Hindawi
Scientific Programming
Volume 2022, Article ID 2752334, 6 pages
https://doi.org/10.1155/2022/2752334

*Research Article*

# Analysis of Improved YOLO Algorithm in English Translation

**Ling Ye[1,2] and Peng Yin[3]**

[1]*School of Foreign Languages, Hubei Engineering University, Xiaogan City, Hubei Province 432000, China*
[2]*Faculty of Education, Universiti Kebangsaan, Bangi, Malaysia*
[3]*Center of Information Technology, Hubei Engineering University, Xiaogan City, Hubei Province 432000, China*

Correspondence should be addressed to Peng Yin; yp_2830@hbeu.edu.cn

Received 30 March 2022; Accepted 25 May 2022; Published 4 July 2022

Academic Editor: Punit Gupta

As China becomes more and more international, the number of people traveling abroad is also increasing. The demand for English recognition is becoming more and more vigorous, and traditional translation software is time-consuming, laborious, and less accurate. This article optimizes the target detection model YOLOV3. Firstly, the image is divided into multiple model structures, and the K-means++ clustering algorithm is used to determine the target detection prior frame value and the high frame of the corresponding frame according to the characteristics of the English image. Then, by using K-means++ clustering algorithm to optimize the anchor parameters, the model structure is better adapted to the English identification dataset scene; finally, the feature information extracted by the DarkNet-53 model is spliced to improve the structure of the YOLOV3 convolutional layer, using 3090 graphics card GPU to perform multiscale training and testing. Experimental results show that the improved YOLOV3 algorithm in this paper has a mAP of 0.95 on the English identification dataset and a detection speed of 50fps, which is 0.11 higher than the mAP before optimization. Therefore, optimizing the YOLOV3 algorithm in this article has a good effect. In the future, English translation will become a necessary software program for Chinese people to go abroad.

## 1. Introduction

As our country's economic development becomes more and more international, the number of people going abroad is also increasing. Therefore, many people start to travel abroad, but since domestic social English communication is still in its infancy, few people can communicate directly in English [1]. At the same time, using translation software directly is time-consuming and labor-intensive, which is very cumbersome for those who go abroad for work or business trips. Therefore, there is an urgent need for a software program that can realize translation to meet the needs of people's daily life abroad and improve the effect of traveling abroad.

With the continuous development of deep learning, the target detection technology using deep learning has attracted more and more attention and research by many scholars [2]. When object detection is applied in various directions, its application fields are also getting wider and wider [3]. For example, Xie and Li [4] designed the ILF-YOLOV3 template detection algorithm, which solved the problems of real-time monitoring and management of the on-the-job status of enterprise employees, and the algorithm design improved the accuracy of the model by about 0.8. Feng and Gao [5] designed and proposed a vehicle target shape and position fusion algorithm, which combines stereo vision and lidar technology to perform real-time target detection on vehicles, improving the efficiency and safety of autonomous driving. Lei from Shanghai Jiaotong University [6] proposed a small target algorithm of "detection first and tracking later" under the complex background condition, which further improves the application effect of the algorithm in complex backgrounds. Although the above target detection algorithms have achieved a certain degree of effect, the application field is not a real-time scene and is not suitable for English word detection and translation in scenarios. In view of this, this paper proposes an algorithm model of the improved YOLO algorithm in the English translation scene. This model can better solve the characteristics of slow translation and low accuracy in the English translation scene.

## 2. Related Work

As an important branch in the field of deep learning, the target detection algorithm is mainly divided into two parts, namely, the one-stage and two-stage models, which also have their own advantages and disadvantages. They are mainly distinguished by whether there is a region proposal stage. The one-stage model represents the YOLO (you only look once) series; the two-stage model represents RCNN, SPP-net, Fast-RCNN, Faster-RCNN, and Mask R-CNN. All along, the one-stage detection speed is higher than the two-stage network, but the detection accuracy is indeed lower than the two-stage network. But with the continuous innovation of the YOLO series, YOLOV3 is superior to other classic two-stage models in terms of accuracy and speed.

The starting point of YOLOV3 is optimized from YOLOV2, which solves the problem of YOLOV2's low detection accuracy for small objects. As a small, beautiful, fast, and accurate target detection network, it is highly praised by scientific researchers. The detection model is based on the target detection DarkNet-53 framework, and the input image size of the model is uniformly set to a picture of 416 × 416 × 3. In view of this, the English word recognition system in the English environment proposed in this paper is optimized on the YOLOV3 algorithm model.

## 3. YOLOV3 Algorithm Improvement

*3.1. DarkNet-53 Model Structure.* The YOLOV3 algorithm replaces the DarkNet-19 and ResNet [7] network model structures of YOLOV2 [8] by using DarkNet-53 [9] for feature extraction. The DarkNet-53 feature extraction model consists of 52 convolutional layers and one fully connected layer. The convolution kernels of 3 × 3 and 1×1 are alternately used for convolution operations, and the convolution kernels of 3 × 3 are used for convolution operations. It is to reduce the amount of calculation and the number of parameters. The function of the convolution kernel of 1 × 1 size is to keep the number of input and output channels consistent. The structure of the DarkNet-53 network model is shown in Figure 1, where [1, 2, 4, 8] represents the number of convolution blocks.

*3.2. The Network Improvement of YOLOV3 Algorithm in This Paper.* DarkNet-53 used by YOLOV3 has excellent feature extraction ability compared to the DarkNet-19 network. However, with the deepening of the network model depth, while improving the accuracy and recall rate of the network model, there will also be problems such as disappearance of features, and there is a balance between the performance of the two. Therefore, this paper optimizes the YOLOV3 algorithm structure in order to better adapt to the deep learning feature model of English translation scenarios, reduce the impact of environmental factors on the recognition of characteristic English words, and improve the accuracy and speed of detection. The first two groups of convolutional layers of the 3 YOLO layers are removed, respectively. The optimized YOLOV3 algorithm has a total of 101 layers, including 69 convolutional layers, 23 staggered block layers, 4 feature extraction layers, 2 upsampling layers, and 3 layers. The structure of the optimized network model is shown in Figure 2.

In this paper, the image is first scaled to a 3-channel color image, and the long and wide graphs of the network model. Afterwards, the DarkNet-53 network model is used for feature extraction, and the convolution kernels of 3 × 3 and 1 × 1 sizes are used alternately, thereby avoiding the degradation of some data as the structure of the network model deepens. Also, we reduce the number of convolutional layers, introduce the skip connection of the residual network model, and divide the output three feature dimensions of 77 layers, 87 layers, and 93 layers, respectively, to $13 \times 13 \times 512$, $26 \times 26 \times 768$, and $52 \times 52 \times 384$. Dimensionality reduction is used as the input to the YOLO [10] layer. Therefore, the feature map information of three scales is used for training to obtain the final weight model value, so as to extract the English information in the image.

*3.3. The Parameter Improvement of YOLOV3 Algorithm in This Paper.* The anchor parameter is used as a set of a priori boxes with fixed width and height values in the YOLOV3 network model. The size of the prior frame directly affects the accuracy and speed of target detection. Therefore, when training English words or sentences in images, it is particularly important to set the parameters of the model according to the characteristics of the words and sentences. In order to better adapt to the characteristics of words and sentences in the image and achieve the optimal effect of training, this paper uses the K-means++ clustering [11] algorithm to replace and compare the K-means clustering algorithm in YOLOV3 to perform clustering analysis on image labels [12].

In order to reduce the error caused by the size of the a priori frame itself to the target value, the intersection ratio between English words or sentences and the prior frame is used to replace the original network in the setting of anchor parameters by using K-means++ and K-means clustering algorithms. The size of the objective function value usually represents the deviation between each sample value and the center of the clustering algorithm [13]. Therefore, the smaller the deviation, the better the effect. Among them, the calculation method of the objective function value $M$ is as follows:

$$M = \text{Min} \sum_{b=0}^{i} \sum_{c=0}^{j} \left[1 - IOU_c^b\right], \tag{1}$$

where $b$ represents the target frame of English image labels, $c$ represents the cluster center, $i$ represents the number of datasets, and $j$ represents the number of categories of images.

*3.4. Detection Method.* The detection environment in this paper is an English-speaking detection environment for experiments and is optimized in a YOLOV3-based

| Type | Filters | Size | Output | |
|---|---|---|---|---|
| Convolutional | 32 | 3×3 | 256×256 | |
| Convolutional | 64 | 3×3/2 | 128×128 | |
| Convolutional | 32 | 1×1 | | |
| Convolutional | 64 | 3×3 | | ×1 |
| Residual | | | 128×128 | |
| Convolutional | 128 | 3×3/2 | 64×64 | |
| Convoluti onal | 64 | 1×1 | | |
| Convolutional | 128 | 3×3 | | ×2 |
| Residual | | | 64×64 | |
| Convolutional | 256 | 3×3/2 | 32×32 | |
| Convolutional | 128 | 1×1 | | |
| Convolutional | 256 | 3×3 | | ×8 |
| Residual | | | 32×32 | |
| Convolutional | 512 | 3×3/2 | 16×16 | |
| Convolutional | 256 | 1×1 | | |
| Convolutional | 512 | 3×3 | | ×8 |
| Residual | | | 16×16 | |
| Convolutional | 1024 | 3×3/2 | 8×8 | |
| Convolutional | 512 | 1×1 | | |
| Convolutional | 1024 | 3×3 | | ×4 |
| Residual | | | 8×8 | |

Figure 1: DarkNet-53 model structure diagram.

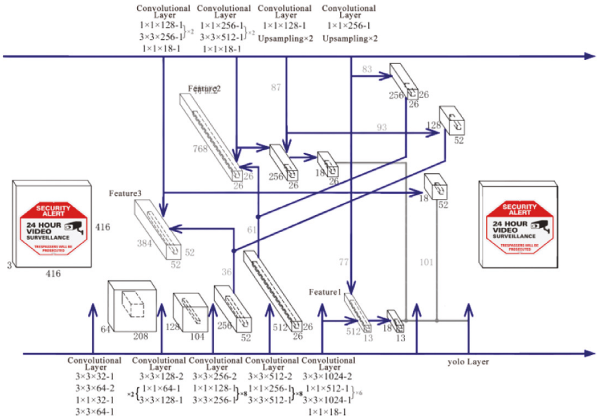

Figure 2: Improved structure diagram of YOLOV3 model.

environment. In order to submit the detection speed and simplify the detection network model, this paper removes the two convolutional layers of YOLOV3, respectively. The detection process includes a total of 6 steps, which are as follows:

(1) Preprocess the English logo images in the training set and use the processed images as the input values of the model.

(2) The input value is sent to DarkNet-53 for feature extraction of the model, and the extracted 77th layer is used as a feature vector value, and upsampling and downsampling are performed at the same time.

(3) The output of the 61st and 83rd layers is feature spliced to obtain the feature information of the second feature map, and an upsampling and convolution operation is performed on it.

(4) Perform feature splicing on the output values of the 36th and 93rd layers to obtain the feature information of the third feature map.

(5) The feature information of the above three feature maps is sent to the YOLO model layer for model training, and the final model weight is generated after the iteration is completed.

(6) Input the test set image in this paper into the model, call the weight value obtained from the model training to perform target detection on the English logo image of the test set, and do not output the detection result; the schematic diagram of the algorithm in this paper is shown in Figure 3.

## 4. Experimental Analyses

*4.1. English Logo Image Dataset.* The selection of training set and test set has an important influence on the learning efficiency of the model, which further affects the accuracy of

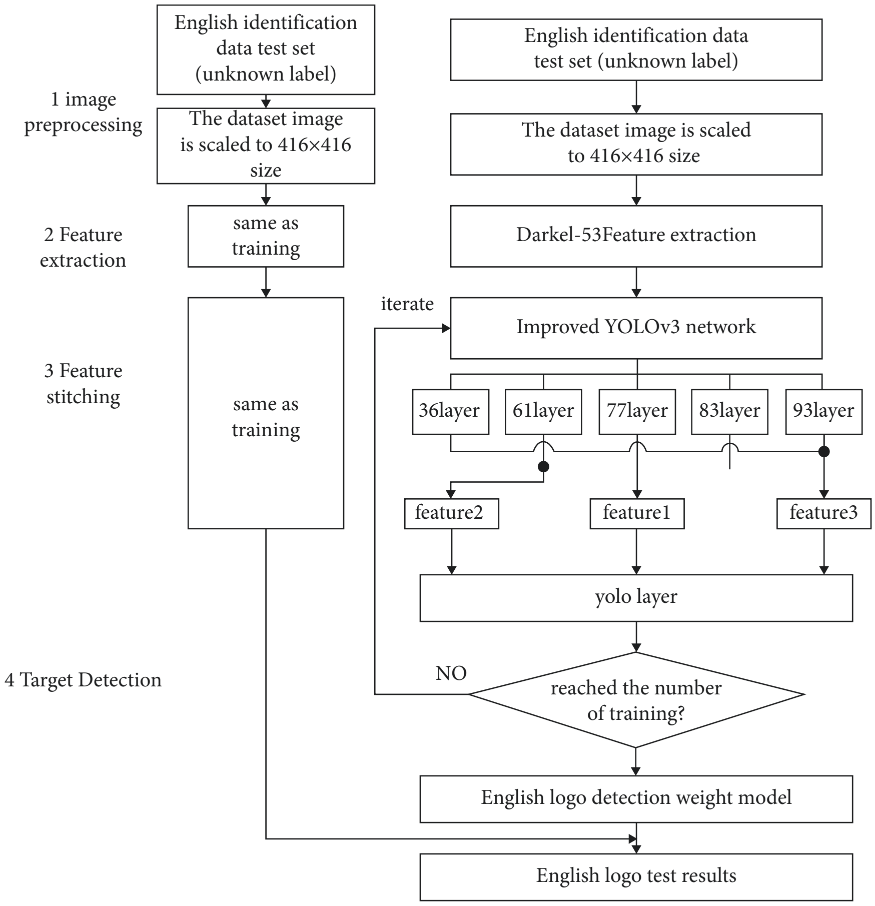
English identification data test set (unknown label)
The dataset image is scaled to 416×416 size
English identification data test set (unknown label)
The dataset image is scaled to 416×416 size
1 image preprocessing
2 Feature extraction
same as training
Darkel-53Feature extraction
3 Feature stitching
same as training
iterate
Improved YOLOv3 network
36layer
61layer
77layer
83layer
93layer
feature2
feature1
feature3
yolo layer
4 Target Detection
NO
reached the number of training?
English logo detection weight model
English logo test results

Figure 3: Schematic diagram of the algorithm in this paper.

the test set [14]. This paper uses the English public dataset provided by the University of Chinese Academy of Sciences, which contains a total of 1300 English label datasets. In this paper, the English labeling dataset is first organized according to the VOC2007 dataset, and the dataset is classified according to the ratio of 9 : 1, of which the training set accounts for 90% and the test set accounts for 10%, and then image label is used to label the dataset. The generated data files are in extensible markup language format.

*4.2. Cluster Analysis of Dataset Labels.* Because the VOC2007 dataset [15] used does not contain datasets related to the English label dataset and because YOLOV3 used in this paper for direct model training will have a certain impact on the training effect, this paper uses the K-means algorithm and the K-means++ algorithm to perform dimensional clustering analysis on the labels of the English label dataset. With the change of the $K$ value, the curve of the objective function $D$ changes as shown in Figure 4. After clustering, the prior orbital width and height are shown in Table 1.

*4.3. Experimental Environment.* This article uses the Ubuntu20.04 operating system, the video card memory size is 24G 3090i type, CUDA version is 8.0, and cuDNN version is 6.0. With the continuous increase of the number of iterations in this article, the loss function curve of the model is shown in Figure 5. The change in size of the ratio is shown in Figure 6.

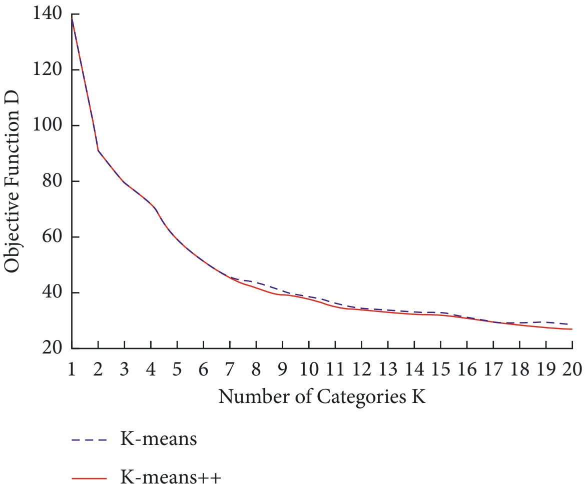

Figure 4: The change curve of the objective function value corresponding to different $k$ values.

It can be seen from Figure 5 that the initial value of the loss function is 1.8. When the number of iterations continues to increase, the value of the loss function also

Table 1: Comparison of prior orbital width and height.

| $K = 7$ | $K = 8$ | $K = 9$ | $K = 10$ | $K = 11$ |
|---|---|---|---|---|
| (6, 9) | (6, 9) | (6, 9) | (5, 12) | (5, 7) |
| (10, 15) | (8, 12) | (9, 14) | (5, 17) | (7, 11) |
| (13, 21) | (11, 17) | (12, 18) | (7, 11) | (10, 14) |
| (19, 30) | (15, 24) | (15, 24) | (10, 14) | (10, 18) |
| (27, 44) | (20, 32) | (20, 32) | (11, 18) | (13, 20) |
| (36, 60) | (26, 43) | (26, 43) | (15, 24) | (16, 25) |
| (141, 10) | (36, 69) | (32, 51) | (20, 32) | (21, 32) |
| * | (141, 10) | (40, 69) | (27, 44) | (26, 43) |
| * | * | (141, 10) | (36, 60) | (32, 51) |
| * | * | * | (141, 10) | (40, 70) |
| * | * | * | * | (141, 10) |

In the table, "*" represents none, which is also a critical value.

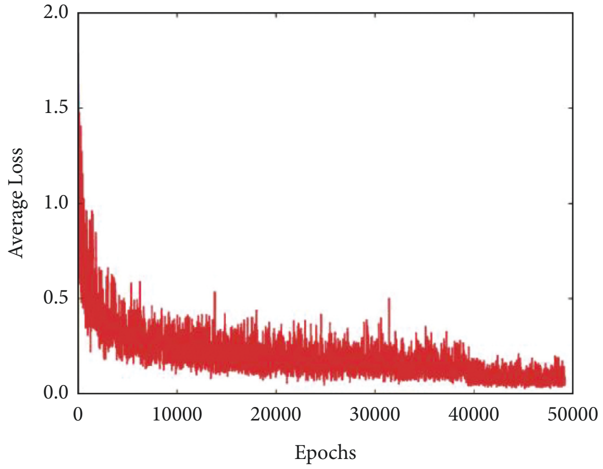

Figure 5: Loss function change curve.

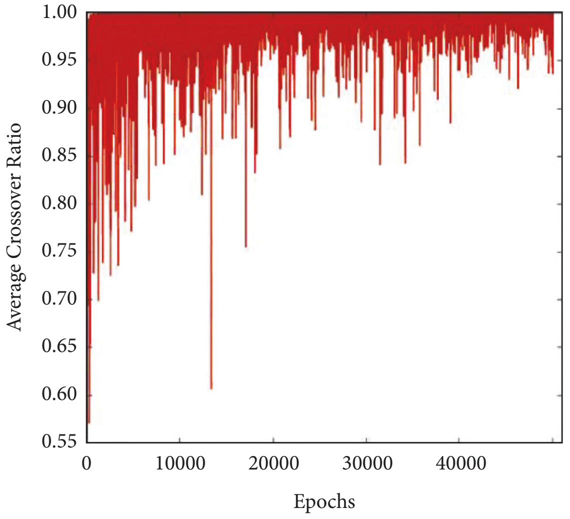

Figure 6: The change curve of the cross-to-bin ratio.

Table 2: Model structure performance comparison.

| Model | Average test time (s) | Average accuracy (%) | mAP (%) |
|---|---|---|---|
| Caltech | * | * | 72.3 |
| VPGNet | * | * | 88.4 |
| YOLOV3-107 | 0.021 | 8.9 | 84.4 |
| YOLOV3-101 | 0.019 | 0 | 89.8 |
| YOLOV3-K-107 | 0.021 | 2.2 | 91.4 |
| YOLOV3-K-101 | 0.019 | 0 | 95.3 |

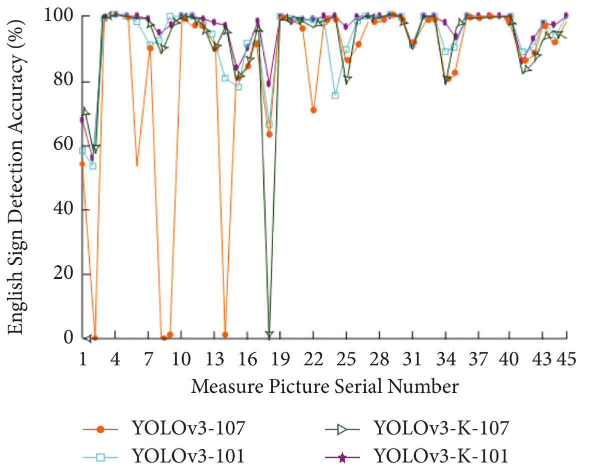

Figure 7: The accuracy curve of the single English logo test set in four different models.

decreases continuously and tends to a stable state when it drops 40,000 times and is basically stable when it drops 50,000 times, reaching convergence. It can be seen from Figure 6 that the initial value of the cross-union ratio is 0.6. As the number of iterations continues to increase, the accuracy of the model is also constantly improving. When the number of iterations reaches 11,000, the model's cross-union ratio is stable above 0.9. The above two figures show that the average loss and average cross-union ratio reach a stable state after iteration, and the model hyperparameter settings in this paper are in line with this experiment.

### 4.4. *Algorithm Performance Comparison.*

The training set and test set used in this paper are a total of 2250, and the data enhancement makes the dataset five times larger. The training set includes 1800 images, and the test set includes 450 images. The hyperparameter weight decay coefficient set in this paper is 0.045, the initial learning rate is 0.001, the threshold size is set to 0.25, the epoch is set to 50000, and the anchor parameters are shown in Table 1. The result values of width and height are obtained after using the K-means++ algorithm.

Compared with our dataset, using "YOLOV3-107 layer," "YOLOV3-101 layer," "YOLOV3-107 layer," and "YOLOV3-K-101 layer" is validated by four experiments.

Using the weights of the feature values generated by each training to test the English logo test set, the test time, missed detection rate, and accuracy are shown in Table 2. Figure 7 shows the accuracy curves of the four different models for the single English label test set.

From Table 2 and Figure 7, it can be seen that the YOLOV3-101 layer of the improved model structure in this paper reduces the detection speed by 2 milliseconds compared to the unimproved model structure YOLOV3-107 layer, increases the mAP by 5.4%, and reduces the probability of showing off the shoulders of the English logo. At the same time, using the K-means++ algorithm for clustering reduces the missed detection rate of the model structure by about 6.7% and increases the mAP by about 7% compared with the model without the K-means++ algorithm for clustering. It is concluded that the clustering effect using the K-means++ algorithm is the best, the final mAP is less than 95.3%, the speed is 50fps, and the probability of missed detection is better reduced.

## 5. Conclusion

With the increasing number of people traveling abroad, the requirements for English translation software are also increasing. This paper improves and optimizes the target detection model YOLOV3 and proposes a YOLOV3-K-101 layer English logo recognition model algorithm. Its main contributions include introducing the YOLOV3 algorithm into the English recognition model, using the K-means++ algorithm to replace the original K-means algorithm and optimize the YOLOV3 anchor algorithm parameters, which increases the mAP of the model by 0.7, and it simplifies the YOLOV3 model structure, and a YOLOV3-K-101 layer model is proposed, which makes the detection speed reach 50fps, reducing the occurrence of missed detections in the English dataset. Due to some limitations of the method proposed in this paper, the generalization ability of the algorithm model will be further improved in the later stage. In the future, the use of deep learning methods to translate English words will be developed to target speed and effectiveness.

## Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Conflicts of Interest

The authors declare that they have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Acknowledgments

This study was supported by the National Foreign Language Teaching and Research Project of 2021 (Research on the Construction and Application of Digital Resources in College English Textbook in the Context of Blended Learning) (grant no. 2021HB0446).

## References

[1] F. Yang, W. Xing, and W. Zhang, "An analysis of the blended teaching mode of College English," *Foreign Language Teaching*, vol. 12, no. 1, pp. 21–28, 2017.

[2] F. Zhou, L. Jin, and J. Dong, "Review of convolutional neural network research," *Journal of Computers*, vol. 40, no. 6, pp. 1229–1251, 2017.

[3] J. Yuan, X. Ma, G. Han, S. Li, and W. Gong, "Research on lightweight disaster classification based on high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 11, p. 2577, 2022.

[4] B. Xie and N. Li, "Research on the detection algorithm of personnel on-the-job status based on ILF-YOLOv3," *Journal of Taiyuan University of Science and Technology*, vol. 42, no. 6, pp. 441–448+455, 2021.

[5] M. Feng and X. Gao, "Research on the fusion algorithm of vehicle target shape and position based on stereo vision and lidar," *Journal of Instrumentation*, vol. 42, 2021.

[6] Y. Lei, *Research on Infrared Small Target Detection and Tracking Algorithm under Complex Background Conditions*, Shanghai Jiaotong University, Shanghai, China, 2006.

[7] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: generalizing residual architectures[J]," 2016, https://arxiv.org/abs/1603.08029.

[8] J. Zhang, M. Huang, X. Jin, and X. Li, "A real-time Chinese traffic sign detection algorithm based on modified YOLOv2," *Algorithms*, vol. 10, no. 4, p. 127, 2017.

[9] H. Wang, F. Zhang, and X. Liu, "Fruit image recognition based on DarkNet-53 and YOLOv3," *Journal of Northeast Normal University*, vol. 52, no. 4, pp. 60–65, 2020.

[10] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers," in *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, pp. 2503–2510, IEEE, Piscataway, NJ, U.S.A, December 2018.

[11] D. Arthur and V. Sergei, *k.-means++: The Advantages of Careful seeding*, Stanford, Stanford, CA, U.S.A, 2006.

[12] G. Hamerly and C. Elkan, "Learning the k in k-means," *Advances in Neural Information Processing Systems*, vol. 16, pp. 281–288, 2003.

[13] J.-G. Sun, J. Liu, and L. Zhao, "Clustering algorithms research," *Journal of Software*, vol. 19, no. 1, pp. 48–61, 2008.

[14] X. Xiaoming, *SVM Parameter Optimization and its Application in Classification [D]*, Dalian Maritime University, Dalian, China, 2014.

[15] B. Jin, L. Cruz, and N. Goncalves, "Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123649–123661, 2020.