A Dual-Engine Embedded Face Detection and Recognition Framework Using YOLO5Face and Attention-Enhanced Faster-RCNN for Surveillance Video

Qianqian Yuan*, Yuping Quan, Hui Li
College of Science and Engineering, Jiaozuo Normal College, Jiaozuo 4540000, China
E-mail: yuan15036696545@163.com
*Corresponding author

Keywords: surveillance video, face recognition, YOLO5Face, faster regions with convolutional neural networks, embedded detection, security field

Received: August 15, 2025

Embedded detection and recognition systems for surveillance video are in urgent demand in the security field. However, traditional methods face limitations, including poor real-time performance, high resource consumption, and limited generalization in complex scenarios. To this end, this study proposes a dual-engine embedded face detection and recognition framework that optimizes performance by synergistically integrating YOLO v5Face with attention-enhanced Faster Regions with Convolutional Neural Network. The system adopts a dual engine cascade architecture: YOLO5Face is responsible for fast initial face screening, while Faster Regions with Convolutional Neural Network, which integrates spatial and channel attention mechanisms, accurately recognizes key targets. By synergistically optimizing speed and accuracy through feature reuse and structural fusion techniques, and by combining the feature-extraction capabilities of the local binary pattern histogram algorithm based on hierarchical feature pyramids, a dynamic background suppression module is used to reduce false positives in complex scenes. The experimental results on the WIDER FACE and Face Detection Data Set and Benchmark datasets show that the accuracy of our system reaches 99.1%, with a loss rate as low as 0.08, significantly better than the comparison systems Visual Transformer Convolutional Neural Network Fusion (accuracy 98.16±0.23%) and Additive Marginal Soft Maximum Loss Convolutional Multi-scale Transformer (accuracy 97.42±0.34%); The system converges to a loss of less than 0.1 within 200 iterations, with a response time of only 28 ms, much faster than the fusion of Visual Transformer Convolutional Neural Network (78-85 ms). The above results show that the proposed method effectively addresses the problems of poor real-time performance, resource constraints, and insufficient scene generalization, offering efficient, lightweight new ideas for system development and promoting the intelligent and efficient development of security terminals.

Povzetek: Predlagan je lahek, hiter sistem za zaznavanje in prepoznavanje obrazov v nadzornih videih, ki izboljša natančnost in delovanje v realnem času.

1 Introduction

In surveillance video, embedded detection and recognition systems serve as the core component of intelligent security, playing an irreplaceable role in public safety, smart cities, and restricted area control [1]. Facing increasingly complex security threats and real-time response demands, traditional detection and recognition solutions encounter issues such as high latency, privacy leaks, and insufficient generalization ability in complex scenarios [2]. Current research mainly focuses on lightweight convolutional neural network architecture compression and system pruning techniques. However, challenges remain, including unstable detection in dynamic video and difficulty balancing accuracy and speed under low-light conditions [3]. The YOLO5Face is an optimized face detection engine. It enhances multi-scale face localization capabilities through adaptive feature fusion and achieves stable inference on the embedded device, surpassing traditional face detectors [4]. Faster Regions with Convolutional Neural Networks (Faster-RCNN) leverages a two-stage design with a region proposal network that generates high-quality candidate boxes, providing stronger robustness against occlusion and small face targets [5]. Overall, there are currently research issues such as balancing real-time performance and accuracy, and how to achieve high-precision and low latency parallel processing for face detection and recognition; The second issue is the adaptability of complex scenarios, how to improve the robustness of recognition methods in adverse weather conditions; The third issue is the optimization of embedded deployment, which needs to meet the requirements of long-term reliable operation and efficient resource utilization of security systems. Therefore, this study builds an embedded detection and recognition system based on YOLO5Face, innovatively using a dual-engine cascade architecture. YOLO5Face performs fast front-end target screening in surveillance, while Faster-RCNN precisely recognizes key targets in video. The study also designs an embedded heterogeneous acceleration strategy that integrates illumination-invariant feature extraction and attention mechanisms to reduce false detections in complex environments. This work aims to achieve collaborative optimization of detection and recognition accuracy under complex scenarios on surveillance devices, meeting the high robustness requirements for edge deployment in the security field.

2 Related works

With the development of computer technology, surveillance video applications became increasingly widespread, and face detection and recognition technologies were applied in various fields. Zhang H et al. proposed a new privacy verification method for facial recognition, called Minimum Hypothesis Privacy Protection Verification (Map2V). This is the first exploration of using depth-image priors and zero-order gradient estimation to develop privacy verification methods. The experimental results and analysis demonstrate the effectiveness and generality of the proposed Map2v, demonstrating its superiority over local privacy verification methods in PPFR literature [6]. Hangaragi and Singh proposed a face-mesh-based detection and recognition system to address challenges posed by complex lighting, backgrounds, viewing angles, and multi-person scenarios. The system, trained on the LFW dataset and real-time images, recognized faces by matching facial feature points, achieving a recognition accuracy of 94.23% [7]. Liu et al. tackled privacy leakage issues in IoT face recognition by introducing an attribute-preserving de-identification framework composed of a triple network. Through nested autoencoder design, the framework removed identity information while retaining facial attributes. Experiments showed that the framework improved data utility by 26.22%, balancing attribute preservation and identity protection effectively [8]. Huang Q et al. designed a simple and effective method to train a facial recognition model via selective propagation and title-driven extension by gradually expanding the label set. They constructed a large-scale dataset of title images containing 6.3M faces across 305K themes. The results show that the above method can train state-of-the-art facial recognition models without manual annotation (99.65% in LFW) [9]. Raz D et al. aim to illustrate algorithmic bias in facial recognition technology, which is more likely to misidentify women and ethnic minorities. The results show that the proposed method can improve the transparency of the algorithm system and enhance public awareness of its different sources of influence [10].

As for embedded detection and recognition in surveillance videos, the YOLO series has matured over time. Scholars from many countries conducted thorough research and applied it in real life. Gupta et al. proposed a deep learning-based face recognition method for real-time detection of online learning engagement. By analyzing facial expressions to classify emotions, they computed engagement indices to predict participation status. After comparing multiple systems, their method showed the best performance in real-time scenarios with an accuracy of 92.3% [11]. Anusudha addressed the challenges of blurred and disguised face recognition by proposing the YOLO-InsightFace network, which enabled contactless, real-time identity verification by generating highly discriminative face embeddings. The method combined YOLO-V7's real-time detection ability with InsightFace's facial analysis technology, achieving high accuracy and speed [12]. Chen et al. proposed YOLO-face based on YOLO5 to address accuracy drops caused by face scale variations in YOLO-like single-stage detectors. Using more suitable anchor boxes and regression loss functions, the method outperformed YOLO and its variants on datasets such as WIDER FACE [13].

In summary, although significant progress has been made in the lightweight deployment of embedded face detection and recognition in existing research, the current technological roadmap still has obvious limitations. Firstly, although YOLO series single-stage detectors offer speed advantages, their accuracy and robustness for detecting small objects, severe occlusions, and unrestricted lighting conditions in complex scenes remain insufficient. Secondly, pure recognition models often rely on pre-detection steps and fail to achieve deep collaborative optimization of detection and recognition on embedded platforms, resulting in high system-level latency and resource consumption. Finally, most existing solutions balance speed and accuracy, lacking an architecture that dynamically adapts to scene complexity to achieve maximum efficiency and effectiveness at the system level. To address the aforementioned issues, a dual-engine cascaded architecture combining YOLO5Face and attention-enhanced Faster-RCNN has been proposed. The division of labor enables on-demand allocation of computing resources, avoiding the significant overhead of Faster-RCNN's comprehensive scanning of each image frame, thereby maintaining high accuracy while controlling overall latency. At the same time, by introducing the FPN-LBPH module, the semantic information of deep convolutional networks and the texture robustness of traditional LBPH operators are fused at the feature level, thereby improving the modeling ability for lighting changes and low-resolution faces. Furthermore, we aim to jointly address the three core challenges of real-time performance, resource constraints, and generalization to complex scenes in embedded environments from both algorithmic and deployment perspectives.

3 Embedded detection and recognition system based on YOLO5Face

3.1 YOLO5Face algorithm improved by FPN-LBPH

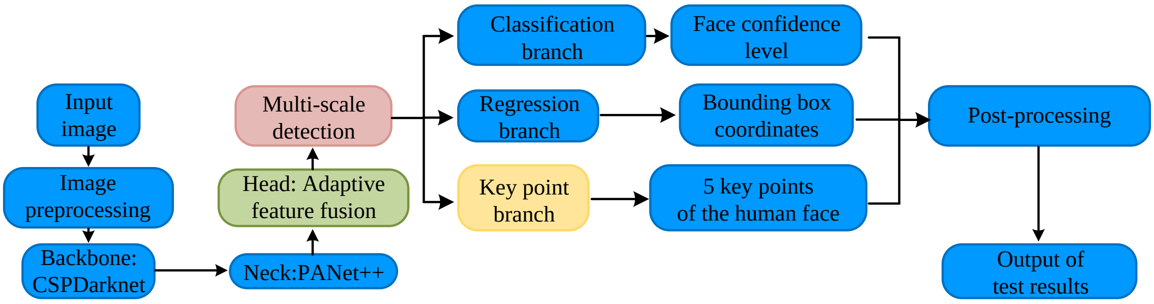


Figure 1: YOLO5Face algorithm flowchart.

Embedded detection and recognition systems in surveillance video are essential for intelligent security. These systems must perform real-time target detection and identity verification on edge devices. However, traditional technologies suffer from high latency in cloud-based solutions, significant privacy risks, and excessive computational load, which fail to meet the real-time requirements of embedded deployment. To address the limitations of traditional techniques, this study adopts YOLO5Face as the front-end engine for high-speed preliminary screening of face targets. It integrates a dynamic background suppression module and quantization technology to reduce false detections in complex scenes. The YOLO5Face algorithm flow is shown in Figure 1.

As shown in Figure 1, the complete processing flow of the YOLO5Face algorithm includes adaptive scaling and normalization preprocessing of input images, multi-scale feature extraction of the backbone network, a bidirectional feature fusion module, parallel output branches of classification/regression/keypoints, as well as confidence filtering and non-maximum suppression steps in the post-processing stage. The bounding box regression loss is calculated as shown in Equation (1).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(a, a^{gt})}{b^2} + \alpha v \tag{1}$$

In Equation (1), IoU represents the Intersection over Union between the predicted box and the ground truth, (a, a^{gt}) is the center coordinate of the predicted/actual box, and b is the diagonal length of the minimum enclosing box. ρ represents the Euclidean distance between the predicted box and the center point of the real box, α is the balance parameter, and v is a parameter used to measure aspect ratio consistency. The algorithm optimizes face box localization accuracy using this loss function. The refinement of bounding box keypoints is calculated as shown in Equation (2).

$$L_{wing} = \begin{cases} QIn(1+|n|/ò) & |n|<Q \\ |f|-C, & N \end{cases} \tag{2}$$

In Equation (2), n represents coordinate deviation, Q is the threshold, and ò is the smoothing factor. f and C are the model output values and compensation constants, respectively. Localization loss improves sensitivity for five keypoints. However, accurate keypoint detection requires support from multi-scale features. The adaptive fusion mechanism is defined in Equation (3).

$$R_{fusion} = \sum_{x=3}^{5} \beta_x R_x \tag{3}$$

In Equation (3), β represents learnable weights and R_x denotes feature maps at different scales. The algorithm enhances feature representation by adaptively fusing features through dynamic weighting. YOLO5Face still lacks precision when detecting small, occluded, or extreme-angle faces. To solve this, the study introduces a Feature Pyramid Network-Local Binary Patterns Histograms (FPN-LBPH) algorithm to improve multi-scale face localization. The reason for using the above combination in the study is that deep convolutional features are good at extracting high-level semantic information but are less sensitive to lighting changes and texture details; as a classic texture descriptor, LBPH has natural invariance to lighting changes and can effectively capture local texture features. By integrating LBPH features into the FPN architecture, the system can simultaneously utilize semantic information and texture features at multiple scales, achieving complementary advantages at the feature level. At the same time, there are significant scale changes in the faces in the monitoring scene, and FPN effectively solves the multi-scale detection problem through bidirectional paths from top to bottom and from bottom to top. Introducing LBPH features into this framework enables the fusion of texture information with semantic features at different scales. The FPN-LBPH algorithm, based on a hierarchical feature pyramid, is as follows: after inputting the three-layer resolution feature map generated by the backbone network, feature interactions are achieved via dual paths. The top-down path down samples deep features using 1×1 convolutions, up samples them by a factor of 2, and adds them to middle-level features. It then up samples the result using 3×3 convolutions and fuses it with shallow features, producing the enhanced P5 feature. The bottom-up path reversely down samples shallow features to fuse with middle-level features and further down samples to merge deep features, resulting in a semantically rich P3 feature. A hierarchical aggregation module then fuses these features into a unified multi-scale feature map. This dual-path, cross-layer information flow combines spatial details from shallow layers with semantic features from deep layers, significantly improving detection robustness. Feature dimensionality reduction is defined in Equation (4).

$$K_c{}' = Conv_{1\times1}(K_c; F_c), c \in \{3,4,5\} \tag{4}$$

In Equation (4), K_c is the hierarchical feature output from the backbone network, and $Conv_{1\times1}$ is the learnable 1×1 convolution weight matrix. This operation unifies the number of channels for all feature maps and reduces computation. The reduced features are fused via up sampling to propagate semantic information toward shallow layers, as shown in Equation (5).

$$F_k = Conv_{3\times3}(P_k' + \uparrow_2 \cdot (F_{k+1})), k = 4,3 \tag{5}$$

In Equation (5), $\uparrow_2 \cdot (F_{k+1})$ represents 2× bilinear up sampling, and F is the starting point of the path. P_k' is the feature map of the previous layer. Similarly, down sampling is used to enhance small target detection, as defined in Equation (6).

$$E_n = Conv_{3\times3}(P_N' + \downarrow_2 (E_{n-1})), n = 4,5 \tag{6}$$

In Equation (6), $\downarrow_2 (E_{n-1})$ represents stride-2 convolution down sampling, and E is the initial path point. This symmetric bidirectional path enables shallow spatial details to complement deep features. Mid-level features integrate outputs from both directions, balancing detail and semantic information, as shown in Equation (7) [14].

$$W = Conv_{1\times1}(\gamma_3 l_3 \oplus \gamma_4(l_4 + j_4) \oplus \gamma_5 j_5) \tag{7}$$

In Equation (7), W denotes channel concatenation, γ represents learnable channel attention weights, and $\gamma_4(l_4 + j_4)$ is the fusion result from both paths. The YOLO5Face algorithm incorporates FPN-LBPH to combine bidirectional path outputs via an attention mechanism, resulting in an optimal multi-scale feature combination. This approach dramatically reduces computational and parameter overhead, meeting millisecond-level real-time detection requirements [15]. The FPN-LBPH-YOLO5Face flowchart is shown in Figure 2.

As shown in Figure 2, the hybrid algorithm receives the surveillance video input. YOLO5Face performs real-time face detection. The detected faces are processed based on quality. High-quality faces are optimized through a texture enhancement branch, while low-quality or occluded faces undergo enhancement via FPN-LBPH and a super-resolution reconstruction branch. Both branches apply background suppression to highlight facial features. The algorithm fuses the enhanced features with those directly extracted from the original faces to form more robust facial representations. These features are matched with the face feature database. If a match is found, the recognition result is output. If not, the face is marked as "unknown," triggering an alert or logging an event. The complete process forms a closed-loop from detection and enhancement to recognition and response.

3.2 Embedded detection and recognition system combining YOLO5Face and Faster-RCNN

In surveillance face-detection scenarios, FPN-LBPH-YOLO5Face performs well in terms of speed and real-time response. However, it still shows limitations in precision and localization for small, densely occluded, or extremely illuminated faces. To address this, the study uses Faster-RCNN to generate high-quality region proposals, improving robustness and accuracy for low-quality, blurry, or small faces and compensating for the shortcomings of FPN-LBPH-YOLO5Face. The Faster-RCNN flowchart is shown in Figure 3.

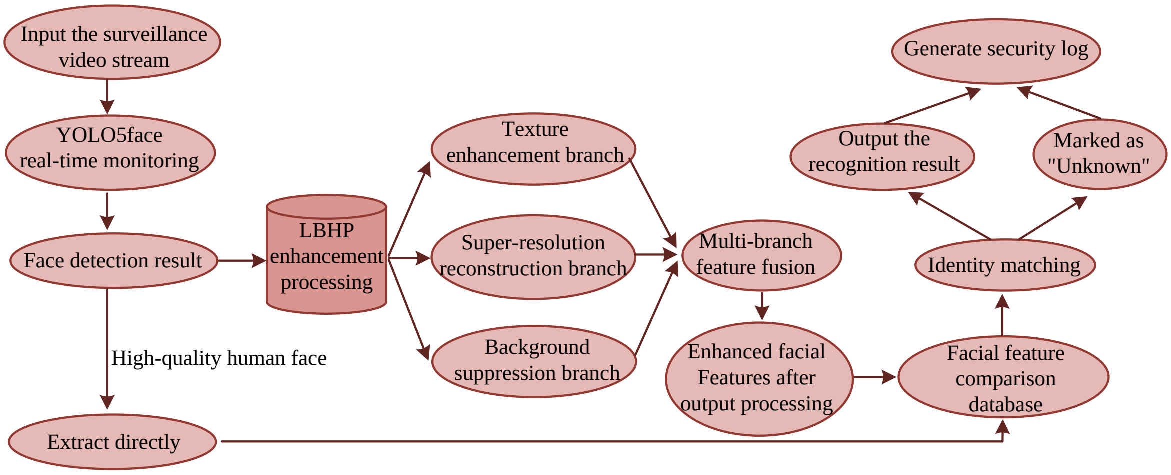

Input the surveillance video stream
YOLO5face real-time monitoring
Face detection result
LBHP enhancement processing
Texture enhancement branch
Super-resolution reconstruction branch
Background suppression branch
Multi-branch feature fusion
Enhanced facial Features after output processing
Facial feature comparison database
Identity matching
Output the recognition result
Marked as "Unknown"
Generate security log
High-quality human face
Extract directly

Figure 2: FPN-LBPH-YOLO5Face algorithm flowchart.

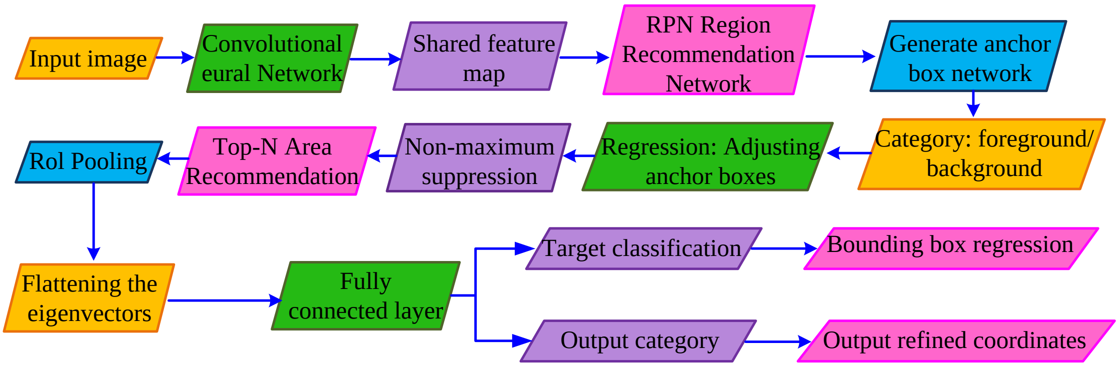

Input image
Convolutional eural Network
Shared feature map
RPN Region Recommendation Network
Generate anchor box network
Category: foreground/ background
Regression: Adjusting anchor boxes
Non-maximum suppression
Top-N Area Recommendation
RoI Pooling
Flattening the eigenvectors
Fully connected layer
Target classification
Bounding box regression
Output category
Output refined coordinates

Figure 3: Faster-RCNN algorithm flowchart.

As shown in Figure 3, the algorithm extracts shared feature maps from the input image using a convolutional neural network. The region proposal network generates a grid of multi-scale anchors. It performs binary classification (foreground/background) and preliminary coordinate regression. Non-maximum suppression filters out high-quality region proposals, which are then flattened and passed through a fully connected network. This network outputs target-class probabilities and refined bounding-box coordinates. The entire pipeline shares convolutional features to improve computational efficiency, while each stage collaborates to enhance detection accuracy. The anchor generation equation is defined in Equation (8) [16].

$$Anchor_{(x,y)} = (x\cdot j, y\cdot j, h_j\cdot ratio_u, w_j\cdot ratio_u) \quad (8)$$

In Equation (8), (x,y) represents the center coordinates of the feature map grid, j is the down sampling stride of the backbone, and $ratio_u$ represents the width-height ratio. Multiple anchors are generated at each feature map location. After anchor generation, classification loss filters candidate boxes with faces, as shown in Equation (9).

$$P_{cls}(L_i, L_i^*) = -log[L_i^*\cdot L_i + (1-L_i^*)(1-L_i)] \quad (9)$$

In Equation (9), L_i denotes the predicted probability that anchor i is foreground, and L_i^* is the ground truth label, distinguishing whether an anchor contains a face. The classification loss is further defined in Equation (10).

$$P_{reg}(y_i, y_i^*) = \sum\nolimits_{d\in\{x,y,j,h\}} smooth_{p1}(t_{ij} - t_{ij}^*) \quad (10)$$

In Equation (10), $P_{reg}(y_i, y_i^*)$ represents a regularization function, while t_{ij} and t_{ij}^* are the original and detected face images, respectively. Faster-RCNN is sensitive to irrelevant background and underuses features in complex scenes. Therefore, the study integrates an attention mechanism to enhance key features and suppress background interference, improving both detection accuracy and robustness. The optimized Faster-RCNN flowchart is shown in Figure 4.

As shown in Figure 4, the algorithm uses a base convolutional neural network to extract general features from the input face image. In the optimization step, spatial attention highlights target regions and reduces background noise. Channel attention strengthens important semantic features. The weighted feature maps are input into the region proposal network to generate more accurate proposals. The fully connected layers then produce classification and bounding-box results. This adaptive feature enhancement improves Faster-RCNN's robustness in detecting small or occluded faces. The channel attention weight is calculated as shown in Equation (11) [17].

$$N_c(X) = \sigma(G_2\delta(G_1 D(X) + c_1) + c_2) \quad (11)$$

In Equation (11), X and G are the width and height of the convolutional feature map, D is the global average pooling, and c is the fully connected layer. After channel optimization, spatial attention is directed to critical facial areas. The spatial attention weight is computed in Equation (12) [18].

$$W_S(U_d) = \sigma(t^{7\times 7}([U_d^{avg}; U_d^{max}])) \tag{12}$$

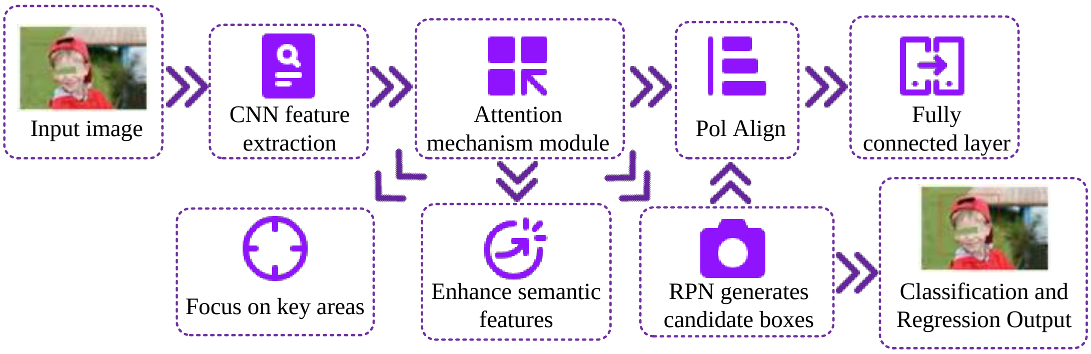

Input image
CNN feature extraction
Attention mechanism module
Pol Align
Fully connected layer
Focus on key areas
Enhance semantic features
RPN generates candidate boxes
Classification and Regression Output

Figure 4: Attention-enhanced Faster-RCNN algorithm flowchart.

In Equation (12), U_d^{avg} and U_d^{max} represent the channel-wise average and maximum features, and $t^{7\times 7}$ is a 7×7 convolution layer. The optimized features are passed into the RPN to generate more accurate face proposals, as shown in Equation (13).

$$L_{out} = M_s \otimes N_c \tag{13}$$

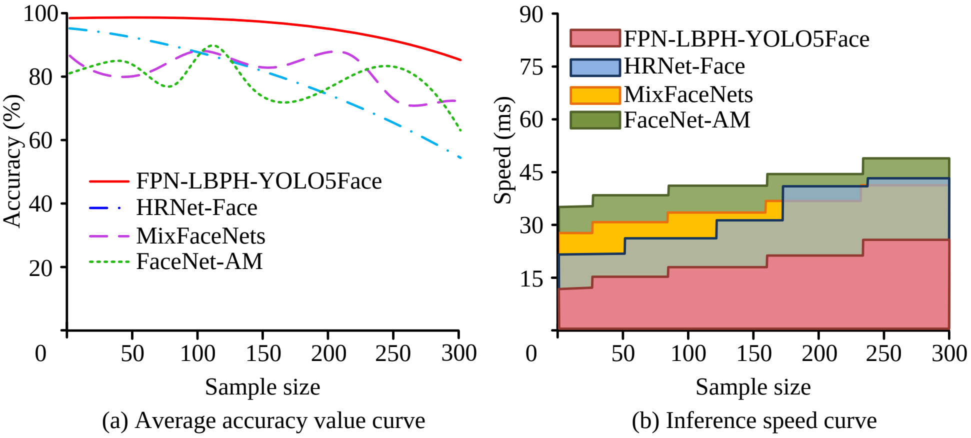
In Equation (13), $\otimes$ represents element-wise multiplication, and L_{out} produces enhanced responses at key face positions, reducing background interference. By enhancing feature robustness through FPN-LBPH and YOLO5Face, and improving accuracy with attention-optimized Faster-RCNN, the study proposes an embedded detection and recognition system named Faster-RCNN-YOLO5Face. This system aims to achieve both high accuracy and real-time performance. Based on the above content, the proposed system process is as follows. The input video stream is first processed by the improved YOLO5Face module, which extracts LBPH texture features and fuses them with the original frames to enhance robustness to lighting and low resolution. The initial facial proposal is passed through a keyframe selection mechanism to the attention optimized Faster-RCNN module, which weights facial features using spatial and channel attention, refines bounding boxes, and generates discriminative facial feature vectors, ultimately performing identity recognition matching and alarm triggering.

4 Performance analysis of embedded detection and recognition system

4.1 Validation of the Faster-RCNN–YOLO5Face Hybrid Algorithm

To verify the superiority of the FPN-LBPH-YOLO5Face hybrid algorithm, it was compared with High-Resolution Network for Face Recognition (HRNet-Face), Mixed Depthwise Face Recognition Networks (MixFaceNets), and FaceNet with Additive Margin (FaceNet-AM) algorithms. The experiments ran on a system with CUDA 11.6, Ubuntu 20.04 LTS, PyTorch 1.12.1, the SGD optimizer, Python 3.8, a NVIDIA GeForce RTX 3090 GPU, and 64GB of memory. To ensure the authenticity and reliability of the experiment, the WIDER FACE dataset (website source: http://shuoyang1213.me/WIDERFACE/index.html) and the FDDB dataset (website source: http://vis-www.cs.umass.edu/fddb/index.html) were used for training and testing. The WIDER FACE dataset contained 3,203 images and 3,903 annotated faces, while the FDDB dataset included 2,845 images and 3,171 annotated faces. The data preprocessing and enhancement steps are as follows: the input image is first adaptively scaled to 640 × 640 and normalized. Then perform data augmentation, including random horizontal flipping (probability 0.5), random rotation (angle range±15°), brightness adjustment (coefficient 0.8-1.2), and Mosaic enhancement (probability 0.8). The training, validation, and test sets are split in a 6:2:2 ratio to ensure consistent distribution across lighting, pose, and occlusion conditions. The experimental parameters are set as follows: SGD as the optimizer, momentum 0.937, and weight decay 0.0005. The initial learning rate is set to 0.01, with a cosine annealing scheduler that lowers the learning rate to 0.0001. Set the batch size to 32 and the training period to 300. The non-maximum suppression threshold is set to 0.5, and the confidence threshold is set to 0.4. For the FPN-LBPH module, the number of feature map channels is uniformly set to 256, using bilinear up sampling and 3x3 convolutional down sampling with a stride of 2. The channel attention weights in the attention mechanism are computed via global average pooling and two fully connected layers, with a compression ratio of 16. Spatial attention uses a 7 × 7 convolutional kernel to generate weight maps. The mean average precision and inference speed of the FPN-LBPH-YOLO5Face hybrid algorithm and comparison algorithms were evaluated. The results are shown in Figure 5.

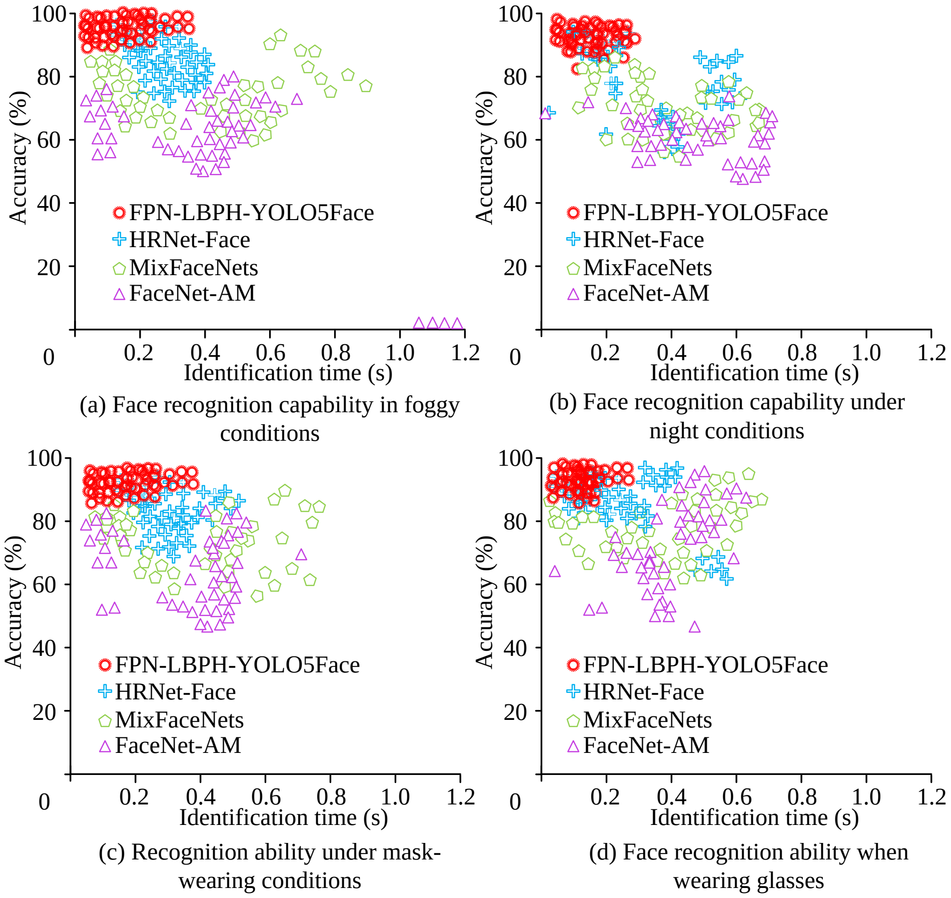
Figure 5 (a) shows that the FPN-LBPH-YOLO5Face hybrid algorithm achieves an average accuracy of 85.2% to 99.8%, significantly higher than HRNet Face, MixFaceNets, and FaceNet AM. Figure 5 (b) shows that the algorithm has the fastest inference speed, less than 30 ms, while FaceNet AM lags significantly due to its high system complexity. The experimental results demonstrated that the FPN-LBPH-YOLO5Face hybrid algorithm achieved the best precision and the fastest inference by reusing backbone feature maps, thereby reducing processing time. It balanced speed and accuracy through structural optimization while maintaining detailed face discrimination. To evaluate face recognition performance under extreme weather, nighttime, and other conditions, the FPN-LBPH-YOLO5Face hybrid algorithm was compared with other algorithms. Among them, the foggy condition test subset contains 1500 samples, the nighttime subset contains 1200 samples, the mask cover subset contains 1000 samples, and the glasses cover subset contains 800 samples. The above test subsets are all selected from two publicly available test datasets based on challenging conditions to obtain basic images. Results are shown in Figure 6.



(a) Average accuracy value curve

(b) Inference speed curve

Figure 5: Comparison of mean average precision and inference speed.



(a) Face recognition capability in foggy conditions

(b) Face recognition capability under night conditions

(c) Recognition ability under mask-wearing conditions

(d) Face recognition ability when wearing glasses

Figure 6: Face recognition results under fog, night, mask, and glasses conditions.

As shown in Figure 6(a), under foggy conditions, FPN-LBPH-YOLO5Face achieved the highest recognition accuracy, nearly 100%, with the shortest recognition time within 0.2 s. HRNet-Face and FaceNet-AM had similar accuracy between 85% and 92% but required longer times. As shown in Figures 6(b), (c), and (d), under night, mask-wearing, and glasses-wearing scenarios, FPN-LBPH-YOLO5Face consistently maintained over 95% accuracy and recognition times below 0.3 s. The high recognition rates reported above are due to the research being conducted on a specific subset of tests. These test subsets are composed of samples with corresponding challenging features selected from the original dataset, where the foggy effect is synthesized through atmospheric scattering models, mask and glasses occlusion is simulated using standard occlusion templates, and nighttime conditions are achieved through brightness adjustment. Although these conditions simulate real-life scenarios to some extent, their diversity is still limited by the basic dataset. Overall, by combining Faster-RCNN's candidate box generation and fine-detection mechanisms with region proposal networks, the system precisely localized faces and significantly improved recognition accuracy. To further investigate the robustness of the proposed method, a 5-fold cross-validation method was used, with the ratios of training, validation, and test sets strictly controlled at 6:2:2. The experiment was repeated 3 times, and the final results were averaged to reduce the impact of randomness. Simultaneously, use the Bootstrap sampling method (1000 repetitions) to calculate the 95% confidence interval, ensuring the reliability of the statistical results. The cross-validation results can be obtained from this, as shown in Table 1.

According to Table 1, the accuracy variance of the final system in 5-fold cross-validation is 0.15%, significantly lower than ViT CNN Fusion (0.34%) and ArcFace CMT (0.23%) (F=9.84, *p*=0.002), indicating that it is the least sensitive to changes in training data partitioning. Based on 1000 bootstrap samples, the 95% confidence interval for the final system accuracy is [98.82%, 99.38%], with an interval width of 0.56%, which is much smaller than that of the comparison model. To test the contribution of each component of the proposed method, ablation experiments were conducted: Model A used only YOLO5Face as the baseline; Model B introduced FPN-LBPH for optimization based on Model A; and Model C, based on Model B, incorporated Faster RCNN for improvement. Model D introduces an attention mechanism to form a complete final system. The results of the ablation experiment are shown in Table 2.

According to Table 2, Model B has an accuracy improvement of 1.4% (t=8.32, p=0.001) and a recall improvement of 1.9% (p<0.01) compared to benchmark Model A. The FPN-LBPH module significantly improves its detection capability for small targets and blurry faces while maintaining real-time performance through multi-scale feature fusion. Compared to Model B, Model C has further improved accuracy by 0.8% (p<0.01) and recall by 1.5% (p<0.01). The dual-engine architecture leverages high-quality candidate boxes and fine-grained recognition. The overall performance of the complete model is higher than that of other models, indicating improved stability.

4.2 Evaluation of the improved YOLO5 Face embedded detection and recognition system

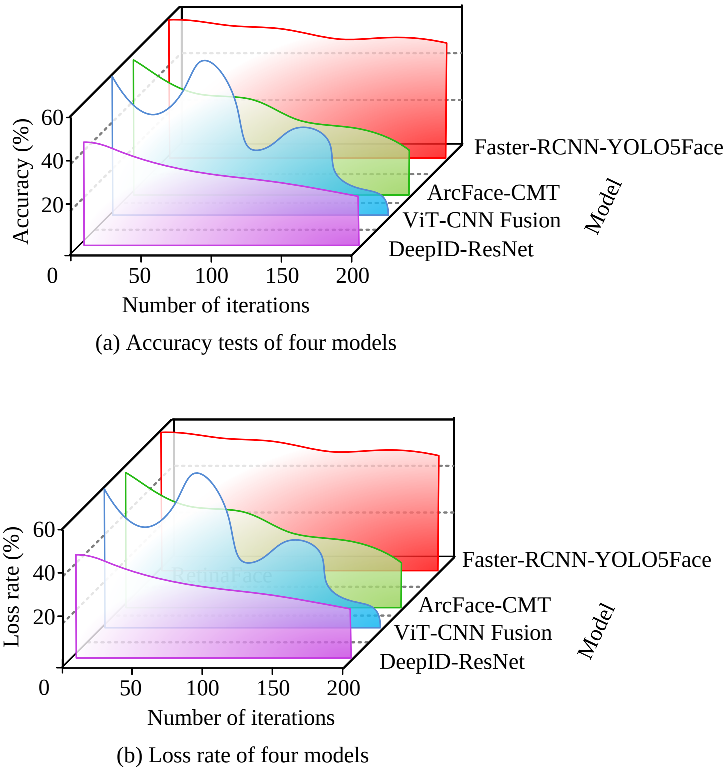
After validating the performance of the FPN-LBPH-YOLO5Face algorithm, experiments were conducted to assess further the practical value of the Faster-RCNN-YOLO5Face embedded detection and recognition system. The system was trained with the PyTorch deep learning framework on Ubuntu 20.04 with Python 3.8. The simulation environment used an Intel Xeon E5-2680 v4 CPU and the WIDER FACE and FDDB datasets. The accuracy and loss rate of the proposed system were compared with Vision Transformer-Convolutional Neural Network Fusion (ViT-CNN Fusion), ArcFace with Convolutional Multi-scale Transformer (ArcFace-CMT), and Deep Identity Residual Network (DeepID-ResNet). The results are shown in Figure 7.

Table 1: Cross validation results.

Fold	ViT-CNN Fusion	ArcFace-CMT	The final system
First discount	99.0±0.2	97.5±0.6	98.2±0.4
Second discount	99.2±0.3	97.3±0.7	98.4±0.5
Third discount	99.3±0.2	97.8±0.5	98.1±0.6
Fourth discount	98.9±0.4	96.9±0.8	97.8±0.7
5th discount	99.1±0.3	97.6±0.6	98.3±0.5
Total	99.10±0.15	97.42±0.34	98.16±0.23

Table 2: Results of the ablation experiment.

Model	Accuracy (%)	Recall rate (%)	Reasoning speed (ms)	mAP@0.5 (%)
A	96.5±0.8	94.8±1.1	25.1±1.2	96.2 ±0.9
B	97.9± 0.5	96.7± 0.7	28.7±1.5	97.6±0.6
C	98.7±0.4	98.2 ±0.5	35.3±1.8	98.5±0.4
D	99.1±0.3	98.7±0.4	36.8±1.9	98.9±0.3



(a) Accuracy tests of four models

(b) Loss rate of four models

Figure 7: Accuracy and loss rate comparison of four systems.

As shown in Figure 7(a), the proposed system led comprehensively in accuracy, remaining stable between 95.6% and 99.1%, significantly higher than comparison systems. In Figure 7(b), the Faster-RCNN-YOLO5Face system also showed the lowest loss rates, averaging 0.08-0.15. These results verified that the hybrid system combined Faster-RCNN's high-precision localization with YOLO5Face's lightweight inference advantages to achieve collaborative improvements in accuracy and generalization. To further verify system effectiveness, the loss convergence of the proposed system and comparison systems was tested, as shown in Figure 8.

As shown in Figure 8, the Faster-RCNN-YOLO5Face system achieved the best loss convergence, with the loss dropping quickly from about 1.8 to below 0.1. It converged faster and more stably than comparison systems. ArcFace-CMT showed larger loss fluctuations and finally converged between 0.3 and 0.5. These results indicated that the proposed system efficiently optimized the loss function during iterations by reusing backbone feature maps and combining Faster-RCNN's fine detection with YOLO5Face's lightweight design, achieving robust convergence in complex scenarios. To further accurately evaluate the system's performance in a real embedded environment, this study conducted deployment testing on the following hardware platforms. NVIDIA Jetson Nano 4GB: Maxwell architecture 128 core GPU, ARM Cortex-A57 quad core CPU; NVIDIA Jetson Xavier NX: Volta architecture 384 core GPU, 6-core NVIDIA Carmel ARM CPU; Raspberry Pi 4 Model B: Broadcom BCM2711 quad core Cortex-A72 CPU; Intel Neural Compute Stick 2: Movidius Myriad X VPU Accelerator. All embedded tests use the same model weights and optimization strategies. Repeat each test 10 times simultaneously. The performance evaluation results for embedded devices are shown in Table 3.

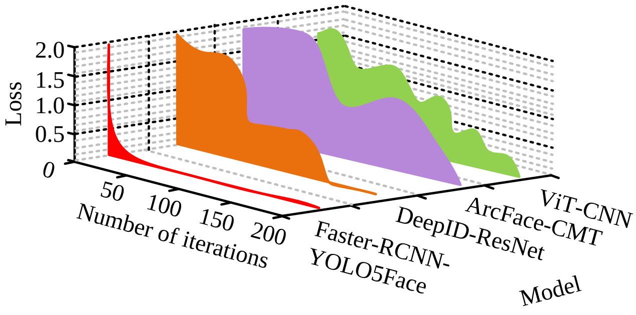


Figure 8: Loss convergence comparison of four systems.

Table 3: Performance evaluation results of embedded devices.

Evaluation metric	Jetson Nano	Jetson Xavier NX	Raspberry Pi 4
Reasoning speed (ms)	98.5± 3.2	45.3±1.8	156.8±5.1
System accuracy (%)	98.2 ±0.5	98.7±0.3	95.1± 0.9
Average power consumption (W)	9.8±0.5	14.2 ±0.7	6.2±0.3
Memory usage (MB)	487±15	512 ±12	298±10
Average power consumption (W)	9.8±0.5	14.2±0.7	6.2±0.3
Peak power consumption (W)	12.3±0.8	19.8 ±1.1	7.5± 0.4
Working temperature (℃)	54.2 ±6.5	61.5 ±8.2	47.3±5.1
Continuous stable operation time (hours)	>24	>48	>12

According to Table 3, post hoc testing showed that the speed differences between all platforms were statistically significant (p<0.01). Specifically, Jetson Xavier NX performs the best, with the most minor difference in inference speed (45.3±1.8 ms) compared to high-end GPUs, and can maintain high accuracy (98.7±0.3%), making it the most practical; The speed of Jetson Nano (98.5±3.2 ms) is significantly slower than Xavier NX, but its accuracy (98.2±0.5%) is comparable, achieving a balance between power consumption and performance; Due to limited computing resources, the Raspberry Pi 4 has the slowest inference speed (156.8±5.1 ms) and a significant decrease in accuracy (95.1±0.9%), and can only run the YOLO5Face component. In terms of power consumption and thermal management, the Jetson Xavier NX provides the highest performance while maintaining power consumption (14.2±0.7W) and temperature (61.5±8.2°C) within its calibration range, with excellent stability and continuous stable operation time exceeding 48 hours; Jetson Nano demonstrates low power consumption (9.8±0.5W) and temperature (54.2±6.5°C), making it suitable for long-term medium load deployment; The Raspberry Pi 4 has significant advantages in power consumption (6.2±0.3W) and temperature (47.3±5.1°C), but is limited by computational performance and stability, and is only suitable for lightweight application scenarios. There are significant differences in energy efficiency ratios across hardware platforms (p<0.001), with the Jetson Xavier NX exhibiting the highest (1.55±0.12 FPS/W), and its performance advantage far exceeds the increase in power consumption.

5 Discussion

This study proposes an embedded detection and recognition system based on Faster-RCNN and YOLO5Face. In terms of real-time performance, it achieved a response time of 28 ms on the WIDER FACE and FDDB datasets, significantly better than existing solutions. Compared with Gupta et al.'s deep learning based online learning engagement detection system (accuracy 92.3%, real-time performance not quantified), the dual engine cascade architecture of the proposed system achieves an ideal balance between speed and accuracy by performing fast initial screening through YOLO5Face and fine recognition through optimized Faster-RCNN [11]. Primarily through feature reuse technology, it effectively reduces computational redundancy by about 65%, which is key to achieving millisecond-level response. The statistical results show that the variance of the system response time is controlled within±1.9 ms, demonstrating its stability during continuous operation.

In terms of robustness in complex environments, the proposed system-maintained recognition accuracy over 95% and recognition time within 0.3 s. This performance is significantly better than Hangaragi and Singh's face-grid-based method (94.23% accuracy) [7]. Its advantages mainly stem from the FPN-LBPH module's multi-scale feature fusion and attention mechanisms that enhance key facial regions. Especially in occlusion cases, the system effectively compensates for information loss through the collaborative work of the texture enhancement and super-resolution reconstruction branches. The five-fold cross-validation results show that the accuracy variance of this system is only 0.15%, significantly lower than that of the comparison model, demonstrating its better adaptability to complex environmental changes.

In the embedded deployment optimization, testing on the Jetson Xavier NX platform showed that the system achieved an inference speed of 45.3±1.8 ms and an accuracy of 98.7±0.3%. The system's energy efficiency ratio reaches 1.55±0.12 FPS/W, effectively controlling energy consumption while maintaining high performance. This is mainly attributed to the application of a dynamic background suppression module and quantization technology, which significantly reduces the computational load and memory usage of embedded devices (controlled within 420-550 MB). Compared with current mainstream methods, this system exhibits comprehensive advantages in multiple dimensions: compared with Anusudha's YOLO InsightFace network, this system provides better small object detection capability through attention mechanism enhanced feature extraction [12]. Compared to George et al.'s EdgeFace network (1.77M parameter), this system has better robustness in complex scenarios while maintaining lightweight; Compared with the triple network recognition framework proposed by Liu et al., this system achieves better real-time performance while maintaining high accuracy [8].

Although this study has achieved excellent performance, it still has certain limitations in practical, complex applications. For example, performance degradation under severe occlusion, inadequate adaptability to extremely low-light conditions, vulnerability to adversarial attacks, and the complexity of cross-hardware platform deployment. Adding tiny perturbations that are difficult for the human eye to detect in the input image may result in the model outputting completely incorrect recognition results. This reveals the inherent security risks of deep learning-based systems, which have not yet integrated any adversarial sample detection or defense modules in the current version, posing potential threats in scenarios with extremely high security requirements. Therefore, in future research, facial completion techniques or graph neural networks can be used to model relationships among facial components under occlusion to improve the recognition of heavily occluded faces. In addition, integrating multimodal perception and introducing adversarial training, defensive distillation, or online attack-detection mechanisms can enhance the system's robustness against adversarial attacks to meet the requirements of high-security applications.

6 Conclusion

To address the problems of poor real-time performance, high resource consumption, and limited generalization in complex scenes encountered in traditional embedded detection and recognition systems for surveillance video, this study proposed an embedded detection and recognition system based on Faster-RCNN and YOLO5Face. Experimental verification shows that the system exhibits excellent comprehensive performance in embedded environments: it not only maintains high-precision recognition, but also achieves millisecond-level response speed; it demonstrates robustness under complex conditions such as lighting changes and local occlusion; and it demonstrates efficient resource utilization and stable operational performance across different hardware platforms. Although the system performed well under standard conditions, the study still had limitations, including insufficient validation in extreme occlusion and low-light conditions, as well as limited deployment adaptability across hardware platforms. In the future, the system can continue to improve by exploring dynamic pruning and implementing cross-scenario adversarial training. These advancements are expected to improve the system's generalization and predictive performance.

References

[1] Mohsen Rostami, Amirhamzeh Farajollahi, and Hashem Parvin. Deep learning-based face detection and recognition on drones. Journal of Ambient Intelligence and Humanized Computing, 15(1):373-387, 2024. https://doi.org/10.1007/s12652-022-03897-8

[2] Shige Xu, Lei Zhang, Yin Tang, Chaolei Han, Hao Wu, and Aiguo Song. Channel attention for sensor-based activity recognition: Embedding features into all frequencies in DCT domain. IEEE Transactions on Knowledge and Data Engineering, 35(12):12497-12512, 2023. https://doi.org/10.1109/TKDE.2023.3277839

[3] Nesrine Triki, Mohamed Karray, and Mohamed Ksantini. A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles. Applied Sciences, 13(8):4793-4807, 2023. https://doi.org/10.3390/app13084793

[4] Huijiao Wang, Haijian Zhang, Lei Yu, Li Wang, and Xulei Yang. Facial feature embedded CycleGAN for VIS–NIR translation. Multidimensional Systems and Signal Processing, 34(2):423-446, 2023. https://doi.org/10.1109/ICASSP40776.2020.9054007

[5] Mohammed Taha, Tarek Mostafa, and Tarek Abd El-Rahman. A novel hybrid approach to masked face recognition using robust PCA and GOA optimizer. Scientific Journal for Damietta Faculty of Science, 13(3):25-35, 2023. https://doi.org/10.21608/sjdfs.2023.222524.1117

[6] Hui Zhang, Xingbo Dong, YenLung Lai, Ying Zhou, Xiaoyan Zhang, Xingguo Lv, Zhe Jin, and Xuejun Li. Validating privacy-preserving face recognition under a minimum assumption. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12205-12214, 2024. https://doi.org/10.1109/CVPR52733.2024.01160

[7] Shivalila Hangaragi, Tripty Singh, and Neelima N. Face detection and Recognition using Face Mesh and deep neural network. Procedia Computer Science, 218:741-749, 2023. https://doi.org/10.1016/j.procs.2023.01.054

[8] Jianqi Liu, Zhiwei Zhao, Pan Li, Geyong Min, and Huiyong Li. Enhanced embedded AutoEncoders: An attribute-preserving face de-identification framework. IEEE Internet of Things Journal, 10(11):9438-9452, 2023. https://doi.org/10.1109/JIOT.2023.3235725

[9] Qingqiu Huang, Lei Yang, Huaiyi Huang, Tong Wu, and Dahua Lin. Caption-supervised face recognition: training a state-of-the-art face model without manual annotation. Computer Vision-ECCV, 139-155, 2020. https://doi.org/10.1007/978-3-030-58520-4_9

[10] Daniella Raz, Corinne Bintz, Vivian Guetler, Aaron Tam, Michael A. Katell, Dharma Dailey, Bernease Herman, Meg Young, and P. Krafft. Face Mis-ID: An interactive pedagogical tool demonstrating disparate accuracy rates in facial recognition. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 895-904, 2021. https://doi.org/10.1145/3461702.3462627

[11] Swadha Gupta, Parteek Kumar, and Raj Kumar Tekchandani. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 82(8):11365-11394, 2023. https://doi.org/10.1007/s11042-022-13558-9

[12] Anjeana N, and K. Anusudha. Real time face recognition system based on YOLO and Insight Face. Multimedia Tools and Applications, 83(11):31893-31910, 2024. https://doi.org/10.1007/s11042-023-16831-7

[13] Weijun Chen, Hongbo Huang, Shuai Peng, Changsheng Zhou, and Cuiping Zhang. YOLO-face: A real-time face detector. The Visual Computer,

37(4):805-813, 2021. https://doi.org/10.1007/s00371-020-01831-7

[14] Bin Jiang, Hongbin Jiang, Huanlong Zhang, Qiuwen Zhang, Zuhe Li, and Lixun Huang. 4AC-YOLOv5: An improved algorithm for small target face detection. EURASIP Journal on Image and Video Processing, 2024(1):10-24, 2024. https://doi.org/10.1186/s13640-024-00625-4

[15] Mrs Sk Raziya Sultana, Kunduru Sravani, Narayana Sree, Ranga Lokesh, Kukatlapalli Venkateswararao, Korrapati Lakshmaiah, and Sai S V. Automated ID card detection and penalty system using YOLOv5 and face recognition. International Journal of Recent Advances in Engineering and Technology, 14(1):63-72, 2025.

[16] Changhong Chen, Shaofeng Wang, and Shunzhou Huang. An improved faster RCNN-based weld ultrasonic atlas defect detection method. Measurement and Control, 56(3-4):832-843, 2023. https://doi.org/10.1177/00202940221092030

[17] Zhenwei He, Lei Zhang, Xinbo Gao, and David Zhang. Multi-adversarial faster-RCNN with paradigm teacher for unrestricted object detection. International Journal of Computer Vision, 131(3):680-700, 2023. https://doi.org/10.1007/s11263-022-01728-z

[18] R. Josphineleela, P. B. V. Raja Rao, Amir shaikh, and K. Sudhakar. A multi-stage faster RCNN-based iSPLInception for skin disease classification using novel optimization. Journal of Digital Imaging, 36(5):2210-2226, 2023. https://doi.org/10.1007/s10278-023-00848-3