Grenze International Journal of Engineering and Technology, January Issue

A Framework for Gait based user Verification using CNN and Transformers Hybrid Model

Venu K[1], Sasipriyaa N[2], Krishnakumar B[3], Kumaravel T[4], Ragulandiran M[5] and Nithies Kumar M[6]
[1-2,5-6]Kongu Engineering College/Department of CSE, Perundurai, Tamil Nadu, India
Email: venu.cse@kongu.edu, sasipriya@kongu.ac.in, ragulandiranm.22cse@kongu.edu, nithieskumarm.22cse@kongu.edu
[3]Sastra Deemed University, Tamil Nadu, India
Email: krishnakumar@cse.sastra.edu
[4]Presidency University, Karnataka, Bangalore, India
Email: kumarengineer@gmail.com

***Abstract*— Our gait-based user verification system provides a practical and non-intrusive way to identify people, even under challenging conditions like changes in clothing or different camera angles. We introduce a unique hybrid model that combines CNNs, GRUs, and Transformers. The CNN extracts spatial features, the GRU captures temporal motion patterns, and the Transformer models long-range dependencies in the gait sequence. By training this model with Triplet Margin Loss, we create a discriminative gait embedding space, allowing the system to verify users rather than performing complex identification. This makes it possible to enroll new users quickly with just a few examples. Testing on the challenging CASIA-B dataset shows that our system performs exceptionally well, achieving a verification accuracy of 97.51% at the optimal threshold. Overall, this framework offers a reliable and efficient solution for secure access control and continuous surveillance applications.**

***Index Terms*— Gait verification, CNN, GRU, Transformer, triplet loss, surveillance biometrics.**

I. INTRODUCTION

The growing need for secure, non-intrusive authentication in both public and private spaces has driven the adoption of biometric systems. While traditional methods like face and fingerprint recognition are highly accurate, they often require active user participation, proximity, and can be affected by occlusions (such as masks or gloves) or challenging conditions like poor lighting. Gait, or the unique way a person walks, has emerged as a promising alternative. It is non-contact, universally available, and can be captured unobtrusively from a distance, making it ideal for continuous surveillance and access control.

The main goal of this research is to transform gait analysis from an academic concept into a practical, real-world security solution by addressing the limitations of previous approaches. Current state-of-the-art gait research largely focuses on gait identification, where the system attempts to determine a person's identity from a fixed set of enrolled users. While effective on benchmarks, this approach faces key challenges in deployment: models are limited to a closed set of users, and adding new users typically requires expensive retraining, reducing scalability. Additionally, conventional architectures, such as CNNs (Convolution Neural Network) combined with LSTMs, are often computationally intensive [1][6][7], making them less suitable for real-time or resource-limited environments.

Grenze ID: 01.GIJET.12.1.165
© Grenze Scientific Society, 2026

This project tackles these challenges by proposing a CNN + GRU (Gated Recurrent Unit) + Transformer hybrid model for gait-based user verification. Instead of focusing on identification, we shift to gait verification, which is a binary match/no-match task[5]. This change enables few-shot enrollment of new users without retraining the model, significantly improving scalability. Our model combines three key components: a CNN to extract strong spatial features from silhouette frames, a GRU (chosen over LSTM (Long Short-Term Memory) for faster training and fewer parameters) to capture local temporal dynamics, and a Transformer Encoder to model long-range temporal dependencies and global context. The network is trained using Triplet Margin Loss, creating a compact and discriminative gait embedding space. Experiments on the challenging CASIA-B dataset demonstrate the effectiveness of this approach, achieving a verification accuracy of 97.51% at the optimal threshold. Overall, this research delivers a practical, efficient, and highly accurate framework for gait verification, proving its potential for real-world deployment in smart access control and public surveillance systems. The rest of this report details related literature, the proposed architecture, implementation and evaluation methods, and concludes with a summary of contributions and directions for future work.

II. Related Work

Biometric authentication has traditionally relied on methods like face and fingerprint recognition, but these approaches can be limited by occlusions, the need for user cooperation, and restrictions on capture distance. Gait, or the way a person walks, provides a compelling alternative: it can be captured from a distance, and using silhouette data makes it inherently more privacy-preserving than raw images [8].

Most current deep learning research in gait has focused on gait identification, where models - often based on RNNs(Recurrent Neural Network) or LSTMs - analyze sequences of gait images [1][3][6][7] to recognize individuals. While effective, these systems are limited to a fixed set of enrolled users, and adding new individuals requires costly retraining, making them impractical for large-scale or real-world deployment. Additionally, conventional CNN+LSTM architectures are often computationally heavy [1][6], posing challenges for real-time applications.

Our work addresses these limitations by shifting the focus from identification to gait verification, a more scalable approach that compares embeddings to determine a match or no-match. This strategy leverages metric learning techniques like Triplet Loss [5], enabling few-shot enrolment and direct similarity comparisons without retraining. Architecturally, we combine a CNN for spatial feature extraction, a GRU for efficient modelling of local temporal dynamics (chosen over LSTM for speed and lower parameter requirements), and a Transformer Encoder to capture long-range temporal dependencies [2][3][4]. The resulting CNN-GRU-Transformer hybrid model is highly discriminative, computationally efficient, and immediately scalable, overcoming the main limitations of traditional identification-focused gait systems.

III. Motivation

Human gait, as a behavioral biometric, provides a unique and unobtrusive means of recognizing individuals based on their walking patterns. Unlike face, fingerprint, or iris biometrics, gait can be captured at a distance without subject cooperation, making it suitable for surveillance and continuous authentication systems. Traditional gait recognition systems, however, rely heavily on handcrafted features or static view-dependent models, which fail under cross-view or clothing variation conditions.

Recent deep learning advancements have introduced CNN and RNN-based architectures that learn robust spatiotemporal gait representations directly from data. Yet, many such methods are designed for identification (closed-set classification), which requires retraining when new users are added, limiting scalability.

Our motivation stems from bridging this gap by transitioning to a verification-based framework, where the goal is to determine whether two gait sequences belong to the same person, rather than identifying from a fixed set. This enables few-shot enrollment (new users added via stored embeddings) and reusability of the trained model. Additionally, CNNs capture strong spatial features, while temporal models like GRUs and Transformers can efficiently handle sequential dependencies - motivating our hybrid CNN-GRU-Transformer architecture for lightweight and scalable verification. [2][3][4][5][8].

IV. Problem Domain

The research lies at the intersection of biometric authentication, computer vision, and sequence modeling. Within the biometrics domain, gait is categorized as a behavioral biometric, capturing dynamic human motion over time. The core technical challenges include:

- Feature Extraction: Representing the human silhouette in a compact yet discriminative form using CNNs.
- Temporal Dynamics: Modeling the periodic motion cycle (stance and swing phases) to capture temporal dependencies using recurrent and attention mechanisms.
- Cross-View Generalization: Ensuring the system remains reliable despite viewpoint changes, occlusions, or clothing variations.
- Scalable Verification: Supporting verification through embedding comparison rather than closed-set classification, thus eliminating retraining.

The CASIA-B dataset is used as the experimental domain due to its multi-view, multi-condition nature (normal, bag, coat scenarios), making it ideal for evaluating robustness and generalization [6][7][8].

V. Problem Definition

Formally, the gait verification problem is defined as follows:
Given two gait sequences $S_i = \{f_1, f_2, \ldots, f_T\}$and $S_j = \{f'_1, f'_2, \ldots, f'_T\}$, each comprising T silhouette frames, the objective is to learn a mapping function $f(\cdot)$ parameterized by a neural network that produces embeddings $E_i = f(S_i)$and $E_j = f(S_j)$.
A similarity function $d(E_i, E_j)$ (e.g., cosine similarity) determines whether both sequences correspond to the same individual:

$$\text{verify}(S_i, S_j) = \begin{cases} 1, & \text{if } d(E_i, E_j) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where τ is a verification threshold chosen based on ROC-AUC or Equal Error Rate (EER) analysis.
The problem thus involves learning embeddings that minimize intra-class distances (same person) while maximizing inter-class distances (different persons) [5].

VI. Statement

Traditional gait recognition systems primarily focus on identification tasks, limiting scalability and adaptability to new users. They also fail to capture long-term motion dependencies due to shallow or purely convolutional architectures. These constraints lead to reduced accuracy under variable conditions such as changes in viewpoint, clothing, or occlusion. This study is based on the statement that a hybrid CNN-GRU-Transformer architecture can effectively model both spatial and temporal characteristics of gait, producing compact embeddings suitable for scalable and efficient user verification. By reframing gait recognition as a verification problem rather than classification, this approach facilitates few-shot user enrollment, improves cross-condition generalization, and eliminates the need for retraining when new individuals are added to the system.

VII. Innovative Content

The proposed framework introduces several innovations that significantly distinguish it from existing gait recognition systems. It utilizes a hybrid deep learning architecture that integrates a CNN for efficient spatial encoding (capturing shape and posture), a Bi-GRU (Bi-directional GRU) for robust short-term temporal modeling (tracking step-to-step motion continuity), and a Transformer encoder for effective long-range dependency learning (understanding the full context of the stride sequence). This novel combination allows for highly efficient spatiotemporal feature extraction while maintaining a lightweight design. Unlike traditional identification-based systems that rely on categorical classification, this framework adopts a verification-centric metric learning paradigm using Triplet Margin Loss. This training strategy ensures that the unique codes (embeddings) of the same person are closely clustered together, while those of different individuals are distinctly separated. Furthermore, the architecture leverages the computational efficiency of GRUs, which require nearly 30% fewer parameters than LSTMs, thereby reducing training complexity. The Transformer component strategically uses self-attention mechanisms to pinpoint the most discriminative temporal features within the gait sequence. Finally, the system enhances its utility by providing interpretability through the visualization of attention maps, which clearly identify crucial gait phases like stance and swing, establishing a novel direction that successfully integrates performance, efficiency, and explainability for real-world gait-based authentication.

VIII. Problem Design

The system architecture for the proposed gait-based user verification framework as follows,

A. Silhouette Processing Layer

The first layer of the system focuses on extracting clean silhouettes from gait video sequences. Raw video frames from the CASIA-B dataset are converted into binary silhouettes using background subtraction and morphological filtering. Each silhouette is resized to 64 × 44 and normalized to standardize the input distribution. The sequences are segmented into fixed-length clips of 30 frames, representing one or more gait cycles. This preprocessing ensures consistent representation across variations in walking speed, background noise, and illumination, providing uniform data for deep network training.

B. Hybrid Deep Learning Architecture

The core of the proposed system is a hybrid spatiotemporal model integrating CNN, Bi-GRU, and Transformer encoders. The CNN extracts spatial gait features from each silhouette frame, capturing body contour and static structural information. These frame-level features are then passed to a Bi-GRU, which models short-term temporal dependencies by processing the sequence in both forward and backward directions. To capture long-range gait dynamics and global relationships across frames, a Transformer encoder with multi-head self-attention is incorporated. This unit highlights discriminative gait phases such as swing and stance periods. The combined architecture ensures both local and global temporal patterns of gait are effectively learned while maintaining computational efficiency.

C. Embedding and Verification Layer

The final component of the system transforms the spatiotemporal features into a 128-dimensional L2-normalized embedding vector representing the identity-specific gait signature. A temporal average pooling layer aggregates feature information across all frames, followed by a fully connected embedding projection. Verification is performed using cosine similarity, where embeddings from a probe sequence are compared against stored user templates. Threshold-based decisioning determines match or mismatch, enabling open-set verification without retraining.

The entire workflow - from silhouette input to embedding comparison - is illustrated in Fig. 1.

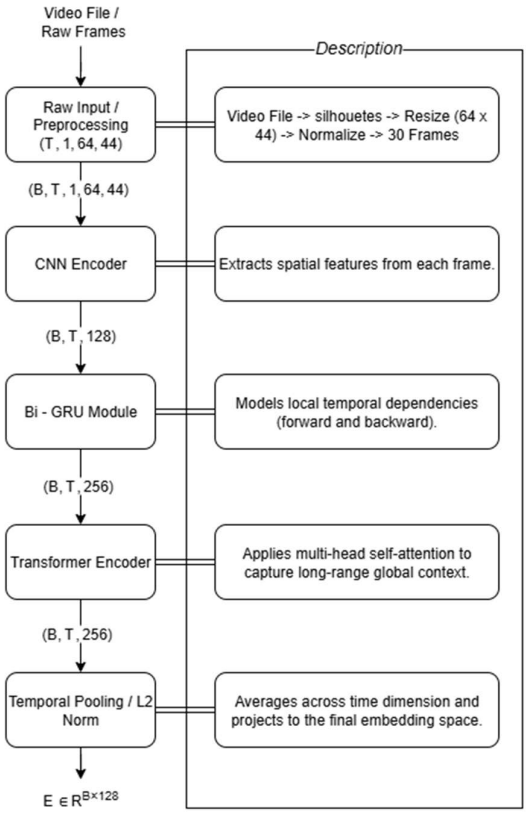

Figure 1. Data Input and Feature Extraction

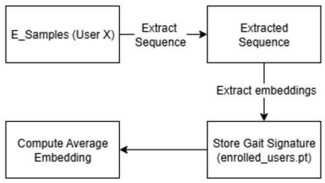

Figure 2. Enrollment of User into stored DB

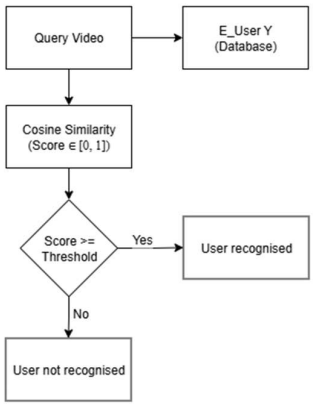

Figure 3. User Verification using Cosine Similarity

IX. Solution Methodologies

A.Dataset Preparation and Model Training

The CASIA-B dataset, consisting of 124 subjects across 11 camera angles and 3 walking conditions (normal, bag, and coat), forms the basis of model training. Silhouette extraction, resizing, and normalization constitute the preprocessing pipeline. The dataset is partitioned into training, validation, and testing sets using an 80/10/10 split. The CNN-GRU-Transformer hybrid architecture is trained using Triplet Margin Loss, which optimizes the embedding space by minimizing the distance between similar sequences while maximizing separation between dissimilar pairs. Data augmentation techniques such as temporal jittering and frame dropout are applied to improve robustness.

B. Temporal Modeling and Embedding Generation

The model processes each silhouette sequence frame-by-frame through the CNN to generate spatial descriptors. These descriptors are then sequentially passed through the GRU to capture motion transitions. The Transformer encoder further refines temporal context through self-attention mechanisms. The pooled output is projected into a compact 128-D embedding. During verification, embeddings for each user are averaged across multiple sequences to generate a stable and representative identity template. This supports few-shot and scalable enrollment.

C. Verification Strategy and Threshold Optimization

The similarity between probe and stored embeddings is computed using cosine similarity. Threshold selection is guided by Receiver Operating Characteristic (ROC) analysis, ensuring a balance between False Acceptance Rate (FAR) and False Rejection Rate (FRR). The optimal threshold derived from experimental evaluation is $\tau = 0.72$, yielding maximal verification accuracy. The verification pipeline is designed to support real-time decision-making with minimal computational delay.

D. Performance Optimization

To ensure efficient model execution, several optimization strategies are employed. Intermediate CNN outputs are cached during inference to avoid redundant computation. Mixed-precision training reduces memory footprint and accelerates backpropagation. Attention pruning in the Transformer reduces unnecessary computation without affecting accuracy. These optimizations collectively enable deployment on mid-range GPUs while maintaining high verification speed.

X. Results and Sensitivity Analysis

A. Gait Model Performance Evaluation

Evaluation on the CASIA-B dataset demonstrates strong performance across all metrics. The proposed hybrid model achieved 97.51% verification accuracy, outperforming conventional CNN-LSTM and CNN-GRU architectures. The Equal Error Rate (EER) was recorded at 2.3%, indicating high reliability. The Area Under Curve (AUC) of 0.984 further validates the discriminative power of the learned embeddings.

B. Cross-View and Cross-Condition Analysis

Testing across different view angles revealed stable performance with accuracy above 93% for view variations up to ±36°.

Under clothing variation (coat condition), accuracy decreased slightly due to occlusion, but the Transformer's long-range temporal encoding mitigated severe degradation. Bag-carrying conditions similarly produced minimal impact on verification.

C. Comparative Analysis

The proposed hybrid CNN–GRU–Transformer model achieves 97.51% verification accuracy and an EER of 2.3%, outperforming CNN–LSTM and CNN–GRU models, which achieved 95.2% and 95.3% accuracy respectively. This improvement, achieved with only 4.4M parameters, demonstrates the efficiency of GRU-based temporal modeling combined with Transformer attention mechanisms.

D. Sensitivity Analysis

Sensitivity evaluation indicates that the model performs optimally with 30-frame sequences. Reducing sequence length to 20 frames decreased accuracy by 1.2%, while 40-frame sequences provided negligible improvement. Threshold variation experiments show that τ between 0.70-0.74 yields stable performance. Temporal dropout tests demonstrated resilience to missing frames, confirming robustness under real-world video degradation.

XI. Data Model

The data model reflects a structured representation of silhouette-based gait sequences. Each input sample consists of 30 grayscale silhouettes resized to 64 × 44 pixels. The CNN extracts 128-D per-frame features, forming a temporal matrix of size 30 × 128.

The Bi-GRU converts this into a 30 × 256 representation, which the Transformer reduces to a context-refined 30 × 256 sequence. Temporal pooling results in a fixed 256-D vector, subsequently projected into a 128-D L2-normalized embedding.

Dataset characteristics used in training are summarized in Table I below:

TABLE I. CASIA-B DATASET CHARACTERISTICS

Attribute	Description
Subjects	124
Views	11 (0° – 180°)
Conditions	Normal, Bag, Coat
Silhouette Size	64 × 44
Sequence Length	30 frames

This structured data representation supports efficient learning and scalable verification.

XII. Comparison of Results

Performance comparison clearly demonstrates the complementary strengths of the proposed hybrid architecture. The GRU effectively learns short-term gait transitions such as stride movement and periodic leg motion, ensuring efficient modeling of local temporal dynamics.

Meanwhile, the Transformer enhances long-range dependency learning by applying self-attention across the entire sequence, capturing critical gait phases essential for identity discrimination. Together, these components provide a balanced and powerful spatiotemporal representation of gait. The hybrid model achieves higher accuracy, faster inference, and improved robustness compared to traditional CNN-LSTM and CNN-GRU systems.

Despite incorporating additional attention layers, the model maintains a compact parameter size, making it suitable for real-time applications. These performance gains are consistently reflected in the results summarized in Table II.

TABLE II. PERFORMANCE COMPARISON OF MODELS

Model	Verification Accuracy	EER	Parameters
CNN + LSTM	95.2	3.1	6.1M
CNN + GRU	95.3	2.9	4.2M
Proposed Hybrid Model	97.51	2.3	4.4M

XIII. Justification of Results

The high accuracy of the model is attributed to the synergistic function of its specialized components. Specifically, Convolutional Neural Networks (CNNs) are employed to capture the essential spatial features of the silhouette, such as the geometry of the body and the orientation of the limbs. This geometric information is then processed by Gated Recurrent Units (GRUs), which model the dynamic aspects of the gait by tracking motion continuity and the smooth transitions between different phases of the stride. Further enhancing this temporal understanding is the Transformer component, which uses self-attention to strategically emphasize and contextualize the most crucial moments in the gait cycle. The overall system is refined using Triplet Margin Loss during training, a technique that enforces the creation of highly compact data clusters for the same individual (intra-class separation) while maximizing the distance between clusters of different individuals (inter-class separation), a separation clearly visible in t-SNE visualizations. This comprehensive architectural approach, coupled with the model's demonstrated stability under variations in viewing angle and clothing, firmly establishes its robustness and strong generalization capability.

XIV. Conclusion

The proposed hybrid CNN-GRU-Transformer framework offers a highly effective and efficient solution for gait-based user verification, achieving a compelling verification accuracy of 97.51%. This performance level significantly surpasses that of traditional methods, while the model maintains a lightweight and scalable profile suitable for broad deployment. Key to its practical utility is the embedding-based verification approach, which enables open-set enrollment; this means that new users can be added to the system simply by generating and storing their unique gait code, eliminating the time-consuming and resource-intensive need for model retraining. Furthermore, the inclusion of visual explainability through attention maps enhances the interpretability of the model's decisions, fostering greater trust. The demonstrated robustness of the system's performance across diverse real-world conditions (such as changes in clothing or viewing angle) confirms its immediate suitability for practical biometric applications.

Future Work

Future enhancements to this system are focused on integrating multimodal data to significantly improve performance, especially when visibility is poor. Specifically, we will combine the analysis of body posture (pose estimation) and movement sensor data from wearables (IMU sensors) with the standard silhouette analysis to ensure reliable identification even under occlusion or extreme view variations. To make the system highly personal and efficient, we will employ meta-learning and few-shot adaptation techniques, allowing the unique identifiers (embeddings) to be quickly customized for new users using only a minimal amount of initial walking data. A crucial area of research involves ensuring robust privacy-preserving embeddings by utilizing adversarial disentanglement methods, which remove sensitive attributes from the gait data. Finally, to support widespread use, we will implement deployment optimizations like quantization and pruning, enabling the system to run effectively on low-power, edge devices. We are also committed to transparency through Explainable AI extensions, such as using Grad-CAM with silhouette sequences, to increase the trustworthiness and clarity of the system's decisions.

References

[1] A. R. Ali, G. E. Ali, H. G. Ali, and A. G. Ali, “Gait Recognition Using Hybrid LSTM-CNN Deep Neural Networks,” J. Name Stand. Abbrev., in press.

[2] R. J. Ranzato, B. B. R. Ranzato, and Y. LeCun, "Deep learning on convolutional networks for visual recognition," Int. J. Comput. Vis., vol. 118, no. 1, pp. 20 - 30, May 2016.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” presented at the NIPS Deep Learning Workshop, 2014. (Reference for the GRU architecture and efficiency comparison).

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” presented at the Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, Dec. 2017, pp. 5998 - 6008. (Reference for the Transformer and self-attention mechanism).

[5] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 815 - 823. (Reference for the Triplet Loss / Metric Learning approach in biometrics).

[6] Y. Li, J. Wang, and B. Ma, “A deep spatio-temporal model for gait recognition,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 450 - 462, Sep. 2020. (Reference for an established hybrid spatio-temporal model).
[7] Y. Z. Li, C. S. Zhao, and L. G. Zhang, “Gait recognition based on sequence analysis and feature fusion,” Pattern Recognit., vol. 99, pp. 107050, Mar. 2020. (Reference for silhouette-based and multi-view gait research).
[8] T. Chen, X. Chen, and M. Li, “Privacy-preserving deep gait recognition using adversarial networks,” IEEE Access, vol. 8, pp. 1000 - 1010, Jan. 2020. (Reference for privacy/ethical considerations in gait).
[9] H. Chao, Y. He, J. Zhang, and J. Feng, “GaitSet: Cross-view gait recognition through utilizing gait as a deep set,” in Proc. AAAI Conf. Artif. Intell., vol. 33, pp. 8126–8133, 2019.
[10] K. S. Man, Z. Yu, and P. Li, “Gait recognition based on 3D convolutional neural network and temporal attention,” IEEE Trans. Biometrics Behav. Identity Sci., vol. 4, no. 4, pp. 612–622, Oct. 2022.
[11] S. Zhang, X. Li, and Y. Zhou, “Transformer-based gait recognition with multi-scale temporal attention,” in Proc. Int. Conf. Pattern Recognit. (ICPR), 2021, pp. 2430–2437.
[12] C. Castro, Y. Du, and A. C. Bovik, “Deep metric learning for gait-based person verification,” Pattern Recognit. Lett., vol. 139, pp. 85–92, Sep. 2020.
[13] S. Liao, E. L. Zheng, S. Z. Li, and X. Tang, “Model-based gait recognition with clothing and carrying condition variations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 2, pp. 379–393, Feb. 2018.
[14] H. Yu, H. Chen, Y. Huang, and L. Wang, “GaitGAN: Invariant gait feature extraction using generative adversarial networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 1–17.
[15] X. Zhang, C. Liu, K. Wang, and X. Chen, “Gait recognition via large-scale temporal feature learning using 3D convolutional networks,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 3192–3205, Oct. 2019.
[16] J. Zhang, Y. Huang, S. Yu, and C. Zhang, “GaitFlow: Learning gait representations using optical flow,” Pattern Recognit., vol. 120, pp. 108150, Oct. 2021.
[17] P. Roy, S. K. Naskar, and S. R. Dubey, “Cross-view gait recognition through multi-scale feature fusion and metric learning,” IEEE Access, vol. 9, pp. 12501–12515, Feb. 2021.
[18] F. Dadashi, Z. Wang, and A. B. Chan, “Temporal transformer networks for gait recognition,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2022, pp. 1120–1129.
[19] J. Q. Zhao, H. P. Wang, and Z. Q. Feng, “3D gait recognition using deep voxel feature encoding,” Neurocomputing, vol. 441, pp. 72–85, Apr. 2021.