Liu *et al. Heritage Science* (2023) 11:40
https://doi.org/10.1186/s40494-023-00883-x

Heritage Science

RESEARCH

Open Access

Research on Chinese traditional opera costume recognition based on improved YOLOv5

Kaixuan Liu*, Kai Lin and Chun Zhu

Abstract

In order to protect the cultural heritage of opera costumes, establish visual labels for opera costumes, accelerate the establishment of a database for opera costumes, and increase the dissemination of opera culture, we propose an improved You Only Look Once (YOLO) v5-based opera costume recognition model for opera costumes with a wide range of styles, rich colors, and complex stage environments. By adding Coordinate Attention (CA) mechanism to the backbone of YOLOv5, the network can focus on more interesting information when extracting features; replacing the original feature pyramid module with a weighted bidirectional feature pyramid module in the Neck part to achieve efficient fusion of features; replacing the original loss function GIOU with DIOU to improve the detection accuracy and convergence speed. The average detection accuracy of the improved YOLOv5 model reaches 86.3% and its inference speed reaches 28 ms per frame through experiments on the homemade Chinese costume dataset, which improves the average detection accuracy by 3.1% compared with the original model, and has good robustness in detecting complex scenes such as covered targets, light-colored costumes, cross targets, dense targets and different angles. The model meets the requirements for accuracy and real-time costume recognition in complex theatrical environments.

Keywords Chinese opera, Opera costume, Costume image recognition, YOLOv5, Machine learning

Introduction

Chinese opera costumes, which can be traced back as far as the Ming dynasty in China, referring to the costumes worn in traditional Chinese theatrical performances, are designed to help the opera portray the external image of the characters. Chinese opera costumes have a variety of styles, ornate patterns, rich colors, and exquisite craft and have an important place in stage art with a solid aesthetic flavor. However, as a general audience, it isn't easy to distinguish among the categories to which the opera costumes belong.

*Correspondence:
Kaixuan Liu
L40260611@hotmail.com
School of Fashion and Art Design, Xi'an Polytechnic University, Xi'an 710048, China

There are five categories of Chinese opera costumes, namely MangPao (蟒袍), Pei (帔), Kao (靠), ZheZi (褶子), and Yi (衣) [1]. Among them, high-status figures such as emperors and generals wore MangPao, and it can also be divided into the GuanMang and the LongMang, etc. Officials and their dependents wore Pei at all levels on domestic occasions. Military generals wore Kao. ZheZi was a long casual shirt with a slanting collar. Yi was a generic term for all the other costumes except the previous four, and it can also be divided into the JianYi and the ShengYi, etc.

With the development of Internet technology, there are many ways to preserve cultural heritage [2, 3]. Promoting the overall digitization of opera culture is the inevitable choice to realize the inheritance and development of opera [4]. Faced with numerous categories of opera costumes, an accurate and effective classification model can reduce labor costs and improve image classification

and detection [5–7]. At present, most scholars have studied opera costumes from costume restoration [8], color analysis [9, 10], secondary design [11–13] and aesthetic perspective [14], etc. In order to help people better recognize the opera costumes, the most direct way is to use three-dimensional (3D) virtual visualization technology to restore the costumes. Clothing virtual restoration should focus on color, pattern, silhouette and other aspects [15]. In order to extract the main color features of opera costumes, the color clustering algorithm is also used in the analysis and recognition [16]. In the field of traditional computer vision, Jia et al. used the spatial moment algorithm to simply compare the silhouettes of opera costumes [17]. But it can only be used to compare the contours with exactly the same position, size and angle. The above research focuses on the analysis of the appearance characteristics of opera costumes, which is unable to realize real automatic recognition.

However, we have adopted an opera costume classification method based on deep learning, which can not only achieve real-time detection, but also achieve high classification accuracy. We innovatively modified the model structure of the original algorithm, which makes it possible to improve the accuracy of detection as well as solves the recognition problem of complex drama scenes. The aims of research on more effective classification are as follows: (i) Help opera costumes establish corresponding visual labels and accelerate the construction of a database for opera costumes; (ii) Provide a more effective way to retrieve the corresponding opera costumes and help people understand the costumes more easily; (iii) To play a positive role in the inheritance, preservation, and promotion of opera costumes.

In recent years, the development of computer vision has provided the possibility of digitizing culture [18–20]. Computer image recognition algorithms can be divided into traditional computer vision algorithms and computer vision algorithms based on deep learning [21]. For traditional computer vision algorithms, the key is the feature extraction algorithm. For example, canny edge detection [22], Histograms of Oriented Gradients (HOG) [23], corner point detection [24], and Local Binary Pattern (LBP) [25], etc. They must distinguish between the different characteristics of different classes of detection objects. Some scholars have conducted the following studies on garment image classification using traditional computer vision algorithms: Aulia et al. improved the performance of garment image recognition by classifying apparel images based on the Region Of Interest (ROI) segmentation mechanism of directional gradient histogram [26]. Lorenzo-Navarro et al. conducted an experimental study on LBP and HOG features of garment images, using two classifiers, Support Vector Machine (SVM) and Random Forest, for classification [27]. Bossard et al. proposed a random forest-based multiclass learner to classify 15 garment categories [28]. Yu et al. proposed a nested edge detection algorithm combined with canny to achieve garment pattern contour extraction and accurate tailoring [29].

All the above traditional algorithms have made excellent progress in garment image recognition problems, but they rely on high-quality input images and artificial feature screening. However, for opera costumes, there are many types of costumes with rich colors and ornate patterns, which in turn brings problems to computer vision in that the number of features is too many, the computation becomes complicated, and the accuracy decreases. It usually takes a lot of effort to do the classification judgment on the costume features artificially. The use of deep learning-based methods can better extract the feature map automatically and dig deeper feature information.

There are many kinds of computer vision classification models based on deep learning, such as one-stage models with faster detection speed [30, 31] and two-stage models with high recognition accuracy [32, 33]. These deep learning algorithms are helpful for objects with many features like costumes for their classification and retrieve. Many scholars have used deep learning in their image classification studies: Ren et al. improved the backbone section in the network based on You Only Look Once (YOLO) v5 by combining ResNet to achieve high-accuracy garment classification [34]. Yin et al. introduced an attention mechanism in YOLOv3 to improve the accuracy of the algorithm, which can effectively accomplish the task of real-time pig detection [35]. Iannizzotto et al. used a computer vision system based on deep neural networks (DNNs) to detect whether workers are wearing Personal Protection Equipment (PPE) [36]. Xiang et al. used the Inception-ResNet V1 model and L-Softmax to represent images and identify attributes of garment based on the Region-Convolutional Neural Networks (R-CNN) framework [37].

There are fewer studies on image recognition of traditional opera costumes, and the reasons for this are the following three points: (1) There are fewer image data about opera costumes, which makes the costume detection and classification task challenging; (2) The traditional opera costumes have many styles, complex detail designs, rich colors, and patterns, which are hard to be recognized; (3) The actors performing on stage are dynamic. The classification of the costumes is often affected by movement blocking as well as by changes in lighting, etc. To address the above problems, YOLOv5 is chosen as the original network and improved, and a Chinese opera costume dataset is constructed. The

advantages of YOLOv5 are the following three points: (i) The Input part comes with Mosaic data enhancement, which can improve the robustness of the model in the case of limited data; (ii) The network structure has a high degree of freedom and can be designed and improved according to its dataset; (iii) The network is a lightweight structure with fast inference. It is capable of fulfilling the function of real-time detection.

In this paper, we introduce Coordinate Attention (CA) mechanism module to the original YOLOv5 network structure to improve the problem of missed detection of occluded targets and small targets, to adapt to different complex scenes and different lighting conditions for the recognition and classification tasks of opera costumes. Meanwhile, the original feature pyramid module is replaced by a weighted bidirectional feature pyramid module in the Neck part to achieve efficient feature fusion; What's more, by replacing the original loss function GIOU with DIOU, to improve the detection and recognition effect of the network.

Costume recognition model based on improved YOLOv5 network

YOLO is a single-stage target detection and classification algorithm based on deep learning, which has strong generalization ability, high detection accuracy and fast inference. There are four versions of YOLOv5, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The weights of these four versions, as well as the width and depth of the model, increase sequentially. The larger their models, the greater the accuracy will increase, but at the same time, it will bring the problem of a slow running rate. To meet the demand for real-time costume detection, we chose the YOLOv5s model, which can be divided into four parts: Input, Backbone, Neck, and Prediction. Figure 1 shows the general flow.

Input section

The operations performed in the Input section are Mosaic data augmentation, adaptive anchor box calculation, and adaptive image scaling of the images.

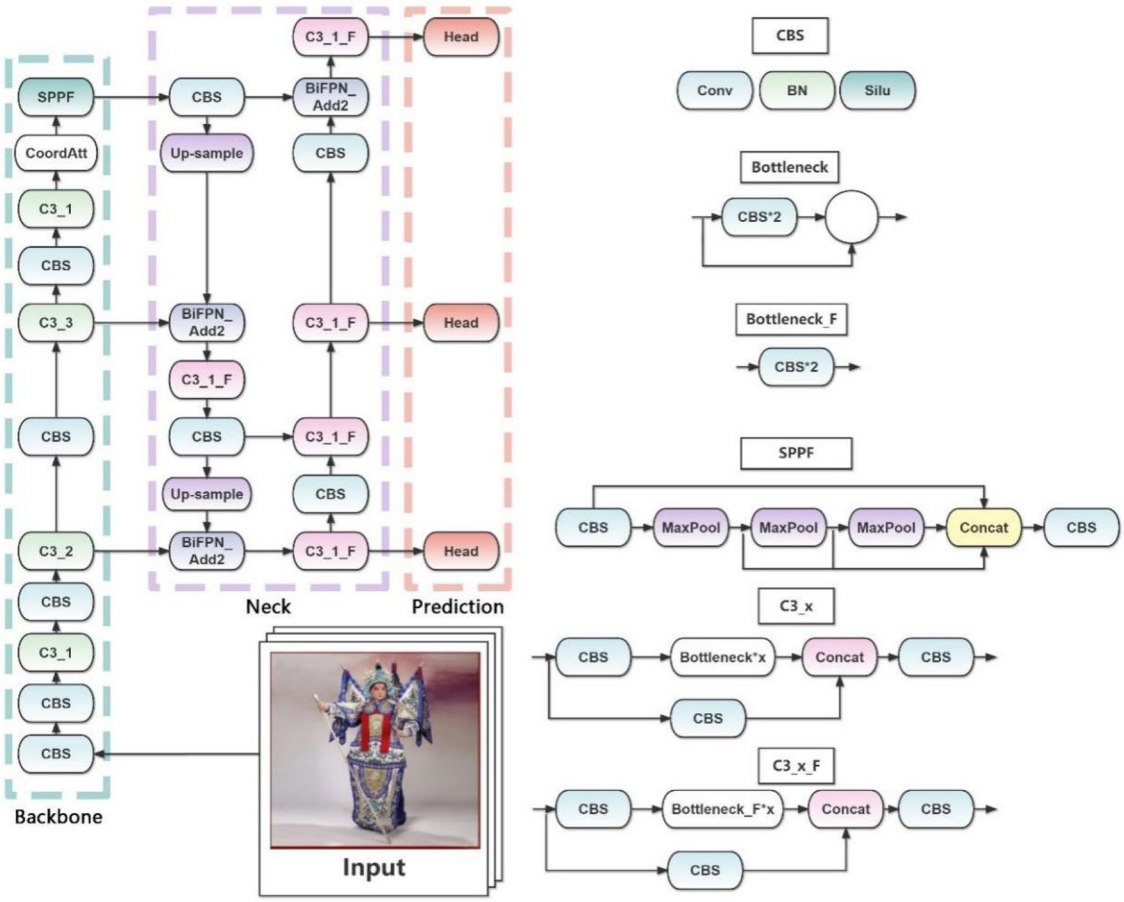

Fig. 1 Improved YOLOv5 network structure

For a large neural network, if the number of samples is smaller than the number of features, overfitting is likely to occur [38], so Mosaic data augmentation is introduced (as shown in Fig. 2). It stitches four images in a randomly scaled, which are labeled previously in the process of making the opera costume dataset, randomly cropped and randomly lined up manner, which can increase the number of small targets in the dataset and thus improve the model's ability to detect small targets. During the model's training, YOLOv5s automatically calculates the best anchor box values in training sets for different categories (as shown in Fig. 3). The network compares the predicted boxes with the real boxes, calculates the difference between them, and then updates the iterative network parameters in reverse. The adaptive scaling of the image is done by scaling the original image to a uniform standard size ($3 \times 640 \times 640$) and then feeding it into the network for detection.

Improved backbone section

The improved Backbone section consists of Convolution, Batch normalization, Silu activation function (CBS), C3_1, CoordAtt and SPPF (spatial pyramid pooling fast). It serves to extract the target features. In the CBS module, the image data is output after convolution, batch normalization and the Silu activation function. C3_x consists of CBS as well as Bottleneck, which uses Concat to combine the features. SPPF connects the CBS and Maxpool to the Concat. Unlike the original network, the attention mechanism CoordAtt module, or CA module, is added to the Backbone in this study.

The variety of opera costumes is challenging to distinguish, even giving the task to the general audience. In the last two years, many researchers have proposed the concept of attentional mechanism [39]. The addition of the attention mechanism helps the network to ignore useless information, focus on crucial information, and improve recognition efficiency and quality [40]. Several

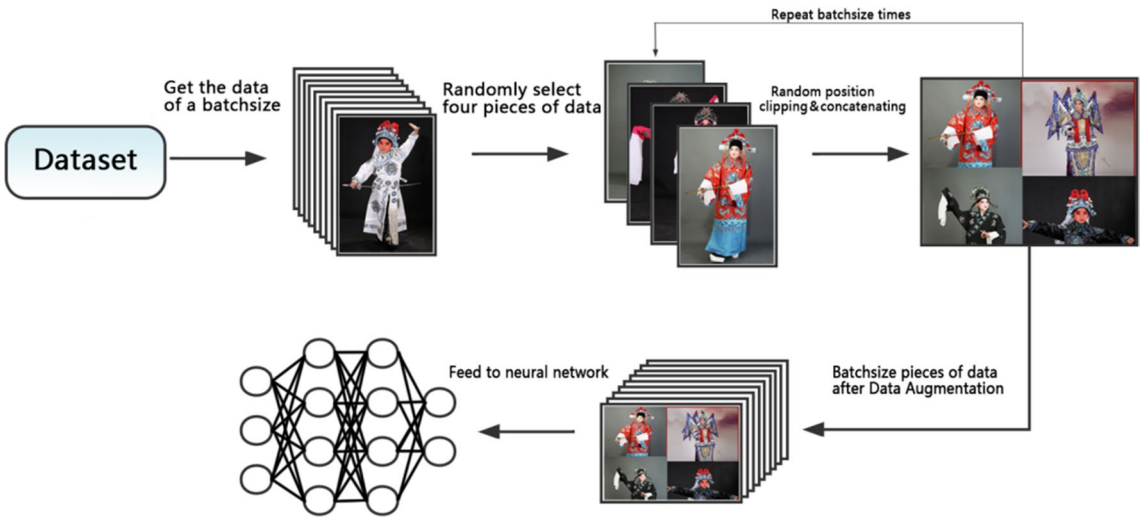

Fig. 2 Mosaic data augmentation

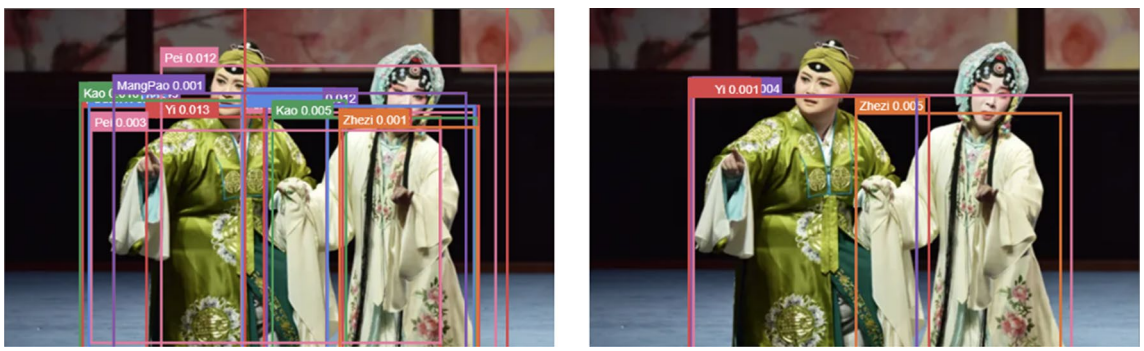

Fig. 3 Adaptive anchor box

attention mechanisms currently exist. The SE module significantly improves recognition performance, but they usually ignore the location information, which is very important for generating attention maps [41]. The CA attention mechanism embeds location information, giving the neural network more interesting data while extracting image features.

To capture precise location information, the CA module performs global pooling in the process of information embedding with the following equation:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{1}$$

where, x represents the input image, c represents the channel of image, and Z_c represents the aggregated feature map, which is essentially a two-dimensional array. When using the pooling kernel (H, 1) and (1, W) to encode the features of the input Xc in the i direction and the j direction, a pair of direction-aware feature maps composed of two-dimensional arrays will be obtained. Z_c gives good access to the global sensory field and helps the network locate the target of interest more accurately. Subsequently, concatenate z^h and z^w with the following equation:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{2}$$

where, z^h and z^w represents the aggregated feature map in the horizontal and vertical direction. Where [,] represents stitching together the feature maps in width and height directions. F_1 is the convolutional transform function that reduces the feature dimension to the original c/r, where r is used to control the reduction rate. Where δ is the nonlinear activation function, f represents the spatial information's feature mapping when encoding the vertical and horizontal directions. Then, f is divided into two separate tensors f^w and f^h, along the horizontal and vertical directions. f^w and f^h are transformed to have the same number of channels as the input X using two 1×1 convolutional variations F_h and F_w. The formula is as follows.

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{3}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{4}$$

where, σ is the sigmoid activation function, g^h and g^w represent the attention weights of the feature map in the height and width directions, respectively. The results of the CA module are expressed as

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{5}$$

where, y_c is the output of the CA module, c represents the channel of image, and x_c represents the input image. Where $g^h{}_c$ and $g^w{}_c$ represent the attention weights of the feature map in the height and width directions.

After the CA mechanism, a pair of direction-aware feature maps are formed. They can be complementarily applied to the input feature maps to enhance the target of interest. As shown in Fig. 4, the upper side is the input image, and the lower side is the attention heat map after the output of the attention mechanism. The two sets of images to the left are pure backgrounds, and the two to the right are complex backgrounds. The redder part of the heat map indicates the more important information of interest to the neural network. The bluer part indicates that the information around is irrelevant to the classification.

Improved neck section

YOLOv5s adopts the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) feature pyramid structure in the Neck part. FPN uses a top-down channel to fuse the features for prediction. PAN uses a bottom-up channel to fuse the features conveyed by FPN with localization features. It can effectively solve the information retention of small objects in target detection. However, we found that different input features tend to have unequal contribution rates. Therefore, we adopt the Bi-directional Feature Pyramid Network (BiFPN) weighted bidirectional feature pyramid network structure in the Neck part. Three different network structures are shown in Fig. 5 [42]. BiFPN is based on PAN, which is changed from the initial one-way connection to a cross-scale two-way connection. It introduces learnable weights to judge the importance of different features, enhancing feature fusion and improving detection accuracy.

Improved prediction section

The loss function used by the YOLOv5s prototype in the Prediction section is $Loss_{GIOU}$, whose equation is shown below.

$$Loss_{(GIOU)} = 1 - IoU + \frac{|A^c - u|}{|A^c|} \tag{6}$$

where, IoU (Intersection-over-Union) refers to the intersection rate of the candidate box and the real box, A^c is the intersection area of the predicted box and the real box, and u is the merging area of the predicted box and the real box. The original loss function solves the gradient problem existing in IoU. However, there are also

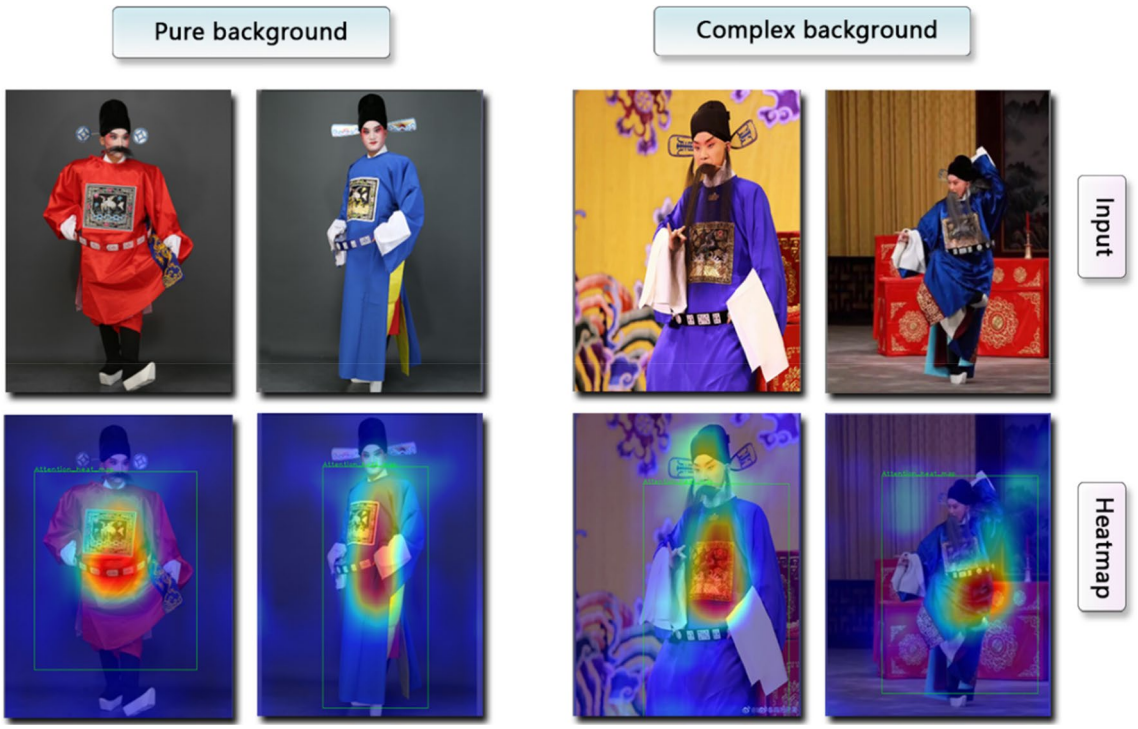

Fig. 4 Attention heat map

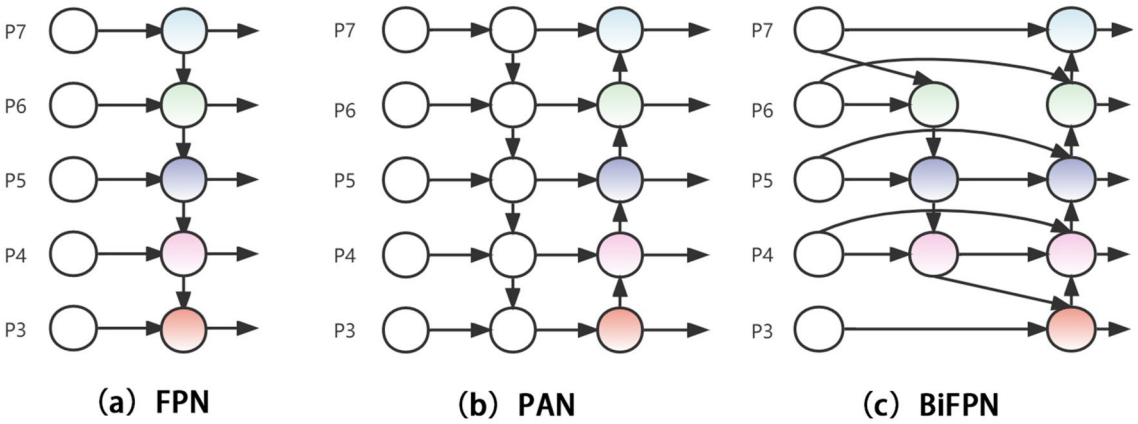

Fig. 5 FPN, PAN, BiFPN structure

the following problems: (i) GIOU will be infinitely close to *IoU* when the predicted box is of the same height as the real box and at the same level. (ii) The convergence speed is slow, and the regression is not accurate enough. To solve the above problems, we introduce DIOU to improve [43]. Its formula is shown as follows.

$$Loss_{DIOU} = 1 - IoU - \frac{\rho^2\left(b, b^{gt}\right)}{c^2}. \tag{7}$$

where, b denotes the centroid of the predicted box, b^{gt} denotes the centroid of the real box, ρ denotes the Euclidean distance between the two centroids, and c denotes the minimum diagonal distance of the region that can contain both the predicted box and the real box. DIOU enables the target detection to focus on the central location by introducing centroids and has a faster convergence rate compared to GIOU and *IoU*, and the accuracy is improved.

Experiment and result

Model *training*

The opera costume data used in this experiment is mainly from field research and web crawlers, consisting of 3000 pieces. Some samples are shown in Fig. 6. The five types of opera costumes mentioned in the introduction were manually labeled using the LabelImg tool. The data set was divided into a training set, and a validation set according to the ratio of 70% and 30%. Finally, we stored the data in the format of the YOLO data set.

The experimental environment of this paper is Windows 10 system, compiled language is Python 3.8, using Pytorch 1.8.1 as a deep learning framework, CUDA version is 10.1. the models involved are run on NVIDIA GeForce RTX2060 GPU. The parameters are set: initial learning rate is 0.01, momentum factor is 0.937, weight decay factor is 0.0005, epoch is 300, and batch size is 16.

The evaluation indexes of model performance are mean Average Precision (mAP), including mAP0.5 and $mAP0.5_{:0.95}$, and the confusion matrix. Among them, mAP0.5 indicates the average detection accuracy of all detection categories when the *IoU* threshold is 0.5; $mAP0.5_{:0.95}$ calculates the average detection accuracy of *IoU* between the threshold 0.5 and 0.95 in steps of 0.05. In general, the higher the *IoU* threshold, the higher the regression capability of the model is required, the higher the detection index at high thresholds, and the more closely the detection results of the model match the actual target. The mAP formula is as follows.

$$mAP = \frac{1}{m}\sum_{i=1}^{m}\sum_{i=1}^{n-1}(r_{i+1} - r_i)p(r_{i+1}) \tag{8}$$

where, m represents all categories, n represents the interpolation points, p represents the precision value, and r represents the recall value corresponding to precison.

In this training model, both the YOLOv5s model and the improved YOLOv5 model (YOLOv5-CBD) use the same data set and parameter settings. The comparison curves are shown in Figs. 7 and 8.

The comparison curve shows that mAP0.5 and $mAP0.5_{:0.95}$ increase with the epoch, and the curve gradually converges from oscillation to convergence when the epoch exceeds 250. Ultimately, YOLOv5s-CBD is higher than YOLOv5s in both mAP0.5 and $mAP0.5_{:0.95}$. The mAP0.5 value for YOLOv5-CBD is 0.863 and for YOLOv5s is 0.832. The $mAP0.5_{:0.95}$ value for YOLOv5-CBD is 0.626 and for YOLOv5s is 0.607.

The confusion matrices for YOLOv5s and YOLOv5-CBD are shown in Fig. 9, respectively. Where *True* indicates the true distribution of a category and predicted indicates the predicted distribution of a category. The confusion matrix shows that YOLOv5-CBD has a higher true positive classification for each category than YOLOv5s, which means that YOLOv5-CBD performs better in precision and recall.

The F1 score for YOLOv5s and YOLOv5-CBD are shown in Fig. 10. It can be regarded as a harmonic mean of model precision and recall, with a maximum value of

Fig. 6 Sample images

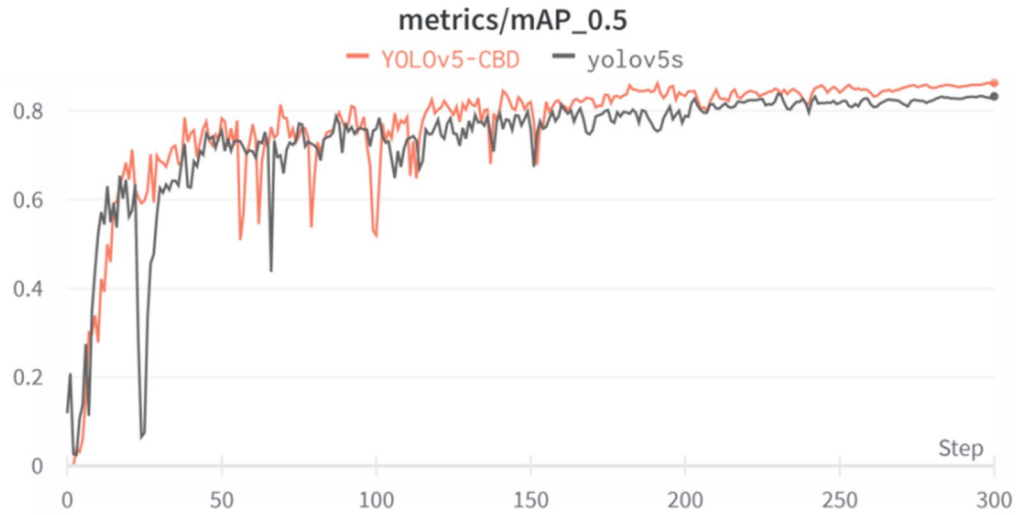

Fig. 7 Comparison of mAP0.5 value between YOLOv5-CBD and YOLOv5s

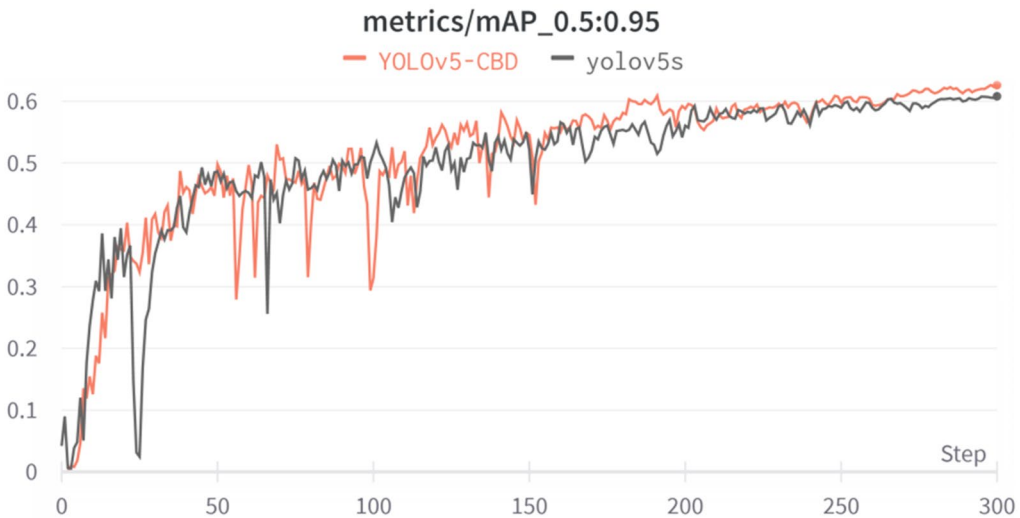

Fig. 8 Comparison of mAP0.5:0.95 value between YOLOv5-CBD and YOLOv5s

1 and a minimum value of 0. Figure 10 shows that when the confidence is 0.241, the F1 score of YOLOv5-CBD reaches the maximum value of 0.81. When the confidence is 0.203, the F1 score of YOLOv5s reaches the maximum value of 0.74, which is lower than the F1 score in YOLOv5-CBD.

Ablation experiment

To verify the optimization effect of the three improvement strategies on the theater costume dataset, we use ablation experiments to test each improvement strategy's effectiveness, and its experimental results are shown in Table 1. Its improvement Model 1 indicates that the attention mechanism is added to the original model, improvement Model 2 indicates that the feature pyramid structure is modified on the original model, and improvement Model 3 indicates that the loss function is modified.

As can be seen from the Table 1, the addition of the CA mechanism increases mAP0.5 by 0.023 percentage points but decreases inference speed by 2 ms; replacing BiFPN to feature pyramid structure increases mAP0.5 by 0.013 percentage points and decreases inference speed by 2 ms; replacing $\text{Loss}_{\text{DIOU}}$ to $\text{Loss}_{\text{GIOU}}$ increases mAP0.5 by 0.01 percentage point and increases inference speed by 2 ms; replacing the three improved strategies into the network structure simultaneously increases mAP by

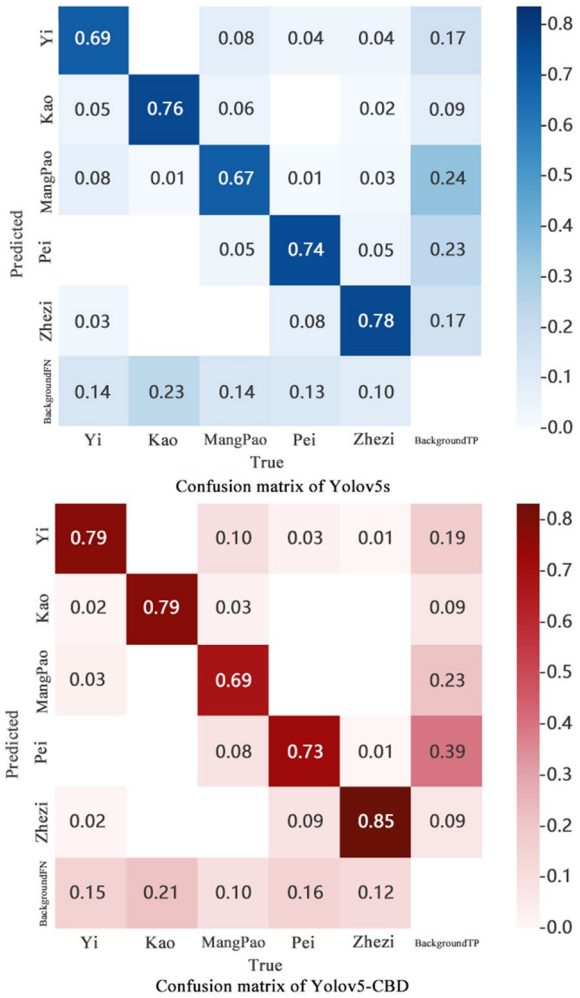

Fig. 9 Confusion matrix

0.031 percentage points and decreases inference speed by 4 ms. The results show that the detection accuracy of the YOLOv5-CBD is improved, but the inference speed is slightly reduced.

Model comparison experiment

To demonstrate the advantages of the improved model in this paper, the YOLOv5-CBD, YOLOv5s, YOLOv5m, YOLOv3, Single Shot multi box Detector (SSD) and Fast-Region CNN(Fast-RCNN) models are used in the same configuration environment and configuration parameters, and the same data set is used for model comparison experiments. Their experimental results are shown in Table 2.

As can be seen in Table 2, compared to YOLOv3, YOLOv5m, Faster-RCNN and SSD, YOLOv5-CBD takes up less memory. This is because it is improved from the network structure of YOLOv5s. YOLOv5s is a lightweight network structure, which has a smaller depth and width than other models. Lightweight structure is more conducive to embedding the network into various platforms.

Compared with the one-stage SSD, its inference time per frame is 25 ms faster, and its mAP0.5 is improved by 0.118. Compared with the two-stage Faster-RCNN, its inference time per frame is 61 ms faster. Compared to YOLOv3, its inference time per frame is 31 ms faster and its mAP0.5 is improved by 0.073. Compared to YOLOv5m, its inference time per frame is 5 ms faster and slightly higher in mAP0.5. Compared to YOLOv5s, whose model weights are almost the same, it is slightly slower in inference speed but improves by 0.029 in mAP0.5. From the results, without increasing the memory, the accuracy of the model has been improved. It can

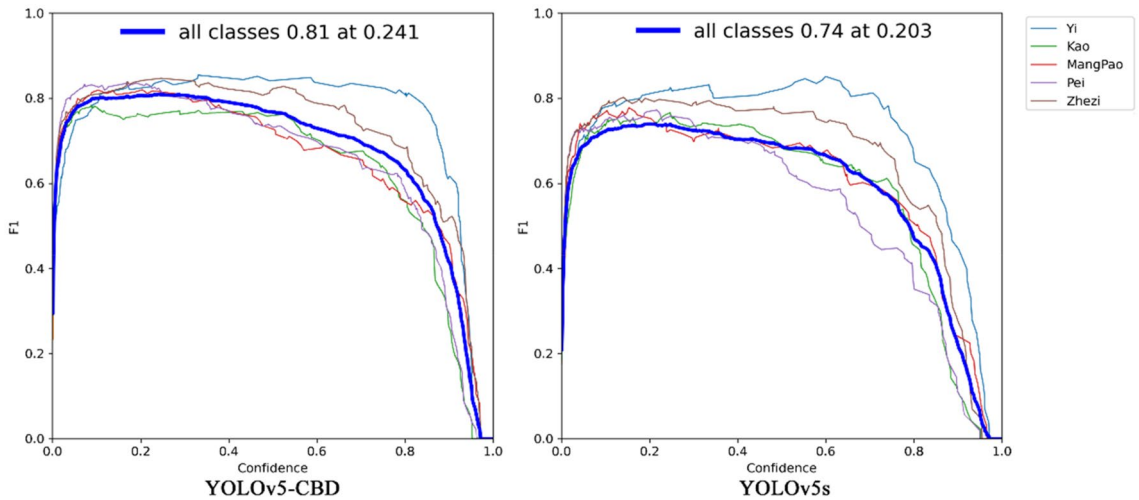

Fig. 10 F1 score

Table 1 Ablation experimental result

Models	Add attention mechanism	Modify feature pyramids	Modify the loss function	mAP0.5	Time per frame/ms
YOLOv5s	×	×	×	0.832	24
Improved Model 1	√	×	×	0.855	26
Improved Model 2	×	√	×	0.845	26
Improved Model 3	×	×	√	0.842	22
YOLOv5-CBD	√	√	√	0.863	28

Table 2 Experimental comparison results

Models	mAP0.5	Time per frame/ms	Memory size/MB
SSD	0.745	53	92.2
Faster-RCNN	0.855	89	146
YOLOv3	0.79	59	231
YOLOv5m	0.852	33	40.2
YOLOv5s	0.832	24	13.7
YOLOv5-CBD	0.863	28	13.4

be seen from the ablation experiment that each modification strategy can help the model improve the accuracy. Although it is slightly slower than yolov5s in terms of reasoning speed, it can still meet the needs of real-time detection.

Discussion

Based on the costume dataset collected in this thesis, combined with the ablation experiment, compared with the original network:

i. Adding the attention mechanism can improve detection accuracy, but the detection speed will be slightly slower. Because the CA module is inserted in the last part of Backbone, which will increase the computation of the network, and after the output, gives the Neck part with the channel and location information of interest, which improves the recognition accuracy.
ii. Replacement feature pyramid also improves the detection accuracy, but the improvement is not as large as the attention mechanism, and the detection speed is slightly slower. The weighted bidirectional feature pyramid introduces learnable feature weights for importance judgment. Cross-scale bidirectional connectivity improves detection accuracy, and feature fusion is enhanced when features are input.
iii. Replacing the loss function GIOU with DIOU improves the detection speed and accuracy. DIOU not only describes whether the predicted bounding box contains the true bounding box but also describes the relative position relationship between them, which enables the target detection to focus on the central position and improves convergence speed.

Combined with the comparison experiment compared with the original network:

i. YOLOv3 is quite different from YOLOv5 in network structure. YOLOv3 lacks Mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling operations at the input. YOLOv5 uses the FPN + PAN structure in the Neck part, while YOLOv3 uses FPN. Therefore, it can be seen that the results of yolov5 are better than those of yolov3 in all aspects.
ii. The Faster R-CNN model belongs to two-stage object detection. Two-stage object detection first generates region proposals, and then uses convolutional neural networks to classify region proposals. Faster R-CNN model can achieve high detection accuracy, but there are cases of slow speed, long training time, take up a lot of memory and relatively high false positives.
iii. The SSD model has lower detection accuracy, slower inference per frame, and large weight files. SSD model uses a pyramid structure, which performs SoftMax classification and position regression on feature maps of different sizes. In contrast, the YOLO model utilizes only the highest layer of feature maps for input during detection.
iv. The YOLOv5m model has relatively higher detection accuracy, slower inference per frame, and larger weight files. YOLOv5m has a deeper and broader model structure than YOLOv5s, which means that the model improves detection accuracy

but also slows down detection speed as it increases the size of the network.

v. The YOLOv5-CDB model provides higher detection accuracy, with slightly slower inference per frame, while keeping the weighted file size almost the same. YOLOv5-CBD is improved based on the YOLOv5s model, which is a lightweight network. With the addition of the attention mechanism, replacement of the feature pyramid and modification of the loss function, the feature extraction capability of the network is improved, which can effectively improve the accuracy of the network by enhancing the problem of missed detection due to small targets or occlusion.

Meanwhile, to verify the feasibility of YOLOv5-CBD, we take some representative photos from the test set. For example, Fig. 11 shows the comparison results of model detection between YOLOv5-CBD and YOLOv5s in different scenes. The left figure represents the detection results of the YOLOv5-CBD model, and the right figure is the detection results of the YOLOv5s model. Figure 11a compares the detection results of the two models for the occluded target. The beard and hand occlude Pei on the far right in the test figure, and it can be seen that there is a missed detection in the right figure, while the left figure identifies it successfully; Fig. 11b compares the detection results of light-colored clothing, showing that the right figure misjudges the light-colored target and identifies the ZheZi as MangPao. In contrast, the left figure can correctly judge the ZheZi; Fig. 11c compares the detection results of crossed targets, in which Pei and ZheZi appear to stage cross. The right figure has a missed detection for crossed targets. In contrast, the left figure can identify them successfully; Fig. 11d compares the detection results of dense targets. The rightmost Yi in the right figure shows a missed detection, while it can be detected in the left figure; Fig. 11e compares the detection results of the side view. The MangPao in side view cannot be detected in the right figure, while it can be detected in the left figure.

To sum up, in recognition of opera costumes facing different complex scenes, the YOLOv5s model can have missed detection and false detection. In contrast, In contrast, the YOLOv5-CBD model has more accurate accuracy and fast inference and can have better robustness in the face of complex theatrical stage backgrounds, reflecting a superior performance.

Conclusion

This paper applies artificial intelligence technology to study the identification of Chinese traditional opera costume. The conclusions are as follows:

i. To solve the problem of identification difficulty and poor identification effect of opera costumes on the real drama stage, we propose an improved YOLOv5-based opera costume classification and recognition method. It has higher recognition and detection performance than other network models and can be carried on embedded platforms and mobile terminals to achieve efficient detection.

ii. By improving the network structure of YOLOv5, it is able to improve the accuracy of detection as well as solve the recognition problem of complex drama scenes. The detection accuracy is improved by adding the CA mechanism to the Backbone section to give the neural network a more focused feature map. By replacing the BiFPN feature pyramid in the Neck section, feature fusion is enhanced to improve the retention of information on small objects in target detection. By replacing the DIOU loss function, the detection accuracy is improved, and the inference time per frame is also decreased. The experimental results show that the improved model has high accuracy and robustness in detecting complex scenes facing obscured targets, light-colored costumes, cross targets, dense targets and different angles. It is 3.1%, 1.8%, 2.1% and 1.6% higher than the original model in mAP0.5, $mAP0.5_{:0.95}$, Precision and Recall, which meet the requirements of accuracy and real-time in opera costume recognition.

iii. The establishment of the opera costume recognition model can establish its corresponding visual label to the opera costume database, speed up the construction of the opera costume database, and play a positive role in developing of the opera costume culture. However, there is still room for improvement in the recognition of opera costumes due to the limitations arising from various styles of costumes, complex stage action performances, uneven lighting and other phenomena. The focus of future theater costume research is still on recognition accuracy as well as inference speed. The recognition accuracy can start by expanding the costume data set and increasing the costume data complexity and richness on the theater stage to improve the generalization ability of the model.

(a) Occlusion target detection

(b) light-colored clothing target detection

(c) Cross target detection

(d) Dense target detection

(e) Side angle target detection

Fig. 11 Comparison of different scene model detection

The inference speed can be accelerated with the help of embedded GPU to improve the real-time performance of detection.

Acknowledgements
Not applicable.

Author contributions
KL developed the research idea. KL and CZ wrote the manuscript. All authors read and approved the final manuscript.

Funding
The work was financially supported by the Humanities and Social Sciences project of the Ministry of Education, China (No. 22YJAZH064), the Later Funded Project of Philosophy and Social Science Research of the Ministry of Education, China (No. 22JHQ008), the National Endowment for the Arts, China (2018-A-05-(263)-0928) and the Youth Innovation Team of Shaanxi Universities, China.

Availability of data and materials
The datasets generated during the current study are available from the corresponding author on reasonable request.

Code availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Received: 9 December 2022 Accepted: 16 February 2023
Published online: 25 February 2023

References

1. Lee Y-S. A study of stage costume of Peking Opera. Int J Costume Cult. 2003;6(1):38–51.
2. Monna F, et al. Deep learning to detect built cultural heritage from satellite imagery—Spatial distribution and size of vernacular houses in Sumba, Indonesia. J Cult Herit. 2021;52:171–83. https://doi.org/10.1016/j.culher.2021.10.004.
3. Narag MJG, Soriano M. Discovering artistic influences of painters from expressionism, impressionism, and surrealism art movements using convolutional neural network. J Cult Herit. 2021;51:182–93. https://doi.org/10.1016/j.culher.2021.08.015.
4. Yang Y, Lei T. The inheritance and future development direction prediction of opera culture based on cloud communication under the background of big data. J Sens. 2022;2022:9. https://doi.org/10.1155/2022/1910766.
5. Wang H, et al. Bodhisattva head images modeling style recognition of Dazu Rock Carvings based on deep convolutional network. J Cult Herit. 2017;27:60–71. https://doi.org/10.1016/j.culher.2017.03.006.
6. Cintas C, et al. Automatic feature extraction and classification of Iberian ceramics based on deep convolutional networks. J Cult Herit. 2020;41:106–12. https://doi.org/10.1016/j.culher.2019.06.005.
7. Hatir ME, Barstuğan M, Ince I. Deep learning-based weathering type recognition in historical stone monuments. J Cult Herit. 2020;45:193–203. https://doi.org/10.1016/j.culher.2020.04.008.
8. Liu K, et al. Archaeology and restoration of costumes in tang tomb murals based on reverse engineering and human–computer interaction technology. Sustainability. 2022;14(10):6232. https://doi.org/10.3390/su14106232.
9. Lee MS, et al. A comparative analysis of the characteristics and images of costume colors in the traditional plays of Korea, China, and Japan. Color Res Appl. 2012;37(4):302–12. https://doi.org/10.1002/col.20673.
10. Kim J-E. Color characteristics of the costumes of the Beijing Opera. J Korean Soc Costume. 2009;59(2):143–53.
11. Zhang J. The application and analysis of opera costume elements in the modern costume design. In: E3S web of conferences. 2021. EDP Sciences. https://doi.org/10.1051/e3sconf/202123704015.
12. Gao T, Kuang L. Feature data extraction algorithm technology on traditional costume innovation. In: International conference on application of intelligent systems in multi-modal information analytics. Berlin: Springer; 2021. https://doi.org/10.1007/978-3-030-74814-2_131.
13. Luo J. Analysis on application of traditional Chinese opera costume elements in contemporary costume design. In: 2017 3rd international conference on economics, social science, arts, education and management engineering (ESSAEME 2017). Atlantis Press; 2017. https://doi.org/10.2991/essaeme-17.2017.88.
14. Shin KS. A study of the costume used in Peking Opera. J Korean Soc Costume. 2010;60:132–50.
15. Liu K, et al. Study on digital protection and innovative design of Qin opera costumes. Herit Sci. 2022;10(1):1–15. https://doi.org/10.1186/s40494-022-00762-x.
16. Hou J et al. The implementation of a Beijing Opera interactive display system based on motion recognition. In: 2021 IEEE international conference on artificial intelligence and computer applications (ICAICA). New York: IEEE; 2021. https://doi.org/10.1109/ICAICA52286.2021.9498245.
17. Jia L, Gao T. Research on computer intelligent image recognition algorithm in Chinese local opera clothing education teaching. In: International conference on cognitive based information processing and applications (CIPA 2021). Berlin: Springer; 2022. https://doi.org/10.1007/978-981-16-5854-9_28.
18. Caspari G, Crespo P. Convolutional neural networks for archaeological site detection—Finding "princely" tombs. J Archaeol Sci. 2019;110:104998. https://doi.org/10.1016/j.jas.2019.104998.
19. Orengo HA, Garcia-Molsosa A. A brave new world for archaeological survey: automated machine learning-based potsherd detection using high-resolution drone imagery. J Archaeol Sci. 2019;112:105013. https://doi.org/10.1016/j.jas.2019.105013.
20. Jalandoni A, Zhang Y, Zaidi NA. On the use of Machine Learning methods in rock art research with application to automatic painted rock art identification. J Archaeol Sci. 2022;144:105629. https://doi.org/10.1016/j.jas.2022.105629.
21. Wiley V, Lucas T. Computer vision and image processing: a paper review. Int J Artif Intell Res. 2018;2(1):29–36. https://doi.org/10.29099/ijair.v2i1.42.
22. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;6:679–98. https://doi.org/10.1109/TPAMI.1986.4767851.
23. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). New York: IEEE; 2005. https://doi.org/10.1109/CVPR.2005.177.
24. Rosten E, Drummond T. Machine learning for high-speed corner detection. In: European conference on computer vision. Berlin: Springer; 2006. https://doi.org/10.1007/11744023_34.
25. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell. 2002;24(7):971–87. https://doi.org/10.1109/TPAMI.2002.1017623.
26. Aulia N, Arnia F, Munadi K. HOG of region of interest for improving clothing retrieval performance. In: 2019 IEEE international conference on cybernetics and computational intelligence (CyberneticsCom). New York: IEEE; 2019. https://doi.org/10.1109/CYBERNETICSCOM.2019.8875636.
27. Lorenzo-Navarro J, et al. Evaluation of LBP and HOG descriptors for clothing attribute description. In: International workshop on video analytics for audience measurement in retail and digital signage. Berlin: Springer; 2014. https://doi.org/10.1007/978-3-319-12811-5_4.

28. Bossard L, et al. Apparel classification with style. In: Asian conference on computer vision. Berlin: Springer; 2012. https://doi.org/10.1007/978-3-642-37447-0_25.
29. Yu N, et al. An improved method for cloth pattern cutting based on holistically-nested edge detection. In: 2021 IEEE 10th data driven control and learning systems conference (DDCLS). New York: IEEE; 2021. https://doi.org/10.1109/DDCLS52934.2021.9455545.
30. Redmon J, et al. You only look once: unified, real-time object detection. Proc IEEE Conf Comput Vis Pattern Recognit. 2016. https://doi.org/10.1109/CVPR.2016.91.
31. Liu W, et al. Ssd: single shot multibox detector. In: European conference on computer vision. Berlin: Springer; 2016. https://doi.org/10.1007/978-3-319-46448-0_2.
32. Ren S, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
33. He K, et al. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. Venice: IEEE; 2017. https://doi.org/10.1109/ICCV.2017.322.
34. Ren F, et al. Research on garment image classification and detection algorithm based on improved deep learning. In: 2022 7th international conference on intelligent computing and signal processing (ICSP). New York: IEEE; 2022. https://doi.org/10.1109/ICSP54964.2022.9778478.
35. Yin D, et al. Pig target detection from image based on improved YOLO V3. In: International conference on artificial intelligence and security. Berlin: Springer; 2021. https://doi.org/10.1007/978-3-030-78615-1_9.
36. Iannizzotto G, Bello LL, Patti G. Personal protection equipment detection system for embedded devices based on DNN and fuzzy logic. Expert Syst Appl. 2021;184:115447. https://doi.org/10.1016/j.eswa.2021.115447.
37. Xiang J, et al. Clothing attribute recognition based on RCNN framework using L-Softmax loss. IEEE Access. 2020;8:48299–313. https://doi.org/10.1109/ACCESS.2020.2979164.
38. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser. 2019;1168(2):022022. https://doi.org/10.1088/1742-6596/1168/2/022022.
39. Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies. New Orleans: Association for Computational Linguistics; 2018. https://doi.org/10.48550/arXiv.1803.02155.
40. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE. 2018. https://doi.org/10.1109/CVPR.2018.00745.
41. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville: IEEE; 2021. https://doi.org/10.1109/CVPR46437.2021.01350.
42. Tan M, Pang R, Le QV. Efficientdet: scalable and efficient object detection. In 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle: IEEE; 2020. https://doi.org/10.1109/CVPR42600.2020.01079.
43. Zheng Z, et al. Distance-IoU loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 2020. https://doi.org/10.1609/aaai.v34i07.6999.

Publisher's Note