Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD36591

From CNNs to Transformers: The Architectural Shift in Face Recognition

Wenge Wu

School of Computer Science and Technology, Anhui University of Technology, Ma'anshan, China
mdown3539@gmail.com

Abstract. For almost the past decade, face recognition has been dominated by convolutional neural networks. Margin-based losses (CosFace, ArcFace) have achieved over 99.6% verification accuracy on classic benchmarks. For a long time, further performance gains were believed to rely solely on larger-scale datasets and more sophisticated loss functions. This perspective has since shifted. Having demonstrated strong performance across general computer vision tasks, Vision Transformers are now adopted for face recognition to directly remedy key CNN deficiencies. This paper explores the architectural evolution of margin-based CNNs, wherein hybrid CNN-Transformer designs and fully Transformer-based systems have emerged, drawing upon a range of relevant studies published between 2024 and 2026, such as LVFace (ICCV 2025), NPTFace and PaCo-FR (CVPR 2026), and the FaceLiVT series for mobile deployment, etc. The findings reveal that the the core issue lies not in Transformers outperforming CNNs on existing benchmarks, but in the shifts in computational models underlying face-representation mechanisms.

Keywords: CNN, Face recognition, Vision Transformer, Face embedding

1. Introduction

Deep learning has excelled in face recognition, with significant advancements through applications like DeepFace, CosFace, and ArcFace [1, 2]. By the time the verification accuracy on LFW reached 99.6%, it was reasonable to believe that the remaining problems were due to the architecture and not merely a lack of more data and better tuning [3].

They were buildings, and the defects were quite obvious. A convolutional filter is by definition a small area of pixels. The network gradually builds the whole picture from the back and forth connections. Most of the object categories work fine, a cat is still a cat even if you cannot see its ears or tail. Faces are not so good. What makes one's face different from another's is the placement of the eyes relative to the bridge of the nose and the width of the jaw set against the cheekbones. These are the attributes. They are not in any of the convolutional feature maps. A CNN will eventually learn them, but it will do so through a chain of local measurements, and the combined behaviour is not immediately known to the loss function [4].

Vision Transformers do the other way around. Self-attention considers the relationships among all pairs of positions at the same time and has all the surrounding context instantly available at each level. It is a better match for faces on paper. ViT training recipes built for ImageNet do not transfer

© 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: 10.54254/2755-2721/2026.LD36591

to face datasets, and the optimisation objective is different if you are interested in embedding geometry rather than class probabilities. Self-attention is quadratic, therefore, high-resolution face processing is too expensive, and the original Design cannot be realised on mobile phones. Fixed-position encoding is a fixed 2D grid, but when a face moves out of this plane due to rotation, the above assumption is no longer valid [5, 6].

Substantial progress has taken place between 2024 and 2026. LVFace believed that ViTs had simply been trained according to the protocols for a different type of problem [5]. Progressive Cluster Optimisation changes the training process according to the structure of face embedding spaces. NPTFace has recast the problem of pose alignment as that of positional encoding, so it does not lose the pixel information that traditional pipelines omit [6]. PaCo-FR has discovered that if the mask strategy is based on facial anatomy, self-supervised pre-training can reduce the gap with fully supervised models and require much less labelled data [7]. FaceLiVT is the combination of CNN and Transformer, and it works on mobile phones with low latency [8, 9].

This review summarizes how face recognition has evolved from CNNs to Transformers, distilling their core design principles and the unresolved issues.

2. The CNN era: achievements and architectural limits

2.1. Margin-based losses and the embedding paradigm

The bottom network maps a face image to a short vector in a hypersphere. Identity is given by cosine similarity of vectors. The Network is a feature extractor, usually ResNet. The loss function will form an outline in the embedding space and bring together vectors that are similar and push apart those that are dissimilar [3, 4].

Margin-based Softmax variants are what achieved this result. CosFace adds a constant offset to the logit of the correct class in cosine space to push the embedding closer to the center of that class. ArcFace moves the penalty to angular space and is therefore in line with the geodesic distance on the hypersphere. It was not Vanilla Softmax or a triplet-based method [4].

What came next was mostly margin engineering. ElasticFace takes the per-sample margins of a Gaussian distribution and increases them for difficult cases and decreases them for simple cases. X2-Softmax and InterFace connected the margin to the sample's angular distance from the class center, and thus the boundaries varied according to the data. There are now some errors, but the fundamental calculation remains the same. CNNs extract local features, pool them by depth, and then output a global embedding, the pipeline is the same, but the embedding space is better shaped [4].

2.2. Bottlenecks in convolutional neural networks

CNNs have been increasing their receptive field with depth. Neurons in the middle layers of deep ResNets only see a small part of the input image. Face recognition is about learning local features of faces, such as edges, skin patterns, etc., but it does not know how these features are organised in space.

CNNs are not rotation-equivariant, they only have rough translation invariance due to pooling. When the face is tilted, the same facial landmark (e.g., the left eye) will have different pixel coordinates. CNNs can do this by using data augmentation and explicit alignment to find the landmarks of faces, warp these faces to a standard position, and then extract features. Warping is not

DOI: 10.54254/2755-2721/2026.LD36591

lossless, some information is lost, and the reconstructed areas are usually blurred or distorted compared with the original. The embedding is also an attribute [6].

To learn about the relations among all the parts of the face, a very deep structure or an explicit attention mechanism over the CNN backbone is needed. Both choices are too expensive and do not solve the problem. Channel Attention (Squeeze-and-Excitation) and Spatial Attention Modules are good, but they operate on already compressed feature maps and cannot recover the spatial information lost in the previous layers [5, 9].

3. Vision transformers for face recognition

3.1. Reasons for transformers and their present-day rise

Divide the image into patches, map each patch to a token, and then apply self-attention to the sequence of tokens in ViT.

One is that identity is constituted by relations, and self-attention models pairwise relationships natively. The other is more subtle: since the spatial information enters through positional embeddings and not the fixed grid of convolutions, you can modify the encoding without altering the pixels. Thus, we do not need to deal with issues such as poses and occlusions in preprocessing [6]. Instead, they can be encoded.

The Problems were actual, not theoretical. ViTs are data-hungry, and the first model had to be trained on hundreds of millions of pictures before it could surpass CNNs. Self-attention is $O(N^2)$ in the number of patches, so it becomes too expensive at high-resolution faces. Training recipes optimised for ImageNet classification have unstable convergence and poor embeddings when used in metric learning. Therefore, for several years, Transformers in French were more promising than those in English [5].

3.2. LVFace: adapting ViT training to face recognition

LVFace and You et al. at ICCV 2025 have both addressed the training problem [5]. The authors found that the standard ViT training was unstable on the in-dataset, and its embeddings were worse than those of CNNs. Diagnosis: CNN-inspired training protocols are not suitable for ViTs because they have different optimisation landscapes. The solution, Progressive Cluster Optimization (PCO), unfolds in three stages.

In the first stage, Negative Class Subsampling (NCS) is employed. NCS does not compute the full softmax over all identities because it is too expensive and noisy for a large number of classes, so it randomly samples a small subset of negative classes in each iteration. Stabilise the first training signal and help the model form an initial cluster structure before learning about the distribution of all classes.

The second is an expectation penalty term. As training continues, the class centroids will move and the cluster boundaries will change. Feature expectation penalties move each centroid to the running mean of the embeddings of its members and avoid cluster collapse in the high-capacity representation space and degenerate solutions of transformers.

The third step is to omit NCS and use full-batch softmax. At this time, the embedding space has been organised, we will further refine the cluster boundaries and improve the accuracy in the full-batch stage. LVFace has set a new standard in this field. By March 2025, it has won the prestigious ICCV 2021 Masked Face Recognition Ongoing Challenge, showing that its representation method is highly generalisable.

DOI: 10.54254/2755-2721/2026.LD36591

3.3. NPTFace: pose alignment in encoding space

Ran and others proposed NPTFace at CVPR 2026, and to solve the problem of pose, they conceptually inverted the standard pipeline [6]. Face recognition would first find some points in the face, then compute an affine or thin-plate spline transformation, and finally show a normalised image to the network. Normalise the position encoding in NPTFace.

The first is that the positional encoding in a Transformer indicates the location of a patch in an image. If you know the head pose, you can warp the positional encoding grid to compensate, thus, the patch containing the left eye will always be given the same positional embedding, whether the face is in front of or in profile. Do not change image pixels.

NPTFace has two kinds of Pose-aligned Positional Encoding (PPE): 2D and 3D. A 2D Rotation Operation can be carried out on the coordinate grid. The 3D model is more complicated, it has been warped in 3D according to the estimated depth and then projected back through learnable view-projection matrices. This is for out-of-plane rotation, the face will appear foreshortened or partially hidden.

The advantage of not using image-space warping is quite obvious from the definition of warping. Change the front view of the face, extend the texture of the skin, and interpolate the pixels and synthetic data in the area that was previously covered. The network sees artifacts, not the true signal. NPTFace does not change the image and only modifies the positional encoding to provide the transformer with the original pixel data while keeping the same semantic location for all poses.

NPTFace has been tested on both low-quality surveillance pictures and high-resolution controlled sets, and new top results have been achieved in all cases. Pose-aware gradient weighting is employed in the training to improve identity discrimination at the class centre. This is a good way to do it, we know that class centres (the "ideal" identities) are generally front-facing, so embeddings from profile views should be pulled closer to these centres than to other profile views.

3.4. PaCo-FR: self-supervised pretraining with facial structure awareness

PaCo-FR was put forward by Xie and others at CVPR 2026 to address the problem of pre-training [7]. Another problem with using ViT for facial recognition is that it needs a large number of labelled images for training. These are costly to obtain and have privacy issues. PaCo-FR explores the possibility of pretraining facial representations in a label-free manner and subsequently fine-tuning the model with limited labeled data.

It is a Masked Image Model (MIM) with patch-pixel alignment. PaCo-FR is not a random-patch masking method, rather, it is structured according to the anatomy of the face. Masks are placed over the entire eye area, mouth, forehead, etc., rather than at various positions. Therefore, the model needs to reconstruct the general structure of the face reasonably and learn meaningful features rather than small details.

The second way is a patch-based codebook. The first basic MIM method directly reconstructs pixels and thus has limited learning of low-level statistics through memorisation, it is not very semantic. PaCo-FR maps the patches to discrete codes in a learned codebook, and the reconstruction target is the correct code assignment. Thus, the learning signal moves from pixel-level accuracy to semantic distinction. Instead of merely learning the average color of facial regions, the model focuses on identifying specific masked facial parts.

The outcome is positive. PaCo-FR only needs to be pre-trained on 2 million unlabeled images, which is about one-tenth of the amount used for general vision foundation models. It can still meet or exceed the results of fully supervised baselines in several facial analysis benchmarks.

DOI: 10.54254/2755-2721/2026.LD36591

3.5. Synthesis: a new Design Space

The three systems have formed a Design Space that is quite different from that of CNNs. A few are quite obvious.

First, training stability is not merely a concern but a fundamental prerequisite. LVFace has found that most of the performance gap for CNN-ViT in FR is a training artifact [5]. If you optimise well, the Transformer Capacity will no longer seem very low. This also has the effect of FR: if the new architecture performs worse than the old one on a metric-learning task, the first hypothesis should be "incorrect training protocol" and not "incorrect model".

Second, spatial reasoning is in feature space. NPTFace took pose alignment, which the field had long treated as preprocessing, and added it to the model [6]. The same idea can be extended to lighting normalisation, expression variation and occlusion, these are all modifications that can be expressed as changes in spatial coordinates rather than pixel values.

Third, facial anatomy is a free pre-training course. PaCo-FR did well and used prior knowledge of anatomy to annotate 2 million images [7]. It is neither as costly as the supervised method nor a general self-supervised method that treats faces the same way. Domain knowledge can be added as structural priors in the pre-training objective and does not need to be obtained from labelled data.

4. Hybrid architectures for edge deployment

Although the above systems have achieved good accuracy, they are too large for mobile apps. Given the requirements of the times for low-latency processing and privacy protection, many researchers have begun to study the Design of lightweight face recognition on devices [8-10].

The FaceLiVT series by Setyawan et al. is the most typical work in this area [8, 9]. FaceLiVTv1 (IEEE ICIP 2025) is a hybrid CNN-Transformer model with Multi-Head Linear Attention (MHLA). MHLA replaces the quadratic self-attention of the conventional Transformer with linear projections of the token sequence to reduce the complexity from $O(N^2)$ to $O(N)$. Reparametrisation token mixer with depthwise convolution and linear projection is used to make FaceLiVTv1 about 8.6 times faster than EdgeFace and around 21.2 times faster than pure ViT.

FaceLiVTv2 added a Light MHLA to reduce the number of layers in the attention mechanism and added affine rescale transformations to keep the diversity of representations among the heads [9]. RepMix is a single module that integrates local CNN features with global Transformer information, and global depthwise convolution is used instead of an embedding stem. The latency of the iPhone 15 Pro's FaceLiVTv2 has dropped by 22% from that of v1, it has been accelerated by 20-41% compared with other hybrid models, and it maintains good accuracy on the LFW, CFP-FP, AgeDB-30 and IJB benchmarks.

Gupta et al. took a different approach, demonstrating that MobileNetV2 with TensorFlow Lite after training quantization can also be used on Android for actual application [10]. The Pipeline is about 70% smaller in size and 60% faster at inference than a typical CNN. Although the upper bound of the accuracy is lower than that of the hybrid design, it is relatively simple and thus suitable for cases where the main problem is development cost.

The general lesson of the above deployment-oriented works is that CNNs and The dichotomy of transformers is false. The best models are relatively small and have added attention modules to focus on different areas of the image. The architectural frontier is not about choosing one side, but rather how to divide the work of the two computation models properly.

DOI: 10.54254/2755-2721/2026.LD36591

5. Conclusion

Face recognition is at the centre of an architectural change. Over the past decade, the CNN-margin loss paradigm has been the subject of research, and now, due to its structural defects, many extensions and, in some cases, transformer-based designs have begun to emerge. LVFace showed that ViTs can surpass CNNs in FR under good optimisation with good optimisation. NPTFace has found that pose alignment is relatively easy to perform in the encoding space. PaCo-FR has found that self-supervised pre-training with facial structure awareness can match supervised results, but it needs far fewer labelled samples. FaceLiVT series have also demonstrated that hybrid CNN-Transformer architectures can achieve the same result on mobile phones.

It is not just the replacement of one backbone with another. It can be seen that the field is starting to consider facial representation in a more dynamic and context-aware manner, and has moved away from the former method of fixed convolutional hierarchy aggregation in space. Now we need to carry out this transition and build architectures that are accurate, efficient, robust and fair at the same time, not be led astray by saturated benchmarks that no longer reflect the real progress of the past few years.

References

[1] El Fadel, N. (2025). Facial recognition algorithms: A systematic literature review. *Journal of Imaging, 11*(2), 58.
[2] Kshirsagar, S., & Chhajed, G. (2026). A systematic review of deep learning methods in face recognition. *AIJFR, 7*(3).
[3] Advances in face recognition: A comprehensive review of feature extraction and dataset evaluation. (2025). *Electronics, 15*(2), 338.
[4] EmergentMind. (2025). Margin loss function in deep learning. *EmergentMind*. https://www.emergentmind.com/topics/margin-loss-function
[5] You, J., et al. (2025). LVFace: Progressive cluster optimization for large vision models in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 11840-11849).
[6] Ran, Z., et al. (2026). NPTFace: Native pose-aligned transformer for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)* (pp. 1204-1213).
[7] Xie, Y., et al. (2026). PaCo-FR: Patch-pixel aligned end-to-end codebook learning for facial representation pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)* (pp. 1214-1224).
[8] Setyawan, N., et al. (2025). FaceLiVT: Face recognition using linear vision transformer with structural reparameterization for mobile device. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 1720-1725).
[9] Setyawan, N., et al. (2026). FaceLiVTv2: An improved hybrid architecture for efficient mobile face recognition [Preprint]. *arXiv*. https://arxiv.org/abs/2604.09127
[10] Gupta, M., Singh, C. R., & Agarwal, G. (2026). Design and performance evaluation of a lightweight face recognition model for android devices. *IRE Journals, 9*(9), 629-635.