symmetry

MDPI

Article

MMF-Gait: A Multi-Model Fusion-Enhanced Gait Recognition Framework Integrating Convolutional and Attention Networks

Kamrul Hasan [1,†], Khandokar Alisha Tuhin [1,†], Md Rasul Islam Bapary [1], Md Shafi Ud Doula [1,2], Md Ashraful Alam [3], Md Atiqur Rahman Ahad [4] and Md. Zasim Uddin [1,*]

[1] Department of Computer Science and Engineering, Begum Rokeya University, Rangpur 5404, Bangladesh; kamrulhasanrony111@gmail.com (K.H.); tuhinalisha@gmail.com (K.A.T.); ai.rasulbapary@gmail.com (M.R.I.B.); md.shafi.ud.doula@ait.ac.th (M.S.U.D.)

[2] Department of Information and Communication Technologies, Asian Institute of Technology, Pathum Thani 12120, Thailand

[3] Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Tokyo 113-8655, Japan; aalam@g.ecc.u-tokyo.ac.jp

[4] Department of Computer Science and Digital Technologies, University of East London, London E16 2RD, UK; mahad@uel.ac.uk

[*] Correspondence: zasim@brur.ac.bd

[†] These authors contributed equally to this work.

Abstract

Gait recognition is a reliable biometric approach that uniquely identifies individuals based on their natural walking patterns. It is widely used to recognize individuals who are challenging to camouflage and do not require a person's cooperation. The general face-based person recognition system often fails to determine the offender's identity when they conceal their face by wearing helmets and masks to evade identification. In such cases, gait-based recognition is ideal for identifying offenders, and most existing work leverages a deep learning (DL) model. However, a single model often fails to capture a comprehensive selection of refined patterns in input data when external factors are present, such as variation in viewing angle, clothing, and carrying conditions. In response to this, this paper introduces a fusion-based multi-model gait recognition framework that leverages the potential of convolutional neural networks (CNNs) and a vision transformer (ViT) in an ensemble manner to enhance gait recognition performance. Here, CNNs capture spatiotemporal features, and ViT features multiple attention layers that focus on a particular region of the gait image. The first step in this framework is to obtain the Gait Energy Image (GEI) by averaging a height-normalized gait silhouette sequence over a gait cycle, which can handle the left–right gait symmetry of the gait. After that, the GEI image is fed through multiple pre-trained models and fine-tuned precisely to extract the depth spatiotemporal feature. Later, three separate fusion strategies are conducted, and the first one is decision-level fusion (DLF), which takes each model's decision and employs majority voting for the final decision. The second is feature-level fusion (FLF), which combines the features from individual models through pointwise addition before performing gait recognition. Finally, a hybrid fusion combines DLF and FLF for gait recognition. The performance of the multi-model fusion-based framework was evaluated on three publicly available gait databases: CASIA-B, OU-ISIR D, and the OU-ISIR Large Population dataset. The experimental results demonstrate that the fusion-enhanced framework achieves superior performance.

Keywords: gait recognition; decision-level fusion; feature-level fusion; hybrid fusion; CNN; vision transformer; attention

check for updates

Academic Editors: José Luis Vázquez-Noguera, Horacio Legal-Ayala and Julio César Mello-Román

Received: 7 June 2025

Revised: 9 July 2025

Accepted: 15 July 2025

Published: 19 July 2025

Citation: Hasan, K.; Tuhin, K.A.; Bapary, M.R.I.; Doula, M.S.U.; Alam, M.A.; Ahad, M.A.R.; Uddin, M.Z. MMF-Gait: A Multi-Model Fusion-Enhanced Gait Recognition Framework Integrating Convolutional and Attention Networks. *Symmetry* **2025**, *17*, 1155. https://doi.org/10.3390/sym17071155

1. Introduction

Biometrics is a technology that identifies individuals based on physiological and behavioral characteristics [1]. Examples of biometric traits for human identification include faces [2], fingerprints [3], irises [4], and gait [5]. Facial recognition identifies individuals by analyzing and extracting facial features from images or videos, but its performance can degrade under conditions such as poor illumination, low spatial resolution, occlusion, or changes in facial expressions [6]. Fingerprint-based identification relies on the unique ridge patterns of fingers and typically requires physical contact with a scanner, making it less feasible in unconstrained or remote environments. Iris-based identification analyzes the complex patterns in the colored ring of the eye, offering high precision but requiring close-up, high-resolution imaging under infrared illumination. In contrast, gait recognition, which identifies individuals based on their walking patterns, can be captured from a distance without the participant's cooperation and works effectively with low spatial resolution [7]. Due to the unconscious nature of gait, it is difficult to disguise, making it suitable for applications such as surveillance systems [8], digital forensics [9], and criminal investigations [10].

Although gait recognition works effectively in controlled laboratory environments, real-life scenarios present numerous challenging factors, known as covariates, that affect recognition accuracy. For example, some covariates are related to the individual, such as carried objects (COs) [11], shoes, and clothing [12], while others are related to the surrounding environment, such as variations in viewing angle [13], occlusions [14,15], walking surfaces, and shadows. All of these covariates significantly degrade the performance of gait recognition [16].

In the early stages, gait recognition relied on handcrafted features, including model-based and appearance-based approaches. Model-based approaches [17–19] have tried to construct a model of the human body and observe the movements of separate body parts. Model-based approaches require high computational resources as they generate a human model, identify key body points, and necessitate high-resolution image sequences. In contrast, appearance-based approaches [5,20,21] analyze the sequence of silhouette frames to extract the spatial or spatiotemporal features for identification. However, appearance-based approaches require less computational power and are relatively easy to implement. Moreover, they demonstrated superior performance in gait recognition. However, they usually rely on a single model for feature extraction and often fail to capture essential features in the presence of covariates.

Recently, deep learning (DL)-based methods, such as convolutional neural networks (CNNs), have demonstrated outstanding performance in classification [22], detection [23], and recognition [24]. This success is largely due to their ability to extract essential features from input samples. Moreover, pre-trained CNN models such as GoogLeNet [25], DenseNet [26], VGG [27], ResNet [28], and EfficientNet [29] have gained popularity among researchers. These models, trained and fine-tuned on large-scale datasets, can extract fine-grained and discriminative features that are critical for obtaining superior performance. Although they all belong to the CNN family, they differ significantly in model depth, computational efficiency, and overall effectiveness.

In addition, the introduction of the transformer-based approach [30] has led to significant advancements in the natural language processing (NLP) domain, featuring an encoder–decoder architecture with a self-attention mechanism. Later on, this concept was implemented in DL using the encoder architecture (i.e., vision transformer (ViT) [31]) because of its versatility and performance. Instead of text, the images are split into non-overlapping patches and sent in sequence, along with the positional embedding layer, to the encoder layer for classification. The encoder's self-attention determines which part of the images the model should focus on.

The aforementioned CNN-based and attention-based models have been increasingly applied to gait recognition in recent years [27,28,32–34]. For example, Mogan et al. [32] proposed VGG16-MLP, incorporating a pre-trained VGG-16 [27] model along with a multilayer perceptron (MLP) to improve recognition accuracy, while Pushpalatha et al. [33] used the pre-trained model of ResNet-50 [28] for gait recognition, achieving superior performance due to its architectural complexity and proper training. Later, an attention-based ViT was employed in Gait-ViT [34], where a pre-trained model was employed for gait recognition. Most research, however, relies on a single model, which hinders their ability to extract the detailed features required to recognize complex spatiotemporal features in the presence of covariates. To minimize the impact of covariates, the models must capture all relevant patterns and accurately extract the necessary features. However, each model involves a specific feature and limitations in capturing subtle changes. Therefore, a multi-model-based approach that combines CNNs and ViT is needed for effective gait recognition.

To address the limitations of a single model, multi-biometric approaches [6,35–39] have been explored by researchers. These approaches are based on fusion strategies that enhance classification, identification, and recognition accuracy, as well as robustness. For example, Kittler et al. [35] initially considered multiple face samples and then employed Bayesian estimation theory to fuse these instances, demonstrating improved identity verification. Mehraj et al. [40] proposed a multi-model biometric approach, where they employed AlexNet [22] and VGG-16 [27] to extract gait features and then combined the features from both models and used support vector machines (SVM) for final identification. Inspired by earlier research, this study implements a multi-model fusion-based framework to improve gait recognition accuracy. Specifically, this study leverages the strengths of five selected pre-trained deep learning models (i.e., VGG-16 [27], ResNet-50 [28], GoogLeNet [25], EfficientNet-B0 [29], and ViT [31]) in an ensemble manner to extract diverse features. Key factors were considered when selecting these models (e.g., number of parameters, complexity, effectiveness, computational requirements, etc.). For example, EfficientNet-B0 [29] enables faster learning with a lower computational cost while providing optimal performance. Similarly, VGG-16 [27] is straightforward to implement. Conversely, ViT [31] offers the advantages of patch embeddings and transformer mechanisms, which enable the extraction of data from each image patch, ensuring that detailed local features are captured. However, a single model still faces several limitations, such as a limited representational capacity and algorithmic assumptions, making it prone to poor generalization. Additionally, single models often face issues with the bias–variance tradeoff and heterogeneous data characteristics, resulting in decreased accuracy and adaptability in practical situations.

To overcome these shortcomings, we introduced a multi-model fusion-based framework that utilizes different DL-based models to capture intrinsic patterns of the input data and leverage the strengths of various state-of-the-art (SOTA) algorithms while mitigating the weaknesses of individual models, which ultimately improves accuracy, generalization, and robustness, even in the presence of covariates. Initially, the proposed framework takes a gait energy image (GEI) (i.e., normalized silhouette sequences over a gait cycle, which are averaged together at pixel levels) as an input. The reason for using GEIs is that they exhibit the left–right symmetry assumption [41,42], particularly in the frontal (0°) or rear (180°) views for the sagittal plane, which is very effective for gait recognition. Moreover, the legs produce two high-energy vertical streaks in the GEIs feature because of motion over the gait cycle. These streaks are often symmetrically placed about the centerline in normal gait [41]. These regions are effective in localizing the dominant motion area, which helps improve gait recognition accuracy in the presence of covariates. In the second step, all the selected DL models are utilized separately to fine-tune them and encode the GEI to extract in-depth spatiotemporal features. Subsequently, to improve recognition accuracy, decision-level fusion (DLF), feature-level fusion (FLF), and hybrid fusion (HF) strategies are applied to merge the strengths of these models. Notably, DLF considers the decision of each model and then employs a majority voting mechanism to ensure a reliable output. Furthermore, FLF addresses diversity and robustness by combining the output features of each model to predict the outcome. Particularly, both FLF and DLF are combined in HF, utilizing the advantages of both fusion approaches to increase the performance score. The significant contributions of this study can be outlined as follows:

- We present a multi-model fusion-based framework that leverages the capabilities of five state-of-the-art (SOTA) deep learning models, including both CNN-based and attention-based architectures, to extract intricate and fine-grained spatiotemporal gait features.
- We employ three separate fusion strategies, feature-level fusion, decision-level fusion, and hybrid fusion, to improve the accuracy of gait recognition. These approaches ensure that the decisions of multiple models are consistent and that their unique features enhance recognition performance.
- We demonstrate our proposed framework on the most popular and challenging publicly available gait databases, the CASIA-B [43], OU-ISIR D [44], and OU-LP [45] datasets, and we attain superior performance.

2. Related Work

2.1. Model-Based Approaches

In the early stages of model-based gait recognition, multiple models [46–48] were developed to construct human-body shapes and analyze motion during walking manually. These approaches extracted essential features from the hips, knees, ankles, and feet for person recognition. For example, Bouchrika et al. [46] proposed a model-based approach [46] that used elliptic Fourier descriptors to parameterize joint motion, capturing the nature of human walking and finally extracting ankle, knee, and hip joints for indoor and outdoor environments. Later, Yoo et al. proposed a method [47] that considered gait silhouette sequences to construct a series of 2D stick-shape representations and then used a neural network algorithm for human recognition.

In addition, some studies [17,49–51] have explored the idea that human skeleton data serves as an ideal input for gait recognition models, with which a camera or depth sensor can capture a person's skeleton data. For example, Preis et al. [49] suggested a gait recognition approach that utilizes the Microsoft Kinect, which provides real-time human skeleton generation and tracking data via an integrated depth sensor. Later, Bari et al. [50] introduced a deep neural network-based gait recognition framework that employed Microsoft Kinect to generate a skeleton gait sequence and extract view- and pose-invariant features, such as joint relative cosine dissimilarity and joint relative triangle area, significantly increasing performance.

Recent advancements in DL have also introduced graph convolutional networks (GCNs) [52], which offer a broad scope of research in skeleton-based gait recognition. Several recent studies [53–56] have already demonstrated the effectiveness of GCNs, achieving superior accuracy in gait recognition tasks. For example, Teepe et al. [53] proposed GaitGraph, which leverages GCNs to extract robust spatio-temporal features from skeleton data and demonstrated superior accuracy on the CASIA-B dataset. Later, they introduced GaitGraph2 [54], which employed multi-branch GCNs and residual networks to extract features and achieved state-of-the-art (SOTA) recognition performance. In addition, Ray et al. [56] proposed a fusion-based multi-modal framework that employed OpenPose, AlphaPose, and HRNet human pose estimation algorithms, and they utilized a residual graph convolutional network (ResGCN) [57] for feature extraction. Model-based approaches, in particular, offer advantages such as robustness against clothing variations, covariates, and cluttered backgrounds by focusing on the underlying body structure. However, they typically require depth sensors or high-resolution image data, and pose estimation errors can degrade recognition performance, especially in uncontrolled environments.

2.2. Appearance-Based Approaches

Appearance-based gait recognition approaches [58–62] acquire gait features from silhouettes or RGB images, which are then used to identify individuals. These approaches are further divided into template-based and sequence-based approaches. Template-based approaches focus on converting a binary gait sequence into a single template image that contains spatiotemporal features. For example, Han and Bhanu [58] proposed a spatiotemporal gait representation known as the Gait Energy Image (GEI), in which a gait cycle is first extracted from a height-normalized silhouette sequence, and the silhouettes are then averaged over time to obtain a single GEI. Some examples are shown in Figure 1. Subsequently, GEIs were extended in multiple ways, notably through methods such as the gait history image (GHI) [59], chrono gait image (CGI) [61], gait entropy image (GEnI) [60], and gait flow image (GFI) [62]. More specifically, GHI [59] captures spatial and temporal features across quarter-cycle gait phases, while GEnI [60] encodes an entire gait cycle into a single template image by computing entropy, demonstrating effectiveness under diverse covariate conditions.

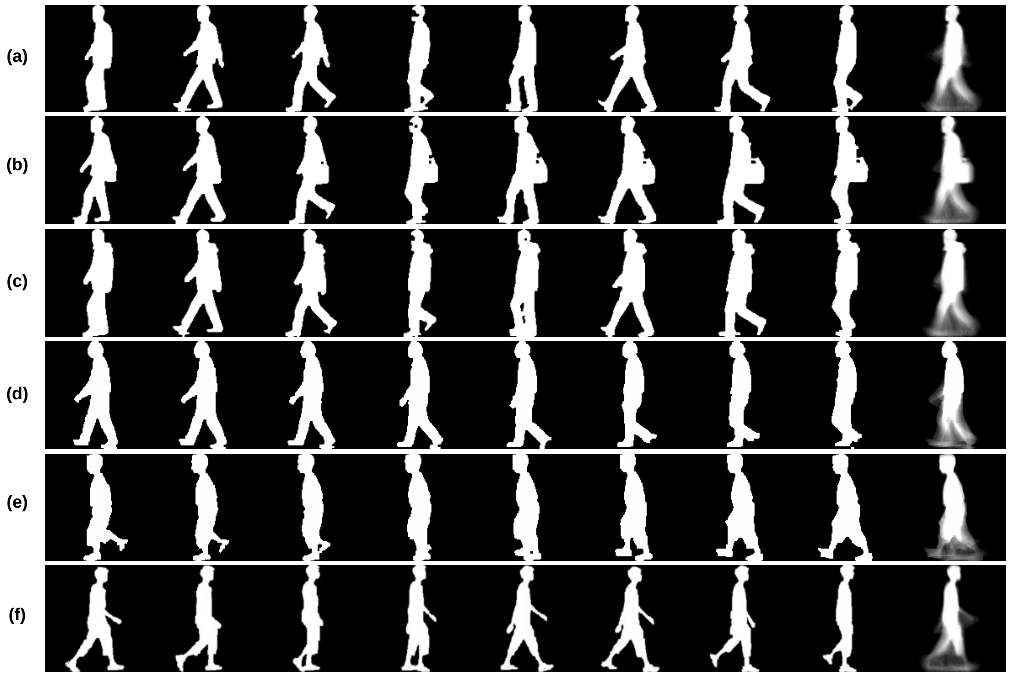


Figure 1. Example of gait silhouette sequences (every third frame of a sequence), with the rightmost image representing the corresponding gait energy image (GEI). Rows (**a**–**c**) are from the CASIA-B dataset, where (**a**) represents normal walking (NM), (**b**) represents walking with a bag (BG), and (**c**) represents walking with a coat (CL). Rows (**d**,**e**) illustrate the DB*high* and DB*low* walking sequences from the OU-ISIR dataset, while the last row (**f**) showcases a walking sequence from the OU-LP dataset.

Furthermore, deep learning (DL)-based approaches have employed these template-based images for gait recognition [63–65]. For example, Xu et al. [63] introduced the Deep Large Margin Nearest Neighbor (DLMNN) approach, which integrates a CNN with nearest neighbor (NN) algorithms, demonstrating competitive performance. Junaid et al. [64] proposed a customized ten-layer CNN that is less susceptible to occlusions and achieved superior gait recognition accuracy. Later, Suthar et al. [65] employed a pre-trained lightweight CNN model (i.e., MobileNetV3Small) that takes GEIs as input, extracts more refined features, and uses various machine learning methods for gait recognition. In particular, processing a single template image requires less storage space and lower computational power. However, such approaches involve certain disadvantages, including limited performance in real-world covariates conditions such as viewpoint changes, clothing variations, or walking speed, due to the reliance on a single model.

Recently, sequence-based approaches [5,20,21,66,67] have been employed for gait recognition, as they can effectively extract spatiotemporal features. For example, GaitSet [20] treated the silhouette sequence as a set and used multiple blocks of 2D CNNs and 2D max-pooling layers to extract spatiotemporal features. These features were then passed through a horizontal pyramid mapping mechanism that splits them horizontally to obtain a stripe-based feature representation. Fan et al. [5] proposed GaitPart, which initially utilized a part-based frame-level feature extractor to split the silhouette into several parts and subsequently employed a micromotion capture module to obtain local spatiotemporal features. These approaches predominantly suffer from limitations in obtaining global representations because they consider only horizontal partition features. Conversely, Chen and Li [66] proposed a dual-branch network, where one branch extracts multi-granularity features from both local and global stances using a selective horizontal pyramid convolutional network, while the other branch systematically investigates correlations between neighboring silhouettes at both pixel and block levels to derive temporal features. Additionally, Uddin et al. [67] proposed a framework combining global, horizontal, and vertical part-based feature extractors in two different pipelines: one employing 3D CNNs and 3D max-pooling layers, and the other consisting of 2D CNNs and 2D max-pooling layers, with all features concatenated together. In particular, sequence-based approaches are gaining popularity due to their effectiveness in 2D and 3D silhouette-based representations. However, most sequence-based approaches require high computational resources, are vulnerable to occlusion and viewpoint variations, and are sensitive to temporal misalignment in gait sequences. In addition, they often require complex training strategies and large datasets. This motivates us to introduce a hybrid fusion-based framework that utilizes GEI-based template images, requiring fewer computational resources and lower-resolution data and offering greater robustness to covariates.

3. Proposed Framework

This paper presents a multi-model fusion-based framework for recognizing individuals based on their gait features. Initially, we obtain the gait energy image (GEI) from a gait cycle of a normalized silhouette sequence by averaging the silhouettes. Subsequently, five state-of-the-art deep learning models, including VGG-16 [27], ResNet-50 [28], GoogLeNet [25], EfficientNet-B0 [29], and the vision transformer (ViT) [31], are fine-tuned and employed to extract intricate spatiotemporal features that are crucial for recognition. More specifically, we utilize three distinct fusion techniques: decision-level fusion (DLF), feature-level fusion (FLF), and hybrid fusion (HF). In DLF, each model first generates predictions for the extracted features, followed by majority voting to determine the final decision. In contrast, FLF combines each model's intermediate features and adds them point-wise to make the final prediction. HF combines both strategies, where pairs of models are first fused at the feature level, followed by decision-level fusion. The proposed framework is illustrated in Figure 2.

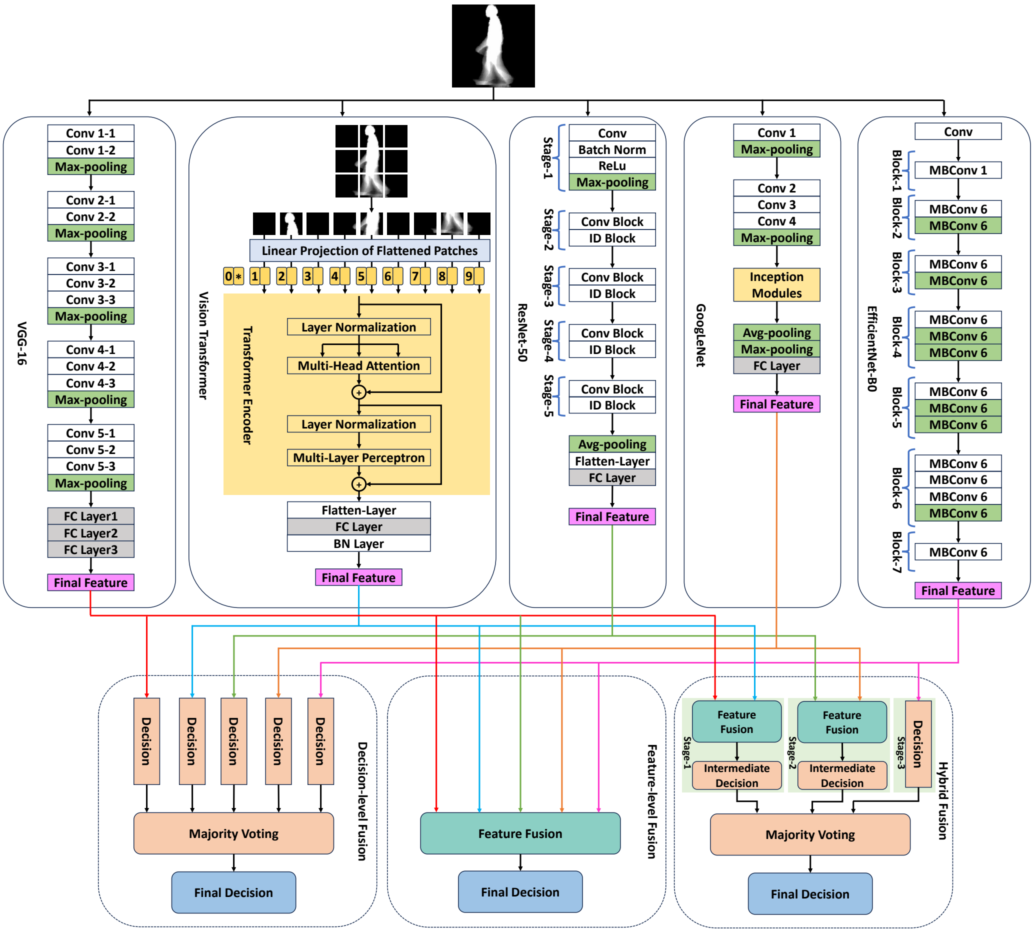


Figure 2. Overview of the proposed multi-model fusion-based gait recognition framework, which consists of three different fusion approaches: decision-level (bottom left), feature-level (bottom mid), and hybrid (bottom right). In this framework, majority voting selects the class as the final decision based on the maximum votes from individual models' decisions. Feature fusion aggregates the final features from multiple models through point-wise addition. The red lines represent the output feature of VGG-16, blue denotes the vision transformer, green denotes ResNet-50, orange denotes GoogLeNet, and pink denotes EfficientNet-B0.

3.1. Gait Energy Image

The gait energy image (GEI) [58] is widely featured for gait recognition models because it effectively retains the spatiotemporal features essential for individual identification. A GEI is obtained by averaging the normalized height and width silhouettes over the corresponding gait cycle, where a gait cycle [68] is calculated from the first contact of one foot with the ground to the next instance of contact with the same foot, encompassing all movement phases during locomotion. GEIs are grayscale images, and they require significantly fewer computational resources compared to RGB images. For a background-subtracted, normalized binary silhouette sequence, the GEI can be computed using the following formula:

$$G = \frac{1}{N}\sum_{t=1}^{N} S_t \tag{1}$$

where N is the total number of binary images in a gait cycle, S_t denotes the silhouette images at time t, and G denotes the generated GEI image. An example of silhouette sequences and their corresponding GEIs is shown in Figure 1.

3.2. Framework Architecture

The proposed multi-model fusion framework aims to improve the accuracy of gait recognition. This framework comprises four models from the CNN family, including VGG-16, ResNet-50, GoogLeNet, EfficientNet-B0, and an attention-based model, ViT. These models are sufficiently trained to demonstrate a correlation between extracted gait features and respective class labels.

VGG-16 [27] is the first model utilized in the proposed framework. Although similar to other CNN-based models, such as AlexNet [22], it features separate convolution layers and different kernel sizes (i.e., 3 × 3). It has thirteen convolutional and three fully connected layers (FCs), comprising a total of 16 weight layers. This network architecture is well known for its effectiveness and simplicity, consisting of 3 × 3 filters throughout the entire architecture. Despite its simplicity, VGG-16 has 138M parameters and performs remarkably well on large-scale datasets for image recognition [27]. Due to its uniform design, featuring all fixed kernel size convolution layers, it offers the advantage of capturing complex hierarchical features, making it versatile for transfer learning. It utilizes an activation function (i.e., rectified linear unit (ReLU)) for its convolutional layers to reduce the risk of vanishing gradient problems. In our work, the input data (i.e., GEI) can be defined as $x \in \mathbb{R}^{H \times W \times C}$, where (H, W) represent the image size, and C represents the total number of channels, respectively. Then, x is fed through the VGG-16 model that gives logits (i.e., unnormalized scores for each class) as follows:

$$y_{VGG}^{logits} = VGG(x) \tag{2}$$

where y_{VGG}^{logits} represents the logits of the image. Later, the argmax function is used to locate the maximum probability of a class whose output can be represented as y_{VGG}^{final}.

ResNet-50 [28] is a deep CNN comprising fifty layers, with the advantage of bottleneck residual blocks, which were the basic version earlier (i.e., ResNet-18 and ResNet-34). The architecture consists of a sequence of residual blocks that contain skip connections, which bypass one or more layers to learn residual mappings instead of explicit mappings. The advantage here is that it delivers proper training on the more complex network by retaining the gradient flow during backpropagation for updating weights. ResNet-50 is composed of convolutional layers, batch normalization, and ReLU activation, organized into 16 bottleneck blocks. For gait image data, with a model, due to its efficacy, compactness, and computational cost, it provides a tremendous performance in a recognition where input x was fed through the model (i.e., ResNet-50), and it gives the logits feature as follows:

$$y_{ResNet}^{logits} = ResNet(x) \tag{3}$$

and then the argmax function is applied to yield the final output (i.e., y_{ResNet}^{final}).

GoogLeNet [25] is another remarkable architecture in the DL family with a more profound architecture that is famous for its efficiency and high performance. The overall network consists of twenty-two layers in total. What makes it different from other conventional models is that the inception module allows it to extract more spatial features at multiple scales simultaneously. Each inception module is a key innovation of GoogLeNet and consists of multiple filters of various sizes (e.g., 1 × 1, 3 × 3, and 5 × 5), as well as max pooling and average pooling layers. The network simultaneously captures features at different spatial resolutions, as these filters are performed in parallel inside the same layer. Later, instead of an FC layer, global average pooling was performed after the last inception module to reduce the spatial dimensions, which is a key factor in reducing the total number of parameters (i.e., approximately 5M) and addressing overfitting issues. In this work, we fed the GEI (i.e., x) through the GoogLeNet model to obtain the logits feature, as shown in the following equation:

$$y_{GNet}^{logits} = GNet(x) \tag{4}$$

where $GNet(.)$ is the fine-tuned GoogLeNet model, and finally, a simple argmax function is employed for the final result, y_{GNet}^{final}.

EfficientNet-B0 [29] is the first model in the EfficientNet group to achieve a high accuracy percentage with a moderate number of parameters and computational cost. Its fundamental innovation is the compound scaling strategy, which simultaneously increases the network's depth, width, and resolution, as opposed to scaling them incrementally. This architecture features Mobile Inverted Bottleneck Convolutions (MBConv), which incorporate depthwise separable convolutions to reduce computational load. First, the input image is fed through a simple convolutional layer and then a series of MBConv blocks, and finally, it ends with a global average pooling layer. With only about 5.3M parameters, this model further improves its learning ability with a swish activation function and a global average pooling layer. The logic behind selecting this model is its accuracy, efficiency, and scalability. For the input GEI image x, the EfficientNet-B0 model gives the final feature (i.e., logits) as follows:

$$y_{ENet}^{logits} = ENet(x) \tag{5}$$

where $ENet(.)$ is the fine-tuned EfficientNet-B0 model, and finally, a simple argmax function is utilized for the final outcome, y_{ENet}^{final}.

Vision Transformer (ViT) [31] is the fifth model incorporated into this study. The ViT architecture comprises three fundamental components: an embedding layer, a transformer encoder, and a multi-layer perceptron (MLP). As shown in Figure 2, input image $x \in \mathbb{R}^{H \times W \times C}$ was divided into 2D patches with a (P, P) resolution, where $N = \frac{HW}{P^2}$ is the total number of patches, and H, W, and C denote the height, width, and number of channels, respectively. Later, patches are flattened and mapped into dimensions D with a trainable linear projection that can be defined as follows:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}} \tag{6}$$

where $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ and $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ and x_p^n is the n-th image patch and $n \in \{1, 2, 3..., N\}$. After that, the embedded images are fed through the transformer encoder.

The transformer encoder consists of L identical encoder blocks with two sub-layers in each encoder block, namely MSA and MLP. The ℓ-th encoder layer takes the input sequence from the previous layers, $\mathbf{z}_{\ell-1}$. First, layer normalization (LN) is performed, and then the normalized features are fed through the MSA layer, as shown in the following formulas:

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1, \ldots, L \tag{7}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1, \ldots, L \tag{8}$$

The encoder's last layer selects the first token in the sequence (i.e., z_L^0) and uses LN to create the picture representation r. The final recognition is performed by feeding r via a minute MLP head, a single hidden layer with a sigmoid function. The logits y_{ViT}^{logits} are the raw, unnormalized scores for each class, obtained through the following equation:

$$y_{ViT}^{logits} = \text{LN}(z_L^0) \tag{9}$$

Here, y_{ViT}^{logits} represents the logits (i.e., final feature), the raw output vector of class scores. These logits are then passed through the MLP for further refinement, and the final results (i.e., y_{ViT}^{final}) are produced.

Feature-level fusion (FLF) is an approach that utilizes multiple models' final extracted features to enhance robustness and improve recognition accuracy by integrating features through point-wise addition or concatenation. In our proposed framework, we conduct point-wise addition of final features (i.e., logits) extracted from all five models: VGG-16 [27], ViT [31], ResNet-50 [28], GoogLeNet [25], and EfficientNet-B0 [29]. The logits of these models are y_{VGG}^{logits}, y_{ViT}^{logits}, y_{ResNet}^{logits}, y_{GNet}^{logits}, y_{ENet}^{logits} from VGG-16 [27], ViT [31], ResNet-50 [28], GoogLeNet [25], and EfficientNet-B0 [29], respectively. As every model involves separate output feature dimensions, we adjusted the final layers of all five models to generate similar output dimensions, ensuring dimensional consistency across various DL model architectures. This same output dimension enabled fusion, eliminating the need for additional projection layers and ensuring smooth fusion processes while preserving each model's feature extraction capabilities. The FLF for gait recognition can be described as follows:

$$y_{FLF}^{logits} = P_{add}(y_{VGG}^{logits}, y_{ViT}^{logits}, y_{ResNet}^{logits}, y_{GNet}^{logits}, y_{ENet}^{logits}) \tag{10}$$

where $P_{add}(.)$ denotes the point-wise addition of the final features from y_{VGG}^{logits}, y_{ViT}^{logits}, y_{ResNet}^{logits}, y_{GNet}^{logits}, and y_{ENet}^{logits}. In addition, y_{FLF}^{logits} denotes fused features, and the final output from this approach is y_{FLF}^{final}.

Decision-level fusion (DLF) represents a sophisticated multi-model fusion strategy that integrates the outputs from various classifiers or models. This method leverages their distinct decisions to yield an ultimate prediction through weighted or majority voting mechanisms. In our proposed framework, we utilize the final output of each of the five fine-tuned classifier models namely, y_{VGG}^{final} from VGG-16 [27], y_{ResNet}^{final} from ResNet-50 [28], y_{GNet}^{final} from GoogLeNet [25], y_{ENet}^{final} from EfficientNet-B0 [29], and y_{ViT}^{final} from ViT [31]. The DLF can be expressed as follows:

$$y_{DLF}^{final} = MV(y_{VGG}^{final}, y_{ViT}^{final}, y_{ResNet}^{final}, y_{GNet}^{final}, y_{ENet}^{final}) \tag{11}$$

where $MV(.)$ represents majority voting, which takes y_{VGG}^{final}, y_{ViT}^{final}, y_{ResNet}^{final}, y_{GNet}^{final}, y_{ENet}^{final} as inputs and provides the final decision as y_{DLF}^{final}.

Hybrid Fusion (HF) represents an advanced approach that integrates DLF and FLF to mitigate their limitations, resulting in improved robustness and accuracy for applications such as human gait recognition, image classification, and biometric identification. Hybrid fusion operates in two phases. First, FLF integrates the salient features extracted from multiple sources into a unified representation. These fused features are then processed through classification or recognition pipelines. Subsequently, DLF combines the predictions from multiple classifiers trained on the fused feature set. This method improves robustness to input variations by integrating decisions from many models, hence ensuring greater resilience to noise and outliers.

In our work, we employed two phases to obtain the final features from HF. Initially, phase-1 consists of three stages; in Stage 1, we fused features from VGG-16 [27] and ViT [31] to acquire y_{FLF1}^{logits}, and in Stage 2 ResNet-50 [28] and GoogLeNet [25] to acquire y_{FLF2}^{logits} and then performed recognition to obtain the output decision from y_{FLF1}^{logits} of $D1$ and from y_{FLF2}^{logits} of $D2$. In Stage 3, we evaluated the prominent output of EfficientNet-B0 [29] and its recognition result (i.e., y_{ENet}^{final}), which we denote as $D3$. Finally, in phase-2, we utilized DLF with these decisions (i.e., $D1$, $D2$, and $D3$) to derive the final output as y_{HF}^{final} as follows:

$$y_{FLF1}^{logits} = P_{add}(y_{VGG}^{logits}, y_{ViT}^{logits}) \tag{12}$$

$$y_{FLF2}^{logits} = P_{add}(y_{ResNet}^{logits}, y_{GNet}^{logits}) \tag{13}$$

$$y_{HF}^{final} = MV(D1, D2, D3) \tag{14}$$

where $P_{add}(.)$ performs the pointwise addition, $MV(.)$ performs majority voting, and y_{HF}^{final} is the final output from HF.

3.3. Loss Function

We employed the cross-entropy loss [69] in our multi-model fusion framework to ensure the proper training of each model. Unlike regression-based losses such as mean squared error (MSE), cross-entropy is explicitly designed for probabilistic models, making it especially efficacious in multi-class recognition issues. This loss estimates the distance between the predicted probability distribution and the original distribution, thereby promoting a sharper alignment of the predicted probabilities with the ground-truth labels. When y is the actual label, and $\hat{y}$ is the predicted label, and then for multi-class recognition, the cross-entropy loss can be defined as follows:

$$CS_{loss} = -\sum_{i=1}^{T} y_i \log(\hat{y}_i) \tag{15}$$

Here, CS_{loss} denotes the cross-entropy loss, T indicates the total number of classes, and $\hat{y}_i$ denotes the predicted probability for class i. The cross-entropy loss rises when the predicted probability deviates from the actual class label. This promotes the model to allocate greater probabilities to the accurate class while minimizing probabilities for the incorrect ones.

4. Experiments and Discussions

4.1. Datasets

CASIA-B [43] is one of the widely used, publicly available gait databases. It consists of 124 subjects, 93 females and 31 males. It comprises 11 viewing angles with 18° intervals from 000° to 180°. There are ten sequences per subject in this dataset: six of them are normal walking sequences (i.e., NM #1-6), two of them carrying a bag (i.e., BG #1-2), and the remaining two wearing a coat (i.e., CL #1-2). This database has 124 × 11 × 10 = 13,640 sequences, each comprising approximately 75 frames on average. Among these frames, for each sequence, a single gait cycle is considered to generate a GEI, which means there are 13,640 GEIs. During the experiments, the dataset is divided into three subsets: 80% for training, 10% for validation, and 10% for testing, with the input normalized GEI image size set to 128 × 128.

OU-ISIR D [44] consists of 185 subjects with 370 sequences observed from the side view. This dataset investigates fluctuations in gait throughout several periods, specifically how gait silhouettes of a similar phase alter throughout a sequence of periods. Based on normalized autocorrelation (NAC), these gait sequences are clustered in two separate groups: DB$_{low}$ and DB$_{high}$. Both groups comprise 100 subjects; DB$_{low}$ has a fluctuating walking gait sequence, whereas DB$_{high}$ has a stable walking sequence.

OU-LP [45] is one of the more extensive gait databases, comprising 4016 subjects ranging in age from 1 to 94 years. The dataset was collected by the Institute of Scientific and Industrial Research (ISIR), Osaka University (OU), Japan. In our experiment, we utilized the latest version of the OU-LP dataset (i.e., Version 2), where each subject has two sequences, designated as A and B, and each sequence comprises four distinct observation angles: 55°, 65°, 75°, and 85°.

4.2. Training Parameters and Test

We considered all the pre-trained versions of VGG-16 [27], ResNet-50 [28], GoogLeNet [25], EfficientNet-B0 [29], and ViT [31] that were trained on ImageNet dataset. Later, we fine-tuned these models on the three datasets (i.e., CASIA-B, OU-ISIR D, and OU-LP). Moreover, we trained all the models for 50 epochs with early stopping options on the best validation accuracy, and patience was set to 5 epochs. The input GEI dimensions were standardized at 128 × 128 for all models to ensure an impartial evaluation, with a batch size of 32 and a learning rate of 0.001, utilizing the Adam optimizer.

Our work utilized four separate GEI datasets, which mostly come with 128 × 128 resolution, and the pre-trained models' default input requirements (224 × 224) [70]. We implemented image resizing through a custom collate function, rather than modifying the model architectures. Specifically, we employed bilinear interpolation to upscale GEI images from 128 × 128 to 224 × 224. This approach preserves the original pre-trained weights and architectural integrity of all models (VGG-16, ResNet-50, ViT, EfficientNet-B0, and GoogLeNet) while ensuring compatibility with our dataset. The resizing operation is performed dynamically during data loading, maintaining the aspect ratio and ensuring consistent input dimensions across all models.

The experiments were conducted on a machine equipped with an NVIDIA GeForce RTX 4090 GPU, running Ubuntu 24.04 and utilizing Python version 3.10.15. Data preprocessing, including silhouette extraction and GEI generation, required approximately 1.5 h for all the datasets. Model training was conducted over 50 epochs, with each epoch taking approximately 2 min, resulting in a total training time of around 1.67 h. The inference time per sample was approximately 25 ms. We evaluated the effectiveness of our proposed framework and other existing models using the Rank-1 identification rate and the receiver operating characteristic (ROC) curve, which indicates the trade-off between the false rejection rate (FRR) of genuine users and the false acceptance rate (FAR) of imposters at varying thresholds. We also computed the equal error rate (EER), which represents the point at which the FAR and FRR are equal and serves as a compact measure of overall performance [6]. Moreover, we assessed additional classification metrics, including accuracy, precision, recall, and F1-score [71].

4.3. Experimental Results

4.3.1. CASIA-B

The Rank-1 identification rate and equal error rate (EER) on the CASIA-B dataset are shown in Tables 1 and 2, with the ROC curve presented in Figure 3a. ViT achieved a Rank-1 rate of 46.81% and an EER of 12.82%, as it sometimes overlooks subtle gait cues, such as leg swing differences or shoulder tilt, which can impact matching with the gallery data. In contrast, GoogLeNet and EfficientNet-B0 achieved comparable Rank-1 rates of 53.62% and 69.06%, with EERs of 48.72% and 30.43%, respectively. VGG-16 and ResNet-50 obtained Rank-1 rates of 62.83% and 89.76%, with EERs of 46.20% and 3.78%, respectively, outperforming other individual deep learning methods. Our proposed DLF, FLF, and HF approaches achieved Rank-1 rates of 90.16%, 90.45%, and 92.07%, with corresponding EERs of 3.02%, 3.68%, and 2.73%, significantly outperforming all standalone DL-based methods. As HF combines both DLF and FLF strategies, it leverages the strengths of larger models, such as VGG-16 and ResNet-50, which are crucial for improving recognition performance and reducing error rates.

Table 1. Average Rank-1 identification rate (%) on CASIA-B, OU-LP, OU-ISIR DB$_{low}$, and OU-ISIR DB$_{high}$ datasets with 128 × 128 resolution of height-normalized GEIs.

Models	CASIA-B	OU-LP	OU-ISIR DB$_{low}$	OU-ISIR DB$_{high}$
VGG-16	62.83	47.34	66.50	49.00
ViT	46.81	69.88	42.50	65.00
ResNet-50	89.76	59.96	73.00	75.00
GoogLeNet	53.62	63.44	51.00	52.00
EfficientNet-B0	69.06	64.34	63.00	64.00
DLF (ours)	90.16	71.55	86.50	91.00
FLF (ours)	90.45	77.80	93.00	90.00
Hybrid (ours)	92.07	87.14	93.50	93.00

Table 2. Equal error rate (EER) (%) on CASIA-B, OU-LP, OU-ISIR DB$_{low}$, and OU-ISIR DB$_{high}$ datasets.

Models	CASIA-B	OU-LP	OU-ISIR DB$_{low}$	OU-ISIR DB$_{high}$
VGG-16	46.20	50.14	47.95	48.02
ViT	12.82	6.79	12.95	13.88
ResNet-50	3.78	10.71	5.13	6.01
GoogLeNet	48.72	8.56	50.46	46.04
EfficientNet-B0	30.43	6.97	37.90	32.47
DLF (ours)	3.02	6.17	3.88	1.92
FLF (ours)	3.68	5.49	1.62	1.18
Hybrid (ours)	2.73	4.08	1.94	1.41

The Rank-1 identification rate for each separate view angle shows significant improvement with our proposed methods. As shown in Table 3, ViT, which struggled with subtle gait cues, achieved the lowest accuracy of 46.81%, with performance varying widely across angles. In contrast, our proposed approaches DLF, FLF, and HF demonstrated superior performance. The DLF approach achieved a mean recognition accuracy of 90.16%, with impressive results at extreme angles, such as 90.32% at 0° and 84.68% at 180°. The FLF approach further improved, reaching a mean accuracy of 90.45%, with consistently high recognition across all angles, including 99.19% at 0°. The HF approach, combining both DLF and FLF strategies, achieved the highest mean accuracy of 92.07%, demonstrating the effectiveness of our fusion-based framework. Overall, our methods significantly outperformed traditional models like VGG-16, ResNet-50, and EfficientNet-B0, which had lower accuracies, particularly at challenging angles such as 72° and 108°, where their performances were noticeably inconsistent.

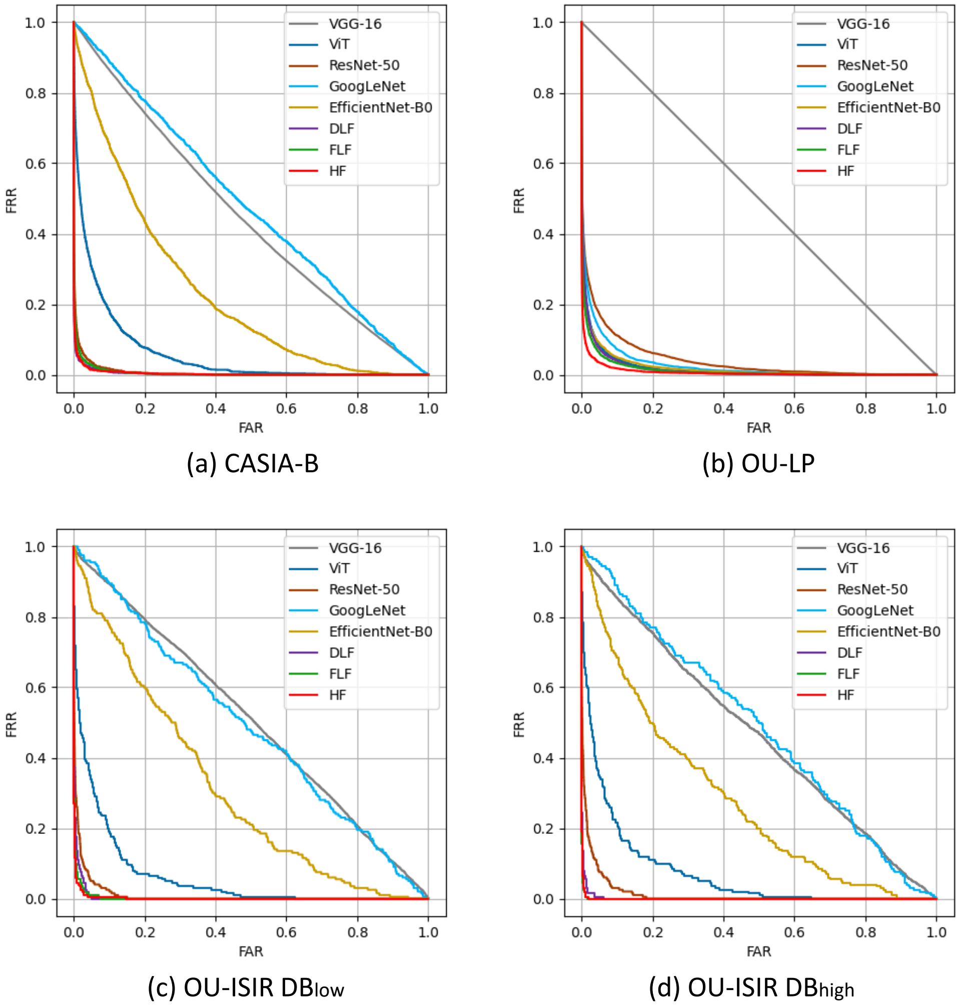
Table 3. Rank-1 identification rate (%) on the CASIA-B dataset with 128 × 128 resolution height-normalized GEIs across all viewing angles.

Models	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
VGG-16	56.10	58.87	53.23	50.00	73.17	75.81	65.85	63.71	57.26	70.97	66.13	62.83
ViT	67.48	50.00	39.52	36.29	44.72	30.65	43.09	45.97	41.13	53.23	62.90	46.81
ResNet-50	88.37	95.16	88.71	82.26	90.24	87.10	90.24	90.32	84.68	94.35	95.97	89.76
GoogLeNet	61.63	62.42	30.65	54.03	50.00	42.42	61.32	59.81	63.81	52.42	51.32	53.62
EfficientNet-B0	64.88	70.48	64.03	70.48	67.32	68.06	69.76	71.29	72.10	68.06	73.23	69.06
DLF (ours)	90.32	87.10	86.29	91.13	91.87	91.94	91.87	88.71	93.55	94.35	84.68	90.16
FLF (ours)	99.19	95.97	90.32	85.48	86.99	85.48	87.80	86.29	87.10	93.55	96.77	90.45
Hybrid (ours)	93.50	91.13	90.32	92.74	91.87	86.29	94.31	92.74	93.55	91.94	94.35	92.07

In addition, the evaluation results, including the metrics accuracy, precision, recall, and F1-score, are presented in Table 4 and Figure 4a. After fine-tuning, VGG-16 achieved 99.27% accuracy along with a 99.31% precision score. Conversely, ViT obtained an accuracy of 99.12% and a precision score of 99.22%, the lowest among the five models, likely due to insufficient inductive bias (e.g., lack of location and translation invariance). We observed improvements across all evaluation metrics for ResNet-50, which achieved 99.56%, 99.64%, 99.56%, and 99.55%, respectively, for accuracy, precision, recall, and F1-score. The lightweight GoogLeNet model obtained 99.34% accuracy, while EfficientNet-B0 achieved 99.63%, the highest among the five models. As EfficientNet-B0 uses MBConv blocks with depthwise separable convolutions, it requires less computational power while preserving representational capacity. In contrast, our proposed DLF attained 99.85% accuracy and 99.88% precision, exceeding all fine-tuned DL models. Furthermore, FLF and HF achieved perfect 100.00% scores across all four metrics, demonstrating that our proposed fusion-based framework can extract detailed spatiotemporal features. DLF demonstrated slightly lower accuracy than FLF and HF, as it primarily relies on majority decisions, rather than confidence scores; consequently, models like ViT and VGG-16, which obtained comparatively lower accuracy such as 99.12% and 99.27%, respectively, can negatively impact DLF performance.

Table 4. Evaluation scores on the CASIA-B dataset with 128 × 128 resolution of height-normalized GEIs.

Models	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]
VGG-16	99.27	99.31	99.27	99.24
ViT	99.12	99.22	99.12	99.10
ResNet-50	99.56	99.64	99.56	99.55
GoogLeNet	99.34	99.36	99.34	99.32
EfficientNet-B0	99.63	99.67	99.63	99.64
DLF (ours)	99.85	99.88	99.85	99.85
FLF (ours)	100.00	100.00	100.00	100.00
HF (ours)	100.00	100.00	100.00	100.00



(a) CASIA-B

(b) OU-LP

(c) OU-ISIR DB$_{\text{low}}$

(d) OU-ISIR DB$_{\text{high}}$

Figure 3. ROC curves for the proposed fusion framework and individual models on the CASIA-B, OU-LP OU-ISIR DB$_{low}$, and OU-ISIR DB$_{high}$. The curves plot the false rejection rate (FRR) against the false acceptance rate (FAR), demonstrating each system's ability to distinguish between genuine and imposter samples.

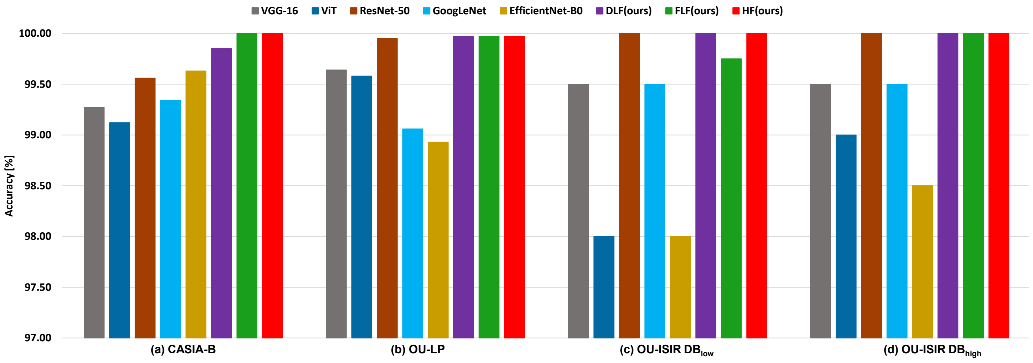


Figure 4. Comparison of the proposed fusion-based method with fine-tuned deep learning (DL) models' (i.e., VGG-16, ViT, ResNet-50, GoogLeNet, and EfficientNet-B0) performances on the CASIA-B, OU-ISIR DB$_{low}$ and DB$_{high}$, and OU-LP datasets.

4.3.2. OU-LP

The Rank-1 identification rate and EER on the OU-LP dataset are shown in Tables 1 and 2, along with the ROC curve in Figure 3b. We can observe that our proposed methods significantly outperform other methods. For example, VGG-16, ResNet-50, and EfficientNet-B0 achieved Rank-1 accuracies of 47.34%, 59.96%, and 64.34%, respectively, while ViT obtained 69.88%, the highest among these traditional models. In contrast, our DLF approach achieved a Rank-1 rate of 71.55%, and FLF and HF further improved accuracy, with a Rank-1 identification rate of 77.80% and 87.14%, respectively. In terms of EER, our methods also performed better than traditional models, with Hybrid achieving the lowest EER of 4.08%, outperforming all other models. FLF and DLF followed closely, achieving EERs of 5.49% and 6.17%, respectively. These results highlight the effectiveness of our proposed multi-model fusion strategies in improving both recognition accuracy and reducing error rates on the challenging OU-LP dataset.

In addition, the evaluation results, including the metrics accuracy, precision, recall, and F1-score, are presented in Table 5, and in Figure 4b, we can observe that VGG-16 obtained an accuracy of 99.64%, along with a precision of 99.45%. ViT works by splitting the images into patches and then using an encoder to extract the feature, and it obtained 99.58%, 99.40%, 99.58%, and 99.46% accuracy, precision, recall, and F-1 score, respectively. However, ResNet-50 obtained 99.95% accuracy and 99.93% F-1 score, making it superior to all the fine-tuned models. However, EfficientNet-B0 failed to properly obtain the feature in more extensive data and obtained only 98.93% accuracy. Another probable reason is that EfficientNet-B0 involves fewer parameters than others, and the OU-LP dataset has higher intra-class and lower inter-class variance, confusing the model in correctly predicting. Similarly, GoogLeNet achieved a 99.06% accuracy and a 98.60% precision score because of its lightweight design, which is insufficient for strong long-range feature modeling with larger datasets. In contrast, our proposed DLF, FLF, and HF methods all achieved 99.97% accuracy, 99.96% precision, 99.97% recall, and a 99.97% F1-score. This demonstrates that our fusion-based approaches consistently outperform individual models. Among the three fusion-based approaches, only HF showed stable results across all three datasets.

Table 5. Evaluation scores on the OU-LP dataset with 128 × 128 resolution of height-normalized GEIs.

Models	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]
VGG-16	99.64	99.45	99.64	99.51
ViT	99.58	99.40	99.58	99.46
ResNet-50	99.95	99.92	99.95	99.93
GoogLeNet	99.06	98.60	99.06	98.75
EfficientNet-B0	98.93	98.40	98.93	98.58
DLF (ours)	99.97	99.96	99.97	99.97
FLF (ours)	99.97	99.96	99.97	99.97
HF (ours)	99.97	99.96	99.97	99.97

4.3.3. OU-ISIR D

The Rank-1 identification rate and EER for the OU-ISIR DB$_{low}$ and DB$_{high}$ datasets are shown in Tables 1 and 2. As expected, all fine-tuned models performed better on DB$_{high}$, which contains more consistent walking sequences, compared to DB$_{low}$, with more varied walking patterns. For instance, VGG-16 achieved a Rank-1 rate of 49.00% on DB$_{high}$ and 66.50% on DB$_{low}$, while ViT obtained 65.00% on DB$_{high}$ and 42.50% on DB$_{low}$. ResNet-50 outperformed all other models with a Rank-1 rate of 75.00% on DB$_{high}$ and 73.00% on DB$_{low}$, showcasing its robustness with residual connections. EfficientNet-B0 and GoogLeNet showed lower performance on DB$_{low}$, with Rank-1 accuracies of 63.00% and 51.00%, respectively. Our proposed DLF, FLF, and HF methods achieved Rank-1 accuracies of 91.00%, 90.00%, and 93.00% on DB$_{high}$, and 86.50%, 93.00%, and 93.50% on DB$_{low}$, respectively, demonstrating the effectiveness and reliability of our fusion-based approaches. Among these, HF outperformed FLF, showing the most stability across all datasets.

In terms of other evaluation metrics such as accuracy, precision, recall, and F1-score, which are presented in Table 6 and Figure 4c,d, ResNet-50 achieved the highest performance across all metrics, with 100.00% accuracy, precision, recall, and F1-score on DB_{high}. VGG-16 also performed well, with an accuracy of 99.63% on DB_{high} but slightly lower results on DB_{low}. ViT showed improved performance on DB_{high}, achieving an accuracy of 99.00% compared to 98.00% on DB_{low}. In contrast, GoogLeNet and EfficientNet-B0 demonstrated lower performance on DB_{low}, as they struggled with the dataset's variability. Our fusion-based methods (DLF, FLF, and HF) consistently outperformed the individual models, achieving perfect scores of 100.00% across all metrics on DB_{high}, with HF providing the most stable results across both datasets. FLF showed a slightly reduced accuracy of 99.75% on DB_{low}, likely due to redundancy or conflicting features extracted via individual models, which may have affected its performance in the noisy dataset.

Table 6. Evaluation scores on the OU-ISIR DB_{low} and DB_{high} dataset with 128 × 128 resolution of height-normalized GEIs.

Models	OU-ISIR DB_{low}				OU-ISIR DB_{high}			
	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]
VGG-16	99.50	99.67	99.50	99.47	99.63	99.67	99.50	99.58
ViT	98.00	98.67	98.00	97.87	99.00	99.33	99.00	98.93
ResNet-50	99.67	99.61	99.83	99.47	100.00	100.00	100.00	100.00
GoogLeNet	99.50	99.67	99.50	99.47	99.50	99.67	99.50	99.47
EfficientNet-B0	98.00	98.67	98.00	97.87	98.50	99.00	98.50	98.40
DLF (ours)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
FLF (ours)	99.75	99.67	99.60	99.77	100.00	100.00	100.00	100.00
HF (ours)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

4.4. Comparison with Previous Studies

The performance of our proposed multi-model fusion-based approach is compared with existing DL-based methods, including GEINet [72], Deep CNN [73], CNN [74], CNN-2 [75], and Deep CNN-2 [76]. For a fair comparison, these methods were also trained under the same settings: 80% for training, 10% for validation, and 10% for testing. The input GEI image size was consistent across all models, set at 128 × 128 pixels for both height and width. The evaluation results, including accuracy, are presented in Table 7 and Figure 5. We can observe that GEINet [72] performs well, achieving an accuracy of 97.65%, while Deep CNN [73] obtained an accuracy of 25.68%, and Deep CNN-2 [76] achieved an accuracy of 86.17% on the CASIA-B dataset. Conversely, there is a significant improvement in CNN [74] and CNN-2 [75], which achieved accuracies of 98.09% and 94.63%, respectively. In contrast, our proposed DLF, FLF, and HF methods achieved the highest accuracies of 99.85%, 100.00%, and 100.00%, respectively. This means that DLF/FLF/HF surpass GEINet by 2.20%/2.35%/2.35%. We believe that the multi-model fusion-based approach can extract in-depth features and has the potential to handle multi-view and multi-sequence datasets like CASIA-B, which includes GEIs from 11 different viewing angles and 10 diverse sequences.

Regarding the comparison on the OU-LP dataset, the experimental results for the proposed multi-model fusion-based approach and existing methods are presented in Table 7 and Figure 5b. As shown, most existing methods struggle to extract accurate gait features, leading to lower accuracy compared to our proposed fusion-based approach. For example, Deep CNN [73] achieved an accuracy of 5.60%, while CNN and Deep CNN-2 [76] obtained 48.32% and 45.52%, respectively. The low accuracy can be attributed to several factors. First, the OU-LP dataset comprises 4016 subjects, which presents a significant large-scale classification challenge. Second, aggressive max pooling operations and relatively small feature maps (i.e., down from 136 × 136 to 5 × 5) likely result in a considerable loss of crucial spatial information, which is essential for capturing fine-grained details in gait recognition. Finally, the limited number of parameters (i.e., 20,932 trainable parameters) is insufficient for capturing the subtle inter-class differences among the 4016 subjects. In contrast, GEINet [72] achieved better accuracy than the other methods, attaining an accuracy of 90.74%. However, we can observe that our proposed fusion-based approaches, DLF, FLF, and HF, demonstrated a superior accuracy of 99.97%, highlighting that our fusion approaches can effectively extract intricate features from a large-scale OU-LP dataset.

In addition, we can see that GEINet [72] achieved a higher accuracy than both Deep CNN [73] and Deep CNN-2 [76], with an accuracy of 99.65% on the OU-ISIR DB$_{low}$ dataset. CNN [74] and CNN-2 [75] significantly improved accuracy, reaching 99.37% and 96.73%, respectively. However, our proposed approach marginally improved accuracy for both DLF and HF, achieving 100.00%. FLF obtained 99.75% accuracy due to the impact of feature fusion, as most models struggled to capture spatiotemporal features because the OU-ISIR DB$_{low}$ dataset contains fluctuating data. Moreover, for the OU-ISIR DB$_{high}$ dataset, our proposed DLF, FLF, and HF approaches surpassed GEINet [72], Deep CNN [73], and Deep CNN-2 [76] by 0.07%, 12.30%, and 3.82%, respectively. Furthermore, we can observe that the fusion-based approaches surpassed CNN [74] by 0.35% and CNN-2 [76] by 10.01%.

Table 7. Comparison of accuracy scores (%) on CASIA-B, OU-LP, OU-ISIR DB$_{low}$, and OU-ISIR DB$_{high}$ datasets with existing established deep learning-based approaches.

Models	**Accuracy [%]**			
	CASIA-B	**OU-LP**	**OU-ISIR DB$_{low}$**	**OU-ISIR DB$_{high}$**
GEINet [72]	97.65	90.74	99.65	99.93
Deep CNN [73]	25.68	5.60	83.81	87.70
CNN [74]	98.09	89.17	99.37	99.65
CNN-2 [75]	94.63	48.32	96.73	89.99
Deep CNN-2 [76]	86.17	45.52	95.21	96.18
DLF (ours)	99.85	99.97	100.00	100.00
FLF (ours)	100.00	99.97	99.75	100.00
HF (ours)	100.00	99.97	100.00	100.00

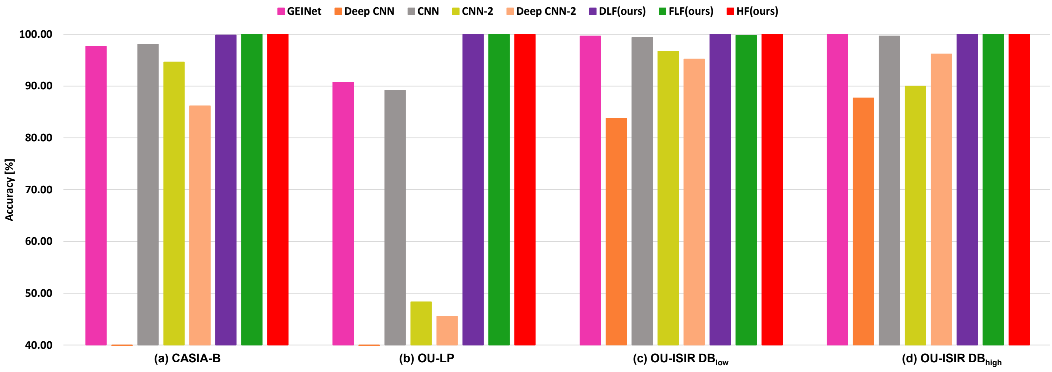


Figure 5. Comparison of the proposed fusion-based method with existing established deep learning (DL) models' performances on the CASIA-B, OU-ISIR DB$_{low}$ and DB$_{high}$, and OU-LP datasets.

4.5. Discussions

Existing fusion approaches, such as decision-level fusion (DLF), feature-level fusion (FLF), produce outputs based on simple majority voting, feature concatenation, or point-wise addition. In contrast, hybrid fusion (HF) introduces several key techniques that differentiate it from these traditional fusion strategies. HF combines the strengths of both DLF and FLF, minimizing the risk of bias that could affect the output. To the best of our knowledge, exploring a hierarchical fusion strategy for gait recognition has not been studied before; we propose a novel parallel three-stage fusion architecture. This architecture integrates FLF in the first stages, followed by DLF in the final stage, as shown in Figure 2. Moreover, our framework incorporates diverse architectural models, including five different DL-based models: VGG-16, ViT, ResNet-50, GoogLeNet, and EfficientNet-B0. Each model represents distinct feature representations, where VGG-16, ResNet-50, GoogLeNet, and EfficientNet-B0 represent convolutional-based features, while ViT focuses on attention-based features. As a result, we observe that, in most cases, the HF approach outperforms existing fusion techniques.

Regarding the CASIA-B and OU-LP datasets, the Rank-1 and EER, shown in Tables 1 and 2, demonstrate that HF outperforms all individual models. On the CASIA-B dataset, HF achieved an impressive Rank-1 rate of 90.67% and a low EER of 2.73%. Similarly, on the OU-LP dataset, HF achieved a perfect accuracy of 99.97% with an EER of 4.08%. Compared to models such as VGG-16 for CASIA-B and OU-LP, as well as ResNet-50, HF consistently provides better performance across all evaluation metrics. By combining DLF and FLF, HF minimises bias and captures more intricate features, making it ideal for large-scale, diverse datasets such as CASIA-B and OU-LP.

4.6. Ablation Study

This section presents ablation experiments on the CASIA-B dataset to determine the optimal number and combination of models for our proposed HF strategy.

4.6.1. Impact of Number of Models Used in Fusion

In the proposed multi-model fusion-based method, selecting the optimal number of models is crucial, as variations in the selected models can affect accuracy. As demonstrated in Table 8, when using two models, a single fusion approach (i.e., FLF) achieves an accuracy of 99.78%. When three models are used, both DLF and FLF are employed, resulting in an accuracy of 99.85%. With four models, both FLF and HF achieve an accuracy of 99.93%. Finally, when all five models are considered, both FLF and HF achieve perfect accuracy (i.e., 100.00%), while DLF achieves a comparable accuracy of 99.85%. This decrease in DLF accuracy is due to the influence of individual models making incorrect predictions, which, through majority voting, may lead to erroneous final predictions.

Table 8. Performance comparisons of the ablation study due to the impact of several models (i.e., in numbers) on our fusion approaches. Here, "-" denotes the absence of accuracy in that approach because of fulfilling the criteria for the correct number of models to apply that fusion approach.

Exp. No.	**No. of Models**	**Accuracy [%]**		
		DLF	**FLF**	**HF**
1	2	-	99.78	
2	3	99.85	99.85	-
3	4	-	99.93	99.93
4	5	99.85	100.00	100.00

4.6.2. Impact of Different Models Combination in Hybrid Fusion

Hybrid fusion consists of decision-level fusion (DLF) and feature-level fusion (FLF), where models initially perform FLF followed by DLF. Finally, all individual decisions are considered for the final prediction, similar to the majority voting mechanism. As shown in Table 9, we considered possible combinations of the selected models in HF, involving three stages. Specifically, stages 1 and 2 perform FLF and generate initial decisions (i.e., D1 and D2); in stage 3, a single model makes its decision (i.e., D3). Finally, all intermediate decisions are aggregated by DLF, with HF providing the final prediction. For the first three combinations, as shown in Table 9 (i.e., Exp. No. 1-3), HF achieved slightly lower accuracy, which are 99.85%, 99.78%, and 99.71%, respectively, due to weaker model combinations. However, HF attained an accuracy of 100.00% in the remaining experiments.

Table 9. Performance comparisons of the ablation study due to different combinations in hybrid fusion (HF). Here, Stage 1, Stage 2, and Stage 3 perform feature-level fusion and give decision (D1), feature-level fusion and give decision (D2), and single model evaluation (D3). In the Final Fusion (i.e., HF), decision-level fusion is employed on D1, D2, and D3.

Exp. No.	Stage 1 (FLF)	Stage 2 (FLF)	Stage 3 (Single Model)	Final Fusion (HF)	Accuracy [%]
1	VGG-16 + ViT	ResNet-50 + GoogLeNet)	EfficientNet-B0	DLF (D1, D2, D3)	99.85
2	VGG-16 + ResNet-50	ViT + GoogLeNet	EfficientNet-B0	DLF (D1, D2, D3)	99.78
3	VGG-16 + GoogLeNet	ViT + ResNet-50)	EfficientNet-B0	DLF (D1, D2, D3)	99.71
4	VGG-16 + EfficientNet-B0	ViT + ResNet-50	GoogLeNet	DLF (D1, D2, D3)	100.00
5	ViT + ResNet-50	VGG-16 + GoogLeNet	EfficientNet-B0	DLF (D1, D2, D3)	100.00
6	ViT + GoogLeNet	VGG-16 + ResNet-50	EfficientNet-B0	DLF (D1, D2, D3)	100.00
7	ViT + EfficientNet-B0	VGG-16 + ResNet-50	GoogLeNet	DLF (D1, D2, D3)	100.00
8	ResNet-50 + GoogLeNet	VGG-16 + ViT	EfficientNet-B0	DLF (D1, D2, D3)	100.00
9	ResNet-50 + EfficientNet-B0	VGG-16 + ViT	GoogLeNet	DLF (D1, D2, D3)	100.00
10	GoogLeNet + EfficientNet-B0	VGG-16 + ViT	ResNet-50	DLF (D1, D2, D3)	100.00

Regarding Exp. No. 4-10, we observe that our proposed HF achieved a perfect accuracy of 100.00%. These combinations paired a CNN and a ViT-based model in both Stage 1 and Stage 2 for feature-level fusion, ensuring robust local fine-grained feature extraction, as well as global context modeling and long-range dependencies due to sequential feature representation using the attention-based model ViT. Moreover, including lightweight models like EfficientNet-B0 or GoogLeNet in Stage 3 ensured efficiency and architectural diversity for the final ensemble. As a result, these experiments (Exp. No. 4–10) consistently achieved an accuracy of 100.00%, validating both the theoretical hypothesis and the empirical performance gains.

5. Conclusions

This paper has proposed a multi-model fusion-based gait recognition framework that leverages the strengths of multiple state-of-the-art deep learning models. Specifically, we employed VGG-16, ResNet-50, ViT, GoogLeNet, and EfficientNet-B0, each contributing unique characteristics to the framework. Four of these models (e.g., VGG-16, ResNet-50, GoogLeNet, and EfficientNet-B0) belong to the family of convolutional neural networks (CNNs), while ViT is an attention-based architecture.

To enhance gait recognition accuracy, we introduced three separate fusion approaches: decision-level fusion, feature-level fusion, and hybrid fusion. Decision-level fusion operates by applying majority voting in an ensemble manner, considering the predictions from all five models. Feature-level fusion fused the feature representations extracted by each model to form a comprehensive feature representation for recognition. Hybrid fusion combines the benefits of both decision-level and feature-level fusion, resulting in a robust framework that effectively exploits the complementary strengths of the models. The framework leverages the strengths of multiple models by incorporating feature and decision-level information, resulting in superior accuracy across different benchmarks.

Author Contributions: K.H.: Conceptualization, methodology, software, visualization, and writing—review and editing; K.A.T.: conceptualization, methodology, software, visualization, and writing—review and editing; M.R.I.B.: dataset processing, software, visualization, and review and editing; M.S.U.D.: dataset processing, software, visualization, and writing—review and editing; M.A.A.: review, editing, and supervision; M.A.R.A.: review, editing, and supervision; M.Z.U.: writing—review and editing and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: We are grateful to Begum Rokeya University, Rangpur, for partially supporting this work.

Data Availability Statement: The data and source code will be made available upon request.

Acknowledgments: We express our gratitude to Lab Semantics for supplying computational resources that substantially aided in the completion of this research.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

References

1. Jain, A.; Bolle, R.; Pankanti, S. *Introduction to Biometrics*; Springer: Berlin/Heidelberg, Germany, 1996.
2. Kak, S.F.; Mustafa, F.M.; Valente, P. A review of person recognition based on face model. *Eurasian J. Sci. Eng.* **2018**, *4*, 157–168. [CrossRef]
3. Besbes, F.; Trichili, H.; Solaiman, B. Multimodal biometric system based on fingerprint identification and iris recognition. In Proceedings of the 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, Syria, 7–11 April 2008; pp. 1–5.
4. Bachoo, A.K.; Tapamo, J.R. A segmentation method to improve iris-based person identification. In Proceedings of the 2004 IEEE Africon, 7th Africon Conference in Africa (IEEE Cat. No. 04CH37590), Gaborone, Botswana, 15–17 September 2004; Volume 1, pp. 403–408.
5. Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; He, Z. Gaitpart: Temporal part-based model for gait recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14225–14233.
6. Uddin, M.Z.; Muramatsu, D.; Kimura, T.; Makihara, Y.; Yagi, Y. MultiQ: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation. *IPSJ Trans. Comput. Vis. Appl.* **2017**, *9*, 18. [CrossRef]
7. Makihara, Y.; Nixon, M.S.; Yagi, Y. Gait recognition: Databases, representations, and applications. In *Computer Vision: A Reference Guide*; Springer: Cham, Switzerland, 2020; pp. 1–13.
8. Bouchrika, I.; Goffredo, M.; Carter, J.; Nixon, M. On using gait in forensic biometrics. *J. Forensic Sci.* **2011**, *56*, 882–889. [CrossRef] [PubMed]
9. Iwama, H.; Muramatsu, D.; Makihara, Y.; Yagi, Y. Gait verification system for criminal investigation. *Inf. Media Technol.* **2013**, *8*, 1187–1199. [CrossRef]
10. Lynnerup, N.; Larsen, P.K. Gait as evidence. *IET Biom.* **2014**, *3*, 47–54. [CrossRef]
11. Uddin, M.; Ngo, T.T.; Makihara, Y.; Takemura, N.; Li, X.; Muramatsu, D.; Yagi, Y. The ou-isir large population gait database with real-life carried object and its performance evaluation. *IPSJ Trans. Comput. Vis. Appl.* **2018**, *10*, 5. [CrossRef]
12. Hossain, M.A.; Makihara, Y.; Wang, J.; Yagi, Y. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognit.* **2010**, *43*, 2281–2291. [CrossRef]
13. Muramatsu, D.; Shiraishi, A.; Makihara, Y.; Uddin, M.Z.; Yagi, Y. Gait-based person recognition using arbitrary view transformation model. *IEEE Trans. Image Process.* **2014**, *24*, 140–154. [CrossRef] [PubMed]
14. Uddin, M.; Muramatsu, D.; Takemura, N.; Ahad, M.; Rahman, A.; Yagi, Y. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Trans. Comput. Vis. Appl.* **2019**, *11*, 9. [CrossRef]
15. Hasan, K.; Uddin, M.Z.; Ray, A.; Hasan, M.; Alnajjar, F.; Ahad, M.A.R. Improving Gait Recognition through Occlusion Detection and Silhouette Sequence Reconstruction. *IEEE Access* **2024**, *12*, 158597–158610. [CrossRef]

16. Rida, I.; Almaadeed, N.; Almaadeed, S. Robust gait recognition: A comprehensive survey. *IET Biom.* **2019**, *8*, 14–28. [CrossRef]
17. Ahmed, M.; Al-Jawad, N.; Sabir, A.T. Gait recognition based on Kinect sensor. In Proceedings of the Real-Time Image and Video Processing 2014, SPIE, Brussels, Belgium, 13–17 April 2014; Volume 9139, pp. 63–72.
18. Deng, M.; Wang, C. Human gait recognition based on deterministic learning and data stream of microsoft kinect. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3636–3645. [CrossRef]
19. Sah, S.; Panday, S.P. Model based gait recognition using weighted KNN. In Proceedings of the 8th IOE Graduate Conference, Okayama, Japan, 28–30 March 2020; pp. 1019–1026.
20. Chao, H.; He, Y.; Zhang, J.; Feng, J. Gaitset: Regarding gait as a set for cross-view gait recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8126–8133.
21. Lin, B.; Zhang, S.; Yu, X. Gait recognition via effective global-local feature representation and local temporal aggregation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14648–14656.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60* , 84–90. [CrossRef]
23. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
24. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
25. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
30. Vaswani, A. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
31. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Mogan, J.N.; Lee, C.P.; Lim, K.M.; Muthu, K.S. VGG16-MLP: Gait recognition with fine-tuned VGG-16 and multilayer perceptron. *Appl. Sci.* **2022**, *12*, 7639. [CrossRef]
33. Pushpalatha, K.; Neha, V.; Prajwal, P.; Ashraf, M.; Chiplunkar, C.H. ResNet-Based Gait Recognition: Leveraging Deep Learning for Accurate Biometric Identification. In Proceedings of the 2023 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Mangalore, India, 13–14 October 2023; pp. 49–54.
34. Mogan, J.N.; Lee, C.P.; Lim, K.M.; Muthu, K.S. Gait-vit: Gait recognition with vision transformer. *Sensors* **2022**, *22*, 7362. [CrossRef] [PubMed]
35. Kittler, J.; Matas, J.; Jonsson, K.; Sánchez, M.R. Combining evidence in personal identity verification systems. *Pattern Recognit. Lett.* **1997**, *18*, 845–852. [CrossRef]
36. Ross, A.A.; Nandakumar, K.; Jain, A.K. *Handbook of Multibiometrics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006; Volume 6.
37. Zhan, H.; Kou, J.; Cao, Y.; Guo, Q.; Zhang, J.; Shi, Y. Human Gait phases recognition based on multi-source data fusion and BILSTM attention neural network. *Measurement* **2024**, *238*, 115396. [CrossRef]
38. Qin, L.; Guo, M.; Zhou, K.; Sun, J.; Chen, X.; Qiu, J. Gait recognition based on two-stream CNNs with multisensor progressive feature fusion. *IEEE Sens. J.* **2024**, 24, 13676–13685. [CrossRef]
39. Li, J.; Zhang, Y.; Zeng, Y.; Ye, C.; Xu, W.; Ben, X.; Wang, F.Y.; Zhang, J. Rethinking Appearance-Based Deep Gait Recognition: Reviews, Analysis, and Insights From Gait Recognition Evolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 9777–9797. [CrossRef] [PubMed]
40. Mehraj, H.; Mir, A.H. Person identification using fusion of deep net facial features. *Int. J. Innov. Comput. Appl.* **2021**, *12*, 56–63. [CrossRef]
41. Makihara, Y.; Sagawa, R.; Mukaigawa, Y.; Echigo, T.; Yagi, Y. Which reference view is effective for gait identification using a view transformation model? In Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), New York, NY, USA, 17–22 June 2006; p. 45.

42. Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. Comput. Vis. Appl.* **2018**, *10*, 4. [CrossRef]
43. Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 4, pp. 441–444.
44. Makihara, Y.; Mannami, H.; Tsuji, A.; Hossain, M.A.; Sugiura, K.; Mori, A.; Yagi, Y. The OU-ISIR gait database comprising the treadmill dataset. *IPSJ Trans. Comput. Vis. Appl.* **2012**, *4*, 53–62. [CrossRef]
45. Iwama, H.; Okumura, M.; Makihara, Y.; Yagi, Y. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1511–1521. [CrossRef]
46. Bouchrika, I.; Nixon, M.S. Model-based feature extraction for gait analysis and recognition. In Proceedings of the Computer Vision/Computer Graphics Collaboration Techniques: Third International Conference, MIRAGE 2007, Rocquencourt, France, 28–30 March 2007; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2007; pp. 150–160.
47. Yoo, J.H.; Hwang, D.; Moon, K.Y.; Nixon, M.S. Automated human recognition by gait using neural network. In Proceedings of the 2008 First Workshops on Image Processing Theory, Tools and Applications, Sousse, Tunisia, 23–26 November 2008; pp. 1–6.
48. Yoo, J.H.; Nixon, M.S. Automated markerless analysis of human gait motion for recognition and classification. *Etri J.* **2011**, *33*, 259–266. [CrossRef]
49. Preis, J.; Kessel, M.; Werner, M.; Linnhoff-Popien, C. Gait recognition with kinect. In Proceedings of the 1st International Workshop on Kinect in Pervasive Computing, New Castle, UK, 18 June 2012; Volume 14.
50. Bari, A.H.; Gavrilova, M.L. Artificial neural network based gait recognition using kinect sensor. *IEEE Access* **2019**, *7*, 162708–162722. [CrossRef]
51. An, W.; Yu, S.; Makihara, Y.; Wu, X.; Xu, C.; Yu, Y.; Liao, R.; Yagi, Y. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 421–430. [CrossRef]
52. Jiang, B.; Zhang, Z.; Lin, D.; Tang, J.; Luo, B. Semi-supervised learning with graph learning-convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11313–11320.
53. Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; Rigoll, G. GaitGraph: Graph Convolutional Network for Skeleton-Based Gait Recognition. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2314–2318. [CrossRef]
54. Teepe, T.; Gilg, J.; Herzog, F.; Hörmann, S.; Rigoll, G. Towards a Deeper Understanding of Skeleton-based Gait Recognition, 2022. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Anchorage, AK, USA, 19–22 September 2021. [CrossRef]
55. Gao, S.; Tan, Z.; Ning, J.; Hou, B.; Li, L. ResGait: Gait feature refinement based on residual structure for gait recognition. *Vis. Comput.* **2023**, *39*, 3455–3466. [CrossRef]
56. Ray, A.; Uddin, M.Z.; Hasan, K.; Melody, Z.R.; Sarker, P.K.; Ahad, M.A.R. Multi-Biometric Feature Extraction from Multiple Pose Estimation Algorithms for Cross-View Gait Recognition. *Sensors* **2024**, *24*, 7669. [CrossRef] [PubMed]
57. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1625–1633.
58. Han, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *28*, 316–322. [CrossRef] [PubMed]
59. Liu, J.; Zheng, N. Gait history image: A novel temporal template for gait recognition. In Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007; pp. 663–666.
60. Bashir, K.; Xiang, T.; Gong, S. Gait recognition using gait entropy image. In Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), IET, London, UK, 3 December 2009; pp. 1–6.
61. Wang, C.; Zhang, J.; Pu, J.; Yuan, X.; Wang, L. Chrono-gait image: A novel temporal template for gait recognition. In Proceedings of the Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Proceedings, Part I 11; Springer: Berlin/Heidelberg, Germany, 2010; pp. 257–270.
62. Lam, T.H.; Cheung, K.H.; Liu, J.N. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognit.* **2011**, *44*, 973–987. [CrossRef]
63. Xu, W. Deep large margin nearest neighbor for gait recognition. *J. Intell. Syst.* **2021**, *30*, 604–619. [CrossRef]
64. Junaid, I.; Ari, S. Gait recognition under different covariate conditions using deep learning technique. In Proceedings of the 2022 IEEE International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 11–15 July 2022; pp. 1–5.
65. Suthar, O.; Katkar, V.; Vaghela, K. Person Recognition using Gait Energy Image, MobileNetV3Small and Machine Learning. In Proceedings of the 2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), Mysuru, India, 10–11 February 2023; pp. 1–6.

66. Chen, Y.; Li, X. Gait feature learning via spatio-temporal two-branch networks. *Pattern Recognit.* **2024**, *147*, 110090. [CrossRef]
67. Uddin, M.Z.; Hasan, K.; Ahad, M.A.R.; Alnajjar, F. Horizontal and Vertical Part-wise Feature Extraction for Coss-view Gait Recognition. *IEEE Access* **2024**, *12*, 185511–185527. [CrossRef]
68. Kharb, A.; Saini, V.; Jain, Y.; Dhiman, S. A review of gait cycle and its parameters. *IJCEM Int. J. Comput. Eng. Manag.* **2011**, *13*, 78–83.
69. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
70. Zuama, L.R.; Setiadi, D.R.I.M.; Susanto, A.; Santosa, S.; Gan, H.S.; Ojugo, A.A. High-Performance Face Spoofing Detection using Feature Fusion of FaceNet and Tuned DenseNet201. *J. Future Artif. Intell. Technol.* **2025**, *1*, 385–400. [CrossRef]
71. Uddin, M.Z.; Shahriar, M.A.; Mahamood, M.N.; Alnajjar, F.; Pramanik, M.I.; Ahad, M.A.R. Deep learning with image-based autism spectrum disorder analysis: A systematic review. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107185. [CrossRef]
72. Shiraga, K.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Geinet: View-invariant gait recognition using a convolutional neural network. In Proceedings of the 2016 International Conference on Biometrics (ICB), Halmstad, Sweden, 13–16 June 2016; pp. 1–8.
73. Alotaibi, M.; Mahmood, A. Improved gait recognition based on specialized deep convolutional neural network. *Comput. Vis. Image Underst.* **2017**, *164*, 103–110. [CrossRef]
74. Min, P.P.; Sayeed, S.; Ong, T.S. Gait recognition using deep convolutional features. In Proceedings of the 2019 7th International Conference on Information and Communication Technology (ICoICT), Kuala Lumpur, Malaysia, 24–26 July 2019; pp. 1–5.
75. Aung, H.M.L.; Pluempitiwiriyawej, C. Gait biometric-based human recognition system using deep convolutional neural network in surveillance system. In Proceedings of the 2020 Asia Conference on Computers and Communications (ACCC), Singapore, 4–6 December 2020; pp. 47–51.
76. Balamurugan, S.; Raj, V.J.; Peter, S.J. Deep features based Multiview gait recognition. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 472–478.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.