Neural Computing and Applications (2023) 35:973–991
https://doi.org/10.1007/s00521-022-07823-z

ORIGINAL ARTICLE

HPFace: a high speed and accuracy face detector

Xiao Ke[1,2] · Wenzhong Guo[1,2] · Xu Huang[1,2]

Received: 7 December 2021 / Accepted: 6 September 2022 / Published online: 24 September 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

With the application of artificial intelligence technology, face detection is now not only concerned with accuracy but detection speed as well. However, most previous works have relied on heavy backbone networks and required prohibitive run-time resources, which seriously restricts their scope for deployment and has resulted in poor scalability. In this study, we used YOLOv5s, which has a good detection rate and accuracy, as the baseline network. First, we added a none-parameter channel attention self-enhancement module to allow the backbone of the network to capture the characteristic features of the face more effectively. Second, a low-level feature fusion module was added to enhance the features of shallow neural layers and then fuse them with the features of deeper layers. Third, a receptive field matching module allows the network's perceptual field to better match the scale of actual faces. Finally, contextual information based on face key points allows the face detector to exclude more cases of error and missed detections. On the most popular and challenging face detection dataset, WIDER FACE, our model performed better than the original network, with improvements of 3.8, 4.4, and 11.6% on the easy, medium, and hard subsets, respectively, and achieved a rate higher than 72 FPS, which meets the real-time requirements.

Keywords Face detection · Feature fusion · Attention mechanism · Receptive field matching · Context information

1 Introduction

One of the goals of face detection [1] is to locate the position of a face in a video automatically. This is an important task in computer vision and has received widespread attention in recent years. With the increasing demand for intelligent landing requirements in areas such as security and attendance, many studies have attempted to improve face detection by deepening the model structures, enhancing data, and expanding labels, among others. However, this can make the algorithm and the model more complex. Moreover, various facial applications, such as face alignment [2], recognition [3], and verification [4], require the face to be detected quickly, which places a high demand on the detection rate. Therefore, a benchmark network that meets the real-time requirements of the application and has a high level of accuracy is required.

Although CNN-based [5] face detection has been widely studied, detecting faces in real scenarios where high variability is present, including the effects of pose, occlusion, expression, appearance, and illumination, remains a challenge. Many of today's methods perform well on publicly available datasets, but do not perform well in terms of detection speed. For example, the DSFD [6] published by Tencent performs well on the WIDER FACE [7] dataset, but only at a rate of approximately 5 FPS, which is not sufficient to meet the real-time requirement. On the other hand, many lightweight networks [8, 9] can meet real-time metrics, but not with high enough accuracy. At this point, the YOLO [10] family of networks has been selected as the preferred networks, because a good balance between speed and accuracy is shown in their midst, and YOLOv5 [11] has the best performance in its class and was, therefore, chosen as the baseline for this study.

Xiao Ke and Xu Huang have contributed equally.

✉ Wenzhong Guo
guowenzhong@fzu.edu.cn

Xiao Ke
kex@fzu.edu.cn

Xu Huang
n190327035@fzu.edu.cn

[1] Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China

[2] Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, Fuzhou 350003, China

Face detection is an intuitive application of traditional object detection [12]. In other words, the face is a special object, so some of the traditional improvements to object detection can also be applied to the face detection task. However, unlike traditional object detection, the face as a class of object has its own characteristics, and these can also be used to improve on the accuracy of detection.

Traditional object detection methods tend to favor one aspect of accuracy or speed, with high accuracy models generally having lower detection rates, and many lightweight networks having good detection rates, but accuracy is the enhancement point that they ignore. The goal of this study was to achieve both, simultaneously. As shown in Fig. 1, the YOLO [10] series has a good balance between these two metrics, and YOLOv5 [11] has the best performance in this series. However, as with many deep neural networks, YOLOv5 still suffers from poor discriminability of the feature maps after multiple layers of convolution and pooling operations [13]; therefore, we draw on the channel attention mechanism [14] and propose a none-parameter channel attention self-enhancement (NCAS) method and a low-level feature fusion (LFF) module to address these problems. In addition, this study addresses the problem of mismatch between the perceptual receptive field [15] of the detector and the face ratio in face detection; we propose a receptive field matching (RFM) module suitable to target of the face specifically, which improves the performance in detecting faces. We combine the above methods and propose an improved method based on YOLOv5s to achieve a SOTA result on the WIDER FACE [7] dataset and ensure that the detection rate meets the real-time requirement.

In summary, we made the following contributions to face detection studies:

1. In traditional neural networks, important feature information is often not learned by the network due to the scattered distribution of weights in the channels. In this paper, an NCAS module is proposed to enable the network to learn more of the expected features in order to improve detection performance. Because NCAS does not carry any activation or convolution structures, it carries no parameters and, therefore, has little effect on the detection speed of the overall module.
2. In existing neural networks, most feature extraction modules apply a square receptive field. However, by analyzing the data, we found that most faces have rectangular aspect ratios, which result in a mismatch between the network receptive field and the facial target. In this paper, we propose an RFM module that uses different scales of perceptual fields for face detection, which can make the network more robust in detecting specific facial targets.
3. The discriminability of the feature map is often poor after the deep neural network has been convolved in multiple layers. In order for facial features with better discriminability and robustness to be learned by the network, we designed an LFF module to fuse the features of the image in the shallow layer of the network with the features from the deeper layer after enhancement, so that the network can eventually learn better quality image feature information.
4. Object detection tasks can generally improve performance by using other object features that are strongly correlated with the target to be detected at the location to aid detection. In this paper, five face key points that are strongly correlated with faces are introduced into the method as background information to aid face detection and achieve a large performance improvement especially in the hard subset of WIDER FACE (Fig. 2).

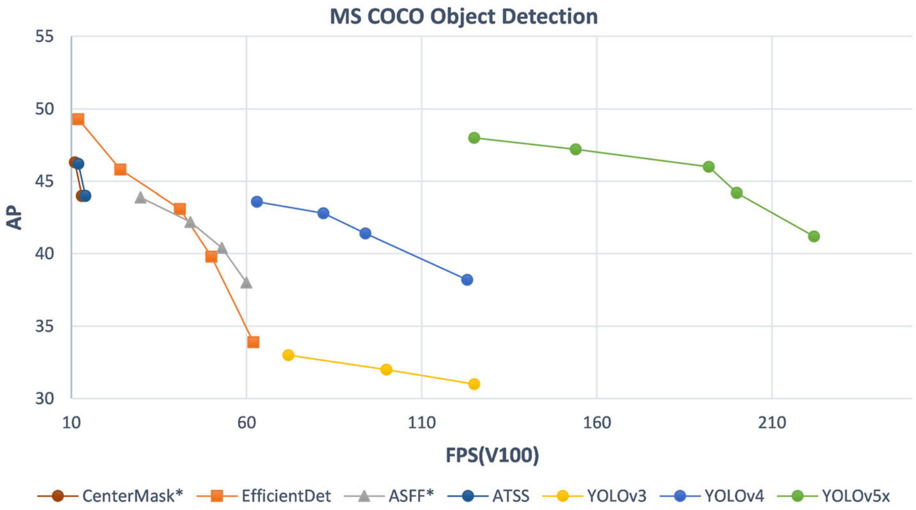


Fig. 1 Comparison of YOLOv5 with other well-known deep neural networks and with the same series of networks. It can be seen that the YOLO series networks perform better in terms of balancing accuracy and detection rate, and YOLOv5 does better than its peer series

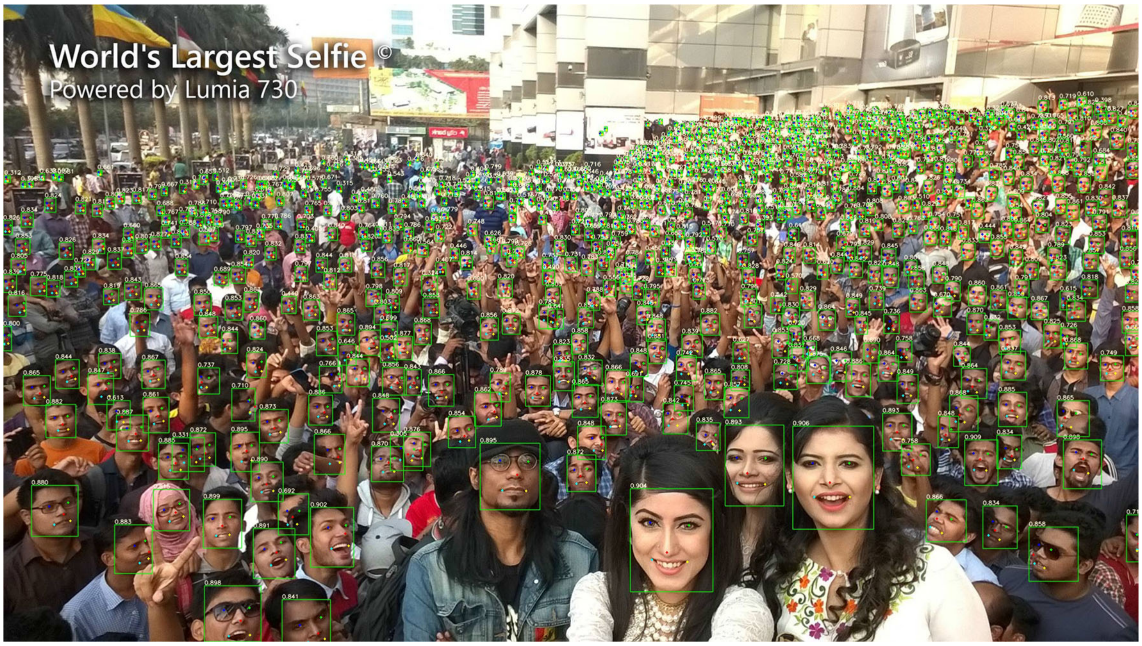


Fig. 2 We validated the final model of this article in the images containing the most faces in the world. The verification results show that the final model in this paper has reached the level of SOTA

2 Related work

Since the 1990s, face detection has been a challenging research field. "Harr-like" features and "AdaBoost" [16] had been used by Viola and Jones first to train face detectors and achieve good accuracy and efficiency. After the 2012 ImageNet [17] classification challenge, CNN-based face detection methods became mainstream, and many new CNN-based methods emerged. A multi-task cascaded network was proposed by MTCNN [2] that can detect faces and their alignment; RetinaNet [18] which used a new loss function "Focal Loss" to solve the problem of serious imbalance in positive and negative samples in one-stage object detection; Faster RCNN [19] (Girshick 2015) solved the problem of category imbalance through two-stage cascade and sampling heuristics deep learning; In addition, CEDN [20] (Wang et al. 2019) is a coupled encoder-decoder network that jointly detects faces and localizes facial key points. The encoder and decoder generate response maps for facial landmark localization. Faceboxes [21] (Lei et al. 2019) propose a novel face detector that consists of the rapidly digested convolution layers and multiple scale convolution layers. The former was designed to enable Faceboxes to achieve CPU real-time speed, while the latter aimed to enrich the features and discretize anchors over different layers to handle faces of various scales. DCFPN [22] is a novel face detector that has a dense anchor strategy and a scale-aware anchor matching scheme to improve the recall rate of small faces with high accuracy at CPU real-time speed. Chen et al. proposed that the STC and STR modules, in their SRN [23] method, should be used to filter the negative samples at the bottom and adjust the position and size of the high-level anchors. Cascade RCNN [24] improves the positioning accuracy of face detection by cascading RCNN with different IoU thresholds. Li et al. [6] proposed a two-shot model structure, DSFD, in which the feature enhancement module was also applied to achieve good accuracy in face detection.

2.1 Feature learning

Before the popularity of deep learning, face detection mainly relied on artificially preset features, such as skin color, the vertical line between the binocular and forehead-chin lines, and others, and used Harr-like features [16], control point set, edge histogram, and other technologies. With the development of deep learning, artificial methods that lack computer guidance are gradually being abandoned. Since the ImageNet classification challenge was won by "Alexnet" [25] by a considerable margin in 2012, deep learning methods led by convolutional neural networks have gradually become mainstream in the field of computer vision. Methods such as CascadeCNN [24] and MTCNN [2] use sliding windows in the first stage to build image pyramids, but they have a slow speed and a low memory utilization rate. The RCNN [4] series works through selective search, obtaining the region proposal and classifying each normalized image region using CNN. The YOLO series presets anchors of various scales on an image, then the image is divided into $S \times S$ grids, and each grid is used as a center point to predict the target position, which results in a faster speed. Later, DSFD and SRN [26] of Tencent introduced a two-shot network structure, which made it possible to supervise the training process of the entire model more effectively.

2.2 Attention mechanism

The attention mechanism [14] was inspired by the fact that humans pay particular attention to a few important localities when observing their surroundings and was created to enable networks to learn important information from different localities and combine them. The attention mechanism has been shown to be an important method for improving the performance of deep learning networks. In particular, a paper published by Google, "Attention is all you need," proposed the idea of replacing CNN and RNN structures entirely with attention mechanism, leading to the proliferation of such methods. The last ImageNet champion model, SE-Net [27], was the first effective channel attention learning mechanism and achieved good results. The subsequent development of attention mechanisms can be divided into two broad areas: augmented feature aggregation and the combination of channel space attention; however, these approaches are based on complex attention modules that increase the complexity of deep learning models.

ECANet [28] that appeared in 2019 was an improvement of SE-Net [27], which reduced the complexity of the model while learning effective channel attention. It was realized with a local cross-channel interaction strategy without dimensionality reduction and considered the K channels of the current channel and its neighbors; the K value could be adjusted adaptively. The proposal of this method has pushed the application and research of the attention mechanism to a new height. We combined facial features and added the improved channel attention mechanism module to the residual structure of the benchmark network so that the network could learn more important parts of the channel and combine them to improve the performance of the network.

2.3 Receptive field

The feature vector at a location in the feature map of a layer in a neural network is obtained from a fixed region on the input of the previous layer, which is called the receptive field. In object detection tasks, the larger the field of perception, the better; this ensures that important feature information is not overlooked before making the final prediction. For anchor-based work, it is necessary to ensure that the anchor strictly corresponds to the field of perception, otherwise the detection performance is seriously affected. There are two main methods for increasing the receptive field:

1) *Pooling* [13] Here, the reduction in image dimensionality can be achieved in a neural network by pooling to increase the perceptual field; however, this method results in a certain loss of image features.
2) *Dilated Convolution* [29] The additional holes are added to the standard convolution layer. The advantage of this is that it increases the perceptual field without pooling the lost information, so that each convolution output contains a larger range of information.
3) *Contextual Information* [30] The basis of computer vision technology is derived from the laws and habits of human observation of objects. When humans observe a target object, they often also rely on the surrounding things to assist their judgment, so in the target detection task, if the target to be detected is understood as the current paragraph to be read, then the surrounding environment, other objects can be seen as contextual information, because in real life, all objects cannot exist alone, but must have some connection with other objects and the environment around them.

Therefore, in the target detection task, the interaction information between the object to be detected and the surrounding objects and environment can usually be used as auxiliary information for the detection of the neural network model, which is called contextual information. Contextual information in computer vision does not refer to other features obtained directly from the appearance of the object to be detected, but rather from feature data of the domain, annotations of other targets, spatial location information or statistical information. However, not all surroundings help to improve the performance of target detection, and the addition of meaningless background noise may even have a negative impact on the performance of the model. Therefore, it is common to use objects, environments, and parts of objects that co-exist with the target to be detected as for contextual information, for example, when doing vehicle detection, the common contextual information is the road, wheels, driver, etc.

3 Method

Among the series of YOLOv5 networks, YOLOv5s, as the lightest version, has the same network structure as YOLOv5, but its model depth multiplier is one-third of the original network, layer channel multiplier is two-fifths of the original network, and the model size is only 0.07 of

YOLOv5, which is more in line with the objective of building a fast and accurate face detection network in this paper; YOLOv5s was therefore chosen as our baseline network, with the addition of improvements to conventional detection and a hurry-up method that treats faces as special targets (Fig. 3).

3.1 None-parametric channel attention self-enhancement

In the development of deep neural networks, the loss of the model decreases as the number of layers of the network increases, and subsequently, saturation has been demonstrated through numerous experiments; if the depth of the network is increased, the loss will increase, leading to a decrease in model accuracy, a phenomenon known as network degradation. When network degradation occurs, a neural network with fewer layers can achieve better training than a deeper network. At this point, if the features from the lower layers are passed to the higher layers, the results would at least be no worse than those of a shallow network. Based on this idea, a residual network was created. However, owing to the method of passing the features of the lower layers to the higher layers through direct mapping, the residual network [31] inevitably made the features learned by the model more confusing. This resulted in the feature extraction layer not being able to extract the effective features of the target to be detected. The benchmark network chosen for this study, YOLOv5, also has many residual structures and, therefore, suffers from the same problems described above.

The above problems are well solved by the advent of the channel attention mechanism. The attention mechanism in neural networks is based on the visual attention of the human eye, which quickly scans a global image to identify the target area that needs to be focused on, generally referred to as the focus of attention. Then, more attention resources are devoted to this area to obtain more detailed information about the target, while suppressing other useless information. Channel attention mechanisms have shown great potential for improving the performance of neural networks; however, many related approaches that have been used to improve the performance of CNNs introduce complex attention modules, which inevitably increase the computational burden and slow down the detection rate. The recent emergence of ECANet [28] seems to be a good solution to this problem, as it achieves a good channel attention mechanism with a small number of parameters through a non-degenerate local cross-channel interaction strategy. However, it still requires parameters, requiring the neural network to calculate gradients during training, thereby increasing training costs, and affecting the rate during detection.

Therefore, to further meet the real-time requirements for this study and inspired by ECANet, we designed a none-parameter channel attention module to achieve the goal of

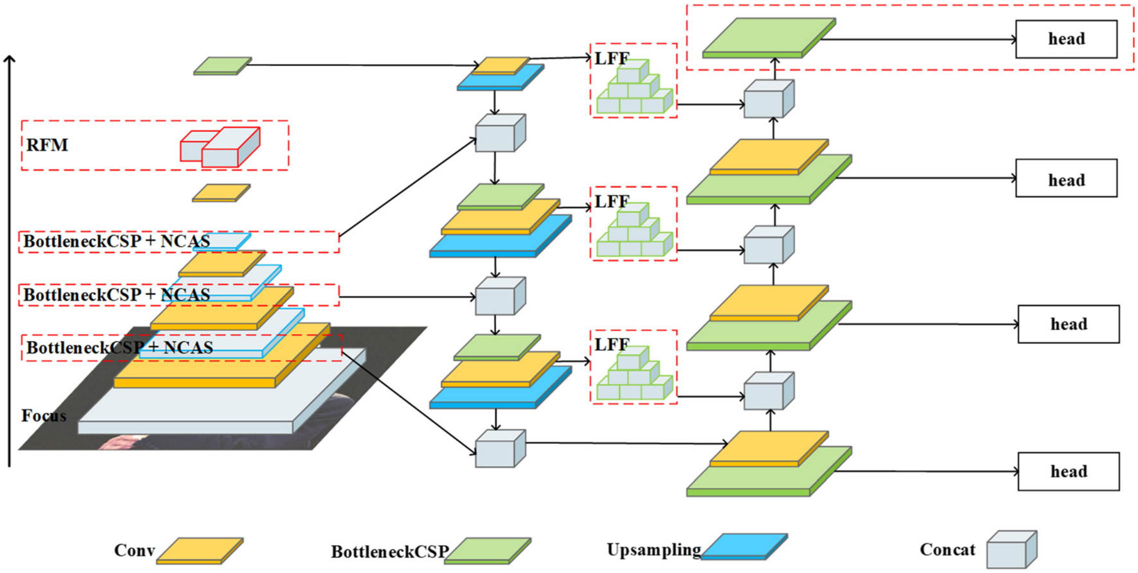


Fig. 3 Improved YOLOv5s model structure diagram. The red dashed boxes indicate the three improvements proposed in this article. First, the image features are passed down through layers such as Focus in the image input model, and then through (1) the Bottleneck + NCAS, using the None-parameter attention module to enable the more useful features to be passed down the backbone of the network. Next, the feature maps are passed into (2) the Receptive Field Matching Module (RFM) that extracts the features; the features that better match the size of the face are passed down. In the next network structure, (3) LFF is responsible for fusing the low-level and deep-level features after enhancement, making the features learned by the network more robust; (4) we added a large-scale output layer that fits the key points of the face at the end of the network layer

improving the detection accuracy without reducing the detection rate. We referred to the channel attention module in ECANet and removed two of the one-dimensional convolutional structures; therefore, the unique parameter was removed, and the input feature map was directly regressed. Then, the feature map was combined with itself through the self-enhancement method to improve the performance of deep learning. We added the improved parameter-free channel attention self-enhancement module to the residual structure of BottleneckCSP in YOLOv5s, to enable the backbone network to learn information that is more in line with the target characteristics. The formula is given as follows:

$$F_{(x)} = SIG(GAP(x)) * x \tag{1}$$

where x is the input feature map, GAP is the global average pooling operation, and SIG is the sigmoid activation function of the globally averaged pooled feature map. Through simple operations and combined with itself, we can make this module no matter in terms of accuracy or speed. The experimental results show that our none-parameter channel attention self-enhancement module has a better effect on improving the performance of deep convolutional neural networks than the channel attention module of ECANet, and because the parameter layer was removed, the computational cost was reduced, and the detection rate was higher.

Figures 4 and 5 show the difference between our module and ECANet and show the way our module is used in the residual structure of YOLOv5 and BottleneckCSP.

3.2 Receptive field matching module

ResNet [32] or VGGNet [33] is currently used by most detection networks as a basic feature extraction module with a square perceptual field. In this study, the benchmark network YOLOv5 was used, and its feature extraction process is also based on a square field. The square receptive field performs well in traditional object detection, but in scenarios with real faces, as in the WIDER FACE dataset, a large proportion of faces are not square: their aspect ratios are greater than 2 or less than 0.5. Therefore, if only the square perceptual field is used for feature extraction, it will affect the detection accuracy (Fig. 6).

The SPP module in the benchmark network of this paper uses a maximum pooling layer of 3 sizes for feature extraction, but its shape is also 1:1, which does not match some of the face proportions, so the above problem also occurs during training and detection. To solve this problem, we borrowed the receptive field enhancement module from SRN and combined the ideas of SRN and SPP modules, using two types of convolutional kernels, 5*3 and 3*5, to provide a rectangular perceptual field that matches the proportion of the face, and with 3*3 and 5*5 pooling operations to provide a square perceptual field, which can make the network's perceptual field both rectangular and square, more suitable for the specific target of the face. This module allows the network to have both rectangular and square receptive fields to better match the face as a

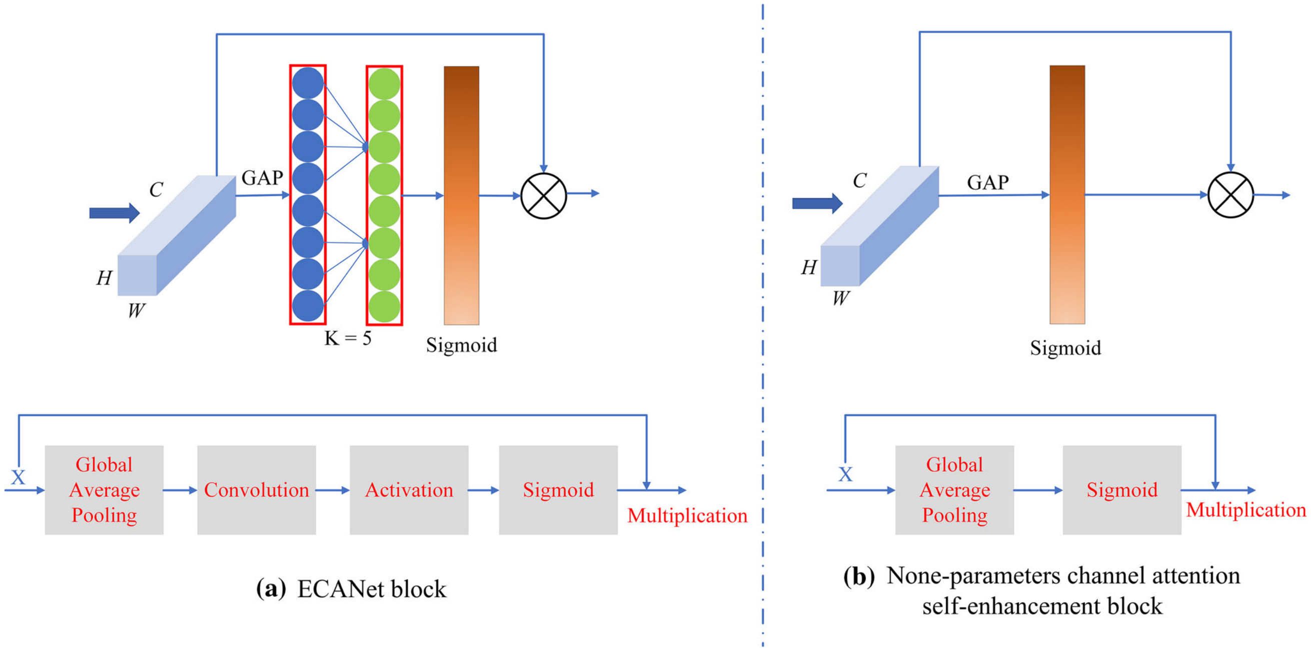


Fig. 4 ECA uses global average pooling (GAP) for channel aggregation and then generates channel weights by performing a fast convolution of size k, which is then multiplied with activation and sigmoid, and a feature map is outputted. NCAS performs the GAP operation directly on the channels, followed by the sigmoid operation and finally multiplied with the original feature map

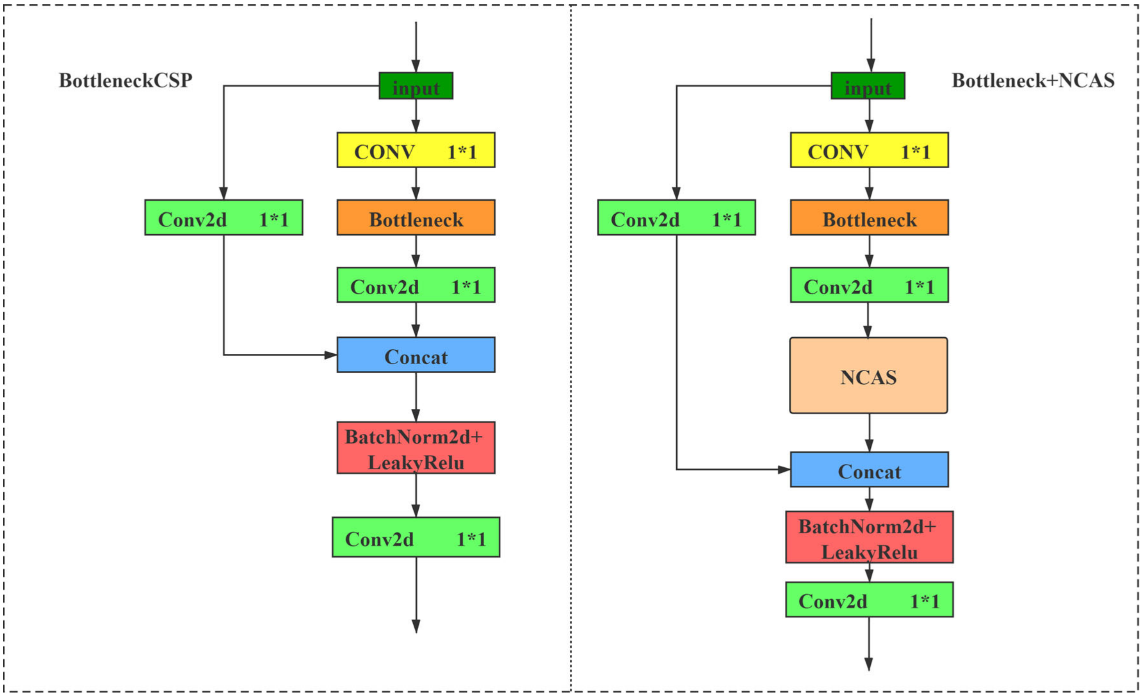


Fig. 5 Schematic diagram of the structure of adding NCAS to the residual structure of YOLOv5s

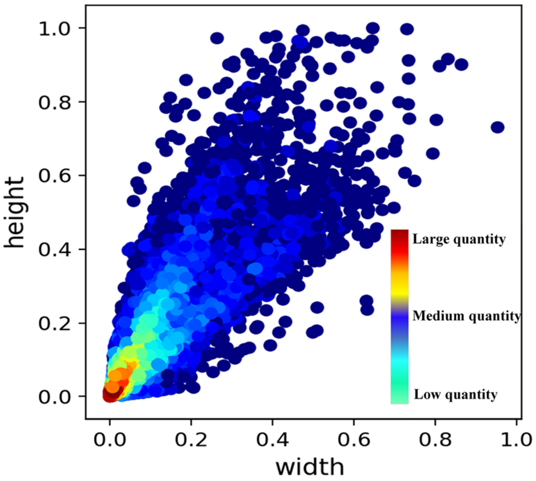


Fig. 6 We normalize and analyze the face labels in the dataset. Corresponding quantity of different colors in the figure are light green, green, light blue, blue, yellow, orange, and red from the least to the most. From the figure, it can be seen that the ratio of the faces that take up more pixels of the image is not 1:1, i.e., the value of height/width is not 1, which means that such faces are not square. The smaller the number of pixels in the image, the closer the ratio of the face is to a square

specific target and is named the receptive field matching module (Fig. 7).

For our method, we reduced the number of channels of the input RFM layer to a quarter of the original and divided them into four branches. After passing through a one-dimensional convolution, the first two branches used 3 * 5 and 5 * 3 ratios. The convolution kernel provided rectangular receptive fields, and then, a convolution kernel layer of 1 * 1 was used for the normalization operation and connected to form a rectangular module. The other two branches, also after 1 * 1 convolution, used ratios of 7 * 7 and 9 * 9 for maximum pooling, and then a convolution kernel layer of 1 * 1 for the normalization operation and connected to form a square module. Then, a “Concat” operation was performed on the rectangular and square modules to connect the two modules in the form of vectors with the corresponding convolution kernels. Finally, the original input feature map was added to make the network learning more consistent with the features of human faces. The structure of the improved RFM module used in this study is shown in Fig. 8.

For the choice of ratios of maxpooling, experimental analysis is also carried out in this paper in the SPP of the benchmark network, and the experimental results are as follows.

From the experimental results in Table 1, we can see that a square receptive field of (9,7,5) is more suitable for the face detection task, while the rectangular receptive field already has a size of 5. In order to adapt the RFM to more sizes of faces, the final maxpooling size was set to (7,9).

3.3 Low-level feature fusion module

The main modules of the neural network structure are convolution, pooling, and activation. This is a standard

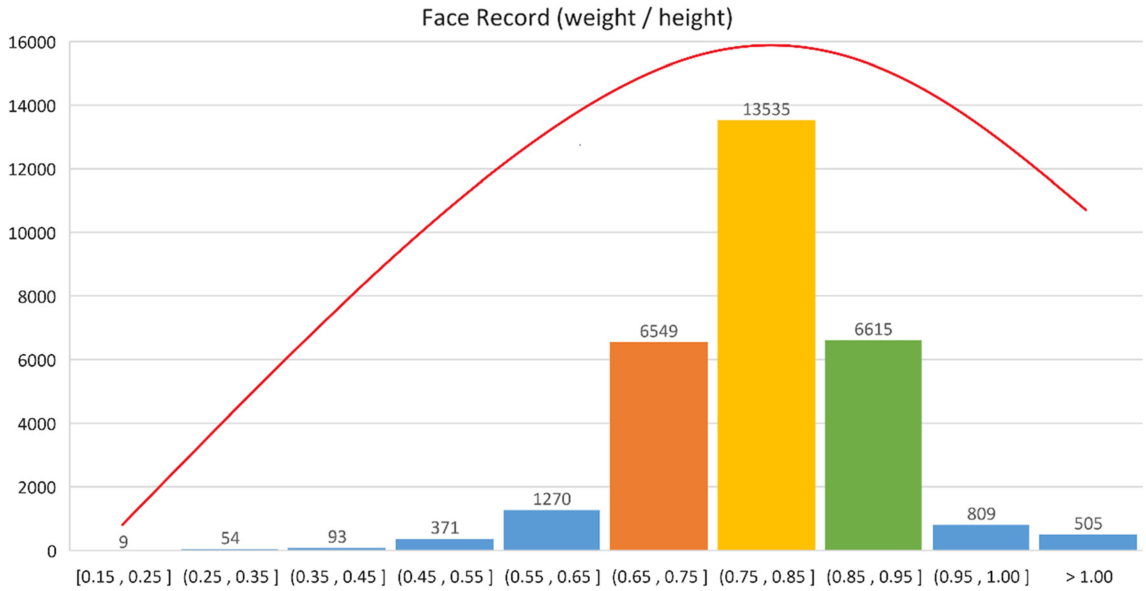
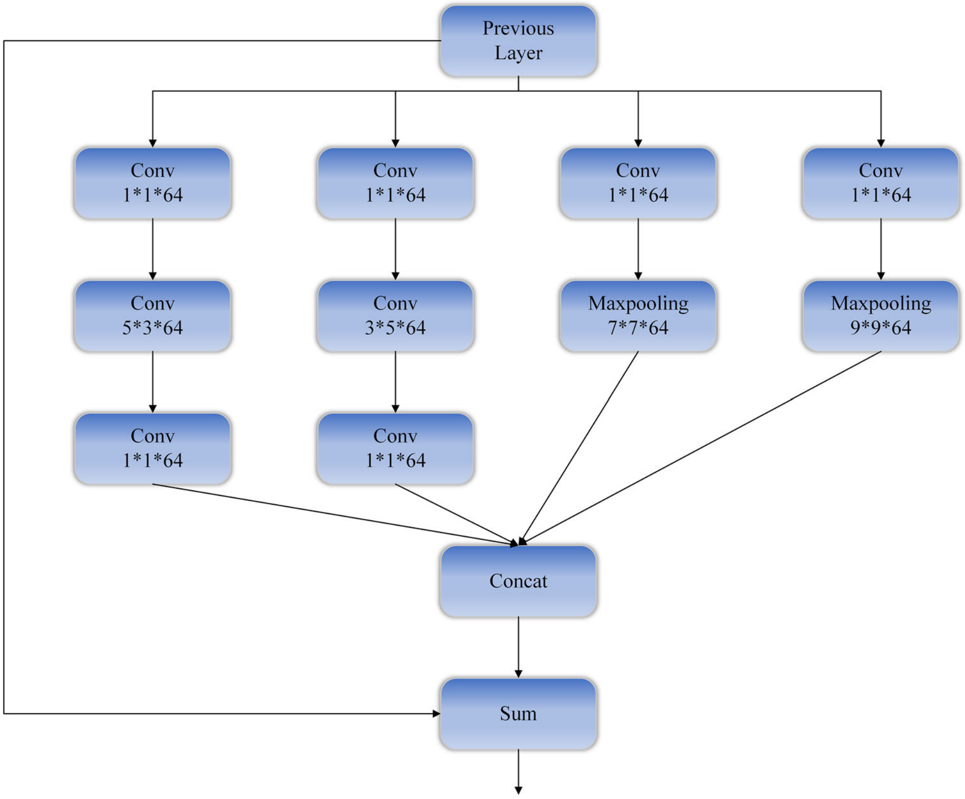


Fig. 7 We analyzed and counted the labels in the dataset. A total of 29,810 human faces were counted and plotted. It can be seen that there are not many faces with a horizontal and vertical ratio of 1:1. Among them, 89.6% of the face ratios are concentrated in the 0.65, 0.95 range, and the proportion of faces in the 0.95, 1.00 range only accounts for 2.9% of all statistical faces. Moreover, the above figure roughly conforms to Gaussian distribution. The mean value of this statistical population data is 0.801, the standard deviation is 0.113, and the variance is 0.012. Therefore, 68.3% of the data is distributed within the range of 0.688, 0.914, which is consistent with the above analysis. It shows that the aspect ratio of face data is mostly not square

Fig. 8 Schematic diagram of the structure of RFM



nonlinear transformation module, where the lower layer of the network learns the edge information of the image, the middle layer of the network learns simple shape information, and the deeper layer of the network learns the shape of the target. The deeper layers of the network can also learn more complex features. If the deep neural

Table 1 Effectiveness of different sizes of maxpooling on the AP performance

Method	Size	Easy (%)	Medium (%)	Hard (%)
The ratios of maxpooling	(13,9,7)	89.11	87.29	72.55
	(13,9,5)	90.41	88.20	72.63
	(9,7,5)	**90.67**	**88.54**	**72.98**
	(7,5,3)	90.12	88.24	71.97
	(3,5,7)	89.57	87.48	71.91

Bold values indicate the best results

network had only one layer, it would mean that the transformations to be learned would be very complex, and difficult to accomplish. However, a more complex the neural network, stacked with modules such as convolution and pooling layers, will result in network degradation as mentioned in Sect. 3.1; a shallow network is able to achieve better results than a deep network. To address this situation, in addition to the residual modules mentioned above, it is possible to combine the features of the shallow network with the features of the deep network, to allow the deep network to learn the feature information in the shallow network.

In our benchmark network YOLOv5, the above-mentioned approach was also adopted, which combined the feature information in the backbone with the feature information in the feature extraction layer to prevent the emergence of network degradation problems. To make the feature information of the low-level network better help the overall network to learn, we propose an LFF module: it passes the more recognizable features in the backbone through the module after enhancement, and it is connected to the feature extraction module, which can merge low-level and deep-level features, enhance feature mapping, and achieve the purpose of making the features learned by the network more recognizable and robust.

Inspired by the feature enhancement module (FEM) in DSFD[6], we designed an LFF module to solve the above problem. The FEM module adopts the form of a dilated convolution for feature enhancement, but the faces studied in this work was small and dense. Dilated convolution has a high probability of not extracting faces completely during each convolution operation. Furthermore, the characteristic information could make the enhancement effect worse. Therefore, we used two-dimensional convolutional kernels instead of dilated convolution in the design of our LFF module. Because faces are relatively small in images, large convolutional kernels cannot extract features well; therefore, we used a 3 × 3 convolution in the first few layers of each ladder structure to scan and extract the feature map, and a 1 × 1 convolutional kernel in the last layer to normalize the feature map. This allowed more robust features to be passed to the deeper feature extraction layers and solved the network degradation problem while further improving face detection accuracy. The structure of the LFF module is shown in Fig. 9.

Specifically, in this module, we divided the feature map into three parts, and then divided the three sub-networks containing different numbers of convolutional layers with different convolution sizes, and finally connected them. To enhance the original image features $IF_{(i,j,l)}$, LFF uses different dimensional information, including the original image features $IF_{(i,j,l)}$ of the upper layer and non-local image features $LF_{(i-\delta,j-\delta,l)}$, $LF_{(i-\delta,j,l)}$,..., $LF_{(i,j+\delta,l)}$, $LF_{(i+\delta,j+\delta,l)}$ of the current layer. In particular, the enhanced image features $EF_{(i,j,l)}$ can be defined mathematically as:

$$EF_{(i,j,l)} = f_{concat}\left(f_{dilation}\left(\sum LF_{(i,j,l)}\right)\right) \tag{2}$$

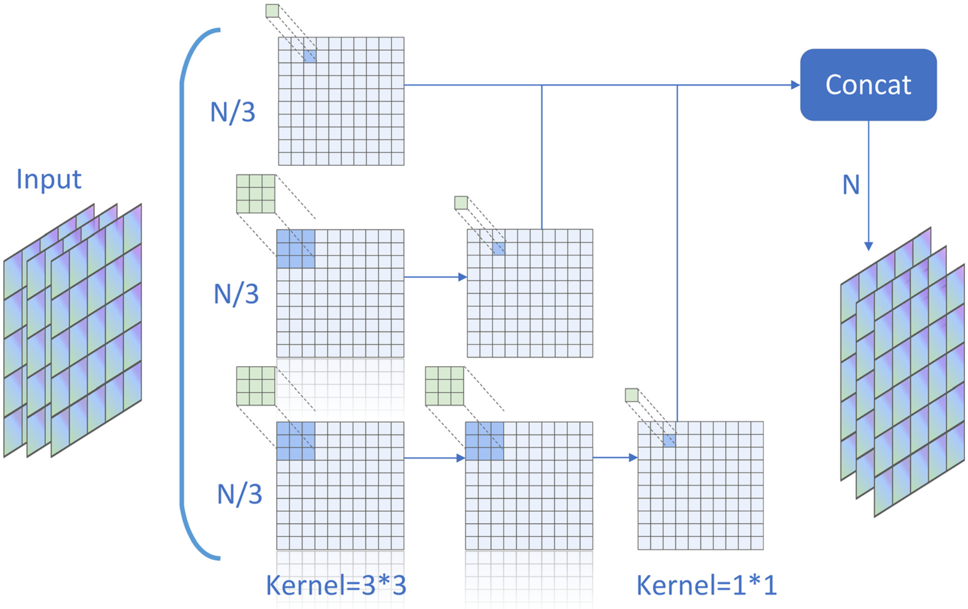
$$LF_{(i,j,l)} = f_{prod}\left(IF_{(i,j,l)}, f_{up}\left(IF_{(i,j,l+1)}\right)\right) \tag{3}$$

where $F_{(i,j,l)}$ represent the coordinates (i, j) in the feature map of the lth layer, f is a set of basic two-dimensional convolution, element generation, or splicing operations.

3.4 Contextual information based on face key points

Computer vision technology is derived from the laws and habits of human observation of objects, and humans also rely on their surroundings to aid their judgment when observing a target object. Therefore, in a target detection task, if the target to be detected is understood as the current passage to be read, then the surrounding environment, other objects can be seen as contextual information. In practice, all objects cannot exist alone, but must have some connection with other objects and the environment around them. Therefore, in target detection, the information between the object to be detected and other objects and the environment around it can usually be used as auxiliary information for neural network model detection, which is called contextual information. Face key points can help the face detector to exclude the more incorrectly detected targets and help it to detect more face targets with less distinctive features.

Fig. 9 LFF structure diagram



Considering that the data in the WIDER FACE dataset are too large, if the context information that is strongly associated with the face is manually labeled, the time cost and labor cost are too high, and it is almost impossible to complete. The face key point detection used in the face alignment task can directly obtain the corresponding position through the existing method, and the accuracy of the face key point detection task has reached extremely high performance. Therefore, on this basis, this section proposes a method for assisting face detection with contextual information based on key points of the face.

Before training, MTCNN was used to detect the key points of the face on the WIDER FACE dataset and add the normalized key coordinates of the face in the annotation file for subsequent training. The effect is shown in Fig. 10.

The loss function used for key points of the face in MTCNN[2] is the L2 loss function. The L2 paradigm is optimized based on the L1 paradigm loss function, but the L2 loss function is not sensitive to small errors, so it is for people who overlap and occlude. The face target cannot fit its key point position well. In this regard, this section introduces the weighted fusion of Wing Loss and the original loss function. The formula of Wing Loss [34] is as follows:

$$L_{wing}(x) = \begin{cases} \omega \ln(1 + \frac{|x|}{\epsilon}) & |x| < \omega \\ |x| - C & otherwise \end{cases} \qquad (4)$$

where ω is a positive number, and its function is to limit the nonlinear loss value between $[-\omega, \omega]$. ϵ is a parameter that constrains the overall curvature of the nonlinear interval. According to formula 4, the constant $C = \omega - \omega \ln(1 + \frac{x}{\epsilon})$ is calculated. Its function is to smoothly connect the area of linear and nonlinear segmented points, making the overall loss function smoother.

4 Experiments

We first analyzed the proposed modules and method in detail to verify the effectiveness of our contributions. Then, we evaluated the final model on common face detection benchmark datasets, including FDDB (Jain and Learner-Miller 2010), and WIDER FACE [7].

4.1 Dataset

Currently, the most popular face detection datasets are FDDB and WIDER FACE[7]. FDDB was released in 2010 and reached saturation under the current development status of deep neural networks, but it is still the dataset used by some face detection methods to verify their contributions. Currently, the WIDER FACE dataset is the most frequently used public dataset for face detection and contains the largest amount of data. It contains 32,203 pictures with 393,703 annotated faces. WIDER FACE is divided into 61 subsets, each of which contains three levels of detection difficulty: "Hard," "Medium," and "Easy". Because the scale, posture, occlusion, expression, lighting, and events of the pictures vary widely, this dataset is complex and represent reality closely; the dataset is divided into a training set, validation set, and test set at 50, 10, and 40%, respectively. After further analysis, we found that the data in "Hard" contains pictures that are defined as "Medium" and "Easy", which indicates that the performance on "Hard" would better reflect the effectiveness of

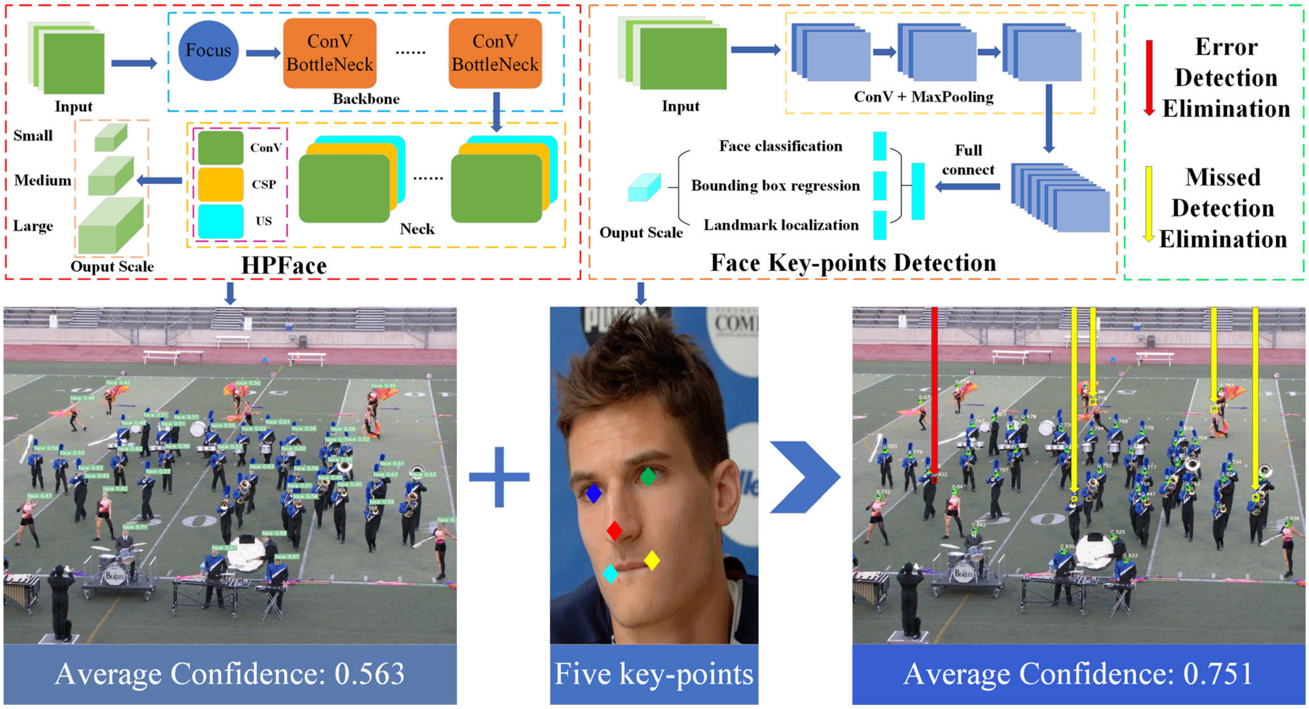


Fig. 10 Schematic diagram of face detection after adding face key points. As can be seen from the diagram, the inclusion of key points allows the face detector to detect many missed faces and also to eliminate false detections. Moreover, the average confidence has increased from 0.563 to 0.751, which can also illustrate the importance of the method in this chapter

different methods. Therefore, we chose FDDB and WIDER FACR as the experimental datasets to verify the method proposed in this paper.

4.2 Implementation details

4.2.1 Overall network design

We used the lightest version of the YOLOv5 series, YOLOv5s, as our benchmark network and added the NCAS module to the BottleneckCSP structure; the SPP module in the original network was replaced by our RFM module; the LFF module is used in this paper to connect the backbone layer to the feature extraction layer. We added both designed modules and the original modules to the benchmark network, in turn, for comparison in each case to demonstrate that our improvements are effective. We also performed ablation experiments on all modules, again demonstrating that our designed modules are suitable and effective for face detection.

4.2.2 Work before verifying the model

For WIDER FACE, we removed the labels from some missing images and images with missing labels and converted them into the data format required by YOLOv5 for training. Before training, we used the K-means [35] clustering algorithm to perform cluster analysis on the face labels in the dataset and finally obtained nine anchor boxes with the most suitable size for the dataset, which were given as follows: [23,26], [137,30], [96,48], [57,109], [172,47], [139,77], [117,126], [176,126], [261,124].

4.2.3 Losses

We conducted experiments based on the benchmark network to compare three loss functions, CIOU, DIOU [36], and GIOU [37]. Based on the experimental results, we chose CIOU as the loss function for our network.

4.2.4 Experimental set

To ensure that each set of experiments was fairly comparable, we used the same parameter settings for all experiments and changed only the components. All models were trained on the WIDER FACE training set and evaluated on the validation set. We conducted experiments on a single NVIDIA GeForce GTX 2080 Super and used the SGD optimizer to train the model. Our initial parameters were as follows: the initial learning rate was 0.01, the momentum was 0.937, the batch was set to 64, the resized input image was 416 × 416, and each experimental round was set to 750 epochs.

4.3 Analysis on HPFace

In this subsection, we discuss the extensive experimental and ablation study performed on the WIDER FACE dataset as a way to evaluate the effectiveness of our proposed modules, thus demonstrating our contributions. The methods include the NCAS module, RFM module, LFF module, and contextual information base on key points of face. For a fair comparison, we used the same parameter settings in all experiments, except for specific changes to the components. All models were trained using the WIDER FACE training set and evaluated on the validation set, ultimately validating the performance of all the modules proposed in this study with official evaluation metrics.

4.3.1 NCAS

First, the NCAS module proposed in this paper enables the network to learn more expected features, enhance the detection performance, and solve the problem of disorder in extracting features from the residual structure, while not affecting the detection speed. To verify the effectiveness of this module, we added the best ECANet of the current channel attention mechanism to the BottleneckCSP structure of the benchmark network and compare it experimentally with NCAS. From Table 2, it can be seen that both methods improved the detection accuracy, but NCAS improved the accuracy more and affected the detection rate less than ECANet. As a result, NCAS improved the detection accuracy by 0.9, 1.0, and 3.5% on the three subsets, respectively, and the FPS only decreased by 1.

4.3.2 RFM

Second, the RFM module proposed in this paper uses different proportional sizes of perception for face detection tasks, which makes the receptive field of the neural network fit the face better and thus improves the accuracy. We conducted comparison experiments using the RFM module designed in this study to replace the SPP module in the benchmark neural network structure. Since we adopted the idea from SRN, we also experimented with RFE, the perceptual field module in SRN. Table 2 shows that our method applied to the benchmark network performs better than RFE, with accuracy rates improving from 90.40%, 88.20%, and 72.60% to 90.90%, 88.86%, and 75.90%, respectively, compared to the baseline network (Table 3).

4.3.3 LFF

Third, the LFF module proposed in this paper makes the final facial features recognized by the network more discriminative and robust by fusing the features first extracted in the feature extraction layer of the network with the deeper features after enhancement and reduces the impact of network degradation on the accuracy rate. We compared the performance of the proposed LFF module with that of the FEM module during experiments. Table 4 shows that both modules helped to improve the performance of the network, but the proposed LFF was more effective, resulting in an increase in accuracy from 90.40%, 88.20%, and 72.60% to 90.30%, 88.80%, and 75.70%, respectively.

4.3.4 Contextual information based on key points

Fourth, the context information based on the key points of the face proposed in this paper enables the network to associate the key points with the face features during the training phase. Therefore, the trained model can use the key points of the face in the image to guess the position of the face. Location, thus, greatly improves model performance. In this paper, a larger scale is added to the output scale of the network structure to better fit the key point positions, among which key point-4 represents the output results of 4 scales. Table 5 shows the experimental results of this article for contextual information. After adding the key points of the face, the overall accuracy rate has been improved, and the network structure of 4 scales is also not small compared with the three scales. The accuracy of

Table 2 Effectiveness of the NCAS module on the AP and FPS performance

Component	FPS	Easy (%)	Medium (%)	Hard (%)	Param (M)
YOLOv5s(baseline)	90	90.40	88.20	72.60	7.091
YOLOv5s+SENet [27]	80	90.42	88.32	74.12	7.695
YOLOv5s+ECANet [28]	84	91.20	89.10	75.50	7.322
YOLOv5s+NCAS(Tanh)	89	91.30	88.93	72.92	7.096
YOLOv5s+NCAS(Mish)	88	91.22	88.89	75.32	7.094
YOLOv5s+NCAS(Relu)	88	91.14	88.87	75.11	7.101
YOLOv5s+NCAS(sigmoid)	89	**91.3**	**89.2**	**76.1**	7.097

Bold values indicate the best results

Table 3 Effectiveness of the RFM module on the AP and FPS performance

Component	Easy (%)	Medium (%)	Hard (%)	FPS
YOLOv5s(baseline)	90.41	88.20	72.63	90
SPP(9,7,5)	90.67	88.54	72.98	88
YOLOv5s+RFE(SRN) [23]	90.50	88.00	75.30	83
YOLOv5s+RFM	**90.90**	**88.86**	**75.90**	82

Bold values indicate the best results

Table 4 Effectiveness of the LFF module on the AP and FPS performance

Component	Easy (%)	Medium (%)	Hard (%)	FPS
YOLOv5s(baseline)	90.40	88.20	72.60	90
YOLOv5s+FEM [6]	**90.80**	88.50	75.30	81
YOLOv5s+LFF	90.30	**88.80**	**75.70**	83

Bold values indicate the best results

Table 5 Effectiveness of the Contextual information based on key points on the AP and FPS performance

Component	Easy (%)	Medium (%)	Hard (%)	FPS
YOLOv5s(baseline)	90.40	88.20	72.60	90
Key points-3	92.80	90.90	82.80	89
Key points-4	**93.70**	**92.40**	**83.80**	87

Bold values indicate the best results

network detection increased from 90.40%, 88.20%, and 72.60% to 93.70%, 92.40%, and 83.80%.

After demonstrating the effectiveness of the four methods proposed in this paper, a full ablation experiment was conducted to verify the effectiveness of the modules as a whole. Table 6 presents the final data for this part of the experiment. As can be seen from the table, the modules proposed in this paper are not only effective individually, but the accuracy increased in combining two or three of the modules, and the best results were achieved by combining all four methods in the network. The accuracy of network detection increased from 90.40%, 88.20%, and 72.60% to 94.20%, 92.60%, and 84.20%, respectively, and a detection rate of 68 FPS on the single NVIDIA GeForce GTX 2080 Super was still achieved, meeting real-time requirements.

As mentioned above, our final model reached the SOTA level results on the WIDER FACE dataset. To better prove this point, this study investigated some existing methods from recent years and compared them in terms of detection rate and detection accuracy. The results of the comparison are presented in Table 7.

From the results shown in Table 7, we can see that some methods achieved higher detection accuracies than our method, but with FPS less than 30, which does not meet the real-time requirement. The methods that achieved FPS greater than 30 and were faster than the model proposed in this paper had insufficient detection accuracy. Therefore, it was established that our method reached the SOTA level.

The final model optimized in this paper achieves a face detection speed of 72 FPS, although there are some gaps in accuracy compared to some SOTA methods with larger models. However, the focus of this paper is to investigate a sufficiently good and fast lightweight face detection model that is fast enough for PC/server, but also for mobile applications in future research to achieve the real-time requirements.

Table 6 Effectiveness of ablation experiments in various modules on the AP and FPS performance

Component	HPFace							
NCAS		✓			✓		✓	✓
RFM			✓		✓	✓		✓
LFF				✓		✓	✓	✓
key points		✓	✓	✓	✓	✓	✓	✓
FPS	90	83	84	82	82	80	79	72
Easy(%)	90.41	92.30	93.80	93.69	93.90	92.25	94.00	**94.20**
Medium(%)	88.2	91.00	92.30	92.32	92.40	91.15	91.19	**92.60**
Hard(%)	72.63	83.20	84.02	84.05	94.17	82.33	84.01	**84.20**

Bold values indicate the best results

Table 7 Comparison of the experimental results of the AP and FPS performance

	Method	Easy subset (%)	Medium subset (%)	Hard subset (%)	FPS
Non-real time	MTCNN [2]	84.80	82.50	59.80	25
	FastFace [38]	83.30	79.60	60.30	22
	Faceness [39]	71.60	60.40	31.50	20
	RetinaFace [40]	95.00	91.90	77.90	14
	Face R-CNN [41]	93.70	92.10	83.10	26
	HAMBox [42]	95.27	93.76	82.75	17
	HR-ResNet101 [43]	92.50	91.00	80.60	2
	DSFD [6]	94.30	91.50	71.40	5
Real time	YOLO-face [44]	89.90	87.20	69.30	38
	APNS+RNet48 [45]	88.30	87.90	76.10	41
	Faceboxes [21]	84.00	76.60	39.50	45
	HPER [46]	88.30	86.80	77.40	44
	JFDA [47]	85.10	82.00	60.70	79
	LightFace [48]	96.30	87.80	52.80	81
	SCRFD [49]	93.78	92.16	77.87	57
	YOLOv5Face [50]	93.61	91.54	80.53	82
	Img2pose [51]	90.00	89.10	83.90	–
	RNNPool [52]	92.00	89.00	70.00	76
Ours	HPFace(without key points)	93.00	90.40	76.50	76
	HPFace(with key points)	94.20	92.60	84.20	72

($>$ 30 FPS for real time)

To compare the method in this study more intuitively with popular face detection methods, they were tested on the validation and test sets of the WIDER FACE dataset. Figure 11 shows the final results.

Similarly, the FDDB dataset was used to test the final obtained model and, after comparing it with many of today's popular algorithms, we can see in Fig. 12 that our method achieves the best results on the FDDB dataset.

4.4 Experimental results display

We show the results of face detection by the final model in various scenarios.

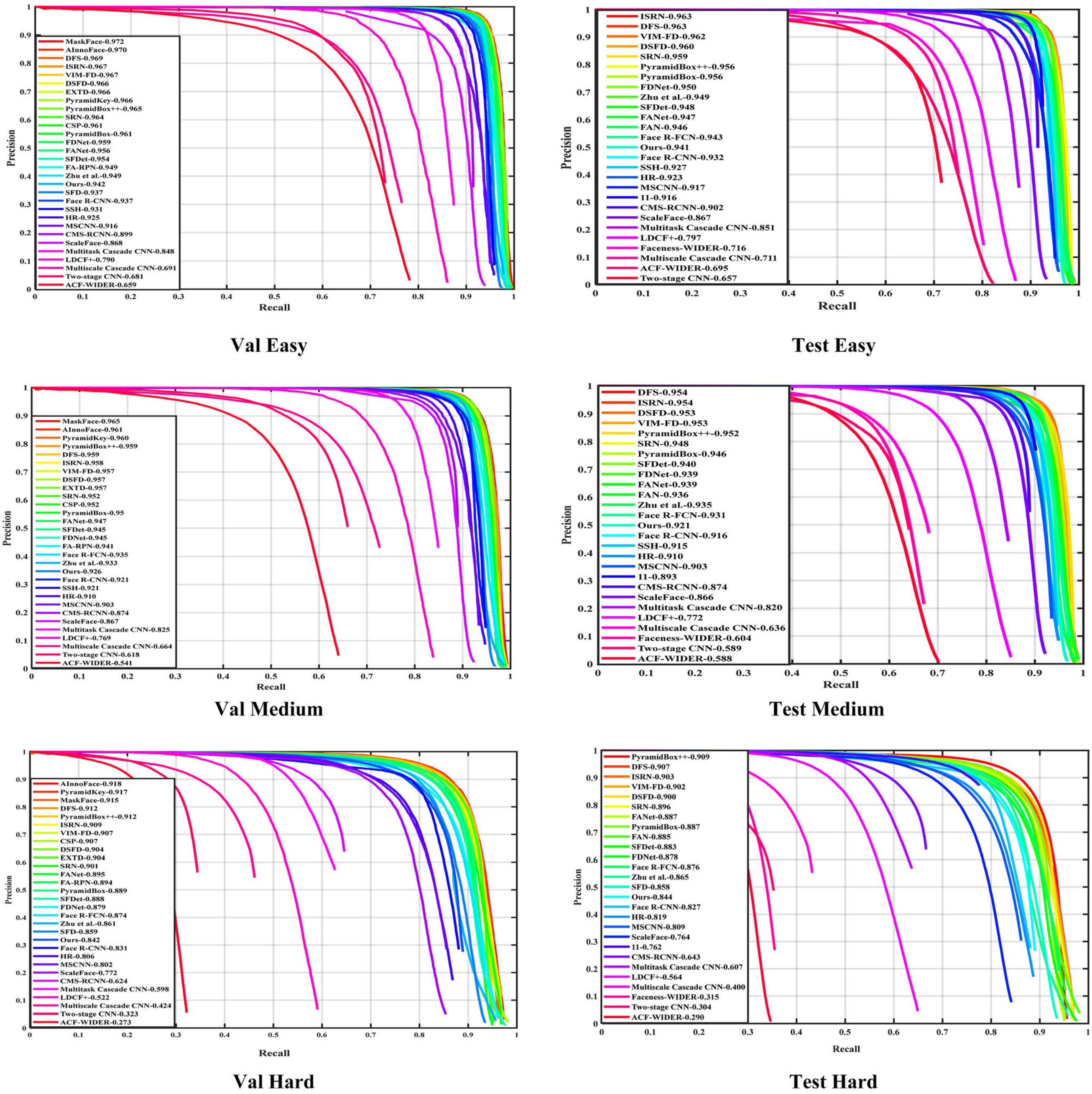
- In the Light section, we show images with low contrast, overexposure, and black-and-white.
- In the Angle section, the faces are rotated at a certain angle and not all the features of the face are shown. On some faces only half of the face is visible.
- In the Blur section, the background of the image contain faces, and some of them are blurred due to focus problems or low resolution.
- In the Occlude section, the faces are partially obscured due to external factors, in which case many important features of the face are not recognized.
- Finally, in the Small section, or the dense face section, all the faces are small in pixel size and close to each other. This is arguably the biggest challenge for the face detection task, as some faces are also obscured.

As can be seen in Fig. 13, the final model proposed in this paper performs well in detecting faces in these challenging scenarios. The introduction of the three modules designed in this study resulted in a significant improvement in the accuracy of the overall network for face detection in various scenarios, and our network also has the advantage of being real-time compared to other methods.

5 Discussion

In this paper, we propose three modules to optimize the face detection task, namely NCAS, RFM, and LFF, for channel attention, perceptual field matching, and feature enhancement, respectively. Our proposed method delivered good results while meeting the real-time requirement. Our final model achieved 58 FPS on a single NVIDIA GeForce GTX 2080 Super graphics card, far exceeding the real-time requirement of 30 FPS.

This study further demonstrated the feasibility of a range of methods, such as the channel attention mechanism and feature enhancement, for optimizing face detection models. Next, we discuss the inspiration derived from other work during this study.



Val Easy

Test Easy

Val Medium

Test Medium

Val Hard

Test Hard

Fig. 11 Precision-recall curves on WIDER FACE validation and testing subsets

5.1 Optimization of neural network structures

During the course of the work on this study, we tried to analyze and experiment with a number of existing methods and found redundant structures in their practical applications. As mentioned previously, the ECANet is popular. Numerous studies demonstrated the effectiveness of ECANet in improving accuracy. However, when we removed the activation and convolutional structures, we found that the new model was still effective in improving the accuracy, and that the speed of detection was improved by reducing the number of parameters in the overall network through the reduction in unnecessary operations. Therefore, we consider trying to simplify our own designed method or other methods with certain rules to obtain better results in the future.

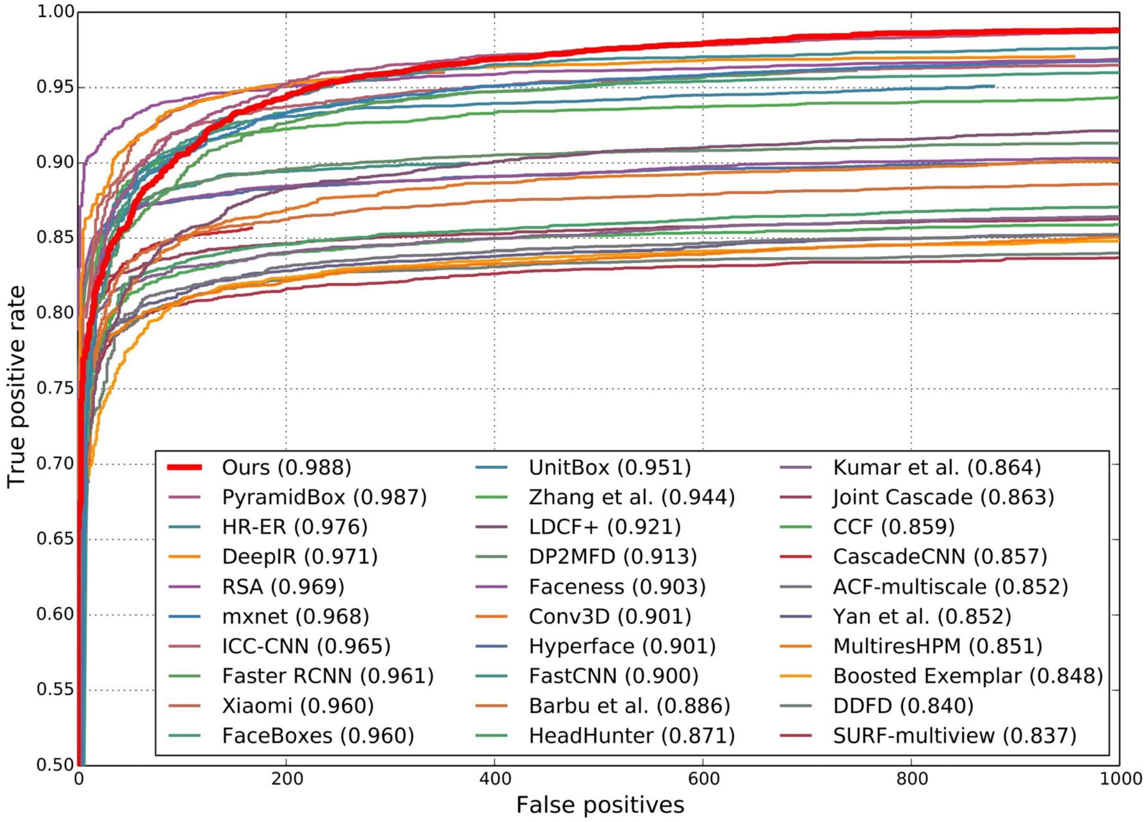


Fig. 12 Precision-recall curves on FDDB

5.2 Plug-and-play module

With the rapid development of deep learning technologies, the concept of “plug and play” is starting to take hold. As with the three modules presented in this paper, the code is simple to implement, and the number of channels of input and output of the module can be adjusted simply for the model to be reused in other models. Such a design can be very convenient for others to reproduce or apply.

5.3 Real time

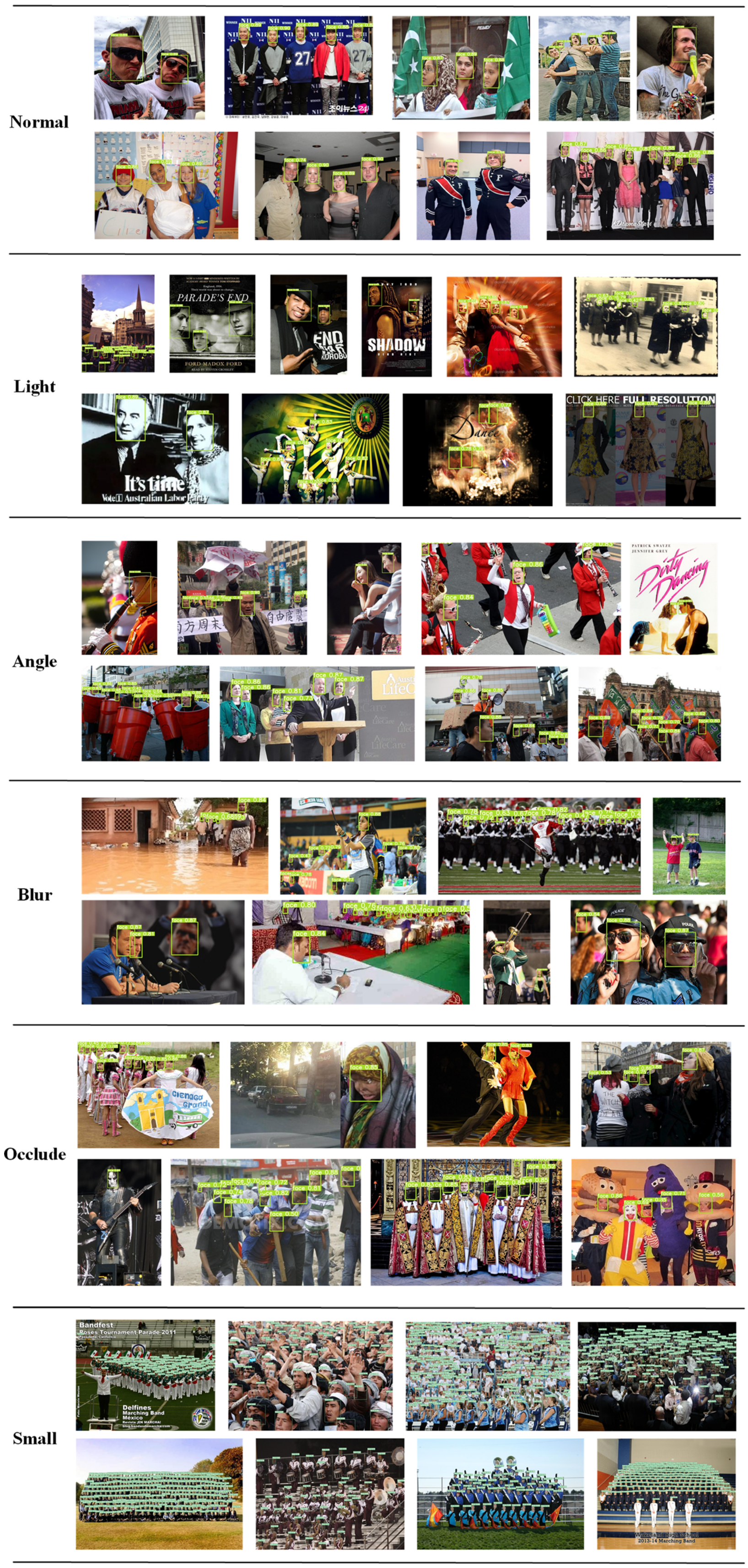
The work in this study was designed based on the criterion of real time and, therefore, has a good application in current “industrialized algorithms” and provides a good basis as a pre-task for face recognition, face alignment, and other tasks. Many current methods focus more on accuracy as a metric and ignore real time, which is detrimental to the application of algorithms in real-life applications. In recent years, the application of algorithms such as face detection to mobile devices has also become a popular research direction, thus requiring lighter weight models and faster detection speeds. This poses new challenges for tasks such as detection and recognition, where accuracy is now the main evaluation criterion, but it is also important to balance accuracy and detection speed. This is the basis of the research presented in this paper.

However, there are still areas for improvement: In designing these methods, we modified the size of the convolution kernel for a facial target, i.e., we made the model more suitable for the face; therefore, applying this method to the detection other types of targets may be less effective. In the future, we hope to develop new ideas and work in the direction of “adaptive network restructuring,” so that methods originally applied to single target detection can be better applied to traditional target detection tasks.

6 Conclusion

In this paper, for the task of face detection, we presented three new modules—the NCAS module, the RFM module, and the LFF module—and applied these three modules to the lightweight version of YOLOv5s. NCAS was used with the residual module in the benchmark network to enable the network to focus more on facial features. RFM replaced the SPP module in the original network, and the single square receptive field module was replaced with a

Fig. 13 Experimental results display diagram



combination of rectangular and square receptive field modules to make the receptive field of the network more compatible with the target of the face and enabled the network to extract complete facial features. The LFF module combined the backbone with the feature extraction module, fusing the shallow features in the backbone with the feature extraction layer after enhancement, and passing the more robust features to the deeper feature extraction layer; this solved the network degradation problem while further improving the face detection accuracy. The context information-assisted detection method based on the key points of the face provides a new reference target for the detection model to detect faces with inconspicuous features, which greatly improves the performance. Our method was extensively tested on the FDDB and WIDER FACE datasets, and the results showed that our method met the SOTA criteria and the criteria for real-time detection compared to other common methods.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61972097 and U21A20472, in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the Natural Science Foundation of Fujian Province under Grant 2021J01612 and 2020J01494, in part by the Major Science and Technology Project of Fujian Province under Grant 2021HZ022007, in part by the Industry-Academy Cooperation Project of Fujian Province under Grant 2018H6010, in part by the Fujian Collaborative Innovation Center for Big Data Application in Governments, and in part by the Fujian Engineering Research Center of Big Data Analysis and Processing.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Ranjan R, Patel VM, Chellappa R (2019) Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intell 41(1):121–135
2. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
3. Jian Y, Lei L, Qian J, Ying T, Zhang F, Yong X (2016) Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Trans Pattern Anal Mach Intell 39(1):156–171
4. Sun X, Wu P, Hoi S (2018) Face detection using deep learning: an improved faster RCNN approach. Neurocomputing 299:42–50
5. Tao Q-Q, Zhan S, Li X-H, Kurihara T (2016) Robust face detection using local CNN and SVM based on kernel combination. Neurocomputing 211:98–105
6. Li J, Wang Y, Wang C, Tai Y, Qian J, Yang J, Wang C, Li J, Huang F (2020) Dsfd: dual shot face detector. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
7. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: a face detection benchmark. In: IEEE conference on computer vision and pattern recognition, pp 5525–5533
8. Liu Y, Lasang P, Pranata S, Shen S, Zhang W (2019) Driver pose estimation using recurrent lightweight network and virtual data augmented transfer learning. IEEE Trans Intell Transp Syst 20(10):3818–3831
9. Xu ZF, Jia RS, Sun HM, Liu QM, Cui Z (2020) Light-yolov3: fast method for detecting green mangoes in complex scenes using picking robots. Appl Intell 50:4670–4687
10. Tack A, Preim B, Zachow S (2021) Fully automated assessment of knee alignment from full-leg x-rays employing a "yolov4 and resnet landmark regression algorithm" (yarla): data from the osteoarthritis initiative. Comput Methods Prog Biomed 205:106080
11. Li S, Gu X, Xu X, Xu D, Dong Q (2021) Detection of concealed cracks from ground penetrating radar images based on deep learning algorithm. Constr Build Mater 273:121949
12. Pal SK, Pramanik A, Maiti J, Mitra P (2020) Deep learning in multi-object detection and tracking: state of the art. Appl Intell 1–30
13. Xu C, Yang J, Lai H, Gao J, Shen L, Yan S (2019) Up-cnn: unpooling augmented convolutional neural network. Pattern Recognit Lett 119:34–40
14. Lu E, Hu X (2021) Image super-resolution via channel attention and spatial attention. Appl Intell
15. Akbarinia A, Parraga CA (2018) Colour constancy beyond the classical receptive field. IEEE Transactions Pattern Anal Mach Intell 40(9):2081–2094
16. Guo H, Li Y, Li Y, Xiao L, Li J (2016) Bpso-adaboost-knn ensemble learning algorithm for multi-class imbalanced data classification. Eng Appl Artif Intell 49:176–193
17. Mishkin D, Sergievskiy N, Matas J (2017) Systematic evaluation of convolution neural network advances on the imagenet. Comput Vis Image Underst 161:11–19
18. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42:2999–3007
19. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149
20. Wang L, Xiang Y, Metaxas DN (2017) A coupled encoder-decoder network for joint face detection and landmark localization. In: IEEE international conference on automatic face and gesture recognition
21. Zhang S, Wang X, Lei Z, Li SZ (2019) Faceboxes: a cpu real-time and accurate unconstrained face detector. Neurocomputing 364:297–309
22. Song G, Liu Y, Jiang M, Wang Y, Yan J, Leng B (2018) Beyond trade-off: accelerate fcn-based face detector with higher accuracy. In: 2018 IEEE/CVF conference on computer vision and pattern recognition
23. Ke W, Chen J, Jiao J, Zhao G, Ye Q (2017) Srn: side-output residual network for object symmetry detection in the wild. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR)
24. Cai Z, Vasconcelos N (2018) Cascade r-cnn: delving into high quality object detection. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
25. Lu T, Yu F, Xue C, Han B (2020) Identification, classification, and quantification of three physical mechanisms in oil-in-water emulsions using AlexNet with transfer learning. J Food Eng 288:110220
26. Chi C, Zhang S, Xing J, Lei Z, Zou X (2019) Selective refinement network for high performance face detection. In: Proceedings of

the AAAI conference on artificial intelligence, vol 33, pp 8231–8238

27. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023
28. Wang Q, Wu B, Zhu P, Li P, Hu Q (2020) Eca-net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
29. Zhang Z, Wang X, Jung C (2019) Dcsr: dilated convolutions for single image super-resolution. IEEE Trans Image Process 28(4):1625–1635
30. Luvizon DC, Tabia H, Picard D (2019) Human pose regression by combining indirect part detection and contextual information. Comput Graph 85:15–22
31. Cao Y, Wu Z, Shen C (2018) Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Trans Circuits Syst Video Technol 28(11):3174–3182
32. Lu Z, Jiang X, Kot CC (2018) Deep coupled resnet for low-resolution face recognition. IEEE Signal Process Lett 25:526–530
33. Yao Q, Wang R, Fan X, Liu J, Li Y (2020) Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. Inf Fusion 53:174–182
34. Feng Z-H, Kittler J, Awais M, Huber P, Wu X-J (2018) Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2235–2245
35. Al-Yaseen WL, Othman ZA, Nazri M (2017) Multi-level hybrid support vector machine and extreme learning machine based on modified k-means for intrusion detection system. Expert Syst Appl 67:296–303
36. Zheng Z, Wang P, Liu W, Li J, Ren D (2020) Distance-IoU loss: faster and better learning for bounding box regression. In: AAAI conference on artificial intelligence
37. Rezatofighi H, Tsoi N, Gwak JY, Sadeghian A, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
38. Zhang H, Wang X, Zhu J, Kuo C (2019) Fast face detection on mobile devices by leveraging global and local facial characteristics. Signal Process Image Commun 78:1–8
39. Yang S, Luo P, Loy CC, Tang X (2018) Faceness-net: face detection through deep facial part responses. IEEE Trans Pattern Anal Mach Intell 40(8):1845–1859
40. Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S (2020) Retinaface: single-shot multi-level face localisation in the wild. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
41. Jiang H, Learned-Miller E (2017) Face detection with the faster R-CNN. In: 2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017), pp 650–657. IEEE
42. Liu Y, Tang X, Han J, Liu J, Rui D, Wu X (2020) Hambox: delving into mining high-quality anchors on face detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 13043–13051. IEEE
43. Hu P, Ramanan D (2017) Finding tiny faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 951–959
44. Chen W, Huang H, Peng S, Zhou C, Zhang C (2020) Yolo-face: a real-time face detector. Vis Comput 37:1432–2315
45. Yu B, Tao D (2019) Anchor cascade for efficient face detection. IEEE Trans Image Process 28:2490–2501
46. Putro MD, Kurnianggoro L, Jo K-H (2021) High performance and efficient real-time face detector on central processing unit based on convolutional neural network. IEEE Trans Industr Inf 17(7):4449–4457
47. Boulkenafet Z, Komulainen J, Hadid A (2017) Face spoofing detection using colour texture analysis. IEEE Trans Inf Forensics Secur 11:1818–1830
48. Luo J, Liu J, Lin J, Wang Z (2020) A lightweight face detector by integrating the convolutional neural network with the image pyramid. Pattern Recogn Lett 133:180–187
49. Guo J, Deng J, Lattas A, Zafeiriou S (2021) Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714
50. Qi D, Tan W, Yao Q, Liu J (2021) Yolo5face: why reinventing a face detector. arXiv preprint arXiv:2105.12931
51. Albiero V, Chen X, Yin X, Pang G, Hassner T (2021) img2pose: face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7617–7627
52. Saha O, Kusupati A, Simhadri HV, Varma M, Jain P (2021) Rnnpool: efficient non-linear pooling for ram constrained inference. Adv Neural Inf Process Syst 33:20473–20484

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.