An End-to-End 3D-CNN Framework for Silhouette-Based Human Gait Recognition

Asif Mehmood[1], Muhammad Rizwan Rashid Rana[2], Zohaib Mehmood Shah[3], Mehran Yousaf[4], Naveed Yousuf[5]
[1]Department of Computer Science, SZABIST University, Islamabad, Pakistan.
[2]Department of Robotic & Artificial Intelligence, SZABIST University, Islamabad, Pakistan.
[3]Institute of Space Technology, Islamabad
[4]Department of Computer Science, Quaid-i-Azam University, Islamabad
[5]Department of Computer Science, National University of Technology, Islamabad
*Correspondence: asifkhattak333@gmail.com

Citation | Mehmood. A, Rana. M. R. R, Shah. Z. M, Yousaf. M, Yousaf. N, "An End-to-End 3D-CNN Framework for Silhouette-Based Human Gait Recognition", IJIST, Vol. 8 Issue. 4 pp 1539-1558, July 2026
Received| June 03, 2026 **Revised|** July 02, 2026 **Accepted|** July 06, 2026 **Published|** July 10, 2026.

Human gait recognition (HGR) is a method of biometric identification that is widely adopted in fields of identification and authentication. Despite the fact that such a method is effective, it suffers from certain limitations caused by common variations, including illumination, carrying-condition, and walking-speed variations. In this paper, a new method called the 3D-CNN model is designed from scratch to overcome this problem. The proposed model uses silhouettes to obtain both spatial and temporal features. The proposed approach is based on some steps, such as acquisition of data, preprocessing, augmentation, designing of the 3D-CNN model, training of the model, and evaluation. The 3D-CNN model comprises four convolution blocks followed by a classification head. Experimental evaluation on the CASIA-C dataset demonstrated a validation accuracy of 98.48%, with a macro precision of 98.22%, macro recall of 97.60%, and macro F1-score of 97.53% by setting the number of epochs to 150. While accuracies of 95.87% and 96.52% were achieved using 100 and 200 epochs,respectively.

Keywords: Human Gait Recognition, Biometric, Computer Vision, and Deep Learning

Introduction:

Authentication and identification of individuals is an important aspect of safety. To achieve this authentication, different methods such as facial recognition, palm print recognition, fingerprint recognition, etc., are used. These methods [1] are considered more efficient because of the uniqueness of the characteristics of individuals that can serve for their identification. Currently, one of the most popular methods is HGR due to the simplicity of data acquisition [2]. HGR is the study of human movement to identify an individual's gait. Gait can be influenced by many factors, including body structure, muscular strength, and habitual actions, which makes it an excellent biometric measure. The advantage of this method is that it does not require subjects to cooperate when it records the data. The length of stride, gait speed, and total body movement can be measured from a distance; this is why it is regarded as a non-intrusive method. Gait recognition is regarded as effective in different applications [3] and domains such as surveillance, security, and medical diagnostics. For example, in the area of surveillance and security, detection of gait can be used for the identification or authentication of a person even in a crowd. In addition, HGR can also be used in the forensic field for the identification of criminals based on their walking style. The medical field also uses gait for the diagnosis of numerous pathologies relating to gait, including walking abnormalities, spinal stenosis, Parkinson's disease, etc. Also, gait recognition can be applied to observe the process of treatment and rehabilitation [4][5][6].

Moreover, gait is very important for the development of assistive devices, enabling the monitoring and improvement of gait abnormalities. The method facilitates detection and recognition of patients' gait, enabling analysis of abnormal patterns of gait [7].

The classical HGR method includes some steps, such as data collection by means of a camera [8]. Then the collected data is processed by various methods. Hence, the need for human cooperation has been mitigated, enabling more accurate and effective measurements.

The recent development in the CNN model with the use of the RNN method has improved the work of identification [9].

The rapid evolution of Deep Learning (DL) systems has greatly advanced HGR through automatic construction of stable spatiotemporal features from silhouette sequences. Specifically, 3D-CNNs showed promising results in learning information about spatial appearance and temporal movement simultaneously in comparison with classical two-dimensional convolutional networks. Moreover, breakthroughs in recent research carried out in the field of gait recognition have shown significant improvement in gait recognition results through the introduction of attention mechanisms, graph convolutional networks, transformer- based architectures, and multi-scale feature learning to enhance the recognition process in complex walking conditions [10][11][12][13][14]. Still, despite all these advancements, the gait recognition task in unconstrained conditions remains difficult because of the existence of factors that affect the result of the gait recognition process, such as changes in viewpoint, clothing, carrying conditions, variance in walking speed, occlusions, illumination variation, etc. Some researchers state that, among other challenges, improving

the resilience of the gait recognition process to these factors is one of the most important tasks in the area of practical gait recognition development. Moreover, although modern DL models proved to be efficient regarding recognition tasks, many of them utilize complex network architectures, attention mechanisms, and complicated feature fusion approaches that significantly complicate the recognition process and prevent its successful implementation in real-time practice [10][11].

The models obtain various important features from raw data. However, there are some drawbacks faced by the existing models due to many variations of activities, such as clothes, walking, and carrying items. The gait features also distort due to the shadow and changes in the view angle. The modern gait recognition methods have problems with getting spatial and temporal data.

Research Objectives:

To address these issues, the paper presents a new model based on the use of a 3D-CNN architecture. The primary objective of this research is to develop an end-to-end 3D-CNN framework for accurate and robust HGR using silhouette sequences. The proposed framework is designed to effectively learn both spatial and temporal gait representations directly from input frames. Specifically, the objectives of this study are as follows:

- To develop a 3D-CNN model capable of learning both spatio-temporal gait features for accurate and robust authentication and identification of a person.
- To evaluate the performance of the proposed 3D-CNN model through comprehensive experiments by setting the learning rate to 0.0001, a batch size of 4, and training durations of 100, 150, and 200 epochs.
- To evaluate the proposed framework on the publicly available CASIA-C gait dataset using different training configurations and analyze its recognition performance.
- To compare the proposed method with recent state-of-the-art gait recognition approaches to demonstrate its effectiveness and practical applicability for biometric identification.

Research Novelty:

There have been various implementations of DL in recognizing human gait. The most innovative of approaches combines the attention concept, graph convolutional networks, transformers, and multi-stream fusion, aiming at improving recognition rates. However, while these methods are indeed promising, they suffer from high computational complexity and the necessity of optimizing the modeling system, which complicates their implementation in real biometric systems.

This research introduces the new concept of an end-to-end 3D CNN architecture that is capable of producing hierarchical spatio-temporal gait features without using any additional technologies or approaches. The proposed system comprises four consecutive 3D convolutional blocks that successfully capture both spatial characteristics and motion effects during the entire process of recognition. Moreover, the new architecture is built in a

comparatively simple manner, but it still demonstrates rather good recognition ability, as evidenced by the test results.

The proposed architecture can be transformed into a complete end-to-end learning system, which includes a preprocessing function, spatio-temporal feature extraction, and hierarchical representation learning. Thus, this proposition proves the possibility of producing certain features in a semi-automatic way without using any specialized behavior cues.

Related Work:

A lot of work has been done by the author previously in the field of HGR. The authors introduced a lightweight device to capture multi-modality data such as ground reaction force (GRF), surface electromyography, and joint motion information [15]. The data acquisition is performed in different scenarios, including walking on stairs, walking on a slope, and standing still. They used a Bi-LSTM model to classify data and attained promising results.

To address the lateral view problem in HGR, a new method based on frontal view analysis is proposed [16]. The method uses multimodal data to represent the features of gait. Initially, the holistic silhouette and optical flow are obtained. Next, the background effect is eliminated using Gait-YOLO, which is based on YOLOv7. Afterward, the feature fusion at the channel and spatial level is performed. Furthermore, the global features are obtained with the help of the gait feature encoder to improve the multi-modal feature fusion. The performance evaluation of the method is carried out by employing CASIA-B and OUMVLP gait, and improved results are achieved compared to recent HGR methods.

In [17], the authors presented an HGR method called RDBA-Net, which is based on the residual block and attention mechanism. The residual block processes the features and concatenates them in the backward direction. The residual block obtains distinct features with the help of attention cues at both the spatial and temporal levels using the silhouette sequence. Moreover, the information flow is improved at earlier layers by employing the backward flow mechanism. The method attained promising results on CASIA-B and OU-MVLP.

In [18], the authors presented a local fuzzy histogram energy image (LFHEI) based HGR method. Initially, the ROI is extracted using the lower body limb, and different sub-regions are used to obtain the optical flow data. Afterward, the feature representation is obtained from the LFHEI, which reflects the characteristics based on the spatial and temporal representations. The features of the descriptors are obtained by employing the deep learning algorithm. After that, the nearest neighbour algorithms are used to map the candidate and query embeddings. CASIA-A and CASIA-B are used to evaluate the method and yield the accuracy of 95.0% and 89.9%, respectively.

A new method [19], which uses the CNN and short-term Fourier transform (STFT), is called STFT-CNN. The data of different walking speeds are acquired with the help of a biomechanical sensor signal mounted on the body. The walking speed data of 22

subjects is processed using a 2D CNN model for classification. An 80:20 ratio is used to evaluate the models' performance, and an accuracy of 91% and 81.40% is obtained for the single-input "IMU-5s" and "GON-NoSeg". In the case of multi-modality data, the attained values for accuracy are 84.20% and 82.80% using "Multi-5s" and "Multi-NoSeg, respectively. To recognize a person from video data, a new HGR method based on DL and moth flame optimization (MFO) is presented [20]. Initially, the two pretrained models are used for feature extraction using a transfer learning approach. Next, both feature vectors are fused by applying the voting-based technique. The feature optimization is performed with the help of a probability-based MFO algorithm. In the final step, various ML algorithms are used for classification. The method is tested with the help of six CASIA-B angles and attained more than 90% accuracy. To deal with the problem of variation in gait identification, a new method is proposed called the improved dense capsule network [21]. Initially, the media pipe pose library is used to transform the silhouettes into skeleton images. Next, for feature extraction from both the skeleton and silhouette images, a CNN model is employed. Additionally, an Improved Dense CapsNet is utilized to integrate the high-level features extracted using a CNN model. To evaluate the method, CASIA-A and CASIA-C silhouettes are used, and the method offered promising results.

A handcrafted feature and ML classifier-based HGR method is presented by the authors to address the challenges of recognition [22]. Initially, the FAST and Harris corner detectors are used to capture important features from input data. The output of these methods is fused and optimized with the help of the K-means clustering algorithm. Finally, the features are subjected to various ML classifiers for recognition. The HGR method evaluation is performed using CASIA-A, and an accuracy of 92.77% is attained.

To overcome the variation problem in the gait video data, a spatial and temporal-based modulation network is used [23]. The spatial features are extracted at different levels of frames with the help of an attention-based CNN model. Furthermore, a Bi-LSTM model is used to extract the temporal features. To improve the segmentation in the video data, an attention mechanism is used. The CASIA-B and OUMVLP are used to evaluate the proposed method, and an accuracy of 91.03% on CASIA-B and 89.10% on the OUMVLP was attained.

A new method called channel topology graph convolution is presented to address the problem of information loss in the spatial and temporal feature-based HGR methods [24]. The topology of nodes is used based on the computation of the unique correlation for each channel. To improve the sensitivity towards the movement of the joint, an attention-based mechanism is used for spatial information. The CASIA-B and OUMVLP-Pose are employed to assess the performance of the method, and promising results are obtained.

In [25], the authors proposed a two-stream network using spatiotemporal feature extraction. They used the CASIA-B dataset and attained promising results. In [26], the authors utilized a two-phase model for gait recognition using a 21-layer CNN model. They used a frontal view of the CASIA-B dataset and attained improved results.

Materials and Methods:

This paper proposed a new HGR technique based on feature extraction using a 3D-CNN model. The technique comprises several stages. These stages include data collection, preprocessing, development of the 3D-CNN model, training the model, and testing the model. The method can be seen in Figure 1, while the upcoming sections explain the methodology in detail.

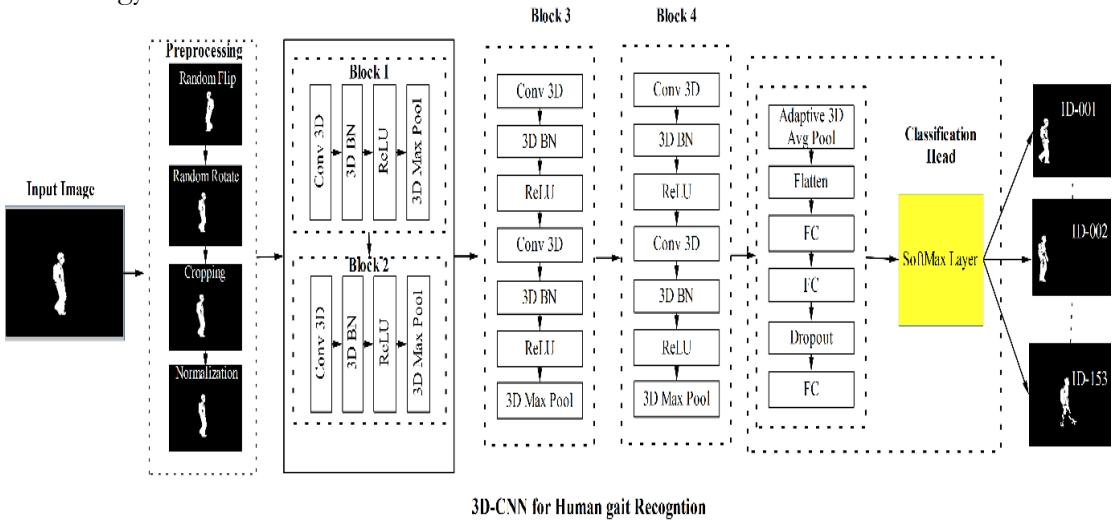

Figure 1. Proposed 3D-CNN Model Architecture for HGR

Dataset Collection:

The CASIA dataset for gait analysis was employed in this work. This dataset comprises multi-view video recordings of human silhouettes. The CASIA-C dataset makes use of infrared sequences and was acquired in a real-world, outdoor climate [2]. The dataset contains walking patterns of 153 people, with variations that include normal walking, slow walking, fast walking, and normal walking while carrying a bag. The overall visualization of the videos is 25fps with a resolution of 352x240. In the gait analysis, the silhouettes of bodies are utilized. In order to retrieve the sequences, a class for the dataset was designed. It automatically inserts the necessary temporal stride into every recorded video instance. The lengths of segments in this case can differ; hence, the positions of frames will change based on the lengths of recorded videos. Each sequence S_i for subject i consists of a set of frames:

$$S_i = [F_1, F_2, \dots, F_{T_i}]\ F_i \in \mathrm{R}^{H \times W} \quad (1)$$

where $S_i$denotes the gait sequence corresponding to the i^{th}subject, $F_t$represents the silhouette frame captured at time step t, $T_i$denotes the total number of frames in the sequence, and $H$and Wrepresent the height and width of each silhouette frame, respectively.

Preprocessing of Input Data:

To achieve better performance, preprocessing has been carried out. To ensure consistency, a clip length and a temporal stride are used before giving frames to the 3D-

CNN model. The length is denoted as L, and a temporal stride is denoted as s. The selection of frames is as follows:

$$S_i' = [F_1, F_{1+s}, F_{2+s} \dots, F_{1+(L-1)s}] \quad (2)$$

where S_i' denotes the preprocessed gait sequence, L is the fixed clip length, s represents the temporal stride used for frame sampling, and $F_{1+(L-1)s}$ denotes the last sampled frame in the selected sequence.

To stabilize the input frames, the last frame of the sequence, denoted as F_{T_i}, is extended by adding redundancy with respect to the frame in the sequence if $T_i < L$. In case $T_i > L$, the sequence will be trimmed to a length of L.

To enhance the generalizing ability of the 3D-CNN model and avoid the overfitting issue, several augmentation techniques have also been adopted with respect to the input frames.

Random Horizontal Flipping:

In the input data, each frame has a 50% chance of being flipped along the horizontal axis. The horizontal flip can be mathematically given:

$$F_j^{flip}(x, y) = F_j(H - x - 1, y), \; x \in [0, \mathrm{H} - 1], y \in [0, \mathrm{W} - 1] \quad (3)$$

Where F_j^{flip} denotes the horizontal flip for an input frame F_j. x and $y$represent the horizontal and vertical pixel coordinates, respectively, while $H$and Wdenote the image height and width.

This augmentation helps the model learn left-right invariance, which is particularly important in gait recognition, as a person may walk with similar patterns from either side.

Random Rotation:

During this phase, all the frames F_j undergo random rotations of θ, where θ is sampled from a uniform distribution. The advantage of this augmentation is that it helps the model to become more robust against factors like slight changes in the viewing conditions, as well as the position change of the object being observed.

$$\theta \sim \mu = -10^0, 10^0 \quad (4)$$

where θ denotes the randomly selected rotation angle sampled from a uniform distribution within the range $[-10^0, 10^0]$.

While the rotated frames can be given as follows:

$$F_j^{rot} = R_\theta(F_j) \quad (5)$$

where F_j^{rot}represents the rotated image, $R_\theta(\cdot)$denotes the rotation operator with angle θ, and $F_j$represents the original input frame.

Random Resized Cropping:

The input images go through random resizing and cropping procedures. The scaling can be calculated in the following way:

$$s_r \sim U(0.9, 1.1) \quad (6)$$

where s_r denotes the randomly selected image scaling factor sampled from the uniform distribution $U(0.9,1.1)$.

After that, the frames are restored to $H \times W$ size. This helps the network give importance to local features and become unaffected by minimal spatial changes.

Data Normalization:

Lastly, for training process stabilization, the pixel values are transformed into a zero-mean and unit-variance as computed below:

$$F_j' = \frac{F_j - \mu}{\sigma}, \mu=0.5, \sigma=0.5 \quad (7)$$

where F_j' denotes the normalized image, F_j represents the original input image, and μ and σ denote the mean and standard deviation used for normalization, respectively.

Normalization is performed to keep the values in the same range and hence make sure that the gradients support the training process. The application of the augmentations makes it possible for the neural network to learn the spatiotemporal features that come from the movement of the limbs.

Model Architecture:

After input data preprocessing, the architecture of a 3D-CNN was developed to extract spatio-temporal characteristics from the data. The architecture consisted of several 3D convolutional layers with ReLU activation and batch normalization so that hierarchical spatiotemporal features are captured.

The proposed 3D-CNN accepts an input tensor of shape (C, T, H, W), where C denotes input channels. While the T denotes the temporal frames, H is the height, and W is the width of the input data, respectively. This representation enables the network to learn spatio-temporal gait information from silhouette sequences. The 3D-CNN consists of four convolutional blocks followed by a classification head.

Block-1 comprises a 3D convolutional layer with 1 input channel and 64 output channels, using a kernel size of 3 × 3 × 3, a stride of 1, and a padding of 1. This layer is followed by BatchNorm3D, a ReLU activation function, and a MaxPool3D layer. The output feature map generated by this block has dimensions of 64 × 20 × 56 × 56.

The second convolutional block expands the feature channels from 64 to 128 using a 3D convolutional layer with the same kernel size, stride, and padding configuration. BatchNorm3D, ReLU activation, and a MaxPool3D layer with a kernel size of 2 × 2 × 2 are subsequently applied. The resulting output feature map is 128 × 10 × 28 × 28.

Block-3 contains two consecutive 3D convolutional layers. The first convolutional layer increases the feature channels from 128 to 256, while the second maintains 256 output channels. Each convolutional layer is followed by BatchNorm3D and ReLU activation, and the block concludes with a MaxPool3D layer of size 2 × 2 × 2. The output of this block is a feature map of size 256 × 5 × 14 × 14.

The final feature extraction stage, Block-4, also consists of two consecutive 3D convolutional layers. The first layer increases the number of channels from 256 to 512, whereas the second preserves the 512-channel representation. Similar to the previous blocks, each convolution is followed by BatchNorm3D and ReLU activation, and a MaxPool3D layer with a kernel size of $2 \times 2 \times 2$ is applied at the end of the block. This block produces an output feature map of size $512 \times 2 \times 7 \times 7$.

After feature extraction, the output is forwarded to the classification head. An AdaptiveAvgPool3D layer first reduces the feature map to a dimension of $512 \times 1 \times 1 \times 1$. The pooled output is then converted into a 512-dimensional feature vector using a Flatten layer. The feature vector is passed through a fully connected layer, followed by a ReLU activation function and a Dropout layer with a rate of 0.5. Finally, a second fully connected layer and a Softmax activation function generate the class probabilities for subject recognition. The complete architecture of the proposed 3D-CNN model is illustrated in Figure 2.

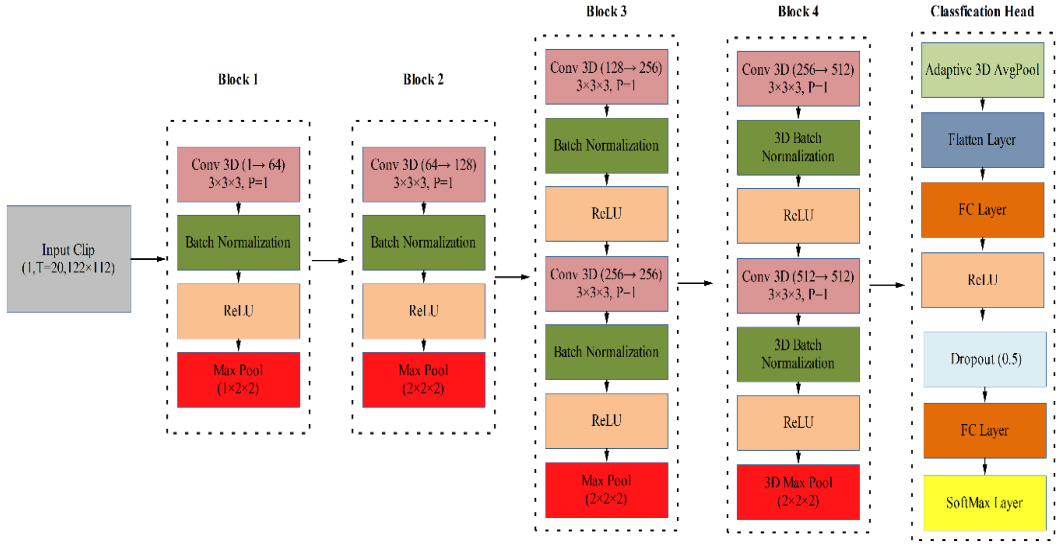

Figure 2. 3D-CNN Model Layer-wise Architecture

Rationale for Selecting the Proposed 3D-CNN Architecture:

The architecture of the suggested 3D-CNN was established to ensure an optimal balance between discriminative spatiotemporal feature learning and performance efficiency. Four convolutional blocks were utilized for gradually learning hierarchical gait representations, where the earlier layers are responsible for obtaining low-level spatial and motion representations, while the deeper layers help in deriving more discriminative semantic gait features. In recent works, it was shown that hierarchical 3D convolutional architectures efficiently model spatial patterns and temporal dynamics, which makes them adequate for gait recognition tasks under difficult covariate conditions such as various viewpoints, clothing-free, and carrying conditions.

The kernel size of 3×3×3 was applied in all the convolutional layers because it represents a good balance between local feature extraction and computational complexity. Prior studies proved that small cubic kernels keep fine spatial data while capturing temporal motion information over the sequential frames, resulting in enhanced efficiency of learning and better generalization in comparison to larger convolutional kernels [11].

The number of feature channels grew from 64 to 512 through the four convolutional blocks so that the network can learn more complex gait representations while keeping the optimization process effective. Batch normalization was implemented after each convolutional block since it assures stabilization of feature distributions, ensures a good speed of convergence, and reduces the internal covariate shift process. The ReLU activation function was selected because of its simplicity and the ability to minimize the vanishing gradient problem related to deep convolutional networks.

Finally, a dropout rate of 0.5 was used before the final classification stage to prevent overfitting. This number has been widely used in recent gait recognition and computer vision models since it provides a good balance between the regularization of the model and the effectiveness of feature extraction from data [13].

Model Training and Evaluation:

The training and validation sets were created from the split of the data into equal parts of 70:30, thus representing all categories. The training process was undertaken with the help of the categorical cross-entropy function utilized for multi-class classifications. For any input sequence S_i, the classification based on the probability distribution among the K classes can be done as follows:

$$y_i' = Softmax(z_i), y_k' = \frac{\exp(z_i,k)}{\sum_{j=1}^{K}(z_i,j)} \quad (8)$$

where y_i' denotes the predicted probability distribution, $z_i$represents the output logits of the network before the Softmax function, $K$is the total number of gait classes, and $y_{i,k}'$denotes the predicted probability for the k^{th} class.

where the unprocessed logit is denoted by z_i for class k. The actual label is used in the one-hot encoding scheme as y_i. The cross-entropy loss is defined as:

$$L_i = \sum_{k=1}^{K} y_{i,k}.\log(y_i', k) \quad (9)$$

where L_i denotes the cross-entropy loss for the i^{th}training sample, $y_{i,k}$ represents the ground-truth label encoded using a one-hot representation, and $y_{i,k}'$denotes the predicted probability corresponding to class k.

While N samples based batch size can be defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}.\log(y_i', k) \quad (10)$$

where Ldenotes the average cross-entropy loss over a mini-batch containing Ntraining samples, $y_{i,k}$represents the ground-truth label, and $y_{i,k}'$denotes the predicted probability for class k.

In case the model predicts the correct class with less probability, it incurs a penalty in the loss function. As such, the loss function pushes the model to predict higher probabilities for the appropriate label of the given class. The model was trained with the Adam optimizer with a batch size of 4, a learning rate of 0.0001 over 100, 150, and 200 epochs. For every epoch, some evaluation parameters were recorded, such as training loss, training accuracy, and validation accuracy.

The detailed algorithm is given below:

Algorithm 1 Proposed End-to-End 3D-CNN Framework

Require: $S,\ L,\ s,\ \alpha,\ B,\ E$
Ensure: $\hat{C}$
$S_p \leftarrow \mathcal{P}(S, L, s)$
for each $S_i \in S_p$ **do**
  $S_i \leftarrow \mathcal{A}(S_i)$
end for
$H_1 \leftarrow \mathrm{CB}_1(S_p)$
$H_2 \leftarrow \mathrm{CB}_2(H_1)$
$H_3 \leftarrow \mathrm{CB}_3(H_2)$
$H_4 \leftarrow \mathrm{CB}_4(H_3)$
$Z \leftarrow \mathrm{AAP}(H_4)$
$z \leftarrow \mathrm{Flatten}(Z)$
$\hat{y} \leftarrow \mathrm{Softmax}\,(\mathrm{FC}\,(\mathrm{Drop}\,(\mathrm{ReLU}\,(\mathrm{FC}(z)))))$
for $e = 1$ to E **do**
  for each mini-batch $B_i \subset S_p$ **do**
    $\mathcal{L} \leftarrow \mathrm{CE}(y, \hat{y})$
    $\theta \leftarrow \mathrm{Adam}(\theta, \nabla\mathcal{L}, \alpha)$
  end for
end for
$\hat{C} \leftarrow \arg\max(\hat{y})$
return $\hat{C}$

In Algorithm 1, S denotes the set of input gait silhouette sequences, while S_i represents the i^{th} gait sequence. The symbols L and s denote the clip length and temporal stride, respectively. The preprocessing and data augmentation operations are represented by $\mathcal{P}(\cdot)$ and $\mathcal{A}(\cdot)$, respectively. The notation CB_i denotes the i^{th} convolutional block of the proposed 3D-CNN architecture, whereas AAP, FC, Drop, and CE denote the Adaptive Average Pooling layer, fully connected layer, dropout operation, and categorical cross-entropy loss function, respectively. The trainable network parameters are represented by θ, and $\hat{C}$ denotes the predicted subject class generated by the proposed framework.

Hyperparameter Optimization and Experimental Settings:

The hyperparameters of the proposed 3D-CNN model were determined by examining different training setups and analyzing the convergence pattern on the validation dataset. Several sets of initial trials were carried out with different learning rates, batch sizes, and dropout rates to find a valid training configuration. Lastly, with regard to the results gained from monitoring the convergence and the validation results, the learning rate of 0.0001, batch size of 4, and dropout rate of 0.5 were chosen as they guaranteed stable optimization and low overfitting. The Adam optimizer was chosen due to its adaptability to the learning process and quick convergence during training of deep networks.

To achieve reproducibility of the experiments, all experiments were conducted under the same random seed, ensuring repeatability of the experiment results in the same computational environment.

The CASIA-C dataset was divided into 70% for training and 30% for validation while keeping the proportion of the classes the same in every gait type. The training process was carried out during 100, 150, and 200 epochs.

Results and Discussion:

This section explains the results achieved by using a 3D-CNN-based HGR by employing the gait dataset CASIA-C.

Experimental Setup:

To evaluate the performance of the proposed model, several performance metrics were employed, including training accuracy and validation accuracy to assess classification performance. In addition, the loss was monitored throughout the training process to evaluate the convergence behavior of the network. After training, training and validation curves were plotted to visualize the learning progression and verify stable model convergence. All experiments were conducted using the Anaconda Jupyter Notebook environment on a personal computer equipped with 32 GB RAM, a 1.5 TB SSD, an Intel Core i5 12th Generation processor, and an NVIDIA RTX 2060 Super GPU with 8 GB memory.

Achieved 3D-CNN Method Results on CASIA-C:

To evaluate the effectiveness of the proposed 3D-CNN-based HGR method, three experiments were conducted by setting the training epochs to 100, 150, and 200. The model performance was evaluated using loss, training accuracy, and validation accuracy.

With 100 training epochs, the proposed HGR model achieved a loss of 0.3513, a training accuracy of 90.19%, and a validation accuracy of 95.87%. Increasing the number of training epochs to 150 improved the model performance, resulting in a loss of 0.1843, a training accuracy of 94.49%, and a validation accuracy of 98.48%. When the number of training epochs further increased to 200, the model obtained a loss of 0.1171, a training accuracy of 95.89%, and a validation accuracy of 96.52%. These results indicate that the best validation accuracy was achieved at 150 training epochs, whereas further increasing the number of epochs reduced the validation performance despite continued improvements in training accuracy and loss, suggesting the onset of overfitting. The quantitative results of all three experiments are presented in Table 1.

Table 1. 3D-CNN Method Results using CASIA-C

No. of Epochs	Loss	Training Accuracy	Validation Accuracy
100	0.3513	90.19	95.87
150	0.1843	94.49	98.48
200	0.1171	95.89	96.52

The results indicate that the trend of improvement in results is also observed in the case of an increase in epochs to 150 from 100. The maximum accuracy achieved is 98.48% for 150 epochs. On the other hand, the accuracy drops to 96.52% when the number of epochs is set to 200. This also indicates that there is still a minor decrease in loss from 0.1843 to 0.1171. Such behavior of the model indicates that it is not always necessary to achieve good results when the loss decreases.

The final results such as macro precision, macro recall, and macro F1-Score were also computed using 150 epochs and obtained values were 98.22%, 97.60%, and 97.53%, respectively.

In order to test the strength and reliability of the proposed 3D-CNN model, experiments were conducted three times. Two experiments involved different random initial conditions, while the third made use of a constant random seed (42) to repeat the experiment. The model produced accuracies of 98.48%, 98.04%, and 97.61%, leading to an average accuracy of 98.04% with a standard deviation of 0.44%. The small variance between the three experiments indicates the reliability of the proposed model.

The training curves are also given in Figure 3 for 100 epochs.

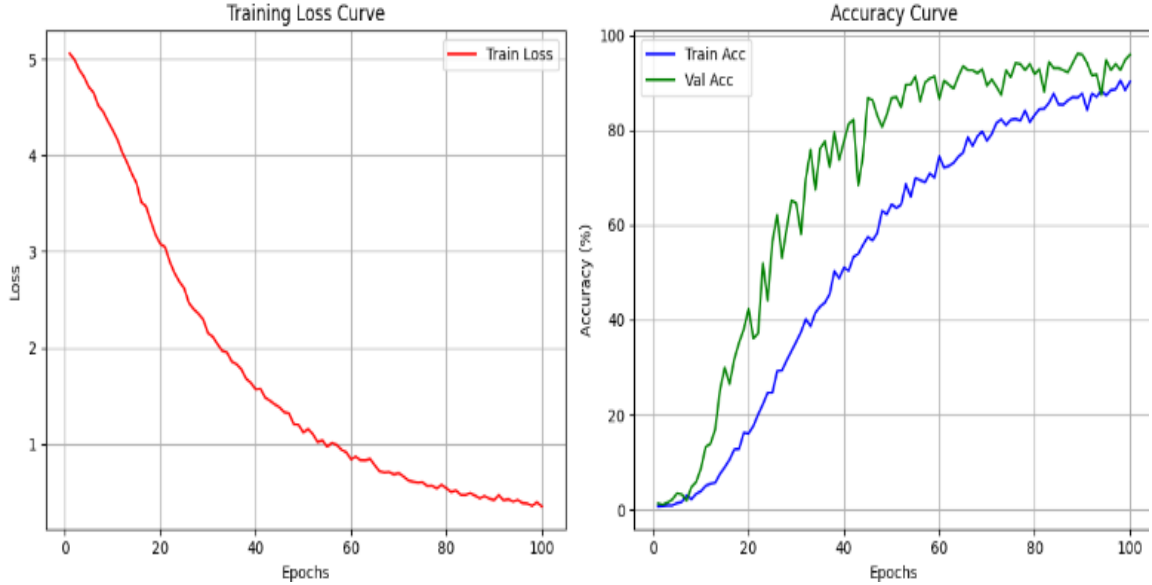

Figure 3. 3D-CNN Method curves by setting 100 epochs

The training curves are also given in Figure 4 for 150 epochs using the proposed 3D-CNN Method.

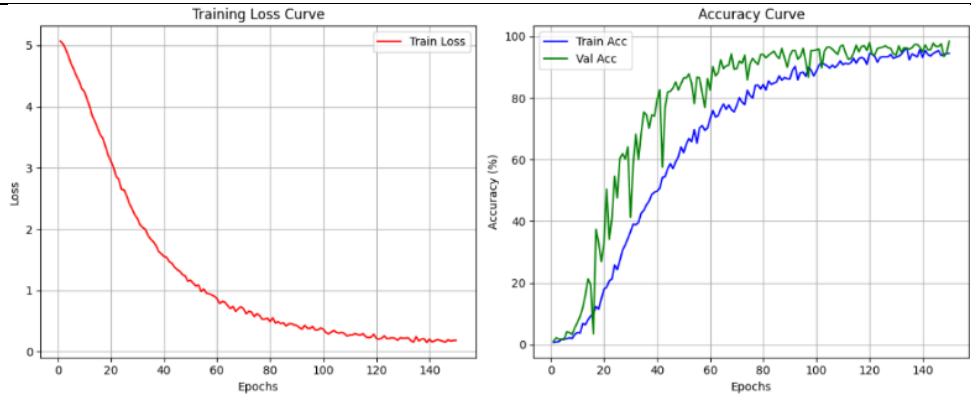

Figure 4. 3D-CNN Method curves by setting 150 epochs

The 3D-CNN method's training curves for 200 epochs are illustrated in Figure 5.

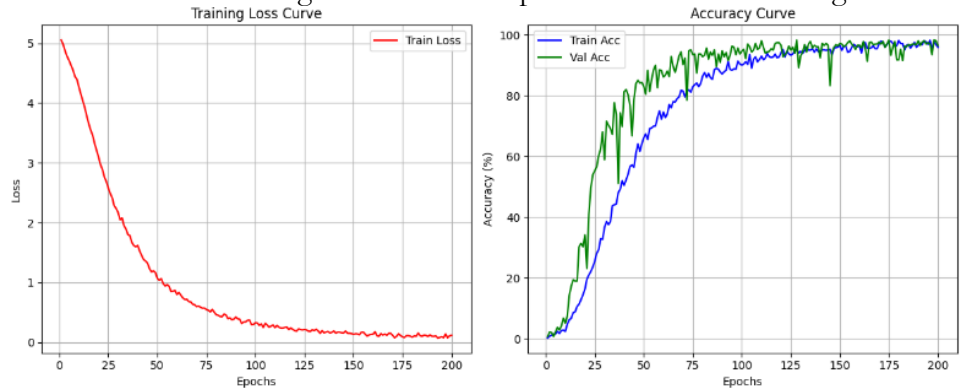

Figure 5. 3D-CNN Method curves by setting 200 epochs

HGR method based on utilizing spatiotemporal characteristics called the Gait Optical Flow Network (GOFN) is introduced by the authors [22]. They presented a multi-scale Inherent Feature Pyramid to fuse the features and used the CASIA-C database to test the method. The processing results yielded 67.90% accuracy. In another method, the authors of [28] proposed a gait recognition method named LGSD, which increased feature similarity with the use of Pairwise Similarity Network (PSN). By employing the CASIA-C database as well, their results were less effective, obtaining only 64.40% accuracy.

In [29], the author introduced an HGR method using a CNN model. They tested their model on CASIA-C and attained 92.79% accuracy, 87.33% precision, 93.23% recall, and 90.18% F1-score, respectively. In [30], the authors used a new few-shot learning ANIL algorithm. They used a meta-training and meta-testing phase and attained an accuracy of 90.96% on CASIA-C.

The proposed 3D-CNN gait recognition method is also evaluated using CASIA-C. The method attained 98.48% accuracy, which is better than the recent gait recognition method, as shown in Table 2.

Table 2. Proposed 3D-CNN Method Comparison with Recent Methods on CASIA-C

Ref	Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Subjects
[31]	GOFI	67.90	-	-	-	100
[28]	LGSD	64.40	-	-	-	100
[29]	CNN Model	92.79	87.33	93.23	90.18	153
[30]	ANIL Algorithm	90.96%	-	-	-	Not-specified
3D-CNN Method (Proposed)		98.04 ± 0.44	98.22%	97.60	97.53%	153

Computational Complexity and Deployment Considerations:

The performance of the proposed 3D-CNN model was evaluated in terms of model complexity, memory requirements, training time, and inference latency. The model consists of 13,669,785 adjustable parameters, which implies a model size of about 52.15 MB, so it can be used on systems equipped with modern GPUs.

The training process lasted about 14,270.89 s (3.96 hours) on the respective hardware. The training process is computationally intensive due to 3D convolutional operations, though it is carried out offline, which does not hinder the capability of real-time applications. The model allowed achieving the average latency of 17.68 ms per gait sequence during inference.

As compared to gait recognition models based on transformers and graph convolutional approaches, which rely on attention mechanisms and graph operations, the proposed architecture is based on a simple hierarchical 3D convolutional structure. So, it is less complex from the architectural point of view while being equally efficient in terms of recognition capability.

Limitations of the Proposed 3D-CNN Model:

The proposed framework has a lot of potential when it comes to performance in recognition; however, it has numerous limitations, which it is important to discuss.

The first issue is that the suggested model has been evaluated on one dataset, CASIA-C, while still providing a variety of walking situations, including normal walking, slow walking, fast walking, and carrying a bag. Due to this fact, the ability of the model to generalize across different datasets was not explored. Therefore, the effectiveness of the model on benchmark datasets, such as CASIA-B, OU-MVLP, Gait3D, or GREW was not evaluated.

There is one more problem. Although the framework has demonstrated outstanding results based on the walking conditions present in the dataset, it has not been tested enough to find out how well it handles a large variety of viewpoints, significant changes in clothing, and strict occlusions. All these factors can change the way gait silhouettes look, thus leading to a decrease in recognition ability in non-structured conditions.

Moreover, the proposed framework has been created and assessed based on the pre-processed silhouette sequences, which made the performance of the model dependent on the extraction of silhouettes and the quality of silhouettes taken from the source.

Discussion:

The findings from the experiments show that the 3D-CNN model could extract distinct gait features from silhouette sequences. Unlike standard 2D networks, which process each image independently, the proposed model carries out spatial and temporal feature extraction through the application of 3D convolutions. This allows the model to better capture walking patterns, thus improving its performance.

The experiments on training conducted over 100, 150, and 200 epochs suggest some interesting patterns in the learning process of the proposed 3D-CNN model. The increase in the number of epochs from 100 to 150 had a dramatic effect on the recognition accuracy of the model, which reached 98.48% from 95.87%, and training loss decreased correspondingly. This effect shows that with more training iterations, the network is able to bring its generalized learning process to a new level. The further increase in the number of epochs to 200 showed a slight drop in the validation accuracy, although there was still a drop in training loss. This indicates that the model started overfitting the training data, though it was particularly effective in remembering the training samples, thus losing its ability to generalize the gait information. Hence, the results of the experiments show that the number of epochs equal to 150 is indeed the balance between an effective learning process and generalization.

The preprocessing approach applied before using the proposed method plays a significant role in its success as well. Utilization of fixed-size silhouette sequences causes the system to properly process walking cycles in a consistent way, while stride duration selection ensures that the flow of motion is preserved within the sequence of recorded frames. This means that various techniques of data augmentation are used in order to evaluate the system in different variations during the training process, such as rotation, cropping, and flipping. Such techniques make the learned capabilities more robust and help avoid the overfitting problem that can lead to poor performance of the system on unseen data.

In this regard, the architectural design is made in a way of providing its four-level hierarchical structure that helps extract the first simple motion structures at earlier stages while more sophisticated movements are being recognized later on. The role of batch normalization is to stabilize the training process, while the dropout process at the classification layer makes the whole system more universal due to an enhanced ability to look less for certain neurons.

The proposed framework is highly effective relative to the current gait recognition methods. Despite the fact that many modern approaches increase recognition capability through the utilization of some means of attention, such as graph convolutional networks, transformer structures, or multi-stream feature fusion. These mechanisms also add complexity, while the proposed 3D-CNN utilizes a simple end-to-end architecture that learns hierarchical spatiotemporal gait representations directly from silhouette images. This requires neither handcrafted feature extraction nor additional feature fusion modules. Through the arrangement of four convolutional blocks, the proposed temporal models

facilitate the gradual accumulation of both low-level motion cues and higher-level gait semantics.

Conclusion:

The HGR technique works based on the ability to identify and authenticate an individual based on the movement of the individual. This paper introduces a novel gait recognition method that utilizes the 3D-CNN model. The proposed method utilizes silhouette frames from the human body for identification purposes. To assess the model's effectiveness, the CASIA-C dataset has been used for subject-level identification, whereby an accuracy of about 98.48% has been achieved, making the new approach more successful in comparison to other gait recognition methods.

Future work will involve testing the 3D-CNN model on other benchmark gait datasets, including CASIA-B, OU-MVLP, Gait3D, and GREW. This will aim to evaluate the 3D-CNN model's ability to cross datasets thoroughly. The model will be tested on various datasets with a range of view angles, clothing conditions, and environmental conditions to deliver a complete evaluation of the robustness of the proposed system. Moreover, future research will focus on investigating how the proposed system can perform in cross-dataset recognition, where the model will be learned during training on one dataset and then tested on another to evaluate its ability to adapt to unknown data distributions.

Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this manuscript in the International Journal of Indigenous Science and Technology (IJIST).

Project details. No funding was received.

References:

[1] "Biometric Recognition - an overview | ScienceDirect Topics." Accessed: Jul. 16, 2026. [Online]. Available: https://www.sciencedirect.com/topics/engineering/biometric-recognition

[2] C. Filipi Gonçalves Dos Santos *et al.*, "Gait Recognition Based on Deep Learning: A Survey," *ACM Comput. Surv.*, vol. 55, no. 2, Feb. 2023, doi: 10.1145/3490235;SERIALTOPIC:TOPIC:ACM-PUBTYPE.

[3] "Gait Recognition with Self-Supervised Learning of Gait Features Based on Vision Transformers." Accessed: Jul. 16, 2026. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7140

[4] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis, "Gait recognition: A challening signal processing technology for biometric identification," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 78–90, 2005, doi: 10.1109/MSP.2005.1550191.

[5] Munish Kumar, Navdeep Singh, "Gait recognition based on vision systems: A systematic survey," *J. Vis. Commun. Image Represent.*, vol. 75, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320321000249

[6] "A wearable system for gait training in subjects with Parkinson's disease - PubMed." Accessed: Jul. 16, 2026. [Online]. Available:

https://pubmed.ncbi.nlm.nih.gov/24686731/

[7] Michal Balazia, Konstantinos N. Plataniotis, "Human gait recognition from motion capture data in signature poses," *IET Biometrics*, 2016.

[8] Rijun Liao, Shiqi Yu, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132031930370X

[9] A. J. Muhammad Zeeshan Arshad, "Gait Events Prediction Using Hybrid CNN-RNN-Based Deep Learning Models through a Single Waist-Worn Wearable Sensor," *Sensors*, vol. 22, no. 21, p. 8226, 2022, doi: https://doi.org/10.3390/s22218226.

[10] J. Li *et al.*, "Rethinking Appearance-Based Deep Gait Recognition: Reviews, Analysis, and Insights From Gait Recognition Evolution," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 6, pp. 9777–9797, 2025, doi: 10.1109/TNNLS.2025.3526815.

[11] Vaishnavi Munusamy, Sudha Senthilkumar, "Emerging trends in gait recognition based on deep learning: a survey," *PeerJ. Comput. Sci.*, vol. 10, 2024, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39145199/

[12] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets, and Challenges," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 7, no. 2, pp. 270–292, 2025, doi: 10.1109/TBIOM.2024.3486345.

[13] Pınar Güner Şahan, Suhap Şahin & Fidan Kaya Gülağız, "A survey of appearance-based approaches for human gait recognition: techniques, challenges, and future directions," *J. Supercomput.*, vol. 80, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s11227-024-06172-z

[14] Anubha Parashar, Apoorva Parashar, "Advancements in artificial intelligence for biometrics: A deep dive into model-based gait recognition techniques," *Eng. Appl. Artif. Intell.*, vol. 130, p. 107712, 2024, doi: https://doi.org/10.1016/j.engappai.2023.107712.

[15] Haoran Zhan, Jiange Kou, "Human Gait phases recognition based on multi-source data fusion and BILSTM attention neural network," *Measurement*, vol. 238, no. 115396, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224124012818

[16] Muqing Deng, Zebang Zhong, Yi Zou, Yanjiao Wang, Kaiwei Wang & Junrong Liao, "Human Gait Recognition Based on Frontal-View Walking Sequences Using Multi-modal Feature Representations and Learning," *Neural Process. Lett.*, vol. 56, no. 133, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s11063-024-11554-8

[17] Mohammad Iman Junaid, Sandeep Madarapu, "Human gait recognition using dense residual network and hybrid attention technique with back-flow mechanism," *Digit. Signal Process.*, vol. 166, p. 105401, 2025, doi: https://doi.org/10.1016/j.dsp.2025.105401.

[18] Muqing Deng, Yi Zou, "Human gait recognition based on frontal-view sequence using discriminative optical flow feature representations and learning," *Eng. Appl. Artif. Intell.*, vol. 145, p. 110213, 2025, doi: https://doi.org/10.1016/j.engappai.2025.110213.

[19] H. Kuduz and F. Kaçar, "Human gait recognition using STFT-CNN approach based on wearable biomechanical sensor data," *Trans. Inst. Meas. Control*, vol. 1, 2025, doi: 10.1177/01423312251341776.

[20] Muhammad Abrar Ahmad Khan, Muhammad Attique Khan, "BAHGRF3: Human gait recognition in the indoor environment using deep learning features fusion assisted framework and posterior probability moth flame optimisation," *CAAI Trans. Intell. Technol.*, 2024, [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12368

[21] A. K. Jhapate and H. Shrivastava, "Deep Learning Empowered Human Gait Recognition with Improved Dense Capsule Networks," *SN Comput. Sci. 2025 66*, vol. 6, no. 6, pp. 596-, Jun. 2025, doi: 10.1007/S42979-025-04123-W.

[22] J. A. Hamid *et al.*, "CASIA-A Gait: Human Identification Through Gait Analysis With FAST and Harris Corner Detector Algorithms," *Natl. Acad. Sci. Lett. 2025 493*, vol. 49, no. 3, pp. 419–423, Mar. 2025, doi: 10.1007/S40009-025-01631-4.

[23] Mohammad Iman Junaid, Allam Jaya Prakash, "Human gait recognition using joint spatiotemporal modulation in deep convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 105, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320324002785

[24] X. Shi, J. Yun, L. Liu, Z. Ma, X. Liu, and Y. Zhang, "CT-Gait: Human Gait Recognition Based on Convolutional Networks with Channel Topology Map," *Lect. Notes Comput. Sci.*, vol. 15858 LNCS, pp. 27–40, 2025, doi: 10.1007/978-981-96-9805-9_3/SAVE-RESEARCH.

[25] Asif Mehmood, Javeria Amin, "Human gait recognition by using two stream neural network along with spatial and temporal features," *Pattern Recognit. Lett.*, vol. 180, pp. 16–25, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865524000436

[26] A. Mehmood, N. Yousuf, and M. Yousaf, "Gait-CNN21: A Two-Phase Human Gait Recognition Technique for Frontal View Analysis," *2024 3rd Int. Conf. Emerg. Trends Electr. Control. Telecommun. Eng. ETECTE 2024 - Proc.*, 2024, doi: 10.1109/ETECTE63967.2024.10823958.

[27] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," *Proc. - Int. Conf. Pattern Recognit.*, vol. 3, pp. 1000–1003, 2006, doi: 10.1109/ICPR.2006.478.

[28] K. Xu, X. Jiang, and T. Sun, "Gait Recognition Based on Local Graphical Skeleton Descriptor With Pairwise Similarity Network," *IEEE Trans. Multimed.*, vol. 24, pp. 3265–3275, 2022, doi: 10.1109/TMM.2021.3095809.

[29] Zill e Haseeb, Shafaan Khaliq, "Intelligent Security In The Maritime Sector Through Artificial Intelligence-Based Gait Analysis," *Polaris J. Marit. Res.*, 2025, doi: 10.53963/pjmr.2025.001.1.

[30] Hongru Wang, Lingui Xiao, "Few-shot gait recognition based on the Almost No Inner Loop algorithm," *Proc. 5th Int. Conf. Comput. Artif. Intell. Control Eng.*, 2026, doi: 10.1145/3804601.3804717.

[31] Hongyi Ye, Tanfeng Sun, "Gait Recognition Based on Gait Optical Flow Network with Inherent Feature Pyramid," *Appl. Sci.*, vol. 13, no. 19, p. 10975, 2023, doi: https://doi.org/10.3390/app131910975.