



International Journal of Computer Applications in Technology

ISSN online: 1741-5047 - ISSN print: 0952-8091

<https://www.inderscience.com/ijcat>

Research on multi-scale face detection based on graph embedded Swin-Transformer

Fang Wang, Huang Zhong

DOI: [10.1504/IJCAT.2025.10072633](https://doi.org/10.1504/IJCAT.2025.10072633)

Article History:

Received:	27 August 2023
Last revised:	11 May 2024
Accepted:	14 June 2024
Published online:	27 August 2025

Research on multi-scale face detection based on graph embedded Swin-Transformer

Fang Wang*

School of Intelligent Manufacturing,
Wuxi Vocational College of Science and Technology,
Wuxi, Jiangsu, China
Email: fangzihpu@163.com

*Corresponding author

Huang Zhong

Suzhou Triwin EST Co., Ltd.,
Suzhou, Jiangsu, China
Email: Matthew.zhong@triwincorp.com

Abstract: The existing Swin-Transformer has achieved great success in the field of object detection due to its lightweight, hierarchical and efficient long-distance dependency modelling advantages. However, in complex scenes, the impact of scale changes seriously restricts the performance of face detection. Therefore, this article proposes a multi-scale face detection network based on graph embedding Swin-Transformer. This network fully utilises multi-scale features and contextual information, effectively improving the performance of multi-scale face detection. Specifically, we designed a multi-feature fusion module based on Swin Transformer, which effectively integrates deep features of advanced information with shallow texture features. In addition, to effectively model the contextual information of the target, we use embedding graph convolution operations. We conducted extensive comparative experiments on publicly available data sets, and the experimental results showed that the proposed model achieved higher recognition performance.

Keywords: multi-scale face detection; Swin-Transformer; graph convolution embedded; multi-scale feature fusion.

Reference to this paper should be made as follows: Wang, F. and Zhong, H. (2025) 'Research on multi-scale face detection based on graph embedded Swin-Transformer', *Int. J. Computer Applications in Technology*, Vol. 76, Nos. 1/2, pp.94–105.

Biographical notes: Fang Wang received the doctoral degree from Nanjing University of Science and Technology, China. Now, she works in Wuxi Vocational College of Science and Technology. Her research interests include pattern recognition, control engineering etc.

Huang Zhong received the bachelor's degree from Tsinghua University, China. Now, he works in Suzhou Triwin EST Co., Ltd. His research interests include precision instrument.

1 Introduction

With the rapid development of information technology and artificial intelligence, facial-related applications have important applications in daily life and public safety, such as facial fraud detection, travel security checks and mobile payment. The accuracy of facial detection directly affects the effectiveness of these facial-related applications, which makes improving the accuracy of facial detection of great significance (Bartlett et al., 2003; Hjeltnæs and Kee, 2001). The artificial intelligence and machine learning technologies have made automatic face detection and recognition become possible, such as classification or regression for

recognition (Zhu et al., 2016a; 2017b), set classification for video (Gao et al., 2024), domain adaptation for across domain recognition (Liu et al., 2024), novelty detection for identification (Zhu et al., 2023, 2016b). The development of facial detection technology has three stages, based on template matching algorithm, AdaBoost algorithm, and deep learning algorithm. The template matching algorithm uses the facial image as a template, and the facial position is determined by algorithm matching with various positions of the detected image (Rowley et al., 1998). The AdaBoost algorithm is constructed as a strong classifier with high accuracy by combining multiple weak classifiers which may obtain a lower accuracy (Freund and Schapire, 1997).

Moreover, deep convolutional neural networks can learn more robust high-order features of images and achieve higher recognition accuracy in facial detection tasks (Li et al., 2015; Zhang et al., 2016). Recently, an increasing number of new convolutional networks have made significant progress in face detection tasks and gradually become the main algorithm in the field of face detection (Mamieva et al., 2023; Hangaragi et al., 2023).

In the process of facial detection, contextual information related to the face, such as the head, shoulders, body and other parts, is closely related. Enhancing the receptive field of the model detection area can fully mine contextual information, effectively used for small-scale facial auxiliary detection. Hu and Ramanan (2017) trained multi-scale detection templates and fully utilised contextual information to assist the detector in identifying small-scale faces. By expanding the receptive field and utilising parts such as the head and shoulders to assist in improving small-scale facial detection, the training steps of this method are relatively complex and require significant time and space costs. Li et al. (2019) proposed the DSFD algorithm which satisfies the requirements of different step sizes, anchor boxes and receptive field sizes by designing reasonable anchors. Although these models have improved performance in small-scale face detection, the receptive field of convolution operations can only model local contextual information and cannot mine long-distance or global contextual information. In recent years, Transformer has been successfully applied in the field of natural language processing, and it has also been introduced into the field of computer vision. Previous studies have shown that in image processing, compared to convolutional operations, the Transformer's dynamic attention mechanism can capture global contextual information and has strong long-term dependency modelling capabilities. Therefore, the Transformer model has been introduced into various tasks in computer vision and exhibits competitive or even better performance (Khan et al., 2022). In the face detection task, Wen et al. (2022) proposed an encoder-only DETR for face detection without reducing the accuracy of detection, dubbed EOPSA-FACE. To compromise between accuracy and speed, Li et al. (2022) proposed an effective deep model named MobileViT which combines Mobilenetv2 with Vision Transformer to propose a balanced sheep face recognition model. Although these models have achieved good recognition accuracy in face detection tasks, their computational complexity is high and their performance in small-scale face detection needs to be improved. In addition, the Transformer model focuses on long-distance dependencies while neglecting the modelling of local relationships. Therefore, a more efficient Transformer model needs to be proposed for multi-scale face detection.

In addition to using multi-scale information, using shallow feature maps is beneficial for small-scale face detection. Shallow feature maps contain sufficient detailed information, and as the number of layers in this information network increases, the detailed information will gradually be weakened and the effective features will gradually decrease. It is crucial to fully utilise shallow features in small-scale object detection (Naseri et al., 2023). Zhang et al. (2017b)

proposed a scale-invariant face detection network (S3FD), which first extracted features using the VGG16 network, and then fully utilised features of different scales to achieve multi-scale face detection. In order to sufficiently integrate feature maps at the same scale, Tang et al. (2018) proposed pyramid box to enhance the network's feature expression ability through semantic enhancement and feature fusion modules between different layers of the network. Based on S3FD, a feature pyramid method was proposed to supplement the semantic and spatial information in deep and shallow feature maps. In order to further improve the performance of small-scale face detection, Ranjan et al. (2019) proposed a context anchor head ResNet (CAHR), which uses a plug and play context module and emphasises the auxiliary role of context information in small-scale face detection. Therefore, how to efficiently use contextual information and multi-scale has become a research hotspot in the task of multi-scale facial recognition.

To this end, we propose an efficient deep model named graph embedded Swin-Transformer for multi-scale face detection. To model long-rang dependencies with fewer model parameters, we adopt Swin-Transformer module. The Swin-Transformer module adopts a window-based hierarchical structure to address two main challenges in the Transformer architecture: the scaling problem of high-resolution images and the high computational complexity problem. Unlike the ViT which uses fixed-size feature maps, Swin-Transformer builds feature maps by gradually merging adjacent patches from small to large. By utilising hierarchical feature maps, traditional multi-scale prediction can be utilised to overcome scaling problems. In addition, the Swin-Transformer introduces a non-overlapping window partitioning operation. Moreover, to explore the local relationships, we embedded graph convolution into Swin-Transformer module. The graph convolution operation can effectively learn the potential low-dimensional vector representations of graph nodes while maximising the preservation of network topology and other information. After learning new node vector representations, machine learning tools can be used to facilitate and effectively perform various tasks. Simultaneously, graph convolutional networks can model local dependencies with fewer network parameters. Finally, we adopt the feature pyramid multi-scale feature integration mechanism to supplement the semantic and spatial information in the deep and shallow feature maps. The entire model fully utilises contextual information and the multi-scale feature fusion mechanism. To verify the effectiveness of the model, we used the Widerface training set for training, the Widerface validation set for ablation experiments, and the Widerface test set, Fddb, MAFA and RMFD test sets for testing. The experiments proved that the algorithm in this chapter achieved excellent results in multiple data sets.

The remainder of this paper is organised as follows: the related work is summarised in Section 2; different modules in the proposed model are introduced in Section 3; the experimental comparison and analysis are reported in Section 3.5; the last Section 4 is the conclusion.

2 Related works

2.1 Deep learning-based face detection

The traditional methods use features that cannot extract high-level semantic features leading to insufficient generalisation ability and high training difficulty. However, deep learning can extract higher-quality features and improve the accuracy of face detection with end-to-end training (Viola and Jones, 2001; Yang et al., 2014). Li et al. (2015) first combined traditional techniques with convolutional neural networks to propose the cascade CNN algorithm for face detection. This algorithm includes multiple classifiers, which use the features of convolutional networks as input and adopt a cascaded structure. Finally, Non-Maximum Suppression (NMS) is used to merge highly overlapping windows to improve the accuracy of face detection. The most representative face detection method is the multi-task cascaded convolutional network MTCNN (Zhang et al., 2016), which trains multiple scale templates in a cascaded manner to match multi-scale faces. At the same time, by exploring facial landmark information, the results of facial detection are further corrected. However, the high complexity of this method leads to high training difficulty and insufficient real-time detection. Howard et al. (2017) designed a lightweight network structure called MobileNets, which mainly consists of multiple 3×3 , 1×1 convolutional kernel and traditional convolutional kernels. The model greatly reduces computational complexity while ensuring the convolution effect, and improves detection efficiency. In addition, the detection module of the model integrates multiple convolutional features, which can obtain more contextual information and a larger receptive field, thereby improving detection performance. Owing to the advantages of the YOLO network (Shafiee et al., 2017) in real-time object detection, Zhang et al. (2017a) proposed a real-time face detection algorithm (Faceboxes). This algorithm designs fast convolutional layers and multi-scale convolutional layers to extract image features and use the extracted features for multi-scale detection. The detector is fast and can be embedded into mobile devices. Recently, face detection based on deep networks also faces the challenge of further optimising algorithm complexity and computing resource occupation while maintaining high accuracy to meet the edge computing requirements for industrial scenes, security and personal user terminals.

2.2 Transformer-based object detection

Since its proposal in 2017, Transformer and its variants of network structures, such as BERT and GPT-2, have been widely used in the field of natural language processing. Transformer relies on a simple but powerful attention mechanism, allowing the model to selectively focus on certain parts of the input, making inference more efficient. In object detection tasks, the first Transformer-based object detection framework, detection transformer (DETR), outperformed Faster R-CNN in detecting large targets on the COCO object detection data set, but did not perform as well

in detecting small targets as the latter. The self-attention mechanism may ignore the structural information and local relationships in each patch. To alleviate this issue, exploring convolution operations with self-attention mechanism can effectively model the local and global dependencies. Chen et al. (2022) presented a parallel design of MobileNet and transformer, where the former is used for local information processing and the latter is used for global relationship modelling. Peng et al. (2021) proposed a hybrid network deep model named Conformer, in which the convolution operations are introduced for local feature extraction and the self-attention mechanisms are explored for enhanced representation learning. Li et al. (2021) further investigated how to effectively introduce local relationship modelling mechanisms into the Transformer model. The above model effectively improves the performance of object detection by introducing the advantages of convolution operations and complementing the advantages of the Transformer model. However, these models did not consider structural relationships, and they did not propose a better solution for detecting small targets. In order to improve the performance of small object detection, we embed graph convolution operations in the Transformer model. On the one hand, we can model local dependencies, and on the other hand, graph convolution can mine the structural information of the target. In addition, we introduce a multi-scale feature fusion mechanism to effectively improve the performance of small target detection.

2.3 Multi-scale face detection

Convolutional neural networks have achieved great success in face detection, but in complex scenarios such as occlusion, pose changes and large lighting changes, they pose greater challenges to deep models. Among them, how to effectively detect faces with large-scale changes in scale has become a research hotspot. In order to resolve this issue, more and more researchers are proposing new network frameworks and learning mechanisms. The first commonly used strategy is to use feature pyramids to improve the utilisation of multi-scale features. Wang et al. (2017) proposed the Face Attention Network (FAN) (Ranjan et al., 2015), which uses a feature pyramid structure and enhances existing features at each detection layer using an attention network to improve the efficiency of utilising existing effective features. Hu and Ramanan (2017) proposed the FTF algorithm to capture large-scale facial scale changes by processing image pyramids in a scale-invariant manner. They also used a specific scale hybrid detector for facial data distribution to thoroughly study the effective method of using image pyramids to detect minimal targets and facial scale changes. Li et al. (2019) proposed the detection network named Dual Shot Face Detector (DSFD), which uses a feature enhancement module to obtain enhanced feature maps and original features to form dual path features, thereby obtaining stronger feature expression capabilities. Through reasonable anchor design, the step length, anchor box size and receptive field size do not need to meet the principle of proportional intervals. The second strategy is to build a deeper and larger network,

utilising feature maps at different levels to achieve scale invariance. For example, Single Shot Multi-Box Detector (SSD) (Yang et al., 2019) uses VGG16 as the backbone network, while You Only Look Once (YOLO) uses a similar backbone network with more layers. Although this solution performs well in the proportion of a single input image, it requires a larger model size and high computational complexity. The third strategy is to fully explore contextual information. Zhang et al. (2019) proposed the S3FD network. In each detection stage, anchor frames with a single ratio and the same aspect ratio are used, and a wide range of anchor frame intervals is designed for the detection of difficult faces. This is done by designing an anchor frame size based on the effective receptive field involved. In order to improve multi-scale face detection, this article constructs a new feature extraction framework by using Swin-Transformer and graph convolution operations. This framework can effectively improve overall and local morality.

3 The proposed method

3.1 Motivation

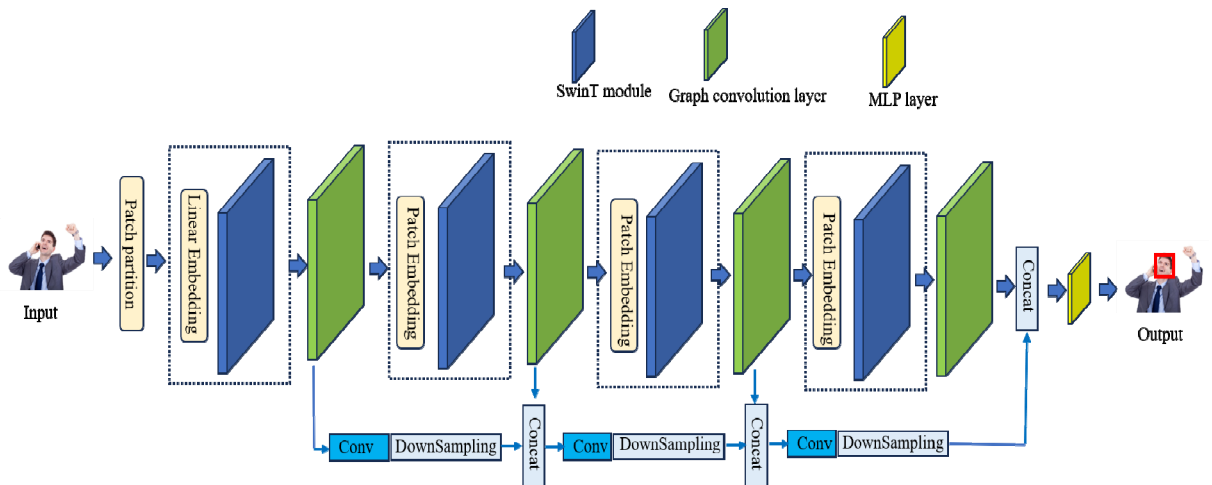
The existing Swin-Transformer model can model long-distance dependencies, effectively alleviating the local receptive field of traditional convolution operations. In small-scale object detection tasks, local contextual information mining has a positive impact on performance improvement. Therefore, combining traditional convolution operations with Transformers can improve the performance of multi-scale object detection, but convolution operations cannot model the structural information of the object. In addition, the graph convolutional network fully utilises the dependency

relationships of neighbouring nodes for feature learning, while modelling the structural relationships of nodes. This article implements local and full dependency modelling by embedding graph convolution into existing Swin-Transformer models. In addition, we will fuse multi-scale features to fully explore the advantages of shallow networks in extracting object detail information. The framework of the proposed model is shown in Figure 1. The proposed model consists of three modules: Swin-Transformer encoding layer, graph convolutional layer and multi-scale feature fusion.

3.2 Swin-Transformer module

Convolutional neural networks have achieved great success in the field of object detection. In convolutional networks, the low layer contains rich texture feature information of the object. However, improving the representation ability of object semantic features, can lead to the loss of small-scale object information by adding a convolutional layer. The Transformer fully mines the long-distance dependencies of features by using multi-head attention mechanisms, it gradually achieves better performance in object detection. However, the existing Transformer models have a huge computational cost and do not have multi-scale prediction capabilities. To alleviate the above two issues, the Swin-Transformer network architecture was proposed. Swin-Transformer introduces a window attention mechanism to greatly reduce computational complexity while achieving global relationship modelling. In addition, by introducing feature fusion and downsampling of the extracted features, the receptive field is increased, thereby improving the extraction of multiple multi-scale features. In this paper, we adopt the Swin-Transformer as a backbone for face feature extraction.

Figure 1 Overview of the proposed framework, in which we fuse features from different Swin-Transformer layers to improve the performance of face detection at different scales. In addition, to improve the performance of small-scale face detection, we introduce graph convolution operation for modelling local dependencies of targets (see online version for colours)



Firstly, we introduce the multi-head attention mechanism. The self-attention module can models the long-range dependencies of different embedding tokens. Specifically, given the input X , the Query Q , Key K and Value V matrices are defined as follows:

$$\begin{aligned} Q &= XW^Q \\ K &= XW^K \\ V &= XW^V \end{aligned} \quad (1)$$

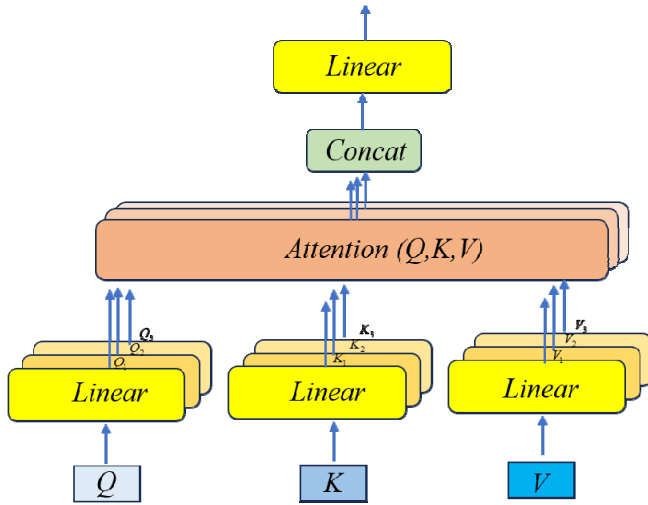
where W^Q, W^K, W^V denote learnable parameter matrices to obtain embedded tokens.

Then, Query Q and Key K are used to calculate the attention matrix, and finally the V combine the attention matrix to produce the final result. The attention matrix is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2)$$

where d is a scaling factor that is used to maintain the stability of gradients during model optimisation. The value of d is the dimension of V , and $\text{softmax}()$ denotes the softmax normalisation function. To explore richer semantic features in different subspace, a multi-head self-attention mechanism can be designed by using multiple one-head self-attention modules.

Figure 2 Overview of the MSA (see online version for colours)



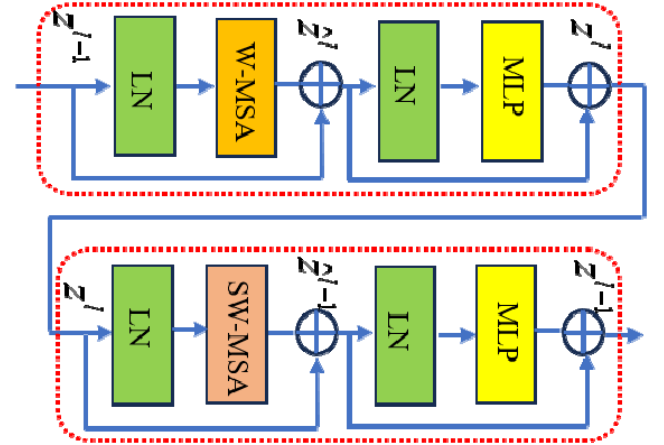
The Multi-head Self-Attention (MSA) is calculated as follows:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \\ \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W_M & \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (3)$$

where W_M, W_i^Q, W_i^K, W_i^V are learnable parameter matrices. The entire process which can be taken as one-head attention mechanism which is shown in Figure 2.

Swin-Transformer introduces the window-MSA (W-MSA) mechanism and shifted-window MSA (SW-MSA) mechanism to reduce the computational complexity and increase the receptive field for multi-scale feature extraction. Swin-Transformer replaces the standard MSA module in the Transformer block with W-MSA and SW-MSA, leaving the other layers unchanged. The standard Transformer architecture calculates the relationship between one token and other tokens, resulting in a quadratic complexity in the number of tokens, making it unsuitable for dense prediction tasks or high-resolution image visual tasks. Therefore, in order to efficiently model, W-MSA is calculated within the local window. The W-MSA module lacks connections between windows, which limits its modelling ability. In order to maintain efficient computation of non-overlapping windows while introducing cross-window connections, SW-MSA is used alternately in continuous swing transformer blocks. Combining the above two operations constitutes the basic Swin-Transformer module, which can propose efficient object feature extraction with fewer parameters. The basic framework of Swin-Transformer is shown in the following Figure 3.

Figure 3 Overview of the basic Swin-Transformer module (see online version for colours)



From the above calculation process in Figure 2, it can be understood that the attention mechanism essentially modulates V through the similarity value between QK . The higher the similarity, the stronger the modulation of V , while the lower the similarity, the weaker the modulation of V . Swin-Transformer in Figure 3 proposes a method for computing attention based on Windows, which is actually more like a training strategy than a disruptive algorithm. This strategy greatly reduces the computational complexity of attention on images, and also designs a hierarchical structure similar to CNN, making it the backbone network for visual tasks.

3.3 Graph convolution module

Although the SwinT module can model long-distance dependencies, the contextual information of the object plays

an important role in small-scale object detection. The existing solution is to combine convolutional operations with MSA, fully utilising the local modelling ability of convolutional operations. However, convolution operations cannot model the structural information of the object context. Therefore, we embed graph convolution operations in the SwinT module. Graph convolution can fully mine the contextual information of the object, while also modelling the structural information of the object. In addition, the graph convolution operation can perform feature extraction with fewer parameters.

The graph is composed of a set of nodes and edges, which can be represented by the symbol $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V}$ represents the set of nodes, $\mathcal{E}$ represents the set of edges and A represents the adjacency matrix. If there is a certain relationship between any two nodes v_i and v_j in graph $\mathcal{G}$, an edge $e_{ij} = (v_i, v_j)$ can be defined between them, where A is a binary matrix and is defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $A_{ij} = A_{ji}$.

Given feature H^l , the output of graph convolution H^{l+1} is defined as follows:

$$H^{l+1} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l \Theta \quad (5)$$

where C is the channel of feature H , $\Theta \in \mathbb{R}^{C \times F}$ is learnable parameters; F is the number of filters; $\tilde{A} = I_N + A$, $\tilde{D} = \sum_j \tilde{A}_{i,j}$, $D_{ii} = \sum_j A_{ij}$, I_N denote identity matrix. By introducing graph convolution operations, we can fully model the contextual information of facial features.

3.4 Multi-scale feature fusion

Traditional convolutional neural networks extract target features through layer-by-layer abstraction. The receptive field of high-level networks is relatively large, with strong semantic information representation ability, which can help us accurately detect or segment targets. However, the resolution of feature maps is low, and the representation ability of geometric information is weak (lacking spatial geometric feature details); The receptive field of low-level networks is relatively small, and their ability to represent geometric details is strong. Although they have high resolution, their ability to represent semantic information is weak. In the task of object detection, small-scale object detection requires the use of feature receptive fields. For example, in YOLOv3 (Mao et al., 2019), the understanding of multi-scale detection is that 1/32 size feature maps (deep layers) have high downsampling multiples, making them suitable for detecting large objects. 1/8 size feature maps (shallow layers) have smaller receptive fields, making them suitable for detecting small objects. The most commonly used

multi-scale feature fusion mechanism is the feature pyramid network, which is a type of multi-scale representation of images and an effective but conceptually simple structure that interprets images with multiple resolutions. A pyramid of an image is a collection of images arranged in a pyramid shape with gradually decreasing resolution, originating from the same original image. It is obtained by cascading down sampling until a certain termination condition is reached before stopping sampling. We compare layer-by-layer images to the higher the pyramid level, the smaller the image and the lower the resolution. In the multi-scale face detection task, we borrowed the feature pyramid network fusion mechanism and constructed a multi-scale feature fusion network structure based on Swin-Transformer.

3.5 Experiments and analysis

In this section, we conducted extensive comparative experiments on two commonly used face detection data sets. In addition, we conducted ablation studies to verify the effectiveness of different modules.

3.6 Face detection data sets and experimental settings

To verify the effectiveness of our proposed model, we conducted extensive comparative experiments on two data sets, Widerface (Yang et al., 2016) and Face Detection Data set and Benchmark (FDDB) (Jain and Learned-Miller, 2010). The Widerface data set includes 32,203 images and 393,703 faces. All facial images were collected from 60 real-life scenarios, including a large number of small-scale faces, dense faces, occluded faces, etc. From the perspective of all facial statistics, nearly 79% of faces are less than 40 pixels, of which 16% are less than 10 pixels. Owing to the abundance of small-scale face samples, this data set is often used as a training and testing data set for small-scale faces. In order to further validate the performance of different algorithms for multi-scale face detection, the data set was divided into three subsets, such as easy, medium and hard. In the easy set, most faces are close to the camera, with less occlusion and sufficient lighting, making face detection more difficult. In the hard set, most scenes are crowded with insufficient lighting, and human faces are not only small in scale but also subject to varying degrees of occlusion. Therefore, the hard set is more challenging in face detection tasks. Compared to Widerface, the face detection task on the FDDB data set is relatively simple. This data set contains 2845 images, with a total of 5171 faces. This data set mainly collects frontal faces to measure the effectiveness of different algorithms on basic face detection. Some samples on the two data sets are shown in Figure 4.

In this section, we conducted extensive comparative experiments on two commonly used face detection data sets. In addition, we conducted ablation studies to verify the effectiveness of different modules.

Figure 4 Some samples of two face detection data sets (see online version for colours)

3.7 Face detection data sets and experimental settings

To verify the effectiveness of our proposed model, we conducted extensive comparative experiments on two data sets, Widerface (Yang et al., 2016) and Face Detection Data set and Benchmark (Fddb) (Jain and Learned-Miller, 2010). The Widerface data set includes 32,203 images and 393,703 faces. All facial images were collected from 60 real-life scenarios, including a large number of small-scale faces, dense faces, occluded faces, etc. From the perspective of all facial statistics, nearly 79% of faces are less than 40 pixels, of which 16% are less than 10 pixels. Owing to the abundance of small-scale face samples, this data set is often used as a training and testing data set for small-scale faces. In order to further validate the performance of different algorithms for multi-scale face detection, the data set was divided into three subsets, such as easy, medium and hard. In the easy set, most faces are close to the camera, with less occlusion and sufficient lighting, making face detection more difficult. In the hard set, most scenes are crowded with insufficient lighting, and human faces are not only small in scale but also subject to varying degrees of occlusion. Therefore, the hard set is more challenging in face detection tasks. Compared to Widerface, the face detection task on the Fddb data set is relatively simple. This data set contains 2845 images, with a total of 5171 faces. This dataset mainly collects frontal faces to measure the effectiveness of different algorithms on basic face detection. Some samples on the two data sets are shown in Figure 4.

This experiment code is conducted Python framework. We use the Adam optimiser to optimise the model parameters. Other super parameters are selected as follows: the drop out of the full connection layer is 0.1, the Learning

rate is $1e-4$ and the batch size is 16. The number of Swin embedded modules is 3. We embed two graph convolutional layers after each Transformer encoder, and the input and output channels of graph convolutional operations are the same. Our experiment was conducted on a Windows 10 system, using Anaconda to build the deep network operating environment. We adopt the Pytorch programming and cuda10 and cudnn11 to accelerate computation. On the hardware side, we use NVIDIA GeForce RTX 3090 with a memory size of 24 GB.

3.8 Results on two data set

To verify the effectiveness of the model proposed, we compared it with different methods on the Widerface data set. The results of comparison methods are shown in Table 1.

Table 1 Accuracy of different models on Widerface data set

<i>Models</i>	<i>Easy (%)</i>	<i>Medium (%)</i>	<i>Hard (%)</i>
ACF (Wang et al., 2015)	64.2	52.6	25.2
LDCF+ (Zhu, 2018)	79.7	77.2	56.4
CMSRCNN (Wang et al., 2017a)	90.2	87.4	64.3
ScaleFace (Stone et al., 2010)	86.7	86.6	76.4
MSCNN (Wang et al., 2021)	91.7	90.03	80.9
SSH (Najibi et al., 2017)	92.7	91.5	84.4
CAHR (Wu et al., 2019)	92.1	90.1	83.2
VGG16SSH (Zhang et al., 2019)	92.1	91.8	84.5
HoiFace (Sun et al., 2018)	94.1	92.4	82.9
Ours	95.2	93.9	86.8

In the Widerface test set, we compare the proposed model with other classic face detection methods, including traditional methods and deep learning methods. We compared three subsets, where the easy subset has a relatively simple face detection task, while the hard subset has a large number of small-scale faces, which can be used to verify the model's ability to detect small-scale faces. From the entire experimental results, it can be seen that models based on deep learning have achieved relatively high performance on all subsets. The traditional machine learning algorithm ACF achieved the lowest performance in three indicators. However, there are differences in the performance of different deep models for face detection at different scales, for example, the accuracy of SSH in small-scale face detection tasks is 84.4%, while the accuracy of HoiFace is only 82.9%. However, in large-scale face detection, HoiFace achieves higher accuracy. The model extracted in this article achieved the highest performance on all three facial scales.

3.9 Results of Fddb data set

On the Widerface data set, we compared our proposed method with other methods in three types of evaluating indicators. In this section, we further validate the effectiveness of our proposed method on the Fddb data set. The task of facial detection on this data set is relatively easy. There is not much variation in facial scale. The results of the entire comparative experiment are shown in Table 2.

Table 2 Accuracy of different models on Fddb data set

<i>Models</i>	<i>DisROC (%)</i>	<i>ROC (%)</i>
FastRCNN (Li et al., 2017)	90.0	61.5
Mxnet (Chen et al., 2015)	96.8	72.2
FasterRCNN (Ren et al., 2015)	96.1	72.4
LDCF+ (Ma and Gao, 2018)	92.1	68.0
Conv3D (Li et al., 2016)	90.1	75.9
MTCNN (Zhang et al., 2016)	86.3	66.8
CCF (Ranjan et al., 2015)	85.9	61.4
CascadeCNN (Li et al., 2017)	85.7	66.8
MultiResHpm (Yang et al., 2015)	85.1	74.7
ODFD (Ghiasi and Fowlkes, 2015)	84.0	66.9
Ours	97.5	85.2

Table 2 presents the results of different methods on the Fddb data set in terms of DisROC and ROC. The two evaluation indicators represent the accuracy under the discrete ROC curve and the continuous ROC curve. It can be seen that under the discrete and continuous curves, the performance of different depth models in conventional face detection tasks is compared. By using two metrics to compare the performance of different algorithms, we can comprehensively analyse the advantages of different models. For example, Mxnet achieves relatively high performance on DisROC, but the performance of ROC is not the best. Our proposed model achieves the highest performance in both indicators.

In summary, our proposed model can construct new feature extraction modules by integrating Transformer and graph convolution operations. This module can model local and global dependencies, enabling it to capture valuable features. Moreover, we introduce multi-scale feature fusion mechanism for a multi-scale face detection recognition.

3.10 Effectiveness of different modules

In our proposed model, we introduce the graph convolution and multi-scale feature fusion modules into the Swin-Transformer framework. In this experiment, we compared the effectiveness of different modules, i.e., graph convolution and multi-scale fusion modules. We conduct all experiments on Widerface and Fddb data sets. The results on both datasets are shown in the Tables 3 and 4.

Table 3 Accuracy of different models on Widerface data set

<i>Models</i>	<i>Easy (%)</i>	<i>Medium (%)</i>	<i>Hard (%)</i>
Transformer	93.2	91.4	93.8
Swin-Transformer	93.8	92.1	85.4
Swin-Transformer+Graph	94.8	93.2	86.2
Swin-Transformer+Multifeature	94.5	93.5	86.0
Ours	95.2	93.9	86.8

Table 4 Accuracy of different models on Fddb data set

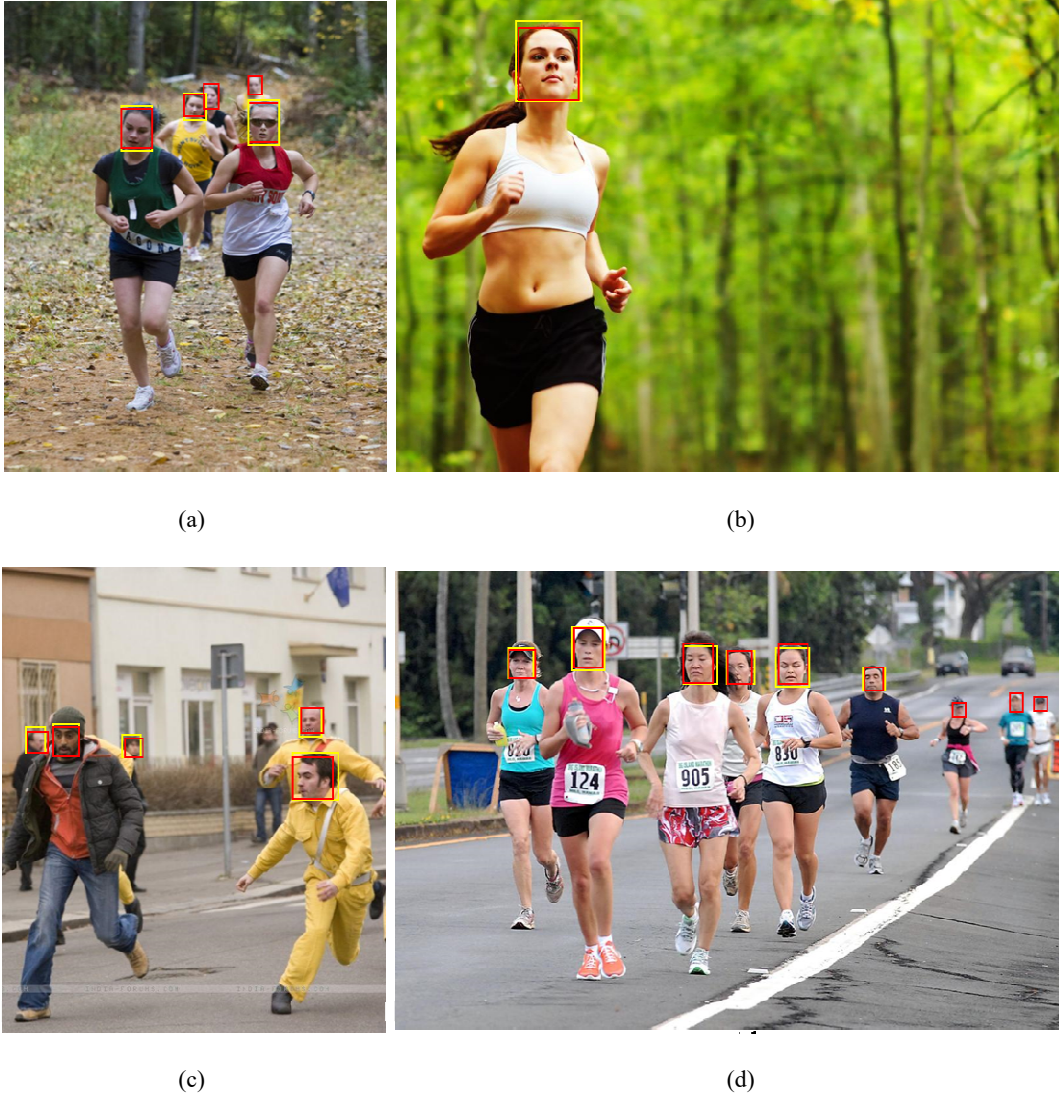
<i>Models</i>	<i>DisROC (%)</i>	<i>ROC (%)</i>
Transformer	95.2	82.4
Swin-Transformer	96.1	83.6
Swin-Transformer+Graph	96.9	84.8
Swin-Transformer+Multifeature	97.1	84.4
Ours	97.5	85.2

We design the baseline: Swin-Transformer. By adding different modules, we construct different models, Swin-Transformer+Graph, Swin-Transformer+Multifeature and Ours. We conduct all experiments on Widerface and Fddb data sets. As the results are shown in Tables 3 and 4, by introducing graph convolution and scale feature fusion modules, the performance can be effectively improved.

3.11 Visualisation results of different methods

To further validate the effectiveness of the proposed model, we compared the visualisation results of the two algorithms on the Widerface test set. To compare the results of multi-scale face detection using different models, we selected different scales of face images from the Widerface test data set, and then used our proposed model and the SSH model for face detection tasks, respectively. The entire result is shown in Figure 5.

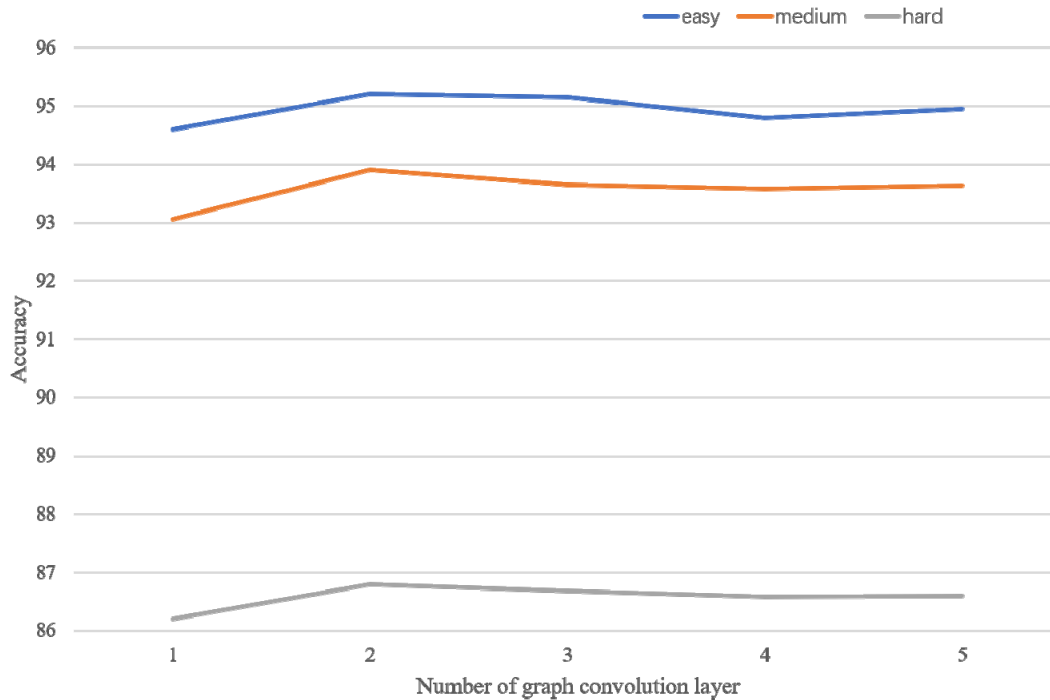
Figure 5 Visualisation results on Widerface data sets. We compared the face detection performance of two deep models at different scales. The red box represents the proposed model detection results, while the yellow box represents the SSH detection results (see online version for colours)



In Figure 5, the red box represents the detection results of our proposed model, and the yellow box represents the detection results of the SSH model. From the results of face detection, both algorithms are effective in detecting facial regions for large-scale faces. As the face scale decreases, SSH algorithms may miss detection, such as Figures 5(a), 5(c) and 5(d). However, our proposed model outperforms the SSH algorithm for small-scale face detection, such as in Figures 5(a) and 5(d). In addition, where there is significant background interference, our proposed model and the SSH model have missed detections, as shown in Figure 5(c).

3.12 Effectiveness of graph convolution layers

To verify the impact of different graph convolutional layers of our proposed model, we conducted comparative experiments on the Widerface datasets data set. We set the number of graph convolutional layers to vary from 1 to 5. All experiments are shown in Figure 6. All experimental results indicate that the model achieved the highest performance when the graph convolutional layer was set to 2. When the number of graph convolutional layers is increased, there is no significant improvement in the performance of the model, but the number of model parameters will increase rapidly.

Figure 6 Effectiveness of graph convolution layers on Widerface data set (see online version for colours)

4 Conclusion

In this paper, we design a novel and effective graph-embedded Swin-Transformer model for multi-scale face detection. To explore facial local relationships, we embed the graph convolution into Swin-Transformer which can model long-range relationships. By integrating the graph convolution and Swin-Transformer, we can effectively model local and global relationships for modelling different scale face object context information. Moreover, we introduce the multi-scale feature fusion mechanism in which the fusion of low-level texture features and high-level semantic features can improve small-scale face detection. To validate the effectiveness of the proposed model, we conducted extensive comparative experiments on two data sets. All results indicate that our proposed model achieved the highest performance in multi-scale face detection tasks. There are two main directions for future work: (1) with the rapid development of Internet of Things technology, the design and mobile deployment of lightweight and urgent deep networks have attracted widespread research interest; (2) with the popularisation of various sensors, the collection of multimodal signals has become more convenient. Therefore, designing a face detection model based on multimodal fusion will be a hot topic in the field of computer vision.

References

- Bartlett, M.S., Littlewort, G., Fasel, I. and Movellan, J.R. (2003) 'Real time face detection and facial expression recognition: development and applications to human computer interaction', *Conference on Computer Vision and Pattern Recognition Workshop*, IEEE, Vol. 5, pp.53–53.
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C. and Zhang, Z. (2015) 'Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems', *arXiv preprint arXiv:1512.01274*.
- Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L. and Liu, Z. (2022) 'Mobile-former: bridging mobilenet and transformer', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5270–5279.
- Freund, Y. and Schapire, R.E. (1997) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and SystemS*, Vol. 55, No. 1, pp.119–139.
- Gao, X., Niu, S., Wei, D., Liu, X., Wang, T., Zhu, F., Dong, J. and Sun, Q. (2024) 'Joint metric learning-based class-specific representation for image set classification', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 5, pp.6731–6745.
- Ghiasi, G. and Fowlkes, C.C. (2015) 'Occlusion coherence: detecting and localizing occluded faces', *arXiv preprint arXiv:1506.08347*.
- Hangaragi, S., Singh, T. and Neelima, N. (2023) 'Face detection and recognition using face mesh and deep neural network', *Procedia Computer Science*, Vol. 218, pp.741–749.

- Hjelmås, E. and Kee, B. (2001) 'Low face detection: a survey', *Computer Vision and Image Understanding*, Vol. 83, No. 3, pp.236–274.
- Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017) 'Mobilenets: efficient convolutional neural networks for mobile vision applications', *arXiv preprint arXiv:1704.04861*.
- Hu, P. and Ramanan, D. (2017) 'Finding tiny faces', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.951–959.
- Jain, V. and Learned-Miller, E. (2010) *Fddb: A Benchmark for Face Detection in Unconstrained Settings Technical Report*, UMass Amherst Technical Report.
- Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M. (2022) 'Transformers in vision: a survey', *ACM Computing Surveys (CSUR)*, Vol. 54, No. 10s, pp.1–41.
- Li, H., Lin, Z., Shen, X., Brandt, J. and Hua, G. (2015) 'A convolutional neural network cascade for face detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5325–5334.
- Li, J., Liang, X., Shen, S., Xu, T., Feng, J. and Yan, S. (2017) 'Scale-aware fast r-cnn for pedestrian detection', *IEEE transactions on Multimedia*, Vol. 20, No. 4, pp.985–996.
- Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J. and Huang, F. (2019) 'Dsfd: dual shot face detector', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5060–5069.
- Li, X., Du, J., Yang, J. and Li, S. (2022) 'When mobilenetv2 meets transformer: a balanced sheep face recognition model', *Agriculture*, Vol. 12, No. 8. Doi: 10.3390/agriculture12081126.
- Li, Y., Sun, B., Wu, T. and Wang, Y. (2016) 'Face detection with end-to-end integration of a convnet and a 3d model', *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp.420–436.
- Li, Y., Zhang, K., Cao, J., Timofte, R. and Van Gool, L. (2021) 'Localvit: bringing locality to vision transformers', *arXiv preprint arXiv:2104.05707*.
- Liu, Z., Wang, T., Zhu, F., Chen, X., Pelusi, D. and Vasilakos, A.V. (2024) 'Domain adaptive learning based on equilibrium distribution and dynamic subspace approximation', *Expert Systems with Applications*, Vol. 249. Doi: 10.1016/j.eswa.2024.123673.
- Ma, Z. and Gao, P-P. (2018) 'Research on the cascade pedestrian detection model based on LDEF and CNN', *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, IEEE, pp.314–320.
- Mamieva, D., Abdusalomov, A.B., Mukhiddinov, M. and Whangbo, T.K. (2023) 'Improved face detection method via learning small faces on hard images based on a deep learning approach', *Sensors*, Vol. 23, No. 1. Doi: 10.3390/s23010502.
- Mao, Q-C., Sun, H-M., Liu, Y-B. and Jia, R-S. (2019) 'Mini-yolov3: real-time object detector for embedded application', *IEEE Access*, 7, pp.133529–133538.
- Najibi, M., Samangouei, P., Chellappa, R. and Davis, L.S. (2017) 'Ssh: single stage headless face detector', *Proceedings of the IEEE International Conference on Computer Vision*, pp.4875–4884.
- Naseri, R.A.S., Kurnaz, A. and Farhan, H.M. (2023) 'Optimized face detector-based intelligent face mask detection model in iot using deep learning approach', *Applied Soft Computing*, Vol. 134:109933.
- Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J. and Ye, Q. (2021) 'Conformer: local features coupling global representations for visual recognition', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.367–376.
- Ranjan, R., Bansal, A., Zheng, J., Xu, H., Gleason, J., Lu, B., Nanduri, A., Chen, J-C., Castillo, C.D. and Chellappa, R. (2019) 'A fast and accurate system for face detection, identification, and verification', *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Vol. 1, No. 2, pp.82–96.
- Ranjan, R., Patel, V.M. and Chellappa, R. (2015) 'A deep pyramid deformable part model for face detection', *Proceedings of the IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, pp.1–8.
- Ren, S., He, K., Girshick, R. and Sun, J. (2015) 'Faster r-cnn: towards real-time object detection with region proposal networks', *Advances in Neural Information Processing Systems*, Vol. 28.
- Rowley, H.A., Baluja, S. and Kanade, T. (1998) 'Neural network-based face detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp.23–38.
- Shafiee, M.J., Chywl, B., Li, F. and Wong, A. (2017) 'Fast yolo: a fast you only look once system for real-time embedded object detection in video', *arXiv preprint arXiv:1709.05943*, 2017.
- Stone, Z., Zickler, T. and Darrell, T. (2010) 'Toward large-scale face recognition using social network context', *Proceedings of the IEEE*, Vol. 98, No. 8, pp.1408–1415.
- Sun, X., Wu, P. and Hoi, S.C.H. (2018) 'Face detection using deep learning: an improved faster rcnn approach', *Neurocomputing*, Vol. 299, pp.42–50.
- Tang, X., Du, D.K., He, Z. and Liu, J. (2018) 'Pyramidbox: a context-assisted single shot face detector', *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.797–813.
- Viola, P. and Jones, M. (2001) 'Rapid object detection using a boosted cascade of simple features', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Vol. 1, pp.I–I.
- Wang, H., Chen, X., Xu, B., Du, S. and Li, Y. (2021) 'An improved mscnn method for underwater image defogging', *Proceedings of the IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, IEEE, pp.296–302.
- Wang, J., Yuan, Y. and Yu, G. (2017) 'Face attention network: an effective face detector for the occluded faces', *arXiv preprint arXiv:1711.07246*, 2017.
- Wang, S., Yang, B., Lei, Z., Wan, J. and Li, S.Z. (2015) 'A convolutional neural network combined with aggregate channel feature for face detection', *Proceedings of the 6th International Conference on Wireless, Mobile and Multi-Media (ICWMMN'15)*, IET, pp.304–308.
- Wen, P., Sun, C., Zhang, S., Luo, Y., Huang, H. and Zhang, J. (2022) 'Eopsa-face: an encoder-only detr for classroom face detection', *Proceedings of the 10th International Conference on Information Systems and Computing Technology (ISCTech)*, IEEE, pp.341–348.
- Wu, T., Liang, D., Pan, J. and Kaneko, S. (2019) 'Context-anchors for hybrid resolution face detection', *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, pp.3297–3301.
- Yang, B., Yan, J., Lei, Z. and Li, S.Z. (2014) 'Aggregate channel features for multi-view face detection', *Proceedings of the IEEE International Joint Conference on Biometrics*, IEEE, pp.1–8.

- Yang, B., Yan, J., Lei, Z. and Li, S.Z. (2015) 'Convolutional channel features for pedestrian, face and edge detection', arXiv:1504.07339.
- Yang, F., Chen, H., Li, J., Li, F., Wang, L. and Yan, X. (2019) 'Single shot multibox detector with kalman filter for online pedestrian detection in video', *IEEE Access*, Vol. 7, pp.15478–15488.
- Yang, S., Luo, P., Loy, C.-C. and Tang, X. (2016) 'Wider face: a face detection benchmark', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5525–5533.
- Zhang, K., Guo, X., He, Y., Wang, X., Guo, Y. and Ding, Q. (2019) 'IMS-SSH: multiscale face detection method in unconstrained settings', *Journal of Electronic Imaging*, Vol. 28, No. 1, pp.013035–013035.
- Zhang, K., Zhang, Z., Li, Z. and Qiao, Y. (2016) 'Joint face detection and alignment using multitask cascaded convolutional networks', *IEEE Signal Processing Letters*, Vol. 23, No. 10, pp.1499–1503.
- Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X. and Li, S.Z. (2017a) 'Faceboxes: a cpu real-time face detector with high accuracy', *IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, pp.1–9.
- Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X. and Li, S.Z. (2017b) 'S3fd: single shot scale-invariant face detector', *Proceedings of the IEEE International Conference on Computer Vision*, pp.192–201.
- Zhu, C., Zheng, Y., Luu, K. and Savvides, M. (2017a) 'CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection', *Deep Learning for Biometrics*, pp.57–79.
- Zhu, F., Gao, J., Xu, C., Yang, J. and Tao, D. (2017b) 'On selecting effective patterns for fast support vector regression training', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 8, pp.3610–3622.
- Zhu, F., Yang, J., Gao, C., Xu, S., Ye, N. and Yin, T. (2016b) 'A weighted one-class support vector machine', *Neurocomputing*, Vol. 189, pp.1–10.
- Zhu, F., Yang, J., Gao, J. and Xu, C. (2016a) 'Extended nearest neighbor chain induced instance-weights for SVMs', *Pattern Recognition*, Vol. 60, pp.863–874.
- Zhu, F., Zhang, W., Chen, X., Gao, X. and Ye, N. (2023) 'Large margin distribution multi-class supervised novelty detection', *Expert Systems with Applications*, Vol. 224. Doi: 10.1016/j.eswa.2023.119937.