




Gait Recognition Using CGAN and EfficientNet Deep Neural Networks

Entesar T. Burges¹^a, Zakariya A. Oraibi²^b and Ali Wali³^c

¹National School of Electronics, University of Sfax, Sfax 3029, Tunisia

²Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, Iraq

³Research Groups in Intelligent Machines, National Engineering School of Sfax, University of Sfax, Sfax 3029, Tunisia

Keywords: Gait Recognition, Deep Learning, EfficientNet, CGAN.


Abstract: The objective of gait recognition is to use a visual camera to identify a person from a distance using a visual camera by their distinctive gait. However, the accuracy of this recognition can be impacted by things like carrying a bag and changing clothes. The framework for human gait recognition system presented in this study is based on deep learning and EfficientNet Deep Neural Network. The proposed framework includes three steps. The first step involves extracting silhouettes. The second step involves computing the gait cycle, and the third involves calculating gait energy. Depending on the conditional generative adversarial networks and EfficientNet Deep Neural Network. In the first step, silhouette images are extracted using Gaussian mixture-based background algorithm. The segmentation of the gait cycle is estimated by measuring the silhouette's bounding box's length and width, then calculating gait energy. Images resulted from the previous stage are used as input to the conditional generative adversarial networks to generate Gait Energy Image (GEI). EfficientNet is employed as an identification discriminator in this work. The suggested framework was evaluated on a challenging gait dataset called CASIA-B, and scored an accuracy of 97.13%. The framework introduced in this paper outperformed techniques in literature in accuracy.


1 INTRODUCTION


The unique strolling style of individuals can be utilized as a natural identifier due to the difficulty of replicating it (Asif et al., 2022). Identifying humans based on their gait is a biometric recognition procedure that leverages stride features to distinguish an individual from a distance without requiring physical contact. This method stands apart from traditional biometric systems, as it allows for recognition at a distance, making it particularly useful in various applications, including surveillance and security. Human aspects such as spatial, static, and temporal characteristics are integral to gait analysis (Iwashita et al., 2014). Unlike other biometric identification methods like fingerprints, facial recognition, iris scans, and palm prints, gait analysis has shown superior discriminative qualities, particularly in dynamic environments where other methods may falter (Li et al., 2022). As a result, it has gained significant traction within the field of Machine Learning (ML), where algorithms are increasingly trained to recognize and classify gait patterns. Recent advancements in real-world applica-

tions, including forensic detection, video monitoring, and crime prevention, have drawn considerable attention to gait recognition systems. These systems have also evolved to include low-contrast gait identification methods, which can be particularly useful in challenging visibility conditions (He et al., 2016; Yu et al., 2017). The ability to identify individuals based on their walking patterns presents a non-invasive alternative to traditional biometric methods. However, gait recognition is inherently a behavioral-based biometric detection method, which can lead to lower accuracy compared to physical biometric methods. This reduced accuracy can be attributed to various covariate factors, including lighting conditions, walking pace, the clothing individuals may be wearing, and variations in viewing angles (Tan and Le, 2021; Yu et al., 2006; Liao et al., 2020; Isola et al., 2017).

These variables significantly affect gait characteristics and make gait recognition a complex task (Dupuis et al., 2013). To address these challenges, researchers have developed various strategies to enhance the extraction and analysis of gait data. These strategies can be broadly categorized into two groups: model-based and appearance-based approaches (Alvarez and Sahonero-Alvarez, 2018; Wang and Yan, 2020). The model-based method re-

^a <https://orcid.org/0000-0002-3307-3498>

^b <https://orcid.org/0000-0003-4965-1579>

^c <https://orcid.org/0000-0002-8423-7923>

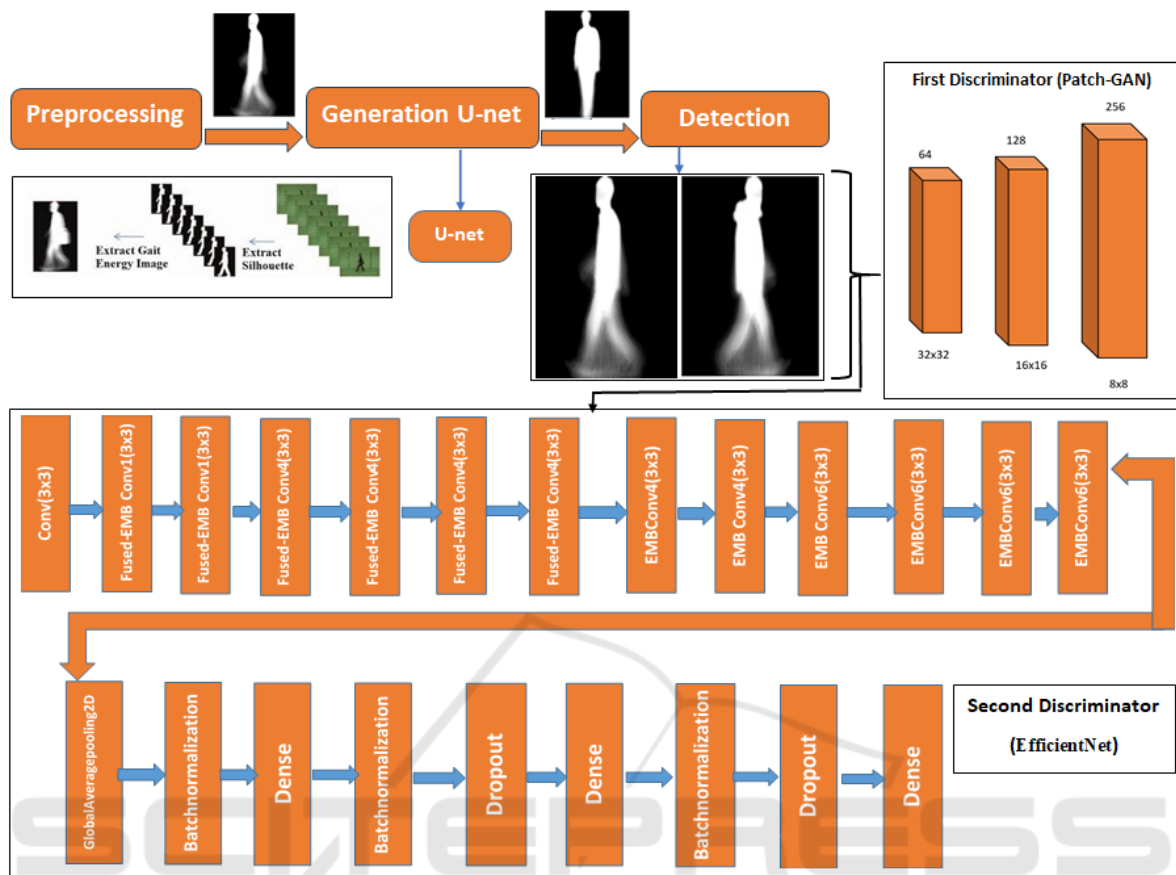


Figure 1: Framework of the proposed methodology.

lies on constructing models of human body shape and motion to categorize gait. However, this approach often suffers from reliability issues due to the aforementioned covariate factors, particularly when extracting gait patterns from low-resolution images. Conversely, the appearance-based method utilizes human silhouettes, which can be retrieved from low-resolution images, though they remain sensitive to changes in illumination and clothing (Tan and Le, 2021, Tan, 2019, Howard, 2017). The work proposed in this paper falls under the second category of gait recognition techniques. Specifically, it employs silhouettes to mitigate distortion caused by varying illumination conditions and differences in clothing colors and textures, as noted in (Ramachandran et al., 2017). During the recognition stage, these silhouettes serve as feature inputs, allowing for effective identification. This approach of using low-resolution images not only demands less computational power but also reduces memory usage without significantly compromising performance (Russakovsky et al., 2015). Such advantages have likely contributed to the rising popularity of model-free approaches within gait recog-

inition methodologies. To tackle the problem of appearance change, we employ a conditional generative adversarial network (CGAN) that generates view-invariant representations. The CGAN architecture consists of a generator modeled as U-Net, accompanied by two discriminators: a patchGAN and an EfficientNet-based Deep Neural Network. This design aims to enhance the robustness of gait recognition across varying conditions. The remaining sections of our paper are organized as follows: Section 2 describes the proposed methodology for recognizing gait using the EfficientNet model. Section 3 presents experimental results obtained from our proposed work. Finally, conclusions and directions for future research are discussed in section 4.

2 THE PROPOSED METHODOLOGY

The proposed EfficientNet model based human gait recognition system relies on multi-views video. In order to apply it, few steps must be performed to pre-

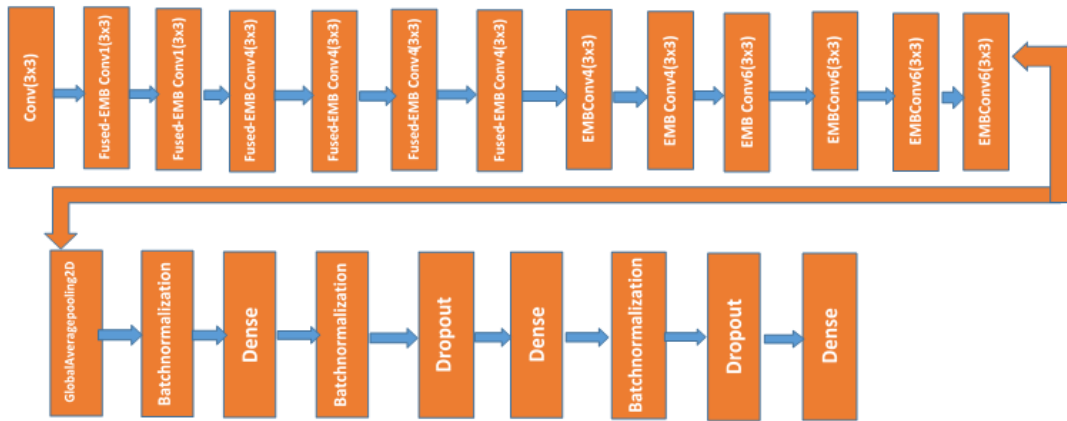


Figure 2: Modified EfficientNetV2L.

Table 1: Specifications of EfficientNet architecture used in this paper.

Layers	Units	Activation Function
EfficientNetV2L	Custom Network	
Dense	120	relu
Dense	120	relu
Dense	1	sigmoid
Batch Normalization	-	-
Dropout	0.2	-

pare the images for the final model. These steps are: Data preprocessing, invariant feature generation, and the final step is classification using EfficientNet. It is worthy to mention that our previous work included the use of a hybrid model of ResNet and CGAN (Talal et al., 2023). In addition, we also proposed a model that uses LSTM and CNN recently (Borges et al., 2024). Both models achieved high prediction on CASIA-B database.

2.1 Preprocessing

Image silhouettes in a single walking cycle must be obtained first. Using the given gait sequence, the method outlined in (He et al., 2016) is applied to produce human silhouettes. Ultimately, all of the images undergo size normalization and horizontal alignment. These images are processed by applying dilation and erosion to remove noise. Next, the bounding box of the silhouette is measured for length and width, and the interval between the two highest lengths is calculated to estimate the gait cycle segmentation. Next, the Gait Energy Image (GEI) is computed using Equation 1 and the samples shown in Fig. 1.

$$GEI(v, w) = \sum_{t=1}^M I(v, w, t) / M \quad (1)$$

Image coordinates are represented as v and w , M

is the number of images in a whole gait cycle, I is the image, and t is the gait cycle frame number.

2.2 Generating Invariant Features

With a U-Net-based architecture, a CGAN model is presented that can convert the representations of gait from any viewpoint and appearance condition into side-view representations under standard conditions.

Input Data: Before the Generative Adversarial Network (GAN) can be trained, the data must first be organized. The Gait Energy Images (GEIs) from all viewpoints in the sequences of normal bag, carrying, and walking and putting on a coat are designated as the source information. The GEIs from side views at a normal walking angle of 90 degrees are designated as the goal data. After that, 40 million source-target representation pairs were collected so that the GAN could be trained.

2.3 Conditional Generative Adversarial Nets

The basic GAN model was observed to not be controlled easily because it does not explicitly incorporate attributes or conditions. To address this, Mirza and Osindero (Yu et al., 2017) proposed Conditional GAN (CGAN) which can guide the classification process. This condition y can be connected to class labels, properties from multiple modalities, or other external data that is provided as an additional input to both the generator and discriminator networks. As a result, this allows the model to generate or classify samples based on the provided condition, making the GAN more controllable and applicable to a wider range of tasks. In summary, the goal of the conditional GAN is to leverage the additional conditioning input y to guide the adversarial training process

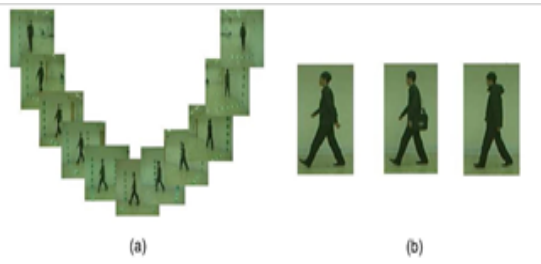


Figure 3: Sample frames of CASIA-B database. Images on the left represent 11 different capturing views. Images on the right represent 3 walking conditions.

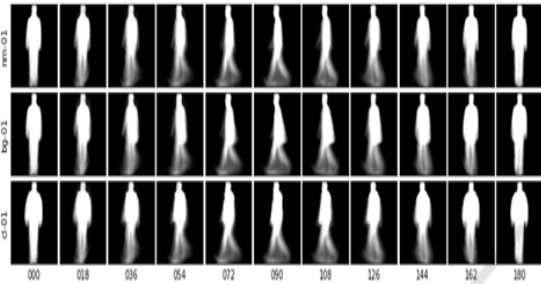


Figure 4: 11 views walking sequences from CASIA-B dataset.

and enable more controlled generation or classification compared to the basic GAN framework.

$$L_{CGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,y}[\log(1 - D(x, G(x, y)))] \quad (2)$$

Whereby the generator G attempts to reduce this function. The discriminator D , on the other hand, seeks to maximize it. Studies previously have also shown that results that are fairly close to reality can be obtained by combining the historical loss with more conventional loss functions.

$$L_{L1}(G) = E_{x,y}[||y - G(x, y)||_1] \quad (3)$$

The definition of final objective can be as follows:

$$G^* = \arg \min_G \max_D L_{CGAN}(G, D) + \lambda L_{L1}(G) \quad (4)$$

λ is a regularizing hyperparameter. For instance, the CGAN produces highly defined outputs when λ is used, but the classification accuracy drops.

2.4 Recognition Stage

EfficientNet deep neural network was developed to address the shortcomings of conventional classification techniques, such as the small sample size issue and the lack of discriminative information in the means of classes. EfficientNet is a family of convolutional neural network models that are designed to

achieve state-of-the-art accuracy with fewer parameters and computational resources compared to other popular models such as ResNet or Inception. The EfficientNet architecture uses a combination of efficient building blocks, including depth-wise separable convolutions and squeeze-and-excitation modules, to optimize the trade-off between model size and accuracy. This makes EfficientNet appropriate for mobile or edge devices which require resource constraining.

2.5 EfficientNetV2L

Introduced in 2021, EfficientNetV2L is a neural network architecture that belongs to the EfficientNetV2 family of models, which was developed to optimize the trade-off between model size, speed, and accuracy in image classification tasks (Tan and Le, 2021). The "L" in EfficientNetV2L stands for "Lite," indicating that it is a more compact and lightweight variant of the original EfficientNetV2 model (Tan, 2019). This design choice aims to provide a more efficient and faster training process while maintaining high levels of accuracy, making it particularly suitable for applications where computational resources are limited. EfficientNetV2L employs several innovative techniques to achieve its objectives. It utilizes depth wise separable convolutions, which reduce the number of parameters and computations compared to traditional convolutions, thus enhancing the model's efficiency (Howard, 2017). Additionally, the incorporation of squeeze-and-excitation blocks allows the model to recalibrate channel-wise feature responses adaptively, further improving its representational power (Hu et al., 2018). The use of swish activation functions contributes to better training dynamics and has been shown to outperform traditional activation functions such as ReLU (Ramachandran et al., 2017). The model has been extensively trained on various image classification datasets, including ImageNet, CIFAR-10, and CIFAR-100, where it has achieved state-of-the-art performance (Russakovsky et al., 2015). Its robust architecture not only excels in classification tasks but also demonstrates considerable effectiveness in transfer learning scenarios, being applicable to other computer vision tasks like object detection and segmentation (Yosinski et al., 2014). These attributes position EfficientNetV2L as a powerful neural network architecture that provides an efficient and lightweight solution for image classification tasks while ensuring high accuracy. Overall, EfficientNetV2L represents a significant advancement in the field of deep learning, addressing the growing demand for models that are both performant and resource-conscious. Its innovative approaches and

impressive benchmark results make it a compelling choice for researchers and practitioners alike.

2.6 The Proposed Architecture

Fig. 1 represents the architecture of our proposed model. The EfficientNetV2L network has been modified by freezing the pertained weights and rebuilding the top by adding four convolution layers (GlobalAveragePooling2D, Batch Normalization, Dropout, and Dense) and an Adam optimizer. Our system uses grayscale images as input, having $150 \times 150 \times 3$ sizes. Then we added our custom network to three convolution layers. The first two layers with the relu activation function, and the last one with sigmoid, see Fig. 2. The details of the network structures are shown in Table 1.

3 EXPERIMENTAL RESULTS

The challenging CASIA-B database (Yu et al., 2006) was captured indoors. As the subject is walking 11 cameras are positioned around the person's left side. The two closest view directions are separated by 18 degrees. Sample frames of the database are shown in Fig. 3. Fig. 4 shows the various views of a walking sequence sample. We applied the experimental approach recommended in (Amin et al., 2021; Alvarez and Sahonero-Alvarez, 2020; Isola et al., 2017) in order to accurately compare the proposed strategy with cutting-edge methods. The dataset is divided in half with the first 62 participants completed six regular, two carrying-bag, and two wearing-coat sequences, making up the training set. In the test phase, the remainder of 62 people were employed. In order to evaluate the variations in view, carrying, and clothing circumstances, sequences denoted as ("nm1"), were considered as gallery set during experiments.

3.1 Model Parameters

The generator and two discriminators make up the two components of CGAN as previously illustrated in Fig. 1. GEI generator makes use of U-Net architecture. Hence, a similar setup to (Liao et al., 2020) was depicted in the experiments which is composed of 2 elements: The encoder and the decoder. In the encoder, there are 4 convolutional layers. Since a U-Net architecture is used, the number of channels are doubled in the decoder. The reason for that is because U-Net concatenates activations between layers i and n . The first layer of the encoder is different from the others in that it does not use the Batch Norm. A convo-

lutional layer and Tanh layer are added after the final decoder layer in order to account for the channel numbers of the output. The reason for utilizing the first discriminator is discriminate fake photos from the real ones. In addition, PatchGAN (Isola et al., 2017) was also employed to penalize specific areas from the image as fake or real. This will allow us to focus on GEI areas that are related to image parts which are more resistant to appearance changes (Dupuis et al., 2013; Alvarez and Sahonero-Alvarez, 2018). The second discriminator is identification, which makes use of EfficientNetV2L architecture. The identification discriminator computes the chance that the data pair belongs to the same individual by using the original gait image sequence and the created output gait image sequence as one training data pair. The output of identification discriminator is either 0 or 1 depending on the training data.

3.2 Training Stage

Hyper parameters were selected carefully in the experiments to ensure superior performance. The optimizer utilized during experiments is Adam while training the generator and the first discriminator. 0.0002 learning rate was used with momentum parameters of $1 = 0.5$, $2 = 0.999$. We found that sufficient performance was achieved after 20 training epochs when utilizing $\beta = 100$. RMS prop optimization with a binary cross entropy loss function and a learning rate of 0.001 was used to train the second discriminator. Since both PatchGAN and U-net were trained from scratch, the weights were derived from a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. Because each participant in the CASIA database only contains a certain amount of sequences. The image size in EfficientNetV2L is $150 \times 150 \times 3$, while in U-net and PatchGAN it is 64×64 . It runs on the Windows 11 OS and the Colab Pro (with GPU) environment using the Python programming language.

3.3 Results and Discussion

The EfficientNetV2L classifier approach on the spatial temporal feature GEI produced by Conditional Generative Adversarial Network (CGAN) is presented in this research. Has been suggested for systems that recognize gaits. The suggested techniques have been statistically evaluated using the CASIA-B dataset. 124 individuals make up the CASIA-B dataset. The dataset's 124 people were divided into 62 groups: one for the training set and another for the testing set. In order to assess the suggested model, the accuracy, F1score, precision, and recall are taken

Table 2: Results of applying our approach.

	Acc.	Prec.	Recall	F1-score
Training data	97.18	0.97	0.97	0.97
Validation data	97.17	0.97	0.97	0.97
Testing data	97.13	0.97	0.97	0.97

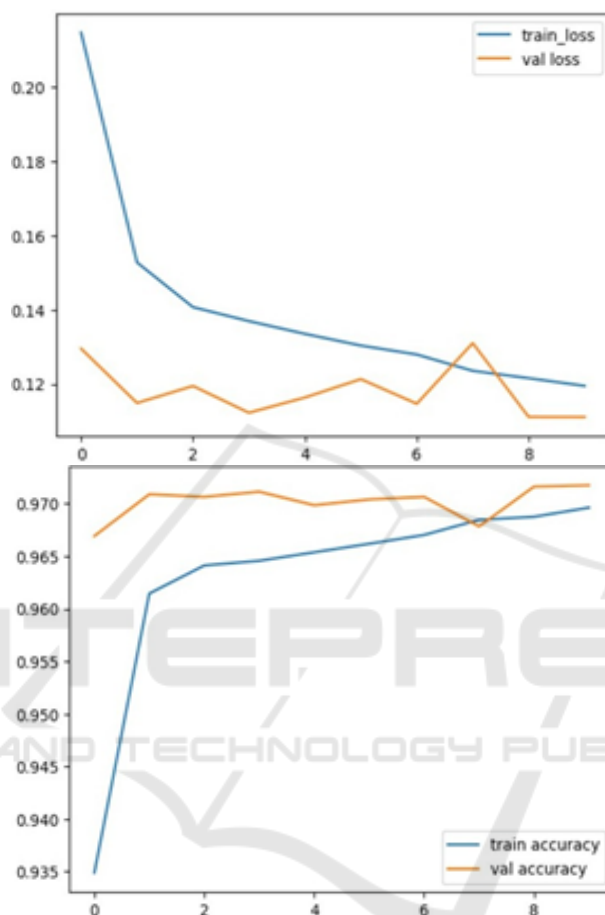


Figure 5: Performance of training stage.

into account (Hossain et al., 2010) as shown in Table 2. Initially, an image segmentation procedure was performed using the Gaussian Mixture methodology, which was found to be effective for fundus gait picture segmentation in previous research (Huang et al., 2021). This technique was used to separate the silhouettes from the RGB gait images. Fig. 1 displays a few of the segmentation outcomes. Following processing, each image is subjected to size normalization and horizontal alignment.

Next, dilation and erosion are used to eliminate any noise from the images. First, the length and width of the bounding box drawn around the silhouette are measured, and then the interval between the two largest lengths is calculated to estimate the gait cycle segmentation that follows. Next, the Gait

Energy Image (GEI) is computed using the samples shown in Fig. 1. To generate invariant characteristics, CGAN was used to develop GEI for a variety of scenarios, generate typical side views photos for multiple viewpoints, create GEIs for carrying a bag and wearing a coat, and then discriminate the first 62 individuals using the first discriminator. A second discriminator, which indicated whether or not the generated GEI belonged to the same person, was used to differentiate the test set.

Based on previously discussed altering neural network settings, the accuracy of the EfficientNetV2L classifier has been evaluated in section 4.D. The testing dataset yielded an accuracy of 97.13%. The performance of our proposed technique during training can be evaluated as in Fig. 5 where training loss and

Table 3: Comparison with the best performing methods.

Authors	Proposed Method	Accuracy
Wang et al. (Wang and Yan, 2020) (2020)	Ensemble learning	92.0%
Wang et al. (Wang and Yan, 2020) (2020)	LSTM	95.0%
Javaria Amin et al. (Amin et al., 2021) (2021)	Conv-BiLSTM	96.0%
Our Proposed Framework (2024)	CGAN + EfficientNet	97.13%

training accuracy are shown after 10 epochs. The total accuracy data acquired, as presented in Table 3, has been used to compare the performance of our proposed strategy with that of existing methods. Based on the comparative analysis presented in Table 8, it is evident that our study is outperforming the other studies included in the list. Wang et al.'s (Li et al., 2020) ensemble learning method for categorizing human gait yielded values of 0.95 and 0.92 CPR, respectively. Wang et al. (Saleem et al., 2021) employed LSTM to learn the sequential patterns of the input images and were able to achieve 0.95 CPR on the CASIA-B dataset. Javaria Amin et al. (Amin et al., 2021) used the Conv-BiLSTM model to generate a CPR of 0.96, of which 0.88 (person with bag) and 0.92 (normal) were reached for the classification of various human kinds.

4 CONCLUSIONS

This work developed a method for a gait identification system based on an EfficientNet classifier and a CGAN architecture using a U-Net. The goal was to overcome appearance variations caused by changes in clothing, carrying conditions, and viewing angles. Since it can be difficult to distinguish between human gait patterns and data-specific quirks, the researchers proposed an algorithm with a generator that creates normal images at a 90-degree angle and two discriminators- one to determine if the images are of the actual person and another to discriminate between fake and real images generated by the generator. The research demonstrated that this design improved the accuracy of gait identification compared to previous approaches, which often struggled with differences in bags and coats. This makes the technology suitable for advanced surveillance applications and other real-world uses. Future work will benchmark performance on larger datasets and explore handling more challenging situations like temporal fluctuations. In addition to improving accuracy across view changes between the probe and gallery sets, expanding the num-

ber of subjects is needed to yield more reliable findings. More sophisticated and powerful models will also be required for effective cross-view gait identification.

ACKNOWLEDGEMENTS

Authors would like to express gratitude to University of Basrah that Entesar B. Talal is associated to for sponsoring the fellowship to acquire the PhD degree.

REFERENCES

- Amin, J., Anjum, M. A., Sharif, M., Kadry, S., Nam, Y., and Wang, S. (2021). Convolutional bi-lstm based human gait recognition using video sequences. *Comput. Mater. Contin.*, 68(2):2693–2709.
- Asif, M., Tiwana, M. I., Khan, U. S., Ahmad, M. W., Qureshi, W. S., and Iqbal, J. (2022). Human gait recognition subject to different covariate factors in a multi-view environment. *Results in Engineering*, 15:100556.
- Burges, E. T., Oraibi, Z. A., and Wali, A. (2024). Gait recognition using hybrid lstm-cnn deep neural networks. *Journal of Image and Graphics*, 12(2).
- Dupuis, Y., Savatier, X., and Vasseur, P. (2013). Feature subset selection applied to model-free gait recognition. *Image and vision computing*, 31(8):580–591.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Hossain, M. A., Makihara, Y., Wang, J., and Yagi, Y. (2010). Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291.
- Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

- Huang, Z., Xue, D., Shen, X., Tian, X., Li, H., Huang, J., and Hua, X.-S. (2021). 3d local convolutional neural networks for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14920–14929.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.
- Iwashita, Y., Ogawara, K., and Kurazume, R. (2014). Identification of people walking along curved trajectories. *Pattern Recognition Letters*, 48:60–69.
- Li, H., Qiu, Y., Zhao, H., Zhan, J., Chen, R., Wei, T., and Huang, Z. (2022). Gaitslice: A gait recognition model based on spatio-temporal slice features. *Pattern Recognition*, 124:108453.
- Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., and Ren, M. (2020). End-to-end model-based gait recognition. In *Proceedings of the Asian conference on computer vision*.
- Liao, R., Yu, S., An, W., and Huang, Y. (2020). A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069.
- Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.
- Saleem, F., Khan, M. A., Alhaisoni, M., Tariq, U., Armghan, A., Alenezi, F., Choi, J.-I., and Kadry, S. (2021). Human gait recognition: A single stream optimal deep learning features fusion. *Sensors*, 21(22):7584.
- Talal, E. B., Oraibi, Z. A., and Wali, A. (2023). Gait recognition using deep residual networks and conditional generative adversarial networks. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1179–1185. IEEE.
- Tan, M. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.
- Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR.
- Wang, X. and Yan, W. Q. (2020). Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International journal of neural systems*, 30(01):1950027.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.
- Yu, S., Chen, H., Garcia Reyes, E. B., and Poh, N. (2017). Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 30–37.
- Yu, S., Tan, D., and Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th international conference on pattern recognition (ICPR'06)*, volume 4, pages 441–444. IEEE.