Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ26566

Application of Deep Learning-Based Face Recognition Method in Robot Vision Systems

Zhengyi Tang

Institute of Energy and Mining Engineering, China University of Mining and Technology, Beijing, Chinaa
2249027615@qq.com

Abstract. In recent years, the development of deep learning has significantly driven the progress of face recognition technology in the field of computer vision. For service robots and embedded intelligent systems, face recognition is not only related to identity authentication, but also affects the implementation effect of functions such as human-robot interaction and behavior analysis, thus largely determining the intelligence level of the system. Although many existing face recognition methods based on deep convolutional neural networks (CNN) have high accuracy, they often have large model sizes and high computational costs, which impose considerable limitations on implementation in scenarios with requirements for real-time performance, energy consumption and deployment space, such as robot platforms. To address this problem, this paper proposes a lightweight face recognition method combining MobileNetV3 and the Convolutional Block Attention Module (CBAM). As the backbone network, MobileNetV3 features a compact structure and low computational cost, which can reduce the model volume and inference latency to a certain extent. On this basis, CBAM is introduced to perform adaptive weighting in both channel and spatial dimensions, enabling the network to focus more on facial regions with richer information during feature extraction, thereby improving feature representation capability and recognition robustness. The model is trained and tested on the public dataset Labeled Faces in the Wild (LFW). The cross-entropy loss function and Adam optimizer are adopted in the training phase, and the model performance is tested under conditions such as complex backgrounds, pose variations and partial occlusion. Experimental results show that while maintaining a small model size, the proposed method achieves higher recognition accuracy than multiple existing lightweight networks, strikes a good balance between accuracy and efficiency, and is suitable for face recognition tasks in robot vision scenarios. Overall, combining lightweight network design with attention mechanisms can provide a practical implementation approach for high-performance face recognition on edge devices.

Keywords: Deep learning, Face recognition, Lightweight network, Attention mechanism, MobileNetV3, Robot vision

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ26566

1. Introduction

With the continuous development of artificial intelligence and deep learning technologies, the perception and cognition capabilities of robot vision systems [1] have been continuously enhanced, and they have been applied in multiple fields such as intelligent security, service robots and human-robot interaction (HRI). For robots, face recognition is one of the important ways for them to understand and recognize humans. This technology has the characteristics of non-contact, easy collection and high user acceptance, so it has obvious advantages in practical applications. Whether it is identity verification, visitor management, or social assistance and personalized services, the introduction of a face recognition module into a robot system can usually further improve the interaction experience and the intelligence level of the system.

Early face recognition methods [2] relied more on handcrafted features, such as Local Binary Pattern (LBP) and Principal Component Analysis (PCA). Such methods can achieve certain effects in simple environments, but their robustness often decreases significantly when encountering illumination variations, occlusion, pose deviation or complex backgrounds. In recent years, convolutional neural networks (CNN) [3] have made great breakthroughs in tasks such as image classification, object detection and face recognition. The end-to-end learning method enables the model to automatically extract more discriminative features, and also provides a reliable basis for building high-precision face recognition systems. However, robot platforms are often limited by computing power, storage space and response time, so it is unrealistic to simply pursue high accuracy. How to make the model lighter and faster while ensuring the recognition effect has become a key point in current research.

Focusing on this problem, this paper constructs a model framework combining lightweight CNN and attention modules [4]. MobileNetV3 [5] is selected as the backbone network because it has a relatively small number of parameters and high computational efficiency, making it more suitable for deployment on mobile terminals and embedded platforms. On this basis, the ideas of CBAM [6] and SE (Squeeze-and-Excitation) [7] are further introduced to adjust the weights of features in the channel dimension and spatial dimension, so as to enhance the model's ability to focus on key facial regions. After such processing, the network can more easily highlight discriminative regions such as eyes, mouth corners and the nose bridge during feature extraction, thus maintaining more stable recognition performance in complex backgrounds.

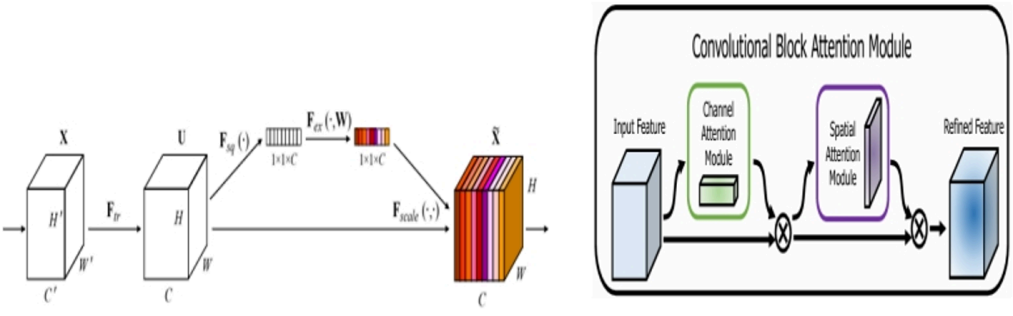


Figure 1. Schematic diagram of the model structure combining lightweight CNN and attention modules. After the backbone network (e.g., MobileNetV3) extracts feature maps, CBAM or SE modules are applied to weight the channel and spatial dimensions respectively to highlight more critical facial regions, and finally face recognition is completed via softmax

DOI: 10.54254/2755-2721/2025.BJ26566

This paper completes the model training and testing pipeline based on PyTorch, and conducts experimental verification on the public dataset LFW. The results show that the proposed lightweight attention fusion model maintains high recognition accuracy while keeping the model complexity at a low level, making it more suitable for practical deployment on edge devices such as robots. In addition, the paper also briefly discusses the expansion value of few-shot learning [8] and continual learning mechanisms [9] in real applications, hoping to provide some reference for the continuous evolution of robot vision systems in the future.

2. Manuscript preparation

2.1. Lightweight neural network architecture

With the increasing demand for mobile deployment and real-time inference, lightweight neural network architectures [10] have attracted more and more attention. Especially in robot vision systems, memory, power consumption and processing capability are often limited. To reduce the computational burden and improve inference speed, researchers have proposed various structural optimization schemes, among which the most representative category is the MobileNet series.

The core contribution of MobileNetV1 is the introduction of depthwise separable convolution [11], which decomposes standard convolution into depthwise convolution and pointwise convolution. After this decomposition, the number of parameters and computational amount can be significantly reduced, while the basic feature extraction capability is retained. On this basis, MobileNetV2 further proposes an inverted residual structure and linear bottlenecks, making cross-layer feature transmission more stable. Its idea is to preserve low-level information through shortcut connections and use linear transformation to alleviate information loss caused by non-linear activation during dimensionality compression. At the stage of MobileNetV3, the model combines neural architecture search (NAS) and the h-swish activation function to optimize the network structure in more detail, thus achieving a better balance among speed, accuracy and energy consumption. At the same time, MobileNetV3 also integrates the SE module for channel-level attention modeling, which also enhances its adaptability in complex vision tasks.

Besides the MobileNet series, lightweight structures such as ShuffleNet and SqueezeNet have also received much attention. ShuffleNet reduces redundant computation and improves parallel efficiency through channel grouping and channel shuffling; SqueezeNet constructs a more compact model through small convolution kernels and aggressive parameter compression strategies, which is suitable for memory-constrained scenarios. Overall, these networks have obvious advantages in lightweight design, but they still encounter a common problem when used for face recognition: the feature representation capability is easily limited, and key facial regions may not be fully extracted. For this reason, how to further enhance the semantic expression capability of lightweight models without significantly increasing computational cost remains an important direction in current research.

2.2. Development and optimization of face recognition models

In recent years, deep learning-based face recognition models have gradually become the mainstream solution in this field. Compared with traditional handcrafted feature methods such as LBP, HOG and Eigenfaces, deep networks [12] can automatically learn more discriminative representations from large-scale face data, so they usually perform better in recognition accuracy and generalization capability. Early DeepFace and DeepID encode face images through multi-layer convolutional

DOI: 10.54254/2755-2721/2025.BJ26566

neural networks, and then complete identity classification using softmax-based classifiers, which laid the foundation for subsequent work.

After that, FaceNet introduced the triplet loss function [13] for feature embedding learning, making samples of the same identity closer in the feature space and samples of different identities farther apart. This idea significantly enhanced inter-class separability. ArcFace further added angular margin to the loss function, and optimized the training objective from the perspective of the cosine angle of feature vectors, making the intra-class distribution more compact and inter-class boundaries clearer. Therefore, it has become one of the most widely used methods in high-precision face recognition. In addition, strategies such as CosFace, SphereFace, Center Loss and Joint Supervision have also improved the model's discriminative capability from the perspectives of loss functions, feature normalization and spatial constraints.

However, although these high-performance methods have achieved near human-level recognition effects on public benchmarks such as LFW and MegaFace, they are often built on large-scale data training and deep complex networks, with strong dependence on computing resources, making practical deployment not easy. For resource-constrained platforms such as robot systems, this problem is more prominent. Therefore, in recent years, more and more studies have turned to the design of lightweight face recognition models [14], hoping to maintain competitive recognition accuracy at lower computational costs. Around this goal, efficient backbone networks such as MobileNet and ShuffleNet have been introduced into face-related tasks, while techniques such as structured pruning, quantization and knowledge distillation have also been widely applied to preserve discriminative features as much as possible and compress model complexity. Some recent studies have also attempted to introduce graph neural networks and hybrid embedding methods into face recognition frameworks to improve the model's robustness under conditions such as age variations, pose variations and expression interference.

2.3. Application of attention mechanisms in face recognition

The attention mechanism [15] is an important direction in image recognition research in recent years. Its basic idea is not complicated: it assigns different weights to different features, allowing the network to allocate more computing resources to information-rich regions while weakening the influence of background noise and redundant responses. This mechanism has shown good effects in tasks such as image classification, object detection and face recognition.

The Squeeze-and-Excitation (SE) module is one of the earlier widely used attention structures. It recalibrates the response of each channel by modeling global channel information, thereby enhancing more effective feature channels. SE has proven its effectiveness in various backbone networks such as ResNet and DenseNet. However, SE focuses more on channel-dimensional modeling and does not explicitly characterize the importance of spatial locations. To address this deficiency, researchers proposed the Convolutional Block Attention Module (CBAM). CBAM sequentially performs channel attention and spatial attention to jointly reweight feature maps in two dimensions. Specifically, the channel attention branch extracts global context using max pooling and average pooling, and then generates channel weights through fully connected layers; the spatial attention branch integrates cross-channel information and generates a two-dimensional attention map through convolution, guiding the model to focus on more informative regions.

The advantage of CBAM is its relatively simple structure and small increase in parameters, so it is very suitable for combination with lightweight networks. This is particularly important for face recognition tasks, because the model not only needs to be accurate enough, but also needs to run on edge devices. When there is occlusion, pose deflection or uneven illumination in the input image,

DOI: 10.54254/2755-2721/2025.BJ26566

the attention mechanism can highlight key facial regions such as eyes, nose and mouth corners, thereby improving local detail discrimination capability and overall recognition stability. For this reason, the combination of lightweight backbone networks and attention mechanisms is gradually becoming a common idea in deployable face recognition systems.

3. Proposed method

3.1. Design of lightweight backbone network

In practical robot applications, computing capability and energy supply are relatively limited, so model design cannot only focus on recognition accuracy, but also take operating efficiency into account. Based on this consideration, this paper selects MobileNetV3 as the feature extraction backbone network. This network has a compact structure and high computational efficiency, and has been widely used in mobile terminal and edge device scenarios. Compared with previous models, MobileNetV3 combines neural architecture search (NAS) and the h-swish activation function to achieve a better balance between performance and efficiency. Its basic unit still takes the inverted residual structure as the core, reduces computational overhead through depthwise separable convolution, and then completes channel fusion through pointwise convolution.

To make MobileNetV3 more suitable for face recognition tasks, this paper makes several targeted adjustments on its basis. First, some redundant convolutional layers are appropriately removed to reduce inference latency. Second, key semantic layers that are more related to facial discriminative information are retained, so that the network can still extract fine-grained regional features such as eyes, nose bridge and mouth corners. Third, by adjusting the convolution kernel size, stride and channel width, the model can integrate receptive fields at different scales to cope with the impacts of pose variations, partial occlusion and illumination differences. In addition, a skip connection strategy is added to the model to enable fuller fusion of shallow texture information and high-level semantic features, which is helpful for improving fine-grained recognition capability.

3.2. Integration strategy of attention mechanism

Although lightweight networks have high operating efficiency, they often fail to focus on truly key regions when the background is complex or the distribution of facial features is uneven, which affects the recognition effect to a certain extent. To solve this problem, this paper embeds the Convolutional Block Attention Module (CBAM) into several stages of the backbone network, hoping to improve the model's responsiveness to important facial regions without adding too much extra overhead.

CBAM consists of two parts: channel attention and spatial attention. The channel attention branch learns the importance of different channels by assigning adaptive weights to the input feature maps, so that semantically stronger responses are strengthened. The spatial attention branch pays more attention to position distribution. It generates a two-dimensional attention map based on the feature maps after average pooling and max pooling, thereby guiding the model to focus more on key positions such as eyes, nose and mouth. In this paper, CBAM is inserted into the middle and high-level units of MobileNetV3 in a lightweight, plug-and-play manner. On the one hand, this can enhance the modeling capability of key semantic regions; on the other hand, it avoids the additional computational overhead caused by large-scale module stacking in shallow layers as much as possible.

DOI: 10.54254/2755-2721/2025.BJ26566

In the experiments, this paper also compares the effects of placing attention modules in shallow, middle and deep layers. The results show that when CBAM is placed in the middle and high-level stages, the discriminative capability of the model is improved more obviously, and the inference speed remains relatively stable. Compared with attention methods that focus on single-dimensional modeling such as SE, CBAM considers both channel and spatial aspects at the same time, so it pays more sufficient attention to key facial structures under occlusion, pose variations and complex backgrounds, and has better overall robustness. Considering the practical demand of edge deployment, this paper controls the number and scale of CBAM, and only adds attention units at key positions to ensure a relative balance between the lightweight property and expressive capability of the model.

3.3. Design of feature encoding and classification structure

After joint feature extraction by MobileNetV3 and CBAM, the network outputs a set of high-dimensional feature maps. To complete identity classification, these features still need further processing. This paper first uses global average pooling (GAP) to convert the output feature maps into fixed-length global feature vectors, and then maps the features to a unified scale space through L2 normalization, so as to make the subsequent classifier training more stable. Then, the normalized feature vectors are input into the fully connected layer to complete identity prediction.

During the training process, this paper takes softmax cross-entropy loss as the main optimization objective, and cooperates with Dropout and batch normalization to improve the stability of the training process and enhance the generalization capability of the model. To reduce the risk of overfitting in multi-class recognition tasks, learning rate decay and early stopping strategies are also introduced during training. Considering the actual needs of the deployment phase, in addition to recognition accuracy, indicators such as model parameter count, memory usage and inference latency are also counted in the experimental evaluation. Experimental results show that the model can maintain a stable operating state on embedded devices such as Raspberry Pi 4B and Jetson Nano, and has competitive recognition performance.

Although methods such as ArcFace or Triplet Loss perform outstandingly in high-precision face recognition, such loss functions have higher requirements for sample construction and training pipelines, and their engineering implementation is relatively complex. Comprehensively considering model reusability, implementation difficulty and deployment requirements, this paper finally selects traditional classification loss as the main training objective, which not only reduces engineering complexity, but also ensures a more stable training process. Experimental results show that the finally trained model can still achieve good recognition performance on challenging datasets such as LFW with a low parameter count and fast inference speed.

4. Experiments

4.1. Experimental setup

The experiments are carried out on the training platform and the deployment platform respectively. Model training is completed on a deep learning server equipped with an NVIDIA RTX 3060 GPU (12 GB), running Ubuntu 20.04 and PyTorch 1.12. Deployment and inference testing are conducted on NVIDIA Jetson Nano, which is equipped with a quad-core ARM Cortex-A57 CPU and a 128-core Maxwell GPU, and is a common type of embedded platform in robot and edge computing scenarios.

DOI: 10.54254/2755-2721/2025.BJ26566

To ensure the fairness and reproducibility of the experiments, the LFW dataset is used for both training and evaluation. This dataset contains more than 13,000 face images from 5,749 identities, with obvious differences in illumination, pose and background. To simulate complex environments closer to real applications, this paper also constructs several challenging test subsets, mainly including occlusion, illumination variations and profile views. All input images are uniformly cropped to 112×112 pixels and aligned with the help of facial keypoints. Data augmentation strategies such as random horizontal flipping, brightness adjustment and random occlusion are used in the training phase to improve the robustness and generalization capability of the model.

The proposed model takes MobileNetV3 as the backbone, and attention modules are added after selected convolutional layers. The Adam optimizer is used for training, with an initial learning rate set to 0.001. The learning rate is scheduled with a cosine annealing strategy within 50 epochs, and the batch size is set to 64. Cross-entropy loss is used for the classification task. In some comparative experiments, ArcFace loss is additionally tried to investigate the possibility of further improving inter-class separability. For the face verification task, this paper uses cosine similarity to compare the normalized feature vectors. The final evaluation indicators mainly include recognition accuracy, model size and inference time on Jetson Nano, which are used to comprehensively reflect the effectiveness and deployability of the model in robot vision scenarios.

4.2. Comparative analysis

To verify the effectiveness of the proposed method, this paper compares it with a variety of representative face recognition models, including ResNet-50, FaceNet, MobileNetV3 and ShuffleNetV2. All models are tested on the same dataset and hardware environment to ensure the fairness of the comparison results as much as possible. Table 1 shows the results of each model in terms of parameter count, recognition accuracy and inference time.

Table 1. Performance comparison between the proposed model and mainstream models on face recognition task (LFW dataset)

Model	Parameters (M)	LFW Accuracy (%)	Inference Time (ms)	Characteristics
ResNet-50	23.5	99.13	270	High precision, high computational cost
FaceNet	22.0	98.70	310	Strong feature expression capability, difficult to deploy
MobileNetV3	3.5	96.50	87	Lightweight and fast, slightly lower accuracy
ShuffleNetV2	3.4	96.41	95	Simple structure, low power consumption
Ours (CNN+CBAM)	4.1	98.67	89	High precision, deployment-friendly

It can be seen from Table 1 that ResNet-50 achieves an accuracy of 99.13% on the LFW dataset, and can be used as a high-performance benchmark model under sufficient computing resources. However, its parameter count reaches 23.5 million, and the inference time per image is 270 ms, which makes it difficult to deploy directly on edge devices. FaceNet also has strong feature learning capability, but due to its complex overall structure, its inference latency on Jetson Nano exceeds 300 ms, which will be greatly limited in practical embedded applications.

In comparison, MobileNetV3 and ShuffleNetV2 have more advantages in terms of lightweight design. With 3.5 million parameters and 87 ms inference time, MobileNetV3 achieves an accuracy of 96.50% on LFW, forming a good balance between model size and speed. ShuffleNetV2 further

DOI: 10.54254/2755-2721/2025.BJ26566

reduces computational complexity through channel grouping and shuffling, achieving an accuracy of 96.41% with an inference time of about 95 ms. Although both types of models are suitable for lightweight deployment, there is still room for further improvement in recognition accuracy.

After adding the attention mechanism on the basis of the lightweight backbone network, the model proposed in this paper shows better comprehensive performance. The parameter count of the model is controlled at about 4.1 million, and the inference time is 89 ms, which is basically close to that of MobileNetV3, but the accuracy is increased to 98.67%, which is significantly higher than other lightweight baseline models, and the gap with ResNet-50 is also controlled within 0.5%. It can be seen that the combination of lightweight networks and attention mechanisms can indeed improve the model's modeling capability for key facial regions without significantly increasing computational overhead, thus achieving a more reasonable balance between recognition accuracy and deployment efficiency.

5. Conclusion

With the continuous expansion of applications such as smart cities, intelligent security and human-robot interaction, face recognition has become a key capability in robot vision systems. Since most of these systems are deployed on embedded or mobile platforms with limited computing power, storage and energy consumption, model design must take both recognition performance and operating efficiency into account. Focusing on this demand, this paper proposes a face recognition model based on MobileNetV3 and attention mechanism, which enhances the feature expression capability of key facial regions while maintaining the overall compactness of the network.

To improve adaptability in complex scenarios, various data augmentation strategies are added during the training process, so that the model can better cope with problems such as illumination variations, occlusion and pose deviation. The network finally outputs normalized feature embeddings and performs identity verification using cosine similarity. Experimental results show that on the LFW dataset and its challenging subsets, the proposed model achieves a recognition accuracy of 98.67% with about 4.1 million parameters and inference time lower than 89 ms on NVIDIA Jetson Nano. Compared with larger-scale models such as ResNet-50 and FaceNet, this method shows more balanced advantages in accuracy, model size and inference speed, and is more suitable for robot vision applications under resource-constrained conditions.

Overall, the combination of lightweight architectures and attention mechanisms is a feasible path to improve face recognition performance on edge devices. The work in this paper demonstrates that without significantly increasing computational cost, the model's ability to focus on key facial structures can be effectively enhanced by reasonably introducing attention modules, thereby improving recognition performance. In the future, combined with few-shot learning [8] and continual learning mechanisms [9], it is expected to further improve the scene adaptability and long-term operation capability of robots in face recognition tasks.

References

[1] Lindner, L., Sergiyenko, O., Rodríguez-Quiñonez, J. C., et al. (2016). Mobile robot vision system using continuous laser scanning for industrial application. *Industrial Robot: An International Journal, 43*(4), 360–369.

[2] Trigueros, D. S., Meng, L., & Hartnett, M. (2018). Face recognition: From traditional to deep learning methods. *arXiv preprint arXiv:1811.00116*.

[3] Li, Z., Liu, F., Yang, W., et al. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems, 33*(12), 6999–7019.

DOI: 10.54254/2755-2721/2025.BJ26566

[4] Woo, S., Park, J., Lee, J. Y., et al. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).
[5] Koonce, B. (2021). MobileNetV3. In *Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization* (pp. 125–144). Apress.
[6] Woo, S., Park, J., Lee, J. Y., et al. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).
[7] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).
[8] Wang, Y. X., & Hebert, M. (2016). Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision* (pp. 616–634). Springer International Publishing.
[9] Thiessen, E. D., Girard, S., & Erickson, L. C. (2016). Statistical learning and the critical period: How a continuous learning mechanism can give rise to discontinuous learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(4), 276–288.
[10] Zhou, Y., Chen, S., Wang, Y., et al. (2020). Review of research on lightweight convolutional neural networks. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)* (pp. 1713–1720). IEEE.
[11] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).
[12] Li, X., Dong, X. L., Lyons, K., et al. (2015). Truth finding on the deep web: Is the problem solved? *arXiv preprint arXiv:1503.00303*.
[13] Cheng, D., Gong, Y., Zhou, S., et al. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1335–1344).
[14] Deng, J., Guo, J., Zhang, D., et al. (2019). Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0–0).
[15] Guo, M. H., Xu, T. X., Liu, J. J., et al. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media, 8*(3), 331–368.