Journal of Engineering and Technological Sciences

Research Paper

An Adaptive Multi-region Fusion Network for Dense Face Detection

Luxia Yang, Chuanghui Zhang, Hongrui Zhang* & Yilin Hou

Department of Computer Science and Technology, Taiyuan Normal University, No. 319 Daxue Street, Yuci District, Jinzhong City, Shanxi Province, Jinzhong 030619, China

*Corresponding author: zhanghongrui@tynu.edu.cn

Abstract

In recent years, face detection has been widely applied in various intelligent monitoring systems. However, missed detections and low detection accuracy present challenges, such as small, blurred, and occluded faces in multi-face detection scenarios. To address these challenges, an adaptive multi-region fusion network is designed for dense face detection. First, in the shallow layers of the network, a multi-scale cross-stage fusion (MC4f) module is designed to replace the C3 module, which solves the issue of gradient explosion or disappearance in deep networks and promotes the effective convergence of the network on small datasets. An adaptive fusion explicit spatial vision centre (AESVC) is then designed between the backbone and neck networks to adaptively fuse local and global features to refine face information and enhance feature representation capabilities in complex tasks. Subsequently, a multi-scale parallel attention mechanism (MSPAM) is proposed to enhance the cross-scale fusion of facial features and reduce the loss of shallow features. Finally, to achieve accurate facial key point localisation and alignment, wing loss and A-loss functions are integrated, which balances the relationship between easy and difficult samples. Compared with the original model, the proposed model increases the mean average precision (mAP) by 1.75, 2.01, and 3.06% for easy, medium, and hard samples, respectively. The experimental results prove the effectiveness of the adaptive multi-region fusion network for dense face detection.

Keywords: *face detection; facial key point localisation; small-blurred faces; multi-scale; adaptive fusion.*

Introduction

Face detection (Kobylkov & Vallortigara, 2024; Naseri et al., 2023; Wang & Deng, 2021) is an important task in computer vision that has progressed significantly with the development of deep learning (Debbouche et al., 2021). As the first step for tasks such as face recognition (Alansari et al., 2023; Cardona-Pineda et al., 2023; Gao et al., 2023), face tracking (Hannuksela, 2022; ImranAhsan et al., 2024; Liu et al., 2022), face alignment (Freitas et al., 2024; Ma et al., 2024; Saadabadi et al., 2024), and expression analysis (Ben et al., 2021), face detection has been extensively researched worldwide (Hioual et al., 2022). In recent years, face detection performance has significantly improved and has been increasingly applied in various fields such as intelligent transportation (Li et al., 2024), security surveillance (Liu et al., 2023), education (Hioual et al., 2022), training, and medical diagnostics. However, face detection in densely populated scenarios still faces challenges, such as occlusion and small-scale face imaging.

Many studies worldwide have explored the field of face detection. He et al. (2023) employed an improved training sample selection (ITSS) approach to identify valuable samples during the training phase. This approach incorporated a residual feature pyramid fusion (RFPF) module to aggregate features with strong semantic resilience, thereby improving face representation across various levels of the feature pyramid. Liu et al. (2024) adopted an improved RetinaFace algorithm that incorporated deformable convolution (DC), feature pyramid networks (FPN), and coordinate attention (CA) mechanisms. With minimal additional computational overheads, the algorithm enhanced the semantic information of the lower-level features, thereby improving the robustness of face detection across different face sizes. Yu et al. (2022) adopted a receptive field enhancement (RFE) module to expand the receptive field to detect smaller faces and reduce the intersection over union (IoU) sensitivity to small object localisation deviations through the normal wasserstein distance (NWD) loss (Xu et al., 2024). The module also introduced a self-supervised equivariant attention mechanism (SEAM) and repulsion loss to address face occlusion issues but neglected the regression of facial key points. Based on the you only look once (YOLO) algorithm, YOLO5Face, proposed by Qi et al. (2022), is an adaptation of the YOLOv5 framework for face detection, adding five facial key point regression heads and using the wing loss function. This algorithm recognises and locates facial features with high accuracy and performs well on the wider face dataset (Chen et al., 2021).

ISSN: 2337-5779

DOI: 10.5614/j.eng.technol.sci.2025.57.3.9

Although the aforementioned face detection research has made significant progress and is maturing, the following challenges remain:

1. Common feature fusion operations do not perform effective filtering or weighted fusion, which may lead to the introduction of redundant information, thereby reducing the expressive power of a model, which is detrimental to the detection of small-scale faces.
2. Some methods enhance the global perception of a model through global features, and local feature information may be lost during layer-by-layer fusion, resulting in the model being unable to focus effectively on key details.
3. Both shallow semantic and deeper features are important; commonly used approaches focus on shallow information, ignoring the relationship between contextual information and more fine-grained features for deeper features.
4. Traditional regression loss, such as L2 or mean squared error, typically perform well when the objects are large; however, when dealing with small objects, the errors are more difficult to control with precision.

To address these challenges, this study aims to improve the accuracy of multi-scale face detection using a lightweight model and proposes an adaptive multi-region fusion network for face detection.

The main contributions of this study are summarized as follows:

1. The YOLOv5n backbone is reconstructed using the designed multi-scale cross-stage fusion (MC4f) module to address the gradient explosion or disappearance issues that lightweight and deep linear networks may encounter and to improve the cross-scale feature fusion capability.
2. An adaptive fusion explicit spatial vision centre (AESVC) is designed between the spine and neck to enhance the global perception of the model by spatially fusing the blocks to establish complementary connections between local and global features and reduce the loss of local feature information.
3. A multi-scale parallel attention mechanism (MSPAM) is proposed to enhance the multi-scale fusion of facial features by focusing on deep features and multi-scale feature information, and reducing the loss of shallow features.

The model introduces facial key point regression and wing loss, ensuring stronger robustness and accuracy, particularly for small objects or challenging facial detection scenarios.

Related Work

YOLO

This study adopts YOLOv5 (Jocher et al., 2022) as the baseline model, with a network architecture consisting of five main parts; an input module (input), backbone network (backbone), feature fusion network (neck), training process (training), and prediction process (prediction).

In the input module, the model uses a 640 × 640 × 3 RGB image as the input. The backbone network module alternates between the Conv BN SiLU (CBS), with batch normalization (BN) and the sigmoid linear unit (SiLU), and C3 modules for feature extraction. In particular, the design of the C3 module draws on the architecture of the cross-stage partial network (CSPNet) (Wang et al., 2020) to effectively extract facial feature information from images. The feature fusion network module combines the architectures of FPN (Gong et al., 2021) and path aggregation networks (PANet) (Liu et al., 2018) to achieve the effective fusion of facial features at different levels. The architectural design and implementation of YOLOv5 have delivered excellent performance in face detection applications.

The inception of the YOLO algorithm (Jiang et al., 2022) marked a significant shift towards whole-image processing. With YOLO, object detection tasks exhibit improved efficiency because images are processed in a single pass. This approach contrasts with selective search methods, leading to significantly faster inference times, which are crucial for real-time applications.

The YOLO algorithm has undergone various iterations, each aimed at refining the performance and efficiency of the model (Chitraningrum et al., 2024). Notable improvements include YOLOv2, which introduced anchor boxes to predict object bounding boxes more accurately, and could detect over 9000 object categories by jointly training on the ImageNet and common objects in context (COCO) datasets (Sharma, 2021). YOLOv3 further enhanced the ability of the algorithm to detect objects at different scales using a multi-scale detection approach. YOLOv4 and YOLOv5 continued this trend, with optimisations for improved speed and accuracy, allowing the model to perform well on lower-power devices while maintaining high performance.

YOLO has been tailored to address the challenges associated with face detection. Enhancements such as improved loss functions for bounding box predictions, integration of additional layers attuned to facial feature detection, and training on extensive face datasets have been key to adapting YOLO for face detection. Research continues to push the boundaries of the applications of YOLO in face detection, balancing the trade-off between real-time processing and high-precision requirements for detecting small or partially occluded faces, as well as operating across a range of scales and environmental conditions.

Face Detection

Face detection is a critical area of research, serving as the cornerstone for various applications such as biometric authentication, surveillance, and social media. Classic face detection techniques, such as Viola-Jones detectors (Rahmad et al., 2020), are the foundation of developments in this area, leveraging features such as Haar cascades. However, these methods struggle with image variations owing to lighting, occlusion, and orientation.

With the emergence of deep learning technologies, convolutional neural networks (CNNs) (Alzubaidi et al., 2021) have become the foundation for advancements in face detection techniques. Proposed architectures, such as region-based CNN (R-CNN) (Xie et al., 2021) and its iterations of fast R-CNN and faster R-CNN (Ren et al., 2015)), have demonstrated significant improvements in accuracy and reliability by incorporating regional information with deep learning.

Further research on face detection is inclined towards creating robust systems that not only perform well in controlled environments but also excel in uncontrolled real-world scenarios that involve challenges such as occluded faces, densely packed faces at multiple scales, and small, blurred faces. Numerous methods have been proposed to address these issues, particularly focusing on the scale, context, and anchors for face detection in various complex scenes. These methods include multi-task cascaded CNN (MTCNN) (Xiang & Zhu, 2017), FaceBoxes (Zhang et al., 2017a), single shot scale-invariant face detector (S3FD) (Zhang et al., 2017b), dual shot face detector (DSFD) (Li et al., 2019), RetinaFace (Deng et al., 2019), RefineFace (Zhang et al., 2020), and the more recent additions of automatic and scalable face detector (ASFD) (Zhang et al., 2020), MaskFace (Yashunin et al., 2020), TinaFace (Zhu et al., 2020), MogFace (Liu et al., 2022), sample and computation redistribution for efficient face detection (SCRFD) (Guo et al., 2021), and YOLO5Face (Qi et al., 2022). YOLO5Face employs the wing loss (Feng et al., 2018) function to locate five facial landmarks, aiding in the supervision of face detection. This method is adopted in this study for the localisation of five facial key points.

Proposed Method

Overview

The proposed model adopted YOLOv5 as the baseline framework. After optimisation, the algorithm achieved a better facial detection performance. The reconstructed YOLOv5nFace network architecture includes the backbone, neck, and head, and is shown in Figure 1.

The backbone network was reconstructed using the MC4f module with a partial network bottleneck and four convolutions to improve the feature extraction capabilities. An AESVC was designed between the backbone and neck of the network to improve the ability of the expressive ability of the model for complex tasks. The MSPAM attention mechanism was proposed to boost the cross-scale fusion of facial features. In the head section, the model conducted multi-scale target detection on the feature maps extracted by the backbone to achieve face detection

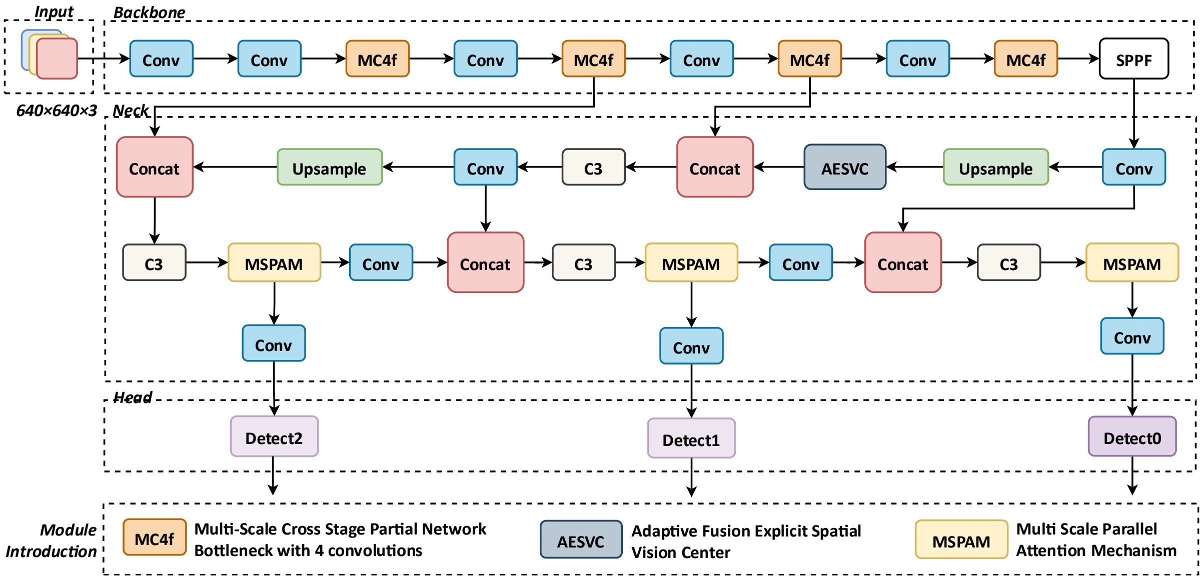
Input
Backbone
640×640×3
Neck
Head
Conv
MC4f
SPPF
Concat
Upsample
C3
AESVC
MSPAM
Detect2
Detect1
Detect0
Module Introduction
MC4f Multi-Scale Cross Stage Partial Network Bottleneck with 4 convolutions
AESVC Adaptive Fusion Explicit Spatial Vision Center
MSPAM Multi Scale Parallel Attention Mechanism

Figure 1 Proposed network architecture.

MC4f

To improve the feature extraction ability of the network and capture more facial features, an MC4f module was designed. Inspired by the C3 and C2f modules, a split operation was used. The number of input tensor channels for each bottleneck structure in the MC4f module was only half that of the previous structure. This significantly reduced the number of calculations, and the increase in the gradient flow also significantly improved the convergence speed and effect. Furthermore, the efficient layer aggregation network (ELAN) concept from YOLOv7 was employed to obtain more abundant gradient flow details while maintaining a lighter framework by increasing the parallelism among the gradient flow branches.

The original C2f module carries the risk of gradient disappearance and explosion. Therefore, residual connections were added to the MC4f module, the convolution operation was integrated into the residual channel, and the main channel was compressed through point-wise convolution to obtain the final output. The design of the residual channel effectively solved the problem of gradient explosion or disappearance in deep networks and promoted convergence on small datasets. The structure of the MC4f module is shown in Figure 2.

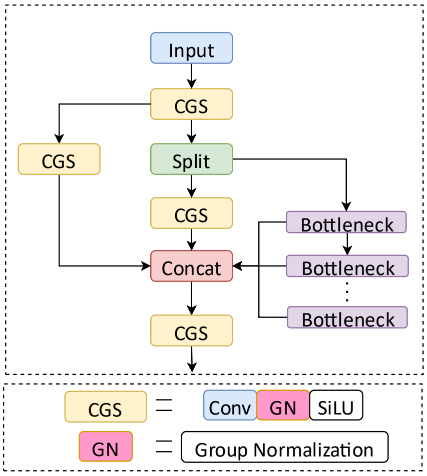

Figure 2 MC4f module with a partial network bottleneck and four convolutions. The full name of CGS is Conv Group Normalization SiLU.

In the original C2f, BN in the CBS module was replaced by group normalization (GN) to produce Conv GN SiLU (CGS) (Wu & He, 2018). BN may produce large errors in small batch training, whereas GN can significantly reduce errors under the same small batch condition; its calculation is independent of the batch size, and it can maintain a stable accuracy under various batch sizes.

In summary, the channel splitting and parallel processing operations of MC4f enhance the model's ability to capture multi-scale features, which helps to identify local details more effectively in difficult scenes. Furthermore, the residual cascade mechanism in the designed structure plays a crucial role in addressing gradient attenuation. It achieves this through optimized skip-connections, which enhance training stability in deep network architectures.

AESVC

In dense face detection scenarios, missed detections frequently occur with small-scale faces owing to the insufficiency of effective features. The explicit vision centre (EVC) module (Quan et al., 2023) aimed to address this challenge, however, it failed to mitigate background noise interference and effectively extract features at various scales. To enhance the detection performance and address these inherent issues in the EVC, an improved AESVC architecture was proposed within the FPN framework. The structure of the AESVC is illustrated in Figure 3.

Compared with the EVC, the following improvements have been made in the AESVC module. First, the CBS module was introduced within the learnable vision centre (LVC) module, replacing the original rectified linear unit (ReLU) activation function in the EVC module with a SiLU. SiLU has non-zero gradients in both the positive and negative domains, which facilitates network-refined data representations.

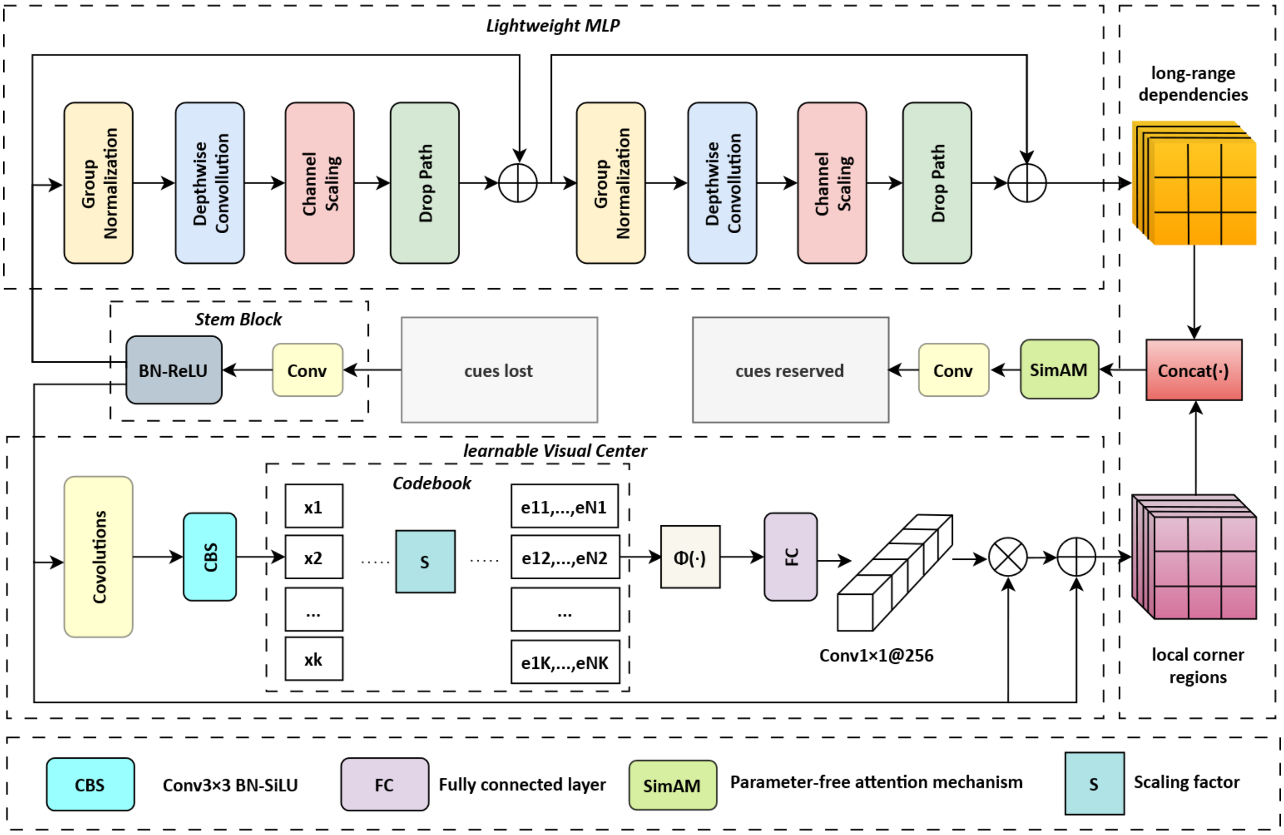

Figure 1 Adaptive fusion explicit spatial vision center, where a lightweight MLP architecture is used to capture the long-range dependencies and a parallel learnable visual center mechanism is used to aggregate the local corner regions of the input image. The integrated features contain advantages of these two blocks, so that the detection model can learn an all-round yet discriminative feature representation.

Second, the global features output by the multi-layer perceptron (MLP) and the local features from the LVC module were adaptively fused through the simplified attention module (SimAM) (Yang et al., 2021). SimAM uses reversible mapping to infer attention weights, thereby assigning proper weights to both global and local features and improving the ability of the model to express itself in complex tasks. Moreover, without increasing parameters, the perceptual scope of the model is enhanced, enabling an explicit increase in the spatial weighting of the global representation along the top-down path; the AESVC module can realize multi-scale feature representations. Experiments have shown that incorporating the AESVC module into the backbone network can improve the detection speed and expressive capability of multi-scale features of the model.

MSPAM

In the context of face detection in complex scenes, negative factors, such as occlusions, unclear images, and small, blurry faces, can significantly affect dense face detection. The convolutional block attention module (CBAM) (Woo et al., 2018) integrates a spatial attention module (SAM) (Tootell et al., 1998) and channel attention module (CAM) (Qin et al., 2021) in a serial structure. Furthermore, it enhances the accuracy and robustness of facial detection.

Although the CBAM places high importance on the CAM, it still affects the features learned by the subsequent SAM. Therefore, to allow both types of attention modules direct access to the original features, this study improved upon CBAM by changing its connection from serial to parallel. This resulted in a MSPAM that does not need to consider the order of spatial and channel attention. The improved MSPAM structure is illustrated in Figure 4.

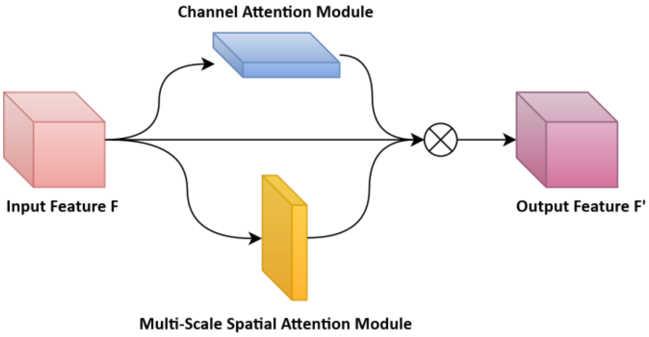

Figure 2 MSPAM structure.

The MSPAM integrates multi-scale convolution. The structure of the multiscale spatial attention module is shown in Figure 5. The input feature map F undergoes global average and maximum pooling separately in the channel dimension. After concatenation and fusion, the input feature map F is then input into a multi-scale convolutional pathway. By combining features from different receptive fields, fine local details and global features are captured. The structure includes a standard 1 × 1 convolution layer and three depth-wise separable convolutions of sizes 3 × 3, 5 × 5, and 7 × 7, respectively. The outputs are then concatenated and fused, passed through a 1 × 1 convolution layer for dimension reduction, and finally normalised using a sigmoid activation function to generate the spatial attention weights M_S . The specific formulae for this process are given by Eqs. (1) and (2):

$$\Lambda(\cdot) = [DWConv_{(1,1)}(\cdot), DWConv_{(3,3)}(\cdot), DWConv_{(5,5)}(\cdot), DWConv_{(7,7)}(\cdot)] \quad (1)$$

$$M_S(F) = \sigma(f^{1\times1}(concat(\Lambda[AvgPool(F); MaxPool(F)]))) \quad (2)$$

where Λ denotes the multi-scale convolutional pathway, $DWConv_{(i,i)}$ denotes the use of a convolutional kernel of size i, and $f^{1\times1}$ represents a convolution operation with a kernel size of 1 × 1.

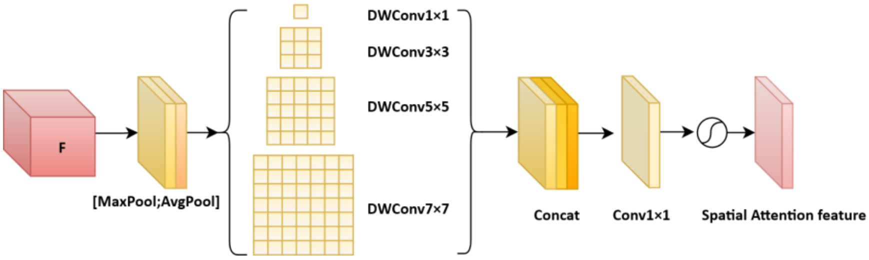

Figure 3 Multi-Scale Spatial Attention Module.

The MSPAM module integrates channel attention and spatial attention to extract multi-scale features from disparate dimensions. Among them, channel attention focuses on semantic-level features, such as channel responses of skin color and texture. In contrast, spatial attention extracts spatial information under different perceptual fields, which uses a multi-scale convolutional kernel to adaptively capture alternative features in occluded regions of the face. Moreover, this module forms a synergy with the lightweight gradient flow optimization of the MC4f module, which together improve the robustness and accuracy of dense face detection.

Face Loss Function

The key points of facial features can be used for facial alignment and recognition. The traditional key points involved 68 distinct locations. This was reduced to five key points using the MTCNN (Xiang & Zhu, 2017), which have been extensively adopted for facial recognition. Drawing inspiration from YOLO5face, these five key points were incorporated into the model in this study. The quality of face key point positioning affects the quality of face alignment and recognition. Common face detectors do not include key points, which were added to the model in this study as regression headers. The localisation output of the face key points was used to align the face image before it was sent to the facial recognition network. Therefore, the processing of difficult samples was optimised while improving positioning accuracy.

L1 and L2 loss functions are commonly used for locating facial key points; however, the L2 loss function used in the MTCNN is not sensitive to small errors. The wing loss function used in this study optimized this problem. The formula for the wing loss is shown in Eq. (3).

$$wing(x) = \begin{cases} w \cdot ln(1 + |x|/e), if\ x < w \\ |x| - C, \quad \text{otherwise} \end{cases} \tag{3}$$

where the range of the nonlinear part is set to a non-negative value w, ranging from $-w$ to w; e is used to limit the curvature to a nonlinear region; and $C = w - w\ln(1 + w/e)$ is a constant used to connect the linear and nonlinear parts of the segmentation smoothly. Compared with the L1 and L2 functions, the wing loss significantly enhances the response in the small-error region close to zero. The loss function for the facial key point vector s is defined in Eq. (4).

$$Wing_{loss} = \sum_i wing(s_i - s_i') \tag{4}$$

The bounding box and wing loss five-point regression loss functions were combined to obtain the FaceLoss function in this study, as shown in Eq. (5).

$$FaceLoss = L_{\alpha-CIOU} + \lambda_L \cdot Wing_{loss} \tag{5}$$

where λ_L is the weight factor of the key point-regression loss function. After blending the loss functions, the model became more sensitive to small errors, thereby significantly improving the detection results for small-scale faces.

Experiments

Implementation and Dataset

All experiments were conducted using a deep-learning framework built on Python 3.8.17 and PyTorch 1.11.0. The operating system used was Ubuntu 20.04, and the GPU model was an NVIDIA GeForce RTX 4090 with 24GB of VRAM.

In the experiments, the size of the training images was set to 640 × 640 pixels, the batch size was set to 64, and the number of epochs was set to 250. The optimization process uses SGD with momentum mechanism. The initial learning rate is calibrated to 1E-2, and the final learning rate is 1E-3, while the weight decay parameter is set to 5E-3. In the first three epochs of the warm-up phase, a momentum value of 0.8 is applied, which is subsequently adjusted to 0.937. The IoU threshold for the NMS operation is set to 0.5. The anchor box sizes for the wider face dataset were calculated using the K-means clustering algorithm, as listed in Table 1.

The wider face dataset is the most extensive image collection for face detection, comprising 32,203 images and over 400,000 faces. The dataset is diverse, containing many small-scale faces and complex scenes and supporting face detection assessment. The wider face dataset is randomly partitioned into training, validation, and testing sets, with allocations of 50, 10, and 40%, respectively. It is organised into three difficulty categories: easy, medium, and hard, with the hard category being the most challenging because it mostly contains small-scale faces in complex scenes. The performance on the hard subset is most indicative of the ability of a face detector to detect small-scale faces. Therefore, the wider face dataset is suitable for validating the designed algorithm in various complex scenes.

Table 1 Anchor Box Sizes for Wider Face Dataset.

Feature Map Size	Anchor Box Size		
160 × 160	4， 5	6， 7	8， 11
80 × 80	12， 15	17， 21	24， 30
40 × 40	33， 42	48， 67	67， 87
20 × 20	110， 155	167， 233	248， 387

Evaluation Metrics

To evaluate the efficiency and precision of face detection, the main evaluation indicators were the mean average precision (mAP) of each category, parameters (params), and detection speed (speed). mAP reflects the overall performance of the algorithm, as shown in Eqs. (6)-(8), where the size of the parameter count indicates the possible application scenarios of the model; and the detection speed is the time required to detect a single image, measured in milliseconds.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$AP = \frac{1}{r}\sum_{i=1}^{r} P_i \tag{7}$$

$$mAP = \frac{1}{SUM}\sum_{j=1}^{SUM} AP_j \tag{8}$$

where *TP* represents the true-positive samples in the detection results, *FP* represents false-positive samples where the detection result is positive for a negative sample, *r* represents all possible values of the recall rate, and *SUM* denotes the total number of categories.

Ablation Study

To validate the feasibility of each step of the improvements in this study, an ablation experiment was conducted, as listed in Table 2.

MC4f vs. C3. The mAP performance of substituting the backbone network C3 module with the MC4f module is listed in Table 2. Based on these findings, it was evident that there were increases of 1.64, 1.89, and 3.30% in the easy, medium, and hard subsets, respectively. This is due to the MC4f module's optimized feature extraction structure and improved normalization strategy. Therefore it has stronger feature extraction capability.

AESVC. The performance enhancements from the AESVC module are listed in the third row of Table 2, where it is evident that the inclusion of AESVC led to mAP increases of 0.89, 1.30, and 2.34% for the easy, medium, and hard subsets, respectively. The enhanced performance of the model on the hard subset can be attributed to the adaptive fusion capability of the AESVC module. Specifically, one of the SimAM attention mechanisms implemented adaptive feature compensation for global and local features in both the channel and spatial dimensions.

MSPAM. The fourth row of Table 2 shows that incorporating the MSPAM into the neck network improved the mAP performance by 0.96, 1.16, and 2.13 for the easy, medium, and hard subsets, respectively. Additionally, in comparison with the CBAM, this module achieved improvements of 0.86, 0.79, and 1.10% across the easy, medium, and hard subsets, respectively. The above performance improvement is due to the parallel attention structure design of the MSPAM module. This structure exploited dual attention to extract original features simultaneously, thus avoiding feature loss.

Table 2 Ablation Experiment

Model	**Evaluation Indicators**			
	Easy	**Medium**	**Hard**	**Params (M)**
YOLOv5n (baseline)	92.09	90.05	80.49	1.886
YOLOv5n+MC4f	93.73	91.94	83.79	2.351
YOLOv5n+AESVC	92.98	91.35	82.83	2.856
YOLOv5n+MSPAM	93.05	91.21	82.62	1.789
YOLOv5n+CBAM	92.19	90.42	81.52	1.951
YOLOv5n+AESVC+MC4f	93.50	91.79	83.80	3.570
YOLOv5n+AESVC+MC4f+MSPAM	**93.84**	**92.06**	**83.55**	**3.099**

Summary. The mAP performance of the model with all aforementioned modules included is listed in the seventh row of Table 2, which shows that the proposed model increased the mAP by 1.75, 2.01, and 3.06% for the easy, medium, and hard subsets, respectively, compared to the baseline. The number of parameters was approximately 1.5 times that of the original model, and the model complexity also increased. The results indicated that the proposed network model exhibited improvements in the detection performance of small-scale, densely packed, and occluded faces. Although the complexity slightly increased, it ensured a balance between a lightweight and accurate model and satisfied the requirements for real-time performance.

Comparisons to the State-of-the-art

Quantitative Results

As shown in Table 3, this study mainly focused on balancing the enhancement of the detection accuracy for hard-to-detect facial models with the need for a lightweight model. A detailed performance comparison analysis was conducted on the leading facial detection algorithms. As shown in Table 3, the size of the model was only 3.570 MB, approximately one fortieth of the size of the DSFD (ResNet152) model, and it increased the mAP values by 0.59 and 12.16% on the medium and hard subsets, respectively. Compared to the YOLO5face (YOLOv5s) model, the proposed model was half the size, with mAP values improved by 0.40% on the hard subsets. Compared to the advanced YOLOv7-tiny, YOLOv8n-face, and YOLOv8s-face algorithms, the mAP values of the proposed model on the hard subset improved by 1.45, 4.55, and 0.45%, respectively. Although the performance of the proposed model on easy and medium subsets could be improved, it ensured the accuracy of hard-to-detect face prediction while being lightweight, being advantageous for lightweight end-terminal deployments. Furthermore, compared to YOLO5face (YOLOv5n), which has a similar lightweight, although the proposed model had a slightly increased parameter count, the algorithm improved the mAP values by 0.23, 0.52, and 3.02% on the easy, medium, and hard subsets, ensuring overall facial detection performance while being lightweight.

Therefore, the proposed model successfully addressed the challenges of achieving a high detection accuracy and a lightweight model for face detection. As listed in Table 3, the experimental results validated the effectiveness of the proposed model. Additionally, the mAP of each subset was improved and a lightweight was produced. This indicates that the proposed model achieved more efficient performance in resource-constrained situations.

Table 3 Comparison Experiments of Mainstream Algorithms.

Model	Evaluation indicators			
	Easy	Medium	Hard	Params(M)
YOLOv5n (baseline)	91.99	89.79	78.89	1.886
RetinaFace (MobileNet0.25)	87.78	81.16	47.32	0.44
YOLO5face (YOLOv5n-0.5)	90.76	88.12	73.82	0.447
SCRFD-0.5GF	90.57	88.12	68.51	0.57
FaceBoxes	76.17	57.17	24.18	1.01
YOLO5face (YOLOv5n)	**93.61**	**91.54**	**80.53**	**1.726**
YOLOv7-tiny	**94.70**	**92.60**	**82.10**	**13.20**
YOLOv8n-face	**94.50**	**92.20**	**79.00**	**-**
YOLOv8s-face	**96.10**	**94.20**	**83.10**	**-**
DSFD (ResNet152)	**94.29**	**91.47**	**71.39**	**120.06**
YOLO5face (YOLOv5s)	**94.33**	**92.61**	**83.15**	**7.075**
Proposed Model	**93.84**	**92.06**	**83.55**	**3.570**

Visualisation Results

This section presents the visualisation results to validate the effectiveness of the proposed model. Visual detection outcomes were assessed on the wider face dataset using the baseline and YOLO5face (YOLOv5s) models, and the proposed model. Notably, the dataset contained multi-face, occluded-face, and blurred-face images. In the detection results in Figure 6, the first column displays the baseline detection outcomes, the second column shows results from the YOLO5face model, and the third column presents results from the proposed model.

1. Results for dense multi-face images

As can be seen from Figure 6, the proposed model achieved satisfactory results for the wider face dataset. In Figure 6(A), a visual comparison between YOLO5face and the proposed model shows that the YOLO5face algorithm had a false detection, which is marked by an orange "×". The proposed model avoided the false detection problem in the case of dense multi-face detection. In addition, it improved the detection confidence of occluded faces.

Figure 6(B) shows denser faces. The proposed algorithm still achieved satisfactory detection results for multi-face detection. Although the confidence was slightly lower than that of YOLO5face, the proposed model accurately detected small-occluded faces that were missed by the baseline and YOLO5face models. The missed face is marked with a red circle. This shows that when dealing with multi-face detection tasks, the proposed model maintained excellent performance and robustness.

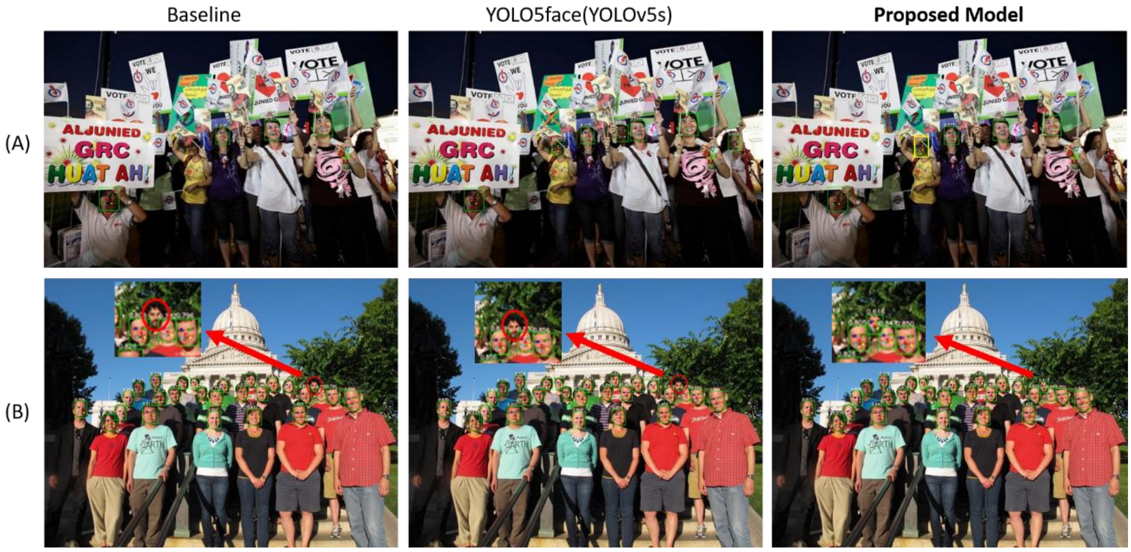

Figure 4 Dense multi-face image results.

2. Results for occluded face images

The two photographs in Figure 7 primarily include occluded face features. Blurred and occluded faces in the two groups of pictures are marked in yellow boxes. As shown in Figure 7(A) and 7(B), the confidence of the proposed model in detecting blurred faces was better than that of the baseline and YOLO5face (YOLOv5s) models. This result proves that the proposed model had a positive effect on the detection of occluded faces. This proves that the proposed model adapted to face detection challenges under various complex conditions.

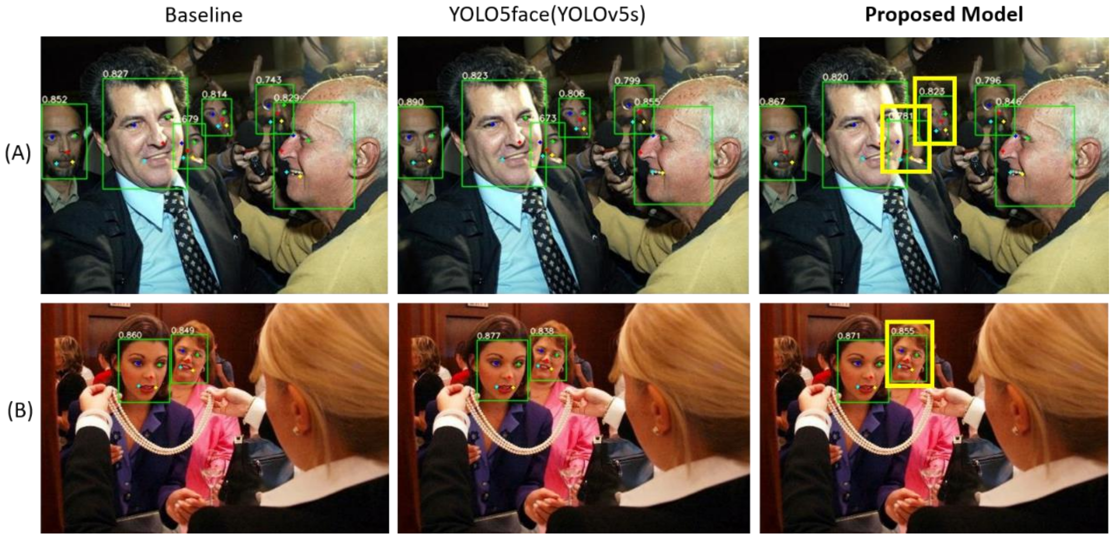

Figure 5 Occluded face image results.

3. Results for blurred face images

The two photographs in Figure 8 were mainly used to test the performance of blurred face detection. In Figure 8 (A), the confidence of the proposed model for the blurred faces marked by the yellow frames was higher than that of the baseline and YOLO5face (YOLOv5s) models. In Figure 8 (B), the baseline model only detected nine faces, while the proposed model and YOLO5face detected 13 faces. Faces missed at the baseline model are marked with red circles. In addition, for the detection of blurred small-scale faces, the detection confidence of the proposed model was better than that of YOLO5face. Blurred faces are marked with yellow boxes.

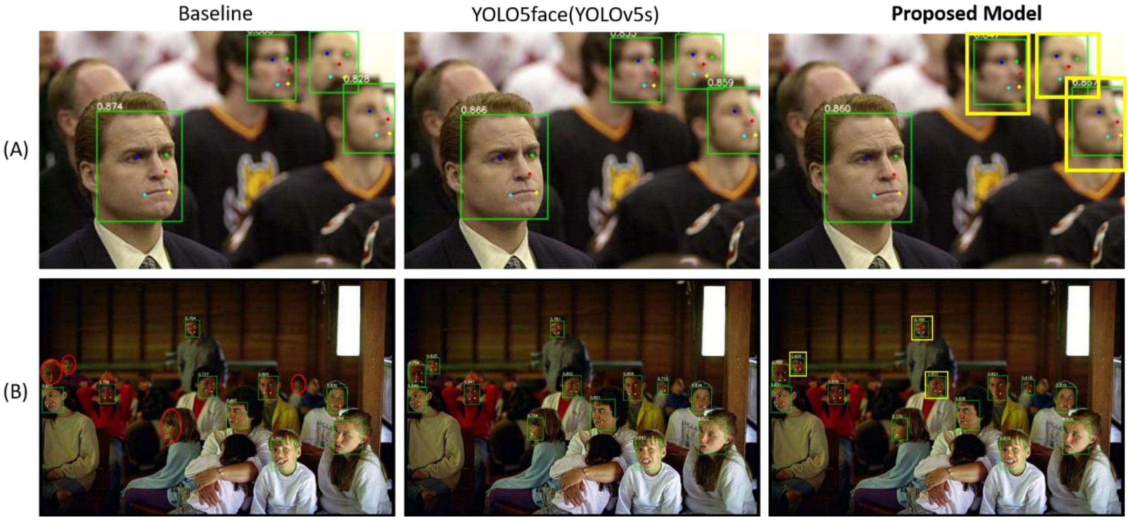

Figure 6 Blurred small-scale face image results.

Through the experimental visualisation comparison, the proposed model not only achieved the positioning of key points of facial features but also validated the adaptability to the challenges of face detection under various complex conditions.

Real-time performance

Real-time testing is typically used to verify the response speed and processing capability of a model in practical applications to ensure that the model can complete tasks within a specified time and satisfy real-time performance requirements. This is particularly critical in tasks such as face detection, where real-time performance is crucial.

As listed in Table 4, the proposed model was compared with the mainstream YOLO models. The proposed model took 26.79 ms to process a single image, which was 5.12 ms faster than YOLO5face (YOLOv5s) and 0.16 ms faster than YOLOv7-

tiny. This further validates that the proposed model improves the detection accuracy while ensuring lightweight design and real-time performance. Although the number of parameters of the proposed model is slightly higher than that of the baseline, YOLO5face (YOLOv5n), and YOLO5face (YOLOv5n-0.5), the real-time performance of the model still maintained a favorable level. This is due to the residual structure in the MC4f module. The existence of this structure allowed running deeper networks without increasing the gradient, so the calculation efficiency was higher while the parameters increase.

Table 4 Comparison of the performance and detection speed of detection models.

Model	Detection Speed (ms）	Params(M)
YOLOv5n(baseline)	20.10	1.886
YOLO5face (YOLOv5n-0.5)	26.11	0.45
YOLOv7-tiny	26.95	13.20
YOLOv8n-face	26.07	-
YOLO5face (YOLOv5n)	24.93	1.73
YOLO5face (YOLOv5s)	31.91	7.06
Proposed Model	26.79	3.57

Limitations and future work

The proposed model addressed the shortcomings of conventional algorithms in detecting occluded faces and indistinct small-scale faces. However, its detection ability was slightly insufficient under the following conditions: (1) Under extremely low-light or high-noise conditions, the face details were lost, making it difficult for the model to extract effective facial features. (2) When the target was in an extreme pose for face detection, false detection and missed detection occurred. To address these issues, we will conduct research from the following two aspects in future work: (1) By jointly training the super-resolution network and the face detector, the detection accuracy of the model can be improved in low-light and high-noise scenes. This will effectively solve the problem of loss of facial details.(2) By training a data augmentation model based on 3D face reconstruction, the limited 2D face data is converted into 3D representation. This approach will enhance the diversity of training samples by incorporating extreme pose variations, thereby improving the robustness of face detection under challenging pose conditions.

Conclusion

In this study, a lightweight adaptive fusion model was presented for detecting small-scale faces. First, the backbone of YOLOv5n was reconstructed using MC4f to address various problems that lightweight and deep linear networks may encounter, such as exploding or disappearing gradients. Second, the AESVC module adaptively integrated features from multiple regions of an image, thereby improving the expressive ability of the model in complex tasks. Third, the MSPAM was proposed which aimed to improve the integration of facial features across different scales while concurrently minimising the degradation of superficial features. Combining wing loss with α-Ciou achieved facial key point localisation, more accurately aligned facial images, effectively balanced easy and difficult samples, and improved the localisation accuracy of small-scale faces. Experiments showed that the proposed model improved the mAP on the hard subset while maintaining a lightweight model. This solved the problem of the missed detection of occluded faces and blurry small-scale faces that exist in traditional algorithms.

In the field of intelligent security monitoring, the ability of security systems to identify people with obstructions and distant blurred faces improves social security. In addition, in the field of driver monitoring, particularly when the face of a driver is obstructed or blurred owing to low light, rain, or other factors, maintaining high-precision detection can effectively ensure the safety of public transportation and drivers. However, in extreme environments, the robustness of detection requires improvement. In the future, image enhancement techniques could be utilised to improve image quality before initiating the primary detection phase, adapt to missing feature extraction in extreme scenes, and improve applicability across various environments.

Compliance with ethics guidelines

The authors declare they have no conflict of interest or financial conflicts to disclose.

This article contains no studies with human or animal subjects performed by authors.

References

Alansari, M., Hay, O. A., Javed, S., Shoufan, A., Zweiri, Y., & Werghi, N. (2023). Ghostfacenets: Lightweight face recognition model from cheap operations. IEEE Access, 11, 35429-35446.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8, 1-74.

Ben, X., Ren, Y., Zhang, J., Wang, S.-J., Kpalma, K., Meng, W., & Liu, Y.-J. (2021). Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 5826-5846.

Cardona-Pineda, D. S., Ceballos-Arias, J. C., Torres-Marulanda, J. E., Mejia-Muoz, M. A., & Boada, A. (2022). Face Recognition-Eigenfaces, 373-397.

Chen, W., Huang, H., Peng, S., Zhou, C., & Zhang, C. (2021). YOLO-face: a real-time face detector. The Visual Computer, 37, 805-813.

Chitraningrum, N., Banowati, L., Herdiana, D., Mulyati, B., Sakti, I., Fudholi, A., Saputra, H., Farishi, S., Muchtar, K., & Andria, A. (2024). Comparison Study of Corn Leaf Disease Detection based on Deep Learning YOLO-v5 and YOLO-v8. Journal of Engineering and Technological Sciences, 56(1), 61-70.

Debbouche, N., Ouannas, A., Batiha, I. M., Grassi, G., Kaabar, M. K. A., Jahanshahi, H., Aly, A. A., & Aljuaid, A. M. (2021). Chaotic Behavior Analysis of a New Incommensurate Fractional-Order Hopfield Neural Network System. Complexity(Pt.31), 2021.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2019). Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641.

Feng, Z.-H., Kittler, J., Awais, M., Huber, P., & Wu, X.-J. (2018). Wing loss for robust facial landmark localisation with convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2235-2245.

Freitas, R. T., Aires, K. R. T., Paiva, A. C., Rodrigo, D. M. S. V., & Soares, P. L. M. (2024). A CNN-based Multi-Level Face Alignment Approach for Mitigating Demographic Bias In Clinical Populations. Computational Statistics, 39(5).

Gao, S., Wu, R., Wang, X., Liu, J., Li, Q., & Tang, X. (2023). EFR-CSTP: Encryption for face recognition based on the chaos and semi-tensor product theory. Information Sciences, 621, 766-781.

Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., & Han, Z. (2021). Effective fusion factor in FPN for tiny object detection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1160-1168.

Guo, J., Deng, J., Lattas, A., & Zafeiriou, S. (2021). Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714.

Hannuksela, J. (2022). Facial Feature Based Head Tracking and Pose Estimation. Department of Electrical & Information Engineering, 7(2), 122.

He, J., Song, X., Feng, Z., Xu, T., Wu, X., & Kittler, J. (2023). ETM-face: effective training sample selection and multi-scale feature learning for face detection. Multimedia Tools and Applications, 82(17), 26595-26611.

Hioual, A., Ouannas, A., Oussaeif, T. E., Grassi, G., Batiha, I., & Momani, S. (2022). On Variable-Order Fractional Discrete Neural Networks: Solvability and Stability. Fractal and Fractional, 6(2), 119.

Hioual, A., Oussaeif, T. E., Ouannas, A., Grassi, G., Batiha, I. M., & Momani, S. (2022). New results for the stability of fractional-order discrete-time neural networks. Alexandria Engineering Journal, 61(12), 10359-10369.

Imran, A., Ahmed, R., Hasan, M. M., Ahmed, M. H. U., Azad, A. K. M., & Alyami, S. A. (2024). FaceEngine: A Tracking-Based Framework for Real-Time Face Recognition in Video Surveillance System. SN Computer Science, 5(5), 609.

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. Procedia computer science, 199, 1066-1073.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Wong, C., Yifu, Z., & Montes, D. (2022). ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. Zenodo.

Kobylkov, D., & Vallortigara, G. (2024). Face detection mechanisms: Nature vs. nurture. Frontiers in Neuroscience, 18, 1404174.

Li, H., Zhao, Y., Mao, Z., Qin, Y., Xiao, Z., Feng, J., Gu, Y., Ju, W., Luo, X., & Zhang, M. (2024). A survey on graph neural networks in intelligent transportation systems. arXiv preprint arXiv:2401.00713.

Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., & Huang, F. (2019). DSFD: dual shot face detector. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5060-5069.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 8759-8768.

Liu, X., Qi, P., Siarry, P., Hua, D., Ma, Z., Guo, X., Kochan, O., & Li, Z. (2023). Mining security assessment in an underground environment using a novel face recognition method with improved multiscale neural network. Alexandria Engineering Journal, 80, 217-228.

Liu, X., Zhang, S., Hu, J., & Mao, P. (2024). ResRetinaFace: an efficient face detection network based on RetinaFace and residual structure. Journal of Electronic Imaging, 33(4).

Liu, Y., Wang, F., Deng, J., Zhou, Z., Sun, B., & Li, H. (2022). Mogface: Towards a deeper appreciation on face detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 4093-4102.

Liu, Z., Ou, J., Huo, W., Yan, Y., & Li, T. (2022). Multiple feature fusion ‐ based video face tracking for IoT big data. International Journal of Intelligent Systems, 37(12), 10650-10669.

Ma, J., Li, X., Li, J., Wan, J., Liu, T., & Li, G. (2024). Quality-aware face alignment using high-resolution spatial dependencies. Multimedia Tools & Applications, 83(14).

Naseri, R. A. S., Kurnaz, A., & Farhan, H. M. (2023). Optimized face detector-based intelligent face mask detection model in IoT using deep learning approach. Applied Soft Computing, 134, 109933.

Qi, D., Tan, W., Yao, Q., & Liu, J. (2022). YOLO5Face: Why reinventing a face detector. European Conference on Computer Vision, 228-244.

Qin, Z., Zhang, P., Wu, F., & Li, X. (2021). Fcanet: Frequency channel attention networks. Proceedings of the IEEE/CVF international Conference on Computer Vision, 783-792.

Quan, Y., Zhang, D., Zhang, L., & Tang, J. (2023). Centralized feature pyramid for object detection. IEEE Transactions on Image Processing, 32, 4341-4354.

Rahmad, C., Asmara, R. A., Putra, D., Dharma, I., Darmono, H., & Muhiqqin, I. (2020). Comparison of Viola-Jones Haar Cascade classifier and histogram of oriented gradients (HOG) for face detection. IOP Conference Series: Materials Science and Engineering, 012038.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28.

Saadabadi, M. S. E., Malakshan, S. R., Dabouei, A., & Nasrabadi, N. M. (2024). ARoFace: Alignment Robustness to Improve Low-Quality Face Recognition. European Conference on Computer Vision, 308-327.

Sharma, D. (2021). Information Measure Computation and its Impact in MI COCO Dataset. 2021 7th International Conference on Advanced Computing and Communication Systems, 1964-1969.

Tootell, R. B., Hadjikhani, N., Hall, E. K., Marrett, S., Vanduffel, W., Vaughan, J. T., & Dale, A. M. (1998). The retinotopy of visual spatial attention. Neuron, 21(6), 1409-1422.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 390-391.

Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14923-14932.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. Proceedings of the European Conference on Computer Vision, 3-19.

Wu, Y., & He, K. (2018). Group normalization. Proceedings of the European conference on computer vision, 3-19.

Xiang, J., & Zhu, G. (2017). Joint face detection and facial expression recognition with MTCNN. 2017 4th International Conference on Information Science and Control Engineering, 424-427.

Xie, X., Cheng, G., Wang, J., Yao, X., & Han, J. (2021). Oriented R-CNN for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 3520-3529.

Xu, H., Wang, L., & Chen, F. (2024). Advancements in Electric Vehicle PCB Inspection: Application of Multi-Scale CBAM, Partial Convolution, and NWD Loss in YOLOv5. World Electric Vehicle Journal, 15(1), 15.

Yang, L., Zhang, R.-Y., Li, L., & Xie, X. (2021). Simam: A simple, parameter-free attention module for convolutional neural networks. International Conference on Machine Learning, 11863-11874.

Yashunin, D., Baydasov, T., & Vlasov, R. (2020). MaskFace: multi-task face and landmark detector. arXiv preprint arXiv:2005.09412.

Yu, Z., Huang, H., Chen, W., Su, Y., Liu, Y., & Wang, X. (2022). Yolo-facev2: A scale and occlusion aware face detector. arXiv preprint arXiv:2208.02019.

Zhang, B., Li, J., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Xia, Y., Pei, W., & Ji, R. (2020). Asfd: Automatic and scalable face detector. arXiv preprint arXiv:2003.11228.

Zhang, S., Chi, C., Lei, Z., & Li, S. Z. (2020). Refineface: Refinement neural network for high performance face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 4008-4020.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017a). Faceboxes: A CPU real-time face detector with high accuracy. 2017 IEEE International Joint Conference on Biometrics, 1-9.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017b). S3fd: Single shot scale-invariant face detector. Proceedings of the IEEE International Conference on Computer Vision, 192-201.

Zhu, Y., Cai, H., Zhang, S., Wang, C., & Xiong, Y. (2020). Tinaface: Strong but simple baseline for face detection. arXiv preprint arXiv:2011.1318.

Manuscript Received: 27 June 2024
1st Manuscript Revision Received: 27 January 2025
2nd Manuscript Revision Received: 7 March 2025
3rd Manuscript Revision Received: 16 April 2025
Accepted Manuscript: 21 May 2025