attention, to reduce the influence of noise on the detection results. In view of the real-time guarantee required for wild Animal detection, the baseline algorithm does not select a two-stage algorithm. According to the corresponding real-time requirements, the YOLO-Animal model uses deep separable convolution, which reduces the number of model parameters and memory occupancy while further improving the detection speed[23].

V. CONCLUSIONS

The rapid development of deep learning and computer vision provides innovative solutions to a series of problems such as biodiversity detection in the wild. However, due to the complexity of the wild environment, the quality of the collected wild animal images will be seriously affected by the light, part of the content is blurred, and the front and back diameter boundary is not clear. In this paper, an efficient and fast wildlife species recognition network based on improved YOLOv5 -- YOLO-Animal is proposed, which integrates ECA attention mechanism and BiFPN structure on the basis of YOLOv5s model to solve the above problems. The YOLO-Animal model aims to extract and integrate features efficiently and fully, make the network pay attention to important channels and spatial information, give different channels or regions different weights, and better identify small target objects. The experimental results show that the detection performance of the proposed algorithm is significantly improved compared with the unimproved baseline algorithm YOLOv5s. Compared with other classical target detection algorithms, the detection accuracy is also improved. In terms of detection speed, in order to meet the requirements of real-time detection, the fusion depth of YOLO-Animal model can be separated by convolution, which further reduces the amount of model parameters and improves the deduction speed of the model. At the same time, the capacity of wildlife dataset used in this study is not sufficient and rich enough, which leads to the failure of the final model to achieve the experimental test effect when it is actually landed. In the future, we will continue to enhance cooperation with a national nature reserve to conduct further research on the direction of wildlife detection based on reality, and strive to improve the detection accuracy and cover more wildlife under the premise of ensuring the detection speed.

ACKNOWLEDGMENTS

Research and demonstration on key Technologies of intelligent monitoring,early warning,protection and restoration of biodiversity in Helan Mountain based on Beidou(2021BEG02007)

REFERENCES

[1] Chen H. D., Ding S. Y., Liu Y. X.. A review of deep learning-based target detection algorithms [J]. Journal of Beijing Union University, 2021, 35(03): 39-46.

[2] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[3] Saxena A, Gupta D K, Singh S. An animal detection and collision avoidance system using deep learning[M]//Advances in Communication and Computational Technology. Springer, Singapore, 2021: 1069-1084.

[4] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[5] Ibraheam M, Li K F, Gebali F, et al. A performance comparison and enhancement of animal species detection in images with various R-CNN models[J]. AI, 2021, 2(4): 552-577.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[7] Chandrakar R, Raja R, Miri R. Animal detection based on deep convolutional neural networks with genetic segmentation[J]. Multimedia Tools and Applications, 2021: 1-14.

[8] Tan M, Chao W, Cheng J K, et al. Animal Detection and Classification from Camera Trap Images Using Different Mainstream Object Detection Architectures[J]. Animals, 2022, 12(15): 1976.

[9] Ukwuoma C C, Qin Z, Yussif S B, et al. Animal species detection and classification framework based on modified multi-scale attention mechanism and feature pyramid network[J]. Scientific African, 2022, 16: e01151.

[10] Verma A, Sangwan V, Shukla N. Wild animal species detection using deep convolution neural network[M]//Recent Trends in Communication and Electronics. CRC Press, 2021: 406-410.

[11] Li Anqi. Research on automatic recognition method of wildlife monitoring images based on convolutional neural network [D]; Beijing: Beijing Forestry University, 2020.

[12] Jiang Fuhao, Sui Chenhong, Ou Shifeng, et al. Wildlife detection based on Swin-Transformer [J]. Artificial Intelligence and Robotics Research, 2021, 10: 281.

[13] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[14] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[15] Wang, Qilong; Wu, Banggu; Zhu, Pengfei; Li, Peihua; Zuo, Wangmeng; Hu, Qinghua (2020). [IEEE 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) - Seattle, WA, USA (2020.6.13-2020.6.19)] 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) - ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. , (), 11531–11539. doi:10.1109/CVPR42600.2020.01155

[16] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[17] Wightman R, Touvron H, Jégou H. Resnet strikes back: An improved training procedure in timm[J]. arXiv preprint arXiv:2110.00476, 2021.

[18] Wu T., Zhang B., Wang Y. F., et al. A review of data cleaning research[J]. Modern Library and Information Technology, 2007, 2(12).

[19] Hendrycks D, Mu N, Cubuk E D, et al. Augmix: A simple data processing method to improve robustness and uncertainty[J]. arXiv preprint arXiv:1912.02781, 2019.

[20] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[21] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[22] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[23] Hannun A, Lee A, Xu Q, et al. Sequence-to-sequence speech recognition with time-depth separable convolutions[J]. arXiv preprint arXiv:1904.02619, 2019.