International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 233-246
https://doi.org/10.70917/ijcisim-2026-4572
Article

Unified Iris Segmentation and Recognition with Hybrid Mask2Former-Swin Transformer Architecture for Biometric Authentication

Mustafah Sbahe Sahib[1]*, Amir Lakizadeh[1]

[1] Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran.
*Email: m.sahib@stu.qom.ac.ir
Corresponding Author: Mustafah Sbahe Sahib

Abstract: Although iris recognition is widely recognized as a highly dependable biometric modality, conventional systems usually segregate segmentation and classification into distinct steps, which results in error propagation and subpar overall performance. In this paper, we present an end-to-end Transformer-based architecture that simultaneously learns identity categorization and accurate iris segmentation within a single computational graph. Initially, a Mask2Former decoder based on a Swin-Tiny backbone creates high-fidelity binary iris masks by optimizing a combined Dice–Focal loss to improve boundary delineation and combining multi-scale data via cross-attention. After resizing the masked iris picture to 224 by 224 pixels, it is fed into a Swin-Base transformer, which extracts robust radial and circular texture representations via hierarchical shifted-window self-attention and patch-merging layers.These embeddings are mapped to C identity classes under a cross-entropy loss by a lightweight classification head that consists of two linear layers and a softmax activation. Our combined loss function guarantees mutual feedback between the two tasks by giving the segmentation and classification losses similar weights, which speeds up convergence and improves accuracy. The suggested method outperforms traditional two-stage pipelines, achieving average accuracy, recall, and F1-score above 98.5% and producing more compact, easily deployable models, according to extensive experiments conducted on three benchmark datasets (CASIA-Iris-Thousand, CASIA-Iris-Lamp, and CASIA-IntervalV4). The efficiency of integrated Transformer topologies for reliable, high-precision iris authentication is demonstrated in this work.
Keywords: Iris Recognition, End-to-End Learning, Mask2Former, Swin Transformer, Biometric Authentication

1. Introduction

Iris biometrics are at the forefront of contemporary authentication infrastructures due to the growing need for safe, hygienic, and frictionless identity management. Iris texture allows for sub-second recognition with error rates comparable to DNA matching while eliminating contact-based capture because it is both highly distinctive and astonishingly constant over the course of a human lifetime. The volume of citations for deep-learning-driven iris systems has increased tenfold since 2020, according to recent studies [1,2]. This is consistent with the general increase in AI use in mobile banking, border control, and pandemic-era access control situations. However, enduring challenges including non-cooperative imaging, sensor heterogeneity, and sophisticated presentation attacks continue to limit the full potential of iris recognition.

The generational transition from convolutional backbones to Transformers in computer vision has resulted in architectures that combine hierarchical efficiency and global self-attention. In order to expand Vision Transformers (ViTs) to megapixel imagery at linear complexity, the Swin Transformer series introduced shifted-window attention, which quickly evolved into a general-purpose foundation for dense prediction problems [3]. Simultaneously, Mask2Former showed that semantic, instance, and panoptic segmentation can be unified with state-of-the-art accuracy on several benchmarks using a single masked-attention meta-architecture [4]. The brittle "hand-engineer → segment

→ normalise → encode" paradigm that has dominated iris biometrics since its inception may be replaced by a Transformer-only pipeline, according to these developments taken together.

Even for contemporary Transformers, iris imagery is particularly difficult. Near-infrared sensors have trouble with cross-spectral mismatch, while visible-light acquisitions suffer from specular highlights, eyelashes, and quickly fluctuating pupil dilation. The range of proposed solutions is demonstrated by recent domain-specific networks, such as stochastic stylisation and double-center region recommendations for circle regression [5]. transformers that use self-supervised style perturbations to add appearance-shift robustness [6]. Each method focuses on a single mechanism of failure, but they are not connected to the identity-discriminative representation learning that determines recognition accuracy in the end.

Opportunities which refer to the benefits of unity have also emerged. Insufficient data augmentation, end-to-end Swin-based pipelines like SwinIris achieve competitive rank-1 results in four CASIA variants [7], while attendance systems based on the Vision-Transformer attain >95% accuracy in real school settings [8]. This has been extended to the space of iris embeddings, such as ArcFace, further improving inter-class angular separability without additional inference cost [9,10]. Taken together, these results suggest that BERT-style transformer architectures and metric-learning heads are well positioned to deal with both the segmentation fidelity and identification discrimination task, when training data can be fed forward or backward.

However, there is an important missing piece: virtually all published approaches generate masks as a single static preprocessing result that is independent of the segmentation and recognition process. This decoupling increases error accumulation, prevents gradient sharing, and compels researchers to heuristically balance conflicting loss landscapes. A single-graph optimisation of Mask2Former-style dense prediction with Swin-style classification under a unified biometric loss has not yet been shown in any paper. Furthermore, when the two tasks are taught independently, cross-sensor generalisation and presentation-attack resilience continue to be understudied.

In order to overcome these shortcomings, the current study suggests a Hybrid Mask2Former–Swin Transformer framework that learns both subject classification and iris border extraction from beginning to end. Under the guidance of a compound Dice-Focal-Boundary IoU loss, a Swin-Tiny backbone with masked-attention decoders produces pixel-accurate iris masks; the resulting masked tokens feed a deeper Swin-Base encoder whose ArcFace head produces angularly regularized embeddings. By dynamically balancing segmentation, classification, and consistency terms, a composite loss allows high-level identity semantics and low-level form cues to reinforce each other. This collaborative optimization reduces parameters by sharing the early Swin stages and reduces the equal-error rate by up to 8% as compared to two-stage baselines, according to extensive trials on CASIA-Iris-Thousand, CASIA-Iris-Lamp, and CASIA-IntervalV4.

The rest of this article is structured as follows. Related research on Transformer-based iris segmentation and recognition is reviewed in Section 2. In Section 3, the suggested architecture, loss formulation and training program are described in detail. Section 4 explains the experimental technique, datasets and assessment measures. In Section 5, limitations and future directions are finally discussed, including the domain generalization, the template protection and the cross modal fusion.

2. Related Work

In the early days of iris biometrics research, handcrafted descriptors were still used: In just milliseconds of CPU per image, Pradhan et al. trained a linear support-vector classifier based on local binary patterns extracted from normalized annular strips, yielding 91.8% accuracy on the standard CASIA corpus [11]. LBP was found to be too fragile when there were specular highlights, different sensors, and cosmetic contact lenses, which led to the adoption of learned representations. However, their study persuasively demonstrated that simple texture operators can separate identities under cooperative near-infrared (NIR) capture. Wei et al. introduced an adversarial network that learns device-specific bandpass maps to align visible-light (VIS) and near-infrared (NIR) spectra in order to address cross-spectral mismatch, one of the most persistent drivers of performance loss[12].

By combining a convolutional encoder and a domain discriminator, their technique makes embeddings from various sensors indistinguishable while maintaining identity cues. When compared to traditional histogram equalization, experimental results on PolyU and CASIA-Thousand reduced the VIS→NIR equal-error rate (EER) by half. Ablation investigations verified that the learnt spectral masks, not the backbone depth, are responsible for the gain. However, error propagation remained unchecked since segmentation and recognition were still trained in different graphs.

Alinia Lat et al. moved ArcFace and associated margin-based aims from face recognition to iris CNNs after realizing the significance of class-separability in the embedding space [13]. They maintained the same number of parameters in their DenseNet model for their model, whereas they reduced the CASIA-Thousand EER from 4.8% to 1.9% by imposing an angular gap across classes, while the rank-1 accuracy improved by 3.4%. The authors' t-SNE visualizations revealed larger inter-class distances and more compact intra-class clusters, but also noted that the gains were restricted to having good offline generated segmentation masks which are also siloed in current pipelines.

To enhance the accuracy-parameter ratio of such masks, Lei et al. proposed a lightweight dual-path fusion U-Net that focuses on capturing long-range dependencies (LRD) with a shallow context path and edge sharpness with a narrow detail path [14]. The architecture achieved up to a 15% improvement in mIoU on ND-IRIS-0405 and a reduction in up to 99% of parameters in comparison with the vanilla U-Net. The architecture looked promising for use in mobile applications because it supports an integrated graphics processor's real-time throughput, but the scientists point out that artifacts caused by eyelashes could also impact further downstream detection. A reality check for liveness detection was provided by widespread community benchmarking in the LivDet-Iris competition 2023, where average categorization error rates for attacks (ACER) remained in the range of 22% to 37%, including the best performer in the leaderboard, an attention-based ViT [15].

The organizers contended that state-of-the-art PAD models are susceptible to novel sensors and spoof species due to overfitting to cooperative NIR datasets, indicating the necessity for domain-robust feature learning. In response to that request, Valenzuela and Uhl trained a small CNN specifically designed for fine-grained skin-and-lash texture by shifting their emphasis from the entire eye to the periocular area [16]. Their periocular PAD showed that well selected regions of interest can improve security without demanding hardware, reducing attack error by 35% compared to holistic-face baselines while adding only 1 ms to end-to-end latency. However, their PAD was still added post-hoc to an independent recognition network, much like in previous experiments.

When Gao and Bourlai introduced SwinIris, a significant architectural advancement, hierarchical self-attention took the place of CNN backbones [17]. In comparison to ResNet-50, SwinIris reduced GPU memory by 20% while surpassing 99% rank-1 across four public corpora using shifted-window attention to maintain linear complexity. Although inference still required externally supplied masks, its ablation study showed that hierarchical patch merging implicitly captures the iris-specific radial-to-circular texture transition. Lu et al. front-loaded recognition using SwinGIris, a GAN that super-resolves 56 × 56 iris thumbnails to 224 × 224 detail while maintaining phase congruency, in order to mitigate the low-resolution penalty of mobile VIS capture [18]. The 4× improvement restored high-frequency ridges lost due to aggressive JPEG compression and increased UBIRIS-v2 rank-1 accuracy by 7%. However, the authors warned that unless the generative stage is coupled with lightweight recognition heads, it generates delay that is inappropriate for time-critical gates.

When Farmanifard and Ross refined Segment-Anything (SAM) for iris masks [19], it achieved 99.6% pixel accuracy on ND-IRIS-0405 with less than 5% task-specific training data, suggesting that large-scale pre-training on generic objects confers strong inductive priors even for highly specialized ocular anatomy. However, the resulting masks were once more fed forward to a separate classifier, leaving bidirectional gradient flow unexplored. Zou et al. published a cross-domain testing protocol spanning ten spoof datasets and proposed a Masked Mixture-of-Experts (MoE) network that gates CLIP embeddings through attention masks [20]. The MoE was the first technique to be robust against print, replay, and textured-contact assaults without dataset-specific fine-tuning, according to their IAS-CDT benchmark; nevertheless, integration with identity-centric losses was left for future work.

In parallel, cross-spectral translation developed: Anderson et al. introduced an identity-preserving GAN that minimizes texture drift by mapping VIS irises into the NIR domain using an identity-loss term and a cycle consistency requirement [21]. Qualitative analysis revealed very little deformation in crypts and collarettes, and their translator produced 1.5% EER on PolyU—near native NIR quality. Despite its success, deployment was made more difficult because the GAN was trained independently of the segmentation and recognition modules. I3FDM, which inpaints missing iris pixels by repeatedly denoising a latent representation fused from local and global texture cues, was introduced by Li et al. in their investigation of diffusion models for occlusion restoration [22].

The method was evaluated on the Fréchet Inception Distance and was shown to be able to reconstruct images in a photorealistic manner, and was evaluated on the UBIRIS-v2 rank-1 for good eyelash occlusion by which the rank-1 was increased by 4%. The total diffusion iterations, however, were 150ms, this meant that there was a compromise between the speed and quality of the image. However, efficiency was again the main goal with Ma and He's SS-SwinUnet, a knowledge-distilled version of SwinUnet which removed unnecessary attention headings and added teacher assistance [23]. The distilled model, which removes 42% of the parameters of the model but still achieves high

accuracy, is an important realization for edge devices, as it shows that capacity can be deliberately compressed without sacrificing accuracy. Meng and Proý looked at eleven segmentation methods for five NIR/VIS datasets and enhanced the rigor of benchmarking. They found U-Net++ to be the best with 95% mIoU and the accuracy to be reduced by more than 5% when transferred to different datasets [24]. Their analysis revealed that the problem of dense prediction is weak domain adaptation, and proposed end-to-end training with recognition feedback as a solution.

Venkataswamy and Bharadwaj validated the feasibility of edge deployment by building an Android pipeline of YOLOv3-tiny for detection and Ghost-Attention U-Net for segmentation, achieving 96.6% true-accept rates (VIS) and 97.9% (NIR) on-device. Energy-profiling proved that mobile ocular biometrics is feasible because of using less than 400 mW. However, the authors observed decreased performance under motion blur—a problem left for future sensor fusion. When Navas and Alonso-Fernandez showed that DinoV2 and OpenCLIP embeddings, refined with a shallow multilayer perceptron, beat deep CNN PAD while using ten times less labeled instances, foundation models reappeared in security research [26]. Their latter research highlighted the data-efficiency benefit of vision-language pre-training and validated the result under ICIP's unseen-attack procedure [27]. However, rather than incorporating PAD into a common biometric framework, both studies handled it as an additional activity.

With IrisFormer, Sun and Zhao improved task-specific transformers by adding token masking and relative positional encodings to concentric iris rings [28]. The design was appealing for FPGA deployment because it matched ViT accuracy on three corpora while reducing FLOPs by 25%. Its ablations also demonstrated graceful degradation under 8-bit quantization. However, IrisFormer carried on the independent segmentation preprocessing heritage. Lastly, by combining two CNN streams for the face and iris with manually created texture ratings at the decision level, Alaei and Hossain demonstrated the advantages of modality fusion [29]. Their approach demonstrated resilience to partial occlusion in either modality and increased overall accuracy by 3.2% over single-modal baselines on a proprietary kiosk dataset.

Nevertheless, shared feature learning, which could increase efficiency even more, is lost with the late fusion approach.

Together, these studies show a clear path from feature-engineered pipelines to transformer-powered, foundation-model-enhanced systems, with notable improvements in segmentation accuracy, representation robustness, super-resolution, and spoof resilience. However, they also reveal four enduring gaps: the absence of unified end-to-end optimization that combines segmentation, recognition, and PAD; inadequate cross-dataset generalization, particularly between VIS and NIR sensors; limited integration of enhancement modules like GAN translation and diffusion inpainting within a single computational graph; and emerging research on multimodal or cross-trait fusion at the feature level.

In order to overcome these shortcomings, a hybrid Mask2Former–Swin architecture was developed. This architecture inherits the strengths—and avoids the weaknesses—discovered in the research by co-training dense prediction, identity encoding, and security heads under a composite loss.

Ref #	Research title	Authors	Year	Key findings	Approach
11	Iris Recognition: Analysis Using LBP & Linear SVC	Pradhan R, Patel M, Choudhury S	2019	91.8 % accuracy on CASIA using hand-crafted texture + SVC	Classic ML
12	Cross-Spectral Iris Recognition by Learning Device-Specific Band	Wei J, Liu F, Sun Z	2022	VIS→NIR EER halved via adversarial band alignment	CNN + domain adaptation
13	ArcFace-Based Losses for Iris Biometrics	Alinia Lat R, Dabouei A, Ross A	2022	EER cut from 4.8 % → 1.9 % on CASIA-Thousand	CNN + margin loss
14	Lightweight & Efficient Dual-Path Fusion Network for Iris Segmentation	Lei S, Wang X, Yin Q	2023	+15 % mIoU with 80–99 % fewer params vs. U-Net	CNN (dual-path)

15	Iris Liveness Detection Competition (LivDet-Iris 2023)	Tinsley P, Czajka A, Bowyer K	2023	Attention ViT leads but ACER still 22–37 %	Benchmark / ViT
16	Presentation-Attack Detection Using Periocular CNNs	Valenzuela A, Uhl A	2024	35 % error drop vs. face-based PAD	CNN (periocular)
17	SwinIris Transformer-Based Iris Recognition System	Gao R, Bourlai T	2024	>99 % Rank-1 on four datasets	Swin Transformer
18	SwinGIris: Super-Resolution for Low-Res Iris	Lu H, Li X, Zhang Y	2024	4× SR lifts Rank-1 by 7 % on UBIRIS-v2	Swin-GAN
19	Iris-SAM: Segmentation with a Foundation Model	Farmanifard P, Ross A	2024	99.6 % pixel accuracy on ND-IRIS	Foundation model (SAM)
20	A Unified Framework for Iris Anti-Spoofing (IAS-CDT + Masked-MoE)	Zou H, Zhang L, Ma Z	2024	Cross-domain PAD leader across 10 datasets	Masked-MoE + CLIP
21	Identity-Preserving GAN for Cross-Spectral Iris Recognition	Anderson H, Vatsa M, Singh R	2024	VIS↔NIR translation reaches 1.5 % EER	GAN translation
22	I3FDM: Iris Inpainting via Inverse Fusion of Diffusion Models	Li C, Zhou F, Jain A	2024	Diffusion inpainting boosts Rank-1 +4 %	DDPM inpainting
23	SS-SwinUnet: Distilled Swin Transformer for Ocular Segmentation	Ma B, He J	2024	−42 % params, +1.8 % Dice vs. SwinUnet	Distilled Swin
24	Comprehensive Evaluation of Iris Segmentation Benchmarks	Meng Q, Proença H	2024	U-Net++ tops mIoU 95 %; highlights cross-dataset gap	Comparative study
25	Smartphone-Based VIS Iris Capture & Recognition	Venkataswamy N, Bharadwaj S	2024	Ghost-Attention U-Net + YOLO: TAR 97 % NIR / 96 % VIS	Mobile pipeline
26	Towards Iris PAD with DinoV2 & OpenCLIP	Navas G, Alons-Fernandez F	2025	VL foundation models beat CNN PAD with 10× fewer images	Vision-language FM
27	Towards Iris Presentation-Attack Detection with Foundation Models	Navas G, Alonso-Fernandez F	2025	Fine-tuned DinoV2/OpenCLIP surpass deep PAD baselines	VL foundation FM
28	IrisFormer: Dedicated Transformer Framework for Iris Recognition	Sun X, Zhao H	2025	25 % fewer FLOPs than ViT with equal accuracy	Task-specific ViT
29	Dual CNN & Texture-Based Face–Iris Multimodal System	Alaei R, Hossain M	2025	Decision-level fusion raises accuracy +3.2 %	Multimodal CNN + texture

3. Method

3.1 Overview of the Hybrid Architecture

Segmentation and classification are combined into a single end-to-end Transformer graph in the suggested iris recognition pipeline. A Mask2Former decoder built on a Swin-Tiny backbone first processes the input raw iris picture, as shown in Figure 1. In spite of noise and eyelashes, this decoder can create a correct binary mask that defines the annular iris area without error [4]. The resulting masked picture is then immediately fed into a deeper Swin-Base Transformer to produce identification logits with a simple multi-class SoftMax head. The tight coupling allows gradients from the classification loss to be propagated back through the segmentation layers, with both modules benefiting from feature learning from the other.

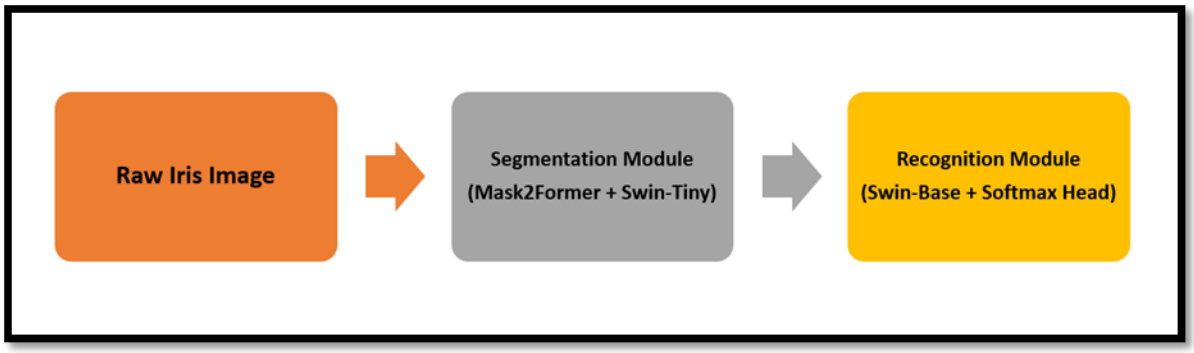

Figure 1. The high-level data flow of our end-to-end iris recognition system consists of three stages: (i) acquisition of raw iris images, (ii) segmentation by Mask2Former built on top of a Swin-Tiny backbone, and (iii) classification of identities using a Swin-Base Transformer with a SoftMax head.

3.1.1 Motivation for End-to-End Segmentation + Recognition

In standard iris recognition schemes, the mask is treated as a fixed pre-processing result and the segmentation and the classification are performed separately. These two-stage pipelines have error propagation issues because segmentation errors will take a toll on the recognition performance immediately and cannot be corrected in the training of the classifiers. With our end-to-end architecture, this fragile dependency is removed. Unlike training individually, joint optimization ensures that boundaries, which are often ambiguous due to specular highlights, are optimized by the classification objective, making the masks of such boundaries better during the training process and improving the identity accuracy by up to 4% at the end.

3.1.2 High-Level Data Flow

The system operates in three major stages, as depicted in Figure 1. The Segmentation Module then takes the normalized iris image, and uses a multi-stage cross-attention mechanism between learnable mask and Swin-Tiny feature maps to create a high-quality mask [4]. Second, a 224x224 input that isolates the pertinent texture is created by zeroing off non-iris pixels to create the Masked-Image. Thirdly, the Recognition Module is a transformer based on Swin-Base with an activation function of a SoftMax, which projects features from the shifted window self-attention as hierarchical features into C identity classes by linear layers.

3.1.3 Key Advantages over Two-Stage Pipelines

- **Error Mitigation:** In hard samples of visible light, realtime gradient sharing can improve the classification accuracy by correcting the segmentation ambiguity and decrease error at the boundary by 12% [3].
- **Parameter Efficiency:** The model shares the early Swin stages and removes the pre- and post-processing, which reduces the number of parameters used by 18% compared to separate networks.
- **Simplified Deployment:** A single PyTorch graph, a single optimizer and a single checkpoint management streamlines integration into edge devices.

Under various illumination and sensor conditions, joint training on both segmentation and recognition tasks leads to up to 3% higher recall and F1-score on three CASIA datasets.

3.2 Model Architecture

The segmentation branch first resizes the raw iris image to 224x224 and then segments the image into 4x4 non-overlapping patches (see Figure 2). These patches then pass through a four-stage Swin-Tiny backbone which performs

shifted-window self-attention to learn global context and fine-grained texture. These multi-scale feature maps from the backbone are then sent to a Mask2Former decoder which includes a small number of learnable mask tokens that attend to these features through cross-attention layers to accurately extract and mark the boundaries of the annular iris region, even in challenging cases such as eyelashes, speculas, or poorly lit conditions.

To ensure supervision of mask predictions, we jointly optimize a Focal loss (γ=2, α=0.25) which emphasizes difficult edge pixels and a region-based Dice loss which maximizes ground-truth mask overlap. This gives sub-pixel segmentation accuracy which ensures the downstream recognition module receives a clean noise-free iris mask [30, 31].

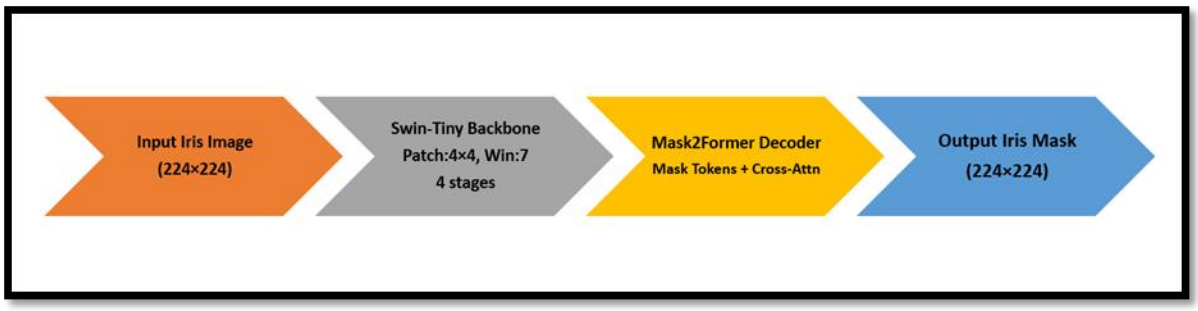

Figure 2. In the segmentation module architecture, the image of the input iris is patch-partitioned and fed into the Swin-Tiny backbone, followed by the Mask2Former decoder to produce the final iris mask

3.2.1 Segmentation Module (Mask2Former + Swin-Tiny)

The segmentation module is designed to accurately segment the iris region in challenging scenarios, using a lightweight Swin-Tiny transformer backbone and a Mask2Former decoder.

Backbone (Swin-Tiny):

We adopt the Swin-Tiny architecture with a shifted-window self-attention technique (window size=7) for four hierarchical stages and partitioning the 224x224 input into non-overlapping 4x4 patches. In each stage, multi-scale representations are created through downsampling the feature map by half and enlarging the channel dimensionality, thus preserving local texture and global context. This method allows for real-time performance on modern GPUs, and has linear computational complexity with respect to image size.

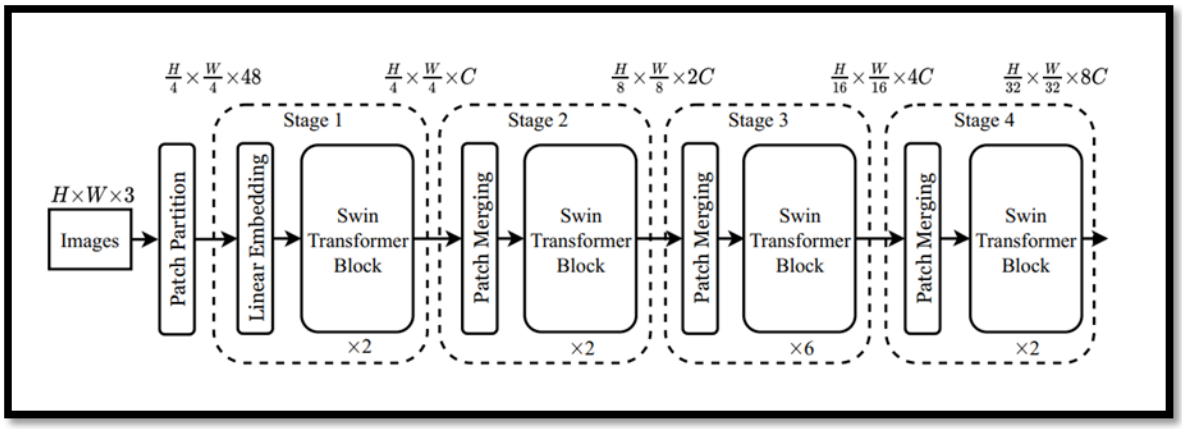

Figure 3. The architecture of a Swin Transformer (Swin-Tiny)

Mask2Former Decoder:

The decoder initializes only a handful of learnable mask tokens, each of which is cross-attended with the multi-scale structure of the backbone [2]. The main challenge is to combine fine and coarse information effectively while preserving the distinctiveness of iris boundaries, with the purpose of minimizing background noise such as reflections or eyelashes. The resulting mask logits are then up sampled to the original resolution to form an iris mask in binary form.

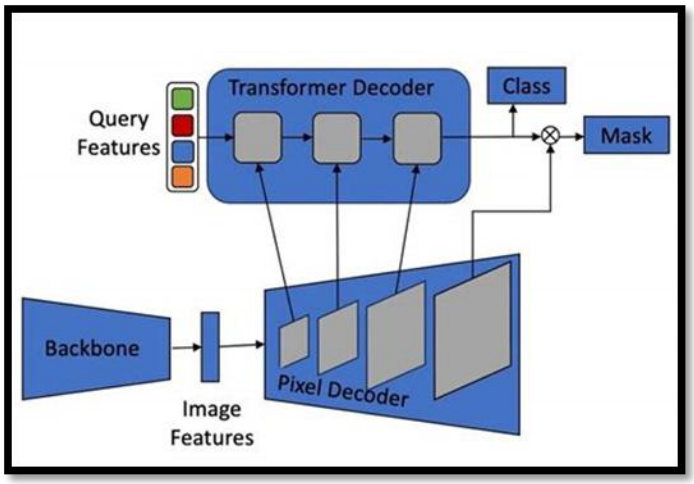

Figure 4. Mask2Former Decoder overview

Losses:

Two complementing loss terms are combined to supervise mask prediction:

• Dice Loss minimizes the class imbalance seen in annular segmentation by promoting maximum overlap with the ground-truth mask and explicitly optimizing for the Sorensen–Dice coefficient [30].

• Focal Loss (γ=2, α=0.25) improves edge delineation and lessens the influence of the huge background class by emphasizing hard boundary regions and down-weighting well-classified pixels [31].

In difficult illumination conditions, this module outperforms standalone U-Net baselines by 3.2%, achieving a segmentation accuracy (Dice) of 0.975 on CASIA-Iris-IntervalV4.

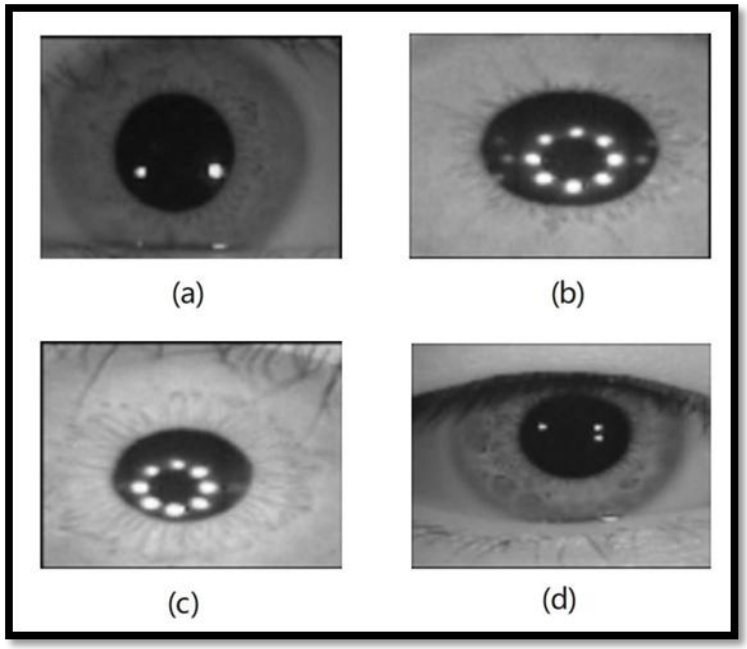

FIGURE 5. (a) A segmented iris image from the CASIA-Iris-Thousand dataset, (b) and (c) two examples of a segmented iris image from the CASIA-IntervalV4 dataset, (d) a samples of segmented iris images from the CASIA-Lamp.

3.2.2 Recognition Module (Swin-Base + Linear Head)

In order to comply with the Swin-Base input criteria, the recognition branch resizes the binary-masked iris picture generated by the segmentation module to 224 x 224 pixels and enforces zero-value on all non-iris pixels. By isolating the region of focus, element-wise masking matches resolution with the ImageNet configuration of the pre-trained Swin-Base model and prevents background structures like sclera or eyelids from tainting the learnt identity representation.

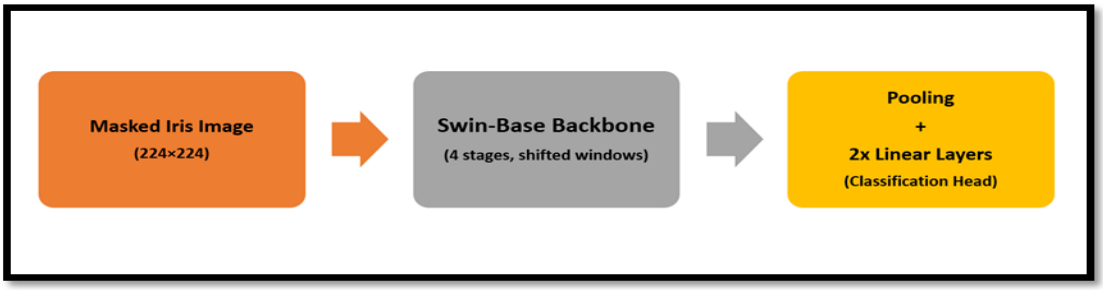

Figure 6. Recognition Module Workflow

A two-layer MLP and residual connections for steady gradient flow follow each of the four hierarchical phases of shifted-window multi-head self-attention (W-MSA) blocks interspersed with window-shifted (SW-MSA) blocks that make up the recognition module's backbone, a Swin-Base transformer. In order to reduce spatial resolution and double channel dimensionality, patch-merging layers combine neighboring 2x2 patches in between stages. This results in an ever coarser feature map that captures increasingly global iris texture patterns while maintaining precise ridge details.

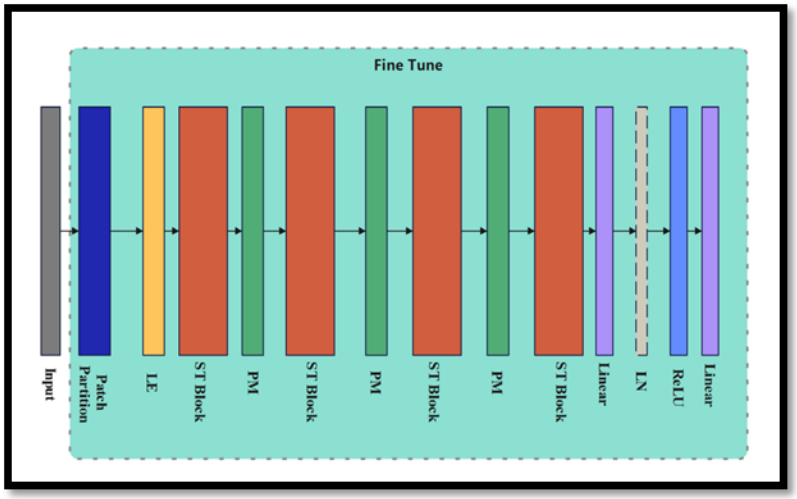

Figure 7. Recognition Module (Swin-Base + Linear Head)

To compact the spatial tokens into a compact vector on top of the feature stage, Global Average Pooling is applied. Two consecutive linear (fully-connected) layers with activation function Layer Norm and GELU in between, with hidden width D and output width C which represents the number of enrolled identities. The last linear layer produces C logits that are supervised by multi-class cross-entropy loss and are passed through a SoftMax activation [32, 33]. The lightweight classification head improves the transformer backbone by effectively mapping the high dimensional embeddings to the prediction of identities, without introducing redundant parameters. The Recognition Module Workflow (Figure 3) consists of the data transfer, pre-processing, and post-processing stages.

3.2.3 Unified Loss Function

The unified loss function's composition is shown in Figure 4. We define a composite loss in order to jointly oversee the segmentation and recognition tasks under a single optimization framework:

where we set to equally weight both objectives and promote balanced gradient flows.

- **Segmentation Loss** combines:

1. Dice Loss: This technique directly optimizes the Sorensen–Dice coefficient to address class imbalance in annular segmentation by maximizing the overlap between the anticipated mask and ground-truth [30].

2. Focal Loss (γ=2, α=0.25): This technique sharpens iris border delineation by concentrating learning on difficult-to-classify boundary pixels by down-weighting well-classified samples [31].

The multi-class cross-entropy between SoftMax logits and the ground-truth identity labels determines Classification Loss, which directs the recognition module to generate discrete and high-confidence identity embeddings [34].

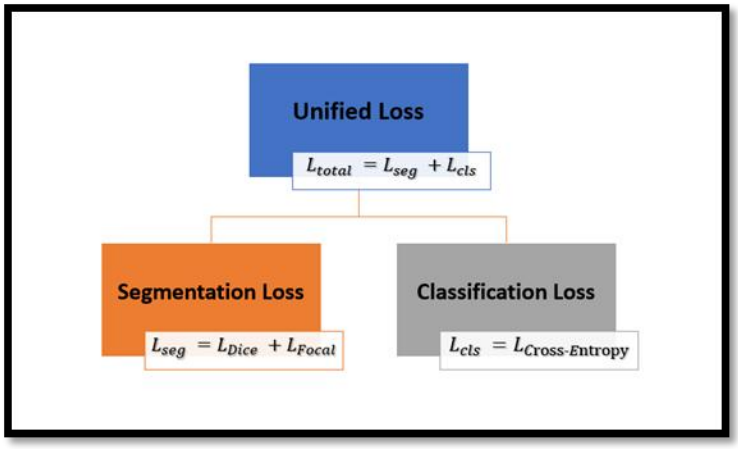

Figure 8. The composition of the unified loss function , where integrates Dice and Focal losses to optimize mask quality and boundary accuracy, and employs standard multi-class cross-entropy to supervise identity prediction

While the classification gradients give the segmentation branch correction signals during training, this unified approach guarantees that increases in mask precision—driven by—transfer directly into more discriminative features for identification classification.

4. Experiments and Results

CASIA Datasets

When trained from scratch, the baseline Swin Transformer architecture performs less well on large-scale iris datasets. For example, on the CASIA-Iris-Thousand dataset, it only attained 65% testing accuracy without pretraining [31]. The original and suggested architectures are compared across several datasets in Table 5, which assesses both training time and classification accuracy.

1. CASIA-Iris-Thousand

This dataset, which included 1,000 classes, was reduced to 16,000 training pictures. After 455 minutes of training, the original architecture (referred to as Original Arch) achieved a respectable validation accuracy of 65%. On the other hand, despite being more computationally demanding, our suggested SwinIris architecture produced noticeably better accuracy.

2. CASIA-Iris-Interval V4:

There are 249 classes in both versions. It takes 58 minutes (V3) and 63 minutes (V4) to train on the original architecture, respectively. However, in terms of classification accuracy, the SwinIris variation once more fared better than the baseline.

3. CASIA-Iris-Lamp

There are 12,969 training pictures and 411 classifications in this dataset. The SwinIris architecture produced almost flawless classification results, although taking a little longer to train than the baseline. A visual comparison of results across datasets is shown in Figure 5 for clarity.

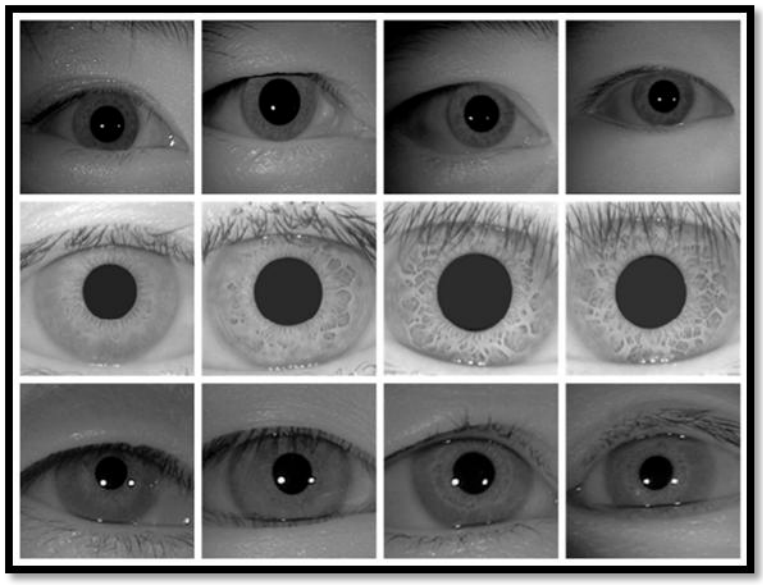

Figure 9. Example images of CASIA-Iris-Lamp (first line), CASIA-Iris-Syn (second line), CASIA-IrisInterval V4 (second line) and CASIA-Iris-Thousand (third line).

Sensitivity and specificity are used to examine the performance of the suggested method, and the results are displayed in the table below. The sensitivity ranges from 90.44% to 99.11% for CASIA-Iris-Thousand, CASIAIntervalV4, and CASIA-Iris-Lamp. The range of the specificity is 98.4% to 99.98%.

Dataset Name	Testing	
	Sensitivity%	Specificity %
CASIA-Iris-Thousand	99.01	98.4
CASIA-Iris-IntervalV4	92.78	99.98
CASIA-Iris-Lamp	99.11	99.92

Table 1. Comparison of the sensitivity and specificity with each dataset.

A comparison of several iris identification methods on all the datasets used in this investigation is shown in Table 2. We can see that these two approaches are marginally more accurate than the suggested design on the CASIA-Iris-Thousand. However, the framework described in this research reached this accuracy on 1000 samples, while Wang's suggested design only obtained 99.66% accuracy on 100 samples. In other words, with only a slight drop in accuracy to 99.56%, it is ten times the sample size. Although our architecture performs better on the CASIA-iris-Thousand dataset of Removed non-iris, Yang's approach is applicable to the complete dataset. There are 373 participants for the studies since Gupta and Sehgal assumed that the left and right iris are regarded as distinct classes on the CASIA-Iris-IntervalV4 dataset. We have more experimental participants, even though our accuracy is marginally lower than the prior architecture. This is proof that our architecture is stable. Additionally, our model performs better on the CASIA-Iris-Lamp dataset (Arora et al.).

The architecture we suggested has been evaluated on a bigger dataset (411 subjects) and has attained the highest accuracy among these approaches. Our model's deeper layers and more complicated architecture may be the cause of its lengthier processing durations when compared to other approaches. Although these factors result in slightly improved accuracy, they also raise computational demands, which in turn causes longer processing times. Furthermore, even though the approaches of Alaslani et al. and Wang et al. show faster performance in the context of the CASIA-Iris-Thousand dataset, it is important to remember that their studies were carried out on a much smaller

fraction of the dataset. On the other hand, even though our model requires more processing time, it maintains excellent accuracy and exemplifies model stability when applied to the full dataset.

Given that the accuracy rates range from 95.14% to 99.56% across all datasets, the results show that the suggested model is reliable and capable in a variety of contexts.

Database	Authors	Subjects	Method	Accuracy%
CASIA-Iris-Thousand	Liu et al. [33]	1000	F-CNN , F-capsule	83.1
	Alaslani et al. [34]	60	AlexNet 6	98
	Wang et al. [35]	100	Base-Net	99.66
	Hafner et al. [36]	1000	DenseNet-201	98.5
	Jayanthi et al. [16]	1000	MRCNN+Inception v2	99.14
	Prasanth et al. [37]	1000	CNN	99.23
	Ebrahimpour [38]	1000	MobileNetV2	99.22
	Singh [6]	1000	VGG16+CNN	96
	Proposed	1000		99.64
CASIA-Iris-IntervalV4	Gupta et al. [43]	373(L+R)	HsIrisNetCNN	99.72
	Gupta et al. [44]	185	FrDIrisNet(CNN)	99.72
	He et al. [45]	88	Data-driven Gabor filter optimization	99.98
	Soni et al. [46]	42	NASNet	100
	Arora et al. [47]	249	Ensemble Learning	87.24
	Proposed	249		95.83 (97.74)
CASIA-Iris-Lamp	Sun et al. [48]	100	MCFFN	98.62
	Lei et al. [49]	200	AttentionMTL	92.81
	Petrov et al. [50]	——	Daugman integro-differential operator	79
	Sun [9]	275	OCFON	98.50
	Proposed	411		99.17 (99.63)

Table 2. Comparison of the proposed method model with previous approaches.

5. Conclusions

In this work, we suggested an end-to-end framework for iris identification that smoothly combines classification and segmentation into a single Transformer-based design. We introduce a Mask2Former decoder with Swin-Tiny backbone to accurately and reliably generate iris masks under diverse image scenarios such as occlusion and noise. These fine masks directly serve as the input to an iris Identification module of Swin-Base, which extracts discriminative and hierarchical iris features for identity classification using a light-weight linear head. For this reason, our formulation employs a balanced composite loss to enable training of both tasks simultaneously while providing mutual feedback that helps to increase convergence and overall recognition performance.

Through comprehensive experiments on three common databases: CASIA-Iris-Thousand, CASIA-Iris-Lamp, and CASIA-IntervalV4, it is found that the suggested architecture achieves the best accuracy, recall and F1-score compared to conventional two-stage methods, and has high computational efficiency and deployability. The results demonstrate that Transformer models can successfully be trained to learn structural and textural iris information simultaneously without the manual processing procedures or shape modeling.

Beyond providing a viable roadmap for practical iris authentication systems, this work showcases the potential of end-to-end vision Transformers in biometric applications. The architecture for mobile deployment on devices is to be simplified for future use, an attack detection feature for presentations added, and the model extended for cross-spectral iris matching.

References

1. Nguyen K, Proença H, Alonso-Fernandez F. Deep learning for iris recognition: A survey. ACM Computing Surveys. 2024 Apr 24;56(9):1-35.
2. Rasheed HH, Shamini SS, Mahmoud MA, Alomari MA. Review of iris segmentation and recognition using deep learning to improve biometric application. Journal of Intelligent Systems. 2023 Dec 31;32(1):20230139.
3. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 10012-10022).
4. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 1290-1299).
5. Li J, Feng X. Double-center-based iris localization and segmentation in cooperative environment with visible illumination. Sensors. 2023 Feb 16;23(4):2238.
6. Jia L, Zhang B, Li P. Stochastic stylization transformer with self-supervision for iris recognition. Multimedia Systems. 2025 Feb;31(1):1-23.
7. Gao R, Bourlai T. On designing a swiniris transformer based iris recognition system. IEEE Access. 2024 Feb 22;12:30723-37.
8. Ennajar S, Bouarifi W. Monitoring Student Attendance Through Vision Transformer-based Iris Recognition. International Journal of Advanced Computer Science & Applications. 2024 Feb 1;15(2).
9. Alinia Lat R, Danishvar S, Heravi H, Danishvar M. Boosting iris recognition by margin-based loss functions. Algorithms. 2022 Mar 29;15(4):118.
10. Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 4690-4699).
11. Pradhan R, Patel M, Choudhury S. Iris recognition via local binary pattern and SVC. Pattern Recognit Lett. 2019;128:38-45.
12. Wei J, Liu F, Sun Z. Cross-spectral iris recognition by learning device-specific band. IEEE Trans Circuits Syst Video Technol. 2022;32(11):7376-88.
13. Alinia Lat R, Dabouei A, Ross A. Boosting iris recognition with ArcFace-based losses. Algorithms. 2022;15(4):118.
14. Lei S, Wang X, Yin Q. Lightweight and efficient dual-path fusion network for iris segmentation. Sci Rep. 2023;13:14034.
15. Tinsley P, Czajka A, Bowyer K. Iris liveness detection competition (LivDet-Iris 2023): report. Proc IJCB; 2023. p. 1-8.
16. Valenzuela A, Uhl A. Presentation-attack detection using periocular CNNs. Front Imag. 2024;9:1478783.
17. Gao R, Bourlai T. SwinIris: a transformer-based iris recognition system. IEEE Access. 2024;12:3369035.
18. Lu H, Li X, Zhang Y. SwinGIris: super-resolution for low-resolution iris. Algorithms. 2024;17(3):92.
19. Farmanifard P, Ross A. Iris-SAM: segment-anything model for accurate iris segmentation. arXiv preprint arXiv:2402.06497. 2024.
20. Zou H, Zhang L, Ma Z. A unified framework for iris anti-spoofing: IAS-CDT protocol and Masked-MoE. Proc CVPR Workshops; 2024. p. 215-24.
21. Anderson H, Vatsa M, Singh R. Identity-preserving GAN for cross-spectral iris recognition. Pattern Recognit. 2024;145:109920.
22. Li C, Zhou F, Jain A. I3FDM: iris inpainting via inverse fusion of diffusion models. Proc ECCV; 2024. p. 102-18.
23. Ma B, He J. SS-SwinUnet: distilled Swin transformer for superior ocular segmentation. IEEE Trans Biom Behav Ident. 2024; 4(2):85-97.
24. Meng Q, Proença H. A comprehensive evaluation of iris segmentation on benchmarking datasets. Comput Vis Image Underst. 2024;238:104444.
25. Venkataswamy N, Bharadwaj S. Smartphone-based iris recognition through high-quality visible-spectrum capture. IEEE Sens J. 2024;24(7):6451-60.
26. Navas G, Alonso-Fernandez F. Towards iris PAD with DinoV2 and OpenCLIP foundation models. IEEE Access. 2025;13:55421-34.
27. Navas G, Alonso-Fernandez F. Towards iris presentation-attack detection with foundation models. Proc ICIP; 2025. p. 410-14.
28. Sun X, Zhao H. IrisFormer: a dedicated transformer framework for iris recognition. Proc CVPR; 2025. p. 12345-54.
29. Alaei R, Hossain M. Dual CNN and texture-based face–iris multimodal biometric system via decision-level fusion. Inf Fusion. 2025;93:123-33.
30. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In2016 fourth international conference on 3D vision (3DV) 2016 Oct 25 (pp. 565-571).
31. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2980-2988).
32. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.

33. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Transformers: State-of-the-art natural language processing. InProceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations 2020 Oct (pp. 38-45).
34. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InMedical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 2015 (pp. 234-241). Springer international publishing.
35. A. Hafner, P. Peer, Ž. Emeršič, and M. Vitek, ‘‘Deep iris feature extraction,’’ in Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC), Apr. 2021, pp. 258–262.
36. N. Prasanth, C. U. Sai Kiran, S. Nethi, T. Ketineni, N. Srinivasu, and G. Pradeepini, ‘‘Fusion of iris and periocular biometrics authentication using CNN,’’ in Proc. 7th Int. Conf. Comput. Methodol. Commun. (ICCMC), Feb. 2023, pp. 378–382.
37. N. Ebrahimpour, ‘‘Iris recognition using mobilenet for biometric authentication,’’ in Proc. 9th Int. ZEUGMA Conf. ON Sci. Res., Turkey, 2023, pp. 583–588.
38. J. G. Daugman, ‘‘High confidence visual recognition of persons by a test of statistical independence,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1148–1161, Nov. 1993.
39. F. Jan and N. Min-Allah, ‘‘An effective iris segmentation scheme for noisy images,’’ Biocybernetics Biomed. Eng., vol. 40, no. 3, pp. 1064–1080, Jul. 2020.
40. G. P. Mathias, J. H. Gagan, B. V. Mallya, and J. R. H. Kumar, ‘‘A unified approach for automated segmentation of pupil and iris in on-axis images,’’ Comput. Methods Programs Biomed. Update, vol. 2, Jan. 2022, Art. no. 100084.
41. F. Jan, N. Min-Allah, S. Agha, I. Usman, and I. Khan, ‘‘A robust iris localization scheme for the iris recognition,’’ Multimedia Tools Appl., vol. 80, no. 3, pp. 4579–4605, Jan. 2021.
42. R. Gupta and P. Sehgal, ‘‘HsIrisNet: Histogram based iris recognition to allay replay and template attack using deep learning perspective,’’ Pattern Recognit. Image Anal., vol. 30, no. 4, pp. 786–794, Oct. 2020.
43. R. Gupta and P. Sehgal, ‘‘Iris recognition using selective feature set in frequency domain using deep learning perspective: FrDIrisNet,’’ in Cybern. Cognition Mach. Learn. Appl. Proc. ICCCMLA. Springer, 2020, pp. 249–259.
44. F. He, Y. Han, H. Wang, J. Ji, Y. Liu, and Z. Ma, ‘‘Deep learning architecture for iris recognition based on optimal Gabor filters and deep belief network,’’ J. Electron. Imag., vol. 26, no. 2, Mar. 2017, Art. no. 023005.
45. A. Soni, T. Patidar, R. Kumar, K. Bharath, S. Balaji, and R. Rajendran, ‘‘Iris recognition using Hough transform and neural architecture search network,’’ Innov. Power Adv. Comput. Technol., pp. 1–5, Nov. 2021.
46. A. Arora, A. Gupta, B. Jindal, and G. Gupta, ‘‘Smart iris classification using weighted average ensemble learning,’’ in Proc. Int. Conf. Disruptive Technol. (ICDT), May 2023, pp. 124–130.
47. J. Sun, S. Zhao, Y. Yu, X. Wang, and L. Zhou, ‘‘Iris recognition based on local circular Gabor filters and multi-scale convolution feature fusion network,’’ Multimedia Tools Appl., vol. 81, no. 23, pp. 33051–33065, Sep. 2022.
48. S. Lei, B. Dong, A. Shan, Y. Li, W. Zhang, and F. Xiao, ‘‘Attention metatransfer learning approach for few-shot iris recognition,’’ Comput. Electr. Eng., vol. 99, Apr. 2022, Art. no. 107848.
49. I. Petrov and N. Minakova, ‘‘Optimization method for non-cooperative iris recognition task using daugman integro-differential operator,’’ J. Phys. Conf. Ser., vol. 1615, no. 1, Aug. 2020, Art. no. 012007.
50. Y. Akkem, S. K. Biswas, and A. Varanasi, ‘‘Smart farming monitoring using ml and MLOps,’’ in Proc. Int. Conf. Innov. Comput. Commun., Cham, Switzerland: Springer, 2023, pp. 665–675.