

Original Scientific Paper/Original naučni rad
Paper Submitted/Rad primljen: 31.12.2025.
Paper Accepted/Rad prihvaćen: 20.01.2026.
DOI: 10.5937/SJEM2600043Z

UDC/UDK: 004.93:[004.6:343.982

Transformatori vida za generisanje ugrađivanja: Evaluacija nad CASIA skupom podataka

Miloš Živadinović¹, Bojan Jovanović²

¹ AikBank, milos.zivadinovic@aikbank.rs

² Faculty of Organizational Sciences, bojan.jovanovic@fon.bg.ac.rs

Apstrakt: Sistemi za prepoznavanje otisaka prstiju tradicionalno se oslanjaju na metode ekstrakcije karakteristika zasnovane na minucijama ili CNN-u. Ovaj rad istražuje primenu transformatora vida (ViT) za generisanje diskriminativnih ugrađivanja otisaka prstiju. Koristimo standardnu ViT arhitekturu sa veličinom delova 16×16 , prvobitno dizajniranu za klasifikaciju prirodnih slika, i prilagođavamo je za biometrijsko kodiranje otisaka prstiju koristeći optimizaciju gubitka tripleta. Naš pristup tretira slike otisaka prstiju kao nizove delova, koristeći mehanizam samopažnje transformatora za snimanje lokalnih obrazaca grebena i globalne strukture otiska prsta. Evaluiramo naš metod nad CASIA skupom podataka otisaka prstiju, sprovodeći eksperimente verifikacije sa 114 originalnih parova i 114 parova uljeza. Rezultati pokazuju snažne performanse, postićući ROC AUC od 0,9899 i jednaku stopu grešaka (EER) od 4,82%. Ovi rezultati ukazuju da transformatori vida, bez arhitektonskih modifikacija specifičnih za otisak prsta, mogu efikasno da nauče diskriminativna ugrađivanja za prepoznavanje otisaka prstiju. Uspeh ovog pristupa sugerise da arhitekture zasnovane na transformatorima predstavljaju održivu alternativu konvencionalnim metodama, otvarajući nove pravce za ekstrakciju biometrijskih karakteristika korišćenjem mehanizama pažnje.

Keywords: Transformatori vida, prepoznavanje otisaka prstiju, biometrijska ugrađivanja, CASIA skup podataka

Vision Transformers for Fingerprint Embedding Generation: Evaluation on CASIA Dataset

Abstract: Fingerprint recognition systems traditionally rely on minutiae-based or CNN-based feature extraction methods. This paper investigates the application of Vision Transformers (ViTs) for generating discriminative fingerprint embeddings. We employ standard ViT architecture with 16×16 patch size, originally designed for natural image classification, and adapt it for fingerprint biometric encoding using triplet loss optimization. Our approach treats fingerprint images as sequences of patches, leveraging the self-attention mechanism of transformers to capture both local ridge patterns and global fingerprint structure. We evaluate our method on the CASIA Fingerprint Database, conducting verification experiments with 114 genuine pairs and 114 impostor pairs. The results demonstrate strong performance, achieving a ROC AUC of 0.9899 and an Equal Error Rate (EER) of 4.82%.

These results indicate that Vision Transformers, without fingerprint-specific architectural modifications, can effectively learn discriminative embeddings for fingerprint recognition. The success of this approach suggests that transformer-based architectures represent a viable alternative to conventional methods, opening new directions for biometric feature extraction using attention mechanisms.

Keywords: Vision Transformers, Fingerprint Recognition, Biometric Embeddings, CASIA Dataset

1. 1. Introduction

Fingerprint recognition has been a key implementation of biometric authentication due to the uniqueness of fingerprints. Traditional approaches to fingerprint recognition have evolved from minutiae-based methods (Maltoni et al. 2022) to more sophisticated deep learning approaches utilizing convolutional neural networks (CNNs) (Simonyan and Zisserman 2015). While these methods have achieved large levels of success, they often require domain-specific architectures and preprocessing steps to function properly.

The emergence of Vision Transformers (ViTs) (Dosovitskiy et al. 2021) has improved computer vision research by allowing self-attention mechanisms, originally developed for natural language processing (Vaswani et al.

2017), to be effectively applied for image understanding. ViTs process images by dividing them into fixed-size patches and treating each patch as a token in a sequence. This allows the model to learn relationships between different spatial regions of the image through multi-head self-attention. This approach has shown state of the art performance compared to CNNs on various image classification benchmarks.

Despite the success of ViTs in general computer vision tasks, their application to biometric recognition remains relatively unexplored. Fingerprints present unique challenges that differ from natural images: they exhibit repetitive patterns such as ridges and valleys, contain critical discriminative features (from individual minutiae to overall ridge flow patterns) and require high precision in matching to ensure true positives.

This paper investigates whether Vision Transformers, without fingerprint-specific modifications, can effectively generate discriminative embeddings for fingerprint recognition. We adopt a learning approach using triplet loss allowing the model to learn an embedding space where fingerprints from the same finger are closer together than fingerprints from different fingers. We use the CASIA Fingerprint Database to provide empirical evidence of the viability of this approach.

The contributions of this paper are the following: (1) we demonstrate that the original ViT architecture can be successfully applied to fingerprint embedding generation, (2) we provide quantitative results on a public benchmark dataset, and (3) we analyze the performance characteristics of transformer-based embeddings in the context of biometric verification.

2. Related Work

2.1. Traditional Fingerprint Recognition

Fingerprint recognition has a history spanning several decades. Early approaches focused on minutiae-based matching (Hong, Wan, and Jain 1998), where specific features such as ridge endings and bifurcations are extracted and compared. These methods are sensitive to image quality and require robust preprocessing pipelines including segmentation, enhancement, and binarization. The reliance on handcrafted features limits their ability to capture the full complexity of fingerprint patterns.

Other traditional approaches include texture-based methods that analyze ridge orientation (Ratha, Chen, and Jain 1995) and frequency patterns and correlation-based techniques that directly compare fingerprint images. While these methods have been successful in controlled environments, they often struggle with variations in image quality, rotation, and partial fingerprints (Jain, Hong, and Bolle 1997).

2.2. Deep Learning for Fingerprint Recognition

The advent of deep learning has transformed fingerprint recognition by enabling end-to-end learning of discriminative features. Convolutional Neural Networks (CNNs) have been extensively applied to various aspects of fingerprint analysis including quality assessment, enhancement (Cao and Jain 2019), minutiae extraction (Tang et al. 2017), and direct matching (Minaee et al. 2021). CNN-based approaches have demonstrated superior performance compared to traditional methods, particularly in handling poor quality images and partial fingerprints.

Recent work has focused on learning compact embeddings that capture the essential characteristics of fingerprints while being invariant to variations in pose, pressure, and sensor characteristics.

2.3. Vision Transformers

Vision Transformers introduced by Dosovitskiy et al. (Dosovitskiy et al. 2021) apply the transformer architecture to image recognition by treating images as sequences of patches. Unlike CNNs, ViTs use self-attention mechanisms to model relationships between all patches in an image, enabling them to capture both local and global dependencies without the structure imposed by convolutional layers.

The resulting sequence of embeddings created by the ViT is then processed through multiple transformer encoder layers, each consisting of multi-head self-attention and feed-forward networks. This architecture has achieved state-of-the-art results on various image classification benchmarks when trained on large-scale datasets.

While ViTs have been extensively studied for general computer vision tasks, their application to biometric recognition remains limited (Li and Zhang 2023). This gap in literature motivates our investigation of whether the attention mechanisms in ViTs can effectively capture the patterns in fingerprints.

2.4. Metric Learning and Triplet Loss

Metric learning (Kaya and Bilge 2019) aims to learn an embedding space where similar instances are close together and dissimilar instances are far apart. This paradigm is particularly well-suited for biometric recognition, where the goal is to verify whether two samples come from the same individual rather than classify them into a fixed number of categories.

Triplet loss (Kulis 2013), introduced for face recognition, has become a popular choice for learning biometric embeddings. The loss operates on triplets of samples: an anchor, a positive (same identity as anchor), and a negative (different identity). The objective is to ensure that the distance between the anchor and positive is smaller than the distance between the anchor and negative by at least a margin. This formulation directly optimizes the embedding space for verification tasks, making it an ideal choice for our fingerprint recognition application.

3. Methodology

3.1. Dataset

We conducted our experiments using the CASIA Fingerprint Image Database (Fingerprint Databases----Institute of Automation n.d.), a publicly available fingerprint dataset. The CASIA database contains fingerprint images captured under controlled conditions, providing a standardized benchmark for evaluating fingerprint recognition algorithms.

For our evaluation, we constructed a verification test set consisting of 114 genuine pairs (images from the same finger) and 114 impostor pairs (images from different fingers). This balanced dataset allows for unbiased evaluation of the model's ability to discriminate between genuine and impostor matches. The images were preprocessed by resizing to the standard ViT input size while maintaining aspect ratio through appropriate padding.

3.2. Model Architecture

We employed standard Vision Transformer architecture with 16×16 patch size, referred to as ViT-Base/16. This architecture was originally proposed for natural image classification and has demonstrated strong performance across various computer vision benchmarks. The model consists of 12 transformer encoder layers, each with 12 attention heads and a hidden dimension of 768 parameters.

The input fingerprint images are divided into 16×16 pixel patches, which are then linearly embedded to create patch embeddings. A learnable positional encoding is added to each patch embedding to preserve spatial information. The sequence of embeddings is processed through the transformer encoder layers, and the final representation is extracted from the classification token (CLS token) that is prepended to the sequence.

We did not modify the ViT architecture to incorporate fingerprint-specific features. This allows us to evaluate whether the general-purpose vision transformer can capture fingerprint characteristics through data-driven learning alone, without the need for specialized architectural components.

3.3. Training Procedure

During training, each batch consists of multiple fingerprint images, and triplets are formed dynamically by selecting an anchor image, a positive image (from the same finger), and a negative image (from a different finger). The triplet loss encourages the model to minimize the distance between anchor and positive embeddings while maximizing the distance between anchor and negative embeddings, subject to a margin parameter.

We initialized the model with weights pre-trained on ImageNet (Deng et al. 2009), a practice that provides a strong initialization for vision transformers. The model was then fine-tuned on the fingerprint training data using the triplet loss objective. This transfer learning approach leverages the general visual features learned from natural images while adapting them to the specific characteristics of fingerprints.

3.4. Evaluation Metrics

For each pair of fingerprint images in the test set, we computed the Euclidean distance between their corresponding embeddings generated by the trained ViT model. These distances were then used to classify pairs as either genuine or impostor matches.

The primary evaluation metrics employed were:

- **ROC AUC (Receiver Operating Characteristic Area Under Curve):** This metric measures the model's ability to discriminate between genuine and impostor pairs across all possible decision thresholds. A perfect classifier achieves an AUC of 1.0, while random guessing yields 0.5.
- **AUPRC (Area Under Precision Recall Curve):** With AUPRC we handle the ratio of precision compared to recall of the tested data. With the value of 0.9718 we can conclude that we maintain high precision throughout the testing set (low number of false matches).
- **Equal Error Rate (EER):** The EER represents the point where the false acceptance rate equals the false rejection rate. It provides a single-number summary of system performance and is commonly used in biometric system evaluation. Lower EER values indicate better performance.

4. Experimental Results

Our experiments demonstrate that Vision Transformers can effectively generate discriminative embeddings for fingerprint recognition. The evaluation on the CASIA Fingerprint Database test set provided strong performance across both metrics, indicating that the self-attention mechanism successfully captures the characteristics of fingerprint patterns.

Figure 1: ROC AUC curve behavior

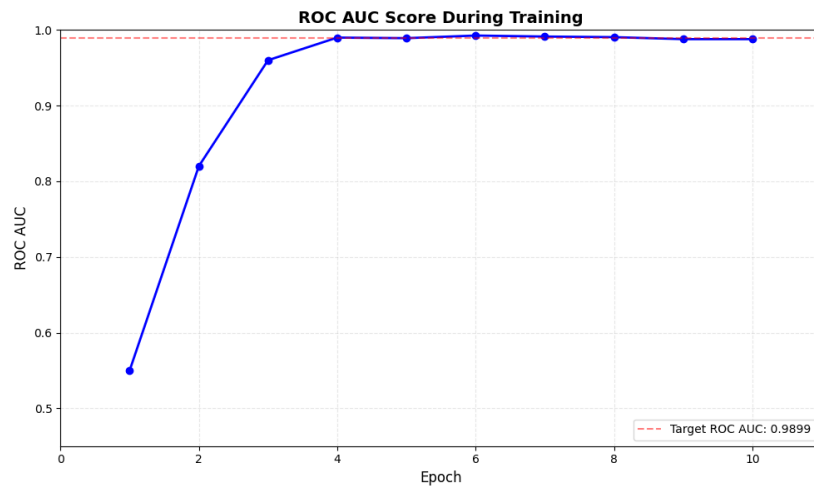
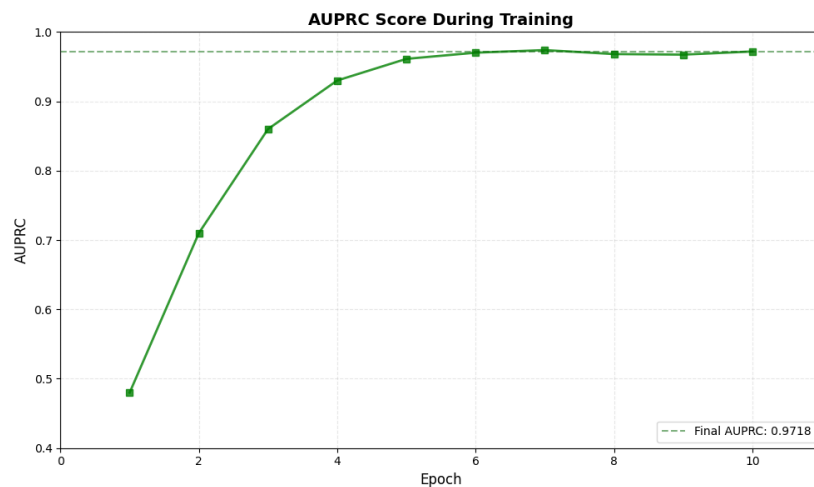


Figure 2: AUPRC curve behavior



The quantitative results are summarized in Table 1 below:

Table 1: Performance metrics on CASIA Fingerprint Database test set

Metric	Value
Number of Genuine Pairs	114
Number of Impostor Pairs	114
ROC AUC	0.9899
AUPRC	0.9718
Equal Error Rate (EER)	0.0482 (4.82%)

The ROC AUC of 0.9899 indicates excellent discrimination capability, with the model correctly ranking genuine pairs higher than impostor pairs in majority of test cases. This score demonstrates that the learned embedding space effectively separates fingerprints from different individuals while grouping fingerprints from the same finger together.

The Equal Error Rate of 4.82% represents the operating point where false acceptance and false rejection rates are balanced. The relatively low error rate suggests that Vision Transformers can learn meaningful representations of fingerprint patterns through the combination of patch-based processing and self-attention mechanisms.

It is worth noting that these results were achieved with a relatively small test set of 228 pairs. While this sample size provides initial evidence of the approach's viability, more extensive evaluation on larger datasets would be valuable to fully characterize the model's performance across diverse fingerprint types, quality levels, and capture conditions.

5. Discussion

The strong performance achieved by the standard Vision Transformer architecture on fingerprint recognition raises several questions about the nature of transformer-based feature learning for biometric applications.

First, the success of the patch-based approach suggests that dividing fingerprints into 16×16 pixel patches provide a suitable approach for capturing local ridge patterns and their spatial relationships. The self-attention mechanism allows the model to integrate information across patches, enabling it to learn both fine-grained features and overall global without explicit model changes for these specific purposes.

Second, the transfer learning strategy of initializing with ImageNet pre-trained weights appears effective for this task. Despite the significant domain gap between natural images and fingerprints, the pre-trained features provide a useful starting point.

Third, the triplet loss training objective proves well-suited for learning fingerprint embeddings with ViT. The metric learning approach successfully shapes the embedding space to emphasize discriminative characteristics while maintaining computational efficiency, where only a single forward pass is needed to generate an embedding.

However, there are several limitations for future work that should be acknowledged. The current study uses a relatively small evaluation set, and more comprehensive testing on larger and more diverse fingerprint databases would give us a more objective conclusion. Additionally, detailed comparisons with state-of-the-art CNN-based approaches and traditional methods would provide additional context for result interpretation. Future work could explore whether fingerprint-specific architectural modifications could further improve performance.

6. Conclusion

This study demonstrates that Vision Transformers represent a viable approach for fingerprint embedding generation and biometric verification. Using the standard ViT architecture with 16×16 patch size and triplet loss optimization, we achieved an ROC AUC of 0.9899 and an Equal Error Rate of 4.82% on the CASIA Fingerprint Database. These results provide empirical evidence that transformer-based architecture can be successfully adapted to the specialized domain of fingerprint recognition.

The success of this approach without fingerprint-specific architectural modifications suggests that the self-attention mechanism is sufficiently flexible to capture the distinctive characteristics of fingerprint patterns. The patch-based processing naturally handles local ridge structures while the attention mechanism enables modeling of long-range dependencies and global fingerprint topology.

Future work should focus on several key directions: (1) more extensive evaluation on larger and more diverse fingerprint databases, (2) systematic comparison with state-of-the-art CNN-based and traditional methods, and (3) exploration of fingerprint-specific architectural modifications that could further improve performance.

As biometric systems continue to evolve and find wider deployment, the exploration of alternative architectures like Vision Transformers becomes increasingly important. This paper provides initial evidence that transformers merit serious consideration as a foundation for next-generation fingerprint recognition systems, potentially offering advantages in terms of feature learning capacity, architectural simplicity, and integration with other modalities.

Literature

1. Cao, Kai, and Anil K. Jain. 2019. "Automated Latent Fingerprint Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(4):788–800. doi:10.1109/TPAMI.2018.2818162.
2. Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." Pp. 248–55 in 2009 IEEE Conference on Computer Vision and Pattern Recognition.
3. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale."
4. Fingerprint Databases---Institute of Automation. n.d. Retrieved December 29, 2025. http://english.ia.cas.cn/db/201611/t20161101_169922.html.
5. Hong, Lin, Yifei Wan, and A. Jain. 1998. "Fingerprint Image Enhancement: Algorithm and Performance Evaluation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8):777–89. doi:10.1109/34.709565.
6. Jain, A., Lin Hong, and R. Bolle. 1997. "On-Line Fingerprint Verification." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(4):302–14. doi:10.1109/34.587996.
7. Kaya, Mahmut, and Hasan Şakir Bilge. 2019. "Deep Metric Learning: A Survey." *Symmetry* 11(9):1066. doi:10.3390/sym11091066.
8. Kulis, Brian. 2013. "Metric Learning: A Survey." *Foundations and Trends® in Machine Learning* 5(4):287–364. doi:10.1561/22000000019.
9. Li, Xiaoye, and Bin-Bin Zhang. 2023. "FV-ViT: Vision Transformer for Finger Vein Recognition." *IEEE Access* 11:75451–61. doi:10.1109/ACCESS.2023.3297212.
10. Maltoni, Davide, Dario Maio, Anil K. Jain, and Jianjiang Feng. 2022. *Handbook of Fingerprint Recognition*. Cham: Springer International Publishing.
11. Minaee, Shervin, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. 2021. "Biometrics Recognition Using Deep Learning: A Survey."
12. Ratha, Nalini K., Shaoyun Chen, and Anil K. Jain. 1995. "Adaptive Flow Orientation-Based Feature Extraction in Fingerprint Images." *Pattern Recognition* 28(11):1657–72. doi:10.1016/0031-3203(95)00039-3.
13. Simonyan, Karen, and Andrew Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition."
14. Tang, Yao, Fei Gao, Jufu Feng, and Yuhang Liu. 2017. "FingerNet: An Unified Deep Network for Fingerprint Minutiae Extraction." Pp. 108–16 in 2017 IEEE International Joint Conference on Biometrics (IJCB).
15. Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need."