GAITAL-EI: A DUAL-BRANCH CNN INTEGRATING GAIT ENERGY IMAGES AND VIDEO SEQUENCES FOR PERSON IDENTIFICATION

Cuong Tran-Chi, Hanh T. M. Tran*, Tien Ho-Phuoc

The University of Danang, University of Science and Technology, Vietnam

*Corresponding author: hanhtran@dut.udn.vn

(Received: May 05, 2025; Revised: June 15, 2025; Accepted: June 20, 2025)

DOI: 10.31130/ud-jst.2025.23(10B).640E

Abstract - Biometric-based identity recognition from images has attracted substantial research interest in recent years, driven by the growing need for reliable and non-intrusive methods of human identification. Among various approaches, gait recognition has emerged as a promising method due to its key advantages, including the ability to identify individuals at a distance and under diverse conditions. In this paper, we propose a deep convolutional neural network, called GaitAL-EI, to learn gait motion features through a dual-branch architecture that processes both gait energy images and video sequences. Our proposed method is trained and evaluated on public datasets. Comparative experiments demonstrate that the proposed approach achieves superior performance compared to the state-of-the-art GaitSet method, highlighting the effectiveness of integrating both static and dynamic gait information.

Key words - Gait Recognition; Convolution Autoencoder; Gait Energy Image.

1. Introduction

Gait recognition [1] is a biometric technique that identifies individuals based on their body shape and walking patterns. Differences in movement and body shape are key to making a person's gait unique. Compared with face recognition, gait recognition has many advantages such as recognition distance, difficulty to spoof, and less affected by camera quality. Due to these advantages, gait represents a promising biometric trait in situations where facial features cannot be captured with sufficient resolution for reliable recognition. Despite the encouraging advancements, gait recognition continues to face a significant challenge: the conditions under which most existing gait datasets are collected differ substantially from those encountered in real-world applications.

Despite facing many challenges, gait recognition with a deep learning approach has become a major research direction, with great potential to become practical applications for biometric needs. Gait recognition can be divided into 2 main methods [2]: model-based [3, 4] and appearance-based gait recognition [5 - 9].

Model-based approaches extract gait features by analyzing the human body structure [3, 4]. These approaches typically follow a two-step process: first, estimating pose information, and then applying models such as LSTM or CNN to extract discriminative features for recognition. They often rely on 2D/3D pose keypoints [3] or a 3D body skeleton [4] constructed from human motion. A key advantage of model-based methods is their ability to address the cross-view problem by rotating the 3D body model, making them less sensitive to variations in view angles, carried objects, or clothing changes. However, these methods are hard to extract key-points accurately, especially from low-resolution images.

Appearance-based methods typically utilize deep convolutional neural networks (CNNs) to encode pedestrian images and identify individuals based on learned gait embeddings. Depending on the type of input data, these methods can be categorized into three groups: template-based, sequence-based, and set-based approaches. Template-based approaches extract features from a single gait image, such as the Gait Energy Image (GEI) [5, 9, 10], using CNNs. Silhouette-based representations like GEIs are straightforward to obtain, even from low-resolution images. Since many motion details are lost due to this abstraction, sequence-based or video-based approaches have been developed to capture more comprehensive motion information. These methods use silhouettes from every frame instead of relying on a single template [11 - 13]. For temporal feature extraction, models such as Long Short-Term Memory (LSTM) networks [6] and 3D Convolutional Neural Networks (3D-CNNs) [7] have been employed. Feng et al. [6] utilized LSTM for cross-view gait recognition, while Xing et al. [7] proposed a 3D-CNN model that jointly captures spatial and temporal features to achieve view-invariant recognition. By leveraging both frame-level spatial features and temporal dynamics, these methods yield more discriminative gait representations. Set-based approaches, such as the method proposed by Chao et al. [8], treat gait as a collection of discrete silhouette images rather than a continuous sequence. In this case, the spatial appearance of the silhouettes serves as a proxy for temporal information. Their model, GaitSet, extracts gait features at the frame level and employs a pooling operation to aggregate them into a comprehensive set-level representation.

Template-based and sequence-based approaches bring certain advantages, are implemented by researchers, and have achieved outstanding performance and accuracy. By treating the gait as a clip and processing it using temporal models, the video processing approach has the strength of extracting small features in the gait movement. As for the approach using energy images, the frequency, cycle and body shape are considered.

In this paper, we propose a two-branch gait recognition framework that processes gait energy images and video sequences in parallel. By combining the complementary

features from both branches, our model learns a gait embedding vector that captures fine-grained walking motion details, gait frequency, and body shape characteristics. The model is trained in two phases using a composite dataset that merges our newly collected gait data with three publicly available gait datasets, thereby covering a wide range of real-world walking scenarios. Comparative results demonstrate superior performance over the GaitSet method.

2. The proposed method

This section provides an overview of the proposed method, called GaitAL-EI, followed by a detailed presentation of its component blocks, including the AL block and the EI block.

Figure 1 illustrates the proposed model for gait recognition that contains two main blocks to effectively extract different motion features of gait. The first block, called Gait Convolutional Autoencoders – Long Short-Term Memory block (AL block), captures walking behavior and the motion characteristics during walking. The second block, known as the Gait Energy Image encoding block (EI block), extracts general information such as body shape and step patterns. The detailed and overall features processed through these two blocks are combined into an embedding vector that describes the gait characteristics of a person. This division into two parallel blocks is inspired by modern high-performance CNN backbones, which tend to integrate dynamic motion features with different approaches. For instance, architectures like ResNet and Vision Transformers, Swin Transformers, often utilize multiple parallel branches by segmenting the input into smaller pieces.

The use of autoencoders as a feature extraction block before the LSTM stems from the desire for each part of the model to learn and perform its specific function effectively: autoencoders learn to encode images, while the LSTM learns to encode gait. When these two blocks are separated during training, the image feature extraction block will not interfere with the temporal feature analysis of the LSTM. Furthermore, BiLSTM is chosen to eliminate the influence of the starting and ending points of the gait on this block.

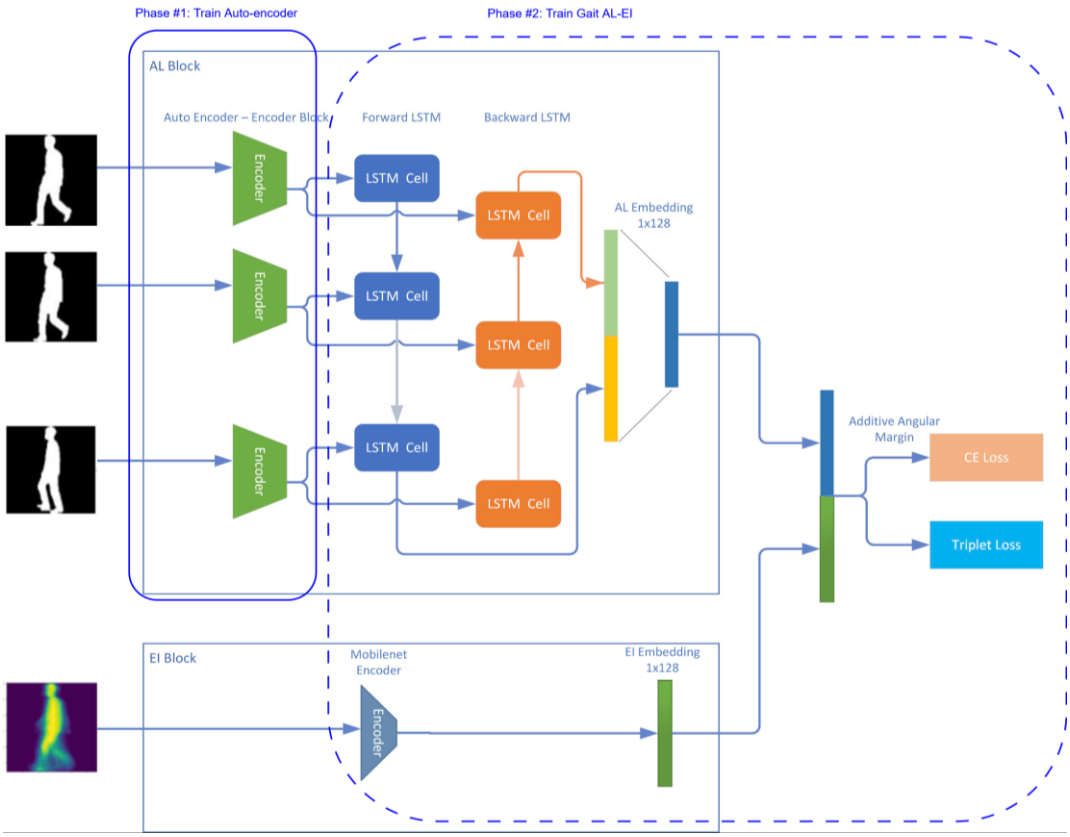


Figure 1. *The block diagram of the proposed GaitAL-EI*

2.1. AL Block

The first branch operates on the principle of extracting low-dimensional feature vectors of individual silhouette images and then feeding the sequence of vectors extracted from the image sequence into the LSTM for analysis and dynamic motion feature extraction, as this sequence has a temporal aspect. This block consists of two main parts: an autoencoder and a Bidirectional LSTM (BiLSTM) [15].

In this work, the encoder consists of two convolutional blocks, each followed by a Hard-Swish activation function [16] and a max pooling layer for down-sampling. The decoder includes three transposed convolution layers followed by a Hard-Swish activation layer. Table 1 demonstrates details of all layers in the encoder and decoder parts. The use of the Hard-Swish layer in the autoencoder's architecture provides a performance improvement of 2-3% compared to using

other activation functions such as ReLU or sigmoid [16]. In this Table, *i_o_s_str* illustrates the number of input channels, output channels, kernel size and stride.

Table 1. *The autoencoder's architecture*

Encoder part		Decoder part	
Layer	**Shape** *i_o_s_str*	**Layer**	**Shape** *i_o_s_str*
Conv	1-16-3-3	ConvTranspose	4-16-3-2
Hard-swish		Hard-swish	
MaxPool	16-16-2-2	ConvTranspose	16-8-5-3
Conv	16-4-3-2	Hard-swish	
Hard-swish		ConvTranspose	8-1-2-2
MaxPool	4-4-2-1	Tanh	

The encoder extracts silhouette images of size 112×112 into a feature vector of size 1×324. A walking clip length of 40 frames is chosen for the gait's feature extraction. Consequently, the size of the output feature vector sequence is 40×324. Bidirectional LSTM is employed on this feature vector sequence to extract an embedding vector that effectively captures the dynamic motion of gait through a sequence of 40 silhouette images. BiLSTM networks are widely employed in natural language processing tasks due to their ability to capture contextual information in both forward and backward directions. This bidirectional flow enables the model to relate past and future information, which is particularly beneficial for recognizing cyclical patterns such as those present in gait sequences.

After using the BiLSTM block to extract features from the feature vectors frame by frame, a 1×128 dimensional vector is obtained. It is called the AL embedding vector, described in Figure 1. This vector is extracted to depict dynamic motion features of the gait. By keeping these two blocks separate during training, the silhouette image feature extraction block does not interfere with the LSTM's temporal feature analysis. Conversely, this separation allows LSTM to operate without being influenced by the starting and ending points of the gait. BiLSTM is chosen to mitigate the impact of these points on the model. The architecture of the BiLSTM model is described in Table 2.

Table 2. *The architecture of the Bidirectional LSTM model with its input and output shapes*

Layer	Input shape	Output shape
BiLSTM	40×324	1×256
Fully connected	1×256	1×64
ReLU	1×64	1×64
Fully connected	1×64	1×128

2.2. EI Block

The second branch is the EI block, where the sequence of input silhouette images is aggregated through summation to create a Gait Energy Image [5, 9, 10]. This image contains overlapping residual images of the gait; thus, it is also unaffected by the starting and ending points of the gait, ensuring that the entire model is not affected by the timing of the start and end.

The Gait Energy image contains two pieces of information: first, the shape of the body while walking, and second, the amplitude and frequency of the movements during walking. The Gait Energy Image (EI) is created using the average of the frames in the clip [5]. It is described in Equation 1 and Figure 2.

$$EI(x,y) = \frac{1}{N}\sum_{t=1}^{N} G_t(x,y) \quad (1)$$

in which G_t is the t^{th} silhouette image in the sequence $N = 40$ is chosen after experiments.

To perform frame addition to obtain an *EI* as in Equation 1, the images must be cropped and re-centered on the body by identifying a bounding rectangle around the subject. As observed in the Gait *EI* in Figure 2, it encapsulates key information about human shape, motion ghosting effects, and body vibration frequencies, represented through transitional regions.

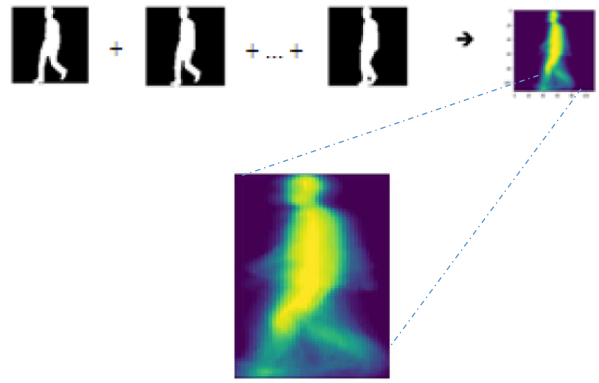


Figure 2. *A Gait Energy Image*

The blurred region extending outward from the object's center is known as the transition region. The first type of transition region represents body vibrations or parts that remain relatively stationary with respect to the body's center of mass during walking (Figure 3). The second type reflects the frequency and trajectory of moving parts relative to the body's center of mass (Figure 4).

Figure 3. *The transition domain of the stationary parts (Head and body)*

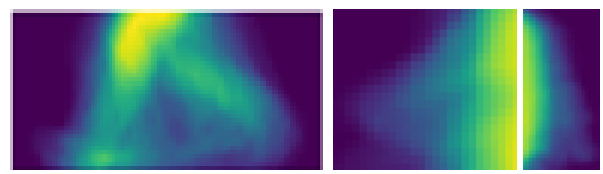


Figure 4. *The transition domain of moving parts (legs and arms)*

These two domains, together with the overall image describing the body shape, create a characteristic of a person's gait. To extract this full feature, we use a MobileNetv2 network [16] that has removed the last fully

connected layers. The MobileNet encoder will take a Gait *EI* image as its input; through the full convolutional layers of the network, a 128-dimensional vector is extracted at the output. This vector, named as the EI embedding, describes the most general features of the body while walking.

The outputs of the two blocks are concatenated to produce a unique vector that describes the gait. This vector is called the gait embedding vector, as described in Figure 1.

3. Training and Inference procedure

The training procedure of our proposed model is divided into 2 phases: (1) Training the autoencoders for the AL block and (2) Training the remaining parts of the GaitAL-EI model. These two steps are demonstrated with a solid line box and a dashed line box in Figure 1. The details of the two phases are described in subsections 3.2 and 3.3. Then the inference step is mentioned in subsection 3.4.

3.1. Dataset

CASIA-B dataset [17] is a popular gait dataset, containing 124 subjects, three walking conditions and 11 views. The walking conditions contain normal (NM), walking with a bag (BG) and wearing a coat or jacket (CL). In our experiment, we employ medium-sample training (MT) in which the first 62 subjects are used for training and the remaining 62 subjects are left for testing. In the test sets, the first four sequences of the NM condition (NM #1-4) are kept in the gallery, and the remaining six sequences are divided into three probe subsets, i.e., NM subsets containing NM #5-6, BG subsets containing BG #1-2 and CL subsets containing CL #1-2.

We aggregated multiple gait datasets - CASIA-B [17], HID Gait [18], and Outdoor Gait [19] - to compile a substantial training set, aiming to develop a model with high accuracy for practical gait recognition scenarios. In total, we employed 306,974 silhouette images of size 112x112, collected from three datasets, to train the autoencoder in the first phase. Moreover, these datasets are combined with our collected data for training the second phase. These two phases are mentioned in Sections 3.2 and 3.3.

3.2. Training Encoder

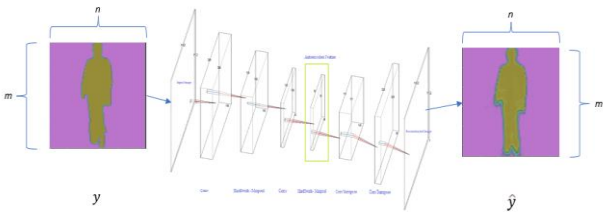


Figure 5. *Autoencoder's input and output*

The encoder in the Gait AL block is trained using an Autoencoder architecture mentioned in Section 2.1, where the loss function is defined as the pixel-wise difference between the input image y and its reconstructed output $\hat{y}$. This loss function is formulated in Equation (2) in which m, n are the image height and width. Figure 5 describes the model's architecture with its input and output. The autoencoder is trained for 20 epochs with Adam optimization [20] and the learning rate of 10^{-3}.

$$MSE = \frac{1}{m \times n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} (y(i,j) - \hat{y}(i,j))^2 \qquad (2)$$

3.3. Training GaitAL-EI

After training the encoder in the AL block during the first phase, we freeze its weights. The remaining parts of the proposed model, which are BiLSTM, and the MobileNet encoder, are then trained using the Triplet Loss function [21]. Moreover, to enhance class separability, Additive Angular Margin algorithm [22] is also employed in this work.

This block was trained and evaluated on a dataset that combines our collected dataset, CASIA-B, HID Gait, and Outdoor Gait datasets. The combined set includes 473 people, 19,796 clips and 1,551,065 video frames. This collection set is divided into two subsets: the train and the validation subsets. Table 3 illustrates the number of persons and clips for each subset.

Table 3. *Details of the combined set for training and validating Gait AL-EI*

Subset	#persons	# clips
Train	400	15,849
Validation	73	3,947

3.4. Gait Identification with GaitAL-EI embedding vectors

Gait matching within the database is the subsequent step in recognition. Given a gait video, after a feature vector is extracted using the trained model, the challenge lies in identifying the most similar vector within the database to perform identification. Performing the same similarity measure as in the GaitSet method, Euclidean distances between the embedding vector of a test sample in probe and the embedding vectors in the gallery are calculated. The smallest distance means the highest similarity of the test sample and its corresponding sample in the dataset. Therefore, it is used to give the identity of the input sequence.

The proposed gait recognition is evaluated with CASIA-B to compare with state-of-the-art methods such as GaitSet. For CASIA-B, we follow the same evaluation approach as in the GaitSet method to divide the gallery-proble set. Table 4 illustrates the number of persons and video sequences used for training GaitAL-EI and evaluating the performance of the proposed gait recognition system.

Table 4. *CASIA-B sets for gait recognition*

Subset	# Persons	# Sequences
Train	62	
Test	62	Gallery: 248 Probe: 372

4. Experiments

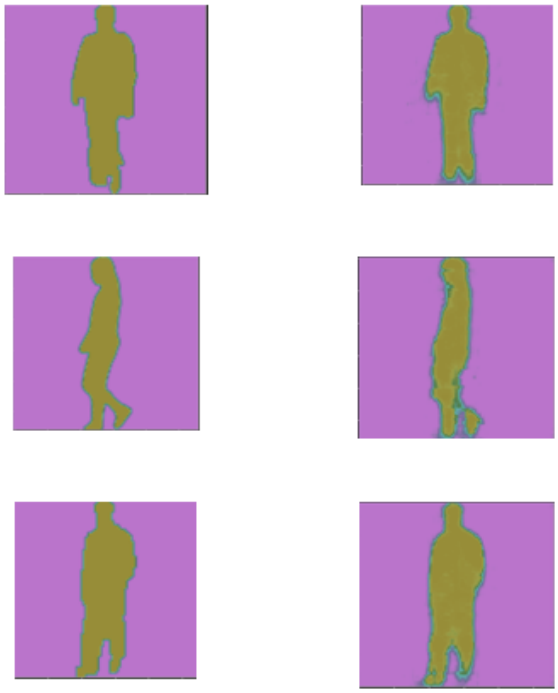
4.1. Evaluation measurements

To evaluate the auto-encoder, the mean square error is employed. Equation 2 illustrates this error. Moreover, recognition accuracy is employed to evaluate the

performance of our proposed gait recognition model. Recognition accuracy is defined as the ratio of correctly identified sequences to the total number of test sequences whose identities are included in the database.

4.2. Evaluation of the autoencoder in AL block

After training, the autoencoder is evaluated with the MSE measurement. Figure 6 illustrates some samples of original silhouette images and their reconstructions. As can be seen in the figure, the model reconstructs the upper body well, while some errors are obtained in the lower part of the human body, especially in the leg.



Original **Reconstruction**

Figure 6. *Autoencoder's results on three test silhouette images*

Table 5 presents the MSE distribution for 4,000 test set images, categorized into five levels: very good (<0.05), good (0.05–0.1), normal (0.1–0.15), bad (0.15–0.2), and very bad (>0.2). The visual results for three sample images are shown in Figure 6.

Table 5. *Autoencoder resilience rating*

MSE range	Amount	Percent
0-0.05	3089	76.46%
0.05-0.1	741	18.34%
0.1-0.15	171	4.23%
0.15-0.2	39	0.96%
>0.2	0	0%

It can be seen from Table 5 that the trained autoencoder has good reproducibility, with most reconstructed images (76.46%) having a very low MSE (0–0.05), indicating high accuracy in reconstruction. Another 18.34% of the images fall within the 0.05–0.1 range, which is still relatively small. 95% of the reconstructed images have an MSE below 0.1, showing that most images maintain a high quality. Only 4.23% of the images fall within the 0.1–0.15 range, while a mere 0.96% have errors between 0.15 and 0.2.

4.3. Evaluation of gait recognition with Gait AL-EI

We compare our proposed method with GaitSet [8]. Table 6 illustrates comparative results of our proposed model on the CASIA-B dataset with the same testing and evaluation approaches as GaitSet [8]. As can be seen in this table, in cases of normal walking, our model outperforms GaitSet consistently across most angles. The highest accuracy of 97.6% is achieved at 144°, and the lowest at 36° (92.7%). Moreover, with walking with a bag (BG), our model again outperforms GaitSet in most view angles, especially in frontal and oblique views, for example, at 90°, our method gives 93.5% compared to 76.7% of GaitSet. However, our model underperforms in coat-wearing scenarios (CL), possibly due to occlusion or silhouette distortion caused by loose clothing. However, this is the most challenging condition, where both models show reduced performance.

Table 6. *Comparative results on MT setting of CASIA-B dataset*

Gallery NM#1-4		$0^o - 180^o$											*Mean*
Probe	**Model**	0^o	18^o	36^o	54^o	72^o	90^o	108^o	126^o	144^o	162^o	180^o	
NM #5-6	GaitSet	89.7	**97.9**	**98.3**	**97.4**	92.5	90.4	93.4	**97.0**	**98.9**	95.9	86.6	94.3
	Ours	**96.8**	96.8	92.7	94.3	**96.7**	**97.5**	**97.6**	96	97.6	**96.8**	**96.0**	**96.3**
BG #1-2	GaitSet	79.9	**89.8**	**91.2**	**86.7**	81.6	76.7	81.0	**88.2**	**90.3**	**88.5**	73.0	84.3
	Ours	**87.0**	78.9	72.4	85.9	**91.9**	**93.5**	**91.1**	84.7	87.1	82.1	**86.3**	**85.5**
CL #1-2	GaitSet	**52.0**	**66.0**	**72.8**	**69.3**	63.1	**61.2**	**63.5**	**66.5**	**67.5**	**60.0**	**45.9**	**62.5**
	Ours	35.5	31.4	42.7	44.4	**63.9**	58.9	60.1	49.2	41.4	47.6	36.3	46.5

5. Conclusion

This paper proposes a gait recognition model, called GaitAL-EI, using two branches to extract the dynamics of human gait using sequences of silhouettes and gait energy images. Comparative results on the public dataset show superior performance under normal walking and walking with bag conditions, making it suitable for general and partially occluded gait scenarios. However, further improvement is needed for wearing coat or jacket conditions, possibly employing clothing-invariant feature extraction or multi-modal inputs.

Acknowledgments: This work was supported by Fujikin Fund and The University of Danang - University of Science and Technology, code number of Project: T2023FSF-02-04.

REFERENCES

[1] Sarkar, Sudeep, P. Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," IEEE transactions on pattern analysis and machine intelligence 27, no. 2, 2005, pp. 162-177. https://doi.org/10.1109/TPAMI.2005.39

[2] Zhang, Shaoxiong, Yunhong Wang, Tianrui Chai, Annan Li, and Anil K. Jain, "Realgait: Gait recognition for person re-identification," *arXiv preprint arXiv:2201.04806,* 2022.

[3] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3D representations and a benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20228–20237

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

[5] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006, https://doi.org/10.1109/TPAMI.2006.38.

[6] S. Tong, Y. Fu, X. Yue, and H. Ling, "Multi-view gait recognition based on a spatial–temporal deep neural network," *IEEE Access*, vol. 6, pp. 57 583–57 596, 2018.

[7] W. Xing, Y. Li, and S. Zhang, "View-invariant gait recognition method by three-dimensional convolutional neural network," *J. Electron. Imaging*, vol. 27, no. 1, pp. 013015, Jan. 2018, https://doi.org/10.1117/1.JEI.27.1.013015.

[8] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 8126–8133, 2019, https://doi.org/10.1609/aaai.v33i01.33018126.

[9] K. Wang, L. Liu, W. Zhai, and W. Cheng, "Gait recognition based on GEI and 2D-PCA," *Chin. J. Image Graph.*, vol. 14, no. 4, pp. 695–700, 2009.

[10] S. C. Bakchy, M. R. Islam, M. R. Mahmud, and F. Imran, "Human gait analysis using gait energy image," *arXiv preprint* arXiv:2203.09549, 2022.

[11] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu and J. Zhou, "Gait recognition in the wild: A benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (ICCV), 2021, pp. 14789–14799, https://doi.org/10.1109/ICCV48922.2021.01452.

[12] J. E. Boyd and J. J. Little, "Gait recognition, silhouette-based," in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Boston, MA: Springer, 2015, pp. 813-820. https://doi.org/10.1007/978-0-387-73003-5_263.

[13] A. Sokolova and A. Konushin, "Methods of gait recognition in video," *Program. Comput. Softw.*, vol. 45, no. 4, pp. 213–220, 2019, https://doi.org/10.1134/S0361768819040091.

[14] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," *arXiv preprint* arXiv:1801.02143, 2018.

[15] A. Howard, M. Sandler, G. Chu, L.C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, and Q.V. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (ICCV), 2019, pp. 1314–1324, https://doi.org/10.1109/ICCV.2019.00140.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), 2018, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.

[17] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *Proc. IEEE Int. Conf. Image Process.* (ICIP), 2011, pp. 2073–2076.

[18] A. Y. Johnson, J. Sun, and A. F. Bobick, "Predicting large population data cumulative match characteristic performance from small population data," in *Proc. Audio- and Video-Based Biometric Person Authentication* (AVBPA), vol. 4, 2003, pp. 821–829, https://doi.org/10.1007/3-540-44887-X_95.

[19] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, pp. 1–14, 2018.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint* arXiv:1412.6980, 2014.

[21] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Similarity-Based Pattern Recognition (SIMBAD)*, Copenhagen, Denmark, Oct. 2015, pp. 84–92, https://doi.org/10.1007/978-3-319-24261-3_7.

[22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), 2019, pp. 4690–4699, https://doi.org/10.1109/CVPR.2019.00482.