



Different gait combinations based on multi-modal deep CNN architectures

Büşranur Yaprak¹ · Eyüp Gedikli²

Received: 27 October 2023 / Revised: 28 February 2024 / Accepted: 3 March 2024 /
Published online: 13 March 2024
© The Author(s) 2024

Abstract

Gait recognition is the process of identifying a person from a distance based on their walking patterns. However, the recognition rate drops significantly under cross-view angle and appearance-based variations. In this study, the effectiveness of the most well-known gait representations in solving this problem is investigated based on deep learning. For this purpose, a comprehensive performance evaluation is performed by combining different modalities, including silhouettes, optical flows, and concatenated image of the Gait Energy Image (GEI) head and leg region, with GEI itself. This evaluation is carried out across different multimodal deep convolutional neural network (CNN) architectures, namely fine-tuned EfficientNet-B0, MobileNet-V1, and ConvNeXt-base models. These models are trained separately on GEIs, silhouettes, optical flows, and concatenated image of GEI head and leg regions, and then extracted GEI features are fused in pairs with other extracted modality features to find the most effective gait combination. Experimental results on the two different datasets CASIA-B and Outdoor-Gait show that the concatenated image of GEI head and leg regions significantly increased the recognition rate of the networks compared to other modalities. Moreover, this modality demonstrates greater robustness under varied carrying (BG) and clothing (CL) conditions compared to optical flows (OF) and silhouettes (SF). *Codes available at* <https://github.com/busrakckugurlu/Different-gait-combinations-based-on-multi-modal-deep-CNN-architectures.git>

Keywords Gait recognition · Multi-modal deep CNN · Gait Combination · GEI · Silhouette

✉ Büşranur Yaprak
busra.kucukugurlu@gumushane.edu.tr

Eyüp Gedikli
gediklie@ktu.edu.tr

¹ Department of Software Engineering, Gümüşhane University, 29100 Gümüşhane, Turkey

² Department of Software Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey

1 Introduction

Gait recognition is a computer vision process used to recognize or identify people based on body shape and gait style by remotely acquiring biometric data. The gait style contains the behavioral and physical characteristics of a person [1]. Compared to physical biometrics such as face, fingerprint, iris, and voice, gait (1) gets without any user collaboration or a scanner, (2) achieves success from long-distance and low-resolution data and (3) is hard to fake. Nowadays, the widespread use of cameras for video surveillance makes gait recognition a useful tool for 'real world' applications such as social security, crime prevention, and forensic identification.

Besides its advantages, gait is not as robust as other popular physical biometrics. Recognition performance drops drastically due to 1) the inconspicuous inter-class differences between different people, and 2) the large intra-class variations such as view angle, carrying, and clothing conditions from the same person [2]. To deal with these problems, two kinds of gait recognition approaches are studied: model-based and appearance-based. Model-based approaches are based on human body structure and movements. These approaches generally fit a human model to the input image and then extract movement patterns. They are robust against view angle or appearance-based variations. However, their computational costs are high, and they usually rely on multiple high-resolution cameras [10]. On the other hand, appearance-based approaches directly extract discriminative features from the input image, they do not need to fit a model and do not require high-resolution images [30]. Therefore, these approaches are more suitable for outdoor surveillance. Numerous studies on appearance-based gait recognition have focused on useful representations to compress a gait cycle into one image, e.g. Gait Energy Image(GEI) [3], Chrono Gait Image (CGI) [4], Gait Entropy Images (GENI) [5], Gait Flow Image(GFI) [6], Frame Difference Energy Images-(FDEI) [7], and Period Energy Images(PEI) [8]. In the literature, these representations are frequently used as input data for deep learning networks. GEI is the most popular feature among them, with its low computational cost, easy implementation, and high efficiency. Wu et al. [14] proposed deep convolutional neural network (CNN), which automatically learns the most discriminative gait features and predicts similarity between a pair of GEIs. Wang et al. [45] achieved gait recognition based on the relationship between GEIs and various parts of the human body, the output of the non-local neural networks is horizontally segmented into three sections. Elharrouss et al. [46] utilized multi-task convolutional neural network models and GEIs to estimate the angle and recognize the gait. Silhouettes and optical flows [12] are also frequently used as input data to extract features in many different CNN-based network architectures. Hou et al. [47] proposed a novel network named Gait Lateral Network (GLN) which can learn both discriminative and compact representations from the silhouettes and Castro et al. [48] proposed AttenGait, for gait recognition equipped with trainable attention mechanisms that automatically discover interesting areas of the optical flow data. Recently, multi-modal network structures [15–17] for gait recognition have achieved remarkable results. However, the recognition rates drop drastically under cross-view angle and appearance-based variations, especially for the CASIA-B dataset. These network structures aim to build a richer and more compact gait representation by combining or fusing information from many input modalities. For this reason, in this study, three different modalities that give the best combination with GEI are tried to be found on three different CNN architectures: EfficientNet-B0 [20], MobileNet-V1 [21], and ConvNeXt-base [49]. The main framework of our study is as follows:

- We evaluate the performance of different modalities, including silhouettes, optical flows, and concatenated image of the GEI head and leg regions, to determine the best combination with the GEI, as well as the leg and head regions separately.
- We evaluate the success of three different multi-modal CNN architectures, based on EfficientNet-B0, MobileNet-V1, and ConvNeXt-base that best improve the performance of the modalities combinations.
- We examine the success of twelve different situations under cross-view angle and appearance-based variations, carrying, and clothing conditions. We also consider this for the identical view.
- We investigate how the fusion of modalities features can help gait recognition rate under cross-view angle and appearance-based variations.
- We evaluate the robustness of the last modality, the concatenated image of the GEI head and leg regions, and its combination with GEI, particularly against appearance-based variations.

The rest part of the paper is organized as follows. Section 2 presents the related work. Section 3 describes the proposed method. Experiments and results are presented in Section 4. Finally, last section concludes the paper.

2 Related work

Gait recognition approaches can be grouped into two main categories: model-based and appearance-based. Model-based approaches first model the human body using 2D or 3D body structures and then generate discriminative features by utilizing this model [22–24]. For example, Ariyanto and Nixon [22] used a structural model including articulated cylinders with 3D Degrees of Freedom (DoF) at each joint to model the human lower legs. Feng et al. [24] used the human body joint heatmap extracted from a RGB image instead of using a binary silhouette to describe the human body pose. The heatmaps were sent into the long short-term memory (LSTM) network to extract temporal features. Liao et al. [25] recently proposed PoseGait exploiting human 3D pose estimated from RGB images by CNN as the input feature for gait recognition. Model-based approaches are robust to cross-view angle variations and appearance-based variations such as carrying and clothing conditions. However, 2D and 3D human body modeling processes are complex and also require cameras to capture the video to fit the model well enough.

Appearance-based approaches generally aim to extract gait representations directly from raw input data. These approaches can be further divided into three main categories. The first category proposes spatio-temporal templates. These approaches encode or compress a gait cycle into one image [3–8]. The most popular template in this category is GEI [3]. The second category aims to extract discriminative features from original silhouettes [11, 13, 26]. Chao et al. [9] proposed a network named Gaitset that extracted the frame-level and set-level features from independent silhouettes. Fan et al. [1] presented a temporal part-based framework that consists of two designed components: a Frame-level Part Feature Extractor (FPFE) and a Micromotion Capture Module (MCM). The FPFE is for learning the part-level spatial features, and the MCM is for modeling the local micro-motions features and the global understanding of the entire gait sequence. Zhang et al. [10] proposed an effective spatial–temporal feature learning model with an LSTM attention model based on horizontally divided silhouette images for gait recognition. The last category is

generative approaches, which often transform different representations of gait in different view angles or conditions into a common view angle or condition [29]. Yu et al. [27] employed the Stacked Progressive Autoencoders (SPAЕ) to transform GEI from any given view angle and condition to the same view angle and normal walking condition. In another study developed for a similar purpose using generative adversarial networks (GAN) [28]. He et al. [8] proposed a multi-task generative adversarial network (MGANs) to transform view-specific gait features to another view, based on the assumption that gait images with view variations lie on a low-dimensional manifold [30].

Recently, several different approaches fuse different modalities features at the same time for recognition. For example, Hoffman et al. [15] combined information from a visual RGB image sequence, a depth image sequence, and a four-channel audio stream for multi-modal gait recognition. Castro et al. [16] proposed a unified approach for using audio, depth, and visual information for gait, gender, and shoe type recognition. Zhao et al. [31] presented a multi-modal network, mmGaitSet, using the Gaitset as the backbone to mine shape-based body features from gait silhouettes and pose-based part features from 2D pose heatmaps. In recent studies, Jimenez et al. [17] proposed a multi-modal network UGaitNet handles and combines four modalities for gait recognition: optical flow, gray, depth, and silhouettes. In addition, this network is robust to missing modalities.

In this study, in addition to the optical flow and silhouette modalities most frequently used for gait recognition, the concatenated image of the GEI head and leg regions is obtained as another modality, and the performance of the combination of GEI with these modalities under cross-view angle and appearance-based variations is investigated separately. Furthermore, an assessment of the individual efficacy of the GEI head and leg regions is conducted. This investigation is carried out on three distinct multi-modal CNN architectures based on EfficientNet-B0 [20], MobileNet-V1 [21], and ConvNeXt-base [49].

3 Proposed methods

In the proposed methods, rather than relying solely on a single modality, we utilize a combination of several modalities. We investigate the impact of different modality combinations with GEI on recognition success. To ensure a fair comparison, we evaluate these different modality combinations across various multi-modal deep CNN architectures, including EfficientNet-B0, MobileNet-V1, and ConvNeXt-base.

GEI is selected as the primary modality due to its frequent usage in the literature and its high efficiency. Subsequently, the recognition performances of combinations of the most frequently used input data silhouettes, after GEI, optical flows, and the concatenated image of the GEI head and leg regions with GEI are examined across multi-modal CNN architectures. Additionally, the combination of head and leg regions is also evaluated. The implementation details are presented in the following part of this section.

3.1 Utilized modalities

In this section, we will introduce selected gait representation modalities to feed directly into network architectures. These are GEIs, silhouettes, optical flows, and the concatenated image of the GEI head and leg regions modalities, respectively.

A GEI [3] is obtained by averaging properly aligned silhouettes of gait sequences. Given the preprocessed binary gait silhouette images $B_t(x, y)$ at time t in a sequence, the gray-level GEI is defined as Eq. (1):

$$G(x, y) = \frac{1}{n} \sum_{t=1}^n B_t(x, y) \quad (1)$$

where n is the total number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number in the sequence, and x and y are values in the 2D image coordinate. Some GEI examples from the CASIA-B [34] dataset with cross-view and appearance-based variations are shown in Fig. 1a.

The other modality, silhouette, has been the most frequently used input data in state-of-the-art methods [1, 9, 10]. It takes up a small proportion of an image. The effect of the background and distance factor, caused by the view angles, needs to be reduced. For this reason, silhouettes taken directly from the CASIA-B [34] dataset are aligned based on methods in Takemura et al. [35]. Some examples of silhouette cropping are presented in Fig. 1b.

Using optical flow as input data for action recognition [36] with CNN demonstrated high performance. This is the same for gait recognition [12]. Therefore, The Farneback

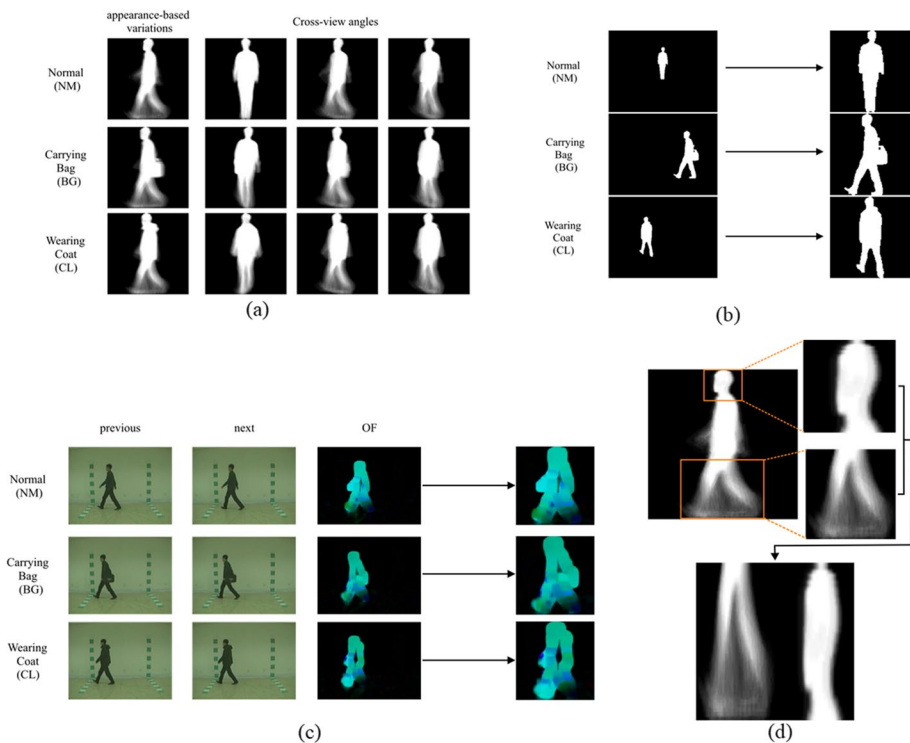


Fig. 1 Four different modalities used for the inputs of network architectures by utilizing the CASIA-B [34] dataset. **a** Some GEI examples, **b** Obtaining the silhouettes (SHs), **c** Obtaining the optical flows (OFs), **d** Concatenated image of the GEI Head and Leg regions(HConL)

[37] method is utilized to obtain the third modality optical flow. The optical flow is calculated between two consecutive RGB gait frames and the final image is aligned like silhouettes [35]. The process of computing optical flow is detailed in Fig. 1c.

Gait recognition by dividing the human body into horizontal parts has been used in previous studies. Rida et al. [50] partitioned the human body into 4 parts based on group Lasso of motion to select the most discriminative human body parts. Rokanujjaman et al. [51] divided the human body into 5 parts and asserted the importance of the head, waist, and leg regions depending on positive or negative effects for recognition. Zhang et al. [10] partitioned the human body into four different horizontal parts and trained multiple separate CNNs for each local part. Moreover, various previous studies have proven that the GEI leg region is highly distinctive in recognition. Choudhury et al. [38] argued that compared to other parts of the human body, the limb region of the GEI better captures discriminative information and is least affected by most carrying conditions and clothing variations. Bashir et al. [39] proposed a GENI to distinguish dynamic gait information and static shape information contained in a GEI. In the study, it is clear that the most dynamic areas in the GEI result images with the feature selection mask applied are the arm and leg regions. In another study [40], an adaptive outlier detection method is proposed to mitigate the impact of clothing on human gait recognition. This approach tries to detect and eliminate clothing effects on silhouettes. It is seen that the leg regions are discriminative in the outlier detection applied result images. When considering variations in clothing and carrying, based on previous studies, only two parts of the human body are selected: the head and leg regions. The selection of these parts takes into consideration the high distinctiveness of the leg region and the low probability of complete coverage of the head region. Subsequently, these regions of GEI are concatenated into a single image for robustness under clothing and carrying condition variations. The concatenating process of the image of the GEI head and leg regions is shown in Fig. 1d.

When considering the modalities evaluated in Fig. 1, GEI is one of the most effective features in gait recognition, it contains the entire walking cycle in a single image. While several studies [32, 33] have shown its efficiency and stability, it is important to acknowledge the information loss inherent in the averaging process. For silhouette(SH) modality, although they are commonly used in gait recognition, they are susceptible to variations in body shape, clothing, and environmental factors [55] such as changes in illumination and dynamic backgrounds. Optical flow (OF), on the other hand, describes the motion between video frames, which makes the recognition easier. However, since it offers a global approach to the human body, such as GEI and SHs, it is likely to be affected by appearance-based variations. Lastly, the concatenated image of the GEI head and leg regions(HConL) presents a local approach, particularly enhancing robustness against variations in clothing and carrying conditions. Consequently, the combination of GEI with modalities exhibiting distinct advantages and disadvantages holds promise for achieving a higher recognition rate.

3.2 Network structures

The CNN architectures are the most popular deep learning framework. They are designed to work specifically on images and achieve remarkable success by extracting spatio-temporal features with a minimum number of parameters. EfficientNet, MobileNet, and ConvNeXt are CNN-based models that have been tested on the ImageNet [41] dataset and have achieved high accuracy results compared to state-of-the-art CNN models. While

EfficientNet-B0 has demonstrated impressive performance in gait recognition [52] using RGB video frames, there remains a gap in evaluation concerning silhouette data. Additionally, the success of MobileNet [53] and ConvNeXt [54] in action recognition has been discussed, but their performance in gait recognition has not yet been assessed. Consequently, for comprehensive performance assessment, three distinct CNN-based network models, namely EfficientNet-B0, MobileNet-V1, and ConvNeXt-base, have been selected for comparative analysis.

Figure 2 depicts the proposed gait recognition frameworks. This figure shows that two different modalities are given to two separate branches of the multi-modal network structures. These modalities, GEIs and SHs, GEIs and OFs, and GEIs and HConLs, as well as the combination of leg and head regions separately are shown in Fig. 1a, b, c, d respectively. Branches consist of two fine-tuned CNN architectures with the same characteristics. These architectures extract features from modalities. Then the features obtained from the different modalities are combined or fused by an concatenate (CON) process. During this process, two feature vectors of dimension n are vertically concatenated, resulting in a final feature dimension of $2n$. In the network branches, EfficientNet-B0, MobileNet-V1, and ConvNeXt-base are used as fine-tuned CNN network structures, respectively.

To understand the impact on recognition performance, all different modality combinations are provided to deep CNN architectures based on a multi-modal approach. Two separate convolution branches of all CNN networks are utilized for feature extraction from two different modalities, and then these features are combined to obtain a single feature vector. This approach allows us to train two separate convolutional branches of the network simultaneously and is useful for evaluating the success of different modality combinations with cross-view and appearance-based variations. Moreover, the fusion process provides a richer and more compact gait representation [16]. Also, using different CNN architectures allows comparison of the performance of these networks for gait recognition, simultaneously.

For the transfer learning process, the network is first initialized with the weights of the pre-trained model, which was trained on the ImageNet dataset. The network is then fine-tuned on the gait recognition dataset. Transfer learning provides more efficient network usage and high-performance training since training is performed on the rather large dataset ImageNet.

4 Experiments and results

4.1 Datasets and metric

4.1.1 CASIA-B

We chose the CASIA-B [34] gait dataset since it contains the original RGB video frames and a large number of video sequences. It is also commonly used. It contains 124 subjects and 11 views (0° , 18° , ..., 162° , 180°) with fixed 18-degree spacing. There are 10 sequences for each subject, 6 sequences for normal walking (NM), 2 sequences for walking with bag (BG), and 2 sequences for walking with coat (CL). For the experimental settings, the first 74 subjects for all conditions are used for training and the rest of the 50 subjects for testing. In the test set, the first 4 sequences of the NM condition (NM #1–4) are regarded as the gallery and the remaining 6 sequences NM #5–6, BG #1–2, and CL #1–2 are used as probe sets, respectively as shown in Table 1.

Fig. 2 Framework of proposed architectures. **a** GEIs and SHs **b** GEIs and OFs **c** GEIs and HConL **d** GEI Leg and GEI Head regions separately are passed through the multi-modal fine-tuned CNNs, and features are fused

4.1.2 Outdoor-gait

Outdoor-Gait [43] is a comprehensive dataset with complex outdoor backgrounds. It contains 138 people with 3 different clothing conditions (NM: normal, CL: with coat, BG: with bag) in 3 distinct scenes (SCENE-1: simple background, SCENE-2: static and complex background, SCENE-3: dynamic and complex background with moving objects). In the experiments, 69 subjects are used as the training set and the remaining 69 subjects are used as the test set. For each condition, there are a minimum of 2 video sequences in both the gallery and the probe sets.

4.1.3 Rank-1

In the experiments, we use average rank-1 accuracies to evaluate the effectiveness of the proposed models. Rank-1 accuracy is determined by the proportion of correctly identified IDs by comparing the probe sequence to all sequences in the gallery (excluding identical view). The average Rank-1 is computed by the ratio of the sum of all Rank-1 values in the specified angle to the total number of angle views (e.g. 11 views for CASIA-B). The average rank-1 accuracy, denoted as C_A , is presented in Eq. (2):

$$C_A = \frac{1}{11} \sum_{a=0}^{10} C_a \quad (2)$$

where C_a is the rank-1 accuracy at the view angle a (excluding identical view) [31].

4.2 Implementation details

Training The inputs of the networks are six different modalities: GEIs, SHs, OFs, HConLs, and GEI Head and GEI Leg parts. They are fed into the networks as GEIs and SHs, GEIs and OFs, GEI Head and GEI Leg parts, and GEIs and HConLs. All modalities (the inputs of all networks) are resized to size the size of 224×224 . Due to the cost of memory and time, the number of training set samples for each of SH and OF is determined as 30. Therefore, 30 SHs and 30 OFs are randomly taken from the SH and OF training sets, respectively.

We use Keras version of Tensorflow [44] for all experiments. The models are trained with Nvidia Geforce Rtx 3060 GPUs and the experimental environment is Windows 10. The Stochastic Gradient Descent (SGD) optimizer is used with the learning rate of 0.0001 and the momentum of 0.9. The output layer has a softmax activation function and cross-entropy loss is used as the loss function.

Testing At the test phase, the similarity measure between gallery and probe features is calculated using cosine similarity. The features obtained from the GEIs and SHs ($\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{SH}}$),

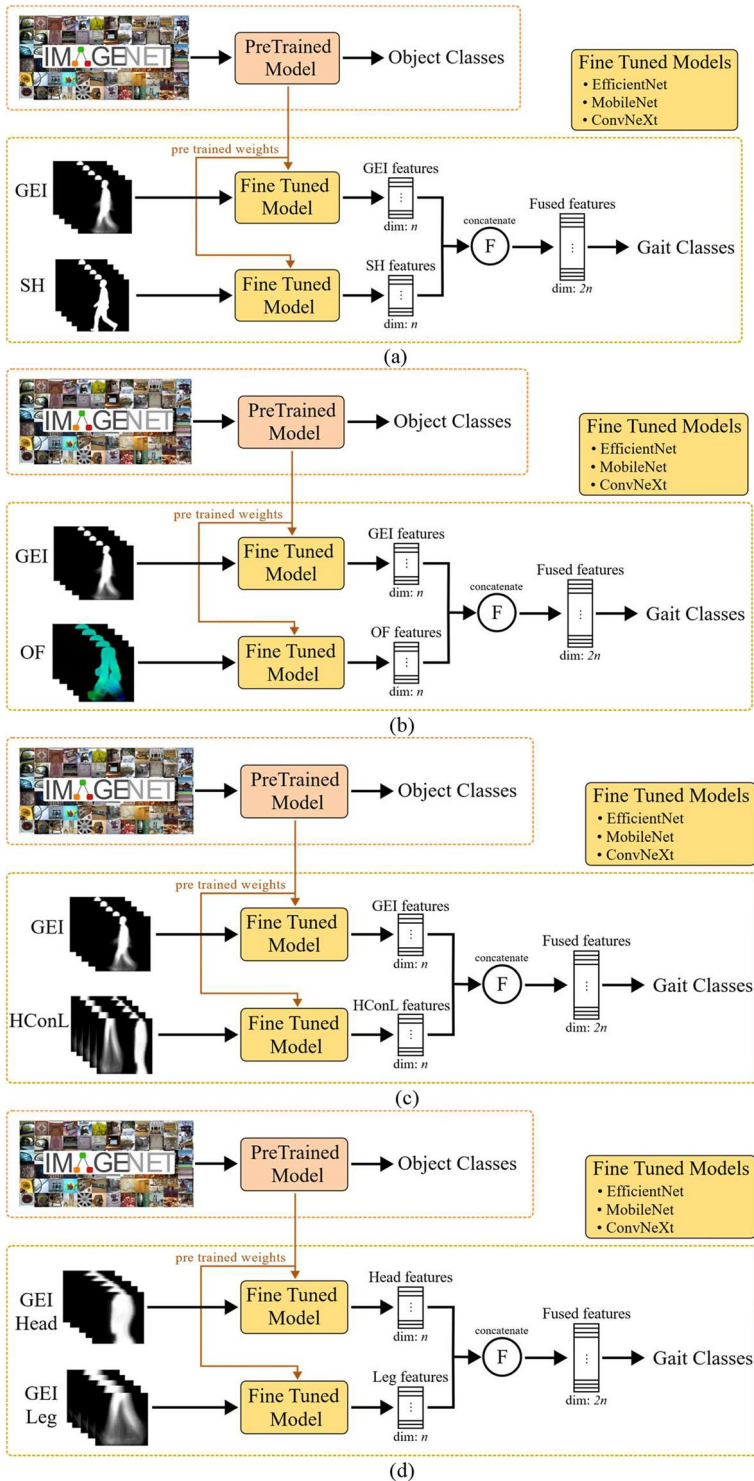


Table 1 Experimental settings on CASIA-B

Training	Test	
	Gallery	Probe
ID: 001–074	ID: 075–124	ID: 075–124
Seqs: NM #01–06, BG #01–02, CL #01–02	Seqs: NM #01–04	Seqs: NM #05–06, BG #01–02, CL #01–02

GEIs and OFs, ($\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{OF}}$), GEIs and HConLs($\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{HConL}}$), and GEI Head and GEI Leg parts($\text{CNN}_{\text{Head}} + \text{CNN}_{\text{Leg}}$) are fused.

4.3 Analysis of different modalities combines

4.3.1 Experiments on CASIA-B Datasets

Experiments under NM variation For a comprehensive comparison, all modality combinations are evaluated on three CNN architectures, specifically EfficientNet-B0, MobileNet-V1, and ConvNeXt-base. Additionally, performance evaluations of VGG16 [18] and ResNet-50 [19] are also presented for the NM variation. When the difference between angles is small, under NM, GEI has a successful performance as a single modality. However, when there is a large difference between angles, the performance drops significantly. This performance achieves individual recognition performances with different modality combinations. The combination of $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{SH}}$, $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{OF}}$, $\text{CNN}_{\text{Head}} + \text{CNN}_{\text{Leg}}$, and $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{HConL}}$ averaged rank-1 accuracy (%) for cross-view angles (excluding the identical views) are detailed in Table 2.

From Table 2, it can be understood that the combination of $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{OF}}$ has a higher accuracy than the combination of $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{SH}}$ for almost all cross-view rank-1 results. This situation can be considered valid for all networks, namely VGG16, ResNet-50, EfficientNet-B0, MobileNet-V1, and ConvNeXt-base. When evaluating network performances, it is clear that MobileNet often has the best results in both modalities, SH and OF and this is also obvious in the mean value. However, the recognition rate of ConvNeXt with OF modality at 0° and 180° view angles significantly exceeds that of MobileNet. When part-based modality combinations are examined, the $\text{CNN}_{\text{Head}} + \text{CNN}_{\text{Leg}}$ combination yields similar results to the $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{SH}}$ and $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{OF}}$ combinations. However, the $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{HConL}}$ combination significantly enhances the performance of all networks except ConvNext. ConvNext achieves its highest performance with the $\text{CNN}_{\text{Head}} + \text{CNN}_{\text{Leg}}$ combination. The comparison results are also shown in Fig. 3.

It can be observed from Fig. 3 that the $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{HConL}}$ combination based on EfficientNet and MobileNet achieves the most successful results among other network-based combinations under the NM variation. This is followed by the $\text{CNN}_{\text{GEI}} + \text{CNN}_{\text{OF}}$ combination. However, this is valid for Convnext in the combination of $\text{CNN}_{\text{Head}} + \text{CNN}_{\text{Leg}}$.

Experiments under BG and CL variations In this section, all models are also tested under BG and CL conditions. However, as in the previous section under the NM conditions, VGG16 and Resnet50 achieve rather poor recognition rates for these variations. Therefore,

Table 2 Comparison of CNN_{GEL} + CNN_{SH}, CNN_{GEL} + CNN_{OF}, CNN_{Head} + CNN_{Leg} (CNN_H + CNN_L) and, CNN_{GEL} + CNN_{HcomL} combinations in terms of averaged rank-1 accuracy (%) for cross-view angles (excluding the identical views), under NM variations

Gallery NM#1–4		0° -180°											mean	
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°		
NM #5–6	CNN _{GEL} + CNN _{SH}	VGG16	60.6	76.2	78.2	75.9	72.3	67.3	71.4	76.0	77.9	76.9	60.3	72.0
	Resnet50	54.7	64.1	68.2	61.1	57.4	51.5	59.2	67.4	69.1	66.5	52.7	61.0	
	EfficientNet	61.8	73.9	78.2	76.7	72.0	67.4	72.6	75.7	77.8	72.7	63.2	72.0	
	MobileNet	62.5	79.1	85.3	84.7	78.0	73.6	77.8	85.1	85.2	79.4	61.0	77.4	
	ConvNeXt	59.4	61.4	65.3	70.4	68.7	62.5	66.9	70.1	68.9	63.6	58.1	65.0	
	CNN _{GEL} + CNN _{OF}	VGG16	60.5	76.0	76.7	75.8	71.7	69.2	72.5	75.9	77.9	80.0	60.2	72.4
	Resnet50	57.0	65.2	70.5	70.9	67.7	62.7	68.2	72.9	75.6	68.8	55.1	66.8	
	EfficientNet	67.4	79.0	79.4	79.7	75.9	75.0	76.4	80.6	82.6	80.1	67.7	76.7	
	MobileNet	67.5	78.8	83.5	85.6	80.6	74.1	80.5	84.7	84.1	79.4	66.5	78.7	
	ConvNeXt	70.9	77.5	80.1	77.8	75.9	72.1	73.5	77.1	80.2	78.4	71.2	75.9	
	CNN _H + CNN _L	VGG16	54.3	66.9	72.4	72.7	66.5	60.6	66.2	71.4	73.4	61.5	45.6	64.7
	Resnet50	57.4	73.5	81.7	80.0	71.6	68.7	73.8	77.6	75.5	71.2	57.9	71.7	
	EfficientNet	62.7	80.0	83.6	78.7	74.5	71.7	76.4	79.8	82.1	73.2	59.1	74.7	
	MobileNet	68.1	82.2	87.6	83.2	78.8	74.8	79.1	80.9	85.1	78.4	61.4	78.1	
	ConvNeXt	70.5	80.8	85.9	86.5	78.9	74.0	76.7	81.5	85.6	80.5	62.7	78.5	
	CNN _{GEL} + CNN _{HcomL}	VGG16	64.0	78.4	82.5	78.4	74.7	70.8	74.8	78.7	82.4	78.3	58.8	74.7
	Resnet50	64.0	80.6	87.2	82.1	73.7	71.6	76.1	82.3	88.0	76.0	60.1	76.5	
	EfficientNet	76.5	88.2	90.7	87.6	82.0	82.5	83.7	87.2	91.5	87.3	71.2	84.4	
	MobileNet	71.0	87.3	90.5	90.1	82.6	81.5	84.7	88.0	90.8	87.1	73.0	84.2	
	ConvNeXt	61.4	72.1	76.9	79.5	74.2	68.6	71.3	78.0	81.2	74.7	60.1	72.5	

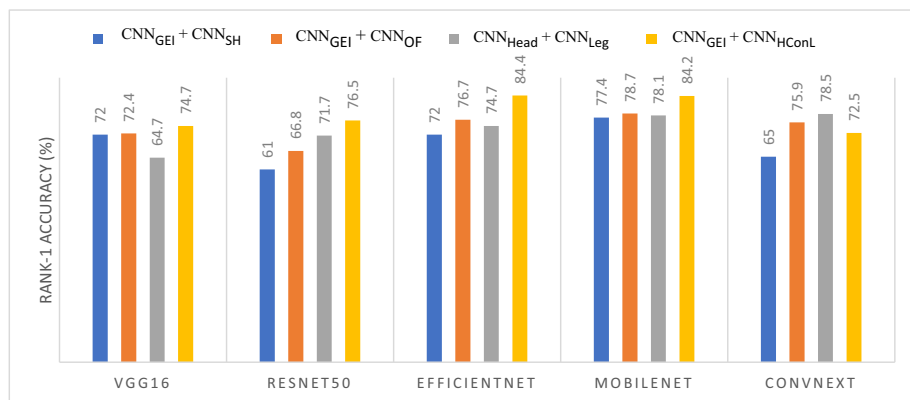


Fig. 3 Comparison of all combinations in terms of average rank-1 accuracy (%), under NM variations for all 11 views (excluding identical-view cases)

only EfficientNet, MobileNet, and ConvNeXt recognition rates are presented for BG and CL variations. The mean rank-1 accuracy (%) of $CNN_{GEI} + CNN_{SH}$, $CNN_{GEI} + CNN_{OF}$, $CNN_{Head} + CNN_{Leg}$, and $CNN_{GEI} + CNN_{HConL}$ combinations under BG and CL variations for cross-view angles (excluding the identical views) are shown in Table 3.

It can be concluded from Table 3 that under BG variation, the $CNN_{GEI} + CNN_{OF}$ combination exhibits higher performance across all networks compared to the $CNN_{GEI} + CNN_{SH}$ combination for almost all cross-view rank-1 results. For these combinations, the performances of EfficientNet and MobileNet are similar; however, MobileNet has achieved the highest mean rank-1 value in both combinations. Considering the combination of $CNN_{Head} + CNN_{Leg}$, the recognition success of all three networks has increased significantly in this combination. Furthermore, for the $CNN_{GEI} + CNN_{HConL}$ combination, these successes continue to increase, although this tendency is reversed for ConvNext. It is observed that under the CL variation, the recognition rates of $CNN_{GEI} + CNN_{SH}$ and $CNN_{GEI} + CNN_{OF}$ combinations based on all networks decrease significantly. However, it is seen that the recognition rate increases slightly in part-based modality combinations. Especially the $CNN_{Head} + CNN_{Leg}$ combination based on MobileNet reaches the highest average rank-1 value under the CL variation. Under all appearance-based variations, namely NM, BG, and CL, the mean rank-1 comparison of each combination is presented in Fig. 4.

When examining Fig. 4, it can be observed that the $CNN_{GEI} + CNN_{SH}$ combination yields the best results when based on MobileNet. This observation is valid for the $CNN_{GEI} + CNN_{OF}$ combination as well. For the $CNN_{Head} + CNN_{Leg}$ combination, MobileNet and ConvNext perform similarly and achieve superior results compared to EfficientNet. The last combination, $CNN_{GEI} + CNN_{HConL}$, achieves the highest performance based on EfficientNet and MobileNet under the NM variation, while under BG and CL variations, it only achieves superior performance with MobileNet. Finally, the optimal outcome is attained with the MobileNet-based $CNN_{GEI} + CNN_{HConL}$ combination for the BG variation, and the MobileNet-based $CNN_{Head} + CNN_{Leg}$ combination for the CL variation.

Table 3 Comparison of $CNN_{GEL} + CNN_{SH}$, $CNN_{GEL} + CNN_{OF}$, $CNN_{H_{Head}} + CNN_{Log}$ ($CNN_H + CNN_L$) and, $CNN_{GEL} + CNN_{H_{CoL}}$ combinations in terms of averaged rank-1 accuracy (%) for cross-view angles (excluding the identical views), under BG and CL variations

Gallery NM#1–4		0° -180°											mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
BG #1–2	$CNN_{GEL} + CNN_{SH}$	EfficientNet	56.5	61.8	62.4	58.6	55.7	50.3	56.2	60.5	61.5	53.9	58.4
		MobileNet	56.1	68.8	70.0	67.4	64.7	59.2	63.8	69.9	69.8	52.6	65.0
		ConvNeXt	52.3	54.8	59.8	61.3	55.7	52.7	52.8	55.6	56.8	49.4	55.5
	$CNN_{GEL} + CNN_{OF}$	EfficientNet	58.1	66.8	63.9	60.0	59.0	54.3	59.4	63.0	65.0	54.4	60.8
		MobileNet	61.7	69.5	71.2	66.0	66.5	60.4	65.4	70.2	68.8	61.8	66.9
	$CNN_H + CNN_L$	ConvNeXt	63.0	69.0	71.6	65.4	62.7	57.9	57.9	64.8	65.8	66.2	64.1
		EfficientNet	58.3	73.0	75.7	73.6	67.9	64.9	68.1	72.5	71.2	69.8	68.2
		MobileNet	65.1	77.7	82.5	77.0	73.1	66.3	69.1	73.8	75.0	73.1	56.8
		ConvNeXt	64.4	76.8	80.3	76.8	74.3	71.4	72.6	77.9	71.8	59.7	73.2
	$CNN_{GEL} + CNN_{H_{CoL}}$	EfficientNet	67.4	75.4	78.6	76.8	73.3	70.0	71.1	78.4	79.6	60.0	73.1
CL #1–2		MobileNet	69.1	79.9	81.6	79.0	74.6	72.2	75.2	81.5	82.2	67.8	76.7
	$CNN_{GEL} + CNN_{SH}$	ConvNeXt	56.6	65.6	68.8	68.9	65.8	60.7	64.2	69.0	71.9	66.4	64.5
		EfficientNet	20.1	26.0	26.8	27.2	22.1	20.2	24.1	25.2	27.1	23.8	23.8
		MobileNet	19.3	29.0	34.9	35.9	36.1	31.5	31.0	33.2	28.9	25.6	19.8
		ConvNeXt	16.5	17.6	24.2	23.1	27.4	29.7	30.5	28.6	27.1	19.2	16.2
	$CNN_{GEL} + CNN_{OF}$	EfficientNet	24.4	28.1	28.6	25.6	20.0	20.9	19.6	20.6	22.1	22.7	19.3
		MobileNet	20.8	31.1	30.6	31.5	27.9	29.9	24.5	25.5	23.1	24.0	19.2
		ConvNeXt	22.0	24.0	28.0	25.0	24.3	22.8	20.6	21.1	22.1	19.1	15.2
	$CNN_H + CNN_L$	EfficientNet	40.5	55.8	59.6	59.0	52.7	53.4	52.9	55.9	50.3	45.3	39.2
		MobileNet	43.8	62.5	65.1	61.2	57.1	55.0	58.2	57.1	53.9	54.1	39.8
	$CNN_{GEL} + CNN_{H_{CoL}}$	ConvNeXt	44.7	56.9	59.4	58.9	59.0	57.9	55.4	58.3	57.0	50.9	39.2
		EfficientNet	30.0	38.9	43.9	42.6	37.0	35.0	35.4	34.4	34.6	26.1	35.7
		MobileNet	36.8	54.8	56.4	52.3	51.9	49.2	47.0	50.1	46.1	43.6	34.5
		ConvNeXt	32.8	40.1	42.6	39.3	41.0	39.5	35.7	36.9	40.4	36.4	29.4

Comparison with the state-of-the-art methods The above experiments have shown that combinations of different modalities achieve good performance, especially with $CNN_{GEI} + CNN_{HConL}$ combination. We compare some proposed combinations with some state-of-the-art methods on the CASIA-B dataset. For this purpose, we organize three comparison groups. The first group of comparisons is made with state-of-the-art method GaitNet [43] which present all the cross-view angle recognition rates, and have the same experimental settings for the NM variation as in Table 1. For comparison, it is selected from the proposed multi-modal networks, $CNN_{GEI} + CNN_{HConL}$ combination based on EfficientNet (Eff+HConL) and MobileNet (Mobile+HConL). Comparison results are presented in Table 4.

The findings presented in Table 4 suggest that the proposed multi-modal networks, Eff+HConL and Mobile+HConL, exhibit performance levels that are very close and comparable to GaitNet. However, it should be noted that when the difference between the angles increased, this situation could not be consistently observed at the mean value. Commonly, according to GaitNet, there is an improvement in performance between angles that are close to each other or symmetric (for example, the gallery angle is 0° and its symmetric is 180° , or the reverse).

The second group of comparisons is made in terms of averaged rank-1 accuracy (%) (excluding the identical view) for all appearance-based variations (NM, BG, CL). The state-of-the-art methods are GEI+PCA [42], GEI-Net [13], DeepCNN [14], GaitNet, and PoseGait [25], which have the same experimental settings as in Table 1. It is chosen from the proposed multi-modal networks for comparison, Eff+HConL and Mobile+HConL for NM variation, Mobile+HConL for BG variation, and $CNN_{Head} + CNN_{Leg}$ based on MobileNet (Mobile+H+L) for CL variation, respectively. Comparison analyzes are shown in Table 5.

It is clear from Table 5 that, Eff+HConL and Mobile+HConL multi-modal networks achieve higher average recognition rates under NM conditions than GEI+PCA, GEI-Net, and a model-based method PoseGait. However, DeepCNN and GaitNet have the best results for average recognition. Mobile+HConL and Mobile+H+L combinations reach the best performance under BG and CL variations, respectively.

The last comparison is made with MGan [8], which is a generative method trained with the first 74 subjects under NM conditions. In Table 6, recognition rates are presented for cross-view angles 54° , 90° , and 126° , as in MGan [8]. According to the results presented in Table 6, the Mobile+HConL multi-modal network achieves the best recognition rate on average for these angles.

4.3.2 Experiments on outdoor-gait datasets

Comparisons of the prepared modality combinations on the Outdoor-gait data set are presented in Table 7. From Table 7, it is clear that the $CNN_{GEI} + CNN_{HConL}$ combination achieves the best performance in all networks. Moreover, the MobileNet-based $CNN_{GEI} + CNN_{HConL}$ combination demonstrates the highest level of success compared to other networks.

Comparison with the state-of-the-art methods Among the prepared modality combinations, MobileNet-based $CNN_{GEI} + CNN_{HConL}$ (Mobile+HConL) and ConvNext-based $CNN_{GEI} + CNN_{HConL}$ (ConvNeXt+HConL) combinations are compared with the appearance-based GaitNet and model-based Human3D [56] methods in Tables 8 and 9, respectively.

Fig. 4 The mean rank-1 comparison of each combination under NM, BG, and CL appearance-based variations. **a** $CNN_{GEI} + CNN_{SH}$ **b** $CNN_{GEI} + CNN_{OF}$ **c** $CNN_{HEAD} + CNN_{LEG}$ **d** $CNN_{GEI} + CNN_{HCONL}$

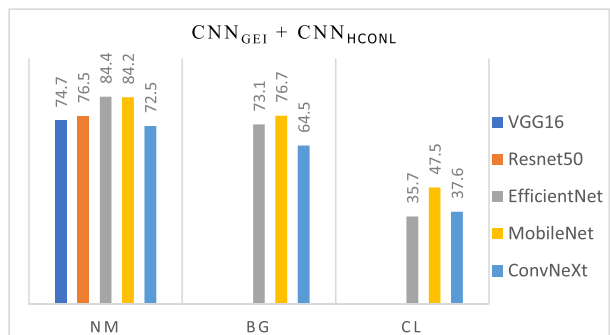
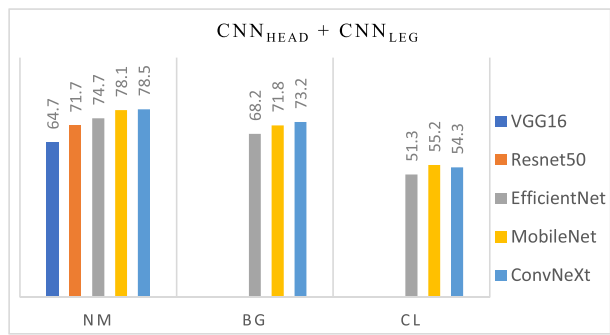
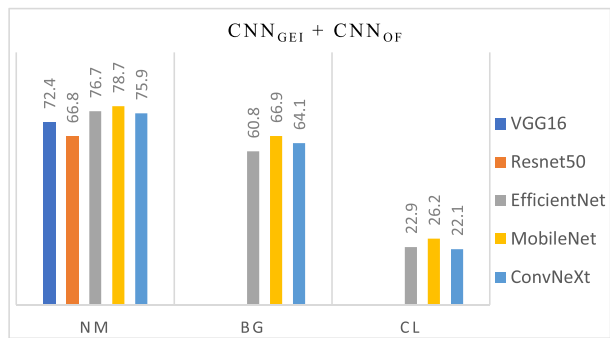
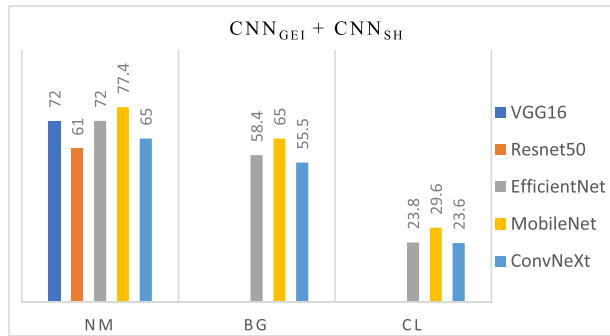


Table 4 Cross-view recognition rates of proposed some multi-modal architectures and GaitNet [43] on CASIA-B, under NM variations

		Probe Angle NM#5–6											mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
Gallery Angle NM #1–4	0°	GaitNet [43]	100	91.0	86.0	77.0	69.0	67.0	63.0	74.0	70.0	84.0	75.6
		Eff + HConL	100	99.0	89.0	74.0	59.0	55.0	65.0	63.0	77.0	89.0	76.5
		Mobile + HConL	100	93.0	86.0	65.0	47.0	50.0	54.0	66.0	73.0	85.0	71.0
18°		GaitNet [43]	98.0	100	100	99.0	87.0	86.0	89.0	88.0	91.0	90.0	91.3
		Eff + HConL	98.0	100	99.0	94.0	78.0	78.0	78.0	80.0	85.0	96.0	88.2
		Mobile + HConL	97.0	99.0	99.0	92.0	73.0	74.0	82.0	78.0	90.0	95.0	87.3
36°		GaitNet [43]	88.0	100	100	100	96.0	90.0	90.0	91.0	85.0	82.0	91.2
		Eff + HConL	86.0	100	100	100	96.0	86.0	92.0	86.0	93.0	91.0	90.7
		Mobile + HConL	86.0	100	100	100	96.0	90.0	91.0	92.0	91.0	85.0	90.5
54°		GaitNet [43]	83.0	97.0	100	100	99.0	97.0	96.0	96.0	95.0	87.0	92.9
		Eff + HConL	70.0	95.0	98.0	100	99.0	98.0	96.0	95.0	85.0	64.0	87.6
		Mobile + HConL	76.0	96.0	100	100	100	99.0	98.0	91.0	77.0	68.0	90.1
72°		GaitNet [43]	79.0	89.0	95.0	97.0	100	100	100	99.0	95.0	79.0	92.5
		Eff + HConL	55.0	76.0	96.0	98.0	100	99.0	94.0	87.0	64.0	54.0	82.0
		Mobile + HConL	49.0	82.0	95.0	99.0	100	99.0	95.0	95.0	64.0	48.0	82.6
90°		GaitNet [43]	78.0	85.0	91.0	95.0	99.0	100	100	97.0	89.0	77.0	91.0
		Eff + HConL	56.0	71.0	94.0	99.0	99.0	99.0	98.0	88.0	68.0	52.0	82.5
		Mobile + HConL	52.0	76.0	92.0	97.0	99.0	99.0	99.0	92.0	65.0	47.0	81.5
108°		GaitNet [43]	73.0	83.0	88.0	95.0	99.0	100	100	99.0	95.0	86.0	91.8
		Eff + HConL	56.0	75.0	91.0	96.0	99.0	99.0	97.0	96.0	73.0	55.0	83.7
		Mobile + HConL	53.0	79.0	91.0	98.0	99.0	99.0	97.0	96.0	81.0	54.0	84.7
126°		GaitNet [43]	80.0	88.0	90.0	94.0	99.0	100	100	99.0	98.0	90.0	93.8
		Eff + HConL	59.0	80.0	88.0	98.0	99.0	98.0	98.0	97.0	88.0	66.0	87.2
		Mobile + HConL	60.0	74.0	94.0	97.0	99.0	99.0	98.0	98.0	94.0	65.0	88.0
144°		GaitNet [43]	75.0	86.0	91.0	92.0	95.0	99.0	99.0	100	100	93.0	92.9
		Eff + HConL	74.0	92.0	95.0	94.0	93.0	90.0	98.0	98.0	95.0	86.0	91.5
		Mobile + HConL	74.0	87.0	92.0	93.0	94.0	94.0	98.0	98.0	96.0	82.0	90.8

Table 4 (continued)

		Probe Angle NM#5–6										mean	
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°		180°
162°	GaitNet [43]	83.0	90.0	90.0	94.0	95.0	96.0	98.0	100	100	100	95.0	94.1
	Eff + HConL	93.0	94.0	85.0	87.0	74.0	75.0	79.0	92.0	98.0	98.0	96.0	87.3
	Mobile + HConL	93.0	95.0	90.0	77.0	72.0	77.0	81.0	95.0	96.0	98.0	95.0	87.1
180°	GaitNet [43]	89.0	86.0	83.0	76.0	74.0	76.0	76.0	81.0	85.0	93.0	99.0	81.9
	Eff + HConL	95.0	88.0	73.0	66.0	55.0	44.0	52.0	68.0	78.0	93.0	98.0	71.2
	Mobile + HConL	94.0	89.0	75.0	59.0	52.0	53.0	58.0	74.0	81.0	95.0	98.0	73.0

Table 5 Comparison of some proposed multi-modal networks with cross-view methods on CASIA-B in terms of averaged rank-1 accuracy (%) (excluding the identical views), under all appearance-based variations

Gallery NM#1–4		0° -180°											mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5–6	GEI + PCA [42]	16.8	24.6	19.9	21.4	26.7	24.6	27.3	24.7	19.5	20.9	21.9	22.6
	GEI-Net [13]	45.8	57.6	67.1	66.9	56.3	48.3	58.3	68.4	69.4	59.0	46.5	58.5
	DeepCNN [14]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitNet [43]	75.6	91.3	91.2	92.9	92.5	91.0	91.8	93.8	92.9	94.1	81.9	89.9
	PoseGait [25]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	Eff + HConL	76.5	88.2	90.7	87.6	82.0	82.5	83.7	87.2	91.5	87.3	71.2	84.4
	Mobile + HConL	71.0	87.3	90.5	90.1	82.6	81.5	84.7	88.0	90.8	87.1	73.0	84.2
BG#1–2	DeepCNN [14]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	PoseGait [25]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	Mobile + HConL	69.1	79.9	81.6	79.0	74.6	72.2	75.2	81.5	82.2	80.2	67.8	76.7
CL#1–2	DeepCNN [14]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	PoseGait [25]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	35.9
	Mobile + H + L	43.8	62.5	65.1	61.2	57.1	55.0	58.2	57.1	53.9	54.1	39.8	55.2

Table 6 Comparison of some proposed multi-modal networks with MGan [8] on CASIA-B, under NM variations

Methods	54°	90°	126°	mean
MGan [8]	84.2	72.3	83.0	79.8
Eff + HConL	87.6	82.5	87.2	85.8
Mobile + HConL	90.1	81.5	88.0	86.5

When Table 8 is evaluated, it is seen that Mobile + HConL outperforms GaitNet in CL variations, namely NM-CL, BG-CL, CL-NM, and CL-BG. At the mean value, the results are comparable. In Table 9, the ConvNext-based $CNN_{GEI} + CNN_{HConL}$ combination showed superior recognition rate in NM-NM, BG-BG, and CL-CL variations. This situation cannot be observed in the mean value. However, the model-based 3DHuman method is more costly due to the 3D human body reconstruction process.

4.3.3 Computation and efficiency analysis

It is evident that among the various combinations prepared in the preceding sections, the MobileNet-based combinations have exhibited superior performance. Mobilenet aims to maximize accuracy while taking care of limited resources. Therefore, it is characterized by its small size, low power consumption, speed, and cost-effectiveness. Consequently, a cost comparison is conducted with the GaitSet and GaitNet methods, considering any combination based on multi-modal MobileNet that is prepared. GaitSet has significantly improved the performance of gait recognition. However, this method is characterized by a complex network architecture, with a large number of parameters and floating-point operations (FLOPs). It has 2.59 M parameters, 8.6G floating point operations (FLOPs), and a final feature dimension of 15872 [57].

Table 7 Comparison of proposed multi-modal networks on Outdoor-Gait dataset in terms of averaged rank-1 accuracy (%) under all appearance-based variations

	Gallery	NM			BG			CL			mean				
		Probe	NM		BG		CL		NM			BG		CL	
CNN _{GEI} + CNN _{SH}	EfficientNet		58.94	48.55	40.34	47.58	63.28	34.42	39.98	35.02	61.59	47.74			
	MobileNet		76.09	67.39	55.31	63.28	81.40	55.19	55.55	54.59	75.12	64.88			
	ConvNeXt		81.16	67.99	58.94	65.94	86.71	51.81	56.88	53.02	84.54	67.44			
CNN _H + CNN _L	EfficientNet		58.94	44.44	40.94	42.75	65.94	27.17	36.83	25.48	63.28	45.08			
	MobileNet		77.54	75.24	72.34	70.17	86.96	58.57	67.87	63.77	85.02	73.05			
	ConvNeXt		80.19	72.95	58.82	66.67	82.13	52.54	52.54	47.58	84.54	66.44			
CNN _{GEI} + CNN _{Hcool}	EfficientNet		77.29	61.95	51.45	64.61	79.23	44.32	49.52	47.22	82.37	61.99			
	MobileNet		87.44	81.52	71.01	76.81	93.24	64.61	67.75	64.37	88.41	77.24			
	ConvNeXt		92.51	79.59	69.08	73.91	95.41	57.49	63.41	54.71	95.17	75.70			

Table 8 Comparison of some proposed multi-modal networks with appearance-based GaitNet method on Outdoor-Gait dataset in terms of averaged rank-1 accuracy (%) under all appearance-based variations

	Gallery Probe	NM			BG			CL			mean
		NM	BG	CL	NM	BG	CL	NM	BG	CL	
GaitNet		96.9	89.1	60.2	92.0	97.1	59.7	58.7	55.4	97.3	78.5
Mobile + HConL		87.4	81.5	71.0	76.8	93.2	64.6	67.7	64.3	88.4	77.2
ConvNeXt + HConL		92.5	79.6	69.1	73.9	95.4	57.5	63.4	54.7	95.2	75.7

Table 9 Comparison of some proposed multi-modal networks with model-based Human3D [56] method on Outdoor-Gait dataset in terms of averaged rank-1 accuracy (%) (under all appearance-based variations)

	Gallery Probe	NM			BG			CL			mean
		NM	BG	CL	NM	BG	CL	NM	BG	CL	
3Dhuman[56]		90.0	83.6	80.9	77.5	90.3	72.6	72.2	66.4	88.8	80.2
ConvNeXt + HConL		92.5	79.6	69.1	73.9	95.4	57.5	63.4	54.7	95.2	75.7

Similarly, GaitNet utilizes a Fully Convolutional Network (FCN) [58] as a backbone network with a high number of parameters. The comparison results of these networks with multi-modal MobileNet (mm-MobileNet) are shown in Table 10.

When examining Table 10, the number of parameters of mm-MobileNet is very close to GaitSet, but the FLOPs are approximately 8 times less than GaitSet. Furthermore, the parameter count is considerably lower compared to FCN-based networks. Additionally, the final feature dimension, crucial for the testing phase, is notably lower than that of GaitSet.

5 Conclusion

Gait recognition approaches often focus on completing the recognition process through a single modality. In this study we investigated the performance of using different modalities jointly for recognition. For this purpose, we combined different most common gait representations using multi-modal network architectures and acquired more robust and rich representations. We utilized four different modalities: silhouettes, optical flows, and concatenated image of the GEI Head and Leg regions

Table 10 Cost analysis of networks. Dim represents the dimension of final feature

Network	FLOPs(G)	Params(M)	Dim
GaitSet	8.59	2.59	15,872
FCN-Alexnet	-	57	-
FCN-VGG16	-	134	-
FCN-GoogLeNet	-	6	-
mm-MobileNet	1.15	3.3	1024 + 1024

(HConL), with the main modality GEI, and evaluated the success of their combination with GEI for cross-view angle and appearance-based variations. To obtain fair and reliable results, we used MobileNet-V1, EfficientNet-B0, and ConvNext-base for the branches of multi-modal networks. We also evaluated the success of these networks for gait recognition. We compared the best performances of all proposed network combinations with some state-of-the-art methods on the CASIA-B and Outdoor-Gait datasets. Experimental results indicated that combinations of different input modality features have different performances depending on cross-view angle and appearance-based variations. In particular, the combination of GEI with HConL based on MobileNet and ConvNext achieved remarkable recognition rate when the difference between angles is small, and the HConL modality increased recognition rates under appearance-based variations. Moreover, the cost of multi-modal networks, particularly MobileNet based, is notably low.

Authors contribution Conceptualization: [Büşranur Yaprak]; Methodology: [Büşranur Yaprak]; Formal analysis and investigation: [Büşranur Yaprak]; Writing—original draft preparation: [Büşranur Yaprak]; Writing—review and editing: [Büşranur Yaprak], [Eyüp Gedikli]; Supervision: [Eyüp Gedikli].

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing of interest The authors have no competing interests to declare that are relevant to the content of this article.

Financial interests The authors declare they have no financial interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, Huang Y, Li Q, He Z (2020) Gaitpart: Temporal part-based model for gait recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14225–14233
2. Zhang C, Liu W, Ma H, Fu H (2016) Siamese neural network based gait recognition for human identification. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2832–2836
3. Han J, Bhanu B (2005) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28:316–322
4. Wang C, Zhang J, Wang L, Pu J, Yuan X (2011) Human identification using temporal information preserving gait template. IEEE Trans Pattern Anal Mach Intell 34:2164–2176

5. Bashir K, Xiang T, Gong S (2009) Gait recognition using gait entropy image. In: 3rd international conference on imaging for crime detection and prevention, pp 1–6
6. Lam TH, Cheung KH, Liu JN (2011) Gait flow image: a silhouette-based gait representation for human identification. *Pattern Recognit* 44:973–987
7. Chen C, Liang J, Zhao H, Hu H, Tian J (2009) Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognit. Lett* 30:977–984
8. He Y, Zhang J, Shan H, Wang L (2019) Multi-task GANs for view-specific feature learning in gait recognition. *IEEE Trans Inf Forensics Secur* 14:1102–1113
9. Chao H, He Y, Zhang J, Feng J (2019) GaitSet: Regarding gait as a set for cross-view gait recognition. In: *Proceedings of the AAAI conference on artificial intelligence*, pp 8126–8133
10. Zhang Y, Huang Y, Yu S, Wang L (2020) Cross-View gait recognition by discriminative feature learning. *IEEE Trans Image Process* 29:1001–1015
11. Wolf T, Babae M, Rigoll G (2016) Multi-view gait recognition using 3D convolutional neural networks. In: *2016 IEEE international conference on image processing (ICIP)*, pp 4165–4169
12. Castro FM, Marín-Jiménez MJ, Guil N, Pérez De La Blanca N (2017) Automatic learning of gait signatures for people identification. In: *International work-conference on artificial neural networks*, pp 257–270
13. Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2016) GEINet: view-invariant gait recognition using a convolutional neural network. In: *Proceedings of the international conference on biometrics*, pp 1–8
14. Wu Z, Huang Y, Wang L, Wang X, Tan T (2017) A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Trans Pattern Anal Mach Intell* 39:209–226
15. Hofmann M, Geiger J, Bachmann S, Schuller B, Rigoll G (2014) The TUM Gait from Audio, Image and Depth (GAID) database. *Multimodal recognition of subjects and traits. J Vis Commun Image Representation* 25:195–206
16. Castro FM, Marín-Jiménez MJ, Guil N (2016) Multimodal features fusion for gait, gender and shoes recognition. *Mach Vis Appl* 27:1213–1228
17. Marín-Jiménez MJ, Castro FM, Delgado-Escañó R, Kalogeiton V, Guil N (2021) UGaitNet: multimodal gait recognition with missing input modalities. *IEEE Trans Inf Forensics Secur* 16:5452–5462
18. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409–1556*
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
20. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International conference on machine learning. PMLR*, pp 6105–6114
21. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704–04861*
22. Nixon MS, Carter JN, Nash JM, Huang PS, Cunado D, Stevenage SV (1999) Automatic gait recognition. In: *Motion analysis and tracking*, pp 3/1–3/6
23. Wang L, Ning H, Tan T, Hu W (2004) Fusion of static and dynamic body biometrics for gait recognition. *IEEE Trans Circ Syst Video Technol* 14:149–158
24. Feng Y, Li Y, Luo J (2016) Learning effective gait features using LSTM. In: *The 23rd International conference on pattern recognition*, pp 325–330
25. Liao R, Yu S, An W, Huang Y (2020) A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recogn* 98:107069
26. Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2017) On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Trans Circ Syst Video Technol* 29:2708–2719
27. Yu S, Chen H, Wang Q, Shen L, Huang Y (2017) Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing* 239:81–93
28. Yu S, Chen H, Reyes G, Edel B, Poh N (2017) GaitGAN: invariant gait feature extraction using generative adversarial networks. In: *IEEE conference on computer vision and pattern recognition workshops*, pp 30–37
29. Zhang P, Wu Q, Xu J (2019) VT-GAN: View transformation GAN for gait recognition across views. In: *International joint conference on neural networks*, pp 1–8
30. Han F, Li ZJ, Shen F (2022) A unified perspective of classification-based loss and distance-based loss for cross-view gait recognition. *Pattern Recogn* 125:108519
31. Zhao L, Guo L, Zhang R, Xie X, Ye X (2022) mmGaitSet: multimodal based gait recognition for countering carrying and clothing changes. *Appl Intell* 52:2023–2036
32. Iwama H, Okumura M, Makihara Y, Yagi Y (2012) The OU-ISIR gait database: Comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans Inf Forensics Secur* 7:1511–1521

33. Makihara Y, Matovski DS, Nixon MS, Carter JN, Yagi Y (1999) Gait recognition: Databases, representations, and applications. In: Wiley encyclopedia of electrical and electronics engineering, pp 1–15
34. Yu S, Tan D, Tan T (2006) 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, pp 441–444
35. Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2018) Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSI Trans Comput Vis Appl* 10:4
36. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on neural information processing systems, pp 568–576
37. Farneback G (2003) Two-frame motion estimation based on polynomial expansion. In: Image analysis: 13th Scandinavian Conference SCIA, pp 363–370
38. Choudhury SD, Tjahjadi T (2015) Robust view-invariant multiscale gait recognition. *Pattern Recogn* 48:798–811
39. Bashir K, Xiang T, Gong S (2010) Gait recognition without subject cooperation. *Pattern Recogn Lett* 31:2052–2060
40. Ghebleh A, Ebrahimi Moghaddam M (2018) Clothing-invariant human gait recognition using an adaptive outlier detection method. *Multimedia Tools Applic* 77:8237–8257
41. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Computer vision and pattern recognition. IEEE, pp 248–255
42. Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identification. *IEEE Trans Pattern Anal Mach Intell* 25:1505–1518
43. Song C, Huang Y, Huang Y, Jia N, Wang L (2019) Gaitnet: an end-to-end network for gait based human identification. *Pattern Recogn* 96:106988
44. Abadi M et al (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*
45. Wang X, Yan WQ (2021) Non-local gait feature extraction and human identification. *Multimedia Tools Applic* 80:6065–6078
46. Elharrouss O, Almaadeed N, Al-Maadeed S, Bouridane A (2021) Gait recognition for person re-identification. *J Supercomput* 77:3653–3672
47. Hou S, Cao C, Liu X, Huang Y (2020) Gait lateral network: learning discriminative and compact representations for gait recognition. In: Springer International Publishing European conference on computer vision, pp 382–398
48. Castro FM, Delgado-Escañó R, Hernández-García R, Marín-Jiménez MJ, Guil N (2024) AttenGait: Gait recognition with attention and rich modalities. *Pattern Recogn* 148:110171
49. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S (2022) A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11976–11986
50. Rida I, Jiang X, Marcialis GL (2015) Human body part selection by group lasso of motion for model-free gait recognition. *IEEE Signal Process Lett* 23:154–158
51. Rokanujjaman M, Hossain MA, Islam MR (2012) Effective part selection for part-based gait identification. In: 2012 7th IEEE international conference on electrical and computer engineering, pp 17–19
52. Khan MA, Arshad H, Khan WZ, Alhaisoni M, Tariq U, Hussein HS, Elashry A (2023) HGRBOL2: human gait recognition for biometric application using Bayesian optimization and extreme learning machine. *Futur Gener Comput Syst* 143:337–348
53. Liu L, Wang X, Bao Q, Li X (2024) Behavior detection and evaluation based on multi-frame MobileNet. *Multimed Tools Appl* 83:15733–15750
54. Fu H, Gao J, Liu H (2023) Human pose estimation and action recognition for fitness movements. *Comput Graph* 116:418–426
55. He C, Li K, Zhang Y, Tang L, Zhang Y, Guo Z, Li X (2023) Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 22046–22055
56. Meng C, He X, Tan Z, Luan L (2023) Gait recognition based on 3D human body reconstruction and multi-granular feature fusion. *J Supercomput* 79:12106–12125
57. Song X, Huang Y, Shan C, Wang J, Chen Y (2022) Distilled light GaitSet: towards scalable gait recognition. *Pattern Recogn Lett* 157:27–34
58. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440