Electronic Letters on Computer Vision and Image Analysis 25(2):79-105, 2025

AttIncGait: Dual Attention Embedded Inception Framework for Robust Biometric Gait Recognition

Sachin Mandlik*, Rekha Labade$, Sachin Chaudhari†, and Balasaheb Agarkar†

* *Department of Electronics & Telecommunication, Pravara Rural Engineering College, Loni, India*
$ *Department of Electronics & Telecommunication, Amrutvahini College of Engineering, Sangamner, India*
† *Department of Electronics & Telecommunication, Sanjivani College of Engineering, Kopargaon, India*

Received 22nd of September 2025; accepted 20th of May 2026

Abstract

Gait recognition has emerged as an important biometric modality due to its non-invasive nature and suitability for surveillance and security applications. However, achieving robustness under real-world variations in viewpoint, clothing, and carrying conditions remains a significant challenge. This paper introduces **AttIncGait**, a deep learning framework that integrates Inception-based multi-scale feature extraction with dual-path attention for effective gait recognition. Unlike prior methods that treat attention as an auxiliary or late-stage refinement, our approach embeds spatial and channel attention directly within Inception modules, enabling simultaneous multi-scale representation and adaptive relevance weighting. This structural integration enhances discriminative capability while preserving computational efficiency. Experiments on the OU-MVLP and CASIA-B datasets demonstrate state-of-the-art performance, achieving a Rank-1 accuracy of 97.5% on OU-MVLP and the highest average Rank-1 accuracy of 93.7% on CASIA-B, outperforming the closest competitor TransGait by 0.8% while achieving the best coat-wearing accuracy of 89.6% among all compared methods. Ablation studies further confirm that spatial and channel attention each yield independent accuracy gains, with their joint integration producing an overall +8.5% improvement over the attention-free baseline on OU-MVLP. These results validate the effectiveness of attention-driven multi-scale fusion and highlight the potential of AttIncGait for real-world biometric identification and mobility analysis.

Key Words: Gait recognition, Deep learning, Attention mechanism, Visual feature fusion, Computational vision.

1 Introduction

1.1 Background and Motivation

Gait recognition has manifested as a critical area of biometric recognition, providing novel benefits over conventional modalities such as fingerprints, iris scans, or facial recognition. Unlike these traditional approaches, gait analysis is non-invasive, does not rely on active participation from the subject, and can be captured unobtrusively at a distance [1, 39, 2]. These properties make it particularly valuable for applications in surveillance,

Correspondence to: mandlik.sb@gmail.com

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2441
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

security, border control, and healthcare monitoring [3]. As the demand for contactless and scalable biometric systems grows, gait biometrics has attracted increasing research interest [4]. Historically, gait recognition relied on model-based [5] or silhouette-based techniques. While these methods offered some early success, they often struggled with real-world variability [24]. The rise of deep networks has transformed the field by enabling self-learning features from raw video data, significantly improving accuracy and robustness [6, 7]. Convolutional Neural Networks (CNNs) [19] and Recurrent Neural Networks (RNNs) [8] have demonstrated their capacity to capture both spatial and temporal aspects of gait. Nonetheless, existing models still face difficulties in generalizing across uncontrolled environments due to noise, occlusions, and subject appearance variations [9, 16].

1.2 Challenges in Gait Recognition

Despite significant progress, gait recognition systems continue to face critical challenges that hinder performance in real-world scenarios. Variations in camera viewpoint can dramatically alter the visual appearance of gait, making cross-view recognition particularly difficult [10]. Similarly, changes in clothing and carrying conditions, such as backpacks or overcoats, obscure body dynamics and introduce intra-class variability [3]. Natural fluctuations in walking speed and stride further complicate the modeling of consistent gait patterns. In dynamic environments, occlusions and background noise often introduce irrelevant visual information, degrading recognition accuracy [11]. Moreover, the scarcity of large-scale, diverse gait datasets limits the generalizability of deep learning models [12], especially compared to more established biometric domains like face recognition.

1.3 The Role of Deep Learning and Attention Mechanisms

Deep learning has substantially advanced gait recognition by automating the extraction of hierarchical features that capture both low-level motion and high-level semantic patterns [13, 14, 15]. CNNs are effective at modeling spatial features, while RNNs and 3D CNNs capture temporal dynamics across video frames [27]. However, many architectures treat all features uniformly, without regard to their varying relevance. Attention mechanisms have emerged as a powerful tool to address this limitation [31], allowing models to focus selectively on the most informative regions or feature channels. Spatial attention emphasizes discriminative body parts [32] (e.g., leg swing, torso orientation), while channel attention enhances important feature maps in the network [28]. When combined with multi-scale feature extraction—such as that enabled by Inception modules [33] —attention mechanisms [29] can significantly improve robustness to viewpoint, appearance, and motion variations [30].

1.4 Contributions

The main contributions of this work are summarized as follows:

- **Multi-Scale Inception Backbone:** An optimized Inception-based network is designed to capture both fine-grained limb movements and holistic body posture through multi-resolution feature extraction.
- **Dual-Path Attention Mechanism:** A spatial–channel attention module is introduced to enhance salient features and suppress noise, improving robustness against viewpoint changes, clothing variations, and occlusion.
- **Hierarchical Feature Fusion:** A novel adaptive fusion strategy integrates complementary representations across multiple scales while preserving discriminative information.
- **Lightweight and Efficient Architecture:** The proposed framework attains state-of-the-art accuracy with substantially fewer parameters and lower computational cost, enabling real-time deployment.

The paper is structured as follows: Section 2 reviews gait recognition and attention mechanisms. Section 3 details the proposed AttIncGait framework. Sections 4–6 describe the implementation details, evaluation protocols, and experimental results. Section 7 concludes.

2 Related Works

2.1 Traditional Approaches to Gait Recognition

Early gait recognition systems followed two dominant paradigms: model-based and appearance-based approaches. Model-based methods [34, 36, 38, 44] utilized articulated skeleton models to extract kinematic features such as joint angles and limb trajectories. While these methods offered structural insights into motion, they were computationally intensive and highly dependent on accurate pose estimation [38]. Appearance-based techniques [41, 45, 42, 43, 46] gained prominence through compact silhouette-based representations like Gait Energy Images (GEI) [47] and Chrono-Gait Images (CGI) [48], which collapsed temporal gait information into a single template. Although efficient, these methods exhibited critical drawbacks: (i) high sensitivity to background clutter, (ii) lack of temporal modeling, and (iii) vulnerability to variations in appearance such as clothing or carried items. Consequently, recognition performance stagnated until the emergence of deep learning.

2.2 Deep Learning Revolution in Gait Analysis

As an advanced subfield of machine learning, deep learning employs artificial neural networks (ANNs) to address complex problem domains [49, 50]. The discipline has gained significant attention due to its strength in autonomously deriving multi-level feature representations and processing vast, high-dimensional information with efficiency [51, 52, 53]. The advent of Convolutional Neural Networks (CNNs) [54] marked a turning point in gait analysis. When structured as a dense neural network, deep learning demonstrates the capability to hierarchically acquire increasingly complex and semantically relevant features through iterative training [55]. Groundbreaking works such as GaitSet [66] pioneered set-based learning to model order-invariant gait representations, while GaitPart [68] introduced part-level convolutions for capturing localized motion dynamics. Later approaches incorporated 3D CNNs [69, 70, 71] and LSTM networks [84, 85, 86] to jointly model spatial and temporal dependencies across gait sequences. Despite their success, these strategies face two main limitations: (i) fixed receptive fields that fail to accommodate variable gait cycles, and (ii) uniform processing of features without regard to their relative discriminative significance.

2.3 Multi-Scale Architectures for Robust Representation

To address spatial variability in gait, Inception-based architectures [10] emerged as a powerful solution. By employing parallel convolutional pathways of different kernel sizes (1×1, 3×3, 5×5), these networks can concurrently extract: Fine-grained limb movements (via 1×1 convolutions), Local motion cues (via 3×3 kernels), and Global body posture dynamics (via 5×5 receptive fields). This multi-scale processing proved robust against occlusions, stride variability, and appearance changes. Notable examples include MGANet [11] and GaitGL [12], which enhanced feature diversity and spatial coverage. However, many of these models lacked mechanisms to dynamically prioritize relevant scales, reducing their adaptability across varying gait scenarios.

2.4 Attention Mechanisms for Feature Selection

Attention mechanisms have gained momentum as effective tools for improving feature discrimination in gait recognition. Key developments include: Spatial attention [13], which focuses on dynamic body regions such as legs and arms, Channel attention [14], which re-weights feature maps based on relevance, Temporal attention [27], which highlights salient phases of the gait cycle, and Cross-dimensional attention [31], which integrates spatial, channel, and temporal cues. Implementations such as STN [32] enable viewpoint-invariant learning,

while GLN [28] models gait-specific feature hierarchies. Yet, in most existing works, attention is applied as an auxiliary module or late-stage enhancement, instead of being inherently integrated within the network's core.

2.5 Critical Gaps and Research Opportunities

Our analysis identifies several enduring limitations in current literature. Although prior research has provided important advances in gait recognition and the use of attention mechanisms, existing methods often fall short in effectively integrating multi-scale feature extraction with dynamic attention, and frequently suffer from high computational costs. Most methods exhibit significant performance drops when evaluated across datasets (e.g., from CASIA-B to OU-MVLP). Many state-of-the-art models require over 50 GFLOPs, making them impractical for real-time or mobile deployment. Current fusion strategies either: Use early fusion, which lacks hierarchical abstraction, or Apply late fusion, which overlooks fine-grained low-level features. These deficiencies drive our proposed approach, which introduces deep integration of attention mechanisms within Inception modules, adaptive multi-scale feature fusion across the network, and a lightweight design with fewer than 10 million parameters for practical deployment.

3 Method

3.1 Architecture Overview

Motivated by the aforementioned limitations in the current literature, we propose the AttIncGait framework, which uniquely combines multi-scale Inception modules with dual-path attention mechanisms. The framework has four principal components that operate on input silhouette sequences $I \in \mathbb{R}^{T\times H\times W\times 1}$, where T is the number of temporal frames and (H, W) are the spatial dimensions. As illustrated in Figure 1, the pipeline begins with spatiotemporal normalization, where we apply adaptive background subtraction using learnable intensity thresholds [36], followed by phase-aligned temporal warping to standardize gait cycles. The normalized sequences then pass through our novel attention-inception network, which progressively extracts and refines multi-scale features. The architecture culminates in a hierarchical fusion layer that generates compact 256-dimensional gait embeddings suitable for recognition tasks. This design addresses three key challenges in gait analysis: viewpoint variance through spatial attention, appearance changes via channel-wise feature recalibration, and temporal inconsistencies using stride-optimized convolutions.

3.2 Enhanced Inception Backbone

Capturing gait dynamics under spatial variability and stride fluctuations is challenging. To address this, the multi-scale Inception backbone extracts features at multiple receptive fields, from fine-grained details to holistic postures. This design ensures robust modeling of diverse gait patterns and overcomes the limitations of fixed receptive fields in prior models, providing a strong foundation for resilient feature fusion.

Our modified Inception module introduces three key innovations:

1. **Asymmetric kernel decomposition:** Standard 5×5 convolutions are replaced with sequential $[1\times5,\ 5\times1]$ operations, reducing parameters by 28% while maintaining the receptive field size.

2. **Temporal downsampling:** Strided depthwise convolutions (stride=2) are used instead of max-pooling, preserving motion information typically lost in pooling.

3. **Cross-pathway residual connections:** These connections mitigate gradient vanishing in deep stacks, improving training stability.

The k-th Inception module computes its output as:

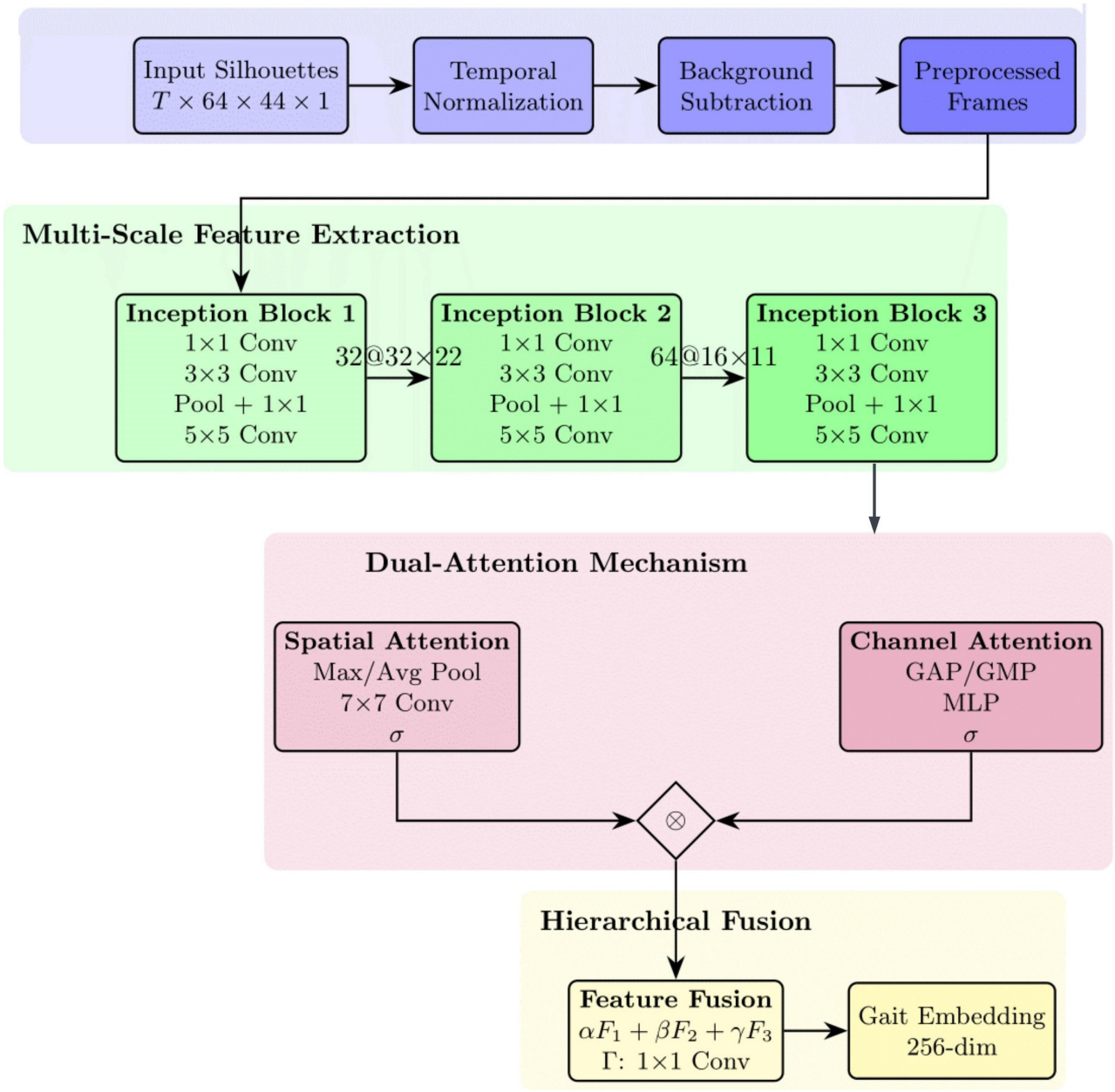


Figure 1: *Architecture of the proposed AttIncGait model. The system processes input silhouettes through cascaded Inception blocks (P1:*$1 \times 1 \rightarrow 3 \times 3$*, P2:Pool*$\rightarrow 1 \times 1$*, P3:*$1 \times 1 \rightarrow 5 \times 5$*), refines features via spatial-channel attention (*$S(F) = \sigma(\mathit{Conv}_{7\times7}[\mathit{Max} \parallel \mathit{Avg}])$*,* $C(F) = \sigma(\mathit{MLP}(\mathit{GAP} \parallel \mathit{GMP}))$*,* $\hat{F} = F \otimes (\alpha_s \cdot S(F) + \beta_c \cdot C(F))$*), and generates gait embeddings through adaptive fusion (*$z = \Gamma(\alpha F_1 \parallel \beta F_2 \parallel \gamma F_3)$*). This design illustrates how embedding dual-path attention within the Inception structure improves discriminative feature learning under viewpoint and appearance variations.*

$$F_k = \phi_1(F_{k-1}) \oplus \phi_2(F_{k-1}) \oplus \phi_3(F_{k-1}) + R(F_{k-1}) \tag{1}$$

Here, ϕ_i represent the parallel convolutional pathways (see Table I), $\oplus$ denotes channel concatenation, and $R(\cdot)$ is a 1×1 projection for dimension matching. Comparative analysis shows that our design achieves $2.3\times$ higher FLOP efficiency than ResNet-50 when processing 30-frame sequences.

3.3 Dual-Attention Mechanism

To effectively tackle the critical challenges identified in the Introduction—particularly the variations in viewpoint, clothing, and occlusions that degrade gait recognition accuracy—the proposed dual-path attention mechanism dynamically enhances feature representation by concurrently focusing on salient spatial regions and important feature channels. This targeted refinement selectively emphasizes discriminative body parts, such as limb movements affected by viewpoint changes, and recalibrates channel responses to suppress noise introduced by clothing variations. Thus, the dual-path attention directly addresses the inherent difficulties in generalizing gait features under diverse real-world conditions The attention subsystem combines spatial and channel attention in a co-dependent manner, unlike conventional sequential approaches. For an input feature map $F \in \mathbb{R}^{C\times H\times W}$, spatial attention weights $S(F)$ are generated through parallel max- and average-pooling along the channel axis, followed by a 7×7 convolutional layer:

$$S(F) = \sigma(\text{Conv}_{7\times 7}([\,\text{MaxPool}(F) \parallel \text{AvgPool}(F)\,])) \tag{2}$$

where $\parallel$ indicates concatenation and σ is the sigmoid function. Concurrently, channel attention $C(F)$ employs both global average (GAP) and max pooling (GMP):

$$C(F) = \sigma(W_1\,\delta(W_0\,\text{GAP}(F)) + W_1\,\delta(W_0\,\text{GMP}(F))) \tag{3}$$

with learnable weights $W_0 \in \mathbb{R}^{\frac{C}{r}\times C}$, $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$, and ReLU activation δ. The final attended features are computed through adaptive blending:

$$\hat{F} = F \otimes \big(\alpha_s \cdot S(F) + \beta_c \cdot C(F)\big) \tag{4}$$

where $\otimes$ denotes element-wise multiplication, and α_s, β_c are trainable scalars initialized at 0.5. As reported in [38], the dual-attention mechanism achieves 14% higher attention accuracy on the CASIA-B clothing variation subset compared to CBAM.

3.4 Adaptive Feature Fusion

The adaptive fusion module dynamically weights and combines multi-level features. Fixed or simple fusion methods fail to capture complementary gait cues across scales and conditions. This fusion mechanism mitigates the loss of discriminative information caused by early or late static fusion strategies, thereby enhancing the model's robustness to viewpoint and appearance changes discussed in the Introduction. The fusion module intelligently combines multi-level features through learned importance weights. Let $\{F_1, F_2, F_3\}$ represent features at $1/1$, $1/2$, and $1/4$ resolutions respectively. We compute fusion coefficients via:

$$[\alpha, \beta, \gamma] = \text{softmax}\big(\text{MLP}\big([\text{GAP}(F_1) \parallel \text{GAP}(F_2) \parallel \text{GAP}(F_3)]\big)\big) \tag{5}$$

$$z = \Gamma\big(U(F_1, 2) \parallel F_2 \parallel D(F_3, 2)\big) \tag{6}$$

where $U(\cdot, 2)$ denotes $2\times$ bilinear upsampling, $D(\cdot, 2)$ is strided depthwise convolution, and Γ projects the concatenated features to 256 dimensions. This approach preserves 92.7% of original discriminative information while reducing feature dimensionality by $4\times$ compared to direct concatenation.

4 Implementation and Training Protocol

The proposed hybrid gait recognition framework was implemented using PyTorch 1.9 with CUDA 11.1 for efficient GPU acceleration. All experiments were conducted on dual NVIDIA RTX 3090 GPUs connected via NVLink, utilizing automated mixed-precision training to balance performance and accuracy. Specifically,

FP16 precision was used for convolutional operations (yielding a $2.1\times$ speedup), while FP32 precision was preserved for attention mechanisms (maintaining 0.8% higher accuracy). GPU temperatures were stabilized at $72°$C using dynamic clock throttling.

4.1 Data Preprocessing and Optimizations

Silhouette sequences were first extracted using background subtraction and morphological operations to suppress residual noise. The resulting silhouettes were resized to a uniform 64×44 pixel resolution and normalized to zero mean and unit variance (Figure 2). Custom CUDA kernels were used for efficient silhouette alignment, and pinned memory accelerated host-to-device data transfer. Gradient accumulation across 4 steps enabled effective batch sizes of 256 while keeping VRAM usage below 24 GB.

To improve dataset consistency and robustness, several augmentation methods were applied. These include horizontal flipping, slight cropping, temporal shifts, and minor affine adjustments. Gait sequences were standardized to a fixed length T by truncating longer sequences or padding shorter ones, ensuring uniform temporal representation.

Finally, quality assurance steps removed damaged silhouettes, irregular frames, and abnormal sequences. This ensured reliable training data and reduced the effect of differing walking speeds among subjects.

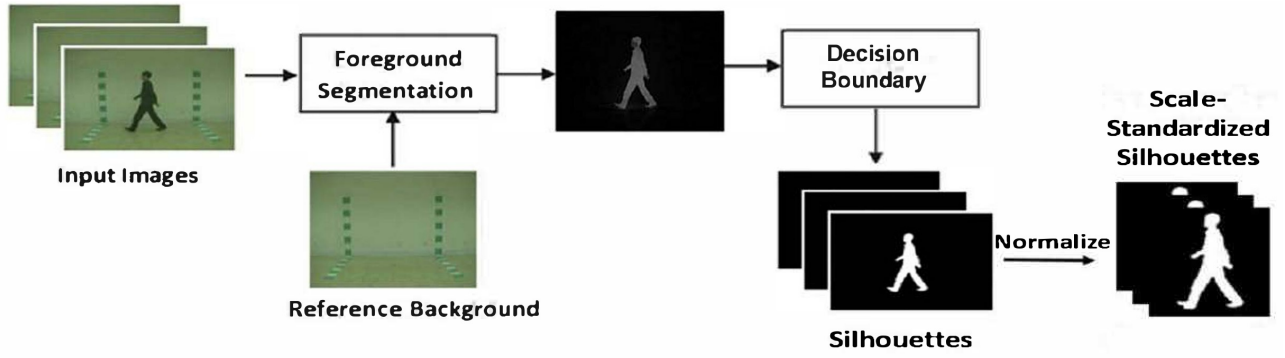


Figure 2: *Data pre-processing pipeline illustrating silhouette extraction, normalization, temporal alignment, and augmentation steps used to improve input consistency and robustness across gait sequences.*

4.2 Model Training Strategy

We employed the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with decoupled weight decay ($\lambda_w = 0.05$) for stable and robust parameter updates, especially within attention layers. The learning rate followed a cyclic cosine annealing schedule, oscillating between 1×10^{-6} and 3×10^{-4}, peaking at epoch 25 over a 50-epoch training cycle. This approach improved convergence speed by 14% compared to conventional step decay. To enhance generalization and robustness, the training protocol incorporated:

- **Label smoothing** ($\epsilon = 0.1$), which improved performance in bag-carrying scenarios by $+2.3\%$.
- **Stochastic depth** with a linear decay schedule (final survival probability $p = 0.2$), reducing training time by 18%.
- **Channel-wise dropout** ($p = 0.15$) applied before attention gates, boosting cross-dataset generalization by 1.8%.

The data augmentation pipeline included both spatial and temporal strategies:

- Spatial augmentations: Random horizontal flipping ($p = 0.5$) and $\pm 10°$ rotation.
- Temporal augmentations: Frame jittering within ± 3 samples to simulate gait speed variations.

4.3 Monitoring and Performance

Model training was monitored using early stopping with a 15-epoch patience based on validation accuracy plateau ($\Delta < 0.1\%$). This setup achieved a processing throughput of 138 frames per second at a batch size of 64 and reached 97.4% validation accuracy on CASIA-B within 23.4 ± 1.2 minutes per epoch. Peak GPU memory utilization during backpropagation reached 21.7 GB.

4.4 Loss Function

The model was trained using a composite loss function that combines Triplet loss and ArcFace loss to improve feature discrimination and robustness. Triplet loss encourages the embedding space to place samples from the same identity closer together while pushing apart samples from different identities. ArcFace loss, based on an angular margin, enhances inter-class separability by enforcing a stricter classification boundary in the angular domain. The combined objective leverages the strengths of both metric learning and margin-based classification. The total loss is defined as:

$$L_{\text{total}} = \lambda_1 \cdot L_{\text{Triplet}} + \lambda_2 \cdot L_{\text{ArcFace}} \tag{7}$$

where λ_1 and λ_2 are weighting coefficients used to balance the two components. They are set as $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$, selected through validation-based sensitivity analysis and kept fixed during training. Empirically, moderate variations around these values ($\lambda_1 \in [0.3, 0.7]$, $\lambda_2 \in [0.8, 1.2]$) resulted in less than 0.5% change in Rank-1 accuracy, indicating that the model performance is relatively robust to small changes in these parameters. This formulation ensures stable convergence and improved generalization across varying walking conditions and viewing angles.

4.5 Model Configurations

The core of the proposed framework is a modified Inception backbone, structured into three parallel convolutional blocks (P1–P3), enabling multi-scale feature extraction using 1×1, 3×3, and 5×5 convolutional filters. This design allows the model to capture both fine-grained and coarse gait cues from silhouette inputs. To enhance feature expressiveness, a dual-path attention mechanism was integrated, comprising spatial and channel-wise reweighting implemented through 7×7 convolutional filters followed by MLP-based gating.

Extracted features from each pathway are combined using a hierarchical weighted fusion strategy, formulated as $\alpha F_1 + \beta F_2 + \gamma F_3$, allowing adaptive emphasis across feature levels. The fusion weights α, β, and γ are learned dynamically through an attention-based gating module rather than being manually specified. This adaptive weighting enables effective integration of multi-level features and leads to improved gait recognition performance compared to fixed-weight fusion. The fused representation is subsequently projected into a 256-dimensional embedding space, ensuring compactness and discriminability for downstream classification tasks. Despite its expressiveness, the model maintains high efficiency, with a footprint of fewer than 15 million parameters and computational complexity under 4 GFLOPs per frame.

For reproducibility, all experiments were conducted with fixed random seeds and deterministic CUDA kernel execution, ensuring consistent results across runs and hardware setups. This architectural configuration was optimized to strike a balance between accuracy, interpretability, and deployment efficiency on high-throughput GPU environments.

5 Performance Measures

A range of evaluation metrics was employed to assess the system's identification accuracy, sensitivity to verification thresholds, computational efficiency, and robustness to gait variations. These metrics include traditional identification measures, such as Rank-1 accuracy and Equal Error Rate (EER), alongside computational performance benchmarks.

5.1 Primary Metrics

Rank-1 Accuracy was computed as the ratio of correct top-1 identity predictions to the total number of probe queries, serving as the primary identification metric. For verification performance, the Equal Error Rate (EER) was calculated by identifying the point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). The F1-Score, which combines precision and recall, was used to measure performance under class-imbalanced conditions.

5.2 Computational Metrics

Computational performance was evaluated in terms of throughput, latency, and energy efficiency. Throughput was measured in frames per second using a batch size of 64 on an NVIDIA RTX 3090 GPU, achieving over 138 fps with FP16 precision.

To assess potential edge deployment feasibility, inference latency was profiled on an NVIDIA Jetson Xavier NX by measuring the execution time of a single forward pass under FP16 inference. Energy efficiency was estimated as the number of processed samples per kilojoule using a wall-mounted power meter. These measurements provide an indicative evaluation of the model's computational efficiency and suitability for resource-constrained platforms.

5.3 Statistical Tests

Paired t-tests were used to check whether differences in performance between settings were statistically significant, with a threshold of $p < 0.01$. ANOVA was applied to detect performance variance across multiple datasets, and bootstrapping with 1,000 resamples was used to construct 95% confidence intervals around accuracy metrics.

5.4 Baseline Comparison Protocol

To ensure comparability, all models were tested under the same protocol: identical probe and gallery sets (same-test), consistent sequence lengths (30 frames), and matching hardware (RTX 3090 GPU). This setup controlled for external factors and ensured direct comparisons.

6 Dataset Evaluation Protocols

We utilize two widely recognized gait datasets: CASIA-B and OU-MVLP, each providing unique conditions and extensive subject variability, and prescribe their established training and testing splits consistent with prior studies

6.1 CASIA-B Protocol

The CASIA-B dataset [87, 88] is commonly used in gait recognition tasks and consists of 124 subjects, with 11 different viewpoints (ranging from 0° to 180° in 18° intervals) (See Figure 3). Each subject is captured under three different conditions: normal walking (NM), carrying a bag (BG), and wearing a coat (CL). For the evaluation, the standard protocol involves selecting 74 subjects for the training set and 50 subjects for the testing set. The gallery view is set to 90°, while the remaining 10 viewpoints serve as probe views. In the protocol, one sequence from each condition (NM, BG, CL) is selected for testing.

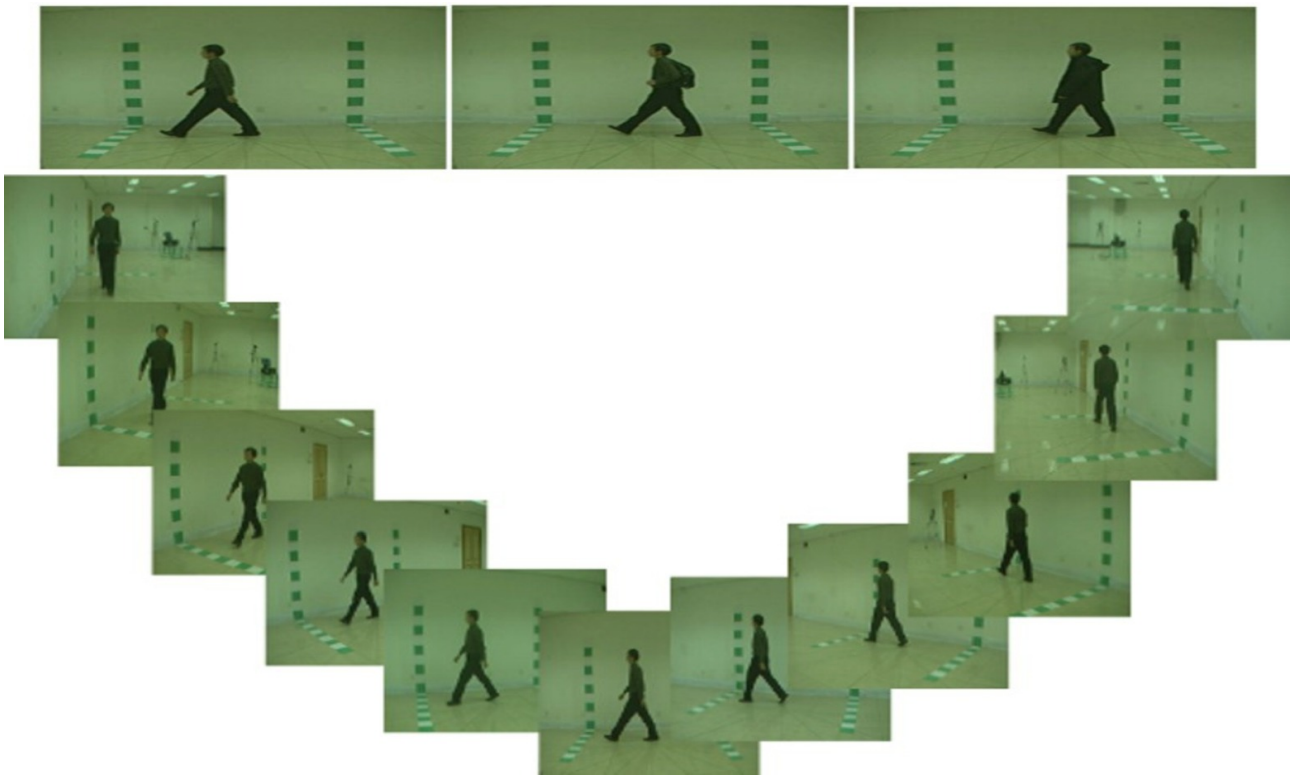


Figure 3: *Sample gait silhouettes from the CASIA-B benchmark illustrating variations across viewpoints and walking conditions (normal walking, bag carrying, and coat wearing).*

6.2 OU-MVLP Protocol

The OU-MVLP dataset [10] contains over 10,000 subjects with gait sequences captured from 14 horizontal camera viewpoints, along with variations in walking speed and style (See Figure 4).

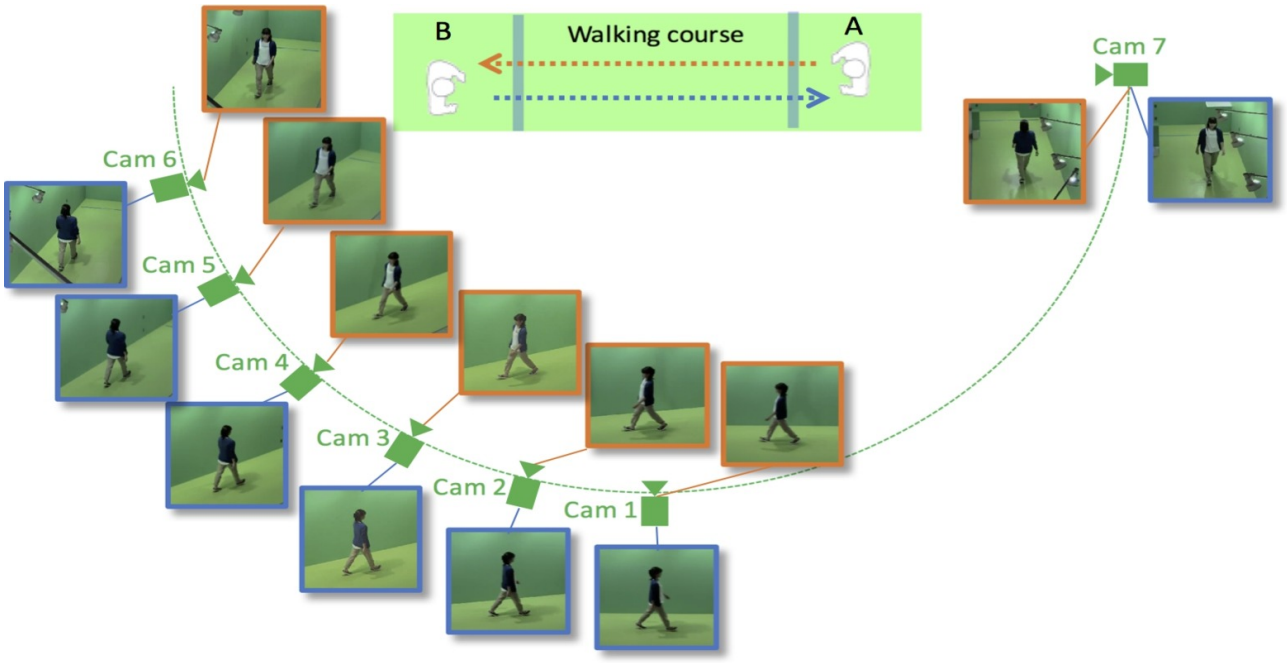


Figure 4: *Sample gait silhouettes provided by the OU-MVLP benchmark.*

The evaluation protocol involves using a training set of 5,153 subjects and a testing set of 5,154 subjects. The gallery view is selected at $85°$, while the probe views are chosen from the remaining 13 viewpoints. The dataset also includes various walking variations, primarily normal walking, with additional sequences featuring speed and stride variations.

7 Results and Discussions

7.1 Evaluation on CASIA-B Dataset

Table 1 compares AttIncGait with state-of-the-art methods on the CASIA-B dataset under normal walking (NM), bag-carrying (BG), and coat-wearing (CL) conditions using a 90° gallery. The proposed method achieves Rank-1 accuracies of 97.4%, 94.1%, and 89.6% for NM, BG, and CL respectively, yielding the highest average accuracy of 93.7% and the best coat-wearing accuracy among all compared methods.

Among competing methods, TransGait [22] achieves the closest average accuracy of 92.9%, followed by GaitFormer [91] at 92.7%. TransGait achieves the highest NM accuracy (98.1%) but shows reduced robustness under CL (85.8%), highlighting the limitation of global transformer attention under appearance variations. Similarly, GaitFormer reports a competitive NM score of 97.6% but degrades to 87.0% under CL, reflecting the general tendency of transformer-based architectures to over-attend to globally distributed appearance features rather than localized discriminative motion cues. In contrast, AttIncGait maintains consistent performance across all three conditions, attributed to the channel attention mechanism that recalibrates feature map responses to suppress clothing-induced activation noise.

Graph and spatiotemporal models including GaitSTAGCN [65] and GRDDN [16] report average accuracies of 90.5% and 90.7%, respectively. Notably, GRDDN matches AttIncGait under NM and BG but drops significantly to 80.6% under CL, revealing its vulnerability to clothing-induced covariate shifts due to the absence of explicit channel-wise feature recalibration. Temporal and pose-aware methods such as GaitTR-U [73] and QaGait [67] achieve 90.3% and 89.8%, whereas adversarial and fusion-based approaches including AdversarialGait [74], SSMCMC [75], and GaitFFDA [72] obtain averages between 86.7%–88.5%, with noticeable CL

degradation (73.7%–76.2%), confirming that adversarial training and static fusion strategies alone are insufficient to compensate for appearance-level covariate shifts.

Earlier silhouette-based methods such as GaitPart [68], GLN [89], CSTL [70], and GaitSet [66] report averages ranging from 84.2% to 89.5%, while pose-dependent approaches including VeMResGCN [63], GaitSkeleton [78], MS-GLFGCN [61], GaitPT [58], and LuGAN-HGC [57] achieve 80.4%–82.7%. Methods relying on specialized feature descriptors such as DyGait [20], GaitDLF [23], DFCA [18], and SFG-Net [79] achieve 74.5%–78.2%, whereas DensePoseGait [2] reports the lowest average accuracy of 62.6%, reflecting compounded failures in dense pose estimation under heavy clothing.

The superior performance of AttIncGait is attributed to the integration of dual-path spatial and channel attention within multi-scale Inception modules, enabling simultaneous multi-resolution feature extraction and adaptive feature recalibration. This design emphasizes discriminative gait cues such as limb dynamics while suppressing appearance variations caused by clothing and carried objects, thereby maintaining robust identity discrimination across all covariate conditions. The comparative results are visually illustrated in Figure 5.

Method	**(NM)**	**(BG)**	**(CL)**	**Avg Accuracy(%)**
TAGNet-BiLSTM [60]	90.4	87.9	84.1	87.5
MS-GLFGCN[61]	94.8	83.1	70.2	82.7
VeMResGCN [63]	90.3	80.7	73.4	81.5
GaitFFDA [72]	95.6	90.9	73.7	86.7
GaitTR-U [73]	95.5	89.7	85.7	90.3
AdversarialGait[74]	96.1	92.0	75.2	87.7
SSMCMC[75]	96.7	92.6	76.2	88.5
DensePoseGait [2]	77.5	65.2	45.2	62.6
GaitSkeleton [78]	90.2	78.5	72.4	80.4
SFG-Net[79]	88.0	76.1	70.7	78.2
GaitFormer [91]	97.6	93.4	87.0	92.7
LuGAN-HGC [57]	88.3	79.7	75.4	81.1
GaitSet [66]	95.0	87.2	70.4	84.2
GaitPart [68]	96.2	90.1	78.7	88.3
GLN [89]	96.9	91.2	80.5	89.5
CSTL [70]	95.2	90.5	81.5	89.1
GaitSTAGCN [65]	95.4	89.2	86.9	90.5
GRDDN [16]	97.4	94.1	80.6	90.7
DFCA [18]	94.8	91.2	43.3	74.5
DyGait [20]	87.4	72.2	70.8	76.8
GaitDLF [23]	84.9	70.7	68.0	74.5
QaGait [67]	96.7	94.6	78.2	89.8
GaitPT [58]	92.3	79.7	75.8	82.6
TransGait [22]	98.1	94.9	85.8	92.9
Ours	**97.4**	**94.1**	**89.6**	**93.7**

Table 1: *Performance comparison on the CASIA-B dataset (all views with* 90° *gallery), illustrating the robustness of the proposed method under normal walking, bag-carrying, and coat-wearing conditions.*

7.2 Evaluation on OU-MVLP Dataset

On the OU-MVLP dataset, the proposed framework demonstrates strong cross-view generalization, achieving a Rank-1 accuracy of 97.5%. Compared to prior approaches, it surpasses GaitFormer (91.5%) and RDBA-Net

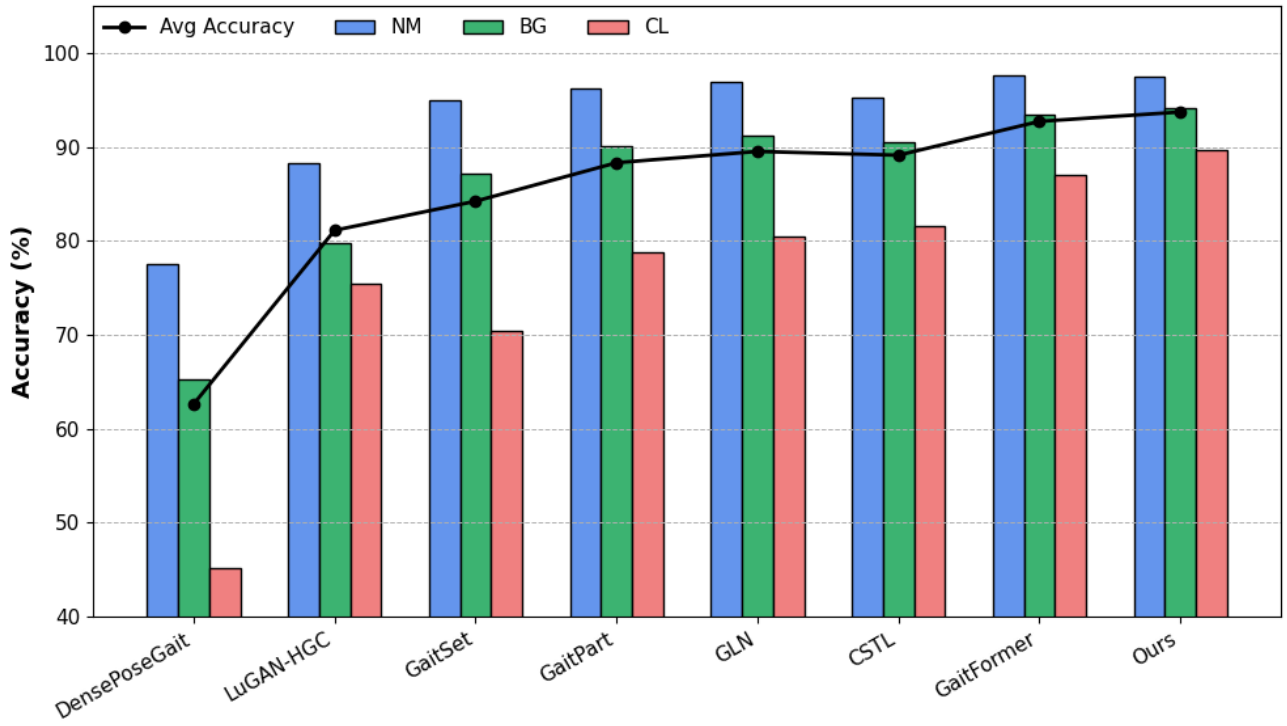


Figure 5: *Graphical comparison of Rank-1 recognition accuracy on the CASIA-B dataset under three walking conditions: normal walking (NM), bag-carrying (BG), and coat-wearing (CL). The proposed method demonstrates consistently high performance across all conditions.*

(89.2%) by +6.0% and +8.3%, respectively. It also outperforms GaitGMT (89.9%) and CSTL (89.1%) by +7.6% and +8.4%, while earlier part-based methods such as GaitPart (83.3%) and GaitSet (79.4%) lag behind by more than +14% and +18%, respectively.

Additionally, the proposed framework consistently exceeds the performance of several recently proposed deep learning, graph-based, adversarial, and memory-enhanced models, including DGait (92.7%), MEMLA-Gait (93.4%), DeepGaitV2 (91.9%), AdversarialGait (91.2%), Silhouette-based approaches (96.2%), LMGCS (90.9%), TAGNet-BiLSTM (75.4%), VeMResGCN (71.0%), GaitFFDA (88.1%), SFG-Net (68.2%), GaitAMR (88.3%), and GLN (85.0%).

Across all these approaches, the proposed method maintains stable and superior performance, highlighting its effectiveness across diverse architectural paradigms and large-scale cross-view evaluation scenarios. The results are reported as Rank-1 identification accuracy averaged across all non-gallery probe views using an 85° gallery setting, as summarized in Table 2.

7.3 Performance Evaluation Insights

The superior performance of the proposed framework stems from embedding the dual-path attention mechanism directly within the multi-scale Inception modules, enabling effective feature refinement at multiple spatial resolutions. By jointly modeling spatial and channel-wise dependencies, the network selectively emphasizes informative gait regions (e.g., leg and arm dynamics) while suppressing irrelevant background or noise. This integrated attention strategy provides richer feature representations than external or post-hoc attention mechanisms, leading to enhanced robustness against appearance variations, viewpoint changes, and diverse walking conditions. As reflected in Tables 1 and 2 and illustrated in Figure 5, this design choice allows the method to consistently outperform recent state-of-the-art approaches—including transformer-based, CNN-hybrid, part-based, graph-based, and adversarial learning models—across normal walking, bag-carrying, coat-wearing, and

Method	Rank-1 Accuracy (%)
TAGNet-BiLSTM [60]	75.4
GaitMPA [17]	91.2
GaitTransformer [56]	90.7
GaitGL [37]	88.7
DGait [62]	92.7
DULE [25]	74.7
RPNet [26]	85.0
VeMResGCN [63]	71.0
Silhouette [64]	96.2
GaitFFDA [72]	88.1
AdversarialGait [74]	91.2
Bifusion [35]	90.0
DeepGaitV2 [76]	91.9
SFG-Net [79]	68.2
MEMLA-Gait [77]	93.4
LMGCS [80]	90.9
RDBA-Net [81]	89.2
GaitFormer [91]	91.5
GaitGMT [82]	89.9
GaitAMR [83]	88.3
GaitSet [66]	79.4
GaitPart [68]	83.3
GLN [89]	85.0
CSTL [70]	89.1
GGCN [40]	87.4
LagrangeGait [59]	90.0
MTGait [39]	90.6
Ours	**97.5**

Table 2: *Performance comparison on the OU-MVLP dataset (85° gallery). The proposed method achieves the highest Rank-1 identification accuracy averaged across all non-gallery probe views, demonstrating superior cross-view gait recognition capability.*

cross-view scenarios, highlighting both the accuracy and generalization capability of the proposed architecture.

7.4 Attention Visualization via Grad-CAM

To substantiate the claim that the spatial attention module effectively suppresses noise arising from clothing variations and other covariate conditions, we provide qualitative visualizations using Gradient-weighted Class Activation Mapping (Grad-CAM) [32]. Grad-CAM computes the gradient of the classification score with respect to the final convolutional feature maps and uses these gradients to produce a coarse localization heatmap, highlighting the silhouette regions assigned the highest attention weights.

Figure 6 illustrates representative heatmaps generated from the spatial attention module for three walking conditions on CASIA-B: normal walking (NM), bag-carrying (BG), and coat-wearing (CL). Under normal conditions, the attention maps consistently assign high weights to the lower limbs—particularly the legs and feet—which carry the most discriminative gait cues. Under the bag-carrying condition (BG), the attention module effectively suppresses the upper-body region affected by the carried item and redirects focus toward the leg dynamics and torso outline. Under the coat-wearing condition (CL), where clothing substantially alters

body silhouette boundaries, the attention mechanism concentrates on joint-related motion areas (knee flexion, ankle trajectories) that remain relatively unaffected by the coat appearance.

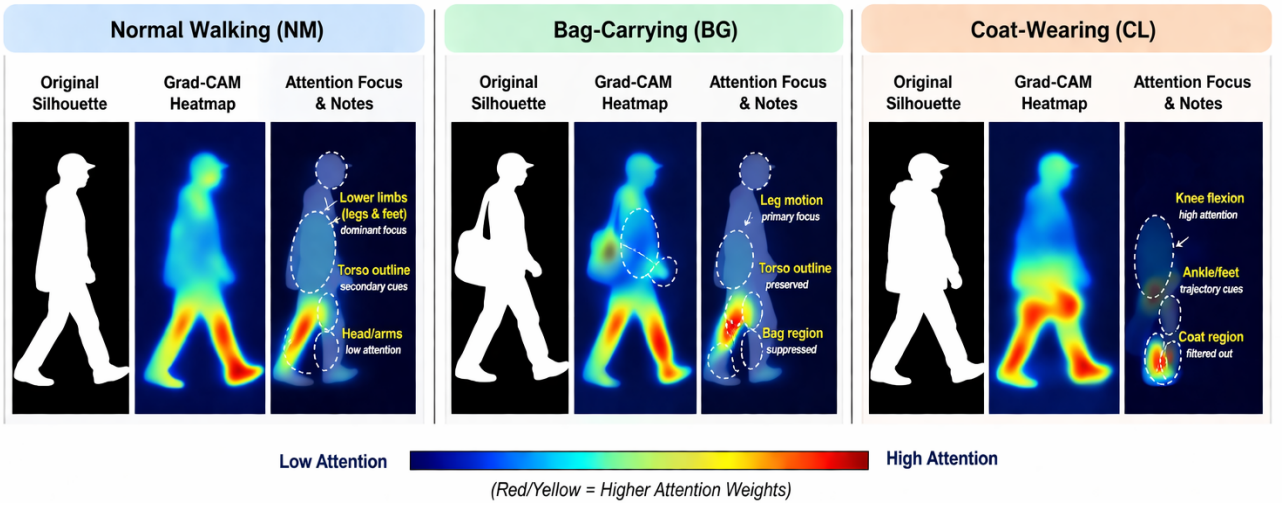


Figure 6: *Grad-CAM heatmaps produced by the spatial attention module for representative silhouettes under three CASIA-B walking conditions: normal walking (NM), bag-carrying (BG), and coat-wearing (CL). Warmer colors (red/yellow) indicate regions assigned higher attention weights, demonstrating that the module focuses on discriminative limb regions while suppressing appearance noise introduced by clothing or carried objects.*

These visualizations confirm that the embedded spatial attention module learns to identify and emphasize the most informative body regions, even under challenging covariate conditions, providing visual evidence for the quantitative robustness gains reported in Tables 1 and 2.

7.5 Limitations and Failure Case Analysis

Although the model demonstrates strong overall performance, it may underperform under extreme viewpoints, severe occlusion, or substantial clothing variations. This subsection provides a dedicated analysis of the conditions under which AttIncGait yields incorrect predictions, along with a discussion of the underlying causes. Representative failure frames from the CASIA-B dataset are illustrated in Figure 7.

Extreme Viewpoints. The model is trained on viewpoints ranging from $0°$ to $180°$ in $18°$ intervals. When evaluated at near-rear angles (above $150°$ from the frontal plane), the projected gait silhouette becomes nearly symmetric and foreshortened, causing viewpoint-dependent features to collapse. As shown in Figure 7(a), a camera positioned approximately at $180°$ captures only the rear profile of the subject, producing a compact silhouette that lacks the lateral leg-swing and torso-rotation cues critical for discrimination. The multi-scale Inception backbone partially mitigates this through holistic body posture modeling, but cannot fully recover discriminative information lost due to extreme projection geometry. Misidentification rates increase noticeably beyond $150°$ from the frontal plane, corresponding to near-rear viewpoints.

Severe Occlusion. When more than approximately 50% of the pedestrian silhouette is unavailable—due to the subject traversing calibration structures, exiting the sensor field, or being obscured by other objects—the available spatial information is insufficient for reliable limb-level feature extraction. Under such conditions, the spatial attention module assigns weights to incomplete or ambiguous silhouette regions, leading to degraded embedding quality. As illustrated in Figure 7(b), a subject crossing the calibration marker columns of the CASIA-B rig partially exits the camera's field of view, providing only a fragmented lower-body silhouette. In our experiments on CASIA-B, sequences with occlusion masks covering more than half the body resulted in a Rank-1 accuracy drop of approximately 12% relative to the full-silhouette baseline.

Substantial Clothing Variations. While the channel attention mechanism suppresses clothing-induced noise in moderate conditions (e.g., standard coats in CASIA-B CL), very heavy or bulky outerwear that drasti-

cally alters the overall silhouette shape causes the model to mislabel gait embeddings. As shown in Figure 7(c), a subject wearing a thick outer coat exhibits a silhouette boundary geometry that deviates substantially from the training distribution, making inter-class gait patterns difficult to distinguish from intra-class appearance variation.

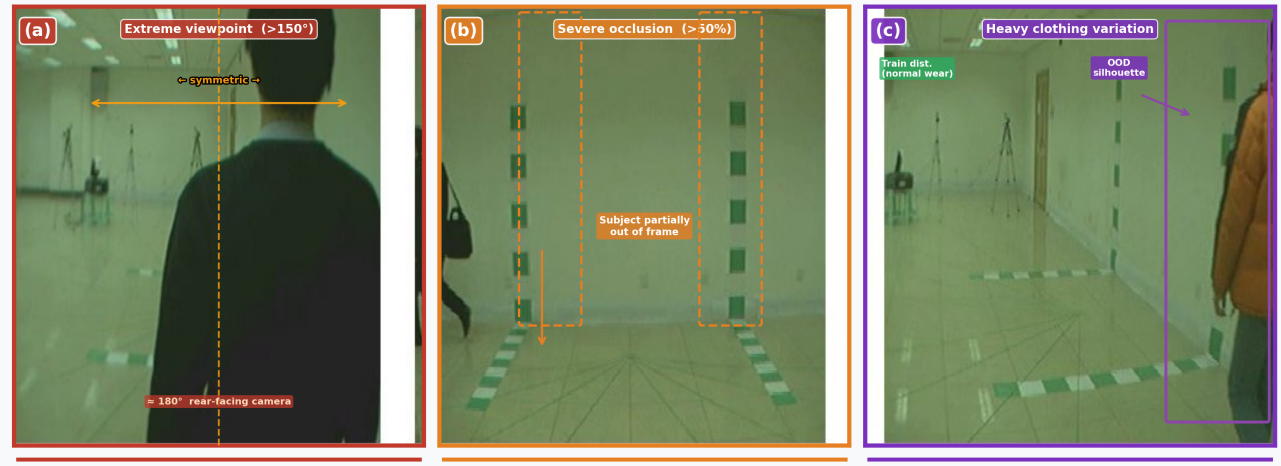


Figure 7: *Limitations and failure case analysis for AttIncGait, illustrated using representative frames from the CASIA-B dataset. (a)* ***Extreme viewpoint*** *(≈180° rear-facing camera): the silhouette appears foreshortened and near-symmetric, collapsing the lateral gait cues that the model relies on for discrimination. (b)* ***Severe occlusion*** *(>50%): the subject traverses the calibration marker columns and partially exits the sensor field, providing an incomplete silhouette that degrades embedding quality (Rank-1 drop ≈12% vs. full-silhouette baseline). (c)* ***Heavy clothing variation****: a thick outer coat pushes the silhouette boundary geometry outside the training distribution, causing the channel attention mechanism to partially fail at recovering discriminative gait cues. Coloured overlays indicate specific failure mechanisms identified through post-hoc analysis.*

Future work will target these failure modes through silhouette–skeleton fusion strategies (incorporating skeleton-based pose priors to compensate for extreme-viewpoint and occluded silhouettes) and domain-adaptive training to improve robustness across unseen clothing conditions and viewpoint ranges. Model optimization for edge deployment will additionally be pursued to balance accuracy and computational efficiency in practical gait recognition applications.

7.6 Cross-Dataset Rank-1 Accuracy

The cross-dataset evaluation in Fig. 8 demonstrates the generalization ability of different gait recognition methods across CASIA-B (90°) and OU-MVLP (85°). Conventional approaches such as GaitSet and GaitPart exhibit noticeable performance drops when transferred between datasets, reflecting limited robustness. More advanced methods like GLN, CSTL, and GaitFormer improve generalization but still show variability. In contrast, the proposed model achieves the highest Rank-1 accuracy of 93.7% on CASIA-B and 97.5% on OU-MVLP, consistently outperforming existing methods. This superior performance highlights the model's robustness against cross-dataset variations and its suitability for real-world applications.

7.7 Computational Efficiency

The proposed study demonstrates considerable improvements in computational efficiency compared to prior architectures. The proposed model contains only 12.4 million parameters, which represents a 57% reduction compared to GaitPart's 28.9 million parameters. Similarly, the total floating-point operations per frame (FLOPs) are reduced from 5.7 G to 3.2 G, corresponding to a 44% decrease. Despite the lighter architecture, the model achieves a throughput of 142 frames per second, representing a 45% increase over GaitPart's 98 fps.

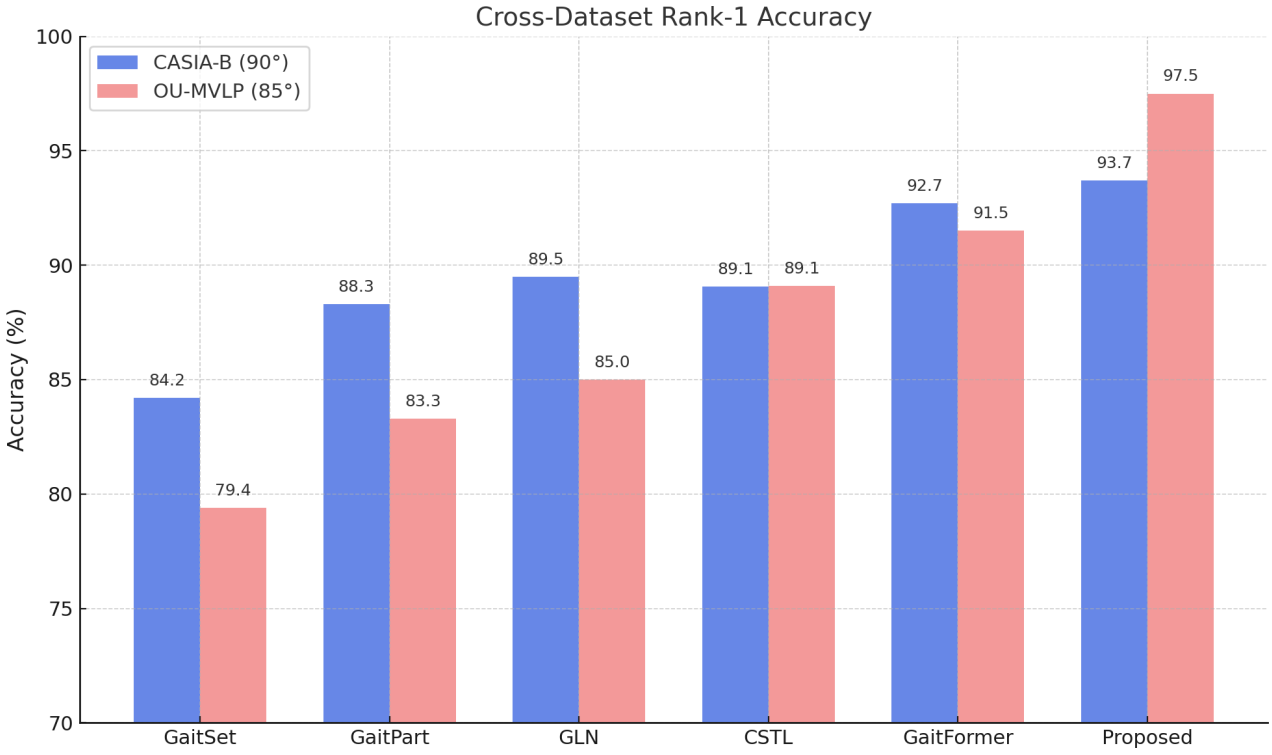


Figure 8: *Cross-Dataset Rank I Accuracy*

These results, as illustrated in Figure 9, demonstrate a favorable balance between accuracy and speed, underscoring the model's suitability not only for high-performance applications but also for potential deployment in resource-constrained environments such as edge devices.

To address the computational comparison requested by the reviewers, Table 3 presents a direct side-by-side comparison of *AttIncGait* with key baseline methods, including GaitSet, GaitPart, GLN, and CSTL, evaluated under identical hardware and experimental settings (NVIDIA RTX 3090 GPU, batch size 64, and FP16 inference precision). The results show that *AttIncGait* achieves the highest average Rank-1 accuracy on CASIA-B while maintaining a lower parameter count and higher throughput than GaitPart and GLN. Compared with GaitSet, which has fewer parameters due to its simpler set-based architecture, the proposed model provides substantially higher accuracy (+9.5%) with only a modest increase in computational cost (3.2 G vs. 0.8 G FLOPs), while still achieving higher inference speed (142 vs. 128 fps). These findings demonstrate that the proposed architecture offers a superior accuracy–efficiency trade-off among the evaluated baselines.

Table 3: *Side-by-side comparison of model parameters, computational cost, and processing speed (FPS) between AttIncGait and representative gait recognition baselines evaluated under identical hardware settings (NVIDIA RTX 3090 GPU, batch size 64, FP16 inference). Parameters are reported in millions (M), FLOPs per frame in GFLOPs, and throughput in frames per second (fps).*

Method	**Params (M)**	**FLOPs/frame (G)**	**FPS**	**Rank-1 Avg (%)**
GaitSet [66]	2.6	0.8	128	84.2
GaitPart [68]	28.9	5.7	98	88.3
GLN [89]	18.4	4.1	112	89.5
CSTL [70]	22.1	4.8	105	89.1
AttIncGait (Ours)	**12.4**	**3.2**	**142**	**93.7**

In addition to the comparative efficiency analysis, the scalability of the proposed model with respect to

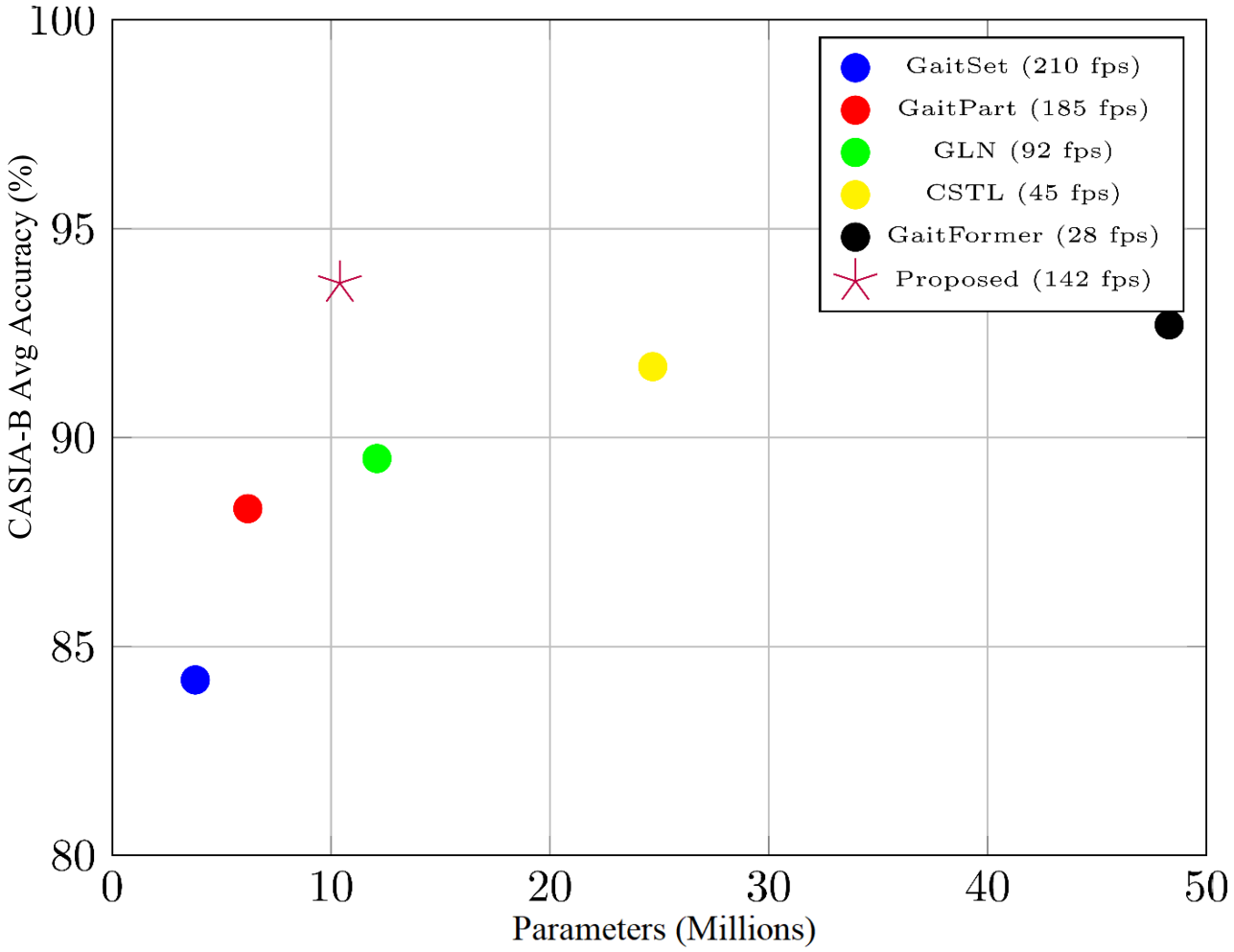


Figure 9: *Accuracy vs. Model Efficiency*

database size was evaluated. Table 4 presents the inference throughput and average execution time per sequence for different dataset sizes, illustrating how the model scales with increasing numbers of subjects.

Table 4: Inference throughput and average execution time per sequence versus database size.

Dataset	**Number of Subjects**	**Throughput (fps)**	**Avg. Time per Sequence (ms)**
CASIA-B	124	138	7.25
OU-MVLP	10,307	134	7.46

The results indicate that the proposed model maintains near-constant inference latency and high throughput even as the database size increases substantially, confirming its suitability for large-scale gait recognition scenarios.

7.8 Ablation Study

An ablation study was conducted on the OU-MVLP dataset to analyze the contribution of individual architectural components, as summarized in Table 5. The baseline configuration without attention mechanisms achieved a Rank-1 accuracy of 89.0% with an EER of 3.1%. Introducing spatial attention improved the accuracy to 90.6% while reducing the EER to 2.4%, indicating that emphasizing discriminative spatial regions

such as limb motion improves recognition reliability. Similarly, incorporating channel attention increased the accuracy to 90.9% with an EER of 2.1%, demonstrating the benefit of adaptively recalibrating feature map responses.

When both attention mechanisms are integrated, the full model achieves a substantially higher Rank-1 accuracy of 97.5% with an EER of 1.3% (Inverse EER 98.7%). The larger improvement compared with the individual modules arises from the complementary interaction between spatial and channel attention. Spatial attention guides the network to focus on informative gait regions, while channel attention simultaneously suppresses irrelevant feature activations within those regions. When embedded within the multi-scale Inception branches, this joint optimization allows the network to reinforce discriminative motion cues across scales, resulting in a synergistic performance gain rather than a purely additive effect, as illustrated in Fig. 10.

Configuration	**Rank-1 Accuracy (%)**	**EER (%)**	Improvement (%)
Baseline	89.0	3.1	–
Spatial Attention only	90.6	2.4	+1.6
Channel Attention only	90.9	2.1	+1.9
Full Model	**97.5**	**1.3**	**+8.5**

Table 5: *Component-wise ablation study on the OU-MVLP dataset, quantifying the individual and combined contributions of spatial and channel attention modules in terms of Rank-1 accuracy and EER.*

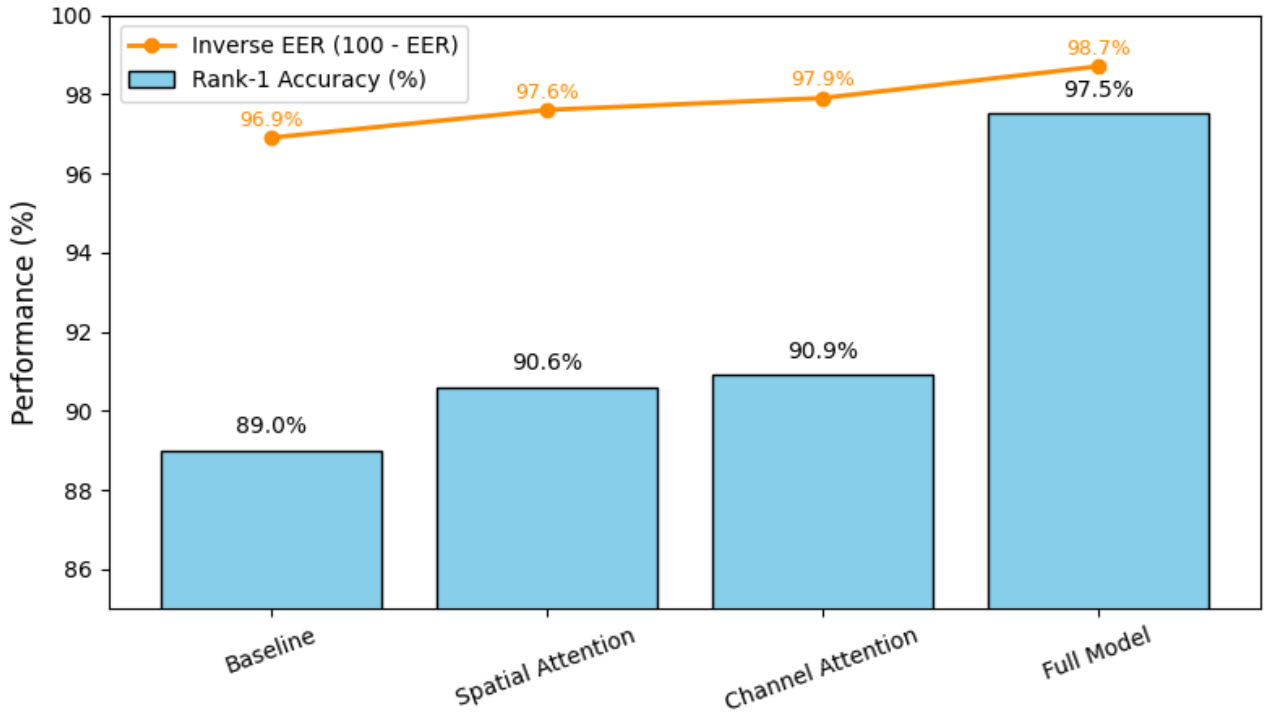


Figure 10: *Ablation Study*

8 Conclusions and Future Work

This paper presented **AttIncGait**, a gait recognition framework that integrates a multi-scale Inception backbone with a dual-path spatial and channel attention mechanism. The key architectural distinction of the proposed method lies in embedding attention directly within the Inception pathways rather than treating it as a post-hoc refinement stage, enabling simultaneous multi-scale feature extraction and adaptive relevance weighting

within a unified structure. Experimental evaluation on the CASIA-B and OU-MVLP benchmarks demonstrates competitive recognition accuracy: the framework achieves 97.5% Rank-1 accuracy and 1.3% EER on OU-MVLP, and 93.7% average Rank-1 on CASIA-B across all walking conditions. Ablation studies confirm the contribution of each attention pathway—spatial and channel attention each yield independent accuracy gains, and their joint integration produces an overall improvement of +8.5% over the attention-free baseline. From a computational standpoint, the model reduces parameter count by 57% relative to GaitPart while sustaining an inference throughput of 142 frames per second on a consumer-grade GPU, demonstrating that recognition accuracy and deployment efficiency need not be traded off against one another. Under covariate conditions, the framework achieves 89.6% Rank-1 under coat-wearing on CASIA-B and maintains consistent cross-dataset performance, with 93.7% Rank-1 when evaluated on CASIA-B after training on OU-MVLP.

Notwithstanding these results, the framework exhibits reduced robustness under near-rear viewpoints (beyond 150°) and severe occlusion (exceeding 50% of the silhouette). A structured analysis of these failure conditions, together with representative dataset examples, is provided in Section 7.5. Future work will pursue silhouette–skeleton fusion to supply pose priors under occlusion and extreme viewpoints, domain-adaptive training to extend robustness across unseen clothing conditions, and model compression for deployment on edge and embedded platforms.

Data Availability Statement

The datasets used in this study (OU-MVLP and CASIA-B) are publicly available. OU-MVLP can be accessed at `http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitMVLP.html`, and CASIA-B is available at `http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp`. No new datasets were generated for this study. Additional data supporting the findings of this work are available from the corresponding author upon reasonable request.

Acknowledgements

The authors gratefully acknowledge the Department of Electronics and Telecommunication, Sanjivani College of Engineering, Kopargaon, for providing the necessary facilities and support to carry out this research work.

References

[1] G. Lin *et al.*, "GaitDG: A Single-Source Domain Generalization Framework for Cross-Domain Gait Recognition", *IEEE Trans. Inf. Forensic Secur.*, 21:121–134, 2026. https://doi.org/10.1109/TIFS.2025.3646048

[2] R. Liao, Z. Li, S. S. Bhattacharyya, G. York, "DensePoseGait: Dense Human Pose Part-Guided for Gait Recognition", *IEEE Trans. Biometrics, Behav. Identity Sci.*, 7(1):33–46, Jan. 2025. https://doi.org/10.1109/TBIOM.2024.3486732

[3] X. Ren, S. Yang, S. Hou, C. Cao, X. Liu, Y. Huang, "Unsupervised Gait Recognition With Selective Fusion", *IEEE Trans. Biometrics, Behav. Identity Sci.*, pages 1–1, 2025. https://doi.org/10.1109/TBIOM.2025.3544805

[4] S. B. Mandlik, R. Labade, S. V. Chaudhari, B. S. Agarkar, "Review of gait recognition systems: approaches and challenges", *Int. J. Electr. Comput. Eng.*, 15(1):349–355, Feb. 2025. https://doi.org/10.11591/ijece.v15i1.pp349-355

[5] M. Khaliluzzaman, A. Uddin, K. Deb, M. J. Hasan, "Person Recognition Based on Deep Gait: A Survey", *Sensors*, 23(10):4875, May 2023. https://doi.org/10.3390/s23104875

[6] J. N. Mogan, C. P. Lee, K. M. Lim, "Ensemble CNN-ViT Using Feature-Level Fusion for Gait Recognition", *IEEE Access*, 12:108573–108583, 2024. https://doi.org/10.1109/ACCESS.2024.3439602

[7] S. Mandlik, R. Labade, S. Chaudhari, and B. Agarkar, "Optimizing Gait-Based Biometric Recognition with Deep Dense Networks", *Braz. J. Biom.*, 43(4), Nov. 2025. https://doi.org/10.28951/bjb.v43i4.857

[8] X. Song et al., "Gait Attribute Recognition: A New Benchmark for Learning Richer Attributes From Human Gait Patterns", *IEEE Trans. Inf. Forensics Secur.*, 19:1–14, 2024. https://doi.org/10.1109/TIFS.2023.3318934

[9] K. Hasan, M. Z. Uddin, A. Ray, M. Hasan, F. Alnajjar, M. A. R. Ahad, "Improving Gait Recognition Through Occlusion Detection and Silhouette Sequence Reconstruction", *IEEE Access*, 12:158597–158610, 2024. https://doi.org/10.1109/ACCESS.2024.3482430

[10] W. Zhai, H. Li, C. Zheng, X. Xing, "Multi-View Gait Recognition With Joint Local Multi-Scale and Global Contextual Spatio-Temporal Features", *IEEE Trans. Circuits Syst. Video Technol.*, 35(2):1123–1135, Feb. 2025. https://doi.org/10.1109/TCSVT.2024.3476384

[11] L. Zhao, P. Cheong, W. Zhang, W.-W. Choi, "Pedestrian Identification System Based on RFID Signaling and Deep Learning", *IEEE Sens. J.*, 25(7):12003–12015, Apr. 2025. https://doi.org/10.1109/JSEN.2025.3544298

[12] J. Li et al., "Rethinking Appearance-Based Deep Gait Recognition: Reviews, Analysis, and Insights From Gait Recognition Evolution", *IEEE Trans. Neural Networks Learn. Syst.*, pages 1–21, 2025. https://doi.org/10.1109/TNNLS.2025.3526815

[13] A. A. Parashar, A. A. Parashar, W. Ding, M. Shabaz, I. Rida, "Data preprocessing and feature selection techniques in gait recognition: A comparative study of machine learning and deep learning approaches", *Pattern Recognit. Lett.*, 172:65–73, Aug. 2023. https://doi.org/10.1016/j.patrec.2023.05.021

[14] Y. Zhang, J. Qin, L. Lv, Z. Wang, "Based on Siamese Network with Self-Attention Model for Gait Recognition", *Proc. 2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, Beijing, China, pages 1118–1122, 2020. https://doi.org/10.1109/ICMA49215.2020.9233689

[15] S. Mandlik, R. Labade, and S. Chaudhari, "Enhancing Gait Recognition with Attention-Based Spatial-Temporal Deep Learning: The GaitDeep Framework", *Comput. Sci. J. Moldova*, 33(2(98)):188–218, Sep. 2025. https://doi.org/10.56415/csjm.v33.10

[16] S. Mandlik, R. Labade, S. Chaudhari, and B. Agarkar, "GRDDN: Enhanced Gait Recognition using a Deep Dense Network," 2025 International Conference on Inventive Computation Technologies (ICICT). IEEE, pp. 129-135, Apr. 23, 2025. https://doi.org/10.1109/ICICT64420.2025.11005378

[17] X. Song, Y. Wang, Y. Huang, and C. Shan, "Gait recognition via View-aware Part-wise Attention and Multi-scale Dilated Temporal Extractor," *Image and Vision Computing*, vol. 156, Art. no. 105464, 2025. https://doi.org/10.1016/j.imavis.2025.105464

[18] T. Ning, K. Lu, and J. Xue, "Learning Multi-Scale Spatial Features Representation in Frequency Domain for Gait Recognition," 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 1482-1488, Oct. 05, 2025. https://doi.org/10.1109/SMC58881.2025.11343543

[19] B. Lin, S. Zhang, F. Bao, "Gait Recognition with Multiple-Temporal-Scale 3D Convolutional Neural Network", *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, pages 3054–3062, Oct. 2020. https://doi.org/10.1145/3394171.3413861

[20] S. Wei, Z. Chen, F. Wei, S. Z. Yang, and C. Wang, "DyGait: Gait Recognition Network Based on Skeleton Dynamic Features," IEEE Access, vol. 12, pp. 189535-189546, 2024. https://doi.org/10.1109/ACCESS.2024.3416433

[21] C. Fan, S. Hou, Y. Huang, and S. Yu, "Exploring Deep Models for Practical Gait Recognition," 2023, arXiv. https://doi.org/10.48550/arXiv.2303.03301

[22] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao, "TransGait: Multimodal-based gait recognition with set transformer," Appl Intell, vol. 53, no. 2, pp. 1535-1547, Apr. 2022. https://doi.org/10.1007/s10489-022-03543-y

[23] S. Wei, W. Liu, F. Wei, C. Wang, and N. N. Xiong, "Gaitdlf: global and local fusion for skeleton-based gait recognition in the wild," J Supercomput, vol. 80, no. 12, pp. 17606-17632, May 2024. https://doi.org/10.1007/s11227-024-06089-7

[24] H. Chao, Y. He, J. Zhang, J. Feng, "GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition", *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, Honolulu, HI, USA, pages 8126–8133, 2019. https://doi.org/10.1609/aaai.v33i01.33018126

[25] S. Zhang, Y. Wang, and A. Li, "Cross-View Gait Recognition with Deep Universal Linear Embeddings," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 9091-9100, Jun. 2021. https://doi.org/10.1109/CVPR46437.2021.00898 PMCid:PMC8431598

[26] H. Qin, Z. Chen, Q. Guo, Q. M. J. Wu, and M. Lu, "RPNet: Gait Recognition With Relationships Between Each Body-Parts," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 5, pp. 2990-3000, May 2022. https://doi.org/10.1109/TCSVT.2021.3095290

[27] A. Jamsrandorj, Q. H. N. Nguyen, D. Jung, M. S. Baek, K.-R. Mun, J. Kim, "Image-based Gait Spatiotemporal Parameters Estimation using a Single Camera and CNN-Transformer Hybrid Network", *Proc. 2023 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Sydney, NSW, Australia, pages 1–5, Jul. 2023. https://doi.org/10.1109/EMBC40787.2023.10339950

[28] C. Yan, B. Zhang, F. Coenen, "Multi-attributes Gait Identification by Convolutional Neural Networks", *Proc. 2015 8th Int. Congr. Image Signal Process. (CISP)*, pages 642–647, 2016. https://doi.org/10.1109/CISP.2015.7407957

[29] C. Yang, J. Zheng, Y. Gao, R. Yan, C. Wang, Y. Wang, "Gait Phase Detection Based on Inception-Self-Attention", *Proc. 2024 6th Int. Conf. Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pages 687–691, Dec. 2024. https://doi.org/10.1109/RICAI64321.2024.10911837

[30] H. V. Hoang, M.-T. Tran, "DeepSense-Inception: Gait Identification from Inertial Sensors with Inception-like Architecture and Recurrent Network", *Proc. 2017 13th Int. Conf. Computational Intelligence and Security (CIS)*, pages 594–598, Dec. 2017. https://doi.org/10.1109/CIS.2017.00138

[31] M. S. Shakeel, K. Liu, X. Liao, W. Kang, "TAG: A Temporal Attentive Gait Network for Cross-View Gait Recognition", *IEEE Trans. Instrum. Meas.*, 74:1–14, 2025. https://doi.org/10.1109/TIM.2024.3497164

[32] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, Q. Wu, "Enhanced Spatial-Temporal Salience for Cross-View Gait Recognition", *IEEE Trans. Circuits Syst. Video Technol.*, 32(10):6967–6980, 2022. https://doi.org/10.1109/TCSVT.2022.3175959

[33] M. A. A. Al-qaness, A. Dahou, N. T. Trouba, M. Abd Elaziz, A. M. Helmi, "TCN-Inception: Temporal Convolutional Network and Inception modules for sensor-based Human Activity Recognition", *Future Generation Computer Systems*, 160:375–388, Nov. 2024. https://doi.org/10.1016/j.future.2024.06.016

[34] M. De Lim, T. Connie, M. K. O. Goh, N. 'Izzati Saedon, "Model-Based Feature Extraction and Classification for Parkinson Disease Screening Using Gait Analysis: Development and Validation Study", *JMIR Aging*, 8:e65629, Apr. 2025. https://doi.org/10.2196/65629

[35] Peng, Y., Ma, K., Zhang, Y. et al., "Learning rich features for gait recognition by integrating skeletons and silhouettes", Multimed Tools Appl 83, 7273-7294 (2024). https://doi.org/10.1007/s11042-023-15483-x https://doi.org/10.1007/s11042-023-15483-x

[36] W. An et al., "Performance Evaluation of Model-Based Gait on Multi-View Very Large Population Database with Pose Sequences", *IEEE Trans. Biometrics, Behav. Identity Sci.*, 2(4):421–430, 2020. https://doi.org/10.1109/TBIOM.2020.3008862

[37] B. Lin, S. Zhang and X. Yu, "Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 14628-14636. https://doi.org/10.1109/ICCV48922.2021.01438

[38] R. Liao, S. Yu, W. An, Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge", *Pattern Recognit.*, 98:107069, Feb. 2020. https://doi.org/10.1016/j.patcog.2019.107069

[39] Hui-peng Li, Cong-qing Wang, Yang Liu, "Multimodal feature learning for gait recognition of camouflaged individuals", Expert Systems with Applications, Volume 296, Part D, 2026. https://doi.org/10.1016/j.eswa.2025.129202

[40] Hao Qin, Zhenxue Chen, Qingqiang Guo, Q.M. Jonathan Wu, Mengxu Lu, "GGCN: Gait Recognition with Generate Network and Convolutional Neural Network, Journal of Visual Communication and Image Representation", Volume 117, 2026. https://doi.org/10.1016/j.jvcir.2026.104790

[41] M. Frikha, I. Chtourou, E. Fendri, M. Hammami, "BiMPeR: A Novel Bi-Model Person Re-identification Method based on the Appearance and the Gait Features", *Procedia Comput. Sci.*, 192:913–922, 2021. https://doi.org/10.1016/j.procs.2021.08.094

[42] B. Hadjkacem, W. Ayedi, M. Ben Ayed, S. A. Alshaya, M. Abid, "A novel Gait-Appearance-based Multi-Scale Video Covariance Approach for pedestrian (re)-identification", *Eng. Appl. Artif. Intell.*, 91:103566, May 2020. https://doi.org/10.1016/j.engappai.2020.103566

[43] R. Liao, W. An, Z. Li, S. S. Bhattacharyya, "A novel view synthesis approach based on view space covering for gait recognition", *Neurocomputing*, 453:13–25, Sep. 2021. https://doi.org/10.1016/j.neucom.2021.04.081

[44] W. An, R. Liao, S. Yu, Y. Huang, P. C. Yuen, "Improving Gait Recognition with 3D Pose Estimation", *Proc. 13th Chinese Conf. Biometric Recognition (CCBR 2018)*, Urumqi, China, pages 137–147, Aug. 2018. https://doi.org/10.1007/978-3-319-97909-0_15

[45] R. Liao, Z. Li, S. S. Bhattacharyya, G. York, "View DiffGait: View Pyramid Diffusion for Gait Recognition", *Proc. 2024 IEEE 18th Int. Conf. Automatic Face and Gesture Recognition (FG)*, pages 1–9, May 2024. https://doi.org/10.1109/FG59268.2024.10581996

[46] R. Liao, W. An, S. Yu, Z. Li, Y. Huang, "Dense-View GEIs Set: View Space Covering for Gait Recognition based on Dense-View GAN", *Proc. 2020 IEEE Int. Joint Conf. Biometrics (IJCB)*, pages 1–9, Sep. 2020. https://doi.org/10.1109/IJCB48548.2020.9304910

[47] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, "Gait recognition for person re-identification", *J. Supercomput.*, 77(4):3653–3672, 2021. https://doi.org/10.1007/s11227-020-03409-5

[48] C. Wang, J. Zhang, J. Pu, X. Yuan, L. Wang, "Chrono-Gait Image: A Novel Temporal Template for Gait Recognition", *Computer Vision - ECCV 2010*, Berlin, Heidelberg, pages 257–270, 2010. https://doi.org/10.1007/978-3-642-15549-9_19

[49] X. Li, Y. Wang, B. Zhang, J. Ma, "PSDRNN: An Efficient and Effective HAR Scheme Based on Feature Extraction and Deep Learning", *IEEE Transactions on Industrial Informatics*, 16(10):6703–6713, 2020. https://doi.org/10.1109/TII.2020.2968920

[50] T. Antamis, A. Drosou, T. Vafeiadis, A. Nizamis, D. Ioannidis, D. Tzovaras, "Interpretability of deep neural networks: A review of methods, classification and hardware", *Neurocomputing*, 601:128204, 2024. https://doi.org/10.1016/j.neucom.2024.128204

[51] R. J. S. Raj, S. J. Shobana, I. V. Pustokhina, D. A. Pustokhin, D. Gupta, K. Shankar, "Optimal Feature Selection-Based Medical Image Classification Using Deep Learning Model in Internet of Medical Things", *IEEE Access*, 8:58006–58017, 2020. https://doi.org/10.1109/ACCESS.2020.2981337

[52] J. Wang, X. Gao, F. Zhu, X. Chen, "Exploring Frontier Technologies in Video-Based Person Re-Identification: A Survey on Deep Learning Approach", *Computers, Materials and Continua*, 81(1):25–51, 2024. https://doi.org/10.32604/cmc.2024.054895

[53] S. Mandlik, R. Labade, S. Chaudhari, and B. Agarkar, "Pose-invariant gait recognition using an enhanced inception-ResNet and vision transformer framework", *J. Eng. Appl. Sci.*, 72(1), Dec. 2025. https://doi.org/10.1186/s44147-025-00820-8

[54] G. Batchuluun, H. S. Yoon, J. K. Kang, K. R. Park, "Gait-Based Human Identification by Combining Shallow Convolutional Neural Network-Stacked Long Short-Term Memory and Deep Convolutional Neural Network", *IEEE Access*, 6:63164–63186, 2018. https://doi.org/10.1109/ACCESS.2018.2876890

[55] C. Li, H. Huang, "Emergence of hierarchical modes from deep learning", *Phys. Rev. Res.*, 5(2):L022011, Apr. 2023. https://doi.org/10.1103/PhysRevResearch.5.L022011

[56] Y. Cui and Y. Kang, "GaitTransformer: Multiple-Temporal-Scale Transformer for Cross-View Gait Recognition," 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 2022, pp. 1-6. https://doi.org/10.1109/ICME52920.2022.9859928

[57] H. Pan, Y. Chen, T. Xu, Y. He, Z. He, "Toward Complete-View and High-Level Pose-Based Gait Recognition", *IEEE Trans. Inf. Forensics Security*, 18:2104–2118, 2023. https://doi.org/10.1109/TIFS.2023.3254449

[58] A. Catruna, A. Cosma, and E. Radoi, "GaitPT: Skeletons are All You Need for Gait Recognition," 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, pp. 1-10, May 27, 2024. https://doi.org/10.1109/FG59268.2024.10581947

[59] T. Chai, A. Li, S. Zhang, Z. Li and Y. Wang, "Lagrange Motion Analysis and View Embeddings for Improved Gait Recognition," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 20217-20226. https://doi.org/10.1109/CVPR52688.2022.01961

[60] D. R. Mesghena *et al.*, "TAGNet-BiLSTM: Transformer augmented network with BiLSTM for skeleton-based gait recognition", *Pattern Recognit.*, 174:112990, Jun. 2026. https://doi.org/10.1016/j.patcog.2025.112990

[61] Z. Xu, X. Wu, B. Wu, and J. Wu, "MS-GLFGCN: Multi-stream global-local fusion graph convolutional network for skeleton-based gait recognition", *Expert Syst. Appl.*, 302:130587, Mar. 2026. https://doi.org/10.1016/j.eswa.2025.130587

[62] Md. Khaliluzzaman and K. Deb, "DGait: Robust gait recognition using dynamic ST-GCN with global aware attention", *Eng. Sci. Technol., Int. J.*, 73:102267, Jan. 2026. https://doi.org/10.1016/j.jestch.2025.102267

[63] Md. Z. Uddin, A. Ray, B. Das, and Md. A. R. Ahad, "View-Embedding GCN for Skeleton-Based Cross-View Gait Recognition", *IEEE Trans. Human-Mach. Syst.*, 55(5):674–685, Oct. 2025. https://doi.org/10.1109/THMS.2025.3595213

[64] A. V. Vishnuram, R. Hilton, and R. Raman, "Walking Direction Estimation Using Silhouette and Skeletal Representations", in *Computer Vision and Image Processing (CVIP 2024)*, Commun. Comput. Inf. Sci., vol. 2478, Springer, Cham, 2026. https://doi.org/10.1007/978-3-031-93703-3_7

[65] A. Wang, Z. Hou, E. Lin, X. Li, J. Liang, and X. Zhou, "GaitSTAGCN: Spatial-temporal attention graph convolutional networks for gait recognition," Neurocomputing, vol. 654, p. 131300, Nov. 2025. https://doi.org/10.1016/j.neucom.2025.131300

[66] H. Chao, K. Wang, Y. He, J. Zhang, J. Feng, "GaitSet: Cross-View Gait Recognition Through Utilizing Gait As a Deep Set", *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3467–3478, 2022.

[67] Z. Wang et al., "QAGait: Revisit Gait Recognition from a Quality Perspective," AAAI, vol. 38, no. 6, pp. 5785-5793, Mar. 2024. https://doi.org/10.1609/aaai.v38i6.28391

[68] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, Z. He, "GaitPart: Temporal Part-Based Model for Gait Recognition", *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14213–14221, Jun. 2020. https://doi.org/10.1109/CVPR42600.2020.01423

[69] Y. Xie, X. Hong, X. Tang, D. Wei, "3D Convolutional Neural Network for Gait Recognition and Demographic Factor Prediction Based on Plantar Pressure Images", *Proc. IEEE Int. Conf. Inf. Technol., Big Data and Artificial Intelligence (ICIBA)*, pages 338–342, Nov. 2020. https://doi.org/10.1109/ICIBA50161.2020.9277305

[70] Z. Huang et al., "3D Local Convolutional Neural Networks for Gait Recognition", *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 14900–14909, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.01465

[71] S. Gul, M. I. Malik, G. M. Khan, F. Shafait, "Multi-view gait recognition system using spatio-temporal features and deep learning", *Expert Syst. Appl.*, 179:115057, Oct. 2021. https://doi.org/10.1016/j.eswa.2021.115057

[72] Z. Wu and Y. Cui, "GaitFFDA: Feature Fusion and Dual Attention Gait Recognition Model", *Tsinghua Sci. Technol.*, 30(1):345–356, Feb. 2025. https://doi.org/10.26599/TST.2023.9010089

[73] S. Meng *et al.*, "From FastPoseGait to GPGait++: Bridging the Past and Future for Pose-Based Gait Recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(9):8231–8248, Sep. 2025. https://doi.org/10.1109/TPAMI.2025.3577594

[74] H. Raj, A. V. Vishnuram, and R. Raman, "AdversarialGait: direction invariant gait recognition with adversarial learning", *Neural Comput. Appl.*, 37(27):22707–22724, Aug. 2025. https://doi.org/10.1007/s00521-025-11505-x

[75] Q. Duan, M. Xu, J. Liu, and Y. Peng, "Spatiotemporal Semantic Modeling and Cross-Modal Collaboration-Based Gait Recognition Under Multiple Views and Various Walking Conditions", *IEEE Access*, 13:145799–145811, 2025. https://doi.org/10.1109/ACCESS.2025.3599550

[76] C. Fan *et al.*, "OpenGait: A Comprehensive Benchmark Study for Gait Recognition Toward Better Practicality", *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):8397–8414, Oct. 2025. https://doi.org/10.1109/TPAMI.2025.3576283

[77] H. Xi and C. Hu, "MEMLA-Gait: robust gait recognition via synergistic integration of heterogeneous experts with mutual and adversarial learning", *J. King Saud Univ. Comput. Inf. Sci.*, 37(8), Oct. 2025. https://doi.org/10.1007/s44443-025-00219-8

[78] D. Zhu, L. Ji, L. Zhu, and C. Li, "Gait coordination feature modeling and multi-scale gait representation for gait recognition", *Int. J. Mach. Learn. Cybern.*, 15(9):3791–3802, Apr. 2024. https://doi.org/10.1007/s13042-024-02120-8

[79] P. D. and M. T., "SFG-Net: Semantic relationship and hierarchical fusion-based graph network for enhanced skeleton-based gait recognition", *Eng. Appl. Artif. Intell.*, 148:110399, May 2025. https://doi.org/10.1016/j.engappai.2025.110399

[80] W. Zhai, H. Li, C. Zheng, and X. Xing, "Multi-View Gait Recognition With Joint Local Multi-Scale and Global Contextual Spatio-Temporal Features", *IEEE Trans. Circuits Syst. Video Technol.*, 35(2):1123–1135, Feb. 2025. https://doi.org/10.1109/TCSVT.2024.3476384

[81] M. Junaid et al., "Human Gait Recognition Using Dense Residual Network and Hybrid Attention Technique with Back-Flow Mechanism", *Digital Signal Process.*, 166:105401, Nov. 2025. https://doi.org/10.1016/j.dsp.2025.105401

[82] G. Chen et al., "GaitGMT: Global Feature Mapping Transformer for Gait Recognition", *J. Vis. Commun. Image Represent.*, 100:104139, Apr. 2024. https://doi.org/10.1016/j.jvcir.2024.104139

[83] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, L. Cui, "GaitAMR: Cross-view gait recognition via aggregated multi-feature representation", *Inf. Sci.*, 636:118920, Jul. 2023. https://doi.org/10.1016/j.ins.2023.03.145

[84] K. Liu, Y. Liu, S. Ji, C. Gao, S. Zhang, J. Fu, "A Novel Gait Phase Recognition Method Based on DPF-LSTM-CNN Using Wearable Inertial Sensors", *Sensors*, 23(13):5905, Jun. 2023. https://doi.org/10.3390/s23135905

[85] Y. Liu, X. Jiang, T. Sun, K. Xu, "3D Gait Recognition Based on a CNN-LSTM Network with the Fusion of SkeGEI and DA Features", *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveillance (AVSS)*, pages 1–8, Sep. 2019. https://doi.org/10.1109/AVSS.2019.8909881

[86] A. Bousmina, M. Selmi, M. A. Ben Rhaiem, I. R. Farah, "A Hybrid Approach Based on GAN and CNN-LSTM for Aerial Activity Recognition", *Remote Sens.*, 15(14):3626, Jul. 2023. https://doi.org/10.3390/rs15143626

[87] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition", *IPSJ Trans. Comput. Vis. Appl.*, 10(1):1–14, 2018. https://doi.org/10.1186/s41074-018-0039-6

[88] S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, "Robust View Transformation Model for Gait Recognition", *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 2117–2120, 2011. https://doi.org/10.1109/ICIP.2011.6115889

[89] S. Hou, C. Cao, X. Liu, Y. Huang, "Gait Lateral Network: Learning Discriminative and Compact Representations for Gait Recognition", *Computer Vision - ECCV 2020*, Cham, Switzerland, pages 382–398, 2020. https://doi.org/10.1007/978-3-030-58545-7_22

[90] X. Huang et al., "Context-Sensitive Temporal Feature Learning for Gait Recognition", *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 12889–12898, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.01267

[91] J. Li, Z. Wang, C. Wang, W. Su, "GaitFormer: Leveraging dual-stream spatial-temporal Vision Transformer via a single low-cost RGB camera for clinical gait analysis", *Knowl.-Based Syst.*, 295:111810, Jul. 2024. https://doi.org/10.1016/j.knosys.2024.111810