基于对抗式师生框架的面向对象鲁棒人脸识别系统设计

2025年5月11日星期日

**Abstract：**

Face recognition technology is one of the most mainstream means of identity authentication and plays a key role in ensuring system security. However, with the rapid development of deepfake image and video generation technology, face recognition systems are facing an increasing threat of forgery. Most of the current mainstream methods rely on deep convolutional neural networks (CNNs) [1]to extract 2D or 3D features of faces, and compare them with the database after alignment to verify the authenticity of the input face. This paper introduces a teacher-student model structure for face recognition: the teacher model is a binary classifier trained on large-scale data, which can efficiently distinguish between real and forged faces, but the model is large and not suitable for direct deployment; the student model simulates the teacher's output on real samples, while maintaining a clear output difference with the teacher model on fake samples. This difference can become a strong forgery detection signal, which helps to improve the system's generalization ability when no forged samples have been seen.[2]

1. **Introduction：**

Since the 1960s, when Bledsoe proposed the first manual face recognition system, the field has attracted widespread attention. In 2014, Facebook’s DeepFace significantly advanced recognition accuracy, achieving 97.35% on static photos and over 91% on video frames[3]. In 2015, Google’s FaceNet and Microsoft’s DeepID further pushed accuracy beyond 99%, making face recognition commercially viable for the first time. Since then, face recognition systems have been widely adopted in scenarios such as security surveillance, smartphone unlocking, airport boarding, and payment authentication. These applications vary from static image matching to real-time video analysis, demanding diverse feature extraction techniques.

In recent years, the rise of deepfake techniques has introduced new security threats to face recognition systems. In response, researchers have started incorporating forgery-aware modules to detect synthetic facial content. Nevertheless, most modern face recognition pipelines still follow the conventional four-step structure: **detect ⇒ Face alignment ⇒ Facial feature vector representation ⇒ Similarity-based classification[3]**, which remains the foundation for both academic research and industrial deployment.

1. **Related work：**

**ResNet**: Most current forgery detection models are still based on deep convolutional neural network (CNN) architectures to extract discriminative features from input images. This type of model has strong feature modeling capabilities by stacking convolution, activation, and pooling modules layer by layer. However, as the network depth increases, the traditional CNN architecture is prone to the vanishing gradient problem: the features extracted by the previous layer gradually weaken during the layer-by-layer transmission process, and are not fully integrated with the high-order semantic information of the next layer, which ultimately leads to a decline in training performance and difficulty in effective network convergence. This phenomenon limits the expressiveness and generalization performance of the model under deeper structures.[4]

To solve this problem, He et al. proposed the residual network (ResNet) structure in 2015 [5]. Its core innovation is to introduce residual connections, which directly add the input of the previous layer to the output of the next layer through "skip connections", thereby alleviating the gradient decay phenomenon in deep networks. Specifically, traditional neural networks attempt to directly learn the mapping from input x to target H(x), while ResNet learns the residual function

F(x)=H(x)-x

and its final output is represented as

H(x)=F(x)+x

ResNet allows the network to learn residual mappings instead of complete transformation functions, which greatly improves the training stability and performance of deep networks. Due to its excellent feature extraction capabilities and good scalability, ResNet has become the default backbone network for many mainstream face recognition systems. It is widely used in advanced models such as ArcFace, MagFace, and CurricularFace, and has demonstrated leading recognition accuracy on datasets such as LFW and MS-Celeb-1M.

**Mesonet**: In addition to the method of deep convolution processing of images through residual functions, mesonet proposed by Afchar et al. takes into account the artifacts and unnatural phenomena that may appear in compressed images and proposes to use a medium-layer convolution model to focus on the edges, blur, and reconstruction traces of the image to identify the authenticity, rather than the deeper authenticity verification of faces, pupils, etc. At the same time, this model introduces the activation layer into the deep convolutional neural network. In the activation layer, they use the activation function ReLU（fig 1.） to weight the feature map obtained by the convolution kernel in the previous layer, retain the part that meets the requirements and convert the unnecessary part to 0, thereby reducing the impact of the disappearance of the steps. [6]

图片包含 游戏机, 工具

AI 生成的内容可能不正确。  
**fig1**[6]

（ReLU is for Nonlinear feature extraction Used to reduce the impact of gradient disappearance  
Sigmoid is It is another activation function. If the output score is less than 0.5, it is changed to 0 to be judged as false. If it is greater than 0.5, it is changed to 1 to be judged as true.）

**GenDet：**However, considering that the model for generating fake images will continue to change with the evolution of technology, if the detection system is to have good adaptability, it is necessary to introduce a large and constantly updated dataset during the training phase. However, in practical applications, such large-scale data-dependent models are difficult to deploy on resource-constrained devices (such as edge devices or embedded systems). To solve this problem, Zhu et al. (2024) proposed a lightweight forgery detection framework GenDet with a teacher-student structure, which is designed to enhance the generalization detection ability of forged images generated by "unseen generators".

In GenDet, the teacher model (Teacher) is a network trained by a traditional binary classifier that can distinguish real images from forged images; the student model (Student) is trained by imitating the output of the teacher model on real images, while deliberately maintaining a large output difference when forged images are input,（fig2） thereby achieving stronger forgery separation capabilities.[2]

图表, 散点图

AI 生成的内容可能不正确。**fig2**[2]

In addition, the authors also designed an adversarial feature augmentation module (Feature Augmenter) to generate more types of forged feature vectors and further enhance the model's ability to separate "unknown forged images".

This method achieved leading performance on both UniversalFakeDetect and GenImage datasets, with an average accuracy of 94.4% and mAP of 95.6%, significantly exceeding traditional CNN detectors and classification methods based on pre-trained features, verifying its lightweight, high discrimination and strong generalization capabilities.

**III. System Analysis and design:**

**3.1 Requirement Analysis：**

Our goal is to extract the mid-level convolutional features of the face in the input image (such as edges, blur, reconstruction traces) through a convolutional neural network and use the learned teacher-student model to detect the extracted features.

**Functional requirement:**

1. Users can upload facial images for identification or verification
2. The system should automatically detect and align input face images
3. The teacher model provides true/false output based on the dataset
4. The student model imitates the output of the teacher model for real images
5. The student model keeps the difference between the output of the teacher model for fake images
6. The discriminator gives a confidence interval of the authenticity of the image based on the output of the student model
7. The discriminator generates the result and saves it in the database
8. The system can be updated in real time based on the face database

**Nonfunctional requirement：**

1. The system maintains a low latency rate
2. The system can be deployed on edge devices
3. Database protects user face information

This section briefly discusses the functions that face recognition systems need to accomplish and ensure the sustainability of the system.

**3.2 Use Case Modeling**

1. **Train Teacher and Student Model**

**Actor:** Admin, System.

**Description:** The administrator specifies the data sets that the teacher model and the student model need to learn. The student model imitates the output of the teacher model on the real data set, and the system dynamically adjusts the learning content.

**Steps:**

1. Administrators provide datasets containing real and fake faces.
2. The system assigns the dataset to the teacher model for learning.
3. The teacher model learns the data set and classifies real and fake faces.
4. The student model imitates the output of the teacher model on real faces.
5. Systematic evaluation of the performance of the student model.
6. The system dynamically adjusts the dataset used to train the robustness of the student model based on the evaluation results.
7. **Deploy and Use Student Model for Inference**

**Actor:** User, System.

**Description:** After the user puts in the image, the system will process the image into a readable format for the student model and let the student model output the result.

**Steps:**

1. The user uploads a face image.
2. The system aligns the face image.
3. Student model extracts features and outputs prediction.
4. The system classifies the image based on learned output patterns.
5. The system outputs the result to the user.
6. The system stores the result in database.
7. **Retrieve Detection Records**

**Actor:** Admin, System.

**Description:** The admin accesses the system to retrieve historical detection results of uploaded facial images, including timestamps, output labels, and confidence scores.

**Steps:**

1. Admin login system
2. Administrator requests test record
3. The system accesses the test records from the dataset
4. The system returns the detection record to the administrator

**图示

AI 生成的内容可能不正确。**

**Fig3**

In the use case diagram(fig3), exporting email as an optional feature extends from the view detection result.

**3.3 Object-Oriented Analysis**

The classes in the class diagram (analysis)(fig4) include：

* User:

System users, upload images and view recognition results.

* Admin:

Administrator, responsible for model training and log viewing.

* FaceImage:

User-uploaded face images, including meta information

* AlignedImage:

Aligned face image as model input.

* TeacherModel:  
  The discriminator used in the training phase to guide the student model learning (only for training).
* StudentModel:  
  Lightweight face recognition model used in the deployment phase
* DetectionResult:

True/false classification results and confidence levels output by the system.

* DectionLog:  
  True/false classification results and confidence levels output by the system.
* TrainingControl:

Controller module that triggers the training process.

* Database:

Persistent components for storing system data such as images, model parameters, and log records.

图示

AI 生成的内容可能不正确。

**Fig4**

**3.4 Object-Oriented Design**

**图示

AI 生成的内容可能不正确。**

**Fig5**

This class diagram in design level(fig5) clarifies the functions that the system will implement, ensures flexible switching between model updates and deployments, and lays the foundation for subsequent system implementation and maintenance.

**3.5 Sequence Diagram**

The face recognition system can be roughly divided into two parts: the training phase and the inference phase.  
The training phase use

1. **Training sequence diagram:**

The training sequence diagram(fig6) is used to describe the interaction sequence and behavior logic of the training phase. The Admin sends a training command to the TrainingController. After receiving the start training command, the TrainingController sends a dataset containing real and fake faces to the TeacherModel. After the TeacherModel is trained, it passes the output of the real face to the StudentModel. The StudentModel tries its best to imitate the output of the TeacherModel when analyzing the real face dataset. Finally, the training results are stored in the Database.

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

**Fig6**

1. **Inference sequence diagram:**

The inference sequence diagram(fig7) is used to describe the logical model of the user and the system in the reasoning stage. The User uploads the picture to the System, the System sends the image to the aligner, and issues an alignment instruction. After the face is processed by the aligner, the aligner transmits the processed face image back to the system. The System sends a prediction instruction to the StudentModel. After the StudentModel returns the predicted score to the System, the System will process the score and send the confidence interval of the image to the User and keep it in the Database.

**图示

AI 生成的内容可能不正确。**

**Fig7**

**IV. specification：**

Admin, as an actor in the training module, actively initiates learning tasks to the system. After receiving the tasks, the System will dynamically arrange a dataset containing real and fake faces for the teacher model to train. The TeacherModel, as a dichotomizer, divides the face images in the data set into real images and fake images and gives corresponding confidence scores, while the StudentModel only imitates the output of the TeacherModel on real images, such as confidence scores. For fake face images, the StudentModel will maintain a large difference from the output of the TeacherModel for fake face images. This difference can make the output of real face images and fake face images on the StudentModel have a greater difference.（fig2）

As the active actor of the inference module, the user sends a face picture in JPG format to the System. The System will align the face picture provided by the User for better judgment by the StudentModel. The StudentModel extracts the mid-level features of the face image as the basis for judgment and gives the confidence interval for this image. The System converts the confidence interval output by the StudentModel into a result that can be understood by Users and Admin, and outputs it.

**V. Conclusion:**

This paper systematically reviews some of the main technologies of face recognition and proposes a detection model of teacher-student model + middle convolution layer. Compared with the traditional detection model, this detection model is lightweight and has a more obvious difference between true and false outputs. At the same time, this is a middle convolution layer model, so the model is also robust to compressed face images. This paper adopts the OOAD process including requirement analysis, OOA, OOD, sequence diagram, and specification, which is a clear separation of responsibilities in the system structure, and is convenient for subsequent module updates and maintenance.

**Reference:**

[1] X. Wang, J. Peng, S. Zhang, B. Chen, Y. Wang, and Y. Guo, “A Survey of Face Recognition,” Dec. 26, 2022, *arXiv*: arXiv:2212.13038. doi: 10.48550/arXiv.2212.13038.

[2] M. Zhu *et al.*, “GenDet: Towards Good Generalizations for AI-Generated Image Detection,” 2023, *arXiv*. doi: 10.48550/ARXIV.2312.08880.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-Level Performance in Face Verification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA: IEEE, Jun. 2014, pp. 1701–1708. doi: 10.1109/CVPR.2014.220.

[4] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” *Artif. Intell. Rev.*, vol. 57, no. 4, p. 99, Mar. 2024, doi: 10.1007/s10462-024-10721-6.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[6] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2018, pp. 1–7. doi: 10.1109/WIFS.2018.8630761.