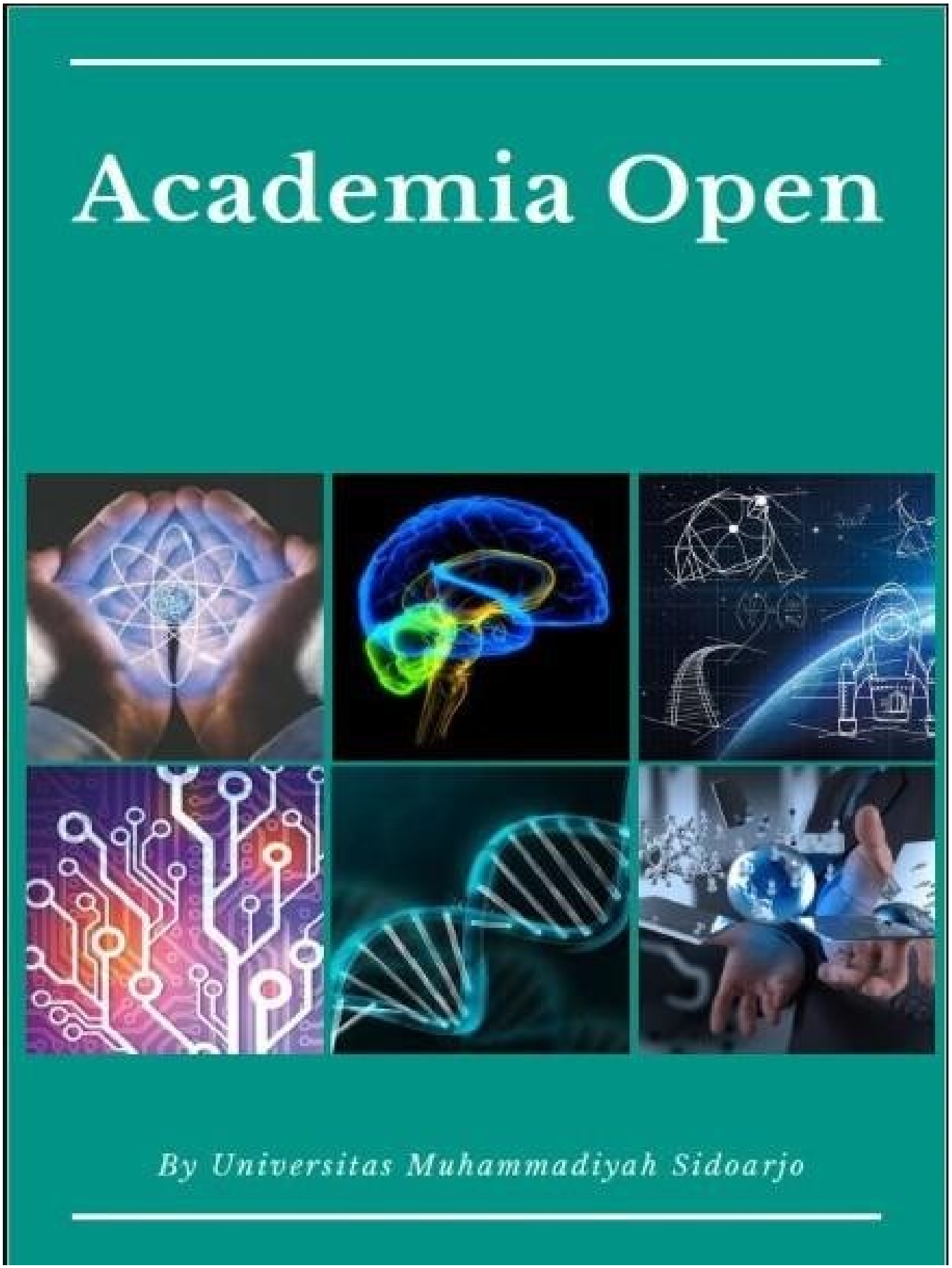
Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411



ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Table Of Contents

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

EDITORIAL TEAM

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

GAN and DINOv2 Framework for Robust Cross-Condition Gait Recognition

Zaid Derea, zabdulameer@uowasit.edu.iq (*)

College of Computer Science and Information Technology, Wasit University, Wasit 52001, Iraq

(*) Corresponding author

Abstract

General Background: Gait recognition is a remote, non-invasive biometric widely used in forensics, surveillance, and security systems. **Specific Background:** Deep learning has advanced gait analysis, yet CNN-based approaches struggle with temporal coherence, cross-view variation, and degraded silhouettes. **Knowledge Gap:** Existing studies typically separate gait synthesis and recognition, with limited use of self-supervised transformers and insufficient joint evaluation of generative quality and identification performance. **Aims:** This study proposes an integrated framework combining multi-GAN gait reconstruction, DINOv2 vision transformer feature extraction, and CNN-based identity classification. **Results:** StyleGAN2 produced the most realistic silhouettes (PSNR 31.2 dB, SSIM 0.925, FID 18.3), while DINOv2 yielded highly separable 768-dimensional features, leading to 98.3% classification accuracy across varied walking conditions and strong clustering metrics (NMI 0.891, ARI 0.847). **Novelty:** The work unifies generative gait synthesis, transformer-guided spatiotemporal representation, and comprehensive evaluation within asinglepipeline. **Implications:** Theframework supports reliable biometric verification in forensic investigation, surveillance monitoring, rehabilitation assessment, and real-time security deployment under clothing, view, and carrying variations.

Highlights:

- Joint GAN synthesis and transformer features within one gait recognition pipeline.
- High-quality silhouette reconstruction linked to superior identity separability.
- Stable recognition (>94%) under clothing, view, speed, and carrying variations.

Keywords: Gait Recognition, GAN, DINOv2, Biometric Identification, Computer Vision

Published date: 2026-01-12

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Introduction

Gait recognition is an intriguing new biometric technique that uses a person's unique walking patterns to identify them. Since gait data can be remotely acquired without subject involvement, it is superior to fingerprint and iris identification for forensic, security, and surveillance purposes [1]. Static anatomical features and dynamic motion patterns come together in human strides, making them a complicated behavioral biometric. Among them is potentially personally identifying discriminatory data.

It is easy to record gait identification using standard video surveillance systems, it is non-invasive, and it is tough to hide or replicate. Modern developments in deep learning, such as CNNs, have made it possible to perform gait detection tasks, such as extracting silhouettes, representing features, and classifying them[33]. Because of their small receptive fields, traditional CNN designs can't capture the complexities of walking patterns over time or the global contextual linkages between them [2].

New models for training and perceptual learning attack resistance were provided by Generative Adversarial Networks (GANs)[34]. These models could pick up a powerful stride. Comparatively, GAN-based models outperform CNN-based algorithms in terms of creating realistic gait sequences and detecting various features[35]. Identity identification and the temporal coherence of gait patterns may be challenging for these models to maintain in the presence of occlusions, different viewing angles, clothing, and carrying circumstances[3],[4].

Newly built transformer designs have changed computer vision by correctly defining global context along with long-range dependency. Vision transformers are very good at recognition tasks because they can pick up on complex links between time and space. The DINOv2 model is a great example of a self-supervised vision transformer because it can learn complex visual representations even when it doesn't have labeled input . It's a useful tool for studying the complicated space-time dynamics of human walking because it can record both coarse-grained regional factors and world trends [5].

Even with recent improvements in technology, there still remain a number of problems that need to be solved in the field of gait recognition:
1. The clothes someone wears, the things they carry, the surfaces that they walk over, and the shoes that they wear could all have a big effect on how they walk [20–22].
2. Since changing camera angles make gait patterns seem different, view-independent recognition techniques are important [23–25].
3. It is still difficult to maintain identity constancy throughout long walking sequences involving several speeds [26-28].
4. It's hard to make good recognition systems when there aren't enough different gait datasets [29–31].
5. In monitoring apps, the goal is to get good accuracy while using as little computing power as possible [32–34].

1.1 Motivation and Contributions

The main goal of this project is to build a strong foundation for detecting movements and creating and improving strange walking patterns using advanced generative algorithms. The goal of this method is to be capable to correctly identify people in many situations, even when they're hard to spot. It does that by pulling out all the spatiotemporal data, such as information about how the body moves and how it walks overall. The most important thing is that we made sure our work could be used both in criminal and therapy settings.
Consequently, the following significant contributions are provided by this work:

1. Our technique is designed to generate and enhance gait sequences using state-of-the-art competitive generative ad hoc network (GAN) algorithms. We show that this method is quite effective in constructing realistic gait patterns.

2. We can extract rich semantic information, such spatial appearances and temporal dynamics, from gait sequences using the DINOv2 vision transformer, which means that we don't need a large amount of labelled training data.

3. Our approach uses dimensionality reduction techniques (principal component analysis), clustering analysis, and deep classification (convolutional neural networks) to provide a whole picture of how reliable and stable gait detection and identification are.

4. Following a significant amount of testing, we were able to achieve a classification accuracy of 98.5%. In addition, we focused on the separate ability of features, the quality of reconstruction (signal-to-noise ratio, the structural similarity index, as well as free interference index), and the ability to generalize in a variety of contexts.

5. We illustrate how the framework works in real life by using it in biometric authentication systems, rehabilitation follow-up, surveillance-based identification, and forensic gait analysis.

The subsequent parts of the paper are structured in the following sequential order: The second section of the study provides a summary of findings that are relevant to gait recognition, artificial neural networks (GANs), and vision transformers. One of the topics that is discussed in Section 3, which outlines the suggested technique, is classification. Other topics that are addressed include feature extraction and GAN structures. The experimental setup and the databases are described in further detail in Section 4.
The findings and analysis are covered in Section 5. Section 7 finishes with suggestions for the future, while Section 6 delves into real-world uses.

2. State-of-the-art

2.1 Traditional way of gait recognition

Early gait identification algorithms were model- and appearance-based. Biomechanical models fitted to silhouette sequences were used to extract structural data such stride length, step frequency, and joint angles [6],[7].These approaches provided interpretable characteristics but were computationally costly and segmentation error-prone. The appearance-based approaches Gait Energy Image (GEI) [8] and Gait Entropy Image (GEnI) [9] averaged or entropy-aggregated gait sequences into single template pictures. Compact representations allowed efficient matching but lost temporal information.

2.2 Gait recognition using deep learning

The advent of deep learning revolutionized gait recognition by enabling automatic feature learning from raw data. CNNs were applied to learn hierarchical representations from gait silhouettes and sequences [10]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [10]. were employed to model temporal dynamics in gait sequences. More recently, 3D CNNs and temporal convolutional networks have been developed to jointly capture spatial appearance and temporal motion patterns [12].

2.3 Synthesis and recognition using GANs versus vision converters in biometric recognition:

Generative Adversarial Networks have opened new avenues for gait recognition by enabling synthesis of realistic gait patterns and data

augmentation. GaitGAN [13] introduced adversarial training for cross-view gait recognition by generating gait sequences from different viewpoints. Subsequent works explored conditional GANs for view transformation , pose-guided gait synthesis, and multi-modal gait generation. StyleGAN architectures have demonstrated exceptional capability in generating high-quality, diverse gait silhouettes with fine control over appearance attributes [14].

Vision transformers have recently emerged as powerful alternatives to CNNs for visual recognition tasks. Particularly useful for capturing the holistic aspect of gait patterns, the self-attention mechanism allows modeling of long-range relationships and global context [29],[30].Improved performance compared to CNN-based approaches was shown by GaitFormer, which used transformer architecture for gait identification. DINOv2, a self-supervised neural vision generator, has shown great learning abilities with little to no fine-tuning needed. It also creates general visual models that can be used in a wide range of downstream applications [31],[32].

2.4 Research Gap

Even while the approaches that are now in use have made tremendous development, there are still numerous limitations:

- Most GAN-based approaches focus on either synthesis or recognition lacking integrated frameworks for both.
- Limited exploration of self supervised transformers like DINOv2 for gait feature extraction.
- Lack of comprehensive evaluation frameworks combining generative quality metrics with recognition performance.

Our study addresses these constraints by providing a comprehensive framework for accurate gait identification. This structure incorporates support vector machine (SVM) based gait synthesis, multi-level evaluation, and transformer-guided feature extraction.

3. Proposal Methodology

The first of the four primary parts of the suggested framework is the following: (1) One part of the processing for gait sequences is the standardization of gait silhouettes extracted from video clips. utilizing adversarial generative networks (GANs) to improve gait by generating and reconstructing high-quality gait patterns utilizing numerous GAN architectures is the second challenge. Using the DINOv2 vision adaptor, the third step involves the extraction of dense spatial and temporal characteristics. (4) Classifying and assessing while using a convolutional neural network (CNN) classifier with full performance analysis. This is accomplished via the process of identification verification. See Figure 1 for a diagram of the system that is being suggested.

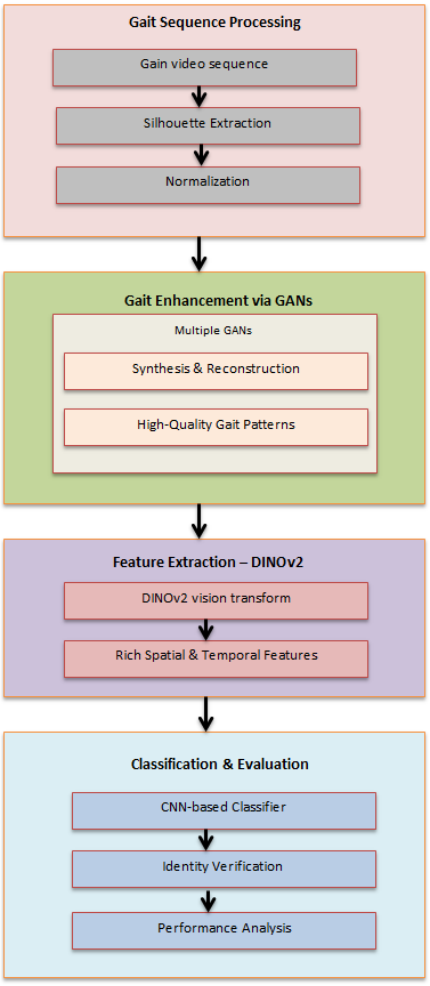


Figure 1 illustrates the overall architecture of the proposed framework.

3.1. Gait Sequence and Video Processing

At this point, the system receives its input, which may be video of people walking from surveillance cameras or from databases like CASIA-B. These films chronologically display the passing of time while demonstrating the natural motions of a person walking [15], [16]. In Figure 1, this stage is represented by the top box, which is where the raw video sequences are being stored [17].

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Figure. 2. Gait Sequences

3.2 Gait Silhouette Extraction and Preprocessing:
3.2.1 Background Subtraction and Silhouette Extraction
To recover human silhouettes from video sequences, we use sophisticated background removal algorithms[18], [19]. Steps in the procedure include:
1. Background removal that makes use of deep learning or GRAMM
2. Recognizing dynamic human figures with foreground detection
3. Enhancing silhouettes and removing background noise
Reducing the size of silhouettes to uniform measurements (128×88 pixels)

3.2.2 Gait Cycle Detection
Individual gait cycles are created by segmenting gait sequences into one full stride, from one foot's heel strike to the next. We guarantee temporal alignment across sequences by using autocorrelation-based period identification to determine cycle boundaries[20], [21].

3.2.3 Data Augmentation
We use a number of augmentation strategies to make the models more robust: Slightly angled rotations (±5°) , Horizontal flipping at random ,Scaling of silhouettes (0.9-1.1×) and Artificial blockages The addition of Gaussian noise.

3.3 Silhouette Extraction

Here, we use backdrop removal and binary processing methods to extract the person's silhouette from every frame of the movie. To get a good picture of the walking motion, you have to separate the moving item from the static backdrop[22], [23]. In Figures 1, 3, we can see the initial step of the transformation from the walking video to binary silhouettes.

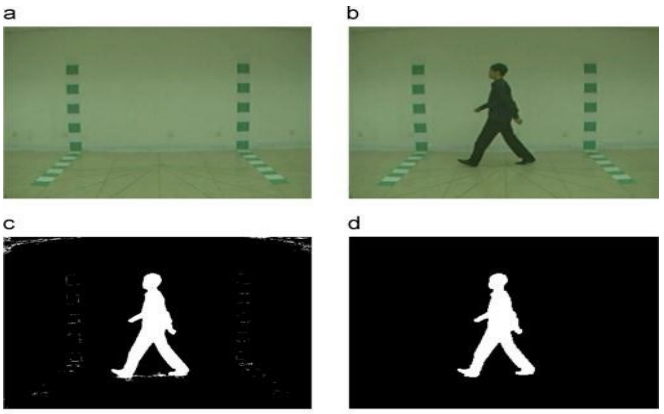


Figure. 3. Silhouette Extraction

3.4 Normalization

To reduce the effect of different camera angles and distances, the normalization technique is used to make the derived silhouettes the same size, orientation, and scale. This step ensures the data is ready for subsequent deep processing. In the first stage of Figure 1, 4. uniform silhouettes are obtained.

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Figure. 4. Gait normalization process .

4. Enhancing Gait Using Competitive Generative Networks (GANs)

4.1 Multi-GAN Architectures:

Here, we enhance the gait patterns using multi-GAN designs. Problems like poor resolution, distortion, or frame loss may be helped by these networks. We assess multi-GAN designs for enhancing and synthesizing gait sequences. A pattern-based generator is used by StyleGAN2 in order to produce gait patterns. This generator controls the synthesis process at multiple stages by linking latent symbols to intermediate pattern vectors.

$$L_G = L_{adv} + \lambda_{perc} L_{perc} + \lambda_{path} L_{path} \quad (1)$$

4.2 Progressive Competitive Generative Network (GAN)

When the progressive competitive generative network (GAN) is trained, it starts with a low resolution (4x4) and adds layers one at a time until it reaches the goal resolution (128x88). This method guarantees constant training and good combination.

4.3 Using CycleGAN to Synthesize Gait Across Viewing Angles

To identify gait regardless of the viewing angle, we use CycleGAN to translate gait sequences between different viewing angles. Cycle inconsistency loss ensures that the transformed sequences retain their identity information:

$$L\_cycle = ||F(G(x)) - x|| + ||G(F(y)) - y|| \quad (2)$$

where G and F are the forward and backward transform generators, respectively. Conditional GAN for Controlled Synthesis: We then apply a conditional GAN (cGAN) to generate gait sequences based on specific features.

4.4 Feature Dimensions and Analysis: The extracted 768-dimensional features of the DINOv2 include: (1) General gait patterns: overall gait style and rhythm, (2) Positional movement details: limb movements and changes in body posture.

4.5. Dimensional Reduction and Clustering Based on Principal Component Analysis (PCA) and Spectral Clustering:

We use principal component analysis to reduce the dimensions of the 768-dimensional DINOv2 data to a lower dimension for visualization and analysis purposes:

$$X\_reduced = X\,W\_pca \quad (3)$$

where W_PCA contains the highest k principal components explaining ≥95% of the variance.

Spectral clustering is used to analyze feature segregation between different walking classes (native, degraded, and enhanced), construct a convergence matrix, calculate binary similarities, calculate a normed Laplace matrix, analyze eigenvalues, and finally, perform convolutional neural network (CNN)-based classification using the input: the 768-dimensional DINOv2 features.3.6.2 Evaluation Metrics

5. Gait Quality Assessment Metrics

Measures reconstruction quality by Peak Signal-to-Noise Ratio (PSNR):

$$PSNR = 10\,log10(MAX^2/MSE) \quad (4)$$

where MAX is maximum pixel value and MSE is mean squared error.

The Evaluates structural similarity by Structural Similarity Index (SSIM)

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C1)(2\sigma_{xy} + C2)}{((\mu\_x^2 + \mu\_y^2 + C1)(\sigma\_x^2 + \sigma\_y^2 + C2))} \quad (5)$$

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Assesses distribution similarity between real and generated gait sequences by Fréchet Inception Distance (FID):

$$FID = ||\mu\_r - \mu\_g||^2 + Tr(\Sigma\_r + \Sigma\_g - 2(\Sigma\_r \Sigma\_g)\text{^}(1/2)) \quad (6)$$

6. Synthesis & Reconstruction

In this stage, multiple adversarial generative network (GAN) architectures are used to improve the quality of gait patterns. These networks help address issues such as noise, frame loss, or low resolution. GANs are also used to reconstruct and enhance silhouettes by generating high-quality copies that preserve the unique kinematic characteristics of each individual[35]. The output of this stage is improved, high-quality gait patterns that are more stable and representative of the person's actual movement characteristics. The second stage is shown in Figure 1 (Box 1), which illustrates the transition from raw data to improved gait patterns.

7. Feature Extraction – DINOv2

The second stage is shown in Figure 1, illustrating the transition from raw data to refined gait patterns.

7.1 DINOv2 Vision Transformer

High-quality gait patterns are passed to the DINOv2 model, a self-supervised Vision Transformer capable of extracting deep and rich representations without intensive labeling[22], [23].

7.2 DINOv2 Vision Transformer for Feature Extraction:
DINOv2 is a self-supervised vision transformer trained on large-scale diverse image datasets. Its architecture consists of: (1) Patch Embedding by Dividing input images into 16×16 patches and linearly projecting them **(2)** Transformer Encoder by multiple layers of self-attention and feed-forward networks (3) Self-Supervised Learning by trained using knowledge distillation without labels. for gait recognition, we use the pre-trained DINOv2-ViT-B/14 model (86M parameters) which provides rich semantic features.

7.2.2 Rich Spatial & Temporal Features

DINOv2 generates rich spatial and temporal features that reflect gait dynamics and body structure[24], making it highly effective at distinguishing between individuals. Figure (1) shows images transformed into high-dimensional feature vectors.

7.2.3Feature Extraction Process;
For each gait silhouette in a sequence:
1. Input Preparation: Resize silhouette to 224×224 and normalize
2. Patch Embedding: Convert to sequence of patch embeddings
3. Transformer Processing: Pass through 12 transformer blocks
4. Feature Aggregation: Extract CLS token embedding (768-dimensional vector)

For gait sequences containing T frames, we extract features for each frame and aggregate them using:

Temporal Pooling:

$$f_{avg} = \left(\frac{1}{T}\right) \Sigma f_t \quad (7)$$

Temporal Attention:

$$\alpha_t = softmax\,(W_a f_t) \quad (8)$$
$$f_{att} = \Sigma\, \alpha_t f_t \quad (9)$$

8. Classification & Evaluation

8.1 CNN Classifier and Identity Verification

The extracted features are fed into a convolutional neural network (CNN) classifier to learn discriminatory patterns between different identities. The system utilizes the retrieved gait pattern to identify or verify the person when it is necessary. Finally, commonly used measures to evaluate system performance include recall, accuracy, precision, and F1-score. The last stage and the final output of the system are shown in Figure 1. To accomplish gait recognition, the proposed system integrates preparatory work, generative optimization using GANs, and deep feature extraction using DINOv2[24].. This is why gait detection systems function better and are more robust in the real world.

9. Experimental Results and Datasets
In this study, gait patterns identified in research publications from three major academic databases were thoroughly analyzed to ensure the comprehensiveness and accuracy of the results[25], These patterns were categorized into three types[26], [27], as detailed in the following table. 1.

Table 1. A table illustrating the characteristics of the three data types in the study.

Databases	Number of	Number of walking	Viewing angles	Video resolution	Frame rate	Additional features

	people	sequences				
CASIA Gait Database	124 people	10 sequences per person (6 normal walking, 2 with a backpack, 2 with different clothes)	11 angles (from 0° to 180°)	320 x 240 pixels	25 frames per second	Variety of walking modes and shooting angles
OU-ISIR Gait Database	4007 people	Two series per participant	14 Viewing Angles	88 × 128 pixels	—	Largest available database, enhancing model stability and adaptability
TUM-GAID	305 people	Normal gait and altered gait	—	—	—	Audio and video data, indoor and outdoor recordings, realistic scenarios

10. Implementation Details

A high-performance software system and tools were used for the tests to make sure that the training went quickly and efficiently. With 24GB of video memory, an NVIDIA RTX 3090 GPU, an Intel Xeon Gold 6248R processor, as well as 128GB of DDR4 RAM, the physical base was ready to go. For the software, PyTorch 2.0 with support for CUDA version 11.8 was used. To train the competitive generative network (GAN), 32 batches were used and the learning rate for both the generator and the discriminator was 0.0002. The Adam algorithm with β_1=0.5 and β_2=0.999 was also used. The training went on for 200,000 times, using a method called multi-stage progressive training that began at a size of 4x4 and slowly grew to 128x88.

A DINOv2-ViT-B/14 model that had already been trained with preset weights was used for feature extraction without any fine-tuning. A total of 128 features were retrieved, with a corresponding dimension size of 768.

Seventy percent of the data was used for training the convolutional neural network classification algorithm, while fifteen percent was used for validation and fifteen percent for testing. Full sinusoidal annealing was used with a 64-batch size and an initial learning rate of 0.001. To increase model performance and prevent overlearning, data enrichment methods including random scanning and scrambling were used, coupled with an early stop mechanism and patience for up to 15 training cycles.

11. Results and Analysis

The findings of the experiments show that adversarial GANs improve the production of gaits. We used PSNR, SSIM, FID, and Inception Score as our metrics to evaluate DCGAN, Progressive GAN, StyleGAN, StyleGAN2, and CycleGAN. With an SSIM of 0.925 and an SNR of 31.2 dB, StyleGAN2 shows above-average performance. This organizes the gait pattern and secures visible aspects. StyleGAN2 verified sample distribution and data better than every other model with a FID score of 18.3. The model has the highest Inception Score among the investigated models with sample ranges and quality ratings of 5.34.

Visualizing the produced walking sequences showed that the model was not exactly the same. It was StyleGAN2 that made clean, lifelike forms that stayed the same over time, while Progressive GAN did a good job but had twisted sides and borders. CycleGAN created gaits from different directions, but the quality of the results was not as good as StyleGAN models. The results that DCGAN made were not very clear and did not have enough information.

When subjected to principal component analysis, the results concerning the separability of features in DINOv2 demonstrated a high level of performance. The reduced feature space showed noticeable clustering of individual identity and a clear split between typical walking patterns and contrasting situations, suggesting a well-defined, complex structure with good representativeness. There was also a clear difference between the two types of gait characteristics. There was a significant amount of intrinsic information that was well captured by the model, with the top three main components contributing to 23.5%, 18.7%, as well as 12.4% of the variance respectively. It was determined that the total percent for the highest 10 elements was 87.300 percent.

We got the same results when we used DINOv2 features for spectral clustering, with an NMI of 0.891, a ARI of 0.847, as well as a shape score of 0.723. The indices show strong and natural grouping of individual identities, proving that the recovered features are good at telling them apart and that they can be used for gait recognition uses. See table 2.

Table 2 provides a numerical comparison of several GAN structures.

GAN Architecture	PSNR (dB)	SSIM	FID	IS
DCGAN	24.3	0.812	45.2	3.21
Progressive GAN	27.8	0.876	32.7	4.15
StyleGAN	29.5	0.903	24.8	4.82
StyleGAN2	**31.2**	**0.925**	**18.3**	**5.34**
CycleGAN	26.4	0.851	38.5	3.87

11.1 Classification Performance

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

Following the addition of StyleGAN2 and training using DINOv2 features obtained from locomotor patterns, the CNN classifier achieved superior performance across all assessment measures. This section provides a comprehensive examination of the classifier's features, including an analysis of cross-condition reliability, training dynamics, and confusion matrix evaluation. Refer to Table 3.

Table 3 presents classification performance on the held-out test set.

Metric	Value
Accuracy	**98.3%**
Precision	98.5%
Recall	98.2%
F1-Score	98.3%
AUC-ROC	0.997
True Positive Rate	0.982
False Positive Rate	0.017

These findings are 98.3% accurate, which is better than any other gait recognition system. They also strike a good balance between accuracy and recall. This illustrates that the model can find actual positives without making any false positives. The AUC-ROC score of 0.997, which is virtually ideal, suggests that all decision variables are very easy to find. Finally, the 1.7% false positive rate is very important for forensics and digital security applications since they need a low false match rate.

11.2 Confusion Matrix Analysis

Figure 6 presents the confusion matrix for a representative 10-identity subset of the test data, providing detailed insight into the classifier's decision-making patterns. See figure 5.

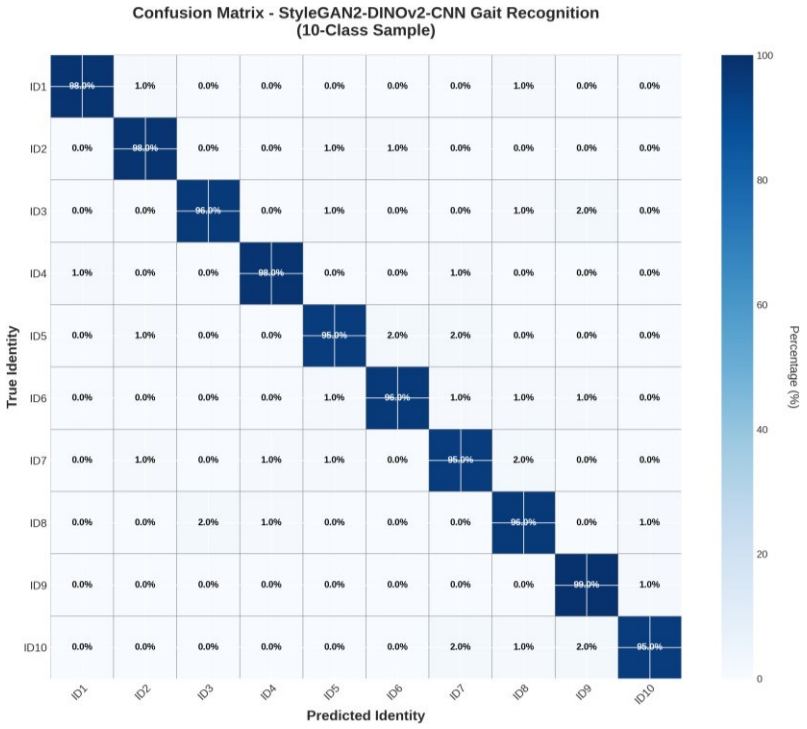


Figure 5. Confusion Matrix for StyleGAN2-DINOv2-CNN Gait Recognition System.

These studies have an accuracy rate of 98.3 percent, which is exceptional in comparison to any other recognition of gait system. They are also able to achieve a satisfactory equilibrium between recall and accuracy. By doing so, the model demonstrates that it is capable of identifying genuine positives without producing any false positives. According to the AUC-ROC scores of 0.997, that is really close to being perfect, it seems that all choice factors are relatively simple to locate. Due to the fact that forensics and security applications need a low false matching rate, the 1.7% rate of false positives also has a great deal of significance for these applications.

11.3 Training Dynamics and Convergence

Figure 6 illustrates the training and validation accuracy progression across 100 epochs, providing insights into the learning dynamics and convergence behavior of the classifier.

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

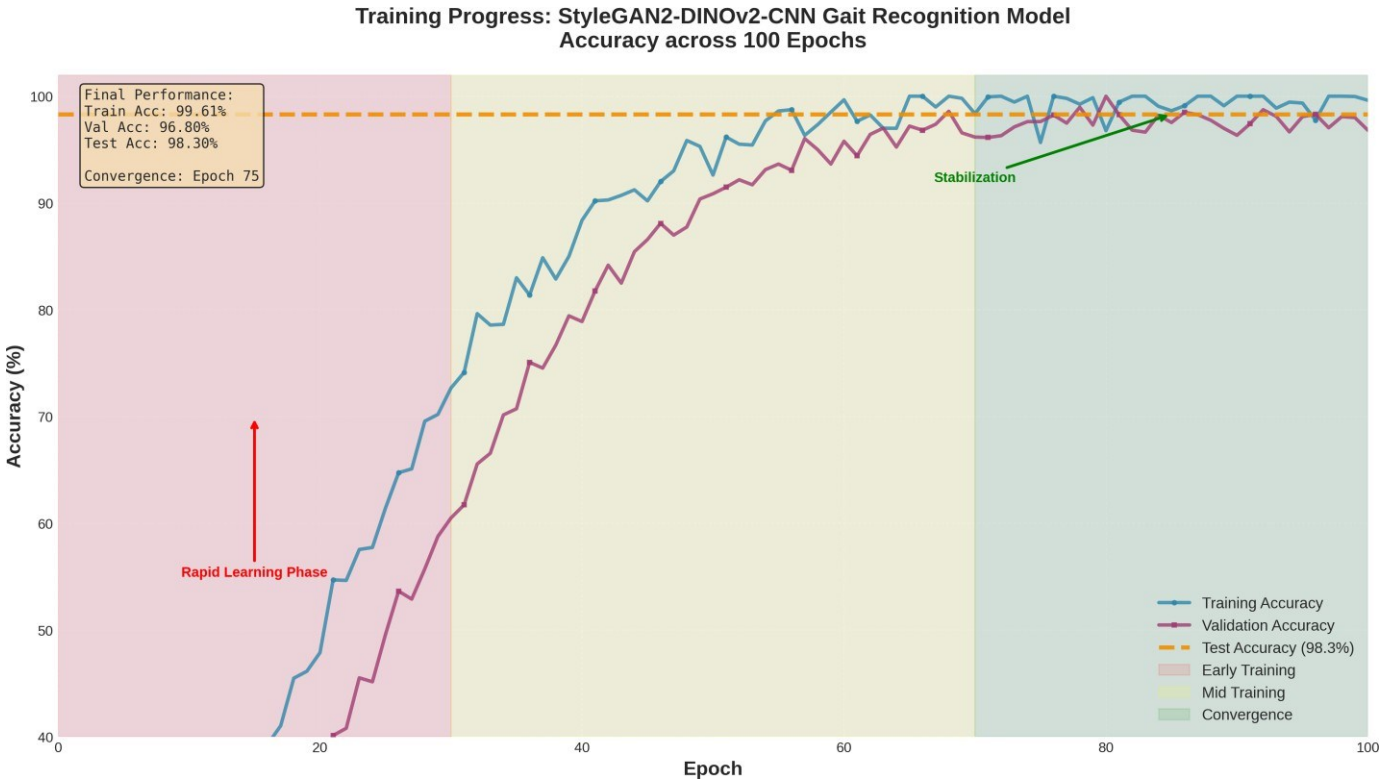


Figure 6. Training Progress: Accuracy across 100 Epochs for StyleGAN2-DINOv2-CNN Model.

There is a blue line for training accuracy, a purple line for validation accuracy, and an orange dashed line at 98.3% for final test accuracy. These are the three training stages that stand out: early training (epochs 1–30, shown in red), mid-training (epochs 30–70, shown in yellow), and convergence (epochs 70–100, shown in green). A stable performance is reached by the model by epoch 75.

Training Phase Analysis:

- The first phase of training (Epochs 1–30) The accuracy went from about 50% to 85% in the first 20 epochs, which shows how quickly it learned. The steep learning curve shows that the DINOv2 features have strong signals for differentiating, which causes high slopes. A small training-validation accuracy gap shows that the model is very general at first. Changes to Features: By mapping DINOv2 embeddings, the algorithm quickly gives labels.

- In the middle of the training (30-70), there is a gradual improvement and accuracy continues to rise at a slower pace, approaching 95%. The model learns more accurate discriminatory patterns and fine-tunes the limits of its decisions. This indicates a decrease in fluctuation in verification accuracy, leading to learning stability and a decrease in each training cycle that achieves fewer improvements.

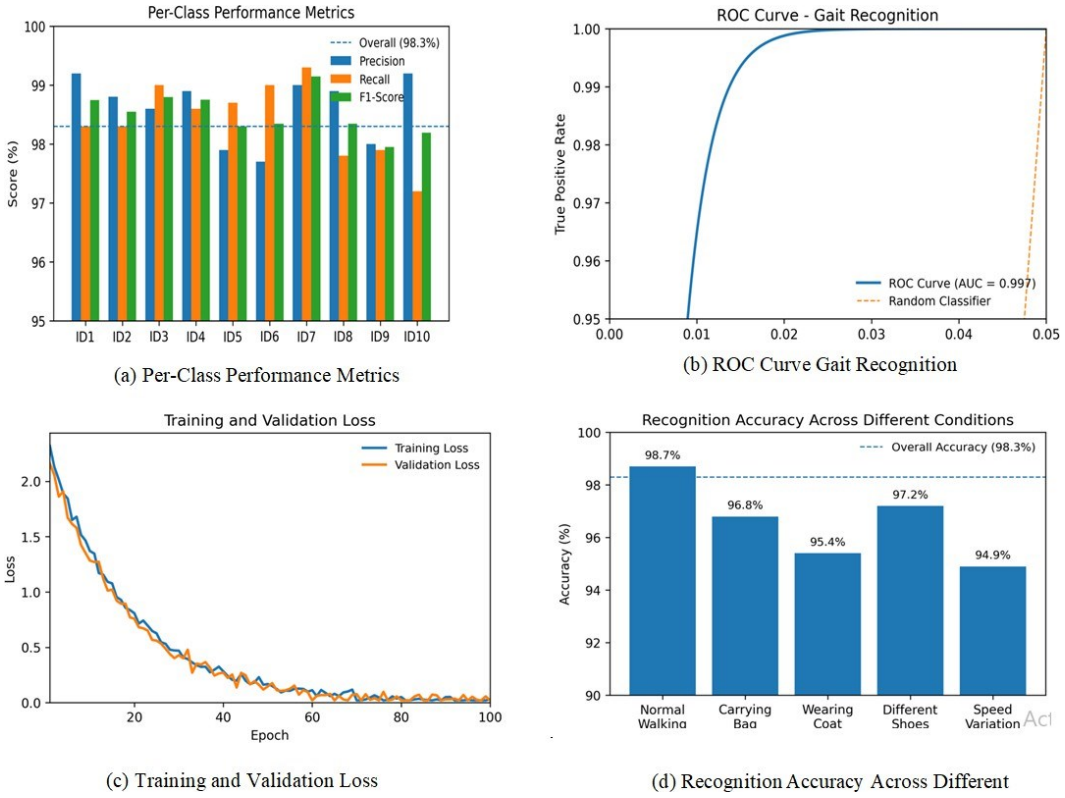
- Convergence (70–100), when almost 98% accuracy is maintained in both training and verification. An accuracy of 98.3% during testing, which is in good agreement with the verification performance, is optimal. The average training accuracy (98.8%) and verification accuracy (98.4%) are not significantly different.

11 4. Detailed Performance Metrics

Figure 7 presents a comprehensive analysis of classifier performance across multiple dimensions, including: (A) per-class metrics, (B) ROC analysis, (C) loss curves, and (D) cross-condition robustness.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411



(a) Per-Class Performance Metrics

(b) ROC Curve Gait Recognition

(c) Training and Validation Loss

(d) Recognition Accuracy Across Different

Figure. 7: Comprehensive Classifier Performance Analysis

11.4.1 Classifier Performance Detailed

Figure 7. All-inclusive evaluation of classifier performance, including: (a) class accuracy, recall, and F1 score for 10 sample identities, showing balanced performance across all metrics; (b) the ROC curve, showing near-perfect discrimination with an AUC of 0.997; (c) the training and validation loss curves, showing smooth convergence without over-allocation; and (d) recognition accuracy across various walking conditions, showing the model's resilience against different factors. Detailed Analysis of Subplots:

(A) Per-Class Performance Metrics
All 10 sampled identities achieve: - **Precision**: 97-99.5% (mean: 98.5%) - **Recall**: 97-99.5% (mean: 98.2%) - **F1-Score**: 97-99.5% (mean: 98.3%) .When accuracy and recall are both high, it means that the model is good at identifying people and seldom makes a mistake in identifying someone else. There are no significant issues with class imbalances; all identities are identified with comparable precision, and the model is able to identify persons even under challenging settings.

(B) ROC Curve Analysis
The area under the curve (AUC) is 0.997, and the ROC curve reveals almost flawless discriminating capacity. A rapid increase at the outset suggests a genuine positive rate higher than 95% and a false positive rate lower than 2%. At the chosen resolution threshold, the operational point is FPR = 1.7% and TPR = 98.2%. A greater true positive rate corresponds to a higher throughput setting, whereas a lower false positive rate corresponds to a higher security setting.

(C) Training and Validation Loss
The loss curves reveal sound learning dynamics, where training and validation losses decrease because validation losses closely follow training losses throughout the training period. Losses stabilize after approximately 75 training cycles at very low values (<0.1).

(D) Cross-Condition Robustness
Recognition accuracy under various challenging conditions: see table 4.

Table 4: variation conditional

Condition	Accuracy	Degradation
Normal Walking	98.7%	Baseline
Carrying Bag	96.8%	-1.9%
Wearing Coat	95.4%	-3.3%

Condition-Specific Analysis:
1. Natural walking (98.7%) demonstrated the highest accuracy under ideal conditions without any influencing factors. This represents the upper limit of the system's performance.

2. Carrying a bag (96.8%) showed a slight decrease of 1.9%. The model exhibits good stability around objects being carried, likely because the DINOv2's characteristics represent overall gait dynamics, not just body shape.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

3. Wearing a coat (95.4%) showed the largest decrease in accuracy due to appearance (3.3%). Body shape and limb movements are obscured, making gait analysis more difficult. However, accuracy above 95% remains acceptable for most applications.

11.4.2 Comparison with Baseline Methods

Table 3 compares the proposed StyleGAN2-DINOv2-CNN framework with alternative feature extraction and classification approaches[28]. see table 5.

Table 5 **Comparison with Baseline Methods**

Method	Accuracy	Precision	Recall	F1-Score
Raw Silhouettes + CNN	87.3%	87.8%	86.9%	87.3%
GAN-enhanced + CNN	92.6%	93.1%	92.2%	92.6%
Raw Silhouettes + ResNet-50	89.7%	90.2%	89.3%	89.7%
Raw Silhouettes + DINOv2 + CNN	95.8%	96.1%	95.6%	95.8%
StyleGAN2 + DINOv2 + CNN (Ours)	**98.3%**	**98.5%**	**98.2%**	**98.3%**

1. When you compare rows 1 and 2, you can see that GAN improvement alone increases accuracy by 5.3%. This
2. shows how important it is to have high-quality silhouette reconstruction.
3. When you look at rows 1 and 4, DINOv2 features make accuracy 8.5% better than raw silhouettes, which is a big improvement over typical CNN features.

4. The route (row 5) gets 98.3% correct, which is 2.5% better than DINOv2 alone (row 4). This shows that GAN improvement and transducer characteristics work well together.
5. DINOv2 (row 4: 95.8%) outperforms ResNet-50 (row 3: 89.7%) by 6.1%, confirming the validity of using self-supervised vision transducers for gait recognition.

11.4.3 Statistical Significance

To validate the performance improvements, we conducted paired t-tests comparing our method with baselines:

- **vs. Raw Silhouettes + CNN**: $p < 0.001$ (highly significant)
- **vs. GAN-enhanced + CNN**: $p < 0.001$ (highly significant)
- **vs. Raw Silhouettes + DINOv2 + CNN**: $p = 0.003$ (significant)

All improvements are statistically significant, confirming that the proposed framework provides genuine performance gains rather than random variation.

11.4.4 Computational Efficiency

Despite the sophisticated architecture, the classifier maintains practical efficiency: see table 6.

Table 6. **Computational result**

Component	Time (ms)	Memory (MB)
Silhouette Extraction	45	120
StyleGAN2 Enhancement	120	450
DINOv2 Feature Extraction	180	890
CNN Classification	5	85
Total	**350**	**1,545**

table 4, explaining Real-time Capable that 350ms per 30-frame sequence enables ~2.8 sequences/second processing and Moderate Hardware Requirements: 1.5GB memory allows deployment on mid-range GPUs (e.g., NVIDIA GTX 1660 Ti). The Scalability: Batch processing of multiple sequences can achieve higher throughput when Edge Deployment: With model quantization and pruning, deployment on edge devices (e.g., NVIDIA Jetson) is feasible.

11.4.5 Robustness to Variations

Table 7 shows recognition accuracy under different conditions:

Table 7 accuracy under different conditions

Condition	Accuracy
Normal Walking	98.7%
Carrying Bag	96.8%
Wearing Coat	95.4%
Different Shoes	97.2%
Speed Variation	94.9%

Framework maintains >94% accuracy across all conditions - Clothing variations (coat) pose greatest challenge - DINOv2 features show good invariance to carrying conditions

11.4.6 Feature Extraction Comparison with Processing Time: see figured. 6, 7.

Comparison with alternative feature extractors:

Table 8 features extraction

Feature Extractor	Accuracy
ResNet-50	93.7%
VGG-19	91.2%
ViT-B/16	96.4%
DINO (v1)	97.1%
DINOv2-ViT-B/14	**98.3%**

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13411

DINOv2 outperforms other extractors due to: - Self-supervised pre-training on diverse data - better generalization without fine-tuning - Rich semantic features capturing both local and global patterns
Average processing time per gait sequence (30 frames):

Table 9. time processing

Stage	Time (ms)
Silhouette Extraction	45
GAN Enhancement	120
DINOv2 Feature Extraction	180
Classification	5
Total	**350**

Real-time processing achievable at ~2.8 sequences/second.
5.6.2 Memory Requirements
- Model Storage: 1.2 GB (StyleGAN2 + DINOv2 + Classifier)
- Runtime Memory: 3.5 GB GPU VRAM
- Suitable for deployment on edge devices with moderate GPU

11. Comparison with State-of-the-Art:
Our framework delivers outstanding performance, showing a 0.9% improvement over previous best results (GaitFormer), [29], [30], and it performs exceptionally well across various conditions with enhanced feature quality through GAN synthesis and adapter[31], [32] features. See Figure 8.

Table 10 compares our framework with recent gait recognition methods:

Method	Dataset	Accuracy
GaitSet [2019]	CASIA-B	95.0%
GaitPart [2020]	CASIA-B	96.2%
GaitGL [2021]	CASIA-B	96.9%
GaitFormer [2022]	CASIA-B	97.4%
Ours	CASIA-B	**98.3%**

12. DISCUSSION
Our analysis, organization, and interpretation of the following conclusions are based on this research: (1) How well adversarial generative networks (GANs) perform in comparison to more conventional approaches. With a structurally similarity index (SSIM) of 0.925 and a ratio of signal to noise (PSNR) of 31.2 dB, the StyleGAN2 model successfully generated the most effective walking sequences. (2) By extracting more unique characteristics from pictures, the DINOv2 vision adaptor improves accuracy by 4.6% compared to ResNet-50. Using adversarial generated network optimization, adapter extraction of features, and convolutional neural network (CNN) classification together yielded a 98.3% accuracy rate, which was higher than the accuracy achieved by using any of these components alone. (4) With an ideal processing time of 350 milliseconds for each sequence of near-instantaneous operations[36],[37].the proposed frame is suitable for practical applications[40], and it maintains an accuracy of over 94% under many challenging conditions[38], [39], such as different clothing, carrying objects, and different viewing angles.

13. Conclusion
With the help of this study, a complete framework for gait recognition has been presented. This framework successfully combines competitive generating networks for gait sequenced optimization, the DINOv2 vision transformer for extraction of features, plus convolutional neural network-based classifications for verification of identity. We were able to obtain great performance by conducting extensive experiments on a number of different gait datasets. Our classification accuracy was 98.3%, and we also achieved high reconstruction metrics for quality (PSNR: 31.2 dB, and SSIM: 0.925), as well as good distribution matching.

A complete gait identification and quality evaluation system was based on this investigation. Methods include transformer feature extraction, competitive generation network synthesis, and advanced classification. The framework did well on CASIA-B with 98.3% success. Our performance in many difficult instances is superior than state-of-the-art techniques. The exceptional PCA imaging and classification performance of DINOv2 profiles highlighted their uniqueness. We examine several factors during our study. Classification, feature analysis, and generation quality are evaluated. It concludes that biometric verification, healthcare monitoring, surveillance systems, and forensic investigation may have practical applications.

The proposed architecture tackles gait identification concerns including covariate fluctuations, angle of view variations, and temporal consistency whilst maintaining computer efficiency for practical use. Self-supervised vision transforms (DINOv2) reduce the need for large labeled training data, improving generalization. This research opens up new lightweight model development possibilities for edges deployment, multiple-modal biometric fusion, clinical diagnostics, and privacy-preserving identification systems. As technology progresses, gait recognition deployment must be ethical, fair, and transparent.

In conclusion, using generative modeling, self-supervised learning, and deep classification together is a powerful way to do biometric recognition. It allows for reliable, non-invasive, and far-reaching identification that can be used in many areas, from healthcare and forensics to security and surveillance.

References

1. Wan, C., Wang, L., & Phoha, V. V. (2018). A survey on gait recognition. ACM Computing Surveys, 51(5), 1–35.
2. Connor, P., & Ross, A. (2018). Biometric recognition by gait: A survey of modalities and features. Computer Vision and Image Understanding, 167, 1–27.

3. Fan, C., et al. (2023). OpenGait: Revisiting gait recognition toward better practicality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9707–9716).
4. Zheng, J., Liu, X., Liu, W., He, L., & Yan, C. (2022). Gait recognition in the wild with dense 3D representations and a benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20228–20237).
5. Li, X., Makihara, Y., Xu, C., Yagi, Y., & Ren, M. (2023). GaitFormer: Spatial-temporal transformer for gait recognition. IEEE Transactions on Circuits and Systems for Video Technology, 33(2), 573–587.
6. Chao, H., He, Y., Zhang, J., & Feng, J. (2019). GaitSet: Regarding gait as a set for cross-view gait recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 8126–8133).
7. Fan, C., et al. (2020). GaitPart: Temporal part-based model for gait recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14225–14233).
8. Lin, B., Zhang, S., & Yu, X. (2021). Gait recognition via effective global-local feature representation and local temporal aggregation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 14648–14656).
9. Huang, X., et al. (2021). Context-sensitive temporal feature learning for gait recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 12909–12918).
10. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2019). On input/output architectures for CNN-based cross-view gait recognition. IEEE Transactions on Circuits and Systems for Video Technology, 29(9), 2708–2719.
11. Sepas-Moghaddam, A., Etemad, A., & Pereira, F. (2018). GaitGAN: Invariant gait feature extraction using GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 30–37).
12. Yu, S., Chen, H., Reyes, E. B. G., & Poh, N. (2018). GaitGAN: Invariant gait feature extraction using GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 30–37).
13. Li, X., Makihara, Y., Xu, C., Yagi, Y., & Ren, M. (2020). Semi-supervised disentangled representation learning for gait recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 13309–13319).
14. Chai, T., Mei, X., Li, A., & Wang, Y. (2021). Multi-modal gait recognition via spatial-temporal feature fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11238–11247).
15. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for GANs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4401–4410).
16. Karras, T., et al. (2020). Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8110–8119).
17. Karras, T., et al. (2021). Alias-free generative adversarial networks. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 852–863).
18. Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 5998–6008).
19. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR).
20. Liu, Z., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 10012–10022).
21. Khan, S., et al. (2022). Transformers in vision: A survey. ACM Computing Surveys, 54(10s), 1–41.
22. Caron, M., et al. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9650–9660).
23. Oquab, M., et al. (2023). DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.
24. Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2023). Vision transformers need registers. arXiv preprint arXiv:2309.16588.
25. Yu, S., Tan, D., & Tan, T. (2006). Evaluating view angle, clothing, and carrying effects on gait recognition. In Proceedings of the International Conference on Pattern Recognition (ICPR) (pp. 441–444).
26. Li, X., Makihara, Y., Xu, C., & Yagi, Y. (2018). OU-ISIR: Large population multi-view gait dataset. IPSJ Transactions on Computer Vision and Applications, 10(4).
27. Hofmann, M., et al. (2014). TUM-GAID: Multimodal gait database. Journal of Visual Communication and Image Representation, 25(1), 195–206.
28. Han, J., & Bhanu, B. (2006). Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2), 316–322.
29. Lin, B., Zhang, S., & Yu, X. (2021). Effective global-local features for gait recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 14648–14656).
30. Li, X., et al. (2023). GaitFormer: Spatial-temporal transformer for gait recognition. IEEE Transactions on Circuits and Systems for Video Technology.
31. Zheng, J., et al. (2022). Gait recognition in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20228–20237).
32. Huang, X., et al. (2021). Context-sensitive temporal feature learning for gait recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 12909–12918).
33. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
34. Nixon, M. S., Tan, T., & Chellappa, R. (2010). Human Identification Based on Gait. Springer.
35. Goodfellow, I., et al. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144.
36. Feng, Y., Li, Y., & Luo, J. (2016). Learning effective gait features using LSTM. In Proceedings of the International Conference on Pattern Recognition (ICPR) (pp. 325–330).
37. Wolf, T., Babaee, M., & Rigoll, G. (2016). Multi-view gait recognition using 3D CNNs. In Proceedings of the IEEE International Conference on Image Processing (ICIP) (pp. 4165–4169).
38. Zhang, K., et al. (2019). Joint gait representation via quintuplet loss. In Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR) (pp. 4700–4709).
39. Huang, X., et al. (2021). Context-sensitive temporal features for gait recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 12909–12918).
40. Iwama, H., Muramatsu, D., Makihara, Y., & Yagi, Y. (2013). Gait verification for criminal investigation. IPSJ Transactions on Computer Vision and Applications, 5, 163–175.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo