Machine Intelligence Research
www.mi-research.net

19(3), June 2022, 247-256
DOI: 10.1007/s11633-022-1327-2

Dense Face Network: A Dense Face Detector Based on Global Context and Visual Attention Mechanism

Lin Song[1] Jin-Fu Yang[1,2] Qing-Zhen Shang[1] Ming-Ai Li[1]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

[2] Beijing Key Laboratory of Computational Intelligence and Intelligent Systems, Beijing 100124, China

Abstract: Face detection has achieved tremendous strides thanks to convolutional neural networks. However, dense face detection remains an open challenge due to large face scale variation, tiny faces, and serious occlusion. This paper presents a robust, dense face detector using global context and visual attention mechanisms which can significantly improve detection accuracy. Specifically, a global context fusion module with top-down feedback is proposed to improve the ability to identify tiny faces. Moreover, a visual attention mechanism is employed to solve the problem of occlusion. Experimental results on the public face datasets WIDER FACE and FDDB demonstrate the effectiveness of the proposed method.

Keywords: Face detection, global context, attention mechanism, computer vision, deep learning.

Citation: L. Song, J. F. Yang, Q. Z. Shang, M. A. Li. Dense face network: A dense face detector based on global context and visual attention mechanism. *Machine Intelligence Research*, vol.19, no.3, pp.247–256, 2022. http://doi.org/10.1007/s11633-022-1327-2

1 Introduction

Face detection is necessary for many face-related applications, such as facial identity recognition[1] and facial expression recognition[2]. As the fundamental problem of computer vision, face detection has achieved remarkable progress, especially with the recent development of convolutional neural networks. Nevertheless, in certain scenarios, such as airports, scenic spots, and concerts, faces are always dense and severely occluded, which significantly reduces the detection accuracy. Therefore, dense face detection is still a challenging issue.

In recent years, dense face detection has been extensively studied since it is very common in many scenarios. Zhang et al.[3] presented a multi-task network, named multi-task convolutional neural network (MTCNN), to jointly address the detection and landmark alignment and keep overall complexity well under control. Hu and Ramanan[4] made a series of analyses for tiny face detection and proposed to process image pyramids in a scale-invariant manner to capture large-scale variations, which greatly improves the detection performance of tiny faces. Najibi et al.[5] proposed a single stage headless (SSH) face detector, which adds context modules on feature pyramids to enlarge the receptive field and further improve the performance of face detection. RetinaFace[6] was improved on the one-stage target detection network RetinaNet[7]. It added a context module and introduced facial landmark regression loss, which achieves better results in both the FDDB[8] and WIDER FACE[9] datasets.

Although face detection algorithms have made great progress in recent years, the improvement in the recall rate of these cases usually brings the risk of high false positives due to limited context information and occlusion in dense faces. The accuracy of dense face detection is not yet ideal.

To solve the problems mentioned above, in this paper, we propose a novel dense face detector based on global context and visual attention mechanism. In order to enhance the model′s contextual reasoning ability, it is necessary to make full use of context information. However, one major problem of multi-scale representation from feature pyramid is that the higher resolution feature maps only have limited global context information to discriminate faces. High-resolution images usually contain more detailed texture features, while low-resolution images contain more spatial context features. Since detailed texture features and global context information are helpful for tiny face detection, we extract the global context information of low-resolution images, and jointly feedback to the high-resolution images in our network. On the other hand, occlusion will cause false detection and compromise accuracy. Therefore, we take advantage of the attention mechanism that can highlight facial features and reduce areas without faces, to improve the detection abil-

Research Article
Manuscript received August 30, 2021; accepted March 1, 2022; published online March 29, 2022
Recommended by Associate Editor Si-Wei Lyu
Colored figures are available in the online version at https://link.springer.com/journal/11633

ity of our network.

The main contributions of this paper are as follows:

1) We propose a global fusion context module based on top-down feedback to fuse multi-scale information, which is helpful for tiny face detection.

2) An improved visual attention module is designed to further improve detection performance by highlighting significant features.

3) The experimental results show that the proposed model based on global context and visual attention mechanism can achieve better performance in terms of accuracy for dense face detection.

2 Related works

Face detection inherits many achievements from general object detection, which are mainly divided into two categories: two-stage methods (e.g., Faster R-CNN) and one-stage methods (e.g., you only look once (YOLO)). The two-stage detection network adopts a proposal and refinement to achieve higher accuracy, but at the same time reduces the detection efficiency. Compared to the two-stage methods, the one-stage method removes the region proposal network and directly regresses the category probability and position coordinates of the object, making it more efficient. In consideration of detection efficiency, we also adopt a one-stage detection framework. Although remarkable improvements have been achieved in face detection, the challenge of locating faces in dense scenes still remains.

2.1 Context modeling

Context is a crucial factor for multi-scale object detection, especially for small objects. Furthermore, sufficient context information can also help us understand the image. Recently, some researchers have indicated the importance of using context information for face detection, especially for localizing and classifying tiny, occluded and blurred faces. Zhu et al.[10] added skip connections to faster RCNN[11] and deployed feature maps of earlier convolutional layers to detect small faces, which achieves higher accuracy. In SSH[5], Najibi et al. applied simple convolutional layers to produce a larger window effect, achieving more efficient context modeling. Xu et al.[12] also used context modules in feature pyramids to expand the receptive field in PyramidBox. Zhang et al.[13] extracted the context information by adopting large filters for every prediction module in single shot scale-invariant face detector (S3FD). In finding tiny faces (HR)[4], Hu and Ramanan made use of large local context in a scale-variant way, and indicated that context is mostly useful for finding low-resolution faces. Although the existing context modules improve the performance of face detection, these models ignore the connection between low levels and high levels, which is important in dense face detection. Low-level features usually have higher resolution for tiny face detection, but lack global context information, which is also very important for tiny face detection. Therefore, we apply a global context fusion module to improve the detection rate of tiny faces.

2.2 Attention

The attention mechanism is derived from the human brain′s observation of things. When the brain observes things, the focus position of the eyes is only a small and important part. Therefore, attention can be interpreted as a means of learning to use global information to selectively highlight significant features and suppress useless features. Attention mechanism has demonstrated its effectiveness in the field of computer vision. Recently, a number of works have attempted to enhance CNN′s performance in detection tasks by taking advantage of attention.

The attention mechanisms in computer vision are mainly divided into channel attention and spatial attention. Channel attention focuses on the meaningful information of the images. It indicates the importance between features and assigns features according to different tasks. The spatial attention mechanism mainly focuses on the position information of the informative part and selectively aggregates the features through the weighted sum. Jaderberg et al.[14] proposed a differentiable module, called spatial transformer networks (STN), which can be inserted into the convolutional networks. It made neural networks actively perform spatial transformation according to the feature map. Hu et al.[15] introduced a new architectural unit called squeeze-and-excitation net (SENet). The SENet can obtain global information by selectively emphasizing informative features and suppressing useless features, which allows the network to perform feature recalibration. In convolutional block attention module (CBAM)[16], Woo et al. proposed an attention module of feed-forward convolutional neural networks, which can infer attention maps sequentially from two independent dimensions, i.e., channel and spatial, to obtain more useful information. In face attention network[17], Wang et al. introduce an anchor-level attention map to improve masked face detection. CBAM is a common attention mechanism in computer vision, but its direct dot product approach may remove some useful context information. Therefore, we first perform an exponential operation and then dots with the feature maps to retain more context information.

3 Proposed approach

As motivated in Section 1, we develop a dense face detector according to the following principles:

1) Processing faces with different scales in multiple feature layers;

2) Making full use of global context information to detect tiny faces;

3) Highlighting the features of the face area and reducing features in areas without faces.

Fig. 1 provides an overview of our algorithm. We present each part of our model in Sections 3.1–3.3.

3.1 Global context fusion module

Convolutional neural networks usually have different semantic information and spatial resolution in different feature layers. Shallow layers usually have a higher spatial resolution, so it is beneficial to the spatial positioning of small objects. However, the lack of semantic information is not conducive to visual classification. On the other hand, deeper layers contain more semantic information, while the spatial resolution is affected.

Most current models adopt a context structure that ignores the connection between the low and high levels. Nevertheless, due to the background of the crowd in an image being messy, especially in a dense crowd, some of the patterns that resemble faces in these groups can be misclassified, especially at the low scales with the low receptive field. Some faces are too small to be detected, and we need contextual information around the face to help with detection. Low-level features have higher resolution and are usually used to detect tiny faces. On the other hand, these features lack the global contextual information which is necessary to detect tiny faces. Therefore, to resolve this contradiction, we extract global context information from higher scales with a larger receptive field and feedback these features to the low layers. As shown in Fig. 1, the global context fusion (GCF) module receives the output of different scales from the feature pyramids. Each scale branch has a context module, and the context module in each layer is also connected with all previous low-resolution scale branches. GCF fuses contextual information from all scales to make predictions. This multi-scale context information fusion process helps to drive global context information to all scale branches and reduce false detection.

Fig. 2 illustrates the internal architecture of the GCF. Each scale extracts features from the current layer of the feature pyramid and receives features from other scale outputs. The features are up-sampled to the same size and fused through the context module of the current scale. S_n represents the scale. S_0 is used to receive the lowest resolution feature output without feedback from other layers, while $S_{n>0}$ of GCF receives feedback from all other scales. These layers fuse the features from the current scale along with multi-scale inputs from higher branches to improve the performance of face detection.

Inspired by SSH[5], we apply an independent context module on each scale. As shown in Fig. 3, the context module has three parallel structures. This structure can enlarge the receptive field to obtain more contextual information. In order to reduce the number of parameters, we use the 3×3 convolution instead of 5×5 and 7×7. These three output layers pass through a concat layer and a leaky ReLU function and obtain the final context information.

This top-down contextual feedback module can locate the face in the scale pyramid more accurately and further mitigates the issues of tiny faces and occlusion in dense face detection.

3.2 Visual attention network

In order to overcome the interference of useless information, we introduce a visual attention model to our framework. Considering the faces with severe occlusion, most invisible parts may have a negative impact on detection. Therefore, in order to ensure the recall rate without increasing the risk of achieving a higher false pos-

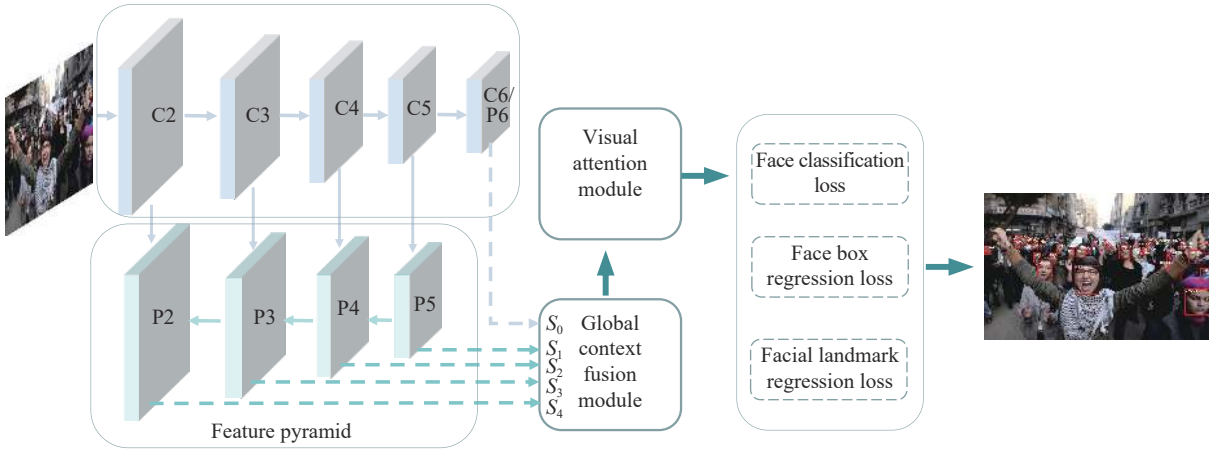

Fig. 1 The proposed framework. Our model is designed based on the feature pyramid with a global context fusion module and visual attention module. Following the attention modules, we calculate a multi-task loss for each anchor. We employ feature pyramid levels from P2 to P6, where P2 to P5 are computed from the output of the corresponding ResNet residual stage (C2 through C5). C2 to C5 is a pre-trained ResNet classification network. The GCF is designed to extract global context features to detect tiny faces, which receives input from each scale of the feature pyramid and performs top-down feedback. $S_n (n = 0, 1, 2, 3, 4)$ represents scale. The visual attention module is used to highlight significant features, which infers attention feature maps from spatial and channel dimensions, respectively.

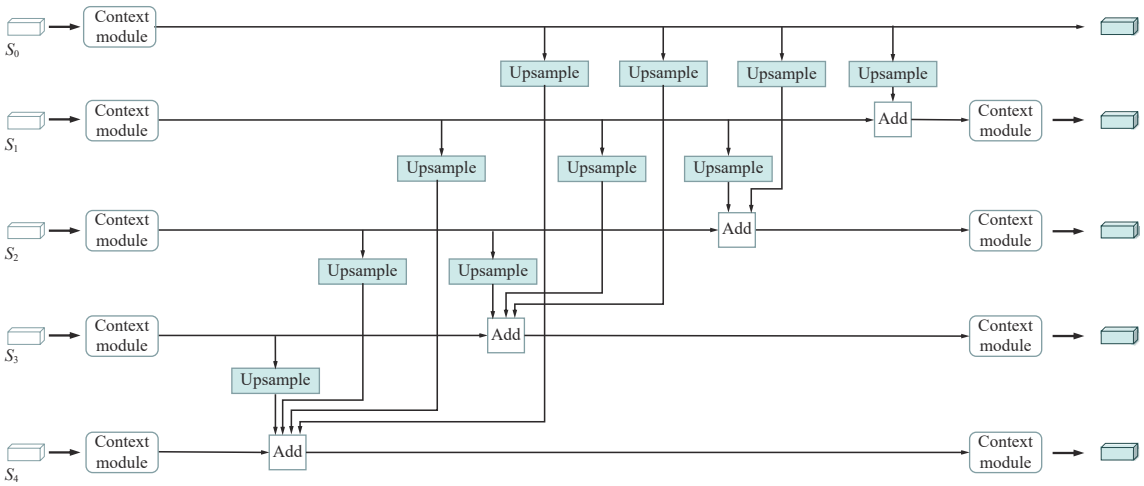

Fig. 2 The architecture of the global context fusion module. $S_n (n = 0, 1, 2, 3, 4)$ represents the scale. Each layer fuses the features from its scale (S_n) along with multi-scale inputs from higher branches to generate context information and the features for the next scale branch.

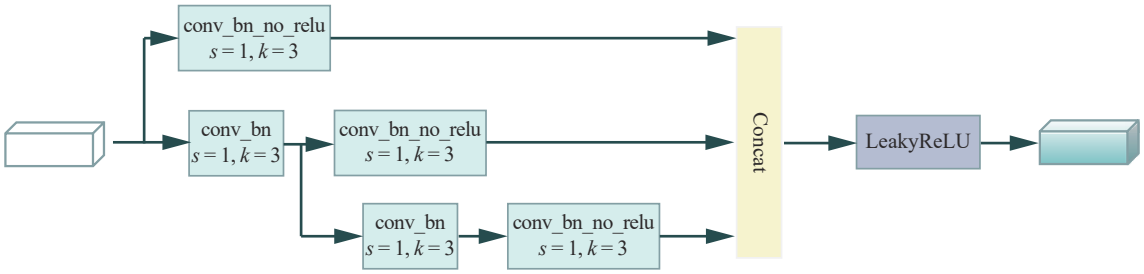

Fig. 3 The internal structure of the context module. The module has three parallel structures, which use 3×3 convolution instead of 5×5 and 7×7 to enlarge the receptive field and reduce the number of parameters simultaneously.

itive rate, we need to guide the model to highlight these visible areas. For this purpose, we use a visual attention, module as shown in Fig. 4. Since we need to highlight the feature of the relevant area, and obtain the location information of the area, inspired by CBAM[16], we apply the visual attention mechanism from two dimensions, channel and spatial.

More specifically, and compared with the original attention model, our visual attention module first performs an exponential operation and then dots with the feature maps. Furthermore, since context information is also important in face detection, our improved attention module can highlight the detection information, and simultaneously retain more context information.

To efficiently compute channel attention, we need to compress the spatial dimension of the input feature map. Average-pooling can achieve the aggregation of spatial information and learn the extent of the target object effectively. At the same time, max-pooling gathers another important clue about distinctive object features to infer finer channel-wise attention. So, we use both average-pooling and max-pooling. As illustrated in Fig. 5, the channel sub-module first utilizes both average-pooling outputs and max-pooling outputs to synthesize spatial information of a feature map, and produces two different spatial context descriptors: F^c_{avg} and F^c_{max}. Then, both descriptors are input into a shared network to generate channel attention map $M_c \in \mathbf{R}^{C\times1\times1}$. The shared network is composed of a multi-layer perceptron (MLP) with one hidden layer. Finally, we use element-wise summation to merge the output features.

Meanwhile, in the spatial sub-module, we use max-pooling and average-pooling to synthesize channel information, producing two 2D maps: $F^s_{avg} \in \mathbf{R}^{1\times H\times W}$ and $F^s_{max} \in \mathbf{R}^{1\times H\times W}$. These two feature maps are then connected and convolved by a convolution layer, generating a 2D spatial attention map.

The improved attention module can further enhance the feature expression of the face area and mitigate the impact of negative information. It can extract the effective information of face more accurately and further improve the model's ability to understand complex scenes. This model can effectively highlight visible parts of occluded faces and suppress cluttered background information to improve the detection rate of occluded faces and reduce the false recognition rate of background information.

3.3 Multi-task loss

We refer to the loss of RetinaFace[6]. In order to improve the efficiency, we only retain the face classification

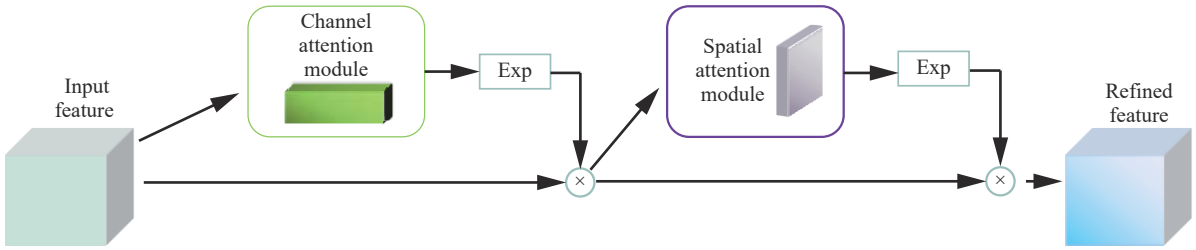

Fig. 4 The overview of the visual attention module. The module has two sequential sub-modules: channel and spatial. The feature map is refined through each module. In order to retain more context information, our attention maps are first feed to an exponential operation and then dot with feature maps.

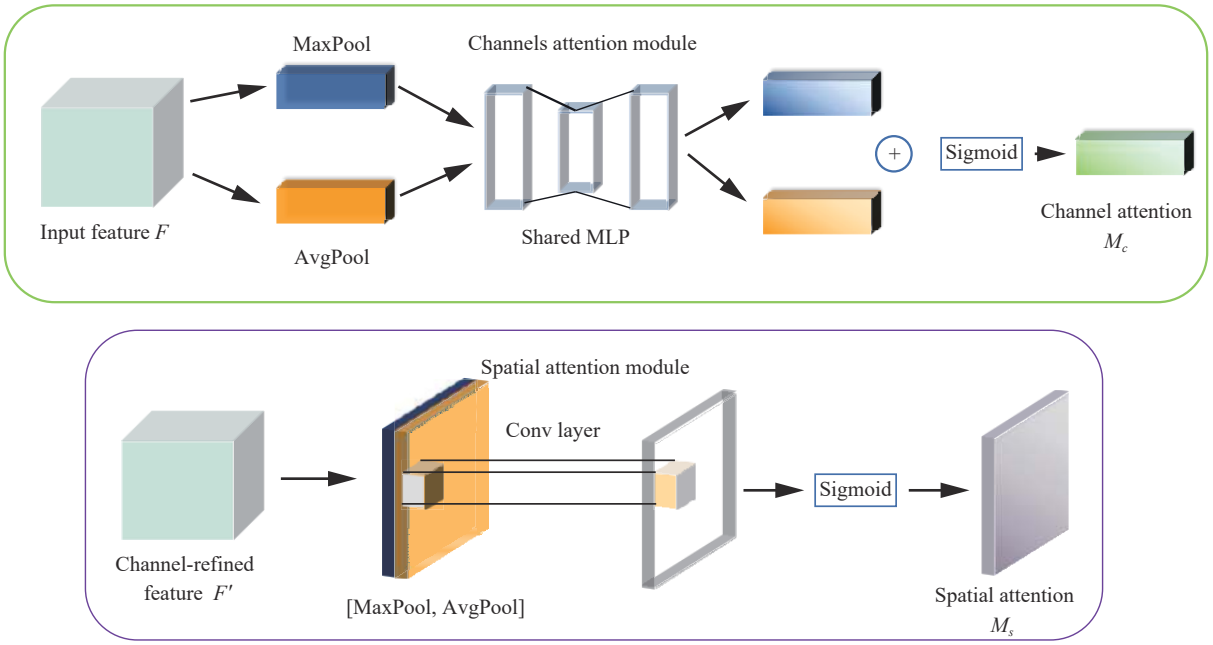

Fig. 5 The structure of each attention sub-module. The channel sub-module uses both max-pooling outputs and average-pooling outputs through a shared network which is composed of MLP with one hidden layer, and then the features are applied to the two channels, respectively. Finally, the attention results of the channels are obtained after the sigmoid function. The spatial sub-module uses channel-refined features that are pooled along the channel axis and forward them to a convolution layer and then we obtain the final spatial attention after the sigmoid function.

loss, face box regression loss, and facial landmark regression loss, remove the dense regression loss, and optimize them. The final multi-task loss is expressed as follows:

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*). \quad (1)$$

1) Face classification loss $L_{cls}(p_i, p_i^*)$, where p_i is the probability that the predicted faces are contained in the anchor and $p_i^* \in (0, 1)$ represents the negative anchor and the positive anchor, respectively.

2) Face box regression loss $L_{box}(t_i, t_i^*)$, where $t_i = \{t_x, t_y, t_w, t_h\}_i$ and $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}_i$ respectively represent the coordinate information of the prediction box and groundtruth related to the positive anchor.

3) Facial landmark regression loss $L_{pts}(l_i, l_i^*)$, where $l_i = \{l_{x1}, l_{y1}, \cdots, l_{x5}, l_{y5}\}_i$ and $l_i^* = \{l_{x1}^*, l_{y1}^*, \cdots, l_{x5}^*, l_{y5}^*\}_i$, respectively represent the predicted five facial landmarks and ground truth related to the positive anchor. λ_1 and λ_2 are the loss-balancing parameters, and we set them to 0.4 and 0.1, respectively.

4 Experiments

4.1 Experimental setting

In this paper, we use MobileNet-0.25[18] and ResNet-50[19] as the backbone to conduct experiments. By employing MobileNet-0.25 as the backbone, our model can achieve real-time on a single GPU. We use the stochastic gradient descent (SGD) optimizer (weight decay at 0.000 5, momentum at 0.9, batch size of 8×4) to train our model. The learning rate starts at 10^{-3}, increasing to 10^{-2} after 5 epochs, then divided by 10 at 55 and 68 epochs. Finally, the training process is completed at 100 epochs.

4.2 Anchor settings

For the anchor setting, we use the same strategy as

RetinaFace. We set our anchors from areas of 16^2 to 406^2 on pyramid levels. In addition, the aspect ratio is set to 1:1. During training, the anchor matches the ground truth when IoU is larger than 0.5 and matches the background when IoU is less than 0.3.

4.3 Data augmentation

According to the statistics of the WIDER FACE, there are around 20% tiny faces and 26% occluded faces. Therefore, the number of training samples in dense scenes may not be sufficient. Thus, we employ the random crop data augmentation and random horizontal flip.

4.4 Evaluation indicators

The measurement indicators in the detection mainly include precision, recall, and average precision (AP). The formula of each indicator is defined as follows:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

where TP represents the number of the positive samples that are predicted as positive samples, FP represents the number of negative samples that are predicted as positive samples, and FN represents the number of positive samples that are predicted as negative samples.

4.5 Datasets

In order to verify the effectiveness of the proposed algorithm, we test it on WIDER FACE[9] and FDDB[8] datasets, respectively. All algorithm models are trained on the WIDER FACE and tested on the FDDB and WIDER FACE.

4.5.1 WIDER FACE

The WIDER FACE[9] dataset contains 32 203 images and 393 703 face bounding boxes with strong variability in expression, illumination, scale, occlusion, and pose. It contains three parts: training set, validation set, and testing set. The validation set, and the testing set are divided into three parts: easy, medium and hard according to the detection difficulties, which can better verify the generalization ability of the model. Due to the strong variability of occlusion, scale and posture, the WIDER FACE dataset is one of the most challenging datasets.

We compare our model with the state-of-the-art detectors like S3FD[13], SSH[5], HR[4] and RetinaFace[6], as shown in Table 1. Our algorithm obtains the best results in all subsets. The results obtained are 96.2% (easy), 95.1% (medium) and 86.7% (hard) for ResNet-50[19], and 92.2% (easy), 89.9% (medium) and 76.7% (hard) for MobileNet-0.25[18]. More specifically, compared with the previous state-of-the-art results, our method improves by 2.3% on the hard set, which contains a lot of occluded and tiny faces. The precision-recall curves are shown in Fig. 6. The abscissa is the recall rate, and the ordinate is the precision rate. The area under the curve is the AP value, and the closer the curve is to the upper right, the better the model performance is. The AP curve also demonstrates the excellent performance of our model, especially in easy subsets. The effectiveness of the proposed algorithm is verified.

Table 1 Testing results of different network models on WIDER FACE

Method	AP (easy) (%)	AP (medium) (%)	AP (hard) (%)
MTCNN[3]	85.1	82.0	60.7
RetinaFace (MobileNet-0.25)[6]	90.7	88.2	73.8
Ours (MobileNet-0.25)	92.2	89.9	76.7
HR[4]	92.3	91.0	81.9
SSH[5]	92.7	91.5	84.4
S3FD[13]	93.5	92.1	85.8
RetinaFace (ResNet-50)[6]	95.5	94.0	84.4
Ours (ResNet-50)	96.2	95.1	86.7

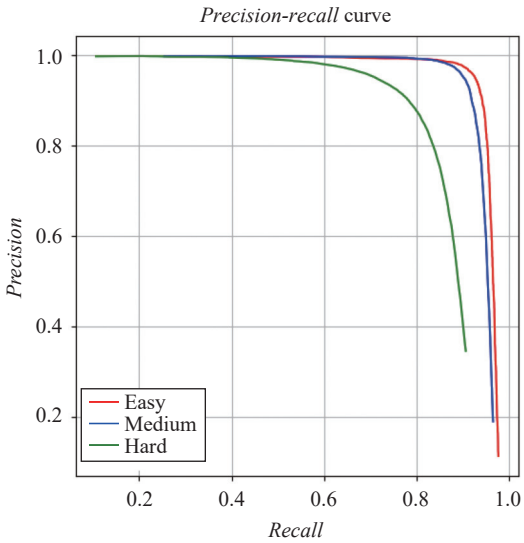

Fig. 6 *Precision-recall* curve on the WIDER FACE

The comparison of the visual results of our method and RetinaFace on the WIDER FACE can be found in Fig. 7. As illustrated in Fig. 7, our algorithm is better than RetinaFace for extremely tiny faces and severely occluded faces in dense scenes. RetinaFace is easy to recognize parts that resemble faces as faces, while our method significantly reduces false detections. Meanwhile, we can find that our method is more robust in the face location

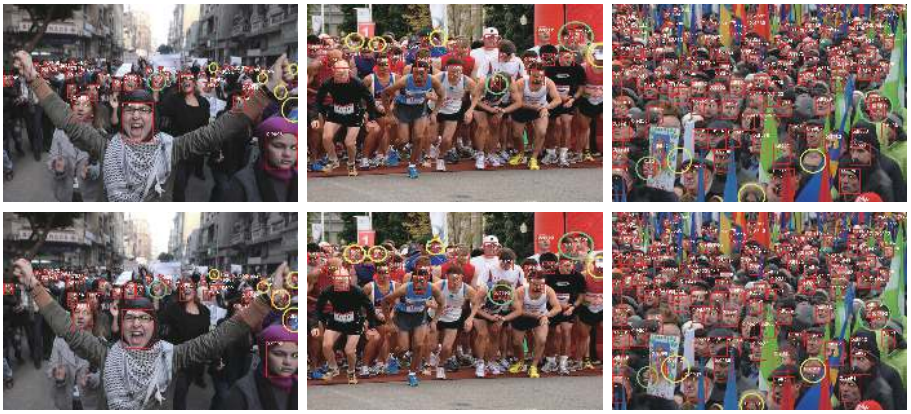

Fig. 7 Comparison of the visualization results on the WIDER FACE[9]. The top row is the RetinaFace[6] testing results, and the bottom row is the testing results of the proposed method. The yellow circles represent whether the face is detected, and the green circles represent correct detection.

for overlapping faces.

4.5.2 FDDB

In order to further verify the effectiveness of the algorithm, we tested it on FDDB[8]. FDDB has fewer faces, but there are also various face states and dense and serious occlusion problems. Our model is trained on Wider Face[9] training set and tested on the FDDB testing set. The results based on average precision are shown in Table 2. DDFD represents deep dense face detector, DP2MFD represents deep pyramid deformable parts model for face detection. Our model significantly outperforms state-of-the-art detectors on the FDDB test set, further showing the promising performance on dense faces.

The visualization results of our model and RetinaFace[6] can be found in Fig. 8. It can also be seen that our algorithm is better than RetinaFace for tiny faces and occluded faces in dense scenes. In addition, our algorithm is also better for complex angle faces such as profile faces.

Table 2 Test results of different network models on FDDB

Method	AP(%)
DDFD[20]	85.0
Cascade CNN[21]	85.6
Fast R-CNN[22]	89.9
DP2MFD[23]	91.7
UnitBox[24]	94.5
Retina face (MobileNet-0.25)[6]	98.7
Ours (MobileNet-0.25)	99.0
Retina face (ResNet-50)[6]	99.2
Ours (ResNet-50)	99.6

4.6 Ablation study

To further verify the proposed network, we conduct additional ablation experiments to examine the effects of the global context fusion module and the visual attention module on face detection performance. As shown in Table 3, along with the global context fusion module, the accuracy has been further improved by 0.4%, 0.8%, 1.5% AP in the easy, medium and hard subsets. From the experiments, we come to the conclusion that the global context fusion module has a better detection effect on a

Fig. 8 Comparison of the visualization results on FDDB[8]. The top row is the RetinaFace[6] testing results, and the bottom row is the testing results of the proposed method.

Table 3 Effectiveness of each strategy

Strategy	Baseline	Our method		
Global context fusion module		√		√
Visual attention network			√	√
Easy	95.5	95.9	95.7	96.2
Medium	94.0	94.8	94.5	95.1
Hard	84.4	85.9	85.5	86.7

small scale, blur, occlusion and overlapping faces. Therefore, it is crucial for improving the accuracy of face detection. The performance is improved by 0.2%, 0.5%, 1.1% AP in the easy, medium and hard subsets, respectively, along with the visual attention mechanism. Fig. 9 is the comparison of visual results with an attention module. As shown in Fig. 9, background like hairs, hands and patterns could be misclassified as a face without the visual attention mechanism, while the addition of an attention module significantly reduces the false detection rate, suggesting that the visual attention mechanism effectively decreases the false positive rate and enables a further improvement.

4.7 Fake face detection

In addition, we also carried out an extended experiment. Since there are some algorithms that can generate fake faces, we also used our model to detect these fake faces and achieved good results. As illustrated in Fig. 10, the top row shows real face images, and the bottom row shows fake face images generated by generative flow (Glow), StyleGAN, progressive growing of generative adversarial nets (PGGAN), Face2Face, and StarGAN, respectively. Nevertheless, our model can still detect the faces and five key points of the faces.

5 Conclusions

In this paper, to address the problem of dense faces

Fig. 9 Comparison of visual results with and without the attention module. The top row shows the detection results without the attention module, and the blue circles represent the parts of error detection. The bottom row shows the detection results with the attention module added, and the blue circles represent the parts of correct detection.

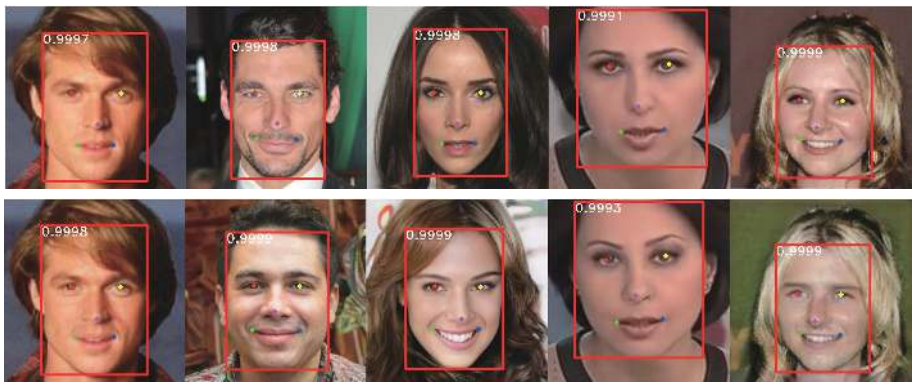

Fig. 10 Detection results on fake faces. The top row shows real face images. The bottom row shows fake face images generated by Glow, StyleGAN, PGGAN, Face2Face and StarGAN, respectively.

with tiny size and serious occlusion, we presented a novel facial detector with a global context fusion module and an attention mechanism, which can significantly improve the accuracy in dense face scenes and not compromise the localization accuracy while obtaining a high recall rate. Our solution outperforms state-of-the-art methods in the current most challenging benchmarks for face detection. Although our detector has achieved good results, further research is needed in the future to improve the efficiency of the detector and apply it in practical scenarios.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61973009).

References

[1] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[2] F. F. Zhang, T. Z. Zhang, Q. R. Mao, C. S. Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3359–3368, 2018. DOI: 10.1109/CVPR.2018.00354.

[3] K. P. Zhang, Z. P. Zhang, Z. F. Li, Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. DOI: 10.1109/LSP.2016.2603342.

[4] P. Y. Hu, D. Ramanan. Finding tiny faces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 1522–1530, 2017. DOI: 10.1109/CVPR.2017.166.

[5] M. Najibi, P. Samangouei, R. Chellappa, L. S. Davis. SSH: Single stage headless face detector. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 4885–4894, 2017. DOI: 10.1109/ICCV.2017.522.

[6] J. K. Deng, J. Guo, Y. X. Zhou, J. K. Yu, I. Kotsia, S. Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. [Online], Available: https://arxiv.org/abs/1905.00641, 2019.

[7] T. Y. Lin, P. Goyal, R. Girshick, K. M. He, P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. DOI: 10.1109/TPAMI.2018.2858826.

[8] V. Jain, E. Learned-Miller. FDDB: A Benchmark for Face Detection in Unconstrained Settings, Technical Report UM-CS-2010-009, University of Massachusetts, USA, 2010.

[9] S. Yang, P. Luo, C. C. Loy, X. O. Tang. WIDER FACE: A face detection benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 5525–5533, 2016. DOI: 10.1109/CVPR.2016.596.

[10] C. C. Zhu, Y. T. Zheng, K. Luu, M. Savvides. CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection. *Deep Learning for Biometrics*, B. Bhanu, A. Kumar, Eds., Cham, Germany: Springer, pp. 57–79, 2017. DOI: 10.1007/978-3-319-61657-5_3.

[11] S. Q. Ren, K. M. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 91–99, 2015. DOI: 10.5555/2969239.2969250.

[12] T. Xu, D. K. Du, Z. Q. He, J. T. Liu. PyramidBox: A context-assisted single shot face detector. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 812–828, 2018. DOI: 10.1007/978-3-030-01240-3_49.

[13] S. F. Zhang, X. Y. Zhu, Z. Lei, H. L. Shi, X. B. Wang, S. Z. Li. S.3FD: Single shot scale-invariant face detector. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 192–201, 2017. DOI: 10.1109/ICCV.2017.30.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 2017–2025, 2015. DOI: 10.5555/2969442.2969465.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, E. H. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020. DOI: 10.1109/TPAMI.2019.2913372.

[16] S. Woo, J. Park, J. Y. Lee, I. S. Kweon. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 3–19, 2018. DOI: 10.1007/978-3-030-01234-2_1.

[17] J. F. Wang, Y. Yuan, G. Yu. Face attention network: An effective face detector for the occluded faces. [Online], Available: https://arxiv.org/abs/1711.07246, 2017.

[18] A. G. Howard, M. L. Zhu, B. Chen, D. Kalenichenko, W. J. Wang, T. Weyand, M. Andreetto, H. Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. [Online], Available: https://arxiv.org/abs/1704.04861, 2017.

[19] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[20] S. S. Farfade, M. J. Saberian, L. J. Li. Multi-view face detection using deep convolutional neural networks. In *Pro-*

ceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, Shanghai, China, pp. 643–650, 2015. DOI: 10.1145/2671188.2749408.

[21] H. X. Li, Z. Lin, X. H. Shen, J. Brandt, G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 5325–5334, 2015. DOI: 10.1109/CVPR.2015.7299170.

[22] R. Girshick. Fast R-CNN. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 1440–1448, 2015. DOI: 10.1109/ICCV.2015.169.

[23] R. Ranjan, V. M. Patel, R. Chellappa. A deep pyramid deformable part model for face detection. In *Proceedings of the 7th IEEE International Conference on Biometrics Theory, Applications and Systems*, IEEE, Arlington, USA, pp. 1–8, 2015. DOI: 10.1109/BTAS.2015.7358755.

[24] J. H. Yu, Y. N. Jiang, Z. Y. Wang, Z. M. Cao, T. Huang. UnitBox: An advanced object detection network. In *Proceedings of the 24th ACM International Conference on Multimedia*, ACM, Amsterdam, The Netherlands, pp. 516–520, 2016. DOI: 10.1145/2964284.2967274.

Lin Song received the B. Sc. degree in measurement and control technology and instrumentation from YanTai University, China in 2019. She is now a master student with Department of Control Science and Engineering, Beijing University of Technology, China.

Her research interests include deep learning and computer vision.

E-mail: songlin@emails.bjut.edu.cn (Corresponding author)

ORCID iD: 0000-0003-2289-7325

Jin-Fu Yang received the Ph.D. degree in pattern recognition and intelligent systems from National Laboratory of Pattern Recognition, Chinese Academy of Sciences, China in 2006. He is now a professor with Faculty of Information Technology and Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, China.

His research interests include pattern recognition, computer vision and robot navigation.

E-mail: jfyang@bjut.edu.cn

Qing-Zhen Shang received the M. Eng. degree in mathematics from Hebei University, China in 2017. She is a Ph. D. degree candidate at Department of Control Science and Engineering, Beijing University of Technology, China.

Her research interests include deep learning and computer vision.

E-mail:shangqingzhen@emails.bjut.edu.cn

Ming-Ai Li received the Ph. D. degree from Beijing University of Technology, China in 2006. She is now a professor with Faculty of Information Technology and Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, China.

Her research interests include brain-computer interface, intelligent control, pattern recognition and implementation of autonomous learning control technology for flexible two-wheeled upstanding robots.

E-mail: limingai@bjut.edu.cn