International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Research on Intelligent Visual Recognition Methods Based on Deep Learning

Haoyu Hu[1], Xinyue Xu[2], Hu Ke[3], Zhang Qing[4], Li Feng*
[1, 2, 3, 4]Anhui University of Finance and Economics
No. 962 Caoshan Road, Bengbu City, Anhui Province, China 233030
*Corresponding author: Li Feng

Abstract*—Intelligent visual recognition system is an important research direction in the field of computer vision and artificial intelligence. In this paper, a set of intelligent visual recognition method is constructed based on deep learning technology, which organically integrates the techniques of YOLO target detection, ResNet image classification, Dlib/FaceNet face recognition, GAN data enhancement, and morphological operations of image preprocessing. Firstly, the system adopts the YOLO model to realize real-time detection and localization of multiple targets in the image; secondly, ResNet, a deep residual network, is used to classify the targets with high accuracy; thirdly, the face feature extraction and recognition is realized by combining the Dlib face detection and FaceNet depth model, in which the ternary loss function is introduced to optimize the embedded representation of the face; at the same time, a realistic face recognition is generated by GAN to generate realistic image samples to expand the training data so as to improve the robustness and generalization ability of the model; finally, morphological preprocessing techniques are used for image enhancement and noise suppression. Experimental results show that the proposed method achieves excellent average target detection accuracy and classification recognition accuracy on several public datasets, where GAN data enhancement effectively improves the model performance in data sparse scenarios, and morphological preprocessing improves the detection effect under low-quality images.*

***Keywords*—** *Deep learning; target detection; image classification; face recognition; generative adversarial networks; mathematical morphology.*

I. INTRODUCTION

In recent years, deep learning has made breakthroughs in the field of computer vision, enabling machines to have visual perception and understanding capabilities close to the human level. This has driven the widespread application of intelligent visual recognition systems in fields such as security surveillance, autonomous driving, and human-computer interaction. However, complex and changing real-world scenarios put forward higher requirements for visual recognition systems, including detection and localization of multiple categories of targets, classification and recognition of fine-grained objects, and accurate recognition of specific face identities. A single algorithm is often difficult to balance speed and accuracy and to handle diverse tasks. For this reason, researchers have begun to explore the integration of multiple deep learning techniques under a unified framework in order to build an intelligent visual recognition system with more superior performance and more comprehensive functions. The work in this paper is based on this idea, integrating the current mainstream target detection, image classification, face recognition and data enhancement techniques to propose an intelligent visual recognition method. Our system adopts YOLO series model for real-time target detection, ResNet with deep residual network for high accuracy image classification, combines Dlib library for face detection and FaceNet model proposed by Google for face feature extraction and recognition, and introduces GAN to synthesize training samples to expand the dataset, and is supplemented with mathematical morphology image preprocessing to improve the quality of images in complex environments. The system is able to improve the image quality in complex environments through the cooperative work of multiple modules. By working with multiple modules, the system is able to analyze the camera video or image data end-to-end, from which it can detect the target of interest, determine its category, recognize the identity of the face, and enhance the input image to improve the overall robustness of the recognition.

II. RELATED WORK

Target Detection: target detection is one of the fundamental tasks in computer vision, aiming to find out all the targets in an image and give their boundary locations and category labels. Before the rise of deep learning, target detection mainly relies on manual features and sliding window classifiers. R-CNN series of methods apply convolutional neural networks to target detection, and achieve accurate detection through candidate region extraction and classification, but the computational overhead is large. In recent years, one-stage detection algorithms such as YOLO and SSD have led a new direction in real-time target detection. Redmon et al. in the first proposed YOLO model, which treats target detection as a single regression problem, realizes end-to-end prediction from image pixels to bounding boxes and categories. The YOLO framework achieves this by dividing an $S \times S$ grid over the image, and each grid directly regresses a fixed number B of bounding boxes and their confidence and category probabilities, thus dramatically improving the detection speed. Subsequently introduced improved models such as YOLOv2/v3 further improve the detection accuracy, e.g., by introducing multi-scale prediction and deeper feature networks. In this paper, we adopt the YOLO algorithm as the target detection module to balance real-time and accuracy.

Image Classification: The task of image classification refers to categorizing an entire image into a predefined certain

Haoyu Hu, Xinyue Xu, Hu Ke, Zhang Qing, and Li Feng, "Research on Intelligent Visual Recognition Methods Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 12, pp. 541-545, 2025.

class. AlexNet proposed by Krizhevsky et al. in 2012 started the boom of deep CNNs for large-scale image classification, and since then networks such as VGG and Inception have been constantly breaking the performance record of the ImageNet dataset. However, as the network deepens, the problems of gradient vanishing and network degradation start to appear. ResNet, a residual network proposed by Kaiming He et al. successfully solves the problem of training ultra-deep networks by adding a "skip connection" of constant mappings between convolutional layers, which enables the network to learn the residual mappings instead of the complete mappings, thus greatly alleviating the problem of gradient vanishing. Due to the excellent performance and easy scalability of ResNet structure, ResNet is selected as the backbone network of image classification module in this paper, which is used to extract the discriminative features of the target and perform classification decision.

Face Recognition: face recognition has long been an important topic in computer vision. Early approaches include face recognition algorithms based on PCA and LDA with artificial features, and Later's LBP features. In the era of deep learning, Facebook's DeepFace and Google's FaceNet models have pushed face recognition accuracy to new heights. FaceNet, proposed by Schroff et al. gets rid of the limitation of classifiers by designing a ternary loss function to measure face similarity directly in the feature space. Specifically, FaceNet uses a deep convolutional network to map face images to a feature vector, and when trained with a reasonable loss function, face features from the same person are closer together and face features from different people are farther apart.Dlib is an open source machine learning library that provides high-performance face detection and alignment. In this paper, we use Dlib's face detector to quickly locate the position of the face in the image and correct the pose, and then extract the 128-dimensional embedding vectors of the face using the FaceNet model, and perform face recognition by comparing the vector distances. Compared with traditional methods, face recognition by deep learning has higher robustness and accuracy, and has been widely used in the fields of identity verification and security monitoring.

Generating Adversarial Networks and Data Augmentation: the performance of deep models is highly dependent on a large amount of diverse training data. data augmentation technique generates new samples by transforming existing data (e.g., flipping, cropping, color perturbation, etc.), which is a commonly used means to enhance the generalization ability of models. In recent years, the Generative Adversarial Network (GAN) proposed by Ian Goodfellow et al. provides a new idea for data augmentation, which consists of two adversarial modules, the generator and the discriminator, and is able to learn the distribution of the training data and generate new samples with the same distribution. In the image domain, models such as DCGAN can generate realistic images that can be used to expand the training set or as part of data augmentation. For example, when there are fewer samples in certain categories, a GAN can be trained to generate synthetic images of that category to balance the data distribution. It has been shown that training with GAN-generated data augmentation can achieve significant performance gains over traditional augmentation in image classification and detection tasks. In this paper, the GAN data augmentation module is introduced into the system to generate and mix the target category images, such as generating samples with different lighting and angles, in order to improve the model's ability to adapt to changes. It should be noted that the quality and diversity of GAN generated samples directly affect the enhancement effect, so training a stable GAN model is also an important part of this system.

Image preprocessing and mathematical morphology: images acquired in complex environments often have problems such as noise, highlights or shadow occlusion, which may affect the recognition accuracy when directly input into the depth model. Traditional image preprocessing methods can improve image quality at the front-end and lay a good foundation for subsequent high-level tasks. Among them, mathematical morphology provides a series of effective image morphological processing operators, including erosion, dilation, open operation, and closed operation. The erosion operation removes noise by shrinking the foreground region, while the expansion operation fills voids and connects broken parts by expanding the foreground region. Their mathematical definitions can usually be expressed as set operations, e.g. given a set A of binary images and a structure element B , the expansion is defined as

$$A \oplus B = \left\{ x \mid \left(\hat{B} \right)_x \cap A \,\mathrm{neq} \varnothing \right\} = \bigcup_{b \in B} (A + b)$$

That is, when the center of the structural element B is located at z, the point z belongs to the inflationary result whenever B has a non-empty intersection with the image A. Conversely, corrosion can be defined as:

$$A \ominus B = \{ x \mid B_x \subseteq A \}$$

Denotes that z belongs to the corrosion result only when all elements of the structure element B are completely contained in A at position z. Based on erosion and expansion, open operations (erosion followed by expansion) can be further defined for eliminating small noises, closed operations (expansion followed by erosion) for filling in regionally fine voids, and so on. In this study, we apply morphological preprocessing for low-quality images before feeding them into the depth model, such as open and closed operations on binarized images to remove isolated noises and smooth target edges to improve the robustness of subsequent YOLO detection and FaceNet recognition.

III. METHODOLOGIES

3.1 YOLO object detection

The YOLO module is responsible for detecting the positions and categories of all targets in the input image in real time. YOLO adopts a single-stage object detection framework, dividing the input image into S × S grids, and each grid predicts B bounding boxes and their category probabilities. To train the YOLO detector, a specific loss function needs to be defined to measure the error between the prediction and the true value. The loss function of YOLO consists of three parts: localization error, confidence error, and classification error. Its

Haoyu Hu, Xinyue Xu, Hu Ke, Zhang Qing, and Li Feng, "Research on Intelligent Visual Recognition Methods Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 12, pp. 541-545, 2025.

structure can be expressed as:

$$L_{\text{YOLO}} = \lambda_{\text{coord}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} [(x_{ij} - x_{ij}^*)^2 + (y_{ij} - y_{ij}^*)^2 + (w_{ij}^{1/2} - w_{ij}^{*1/2})^2 + (h_{ij}^{1/2} - h_{ij}^{*1/2})^2] + \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} (C_{ij} - C_{ij}^*)^2 + \lambda_{\text{noobj}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{noobj} (C_{ij} - C_{ij}^*)^2 + \sum_{i=1}^{S^2} 1_{i}^{obj} \sum_{c=1}^{C} (p_i(c) - p_i^*(c))^2$$

In the above formula, x, y represents the center coordinates of the bounding box, w, h represents the width and height of the bounding box, and * represents the true value; C is the confidence of the bounding box containing the target, $p_i(c)$ represents the probability that the i-th grid predicts that the category is c; $1_{ij}^{obj} = 1$ indicates that the j-th prediction box of the i-th grid detects a target, otherwise it is 0; similarly, $1_{ij}^{noobj} = 1$ indicates that there is no target. The loss function gives different weights (through $\lambda coord$ and $\lambda noobj$) to the coordinate error of the grid containing the target and the confidence error of whether there is a target or not. By minimizing the above loss, the YOLO network learns how to accurately locate and classify multiple targets simultaneously in a single image.

In implementation, we adopt an improved YOLOv3 model as the basis of the detector. YOLOv3 uses the Darknet-53 backbone network to extract features and outputs predictions at three scales to better detect targets of different sizes. During training, we pre-train YOLOv3 on the Pascal VOC dataset and then fine-tune it for the dataset used in this paper. The trained YOLO model can achieve high-precision detection of multiple types of targets such as people and vehicles while maintaining high speed. The output of the YOLO detection module is several bounding boxes with category labels and confidence levels, and this information will be passed to subsequent modules for further identification or processing.

3.2 ResNet image classification

This paper adopts ResNet-50 as the architecture of the classification network, and performs fine-tuning based on the pre-training of ImageNet to adapt to our self-built classification dataset. The functions of the classification module mainly include two aspects: one is to conduct fine classification of the specific target area detected after target detection; the other is to perform scene or attribute classification of the entire image. When training the ResNet classifier, we use cross-entropy loss as the objective function and optimize the network parameters through stochastic gradient descent. Due to the introduction of GAN data augmentation, we were able to expand the samples of minority classes in the training set, effectively alleviating the problem of class imbalance. The training results show that the classification module using residual networks has achieved excellent performance on our dataset, and the Top-1 accuracy has significantly improved compared to the equivalent-depth networks without residual structures.

3.3 Face recognition (Dlib/FaceNet)

The face recognition module consists of two parts: face detection and face representation recognition. First of all, we use the face detector of the Dlib library to locate the face position in the image. Dlib uses linear classifiers and the enhanced Histogram of Oriented Gradients features for face detection. A detection model based on convolutional neural networks can also be selected to achieve higher accuracy. While detecting the face, the face key point location algorithm provided by Dlib can extract 68 feature points of the face for face alignment. We perform affine transformation alignment on the detected original face images based on the position information of the eyes and nose, standardize them to a uniform size, and provide a more stable input for subsequent recognition.

In the face recognition stage, we adopt the FaceNet deep model to extract the face feature vectors. FaceNet uses the Inception architecture to map the aligned face images into 128-dimensional feature embedding vectors. Different from the traditional Softmax classification loss, FaceNet adopts the triplet loss function to directly optimize the distance relationship in the embedding space. The formula for triple loss is:

$$L(A,P,N) = \max\{//f(A) - f(P)//^2 - //f(A) - f(N)//^2 + \alpha, 0\}$$

Here, A represents the anchor sample (A face of A certain person), P represents the positive sample (belonging to the same person as A), N represents the negative sample (not belonging to the same person as A), f (X) is the feature vector extracted by the network, and α is the preset margin. The objective of the loss function is to reduce the distance between the anchor and the positive sample as much as possible during the optimization process, while the distance between the anchor and the negative sample is at least one α greater than the anchor-positive distance. When the latter is still less than the former plus margin, the loss is positive and drives the network to continue optimizing; otherwise, it is considered that the condition has been met. By training on a large number of face triples, this loss prompts the network to learn a highly discriminative feature space, in which faces of the same identity cluster together and faces of different identities are relatively far apart. After the training of the FaceNet model is completed, we can use it to convert any face image into a 128-dimensional feature vector. In the recognition stage, for each face in the input image, we calculate its FaceNet feature vector and then compare it with the face feature vectors of known identities in the database. The common practice is to calculate the Euclidean distance or cosine similarity. If the distance is lower than the preset threshold, it is judged as the same person; otherwise, it is regarded as an unknown face or a mismatch. Because the FaceNet feature has a strong discriminative ability, it can meet the requirements of high-accuracy face verification and recognition in our system. It should be pointed out that in order to achieve better performance, we also utilized GAN data augmentation technology to generate some variant images of human faces to expand the face training set, thereby making the FaceNet feature extractor more robust to pose and expression changes.

3.4 GAN data Augmentation

Haoyu Hu, Xinyue Xu, Hu Ke, Zhang Qing, and Li Feng, "Research on Intelligent Visual Recognition Methods Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 12, pp. 541-545, 2025.

The data augmentation module generates additional training samples based on the Generative Adversarial Network (GAN) to improve the generalization ability of each task model. GAN consists of a generator G and a discriminator D, and both sides are continuously optimized through the game process. The discriminator D(x) learns to determine whether the input sample is real data or generated data; Generator G(z) generates forged samples starting from the random noise z, attempting to deceive the discriminator. The objective functions of the two are usually expressed as the minimax problem:

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p(z)}[\log(1-D(G(z)))]$$

In the optimal case, the data distribution generated by the generator is consistent with the real data distribution, and the accuracy rate of the discriminator drops to 50%. The trained generator can be used to output a continuous stream of synthetic image samples. In our system, different GAN models were trained as needed for data amplification: For general objects, we trained a convolution generator similar to DCGAN to generate new images with a style similar to that of real images; For human faces, we adopted conditional GAN or diffusion models to generate face images with perturbations added on the basis of specific identities. For the datasets of some classification tasks, GAN can also be used to generate category samples that are difficult to obtain. For example, in the animal classification subtask, we used the WGAN-GP model to generate realistic images of some rare animals, thereby making the training set more balanced. The generated expanded samples are generally filtered through manual screening or discriminator scoring before being added to the training set to ensure high authenticity and diversity.

It should be noted that different tasks have different requirements for generating samples. For example, in the task of object detection, in order to maintain the true combination of the background and the object, we may adopt the method of generating a single object image and then superimposing it with the real background to enhance the data; The image classification task can directly use the entire image generated by GAN. Our experiments also found that the performance improvement brought by GAN enhancement is closely related to the quality of the generated samples. If the generator is not adequately trained, resulting in obvious artifacts in the samples, it may interfere with the normal learning of the model. Therefore, when applying GAN enhancement, we ensure that the discriminator loss converges and the generated samples are difficult to distinguish from the real samples in terms of human perception.

3.5 Image preprocessing

We designed an image preprocessing module for problems such as uneven illumination and noise interference, and comprehensively improved the quality of the original image by combining mathematical morphology operations with classical image processing methods: In grayscale images, the contrast is enhanced first by histogram equalization or adaptive histogram equalization. Then, based on the characteristics of the noise, appropriate structural elements are selected to perform the open operation to remove salt-and-pepper noise and the closed operation to fill the internal holes of the target, thereby maintaining the integrity of the shape while denoising. The color image is first converted to the HSV space. The light spot is reduced by median filtering or morphological operations on the luminance channel. Then, the contour is extracted with the help of edge detection and morphological gradient. Finally, it is restored to RGB and input into the depth model.

IV. Experiment and Results

In the face recognition experiment, we trained based on the FaceNet model and adopted the semi-hard triplet mining strategy, achieving a verification accuracy rate of approximately 99.2% on the LFW dataset. This result is very close to the 99.63% benchmark reported in the original FaceNet paper, indicating that the feature representation and discrimination ability of the model have been fully exerted in this implementation. For extreme conditions such as strong light and posture changes, we attempt to add simple morphological smoothing operations at the input end. In the comparative tests of a small number of self-collected samples, the recognition confidence showed a slight improvement after preprocessing, and the misjudgment decreased. However, due to the limited sample size, this phenomenon still requires larger-scale experiments to verify its universality.

To evaluate the gain effect of GAN in scenarios with insufficient samples, we manually deleted 20% of the training images in the CIFAR-10 classification task and used the pre-trained DCGAN to generate an additional 100 synthetic samples for each category for supplementation. The accuracy rate of ResNet trained with this enhanced dataset on the test set increased from 91.0 % to 93.2 %, showing that GAN samples have a moderate positive effect on the model when the data volume is limited. Meanwhile, in the Pascal VOC object detection experiment, we adopted CycleGAN to convert some daytime images into nighttime styles. After expanding the training set, the mAP of the model on the nighttime test set increased by approximately 3.5%. These results indicate that in cases of data distribution offset or difficulty in obtaining real samples, high-quality GAN synthetic images can provide valuable supplements to deep models.

In the morphological preprocessing experiment, we selected 50 night street view images to evaluate the pedestrian detection performance of YOLOv3. Without preprocessing, the model detected a total of 35 pedestrian targets. After applying corrosion denoising and contrast enhancement, the number of detection targets increased to 38, and the missed detection rate decreased by approximately 8.6%. In the noise robustness test of face recognition, we superimposed random noise on some face images, reducing the accuracy rate of FaceNet to 90%. However, after denoising through morphological opening operations, the accuracy rate recovered to 96%. Although the sample size of the above-mentioned experiments is relatively small and insufficient to prove statistical significance, the preliminary results indicate that

Haoyu Hu, Xinyue Xu, Hu Ke, Zhang Qing, and Li Feng, "Research on Intelligent Visual Recognition Methods Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 12, pp. 541-545, 2025.

morphological preprocessing, as an image enhancement method with low computational cost and simple implementation, can bring certain performance improvements to the target detection and recognition model in low-quality or low-light environments and has the potential for deployment in real-time or resource-constrained systems.

V. Conclusion

This study shows that the advantages of different deep learning models can be complemented and integrated through architectural design, thereby meeting the requirements of complex application scenarios. In future work, we believe that with the continuous evolution of deep learning technology and in-depth research on multi-module integration, intelligent visual recognition systems will play a greater role in fields such as security monitoring, autonomous driving, and industrial inspection, and develop in the direction of higher intelligence and wider adaptability.

Acknowledgment

Supported by the Undergraduate Research and Innovation Fund Project of Anhui University of Finance and Economics (Project Approval Number: XSKY24159).

References

[1] Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.

[2] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

[3] Schroff F., Kalenichenko D., Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 815-823.

[4] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative Adversarial Nets. In Advances in Neural Information Processing Systems (NIPS), 2014, 27: 2672-2680.

[5] Gonzalez R. C., Woods R. E. Digital Image Processing (Third Edition). Prentice Hall, 2008.

[6] King D. E. Dlib-ML: A Machine Learning Toolkit. Journal of Machine Learning Research, 2009, 10: 1755-1758.