Clothing attribute recognition algorithm based on improved YOLOv4-Tiny

Meihua Gu (gumh2001@163.com)
Xi'an Polytechnic University
Jie Liu
Xi'an Polytechnic University
Wei Hua
Xi'an Polytechnic University

Research Article

Keywords: Clothing attribute recognition, Res2Net, Multi-scale features, Feature fusion, K-Means clustering

Posted Date: June 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1678879/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Signal, Image and Video Processing on May 16th, 2023. See the published version at https://doi.org/10.1007/s11760-023-02580-5.

Abstract

Aiming at the problem of low accuracy of clothing attribute recognition caused by factors such as scale, occlusion and beyond the boundary, a novel clothing attribute recognition algorithm based on improved YOLOv4-Tiny is proposed in this paper. YOLOv4-Tiny is used as the basic model, firstly, the multi-scale feature extraction module Res2Net is adopted to optimize the backbone network, the receptive field size of each layer of the network is increased, and more abundant fine-grained multi-scale clothing feature information is extracted. Then, the three feature layers of the output of feature extraction network are up-sampled, and the high-level semantic features and shallow features are fused to obtain rich shallow fine-grained feature information. Finally, K-Means clustering algorithm is employed to optimize the anchor box parameters to obtain the anchor box that is more compatible with the clothing object, and to improve the integrating degree between the clothing attribute characteristics and the network. The experimental results show that the proposed method can significantly improve the accuracy of clothing attribute recognition for different scale proportions, occlusion degrees and out-of-bounds degrees, and its mean average prediction(mAP) is improved by 6.75% compared with the original model.

1 Introduction

With the development of computer vision, the analysis and understanding of clothing images is a very active research topic in recent years, and how to detect and recognize clothing images is a current research hotspot, which can be applied to many fields such as clothing image retrieval [1], matching recommendation [2], pedestrian description [3], and workwear detection and recognition [4].

Traditional target detection extracts features such as color, texture and edges of target objects in images by artificial design operator, and then locates and classifies the object [5]. However, the multiple morphologies revealed by garments bring great difficulties to feature extraction, and the existence of multiple garments in a single image with different sizes, occlusions, scaling and viewing angles makes

The rise of deep convolutional neural networks provides new ideas for the recognition of complex targets, among the typical deep learning target detection algorithms, one category is region recommendation based target detection, the representative algorithms are: R-CNN [7], Fast R-CNN [8], Faster R-CNN [9], SPP-NET [10], etc., and the other category is regression based target detection, through using the end-to-end idea, the image is normalized to a uniform size and input into the convolutional neural network, and the category and location information of the target object are predicted by regression, the representative algorithms are: YOLO [11] series, SSD [12] series, etc. In recent years, more and more scholars have carried out research on clothing recognition based on deep learning.

Zhang et al. [13] propose an optimized residual-based convolutional neural network clothing classification algorithm to improve the accuracy of multi-category clothing recognition by adjusting the order of batch normalization layer, activation function layer and convolutional layer in the network, and using a parallel pooling structure of "pooling layer + convolution layer", and replacing the fully connected

layer by the global mean pooling layer. However, this method has low accuracy for clothing recognition with complex background. Lu et al. [14] propose an novel deep residual network model to improve the recognition accuracy of clothing images by improving the arrangement of "BN + ReLU + convolution layers" in the traditional residual block, introducing the attention mechanism and adjusting the structure of the convolution kernel of the network, but the recognition speed is low because the network has more than 58 million parameters. Liu et al. [15] propose a cross-domain clothing retrieval method combined with attention mechanism. Based on deep convolutional neural network, the attention mechanism is introduced to reallocate the weight of different features to enhance the important features and suppress the unimportant features of clothing images, which can effectively deal with the interference of complex background and clothing deformation caused by viewpoint and pose, but the problem of clothing image recognition with occlusion still cannot be solved.

In this paper, according to YOLOv4-Tiny model [16], we propose a novel clothing attribute recognition algorithm. Firstly, the Res2Net module is used to optimize the backbone network, the receptive field size of each layer of the network is increased, and the fine-grained multi-scale clothing feature information is extracted. Then, the three feature layers of the feature extraction network output are up-sampled and fused, the high-level semantic features and shallow features are fused and passed into the FPN network through two feature channels. Finally, K-Means clustering algorithm is employed to optimize the anchor box parameters and obtain the anchor box that is more compatible with the clothing target. The performance of the proposed method is tested on the DeepFashion2 dataset [17], from comprehensive evaluations, it shows that the proposed method achieves high accuracy and outperforms the other compared algorithms.

The remainder of this paper is organized as follows: Section 2 analyzes the related work of YOLOv4-Tiny model. In Sect. 3, we describe the proposed method of the multi-scale feature extraction network, the fine-grained feature fusion and the anchor box parameter optimization. In Sect. 4, we perform the experiments to demonstrate the effectiveness of the proposed method. Finally, Sect. 5 draws conclusions.

2 Related Work

YOLOv4-Tiny network is a simplified version of YOLOv4, which belongs to the lightweight model, with only one tenth of the original number of parameters and faster detection speed. YOLOv4-tiny is characterized by multi-task, end-to-end, attention mechanism and multi-scale. Compared with the other versions of lightweight models, it has significant performance advantages [18].

YOLOv4-Tiny is applied to the recognition task of clothing images, the model is trained and tested on the DeepFashion2 dataset, and the recognition results are shown in Fig. 1. It can be seen that there are problems such as missed detection or incorrect recognition of clothing targets in the images due to the factors such as scale, occlusion and beyond the boundary.

The class activation mapping(CAM) [19] visualization of YOLOv4-Tiny is shown in Fig. 2. The bright part of the thermal diagram represents the high prediction attention intensity of the model. It can be seen that

the bright part of CAM of FPN network layer mainly focuses on the lower part of the dress, indicating that YOLOv4-Tiny algorithm can only extract a small amount of single clothing feature information.

3 Proposed Method

3.1 Optimization ideas

In order to improve the clothing attribute recognition accuracy, we improve YOLOv4-Tiny model as shown in Fig. 3, the optimization idea is as follows:

1.To improve the representation of multi-scale clothing feature information, the multi-scale feature extraction module Res2Net is used to optimize the backbone network of YOLOv4-Tiny to capture both local and global fine-grained clothing attribute features.
2.To retain more shallow features and narrow the semantic and resolution gap between shallow features and deep feature maps, the feature fusion network structure is optimized to fuse more shallow fine-grained feature information into the prediction network.
3.To enhance the fit between clothing attribute features and the network, K-Means clustering algorithm is adopted to cluster and analyze the clothing dataset, and obtain the anchor box that is more suitable for clothing attribute.

3.2 Multi-scale feature extraction optimization

In YOLOv4-Tiny, deepening the depth of the feature extraction network can extract more semantic information, but too many convolution operations will lead to information loss in the feature extraction process and reduction of recognition accuracy. In order to improve the clothing attribute recognition accuracy, the multi-scale feature extraction module Res2Net [20]

is used to optimize the YOLOv4-Tiny backbone network. The structure of the Res2Net module is shown in Fig. 4. The original n channel 3×3 filter is replaced by a series of smaller filter banks with w channel(let n = s×w, where s denotes the scale dimension), different convolution groups are connected layer by layer in a way similar to residual connection, so that the receptive field is constantly changing, so as to extract more fine-grained multi-scale clothing features.

As can be seen in Fig. 4, the small filter groups are connected in a hierarchical residual-like style to increase the number of different scales that the output features can express. The input feature maps are divided into several groups, each group of filters first extracts features from a group of input feature maps, the output features of the previous group are then sent to the next group of filters along with another group of input feature maps, this process repeats several times until all input feature maps are processed. Finally, all feature maps are concatenated and sent to a set of 1×1 filters for information fusion. On any possible path in which the input feature map are transformed into the output feature map, the equivalent receptive field increases after passing through a 3×3 filter, and many equivalent feature scales are eventually obtained due to the combination effect.

This split-mixed connection structure allows the output of Res2Net modules to contain different numbers, sizes, and scales of receptive fields and their combinations. In the structure of Res2Net shown in Fig. 4b, behind the 1×1 convolutional layer, the feature map is divided into s subsets, denoted by x_i, where i∈{1, 2, ..., s}, the number of channels in the subset is 1/s of the original, and each feature map subset x_i has the same spatial size as the original feature map set. Except for x_1, each feature map subset x_i has its corresponding 3×3 convolutional layer, denoted by $K_i()$, and the output of $K_i()$ is defined as y_i. The feature map subset x_i and $K_{i-1}()$ are summed and sent to $K_i()$ together for processing, and y_i is denoted as

$$y_i = \begin{cases} x_i & i = 1; \\ K_i(x_i) & i = 2; \\ K_i(x_i + y_{i-1}) & 2 < i \leqslant \mathrm{s}. \end{cases}$$

1

According to Fig. 4b and Eq. (1), it can be seen that each 3×3 convolution kernel $K_i()$ can receive the feature information of all previous feature map subsets $\{x_j, j \leq i\}$. After each feature map subset x_j passes through a 3×3 convolutional kernel, the output result can have a larger receptive field than x. Because of the combination effect, the output of Res2Net module contains different numbers of different sizes as well as different scales of receptive fields and their different combinations.

The Res2Net module is integrated into the backbone network of YOLOv4-tiny, and the optimized

network is called YOLO-Res2Net, whose structure is shown in Fig. 5. Firstly, the input image of size 416×416 is convolved to generate a feature map of size 208×208, a feature map with size 104×104 is obtained through convolution-normalization-activation function operation. Then, Res2Net module group is used to extract feature from feature map to extract more abundant feature information, Res2block is formed by stacking different number of Res2Net modules. The numbers of module stacks used for the optimized

network in the red dashed box in Fig. 5 are 3, 4, and 6 respectively. Firstly, Res2block1 performs 3 stacks of Res2Net module to obtain a feature layer of size 52×52, then Res2block2 obtains a feature layer of size 26×26 by performing 4 stacks of Res2Net module, and finally Res2block3 obtains a feature layer of size 13×13 by performing 6 stacks of Res2Net module.

To verify the role of the scale dimension in YOLO-Res2Net for clothing attribute recognition, models at different scales are tested on the DeepFashion2 Dataset, the experimental results are shown in Table 1. It can be seen that the clothing attribute recognition accuracy of YOLO-Res2Net improves with the increase of scale dimension, in the case of s equals to 2, 3, 4, residual connection structures between network hierarchies can generate a series of rich equivalent scale sets, and rich feature information is conducive to better recognition accuracy. However, when the scale is 5 or 6, due to the limitation of the test image size, rich multi-scale feature information cannot be extracted and only limited performance can be

improved. Considering the complexity and recognition accuracy of the model, the scale dimension of YOLO-Res2Net is set as 4.

Table 1
Recognition accuracy at different scales

scale	1	2	3	4	5	6
mAP(%)	47.39	49.68	50.96	52.51	52.53	52.56

3.3 Fine-grained feature fusion

In order to make full use of the high-level semantic features and the shallow geometric detail features in the backbone network, and to narrow the semantic and resolution gap between the shallow features and the high-level feature maps, the feature fusion network is optimized. To verify the influence of the fusion of different feature layers on clothing attribute recognition, tests are carried out on the DeepFashion2 dataset, the experimental results are shown in Table 2. The results show that the clothing recognition accuracy is the highest when the three-layer fusion feature is adopted, while when the four-layer or five-layer fusion feature is adopted, more redundant information is fused in clothing attribute feature, and the recognition accuracy is reduced. Therefore, the method of fusing three feature layers is finally adopted, that is, feature layers of size 13×13, 26×26 and 52×52 in the YOLO-Res2Net backbone network are fused to obtain more fine-grained feature information without increasing the network prediction channels, and the feature fusion network structure is shown in Fig. 6.

Table 2
Recognition accuracy of different feature layer fusion

Number of fusion layers	2	3	4	5
mAP(%)	52.23	52.71	52.65	50.31

After the improvement, the fine-grained feature information in the 52×52 feature layer is fused into the prediction network while the prediction of the original model is carried out at 13×13 and 26×26 feature layers. Firstly, the 13×13 feature layer and the 26×26 feature layer are fused by convolution and up-sampling, the deep features after fusion have strong semantic information, but the resolution is very low and the perception of details is poor. Then, on the basis of no additional network prediction channels, convolution and up-sampling are carried out for the 26×26 feature layer, which is fused with the 52×52 feature layer to obtain richer shallow fine-grained feature information. Finally, the fusion features are enhanced by FPN, and the output features are predicted.

3.4 Anchor box parameter optimization

YOLOv4-Tiny divides the input image into several grids, if the center of the actual border of the clothing attribute is in one of the grids, then this grid is responsible for predicting the clothing attribute. Therefore, the rationality of anchor box setting is very important for model performance. YOLOv4-Tiny network

presets 6 groups of generic anchor box parameters: [(10, 14), (23, 27), (37, 8), (81, 82), (135, 169), (344, 319)]. DeepFashion2 dataset is used to identify the clothing attributes, preset anchor box parameters are used in training, the intersection ratio between anchor box and boundary box is less than the threshold value 0.5, leading to many missed checks, so the anchor box needs to be reselected.

K-means clustering algorithm [21] is used to conduct clustering analysis on DeepFashion2 dataset, and cross ratio is used as a metric function to cluster the width and height of clothing target box, the intersection ratio of anchor box and boundary box is increased, and the nearest target is classified. Through continuous learning and iteration, the values of each cluster center are gradually updated until the cluster center remains unchanged. Anchor box parameter optimization can make the model parameters close to the experimental dataset, and reduce the loss while the recognition accuracy is improved. Six groups of new anchor box parameters are generated according to cluster center and data box distribution: [(107, 84), (112, 164), (172, 264), (191, 130), (273, 221), (395, 353)].

4 Experimental Results

The experimental platform uses the Pytorch deep learning framework, and the operating system is Ubuntu 20.04.5, the CPU is Inter Core i5-9400F, and the GPU is Nvidia GTX 2070. 191,916 clothing images from DeepFashion2 are used for training, the training iterations are 100, the batch_size is 32 for the first 50 iterations, and the batch_size is 16 for the second 50 iterations.

4.1 Ablation experiments

The DeepFashion2 dataset is used to train and test YOLOv4-Tiny and the improvement methods, where YOLO-Res2Net represents the model after integrating the Res2Net module into the backbone network of YOLOv4-Tiny, YOLO-Res2Net + represents the model after feature fusion optimization of YOLO-Res2Net, YOLO-Res2Net + + represents the model after optimization of anchor box parameters of YOLO-Res2Net+. Grad-CAM is used to visualize the class activation mapping of the models mentioned above. In the visualization examples shown in Fig. 7, stronger CAM areas are covered with brighter colors. Due to stronger multi-scale ability, fine-grained feature fusion and anchor box parameter optimization, the optimized model based CAM result has more concentrated activation map on the clothing area compared with the original one, which indicates that the proposed optimization methods can improve the feature extraction ability of the network for clothing attributes.

In addition, the recognition accuracy of the model is evaluated by Mean Average Precision (mAP). The ablation experimental results of different models on DeepFashion2 dataset are shown in Table 3. The results show that YOLO-Res2Net + + achieves the highest mAP value, which is 6.75% higher than the original model by optimizing the backbone network structure, feature fusion, and anchor box parameters. At the same time, the validity of the three improved methods is verified respectively.

Table 3
Results of ablation experiments

Neural Network Model	Backbone network optimization	Feature fusion optimization	Achor box parameter optimization	mAP/%
YOLOv4-Tiny	×	×	×	47.39
YOLO-Res2Net	√	×	×	52.51
YOLO-Res2Net+	√	√	×	53.25
YOLO-Res2Net++	√	√	√	54.14

4.2 Qualitative Analysis

To verify the effectiveness of the proposed method for clothing attribute recognition, qualitative analysis is conducted in three aspects: different scales, different degrees of occlusion, and different degrees of out-of-bounds.

Figure 8 show the CAM visualization of different methods for different-scale clothing images, it can be seen that stronger CAM areas are covered with brighter colors. Compared with YOLOv4-Tiny, the proposed method based CAM results have more concentrated activation maps on different scale clothing images, due to stronger multi-scale ability, the proposed method has activation maps that tend to cover the whole clothing target on small-scale and large-scale clothing images, while activation maps of YOLOv4-Tiny only cover parts of the clothing target.

Figure 9 shows the recognition results of the proposed method and YOLOv4-Tiny for clothing images with three different scales, where the first row shows the recognition results of YOLOv4-Tiny, and the second row shows the recognition results of the proposed method. The results show that YOLOv4-Tiny misses short sleeve dress, short sleeve top and skirt in the small-scale clothing images, misses short sleeve top and misidentifies skirt as short sleeved top in the medium-scale clothing images, and misses long sleeved top and short sleeve top in the large-scale clothing images. In contrast, the proposed method can detect and recognize the clothing images with different scales accurately.

Figure 10 shows the CAM visualization of different methods for clothing images with different degrees of occlusion. Compared with YOLOv4-Tiny, the proposed method based CAM results have more concentrated activation maps on the clothing images with different degrees of occlusion. The results show that the proposed method can reduce the influence of occlusion on clothing attribute information extraction. clothing image with different degrees of occlusion. **a** Slight occlusion. **b** Medium occlusion. **c** Heavily occlusion

Figure 11 shows the recognition results of the proposed method and YOLOv4-Tiny for three kinds of clothing images with different occlusion degrees, where the first row shows the recognition results of YOLOv4-Tiny and the second row shows the recognition results of the proposed method. As can be seen, in Fig. 11a, the arms and the leather bag cause slight occlusion to the clothing respectively, causing the false detection of YOLOV4-Tiny. In Fig. 11b, YOLOV4-Tiny mistakenly detected the vest dress as a short sleeve top due to the shielding of the hat, while the shielding of the pants by the leather bag leads to the missing detection. In Fig. 11c, shorts are seriously blocked by the short sleeve top, resulting in missed detection, short sleeve dress is mistakenly detected as skirt because it is seriously blocked by the curtain. In contrast, the proposed method can obtain accurate recognition results for clothing objects with different occlusion degrees.

Figure 12 shows the CAM visualization of different methods for clothing images with different out-of-bounds degrees. It can be seen that the proposed method based CAM results have more concentrated activation maps on the clothing images with different degrees of out-of-bounds.

Figure 13 shows the recognition effects of the proposed method and YOLOv4-Tiny on three clothing targets with different out-of-bounds degrees. The results show that YOLOv4-Tiny mistakenly detects the short sleeve dress in Fig. 13a as a skirt, misses the short sleeve top and skirt, misses the long sleeve top and long sleeve dress in Fig. 13b that are partially out of bounds, and misses the pants and long sleeve dress in Fig. 13c that are out of bounds. In contrast, the proposed method can identify the garments with different out-of-bounds degrees accurately.

4.3 Quantitative analysis

To further evaluate the recognition accuracy of the proposed method, it is compared with several lightweight target detection algorithms such as FBNet [22], GhostNet [23], ShuffleNet [24], and MobileNet [25]. These methods are trained and tested on the DeepFashion2 dataset, and the experimental results are shown in Table 4. The results show that the mAP of the proposed method is improved significantly compared with the other comparison models, and it has the highest recognition accuracy in the clothing attribute recognition task.

Table 4
Comparison of experimental results for different models

Neural Network Model	mAP/%
FBNet	24.18
GhostNet	31.42
ShuffeNet	33.18
MobileNet	39.33
YOLOV4-Tiny	47.39
Proposed method	54.14

5 Conclusion

In this paper, a novel clothing attribute recognition method based on improved YOLOv4-Tiny is proposed. Experimental results show the effectiveness of our method, which can significantly improve the recognition accuracy of clothing images with different scales, different occlusion degrees and different out-of-bounds degrees. The proposed method not only provides the algorithm basis for clothing retrieval, clothing matching and other applications, but also contributes a little to the intelligent clothing industry.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials Upon request.

Competing interests The authors declare that they have no competing interests.

Funding Not applicable.

Authors' contributions Meihua Gu and Jie Liu designed the research, performed the research, Wei Hua analyzed the data, all authors contributed to the writing and revisions.

Acknowledgements The authors would like to thank Jing Feng for fruitful discussions on this work.

References

1. Gupta, M., Bhatnagar, C., Jalal, A, S.: Clothing image retrieval based on multiple features for smarter shopping. Procedia Computer Sci. 125, 143–148 (2018)

2. Zhou, W., Mok, P, Y., Zhou, Y., et al.: Fashion recommendations through cross-media information retrieval. J. Vis. Commun. Image Represent. 61, 112–120 (2019)
3. Qiang, C., Huang, J. Feris, R., et al.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In: IEEE Conference on Computer Vision & Pattern Recognition. pp: 5315–5324 (2015)
4. Ke, X., Liu, T., Li, Z.: Human attribute recognition method based on pose estimation and multiple-feature fusion. SIViP 14, 1441–1449 (2020).
5. Ahmed, K.T., Irtaza, A. & Iqbal, M.A. Fusion of local and global features for effective image extraction. Appl Intell. 47, 526–543 (2017).
6. Aslan, M.F., Durdu, A. & Sabanci, K. Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization. Neural Comput & Applic. 32, 8585–8597 (2020).
7. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp: 580–587 (2014)
8. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp: 1440–1448 (2015)
9. Ren, S., He, K., Girshick, R., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28: 91–99 (2015)
10. He, K., Zhang, X., Ren, S., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9): 1904–1916 (2015)
11. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp: 779–788 (2016)
12. Liu, W., Anguelov, D., Erhan, D., et al.: Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, Cham, pp: 21–37 (2016)
13. Zhang, Z, H., Zhou, C, L., Liang. Y.: An optimized clothing classification algorithm based on residual convolutional neural network. Comput. Eng. Sci. **40**(02): 354–360 (2018)
14. Lu, J, B., Xie, X, H., Li, W, T.: An Improved Clothing Image Recognition Model Based on Residual Network. In: Computer Engineering and Applications, **56**(20): 206–211 (2020)
15. Liu, Y, J., Wang, W, Y., Li, Z, M., et al.: Cross-Domain Clothing Retrieval with Attention Model. In: Journal of Computer-Aided Design & Computer Graphics, **32**(06): 894–902 (2020)
16. Bochkovskiy, A., Wang, C, Y., Liao, H, Y, M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 (2020)
17. Ge, Y., Zhang, R., Wang, X., et al.: Deepfashion2:A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp: 5337–5345 (2019)

18. Jiang, Z., Zhao, L., Li, S., et al.: Real-time object detection method based on improved YOLOv4-tiny. arXiv:2011.04244 (2020)
19. Selvaraju, R, R., Cogswell, M., Das, A., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localiza-tion. In: Proceedings of the IEEE international conference on computer vision. pp: 618–626 (2017)
20. Gao., Shuang, H., et al.: Res2Net: A New Multi-Scale Backbone Architecture. IEEE Trans. Pattern Anal. Mach. Intell. **4**(2): 652–662 (2021)
21. Sinaga, K, P., Yang, M, S.: Unsupervised K-means clustering algorithm. In: IEEE Access, pp: 80716–80727 (2020)
22. Wu, B., Dai, X., Zhang, P., et al.: Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp: 10734–10742 (2019)
23. Han, K., Wang, Y., Tian, Q., et al.: Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp: 1580–1589 (2020)
24. Zhang, X., Zhou, X., Lin, M., et al.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp: 6848–6856 (2018)
25. Howard, A, G., Zhu, M., Chen, B., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)

Figures

Figure 1

Clothing recognition results of YOLOv4-Tiny

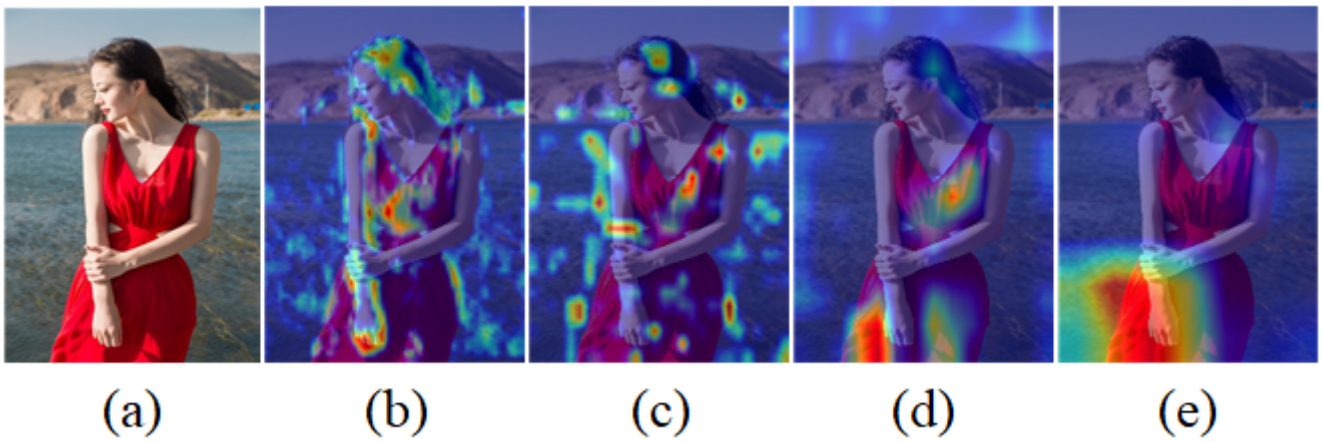


Figure 2

CAM visualization of YOLOv4-Tiny. **a** The original image. **b** Resblock_body1. **c** Resblock_body2. **d** Resblock_body3. **e** FPN network layer

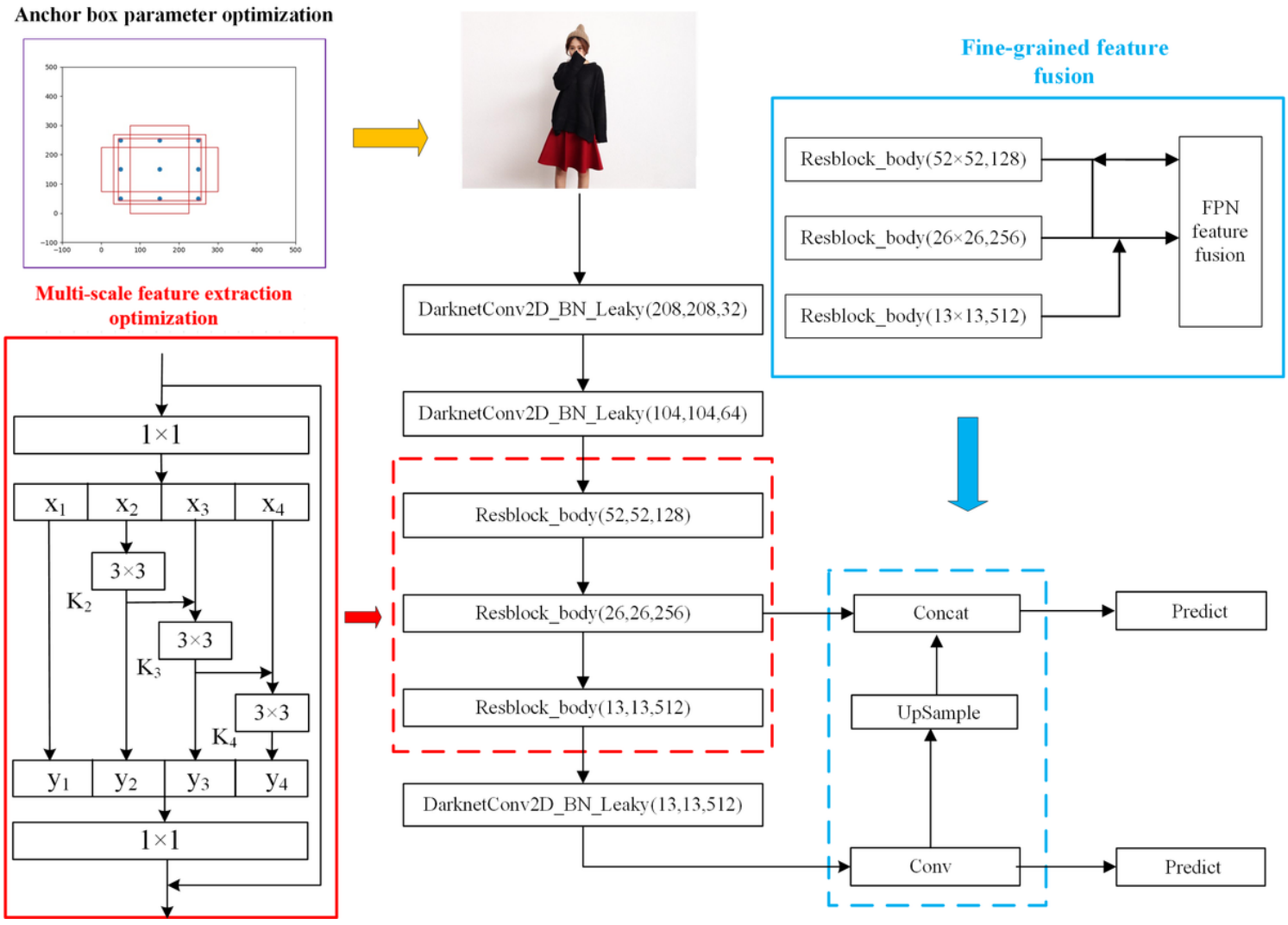
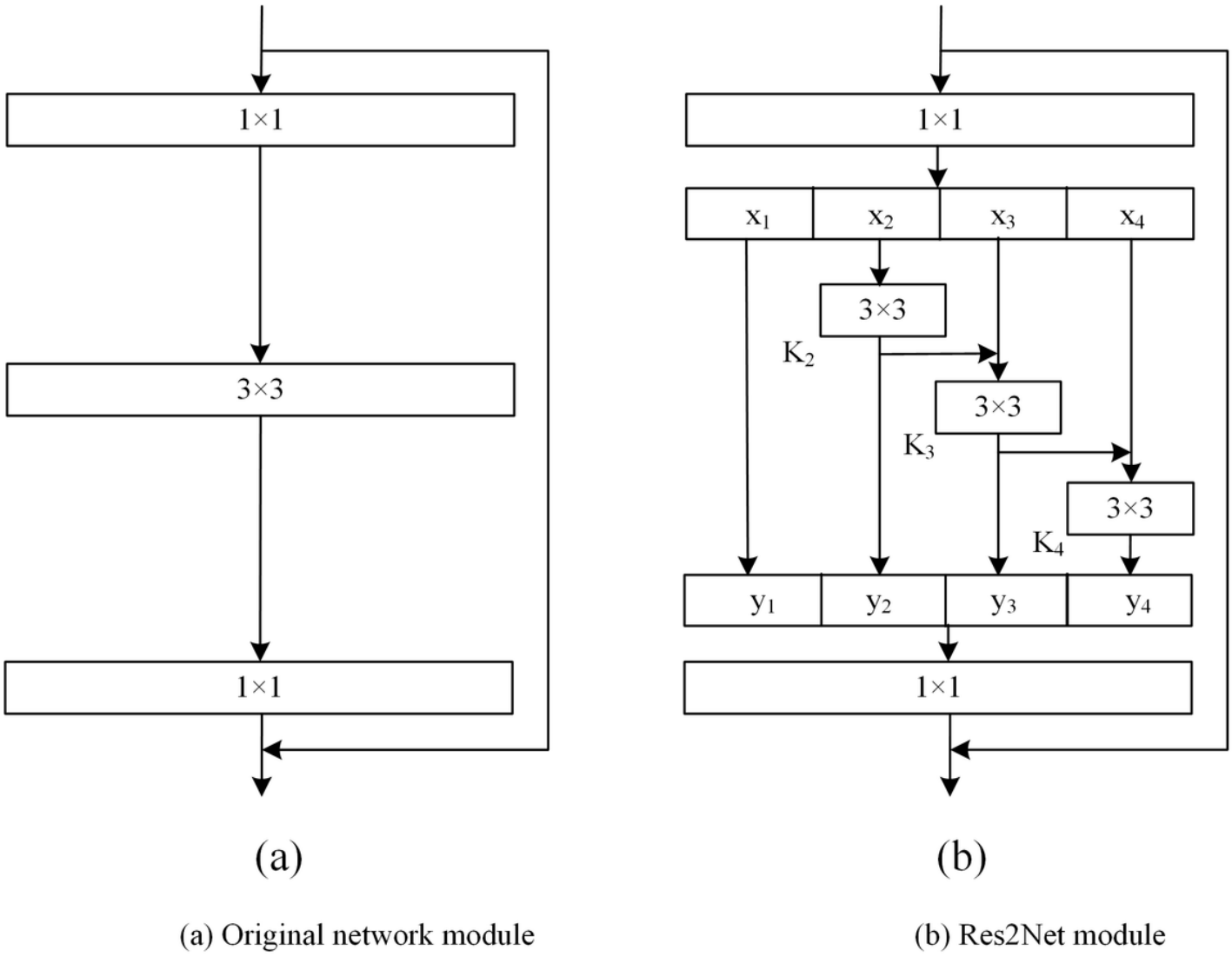


Figure 3

Improved YOLOv4-Tiny model



(a) Original network module

(b) Res2Net module

Figure 4

Res2Net module. **a** Original network module. **b** Res2Net module

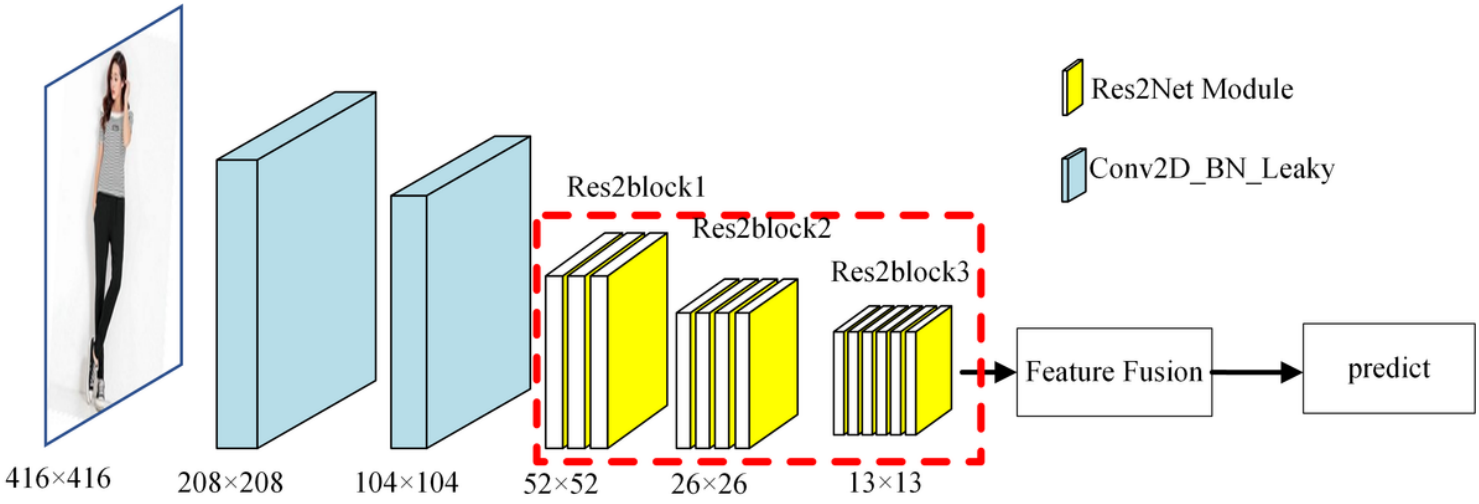


Figure 5

YOLO-Res2Net structure

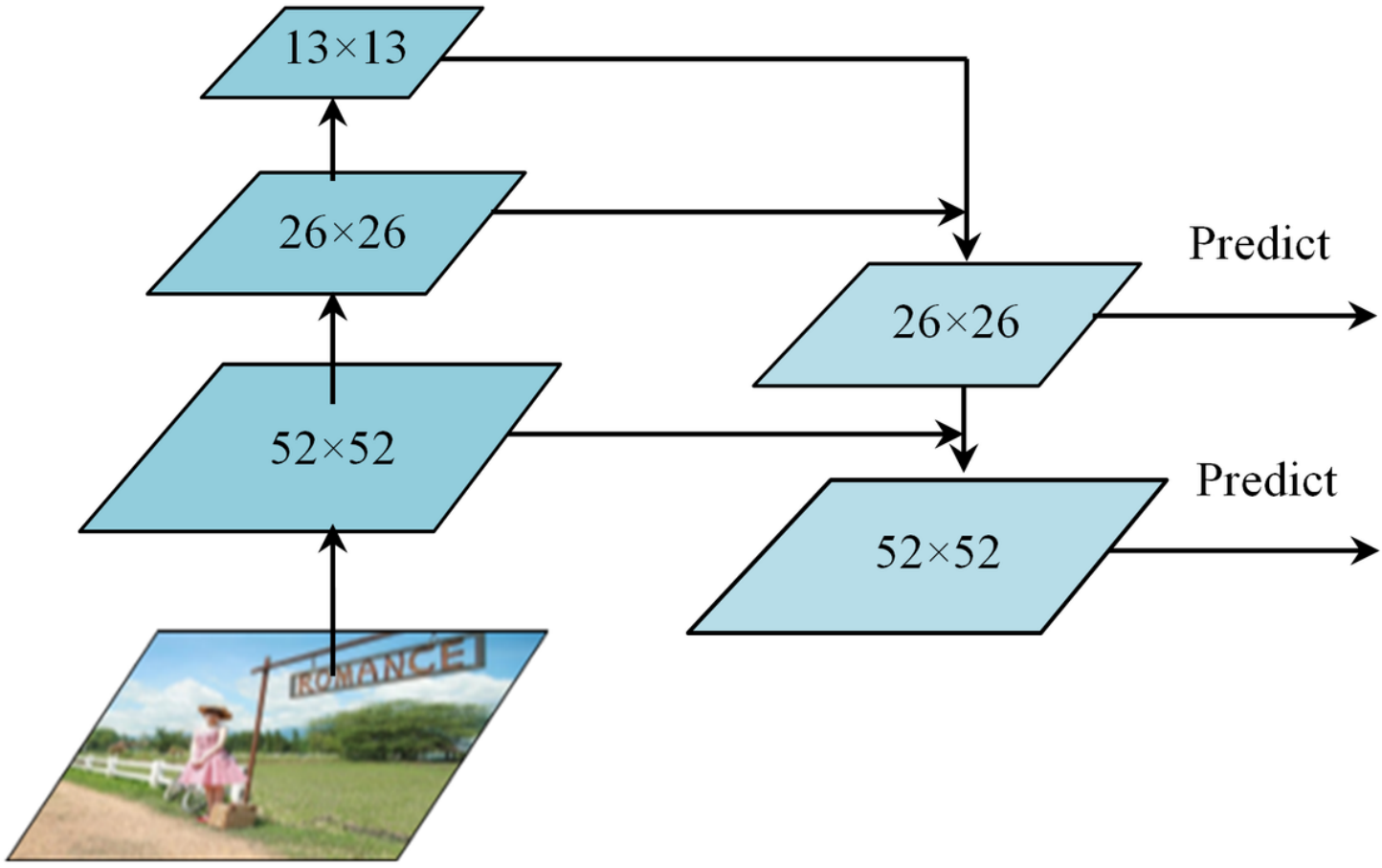


Figure 6

Optimized feature fusion network architecture

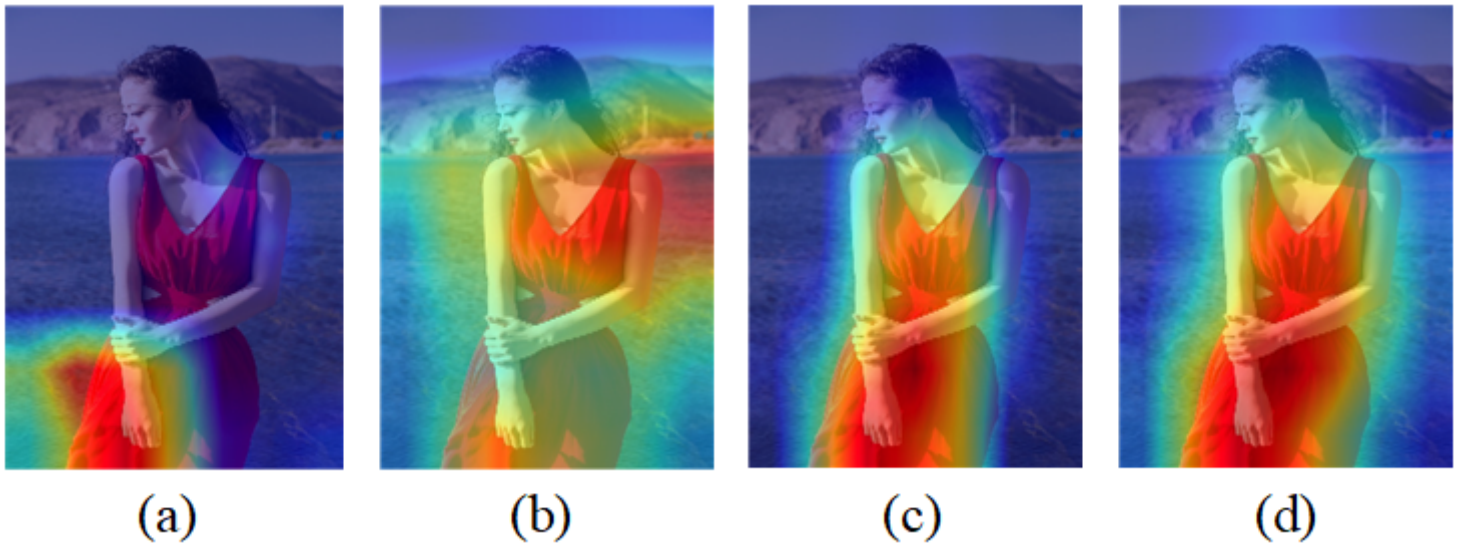


Figure 7

CAM visualization of different models. **a** YOLOv4-Tiny. **b** YOLO-Res2Net. **c** YOLO-Res2Net+. **d** YOLO-Res2Net++

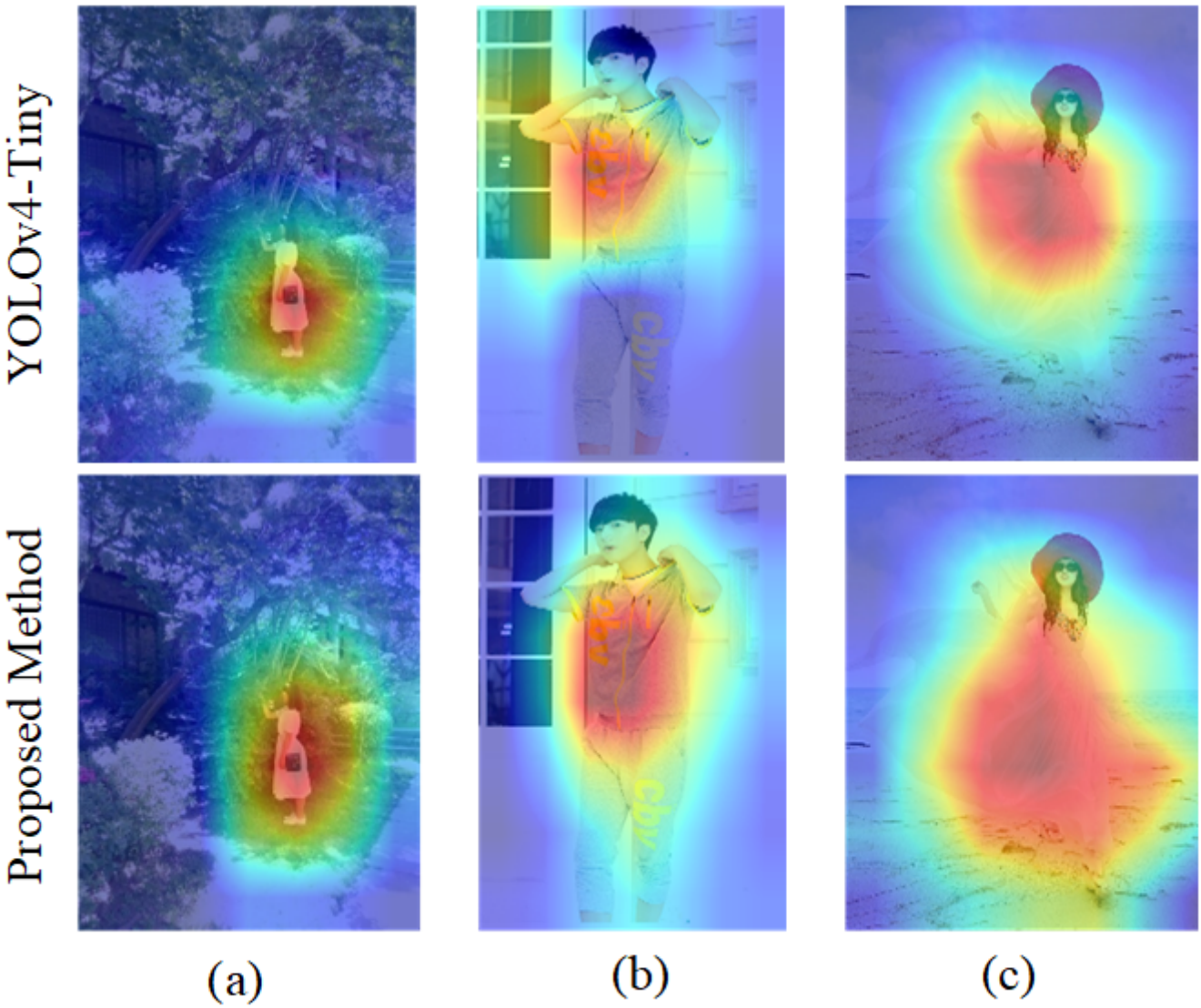


Figure 8

CAM visualization of different methods for different scale clothing images. **a** Small-scale. **b** Medium-scale. **c** Large-scale

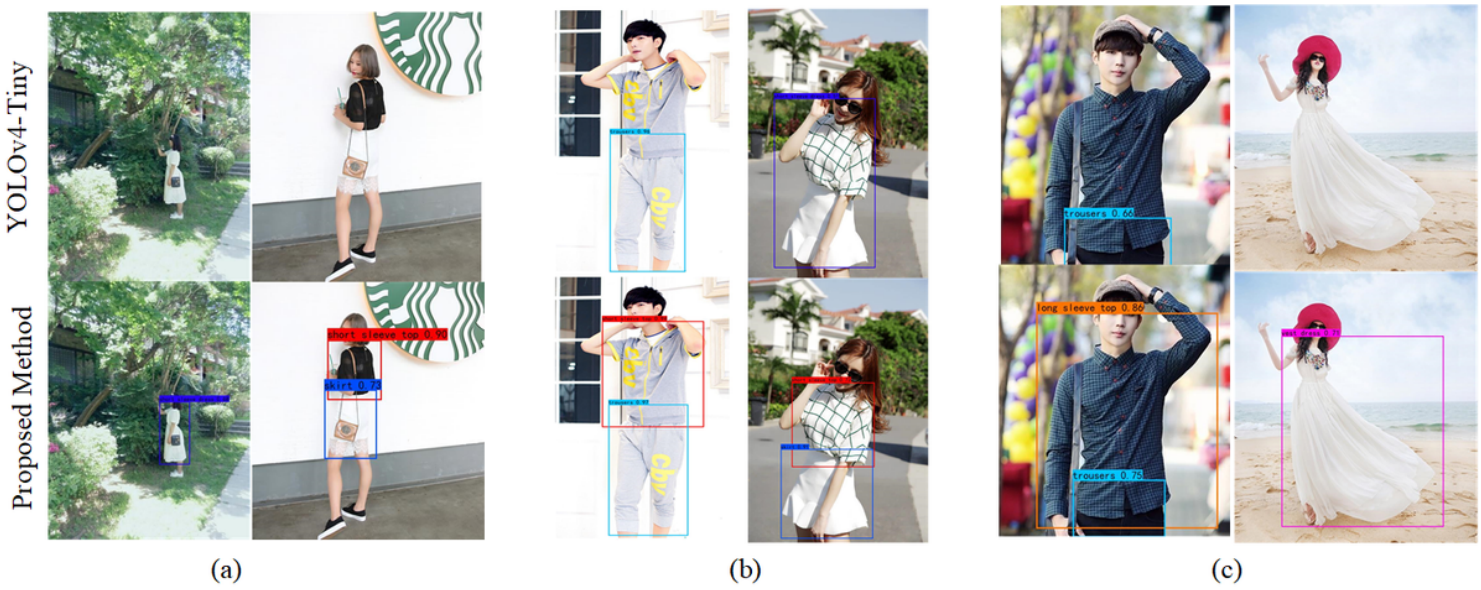


Figure 9

Clothing image recognition results with different scales. **a** Small-scale. **b** Medium-scale. **c** Large-scale

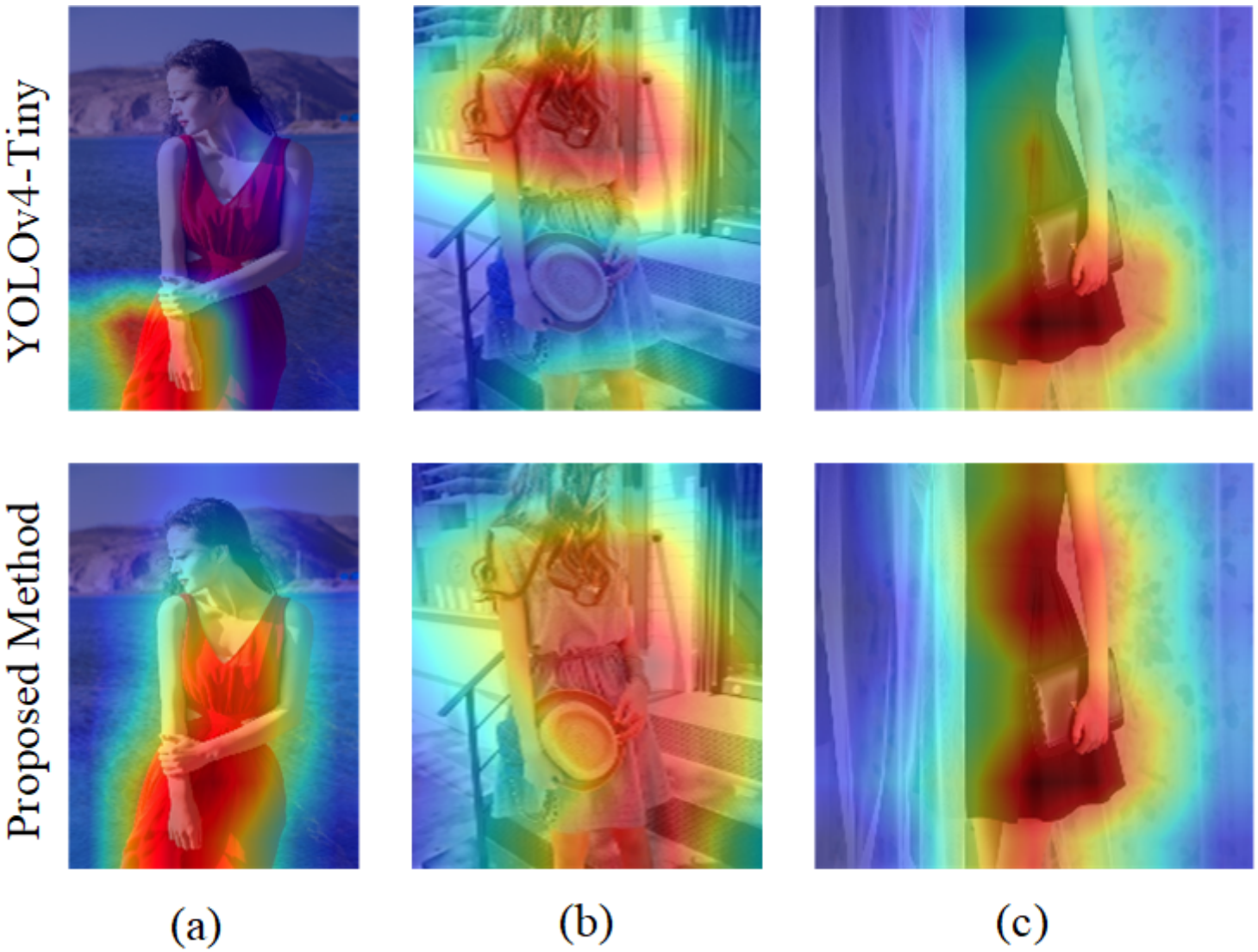


Figure 10

CAM visualization of different methods for clothing image with different degrees of occlusion. **a** Slight occlusion. **b** Medium occlusion. **c** Heavily occlusion

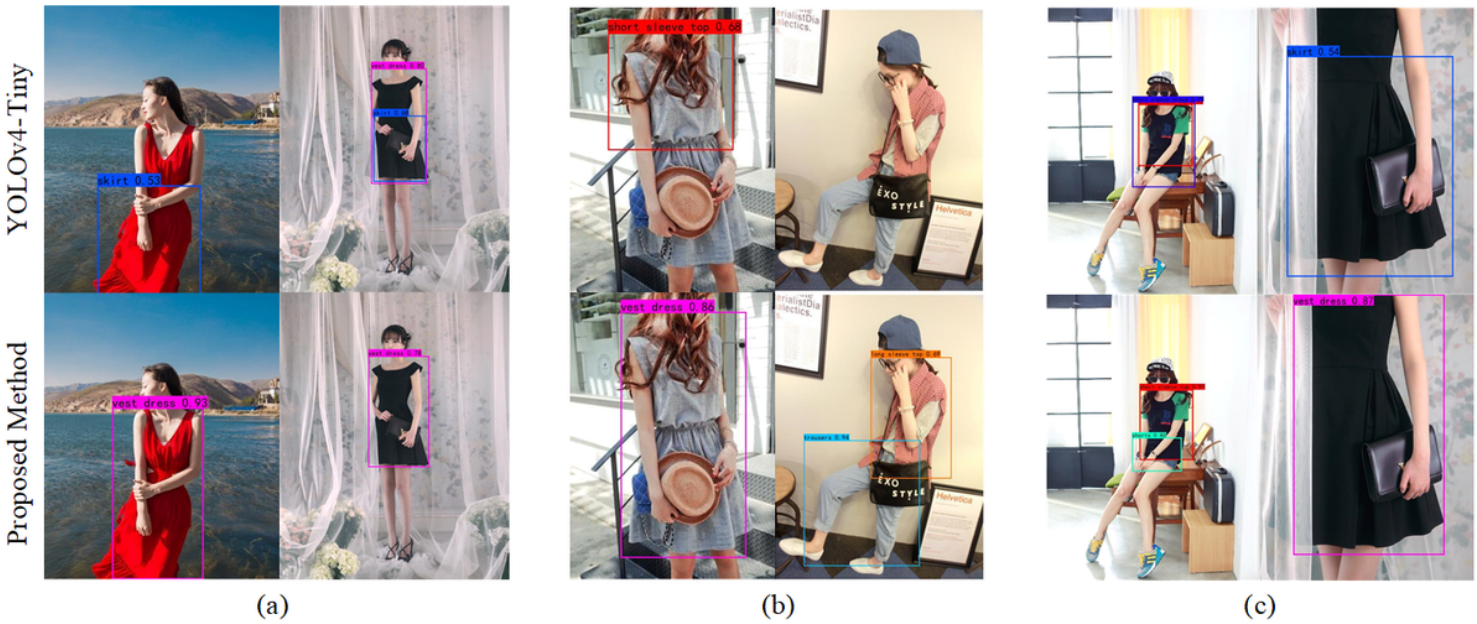


Figure 11

Clothing image recognition results with different degree of occlusion. **a** Slight occlusion. **b** Medium occlusion. **c** Heavily occlusion

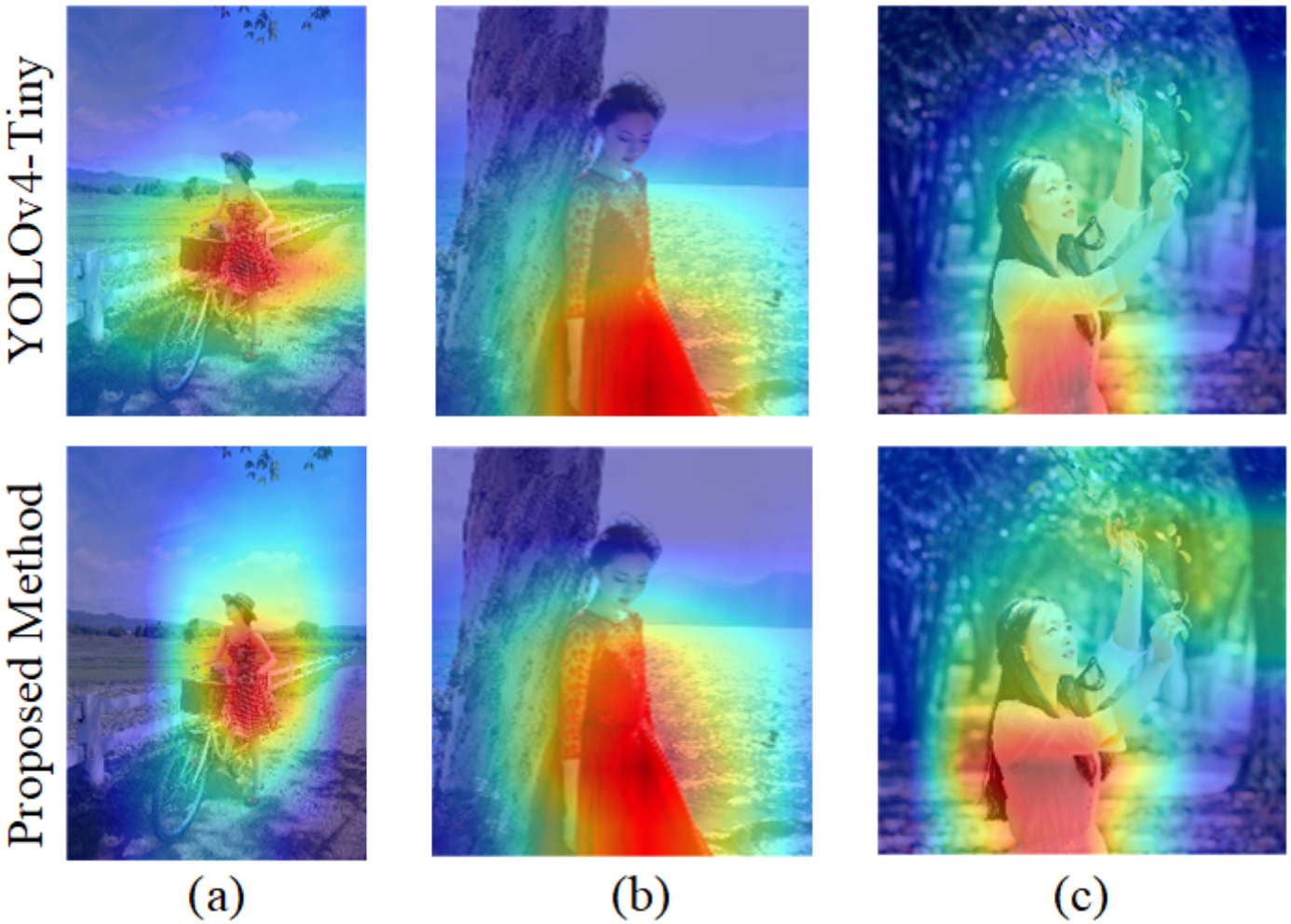


Figure 12

CAM visualization of different methods for clothing image with different degrees of out-of-bounds. **a** Without out-of-bounds. **b** Medium out-of-bounds. **c** Large out-of-bounds

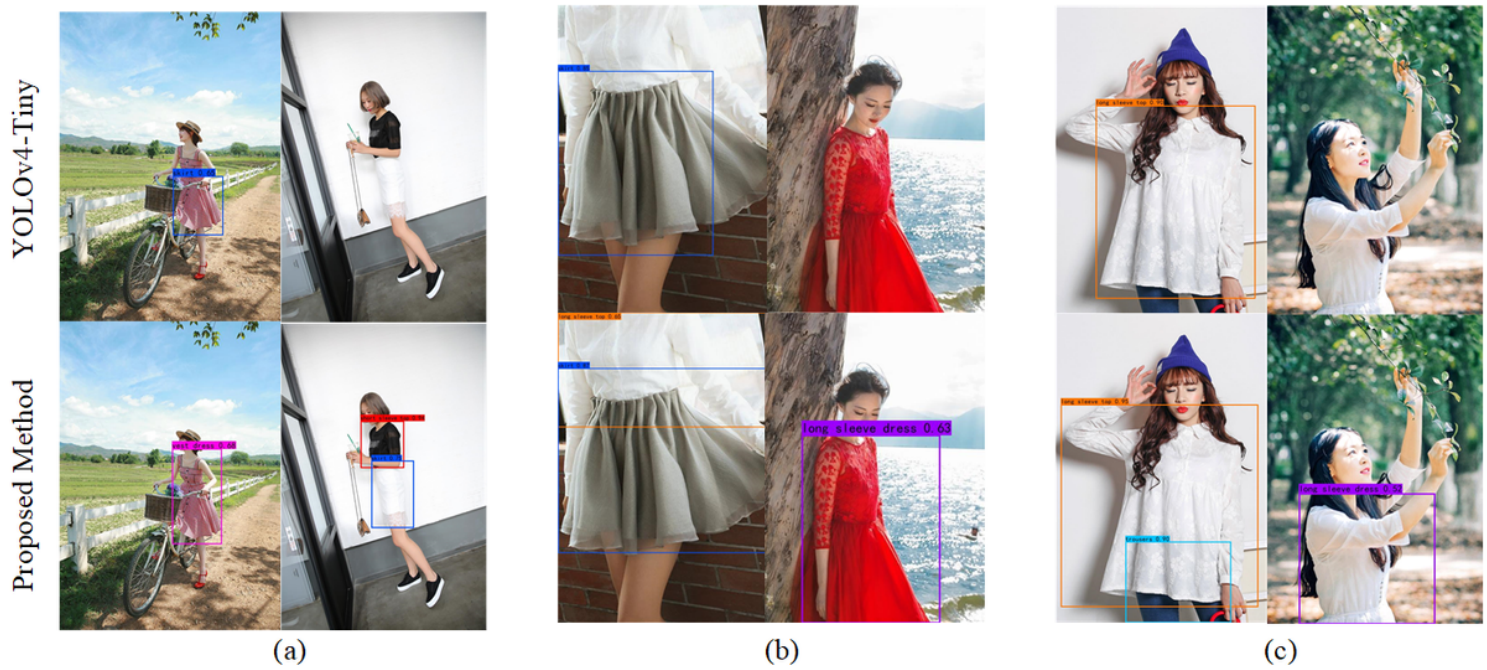


Figure 13

Clothing image recognition results with different degree of out-of-bounds. **a** Without out-of-bounds. **b** Medium out-of-bounds. **c** Large out-of-bounds