JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 11, pp. 3281-3290, Nov. 2024

Check for updates

자기 지도 학습에 기반 한 얼굴 인식 정확도 향상에 관한 연구

손 천 샤[1] · 신 승 수[2*]
[1]동명대학교 컴퓨터미디어공학과 박사과정
[2]동명대학교 정보보호학과 교수

A Study on Improving Face Recognition Accuracy Through Self-Supervised Learning

Chen-Xiao Sun[1] · Seung-Soo Shin[2*]

[1]Doctoral Course, Dept. of Computer and Media Engineering, Tongmyong University, Busan 48520, Korea
[2]Professor, Dept. of Information Security, Tongmyong University, Busan 48520, Korea

[요 약]

얼굴 인식의 경우, 기존 방식은 분산되지 않은 데이터를 다룰 때 한계를 드러내고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 강력한 트랜스포머 아키텍처 하에서 비지도 학습의 잠재력을 활용을 한다. 이를 위해 비디오 자동 처리 방식과 2단계 훈련 모드를 개발 하였다. 이 방법은 보다 효과적인 훈련 과정을 위해 풍부한 비라벨 데이터와 고품질 라벨링된 데이터를 모두 활용하며, 각각 자기 감독 대조 손실과 감독 분류 손실을 활용한다. 실험 결과 다양한 데이터 분포에 대한 일반화 및 정확도 향상 측면에서 제안한 접근법의 우수성을 보여준다. 얼굴 인식을 위한 비지도 학습의 효과를 검증한 본 논문은 특히 분포가 다른 데이터를 처리하는 데 있어 얼굴 인식 기술의 발전에 기여할 수 있을 것으로 기대된다.

[Abstract]

In the context of face recognition, traditional methods have limitations when dealing with out-of-distribution data. To address these challenges, our study leverages the potential of unsupervised training within the transformer architecture. We developed an automatic video processing approach and a two-stage training model. This method utilizes both abundant unlabeled data in the wild and high-quality labeled data to enhance the training process, employing self-supervised contrastive loss and supervised classification loss, respectively. Experimental results demonstrate the superiority of our approach in terms of generalization across diverse data distributions and improved accuracy. This study validates the effectiveness of unsupervised training for face recognition and is expected to contribute to advancements in handling out-of-distribution data.

색인어 : 얼굴 인식, 자기 주도 학습, OOD, 트랜스포머, 대조 학습
Keyword : Face Recognition, Self-Supervised Learning, Out-of-Distribution, Transformer, Contrastive Learning

http://dx.doi.org/10.9728/dcs.2024.25.11.3281

Received 04 September 2024; **Revised** 14 October 2024
Accepted 31 October 2024

***Corresponding Author**; **Seung-Soo Shin**
Tel: ※ 개인정보 표시제한
E-mail: shinss@tu.ac.kr

Ⅰ. Introduction

Face recognition aims to identify or match a person's identity against a gallery of various faces. It is widely used in security, electronic device unlocking, and other fields, demonstrating the importance and significance of its research[1]. Because face recognition systems typically operate in natural settings, they require a high level of robustness, for example, robustness to changes in lighting, makeup, occlusions of key facial features, and so on. Therefore, it is necessary to conduct more in-depth research to address the challenges that may exist in the real world.

Despite existing face recognition models such as ArcFace[2], VggFace[3], and benchmark networks like ResNet[4] and Transformer(ViT)[5] achieving impressive performance on LRW[6] and other datasets, current methods still encounter problems due to being limited to constrained scenarios and closed datasets. These limitations include a lack of diverse racial samples and various image resolutions, or occlusions from masks, hands, sunglasses, and different lighting conditions, leading to significant performance drops on out-of-distribution (OOD) test data[7],[8].

Currently, accurate face recognition in unconstrained scenarios remains a highly challenging task. Therefore, it is essential to propose an effective method to address this issue. This paper proposes a self-supervised pretraining-based face recognition method to address the aforementioned poor OOD performance issues. This is feasible because we also propose an automated data crawling and processing tool to provide large amounts of training data for unsupervised learning. Additionally, the improvement in recognition accuracy is due to further supervised fine-tuning of the model, leveraging existing labeled datasets. Our results show that our method outperforms existing methods in terms of accuracy.

While some methods incorporate dedicated OOD images, such as masking out parts of the face or introducing adversarial examples[9],[10], these datasets do not necessarily reflect the same distribution as natural images. In reality, out-of-distribution scenarios are more likely to involve conditions such as backlit or strongly lit conditions, occlusion by microphones, or unclear photos. The closest work to ours is presented in Hu et al.[10], which introduces an unsupervised facial representation learning method.

While Hu et al.[10] achieves decent performance using unsupervised learning, it has the following drawbacks: it employs margin loss as the optimization objective, which requires manually specifying a margin value and is not conducive to automatically learning the relative similarity of positive samples on large-scale datasets. Additionally, Hu et al.[10] only involves one phase of unsupervised representation learning, although it effectively utilizes large-scale unlabeled web videos, it overlooks the existence of annotated datasets, limiting further accuracy improvements.

The proposed model aims to leverage this abundance of freely available data on the Internet and address the shortcomings in the method Hu et al.[10]. The proposed model utilizes a transformer as the main feature extractor and employs contrastive learning for self-supervised learning during the pre-training phase. In the fine-tuning phase, supervised methods are used for personal identity learning. Additionally, to enable the method to leverage unlabeled massive data, we propose an automated video preprocessing pipeline, which can extract the face images of the same person in a video.

The composition of this paper is as follows. Chapter 2 introduces the datasets used in this paper along with other OOD-targeted methods, and explains a typical self-supervised method. Chapter 3 proposes an automatic data collection pipeline and a self-supervised learning-based face recognition method. Chapter 4 performs comparative evaluations of the proposed model and existing models with a final analysis. Finally, Chapter 5 presents the conclusions.

II. Related Works

2-1 Out-of-Distribution Face Recognition Datasets

In this paper, we utilize CFP-FP[11] and AgeDB-30 [12] as our out-of-distribution (OOD) datasets for performance evaluation. CFP-FP is a face recognition dataset designed to assess whether the current model can effectively learn the correlation between frontal

and highly angled profile faces. This dataset contains a significant number of profile face images, which better represent the distribution of facial images in natural scenes. As for the AgeDB-30 dataset, it captures not only pose variations and head rotations in the wild but also includes many images with occluded facial features.

Considering the existing methods, some of them also tried to solve the performance drop on OOD images. Xiang Xu et al.[9], they design an occlusion-balanced sampling strategy for training the network. Additionally, to ensure the network focuses on the unobstructed parts of faces, which are useful for recognizing identities, they propose an occlusion-aware attention mechanism. This mechanism allows the network to assign higher weights to more discriminative features.

Their results demonstrate that the approach is effective. Another way to deal with OOD data is to reconstruct the occluded parts before performing face recognition. In Fang Zhao et al.[13], they utilized an LSTM-based autoencoder to automatically restore the occluded face patches. This approach yields a certain degree of generalization; however, the quality of reconstruction is not guaranteed and may fail to restore the actual person, potentially introducing additional noise. More importantly, OOD data encompasses more than just occlusions. Our method leverages a large amount of in-the-wild data to more closely approximate the real data distribution.

2-2 Self-Supervised Learning

Self-supervised learning has received considerable attention in recent years due to its powerful potential and has been widely applied in natural language, natural image, audio, and multimodal representation learning[14],[15].

Contrastive learning, as a paradigm of self-supervised learning, has attracted a lot of attention due to its outstanding performance. One of the more typical methods is contrastive predictive coding (CPC)[16]. It mainly constructs positive and negative samples and uses a pretext task to minimize the distance between positive samples and maximize the distance between negative samples. This can be represented by the following formula:

$$L_{cpc} = -E\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right] \quad (1)$$

This loss is called the InfoNCE loss. Given a set of X random samples containing one positive sample from $p(x_{t+k} \mid c_t)$ and $N-1$ negative samples from the distribution $p(x_{t+k})$, the InfoNCE loss aims to maximize the encoding $f(\cdot)$ of the positive samples among all the samples' encodings.

In this paper, we designed a novel network architecture based on insights gained from previous research, while also enhancing the robustness of the dataset. These improvements significantly elevate the model's efficiency compared to existing approaches. A detailed discussion of the network design and dataset enhancements can be found in Chapter III of this paper.

Ⅲ. Improved Self-Supervised Learning-based Face Recognition

3-1 Overview

The full pipeline of our method is shown in Fig. 1, which can be mainly divided into two main stages a self-supervised pre-training stage and a supervised fine-tuning stage. For the first stage, a large amount of image data is required for the training set. Therefore, a prerequisite step called video-data preprocessing is introduced before the first stage. This step produces a vast number of images with pseudo labels, which are determined by the video name (which is unique) and the face tracking number. After obtaining the images, we propose a self-supervised pretraining method to extract the hidden representations and high-level semantics from these large volumes of images.

Then, the trained network is equipped with a classification head (prediction layer) and uses the existing labeled images for supervised fine-tuning. We will first introduce the video-data preprocessing pipeline in Section 3.2, followed by explanations of the two stages in Sections 3.3 and 3.4, respectively.

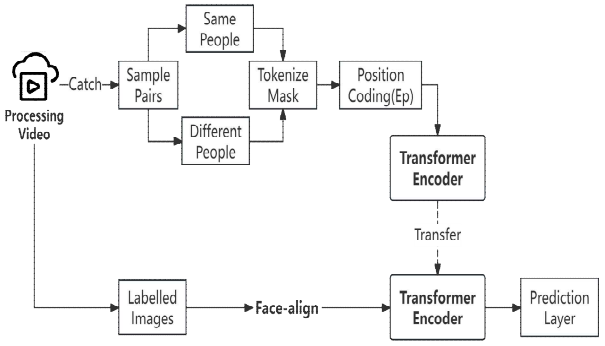


Fig. 1. Overview of proposed face recognition method

3-2 Automatically Video-Data Preprocessing

The aim of automatic video data preprocessing is to automatically process videos without manual labeling, resulting in a series of facial images with identity-related pseudo-labels. The full pipeline is illustrated in Fig. 2. Firstly, we obtain a list of celebrity names from Wikipedia. For each of these names, we constructed several search sentences, including '{name} public interview' and '{name} interview,' where {name} is a placeholder. These search sentences were used as keywords to crawl the YouTube website. The URLs of the searched video URLs were recorded and downloaded. After scrawling and downloading a random video, we then perform scene/shot detection. Subsequently, within different scenes, we conduct a face detection algorithm, retaining detection boxes with an area greater than a threshold T in an image.

If the overlapping portion of detection boxes across consecutive M frames exceeds 50%, it is considered that the detection boxes within this tracking segment belong to the same face. For the same segment, there may be more than one tracked face. To increase the robustness of the preprocessing, a certain tolerance M_t is maintained during frame tracking, allowing for a degree of missed detections or false positives.

After the aforementioned processing, a video will be organized as follows: a series of scene folders, each containing a series of tracking results. Each individual tracking result within a scene is considered to belong to the same person's facial images. Conversely, different tracking results within the same scene are considered to belong to different people's facial images. Due to post-editing of the video, the same person may appear in different scenes, so we do not make identity determinations for different tracking results in different scenes.

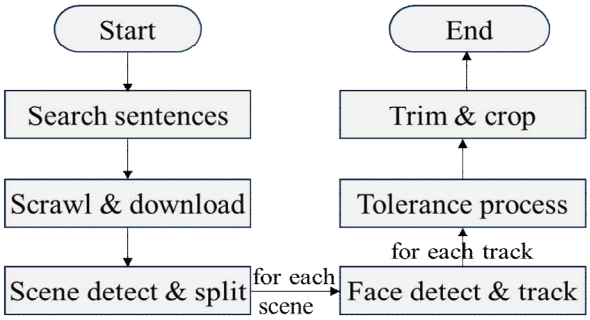


Fig. 2. The processing pipeline of automatically video-data preprocessing

3-3 Self-Supervised Pre-Training

After obtaining the large image data, our method would first utilize self-supervised learning to capture the useful features related to face identity, as seen in Fig. 3. The key to self-supervised training lies in constructing positive and negative sample pairs. For this stage, we define positive sample pairs as images from the same tracking result within the same scene (v_i, v_i^*), where (v_i, v_i^*) is same person's face pictures , it just catches a lot of sample pairs in the video, while negative sample pairs are any pair (v_i, v_k), for $k \neq i$, where (v_i, v_k) is different person's face pictures, it just catches a lot of sample pairs in the video. It's worth mentioning that, due to the lack of manual annotations, it's not possible to accurately obtain labels for positive and negative sample pairs. Therefore, negative samples here may also come from facial images of the same person. This situation tends to diminish as the size of the training dataset increases. To simplify the method introduction, we use one branch as an example. For the input face image v_i, we first split it into equal-sized patches. This step can also be referred to as tokenization, and we use linear projection $\mathrm{Pro}(\cdot)$ to ensure that these tokens are projected to match the dimension of the subsequent transformer layer:

$$v_i = Mask_\alpha(\mathrm{Pro}\, v_i + E_p) \quad (2)$$

The E_p represents the position embedding corresponding to the projected tokens. To increase the diversity of

training samples, we applied a random masking process to tokens with a ratio of α. The masked tokens are then fed into the Transformer The encoder consists of multiple layers of transformer layer, each layer can be abstracted as:

$$y = layer(x;\ MSA,\ LN,\ MLP) \tag{3}$$

Where MSA represents multi-headed self-attention, LN represents layer normalization, and MLP stands for multilayer perceptron blocks containing residual connections. x can represent either the input token or the intermediate representation of the layer.

After transforming by several transformer layers, the final representation x goes through layer normalization and mean pooling to obtain the final embedding :

$$Emb = Pool(x;\ LN) \tag{4}$$

Considering that our task at this stage is to learn the features of facial images, specifically to make the embeddings of the same person's face as close as possible, while keeping the embeddings of different individuals' faces as far apart as possible. Therefore, during training, we input $2N$ images:

$$\{(v_i, v_{i^*}), i = 1, \cdots, N\} \tag{5}$$

The similarity of positive pairs is denoted as s_{i,i^*}, which is computed by the dot product of the normalized embeddings of Emb and Emb^*. As for the negative pairs, for each, we randomly select one image where $v_k(k \neq i)$. Finally, we optimize this transformer encoder by minimizing the following loss function:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{e^{\frac{s_{i,i^*}}{\tau}}}{\sum_{k\neq i} e^{\frac{s_{i,k}}{\tau}} + e^{\frac{s_{i,i^*}}{\tau}}}\right] \tag{6}$$

Where τ is the temperature.

The loss function in equation (6) is contrastive loss, also known as InfoNCE loss introduced in section 2.2. This loss allows the network to learn the relative similarity between positive and negative pairs, thus uncovering the semantic features of facial images related to identity. Because the normalized dot product is equivalent to cosine similarity, the larger the value the more similar it is. Equation (6) needs to maximize the value of each batch diagonal seen in Fig. 4.

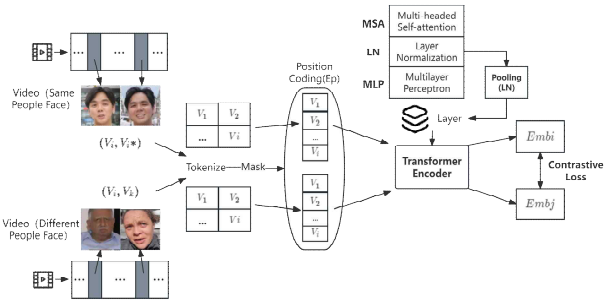


Fig. 3. The sequence of processes of self-supervised learning pre-training

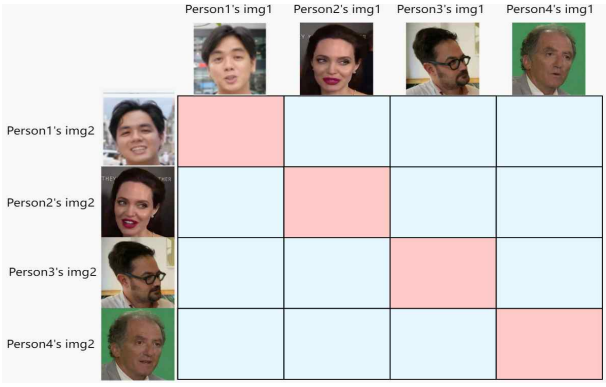


Fig. 4. Pairwise cosine similarity

3-4 Supervised Fine-Tuning

After the first stage of large-scale pretraining, the transformer encoder has become a powerful feature extractor related to identity. However, due to the lack of labels in self-supervised learning, false negative samples are inevitable, which may affect its performance on precise tasks, such as face recognition. Moreover, there are already many available images with identity labels that can be utilized to further enhance the method's performance. Therefore, in the second stage, we propose supervised fine-tuning.

As indicated by the arrows in Fig. 5, during the fine-tuning stage, the initial weights of the transformer encoder are derived from the checkpoint in the first stage, and all parameters are kept trainable during the fine-tuning stage. Additionally, in order to enable the model to classify based on embeddings, we added a prediction layer on top of the transformer encoder to

predict the identity labels corresponding to the input images. The prediction layer consists of a fully connected linear transformation layer and an activation function, with the final recognition loss function being:

$$L_{rec} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\log\frac{e^{PL(emb_{i,c})}}{\sum_{j=1}^{C} e^{PL(emb_{i,j})}} y_{i,c} \quad (7)$$

Where PL represents the prediction layer, C is the total number of identities.

After training the data in the second stage, we can then use the model's output embeddings to infer the test set. The image that needs to be recognized is referred to as a query image, while the existing image-identity pair data in the dataset is referred to as the database. We encode both the query image and all images in the database, then retrieve the image from the database with the highest similarity to the query image embedding. The corresponding identity of the retrieved image is considered to be the identity of the face image that needs to be recognized.

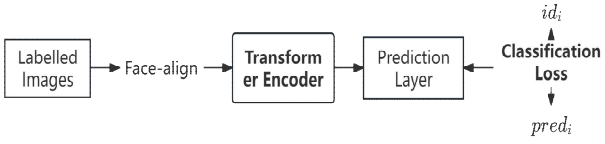


Fig. 5. The sequence of processes of supervised fine-tuning

Ⅳ. Evaluation and Result

In this chapter, we conducted a feasibility and performance evaluation of our method using a public dataset. Furthermore, to assess the relative effectiveness of our approach, we also conducted testing, compared it against other methods under the same hardware environment, and provided a discussion and explanation regarding the results.

4-1 Experiments

1) Dataset Collection and Dataset Labeling

In this paper, for the first stage of self-supervised training, we use publicly available unlabeled images cropped from the Internet. For the final face recognition model, we utilize challenging cross-pose and cross-age datasets, namely CFP-FP and AgeDB, respectively.

The details of the dataset collection are as follows: for self-supervised training, we collected about 5,000 videos. Each video contains several track segments, contributing a total of about 75,000 images. We use 10% of these images as a validation set to select the best model. For face recognition, the CFP-FP dataset contains 3,500 genuine pairs and 3,500 imposter pairs in the evaluation set. The AgeDB comprises 16,488 images across 568 identities, with ages varying from 1 to 101 years. These datasets can be used to evaluate the efficiency and robustness of our model.

In general, our dataset consists of many grayscale photos, as shown in Fig. 6.

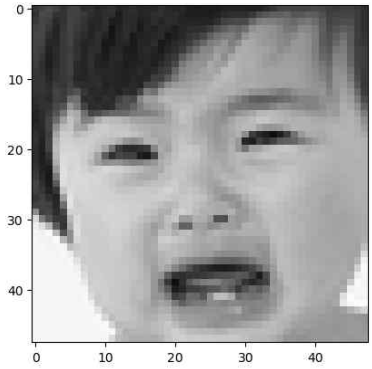


Fig. 6. Dataset composition

2) Dataset Training

The testing environments for testing are described in Table 1.

Table 1. Testing environments

Resources	Specification
GPU	GeForce RTX 2080 Ti
Language	Python 3.8
Libraries	Torch 1.18.1

4-2 Performance Evaluation

In this section, we conduct a series of experiments on the same 2080 Ti GPU hardware to fairly measure the face recognition performance and speed of different methods.

1) Efficiency in Face Recognition Model

We measured the average time (in milliseconds)

taken by different methods to recognize an image on the same hardware equipment, the results of which are shown in Table 2. From Table 2, it can be seen that there is still a gap in efficiency between our method and ArcFace. However, it's still better than Vanilla ViT. To enhance reproducibility in our experiments, it is crucial to detail the computational requirements for each method used in our study. This includes the number of parameters, the number of epochs, training time, and inference time. These metrics provide insights into the efficiency and scalability of each approach.

A description of the performance metrics of our proposed method and other methods is shown in Table 2.

- Inference Speed: This refers to the average time taken by the model to make a prediction on a single input image, measured in milliseconds (ms). Lower values indicate faster inference, which is crucial for real-time applications.
- Number of Parameters: This indicates the total number of trainable parameters in the model. A higher number of parameters can suggest a more complex model, but it also may lead to increased training time and risk of overfitting.
- Training Time: This metric represents the total time taken to train the model, typically measured in hours. It can vary based on the model architecture, the size of the dataset, and the computational resources used.
- Epochs: This denotes the number of complete passes through the training dataset. More epochs may improve model performance but can also lead to overfitting if not managed properly.

By detailing these computational requirements, we aim to improve the reproducibility of our findings. Researchers can use these metrics to better understand the trade-offs between model complexity, training efficiency, and inference speed when selecting methods for facial expression recognition tasks.

Our final accuracy is shown in Fig. 7. As can be seen in the figure, the lowest false alarm rate is only 0.01 while Happy is the most accurate detection rate in this paper, reaching 99%.

Table 2. Performance metrics of difference methods

Method	Inference Speed (ms)	Number of Parameters	Training Time (hours)	Epochs
ArcFace	43	89M	24	100
VGGFace	30	138M	30	50
ResNet34	33	21M	15	100
Vanilla ViT	68	86M	40	200
Proposed Method	67	60M	35	150

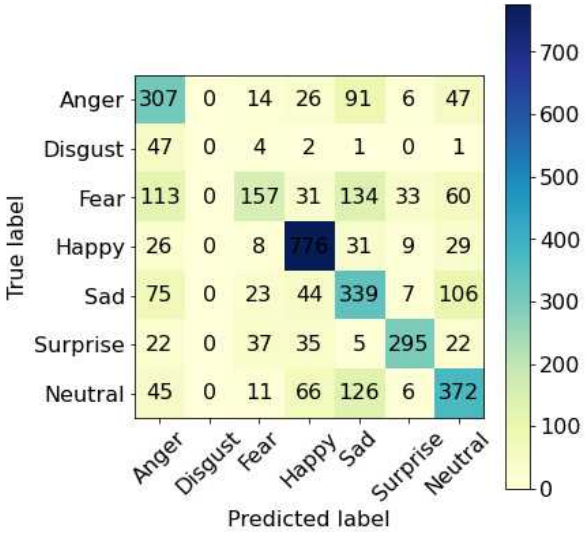


Fig. 7. Accuracy matrix

As show in Fig. 7, Structure of the conclusion matrix: the horizontal axis represents the predicted labels of the model, i.e., the emotions predicted by the classifier. The vertical axis represents the true label, i.e., the actual emotion category. The number in each cell indicates the number of emotions predicted by the classifier into a certain emotion category under that real emotion category. For example, “307” in the first row indicates that the number of times the model correctly predicts anger when the true emotion is “Anger” is 307 times. Color coding: The colors range from light to dark, indicating different values of quantity. The darker the color, the larger the quantity. For example, the darkest color is the number of times the classifier correctly predicted “Happy”, which was 776 times, indicating that this type of emotion is classified very well. Values on the diagonal: The numbers on the diagonal indicate the number of times the model correctly categorized the emotion, i.e., correctly classified it. For example, the classifier

correctly predicted “Happy” 776 times and “Fear” 157 times. Numbers outside the diagonal are misclassifications by the model. For example, the numbers in the first row (where the true emotion is “Anger”) show that the classifier misclassified “Anger” as “Fear” 14 times and “Fear” 6 times when the emotion was “Anger” 6 times it misclassified it as “Happy”. To summarize the classification performance: the classifier performs very well on the “Happy” emotion with 776 correct classifications and fewer misclassifications.

2) Accuracy of the Face Recognition Model

We use the 1:1 verification accuracy (%) evaluation to assess the face recognition performance of different methods on OOD data. The performance of various methods is shown in Table 3. As can be seen, although our method utilizes a transformer as a backbone, it performs significantly better than the vanilla transformer. This improvement can be attributed to our data collection strategy and training process.

Table 3. Comparison of face recognition accuracy

Method	CFP-FP	AgeDB
ArcFace	92.28%	92.12%
VggFace	91.01%	91.39%
Resnet34	90.11%	89.90%
VanillaViT	89.15%	88.27%
Proposed Method	99.12%	98.02%

3) Effect of different parameters on the system as a whole

We enhanced the robustness of the system by adjusting the weights of different parameters in the dataset. We designed an ablation study focusing on five aspects: “age,” “race,” “gender,” “lighting,” and “facial occlusion”. The results indicated that race has a minimal overall impact on the model's accuracy. Age is negatively correlated with accuracy; in simple terms, the younger the individual, the higher the recognition accuracy. This is because adolescents tend to express emotions more vividly and often in an exaggerated manner compared to adults. However, due to strict regulations in various countries regarding datasets involving minors, our study did not specifically focus on data related to underage individuals.

In terms of gender, we found that the facial features of adult males are less significant for the model than those of females. We speculate that adult males experience faster aging of facial muscles due to daily stress. According to the literature[17], females are generally considered to be more sensitive in emotional expression; therefore, incorporating diverse female features significantly benefits the accuracy of the model. Regarding lighting, we confirmed that the model's accuracy is highest when the white balance is set to daylight, whereas accuracy noticeably decreases under other white balance settings. Additionally, facial occlusion is negatively correlated with the model's accuracy; the more occluded the face, the lower the accuracy.

4) Analysis of the Proposed Model

In this section, to verify whether our proposed method can effectively identify faces and the overall efficiency of the process, we performed quantitative measures from two aspects of public challenging

Table 4. Comparison of parameter importance

Parameter	Notation	Mathematical Relationship	Description
Age	A	Accuracy $\propto -A$	For each year decrease, recognition accuracy increases by approximately 5%
Race	R	Accuracy $\approx$ C	The contribution of race to model accuracy is about 2% (minimal impact)
Gender	G	Importance(G_f) > Importance(G_m)	The effect of female features on model accuracy enhancement is approximately 10% to 15%
Lighting	L	Accuracy (L = daylight) > Accuracy($L \neq$ daylight)	When the white balance is set to daylight, model accuracy increases by about 20%
Facial occlusion	O	Accuracy $\propto -O$	For every 10% increase in occlusion, model accuracy decreases by approximately 15%

datasets: recognition speed efficiency and recognition accuracy. The results demonstrate that our model is more competitive compared to other models.

5) Directions for improvement

In our future endeavors, we aim to integrate dynamic lighting conditions, including flicker effects, into the dataset to enhance the model's adaptability to diverse white balance scenarios. Furthermore, contingent upon the availability of advanced computational resources, we intend to explore the utilization of 3D imaging technology for dataset collection. This innovative approach promises to significantly bolster the model's robustness by capturing comprehensive depth information, thereby enriching facial representation across a variety of contexts.

Ⅴ. Conclusion

In this paper, we proposed a novel approach to face recognition based on self-supervised pretraining. The proposed method leverages a transformer architecture and contrastive learning to capture the intrinsic relationship between various faces of the same person, offering robustness regardless of noise or disturbance.

To address the limitations of existing research, this approach implements a promising solution. Firstly, it utilizes large-scale unlabeled image data from the internet to educate the model in identifying face image patterns. Subsequently, high-quality labeled public face recognition data is used to refine the model parameters.

Performance evaluation results demonstrate that our method surpasses other methods, and the accuracy is shown to be high. Looking towards future research, we plan to further enhance the model's capabilities and inspect the impacts of various factors on our results, including the influence of different ethnicities and the system's robustness for a single individual under different makeup and hairstyle conditions, amongst other real-world scenarios.

References

[1] M. Wang and W. Deng, “Deep Face Recognition: A Survey,” *Neurocomputing*, Vol. 429, pp. 215-244, March 2021. https://doi.org/10.1016/j.neucom.2020.10.081

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach: CA, pp. 4685-4694, June 2019. https://doi.org/10.1109/CVPR.2019.00482

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A Dataset for Recognising Faces across Pose and Age,” in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi’an, China, pp. 67-74, May 2018. https://doi.org/10.1109/FG.2018.00020

[4] Z. Luo, J. Hu, W. Deng, and H. Shen, “Deep Unsupervised Domain Adaptation for Face Recognition,” in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi’an, China, pp. 453-457, May 2018. https://doi.org/10.1109/FG.2018.00073

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv:2010.11929v1, October 2020. https://doi.org/10.48550/arXiv.2010.11929

[6] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning Face Representation from Scratch,” arXiv:1411.7923, November 2014. https://doi.org/10.48550/arXiv.1411.7923

[7] H. Phan, C. X. Le, V. Le, Y. He, and A. T. Nguyen, “Fast and Interpretable Face Identification for Out-of-Distribution Data Using Vision Transformers,” in *Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa: HI, pp. 6289-6299, January 2024. https://doi.org/10.1109/WACV57701.2024.00618

[8] H. Phan and A. Nguyen, “DeepFace-EMD: Re-Ranking Using Patch-wise Earth Mover’s Distance Improves Out-of-Distribution Face Identification,” in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans: LA, pp. 20227-20237, June 2022. https://doi.org/10.1109/CVPR52688.2022.01962

[9] X. Xu, N. Sarafianos, and I. A. Kakadiaris, “On Improving the Generalization of Face Recognition in the Presence of Occlusions,” in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle: WA, pp. 3470-3480, June 2020. https://doi.org/10.1109/CVPRW50498.2020.00407

[10] C. Hu, Y. Li, Z. Feng, and X. Wu, “Toward Transferable Attack via Adversarial Diffusion in Face Recognition,”

IEEE Transactions on Information Forensics and Security, Vol. 19, pp. 5506-5519, 2024. https://doi.org/10.1109/TIFS.2024.3402167

[11] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to Profile Face Verification in the Wild," in *Proceedings of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid: NY, pp. 1-9, March 2016. https://doi.org/10.1109/WACV.2016.7477558

[12] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The First Manually Collected, In-the-Wild Age Database," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu: HI, pp. 1997-2005, July 2017. https://doi.org/10.1109/CVPRW.2017.250

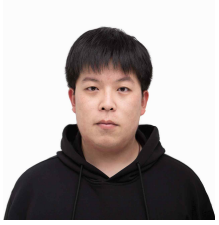
[13] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-Autoencoders for Face De-Occlusion in the Wild," *IEEE Transactions on Image Processing*, Vol. 27, No. 2, pp. 778-790, February 2018. https://doi.org/10.1109/TIP.2017.2771408

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans: LA, pp. 15979-15988, June 2022. https://doi.org/10.1109/CVPR52688.2022.01553

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle: WA, pp. 9726-9735, June 2020. https://doi.org/10.1109/CVPR42600.2020.00975

[16] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv:1807.03748v1, July 2018. https://doi.org/10.48550/arXiv.1807.03748

[17] R. M. Pearson, S. L. Lightman, and J. Evans, "Emotional Sensitivity for Motherhood: Late Pregnancy Is Associated with Enhanced Accuracy to Encode Emotional Faces," *Hormones and Behavior*, Vol. 56, No. 5, pp. 557-563, November 2009. https://doi.org/10.1016/j.yhbeh.2009.09.013



손천샤(Chen-Xiao Sun)

2021년 9월 : University of Mysore (Bachelor)
2023년 10월 : University of Mysore (Master)
2024년 3월 ~ 현 재: 동명대학교 컴퓨터미디어공학과 박사과정
※관심분야 : 영상처리, 얼굴인식, 딥러닝

신승수(Seung-Soo Shin)

2001년 2월 : 충북대학교 수학과 (이학박사)
2004년 8월 : 충북대학교 컴퓨터공학과 (공학박사)
2005년 3월 ~ 현 재 : 동명대학교 정보보호학과 교수
※관심분야 : 네트워크 보안, 딥러닝, IoT, 데이터분석