An Improved Target Detection Algorithm model for Garment image Detection

Chunrui Yang[1],Weiwei Tian[3]
Faculty of Computer Science and Technology
Guangdong University of Technology
Guangzhou, Guang Dong Province, China
1334436753@qq.com

Lichen Zhang[2]
Faculty of Computer Science and Technology
Guangdong University of Technology
School of Information Engineering
Guangzhou Vocational and Technical University of Science and Technology
Guangzhou, Guang Dong Province, China
zhanglichen1962@163.com

***Abstract*—With the rapid development of the Internet platform, users can choose and match clothes according to their personal preferences without leaving home. Merchants have manually sorted and uploaded a large number of clothing images to make it easy for users to shop for clothing online, which consumes a huge amount of labor costs. Such problems can be improved through deep learning related algorithms. However, the conventional deep learning model has a huge amount of computation, resulting in low efficiency of real-time detection of clothing, which limits its application field. Aiming at these theoretical and practical problems, this thesis studies the optimization of clothing image detection and label recognition methods based on deep learning. In view of the real problems of high computational load and slow instant response of existing clothing detection models. This thesis proposes a clothing detection model YOLOv4-GS based on a deep learning framework. Experiments show that compared with the model YOLOv4, this model has a great improvement in detection accuracy and model efficiency. This algorithm first uses the K-means++ clustering method to preprocess the initial dataset DeepFashion2. And construct the GS module based on the deep fusion of Ghost module and SimAM attention mechanism. Then use the GS module to reconstruct the YOLOv4 network to obtain the model YOLOv4-GS, which has higher efficiency and higher model accuracy.**

Keywords-component; Clothing detection; Deep learning; YOLOv4; DeepFashion2; YOLOv4-GS

I. INTRODUCTION

In the direction of garment detection based on deep learning algorithms, many canonical datasets and algorithmic models with significant results already exist, which greatly facilitate the development of garment detection. liu et al. unified the annotation rules for attributes such as color, category and semantics of garments and completed a dataset conforming to this standard called DeepFashion [1]. ge et al. found that the dataset DeepFashion has drawbacks, such as only sparse keypoint annotations for garments in each image, and bounding boxes are estimated based on keypoint markers, which can easily generate noise. To solve the existing problems, Ge proposed DeepFashion2 [2], which has a more comprehensive image annotation and is a new large-scale dataset with wide coverage. It includes a range of tasks such as clothing detection and recognition. ge et al. tried to experiment DeepFashion2 using an improved model based on Mask R-CNN [3] and achieved an mAP value of 66.7%. The huge number of parameters and floating point operations of Mask R-CNN hinders the deployment of such models on low energy devices and limits the application of such models due to the need to apply clothing detection to life.

Alexey et al. [4] conducted a deeper study for DeepFashion2 article. A lightweight model based on the CenterNet [5] algorithm was proposed to address the problem that target detection models are difficult to apply to low-power devices. The model has low computational complexity in the garment detection task and is able to maintain high accuracy while running on low-power devices.Alexey et al. found that the existing solutions, although providing good recognition accuracy, are slow and require a large amount of computational resources. Therefore, a single-stage approach was proposed to overcome this obstacle and provide fast garment detection [6], and the network achieved a running speed of 17 FPS on a smartphone.

Therefore, the study of deep learning algorithms based on lighter weight is of crucial importance to extend the application of detection models in the field of garment detection. Many use Two-stage methods for garment inspection, which have good accuracy but are too slow to achieve real-time results. Those using One-stage methods such as SSD [7] for garment inspection are faster but not as accurate. Given the efficiency disadvantages of Two-stage algorithms, Iandola et al. proposed the lightweight neural network SqueezeNet [8], which performs model compression by reducing the number of parameters.Howard et al. proposed MobilNets [9], which uses point-by-point convolution to construct convolutional layers and achieves better performance.Han et al. proposed the new lightweight network GhostNet with a novel Ghost module that generates more feature maps by a simple linear operation, outperforming existing lightweight networks in terms of efficiency and accuracy. However, since these lightweight networks sacrifice the accuracy of detection, they are not conducive to some real-life applications. the YOLOv4 algorithm achieves a balance of speed and accuracy and is well received. yolov4 has made a great splash since its birth and has been applied to many fields.

With the rapid development of e-commerce technology, the scale of the online market of China's apparel industry is growing in spurts. The sales methods such as direct broadcast with goods allow users to feel the actual wearing

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00014

effect of the clothes more realistically. However, there are many types of clothing, and it is difficult for users to efficiently and accurately discover the type of clothing they like. How to quickly and accurately classify different types of live clothing has important research significance.

In this paper, we try to apply the more efficient target detection model YOLOv4 to clothing detection and improve YOLOv4 by introducing Ghost module and SimAM attention mechanism, reconstructing the whole backbone network, and obtaining YOLOv4-GS model. It is experimentally compared and applied to the garment inspection service.

II. YOLOv4-GS Model

A. Dataset Introduction and Processing

In this paper, DeepFashion2 was chosen as the dataset for training and testing. The dataset contains 13 popular clothing categories and 800,000 clothing samples, including tank tops, suspenders and other clothing types. Each target body in the image has characteristics such as occlusion, scaling, viewpoint, bounding box, dense landmark and pixel mask, and the trained model will have good robustness.

Due to the small sample size of the short-sleeved jacket and suspenders categories in the DeepFashion2 dataset, in order to balance the dataset, some clothing images are intercepted from the network and manually labeled using LabelImg software to generate an XML file. This file contains the size of the images and the coordinate values of the upper left and lower right corners of the target bounding box. In addition we perform data augmentation on these images by rotating and masking them to expand the sample size and thus improve the robustness of the model.

B. Model Description

Considering the efficiency problem of existing garment detection, smaller scale and lower parametric number models are needed, even applying the more efficient target detection model YOLOv4 to garment detection is not the best choice. In this paper, we propose a smaller, less computationally intensive, and more accurate YOLOv4-GS target detection model. While ensuring high efficiency and high accuracy for garment inspection, the model is made lightweight to ensure that it meets the needs of garment inspection. The model takes into account that different anchor frames affect the algorithm accuracy, and the model introduces the K-means++ clustering algorithm to cluster and analyze the target frames in the DeepFashion2 dataset to filter out the appropriate anchor frames as the initial candidate frames to improve the detection accuracy. The model proposes an efficient and low-computation GS module, which is obtained by combining the lightweight Ghost module and the parameter-free 3D attention mechanism SimAM, which can effectively reduce the number of parameters and improve the network feature extraction performance. the Ghost module is used for model scale compression, and the SimAM attention mechanism is to improve the model detection performance without increasing the number of parameters. This algorithm uses the GS module to reconstruct the backbone network CSPDarkNet53 of YOLOv4 and improve it to get the GS-CSPDarkNet53 network. This network has a smaller number of model parameters and computation volume, thus improving the model operation efficiency and achieving the goal of designing a garment detection model with high accuracy, small size and high efficiency.

1) Introduction of K-means++ clustering

In the target detection model, the introduction of the anchor frame transforms the target detection problem into the problem of detecting the presence of a target in a fixed grid and the deviation of the prediction frame from the true frame, and the setting of the prior frame plays a crucial role in the prediction results. For the problem that the K-means [12] algorithm randomly selects the initial points leading to the slow convergence of the algorithm and the clustering results are not very good, the K-means++ algorithm is a good solution to this problem. The algorithm in this paper uses the K-means++ clustering algorithm to generate an adapted prior frame to increase the detection accuracy of the model.

The final obtained a priori boxes conform to the dimensions of a large number of garment types, which can effectively improve the detection accuracy of the model during model training. The algorithm in this paper outputs a total of 3 feature layers, so 9 Anchor boxes are generated. The Anchor boxes corresponding to the width and height of the garment dataset taken by this algorithm are shown in Fig. 1.

Feature map	Scale	Anchors
13*13	Big	(243,168),(325,254),(403,371)
26*26	Middle	(155,198),(202,134),(215,299)
52*52	Small	(76,68),(90,143),(134,87)

Figure 1. Anchors corresponding to clothing dataset

2) Introduction of Ghost module

During the training process of neural networks, a certain number of redundant feature maps are generated. han et al. found that these redundant feature maps can be obtained by mapping some of the feature maps using linear operations. To avoid the redundant feature maps from generating a large number of convolution operations, the normal convolution layer can be divided into two parts: one part is the normal convolution operation, and the other part is the linear operation to generate the feature maps obtained from the convolution operation. This operation is the Ghost module, which reduces the number of parameters and computation of the whole model, as shown in Fig. 2.

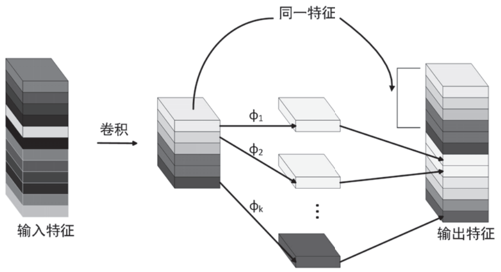

Figure 2. Ghost convolution module

The amount of convolution operation is mitigated by Ghost, but the richness of the feature map generated in this way is also reduced, which leads to the recognition accuracy being affected. In this paper, we propose the GS module, as shown in Fig. 3. The main part of the GS module consists of two Ghost convolutional blocks and the SimAM attention mechanism combined. The complexity of the convolution operation is effectively reduced by using the residual mechanism. Finally, the results of the two parts of the feature map of the same size are feature summed separately to make the feature representation stronger. the GS module enables the model to have lower computational effort and stronger feature extraction capability.

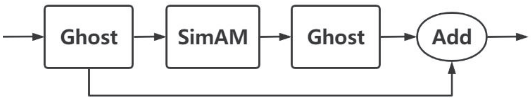

Figure 3. GS module

3) *Introduction of SimAM attention mechanism*

The SimAM parameter-free attention mechanism ensures lightness and efficiency by requiring no additional parameters in the derivation of attention weights for feature maps. It is used to find important neurons by measuring the linear differentiability between neurons and assigning them a higher priority. Using binary labels and adding regular terms, the energy function of each neuron is defined and the minimum energy calculation is shown in Fig. 4.

$$e_t(\omega_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(\omega_t x_i + b_t))^2 + (1-(\omega_t t + b_t))^2 + \lambda {\omega_t}^2$$

$$e_t{}^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda}$$

Figure 4. Energy function and Minimum energy calculation method

4) *Build GS-CSP module*

In order to reduce the model parameters and computation, this paper combines the Ghost module and GS module to get the GS-CSP module, as shown in Fig. 5. One part of the module is two Ghost blocks and GS blocks, and the other part consists of one Ghost block, and finally the results of the two parts of the same scale of the feature map are carried out separately for the feature channel Concat. this module can improve the difference of gradient union, avoid learning duplicate gradient information, effectively alleviate the gradient disappearance, and secondly greatly reduce the amount of model computation.

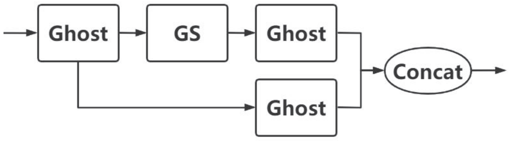

Figure 5. GS-CSP module

5) *YOLOv4-GS model*

In this paper, a YOLOv4-GS model based on YOLOv4 is built, as shown in Fig. 6. The Backbone of the model consists of Ghost blocks and GS-CSP blocks to reconstruct the YOLOv4 backbone network consisting of GS-CSPDarkNet53 network, which has the features of light weight and strong feature extraction capability. The Neck of the model is expanded by the SPP structure to expand the sensory field, pooled using four different scales of maximum pooling kernels respectively, and then feature maps of different kernel sizes are connected as the output, while the PANet feature pyramid structure is used to adapt the target detection at different Levels using parameter aggregation. The Head of the model uses three different sizes of output heads to adapt to the detection of large, medium and small targets.

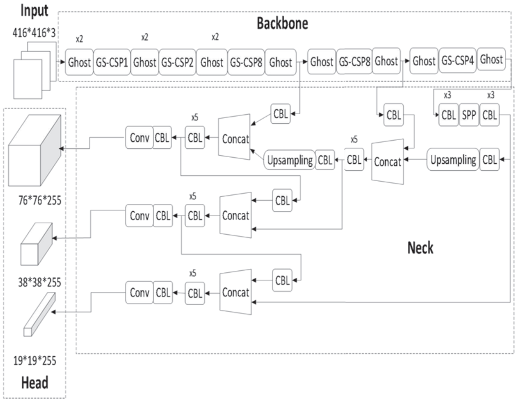

Figure 6. YOLOv4-GS algorithm framework

The SPP structure in the YOLOv4-GS model uses four different scales of maximum pooling processing after the last feature layer of GS-CSPDarkNet53 as a way to increase the network perceptual field and separate the contextual features. the CBL structure is the component in the YOLOv4-GS model that consists of a convolutional layer, a batch normalization layer, and a Leaky Relu activation function. The two structures are shown in Fig. 7.

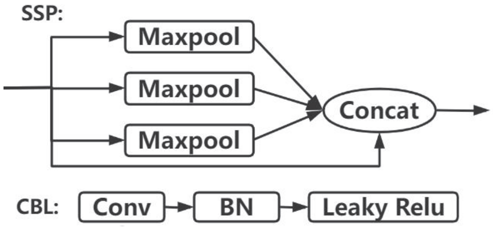

Figure 7. SPP and CBL structure diagram

III. Analysis of Experimental Results

A. *Evaluation Metrics*

The model uses precision, recall, PR (precision-recall) curve, AP, mAP, IOU, etc. to measure.

B. *Quantitative analysis of experimental results*

YOLOv4-GS model training has converged after 100 Epochs of training, train loss, val loss, smooth train loos and smooth val loss.

Through ablation experiments, the GS module and the K-means++ algorithm are introduced in turn, and Fig. 8. focuses on the model's backbone network computational power, backbone network parametric number, model size, and mAP value for comparison. With the introduction of GS module on the basis of YOLOv4 model, the results show that the computational power and the number of parameters are relatively decreased by about 50%, the volume is reduced by 33.12%, and the mAP is improved by 1.6%. With the introduction of the K-means++ algorithm, the mAP is again improved by 0.5%, and the mAP value reaches 73.1%.

Model	Backbone FLOPs	Backbone params	Model weight	mAP
YOLOv4	17.34 GB	26.62M	245.8M	0.710
YOLOv4 + GS	8.96 GB	13.41M	164.4M	0.726
YOLOv4 + GS + K-means++	**8.96 GB**	**13.41M**	**164.4M**	**0.731**

Figure 8. Experimental comparison

As shown in Fig. 9., the AP values of the YOLOv4 and YOLOv4-GS models for 10 clothing classes are demonstrated. In most classes, the AP values of YOLOv4-GS are improved relative to YOLOv4, which verifies that the YOLOv4-GS model proposed in this paper has a strong detection capability.

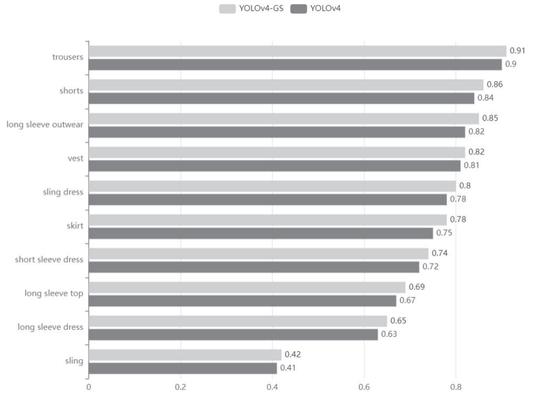

Figure 9. Comparison diagram of YOLOv4 and YOLOv4-GS AP

In this paper, the current mainstream deep learning target detection models SSD, RetinaNet, Faster-RCNN[13], R-FCN and our model are compared, as shown in Fig. 10., and the results show that the model in this paper performs better in terms of Precision, Recall and mAP values. The YOLOv4-GS model obtained after improving the YOLOv4 backbone network with GS module has a large performance improvement and has a good prospect in the field of target detection.

Model	Backbone	Precision	Recall	mAP
SSD[7]	VGG16	0.668	0.611	0.653
RetinaNet[11]	ResNet50	0.684	0.647	0.678
Faster-RCNN[4]	ResNet50	0.688	0.651	0.682
R-FCN[10]	ResNet50	0.721	0.672	0.705
Our model	GS-CSPDarkNet53	**0.763**	**0.695**	**0.731**

Figure 10. Model performance comparison

IV. CONCLUSION

In this paper, we propose a lighter and more efficient YOLOv4-GS target detection algorithm, which solves the problem that the model cannot combine efficiency and accuracy. The algorithm uses K-means++ clustering algorithm for the dataset to obtain better a priori frame parameters. the Ghost module and SimAM attention mechanism are fused to obtain the GS module, which is used to reconstruct the whole YOLOv4 model and obtain a YOLOv4-GS model with small parameters, small size and high performance.

Experimental results on the DeepFashion2 dataset show that the scale of this model is reduced by 33.12% and the mAP is improved by 2.2% compared to the native YOLOv4 algorithm. Among the current target detection networks, the detection performance of the YOLOv4-GS model takes an advantage. The experiments demonstrate that the algorithm of this paper reduces the scale and improves the performance of the deep learning model by introducing the GS lightweight module, which is a good theoretical support for the subsequent application of the deep learning model on mobile devices. The lighter and more efficient model proposed in this paper can play a better adaptation in the field of apparel, and at the same time extends the application area of deep learning in the direction of apparel.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant（No. 61873068）.

REFERENCES

[1] Liu Z, Luo P, Qiu S, et al. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations[C]// Computer Vision & Pattern Recognition. IEEE, 2016.

[2] Ge Y, Zhang R, Wang X, et al. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[3] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988.

[4] Sidnev A, Trushkov A, Kazakov M, et al. DeepMark: One-Shot Clothing Detection[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, 2019.

[5] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.

[6] Sidnev A, Krapivin A, Trushkov A, et al. DeepMark++: Real-time Clothing Detection at the Edge[C]// 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2021.

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

[8] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size[J]. arXiv preprint arXiv:1602.07360, 2016.

[9] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[10] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision.2017: 2980-2988.

[12] Estlick M, Leeser M, Szymanskii J J, et al. Algorithmic Transforms In The Implementation Of K-Means Clustering On Reconfigurable Hardware[J]. Office of Scientific & Technical Information Technical Reports, 2000, 27(3):203-206.

[13] [REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Trans on pattern analysis and machine intelligence, 2017, 39(6): 1137-1149.