JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2026**, 26, 051-055

TÍCH HỢP SWIN TRANSFORMER VÀ MÃ HÓA AES-256 CHO HỆ THỐNG ĐỊNH DANH SINH TRẮC HỌC BẢO MẬT CAO

Nguyễn Duy Quang, Phùng Thế Bảo*, Trần Văn Thọ
Khoa Công nghệ thông tin, Trường Đại học Công thương Thành phố Hồ Chí Minh, Việt Nam
* Tác giả liên hệ: *baopt@huit.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận: 08/03/2026
Ngày hoàn thiện: 23/03/2026
Ngày chấp nhận: 24/03/2026
Ngày đăng: 15/04/2026

TỪ KHÓA

Swin Transformer;
Mã hóa AES-256;
Định danh sinh trắc học;
Chống giả mạo;
YOLOv8;
MiniFASNet.

TÓM TẮT

Bài báo này giải quyết các thách thức cốt lõi của hệ thống định danh sinh trắc học hiện đại: giới hạn về khả năng nắm bắt ngữ cảnh toàn cục của mạng CNN truyền thống, sự gia tăng của các cuộc tấn công giả mạo và nguy cơ rò rỉ dữ liệu nhạy cảm. Chúng tôi đề xuất một kiến trúc hệ thống bảo mật cao, tích hợp Swin Transformer để trích xuất đặc trưng khuôn mặt, YOLOv8 cho phát hiện đối tượng tốc độ cao và MiniFASNet để ngăn chặn giả mạo dựa trên phân tích miền tần số Fourier. Đặc biệt, nghiên cứu chú trọng tính riêng tư thông qua việc triển khai cơ chế mã hóa AES-256 cho vector đặc trưng sinh trắc học. Mô hình được huấn luyện theo chiến lược chuyển giao tri thức (Transfer Learning) với cơ chế đóng băng backbone (Frozen Backbone) trên tập dữ liệu VGGFace2 và kiểm thử trên dữ liệu thực tế. Kết quả thực nghiệm cho thấy hệ thống đạt chỉ số AUC là 0.794, độ chính xác chống giả mạo 96.5% và tốc độ xử lý thời gian thực xấp xỉ 36ms/khuôn mặt trên phần cứng phổ thông. Nghiên cứu chứng minh tính khả thi của việc kết hợp kiến trúc Vision Transformer tiên tiến và cơ chế bảo mật mạnh mẽ trên tài nguyên tính toán hạn chế.

INTEGRATING SWIN TRANSFORMER AND AES-256 ENCRYPTION FOR A HIGH-SECURITY BIOMETRIC IDENTIFICATION

Quang Duy Nguyen, Bao The Phung*, Tho Van Tran
Faculty of Information Technology, Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Viet Nam
*Corresponding Author: *baopt@huit.edu.vn*

ARTICLE INFO

Received: Mar 8th, 2026
Revised: Mar 23rd, 2026
Accepted: Mar 24th, 2026
Published: Apr 15th, 2026

KEYWORDS

Swin Transformer;
AES-256 Encryption;
Biometric Identification;
Anti-spoofing;
YOLOv8;
MiniFASNet.

ABSTRACT

This article addresses the core challenges facing modern biometric identification systems: the limited global context capture of traditional CNNs (Convolutional Neural Networks), the rise of sophisticated spoofing attacks, and the risk of sensitive data breaches. We propose a high-security system architecture that integrates Swin Transformer for facial feature extraction, YOLOv8 (You Only Look Once version 8) for high-speed object detection, and MiniFASNet for anti-spoofing based on Fourier frequency domain analysis. A primary focus of this research is data privacy, achieved through the implementation of an AES-256 (Advanced Encryption Standard) encryption mechanism for biometric feature vectors. The model was trained using a transfer learning strategy with a frozen backbone on the VGGFace2 dataset and validated against real-world data. Experimental results demonstrate that the system achieves an AUC of 0.794, an anti-spoofing accuracy of 96.5%, and a real-time processing speed of approximately 36ms per face on consumer-grade hardware. This study confirms the feasibility of combining advanced Vision Transformer architectures with robust security protocols under limited computational resources.

Doi: https://doi.org/10.61591/jslhu.26.1097
Available online at: https://lhj.vn

1. INTRODUCTION

In the era of digital transformation, biometric authentication plays a pivotal role in information security and digital identity management. Compared to traditional methods such as passwords or magnetic cards, facial recognition has emerged as a superior solution due to its convenience, touchless nature, and seamless integration into real-time surveillance systems. However, the rapid proliferation of facial recognition applications presents three major challenges for contemporary systems: (1) limitations in accuracy and generalization of traditional Convolutional Neural Networks (CNNs) in complex environments; (2) the increasing sophistication of spoofing attacks; and (3) the risk of leaking sensitive biometric feature data.

Regarding object detection-the foundational step in any recognition pipeline-real-time processing speed is of paramount importance. While older two-stage detectors offer high accuracy, they often suffer from significant latency [2]. To address this, one-stage detection architectures such as You Only Look Once (YOLO) have become the new industry standard. Specifically, the YOLOv8 version, with its anchor-free architecture, has demonstrated an exceptional balance between speed and accuracy, providing a robust foundation for subsequent processing stages [1].

However, the core challenge lies in the feature extraction stage. Over the past decade, CNN architectures such as ResNet, VGG, or MobileNet have dominated this field. Despite their high performance, CNNs operate based on local convolution operations, leading to an inherent limitation in the receptive field. This makes it difficult for models to capture global context and long-range dependencies across the face, particularly when subjects are partially occluded, undergo viewpoint changes, or are subject to extreme lighting conditions [4].

To overcome these limitations, the ViT (Vision Transformer) was introduced as a promising alternative, applying the Self-Attention mechanism to model relationships between image regions regardless of spatial distance. Recent comparative studies have indicated that ViTs outperform CNNs in terms of robustness and generalization [4]. However, the quadratic computational cost of global attention in ViTs remains a significant barrier to practical deployment. In this context, the Swin Transformer has emerged as a breakthrough solution. By restricting attention computation to non-overlapping local windows and utilizing a Shifted Windowing mechanism, the Swin Transformer not only reduces computational complexity to linear but also maintains the ability to learn global features through cross-window connections [5]. The efficacy of the Swin Transformer has been validated in various advanced biometric systems, such as iris recognition [5] and multimodal authentication [6], [7].

In addition to accuracy, system security is a vital factor. Facial recognition systems based on 2D (RGB) imagery are highly vulnerable to presentation attacks using printed photos or video replays. Traditional texture-based anti-spoofing methods are often ineffective against high-resolution screens. Therefore, the integration of Fourier frequency domain analysis models, such as MiniFASNet, is essential. Based on the principle that recaptured images often lose high-frequency information and contain anomalous noise, Fourier spectrum analysis allows for the accurate differentiation between live subjects and spoofing artifacts [3].

Finally, the security of feature embeddings is often overlooked in many studies that focus purely on accuracy. A facial feature vector is immutable data; if leaked, users face the risk of permanent identity loss. To address this issue, the application of robust encryption mechanisms is mandatory. The research by Nagaraju et al. [8] emphasizes the importance of combining biometrics with cryptography. In this context, the AES-256 is considered a reliable solution to ensure the confidentiality and integrity of stored data, preventing unauthorized access attacks on the database [8].

Based on the aforementioned analysis, this paper proposes a high-security biometric identification system that integrates the power of Swin Transformer for feature extraction, YOLOv8 for high-speed detection, MiniFASNet for active anti-spoofing, and AES-256 for data security. The main contributions of this research include: Part 1 proposes an optimized pipeline architecture, replacing traditional CNNs with Swin Transformer using an efficient Transfer Learning strategy; Part 2 presents the integration of an anti-spoofing mechanism based on frequency spectrum analysis, enhancing defense capabilities against 2D attacks; Part 3 describes the implementation of the AES-256 encryption process for feature vectors, ensuring user privacy; Part 4 provides comprehensive experimental evaluation conclusions on real-world datasets, demonstrating the feasibility of the system on constrained hardware devices.

2. RELATED WORKS

The development of biometric identification systems has evolved through multiple stages, from traditional image processing methods to modern deep learning models. This section provides an overview of prominent studies related to the three main aspects of the topic: Facial Recognition with Deep Learning, Anti-spoofing, and Biometric Data Security.

2.1. The Shift from CNN to Vision Transformer in Facial Recognition

Over the past decade, CNNs with representative architectures such as ResNet, VGG, MobileNet, and Inception have become the benchmark for computer vision tasks [4]. These models operate based on local convolution filters, extracting features through multiple layers to recognize objects. However, an inherent limitation of CNNs is the restricted receptive field, which makes it difficult for them to capture global context relationships of the face, especially under conditions of occlusion or significant changes in viewpoint [4].

Recently, the emergence of the ViT has marked a major turning point. Unlike CNNs, ViT divides an image into patches and utilizes a Self-Attention mechanism to model

the relationships between these patches, regardless of spatial distance [3]. Experimental comparative studies show that ViT outperforms CNNs in terms of accuracy and robustness in complex facial recognition tasks [4].

To overcome the drawback of high computational costs associated with ViT, the Swin Transformer architecture was introduced. Swin Transformer limits attention computation to local windows and employs a shifted window mechanism to create cross-connections between regions [5]. This architecture has proven effective not only in facial recognition but also in other biometric systems such as iris recognition [5] and multimodal authentication [7].

2.2. Object Detection and Anti-spoofing

Regarding the face detection problem, one-stage detectors such as the YOLO model family have gradually replaced older two-stage methods (R-CNN) due to their superior advantages in real-time processing speed [2]. YOLOv8, the latest version, significantly improves the detection of small and occluded objects thanks to its anchor-free architecture and optimized loss functions [1].

In terms of security, anti-spoofing techniques have evolved from simple texture analysis (LBP, HOG) to using Deep Learning for multimodal analysis (RGB-Depth-IR) [3]. Among these, Fourier frequency domain analysis methods (such as MiniFASNet) are highly regarded for their ability to detect high-frequency artifacts that appear in print or screen-based attacks [3], without requiring expensive depth camera hardware.

2.3. Biometric Feature Data Security

The issue of privacy protection in biometric systems is receiving increasing attention. Storing raw feature vectors (raw embeddings) directly poses an inherent risk of leaking a user's immutable identity information. Recent studies have proposed integrating cryptographic mechanisms into biometric systems [8].

Among these, the AES is currently considered the most reliable solution. Research by Nagaraju et al. [8] has indicated that combining biometric key generation with a multi-round AES cryptosystem enhances security, ensuring the confidentiality and integrity of data even in the event of a database breach.

3. Proposed Method

3.1 General System Architecture

Based on the analysis of the limitations of traditional CNNs in capturing global context and the challenges regarding biometric data security, this study proposes a multi-tier integrated system architecture. This approach combines the power of Swin Transformer for feature extraction, YOLOv8 for high-speed object detection, MiniFASNet for anti-spoofing, and the AES-256 encryption mechanism for privacy protection.

The system is designed as a sequential processing pipeline, ensuring real-time performance and high accuracy. The data flow is processed through four main stages:

1. Acquisition & Detection: Utilizing YOLOv8 to locate faces from the video stream.

2. Liveness Detection: Employing MiniFASNet to analyze the frequency spectrum and eliminate presentation attacks.

3. Feature Extraction: Using Swin Transformer Tiny to convert facial images into feature vectors (embeddings).

4. Security & Matching: Encrypting the vectors with AES-256 before storage and performing matching using the Cosine Similarity metric.

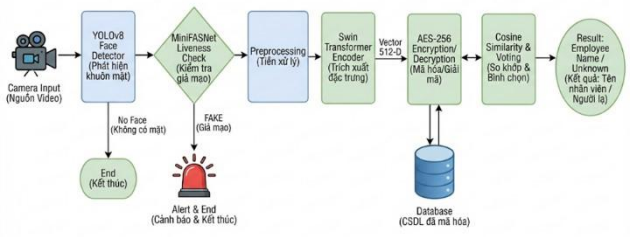

Figure 1. *General Architecture Diagram*

3.2. Face Detection with YOLOv8 (Face Detection)

In modern computer vision systems, accurate object detection is a critical prerequisite for subsequent processing stages [1]. Instead of employing legacy methods such as Haar Cascade or MTCNN, which often struggle with significant tilt angles, this research utilizes YOLOv8.

YOLOv8 implements an Anchor-free mechanism and a Decoupled Head architecture, allowing for the separation of classification tasks and bounding box regression. This enables the model to achieve an optimal balance between speed and accuracy. It is particularly effective for detecting faces across various scales and lighting conditions, effectively overcoming the latency limitations inherent in two-stage detectors.

3.3. Anti-Spoofing Based on Spectral Analysis

With the rise of sophisticated spoofing techniques, methods relying solely on pure visual features (RGB) are no longer sufficient. This research integrates the MiniFASNet model, which operates based on Fourier Frequency Domain analysis. Recent studies on multimodal facial anti-spoofing have indicated that spoofed images (from screens or printed photos) typically suffer from a loss of information in high-frequency bands and exhibit anomalous noise (such as Moiré patterns) compared to real images. MiniFASNet leverages these differences to classify subjects as Live or Spoof with low latency [3]. This component serves as the initial protective layer for the system

3.4. Feature Extraction with Swin Transformer

This is the core component of the system. While traditional CNN architectures (such as ResNet andEfficientNet) are limited by local receptive fields, (ViT) and specifically the Swin Transformer-have demonstrated

superior capabilities in capturing global context and maintaining robustness against facial variations [4], such as occlusions or changes in viewpoint.

Shifted Window Mechanism: The Swin Transformer addresses the computational overhead of the original ViT by restricting the Self-attention mechanism to non-overlapping local windows, combined with a shifted window operation between successive layers.

• W-MSA (Window-based Multi-head Self-Attention): Reduces computational complexity from quadratic to linear relative to the image size [5].

• SW-MSA (Shifted Window-based Multi-head Self-Attention): Facilitates cross-connections between windows, allowing the model to learn global features while maintaining computational efficiency [6].

Extensive comparative experiments have indicated that Vision Transformers not only achieve higher accuracy than CNNs in facial recognition tasks but also exhibit better generalization across diverse datasets [4]. Integrating the Swin Transformer into unimodal or multimodal biometric systems is currently a cutting-edge research trend to enhance system reliability [7].

3.5. Data Security with AES-256

To address the issue of biometric data leakage [8], the system does not store feature vectors in plain text. Instead, the AES-256 symmetric encryption algorithm is applied. Key generation and management are pivotal factors in ensuring the security of the cryptographic system. In the proposed model, the 512-dimensional feature vector extracted is encrypted using AES-256 in CBC (Cipher Block Chaining) mode with a random Initialization Vector (IV). This method ensures that even if the same face is scanned multiple times, the resulting ciphertext will always be different, effectively preventing replay attacks and pattern analysis.

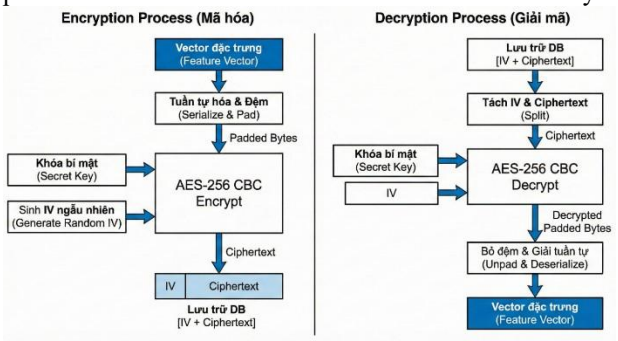

Figure 2. *Feature Encryption Process and Decryption Process*

4. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed system, experiments were conducted to measure three factors: recognition accuracy, anti-spoofing capability, and real-world processing speed.

4.1. Datasets

The study employs a multi-source data combination strategy to ensure the model possesses both generalization capabilities and adaptability to specific environments:

• VGGFace2 (Subset): A subset consisting of 197,699 images of 540 identities was used for pre-training. This data provides significant diversity in poses and ages.

• Self-collected Face Dataset: A dataset manually collected from 120 students and faculty members, containing approximately 500 images. This dataset was split at an 80:20 ratio for fine-tuning and real-world testing, respectively.

• LFW (Labeled Faces in the Wild): Utilized as a benchmark to compare the model's authentication performance against other studies.

4.2. Implementation Details

The experimental process was divided into two distinct phases with different hardware configurations, simulating a real-world deployment workflow from the Cloud down to Edge Devices.

Training Environment: The training process for the Swin Transformer model requires substantial computational resources and was therefore conducted on a cloud platform with data center-grade GPUs.

• Hardware: NVIDIA Tesla T4 GPU (16GB Vram), Intel Xeon CPU (2 vCPUs), 25GB RAM.

• Setup: The model was trained for 15 epochs with a batch size of 128. A "Frozen Backbone" strategy was applied: deep feature layers were frozen, and only the final classification layer was trained using CrossEntropy loss, utilizing the AdamW optimizer with an initial learning rate of 1e-4.

Inference and Deployment Environment: To evaluate real-time operational capability under practical conditions, the system was deployed on a mid-range personal computer.

• Hardware: GPU NVIDIA GeForce RTX 3050 (4GB VRAM), CPU Intel Core i5 Gen 11.

• Software: The system is built on Python 3.10, using the PyTorch 2.0 library and the ONNX Runtime to speed up inference.

4.3. Experimental Results

4.3.1. Recognition Performance

The Swin Transformer (Tiny) model, after the Transfer Learning process, was evaluated on the LFW benchmark and the self-collected Face Dataset. Results are presented through the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) index.

• On the LFW dataset: The model achieved an AUC of 0.794. Although this is lower than State-Of-The-Art (SOTA) models that undergo full fine-tuning on millions of images (such as ArcFace, which reaches > 99%), this result is noteworthy for a Frozen Backbone strategy on a limited dataset [4]. It demonstrates that the Swin Transformer architecture is capable of learning facial features rapidly and effectively.

• On the self-collected Face Dataset: at the optimal threshold setting of 0.31, the system achieved an identification accuracy of approximately 85%. T-SNE clustering diagrams show that feature vectors of the same

identity exhibit distinct clustering, separated from other identities, confirming the feature extraction efficacy of the Shifted Windows mechanism.

4.3.2. Anti-Spoofing Performance Evaluation (Anti-Spoofing Results)

The MiniFASNet model was tested against common spoofing attack scenarios, including Print Attacks and smartphone Replay Attacks.

• Results: The system accurately detected 96.5% of smartphone screen spoofing cases by analyzing the loss of high-frequency information in the Fourier spectrum.

• Advantages: Compared to traditional texture-based methods (such as LBP) [3], MiniFASNet is less affected by environmental lighting conditions and does not require user interaction (Passive Detection).

4.3.3. Inference Speed Evaluation

Speed is a key factor for user experience. Testing on an NVIDIA RTX 3050 GPU with an HD (720p) video stream revealed the average latency of the entire processing pipeline.

Table 1. *Average processing time per stage on RTX 3050*

Stage	Model / Algorithm	Time (ms)
Face Detection	YOLOv8n	~12 ms
Anti-spoofing	MiniFASNetV2	~4 ms
Feature Extraction	Swin Transformer Tiny	~15 ms
Encryption & Matching	AES-256 + Cosine	~5 ms
TOTAL		**~36 ms**

The total processing time is approximately 36ms per face, which is equivalent to a speed of ~27 frames per second (FPS). This result confirms that the system fully meets real-time requirements on consumer-grade hardware, outperforming heavy CNN architectures or legacy two-stage detection methods [2].

5. CONCLUSION

This paper proposes a high-security biometric identification system that integrates Swin Transformer for robust facial feature extraction, YOLOv8 for fast face detection, MiniFASNet for anti-spoofing based on Fourier frequency domain analysis, and AES-256 encryption to protect biometric feature vectors.

By employing a transfer learning strategy with a Frozen Backbone on the VGGFace2 dataset, the Vision Transformer architecture with the Shifted Windows mechanism demonstrates strong capability in capturing global context and long-range dependencies, even when trained on limited data. Experimental results on the LFW benchmark show an AUC of 0.794, while the system achieves 96.5% accuracy in detecting spoofing attacks (print and screen replay). The complete pipeline runs in real time at approximately 36 ms per face on consumer-grade hardware (NVIDIA RTX 3050), confirming its practical deployability without requiring high-end servers.

The integration of MiniFASNet and AES-256 effectively mitigates both presentation attacks and data leakage risks, ensuring confidentiality and integrity of user biometric information. This research validates the feasibility of deploying advanced Vision Transformer models combined with strong cryptographic protection on resource-constrained devices.

Future work will focus on full fine-tuning of the Swin Transformer, expanding the self-collected dataset, and further optimizing the system for mobile and edge devices to approach state-of-the-art accuracy while maintaining real-time performance.

6. REFERENCES

[1] M. Elgendy, Deep Learning for Vision Systems. Manning Publications, 2020.

[2] J. Brownlee, Deep Learning for Computer Vision: Image Classification, Object Detection and Face Recognition in Python. Machine Learning Mastery, 2019.

[3] Z. Yu et al., "Rethinking Vision Transformer and Masked Autoencoder in Multimodal Face Anti-Spoofing," International Journal of Computer Vision, vol. 132, pp. 5217–5238, 2024.

[4] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," Scientific Reports, vol. 14, art. no. 21392, 2024.

[5] R. Gao and T. Bourlai, "On Designing a SwinIris Transformer Based Iris Recognition System," IEEE Access, vol. 12, pp. 30737–30752, 2024.

[6] S. Sharhrah et al., "Multimodal Biometric Authentication using Convolutional Neural Network, Swin Transformer, and Multi-Head Attention with Global Max Pooling," ITM Web of Conferences, vol. 79, art. no. 01052, 2025.

[7] R. Garg, P. Pathak, and M. P. Singh, "A multimodal biometric recognition system based on Fingerprints, Iris and ECG via Swin Transformer and CNN Model," Systems and Soft Computing, vol. 7, art. no. 200369, 2025.

[8] S. Nagaraju et al., "Biometric key generation and multi round AES crypto system for improved security," Measurement: Sensors, vol. 30, art. no. 100931, 2023.