iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83339

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Gait Recognition Using GaitFormer on the CASIA-B Dataset

K. Mokshagna Anurag[1], M. P. S S S Hari Chandra Hlada[2], N. Sai Kishore[3], P. Sai Lalith[4], Dr. P. Suryaprasad[5]
[1, 2, 3, 4]*Dept. of ECE, MVGR College of Engg. (A), Vizianagaram, India*
[5]*Professor, Dept. of ECE, MVGR College of Engg. (A), Vizianagaram, India*

Abstract: ***Gait recognition is a biometric identification tech-nique that identifies individuals based on their walking patterns. Unlike fingerprint or facial recognition, gait can be observed at a distance without requiring the subject's cooperation, making it useful for surveillance and security. This paper presents a gait recognition system based on the GaitFormer architecture, which uses convolutional neural network (CNN) layers for spatial feature extraction and Transformer-based attention for temporal modelling. The system is trained and evaluated on the CASIA-B gait dataset, which contains multi-view silhouette sequences un-der different walking conditions. The input pipeline preprocesses silhouette sequences and computes Gait Energy Images (GEIs) as compact gait cycle representations. The model learns spatial features through convolutional layers and captures temporal dependencies using self-attention. The system was implemented in Python using TensorFlow/Keras on Google Colab. Evaluation results show a test accuracy of 88.89%, Rank-1 accuracy of 88.89%, and Rank-5 accuracy of 100.00%.***

Keywords: Gait Recognition, GaitFormer, CASIA-B Dataset, Transformer, Convolutional Neural Network, Biometric Identification.

I. INTRODUCTION

Biometric identification systems use measurable human characteristics to tell one person from another. Among the different biometric modalities — fingerprints, iris patterns, and facial features — gait recognition stands out because it does not require physical contact or close proximity. A person's walking pattern is shaped by physiological and behavioural factors that are hard to copy, making gait a useful cue for identification [1].

Gait recognition has one clear advantage over other bio-metric traits: it can be captured at a distance using ordinary video cameras, even without the subject knowing. This makes it well-suited for surveillance, forensic analysis, and access control [2]. That said, gait recognition is still a difficult problem because factors like viewing angle, clothing, and carried objects change the appearance of walking and add noise to the data.

Recent work in deep learning — especially CNNs and Transformers — offers ways to handle these issues. CNNs are good at picking up local spatial features from individual frames, while Transformers can model relationships across an entire sequence through self-attention. Combining these two approaches can lead to better gait recognition systems.

This paper presents a gait recognition system built using the GaitFormer architecture and tested on the CASIA-B dataset. The system uses CNNs to extract spatial features from sil-houette frames and a Transformer encoder to learn temporal patterns across the walking sequence.

A. Problem Statement

The goal of this work is to build and test a gait recognition system using the GaitFormer architecture on the CASIA-B dataset. The system should be able to recognise people across different walking conditions by learning useful spatial and temporal features from silhouette sequences.

B. Objectives

The specific objectives are:

1) To preprocess silhouette sequences from the CASIA-B dataset and generate Gait Energy Images (GEIs) as input representations.
2) To implement the GaitFormer architecture integrating CNN-based spatial feature extraction with Transformer-based temporal attention.
3) To train and evaluate the model using standard classifica-tion and retrieval metrics.

4) To compare with a CNN + BiLSTM baseline.
To discuss the strengths and weaknesses of the approach.

II. LITERATURE REVIEW

A. Gait Recognition

Gait recognition methods generally fall into two categories: model-based and appearance-based. Model-based methods try to build a representation of the human body using joint positions or body parameters [5]. Appearance-based methods work directly with the visual look of the walking person, typically using silhouette images [6]. This work follows the appearance-based approach.

B. Gait Energy Image

The Gait Energy Image (GEI) is a compact representation computed by averaging aligned and normalised silhouette frames over one full gait cycle. It encodes both static body shape and dynamic motion information in a single two-dimensional image [6]. Using GEI reduces the amount of input data while keeping the important features of the walking pattern.

C. CNN-Based Methods

CNNs work well for extracting local spatial patterns using convolutional filters. For gait recognition, CNNs can learn features like body shape and limb positions from silhouette frames. GEINet [8] showed that CNNs can handle view-invariant gait recognition, and other studies confirmed that CNN-based models beat traditional hand-crafted feature meth-ods on standard gait datasets [7].

D. Recurrent and Temporal Models

RNNs and their variants like LSTM and BiLSTM have been used to model how gait changes over time [9]. A common setup pairs a CNN for spatial features with an LSTM for temporal modelling. This works reasonably well, but LSTMs can have trouble with long sequences and sometimes miss patterns that span the full walking cycle.

E. Transformer-Based Methods

Transformers were originally designed for language tasks, but they have been applied to vision problems with good results [10]. The self-attention mechanism lets the model look at all positions in a sequence at once, which helps with learning dependencies across many frames. A few recent pa-pers have tried Transformers for gait recognition and reported improvements [11], [12].

F. Related Work

GaitSet [13] treats gait as a set of independent frames and pools them together. GaitPart [14] splits the body into parts and extracts features from each region. GaitGL [15] mixes global and local features. GaitFormer builds on these ideas by using self-attention to handle temporal modelling in a more flexible way.

III. DATASET DESCRIPTION

The CASIA Gait Database – Dataset B (CASIA-B) is one of the most commonly used datasets for gait recognition research. It was collected by the Institute of Automation, Chinese Academy of Sciences [16]–[18]. Table I lists its main properties.

TABLE I
SUMMARY OF THE CASIA-B DATASET

Property	Description
Subjects	124
Walking Conditions	Normal (NM), Bag (BG), Coat (CL)
Sequences/Subject	10 (6 NM + 2 BG + 2 CL)
Viewing Angles	11 (0°–180°, 18° step)
Data Format	Pre-segmented silhouettes

The dataset captures three walking conditions: Normal (NM) walking without additional items, Bag-carrying (BG) walking while carrying a bag, and Coat-wearing (CL) walk-ing while wearing a heavy coat. Each sequence is recorded from 11 camera viewpoints from 0° to 180° at 18° intervals. The silhouette sequences were preprocessed as described in Section IV.

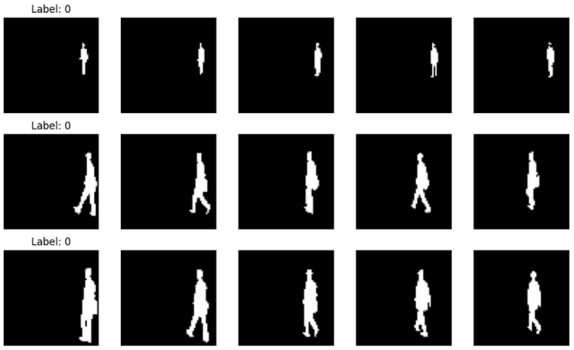


Fig. 1. Sample gait silhouettes from the CASIA-B dataset under different walking conditions and viewing angles.

IV. METHODOLOGY

The proposed pipeline consists of four stages: data prepro-cessing, spatial feature extraction, temporal modelling using attention, and classification.

Fig. 2. Overview of the gait recognition pipeline.

A. *Data Preprocessing*

1) *Silhouette Normalisation:* The CASIA-B dataset comes with pre-segmented silhouettes. Each frame was resized to a fixed resolution, pixel values were scaled to [0, 1], and silhouettes were centred based on the foreground centroid so that the person appears in roughly the same position across frames.
2) *Gait Energy Image Computation:* The GEI is computed by averaging aligned silhouette frames over a complete gait cycle. For T silhouette frames $\{S_1, S_2, \ldots, S_T\}$:

$$\mathrm{GEI}(x, y) = \frac{1}{T} \sum_{t=1}^{T} S_t(x, y) \qquad (1)$$

where $S_t(x, y)$ is the pixel value at position (x, y) in frame t. The GEI captures both static body shape and dynamic motion patterns.

3) *Sequence Preparation:* Sequences are assembled into fixed-length input tensors. Shorter sequences are padded and longer sequences are truncated to maintain uniformity.

B. *GaitFormer Architecture*

The GaitFormer model consists of a CNN-based spatial feature extractor and a Transformer-based temporal attention module.

1) *CNN-Based Spatial Feature Extraction:* The CNN pro-cesses individual silhouette frames or GEIs through convolu-tional layers with activation functions and pooling operations.

For an input frame S_t:

$$\mathbf{f}_t = \mathrm{CNN}(S_t) \in \mathbb{R}^d \qquad (2)$$

where d is the size of the feature vector. The earlier layers tend to pick up edges and textures, while the deeper layers learn higher-level shapes like body parts.

2) Transformer-Based Temporal Attention: The spatial fea-ture sequence $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T$ is processed by the Transformer encoder using multi-{head self-attention} (MHSA).

Positional Encoding: Since Transformers lack inherent sequence ordering, sinusoidal positional encoding is added:

$$PE(pos, 2i) = \sin \frac{pos}{10000^{2i/d}} \quad (3)$$

$$PE(pos, 2i+1) = \cos \frac{pos}{10000^{2i/d}} \quad (4)$$

Self-Attention: The attention mechanism computes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \mathbf{V} \quad (5)$$

where d_k is the key dimension. Using multiple heads lets the model pay attention to different kinds of relationships at the same time:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \quad (6)$$

Each encoder layer also includes a position-wise feed-forward network (FFN) with residual connections and layer normalisation.

3) Classification Head: After the Transformer encoder, the output is averaged using global average pooling and then fed through fully connected layers with softmax to predict the identity:

$$\hat{y} = \text{softmax}(\mathbf{W}_c \cdot \mathbf{h} + \mathbf{b}_c) \quad (7)$$

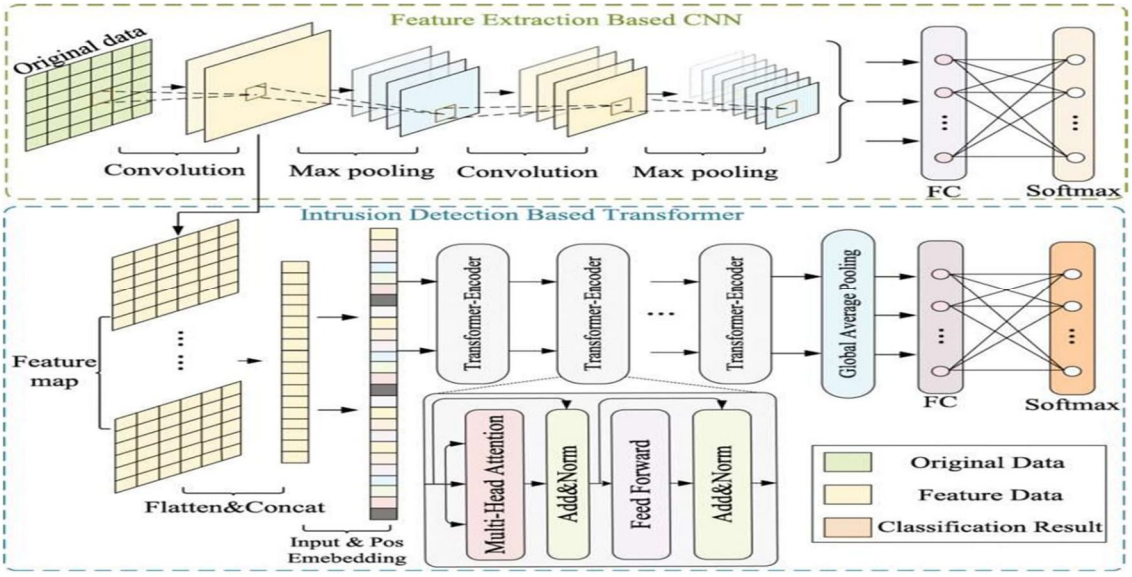


Fig. 3. GaitFormer model architecture.

C. Loss Function and Training

The model is trained using categorical cross-entropy loss:

$$L = -\sum_{i=1}^{N}\sum_{c=1} y_{i,c} \log(\hat{y}_{i,c}) \quad (8)$$

where N is the number of samples, C is the number of classes, and $y_{i,c}$ and $\hat{y}_{i,c}$ are the true label and predicted probability. The Adam optimiser was used with dropout to reduce overfitting.

V. IMPLEMENTATION DETAILS

The system was implemented in Python on Google Colab with GPU acceleration. Table II lists the tools and libraries used.

TABLE II
SOFTWARE TOOLS AND LIBRARIES

Tool / Library	Purpose
Python	Programming language
TensorFlow / Keras	Deep learning framework
OpenCV	Image processing
Matplotlib	Data visualisation
Seaborn	Statistical visualisation
scikit-learn	Evaluation metrics
Google Colab	Execution platform

TensorFlow/Keras was used to build and train the Gait-Former model. OpenCV handled image loading and resizing. Matplotlib and Seaborn were used for plotting training curves and confusion matrices. scikit-learn provided functions for accuracy, precision, recall, F1-score, and the confusion matrix.

The model hyperparameters and training configuration are summarised in Table III.

VI. RESULTS AND DISCUSSION

1) Training Progress

2) Classification Performance

The overall classification performance is summarised in Table IV.

TABLE III
MODEL HYPERPARAMETERS

Parameter	Value
Input Shape	(10, 64, 64, 1)
CNN Filters	32, 64
Attention Heads	4
Key Dimension	64
Dense Units	128
Dropout Rate	0.3
Optimiser	Adam
Loss Function	Sparse Categorical Crossentropy
Batch Size	32
Epochs	30

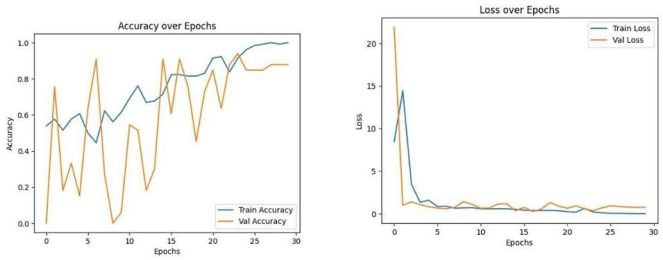


Fig. 4. Training and validation accuracy (left) and loss (right) over epochs.

The test accuracy of 88.89% means the model correctly identified about nine out of every ten test samples. The Rank-1 accuracy matches the test accuracy, meaning the correct identity was the top prediction in 88.89% of cases. The Rank- 5 accuracy of 100.00% means the correct identity always appeared somewhere in the top five predictions, which is useful in practice when a shortlist is reviewed manually.

3) Confusion Matrix

4) Comparison with CNN + BiLSTM Baseline

For contextual comparison, a baseline combining CNN-based spatial extraction with Bidirectional LSTM temporal encoding was also evaluated. Table V presents the results.

TABLE IV
CLASSIFICATION PERFORMANCE ON CASIA-B TEST SET

Metric	Value
Test Accuracy	0.8889 (88.89%)
Rank-1 Accuracy	0.8889 (88.89%)
Rank-5 Accuracy	1.0000 (100.00%)
Precision (Macro)	0.89
Recall (Macro)	0.89
F1-Score (Macro)	0.89

TABLE V
GAITFORMER VS. CNN + BILSTM BASELINE

Metric	GaitFormer	CNN+BiLSTM
Test Accuracy	0.8889	0.7778
Rank-1 Accuracy	0.8889	0.7778
Rank-5 Accuracy	1.0000	1.0000

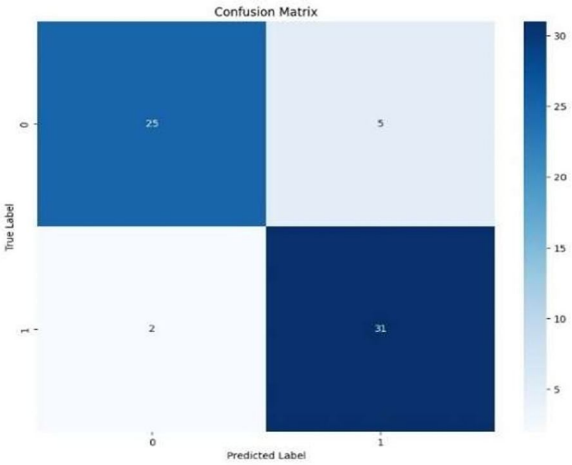


Fig. 5. Confusion matrix for GaitFormer on CASIA-B.

GaitFormer beats the CNN + BiLSTM baseline by about 11 percentage points on both test accuracy and Rank-1 accuracy. This suggests that self-attention handles the temporal side of gait better than recurrent layers in this setup.

5) Discussion

The CNN part of the model picks up spatial features from silhouette frames — body shape, limb positions, and overall appearance. The Transformer attention module then looks across the full sequence to focus on the most informative frames, which helps the model capture patterns that span many time steps.

The Rank-5 accuracy of 100.00% is worth noting: the correct identity always shows up in the top five predictions. This is useful in real-world identification setups where a shortlist is generated and reviewed by an operator.

Limitations: The evaluation was done on a single dataset, so it is not clear how well the model would work on other data or in real-world conditions. The model may not handle extreme variations well, and the Transformer part is computationally expensive, which could be a problem on low-power devices. There are also privacy concerns with gait-based surveillance since people can be identified without their knowledge. Any real deployment would need to follow data protection rules. It should also be noted that some details like exact split ratios and hardware specs are only as reported in the implementation and nothing beyond that is claimed.

VII. CONCLUSION

This paper presented a gait recognition system built using the GaitFormer architecture and tested on the CASIA-B dataset. The model uses CNN layers to get spatial features from silhouette frames and a Transformer encoder to learn temporal patterns in the walking sequence. The data was preprocessed by normalising silhouettes, aligning them, and computing GEIs. The implementation was done in Python with TensorFlow/Keras and run on Google Colab.

On the test set, the model achieved 88.89% accuracy, 88.89% Rank-1 accuracy, and 100.00% Rank-5 accuracy. These numbers show that combining CNNs with Transformer attention works well for gait-based identification on this dataset.

VIII. FUTURE SCOPE

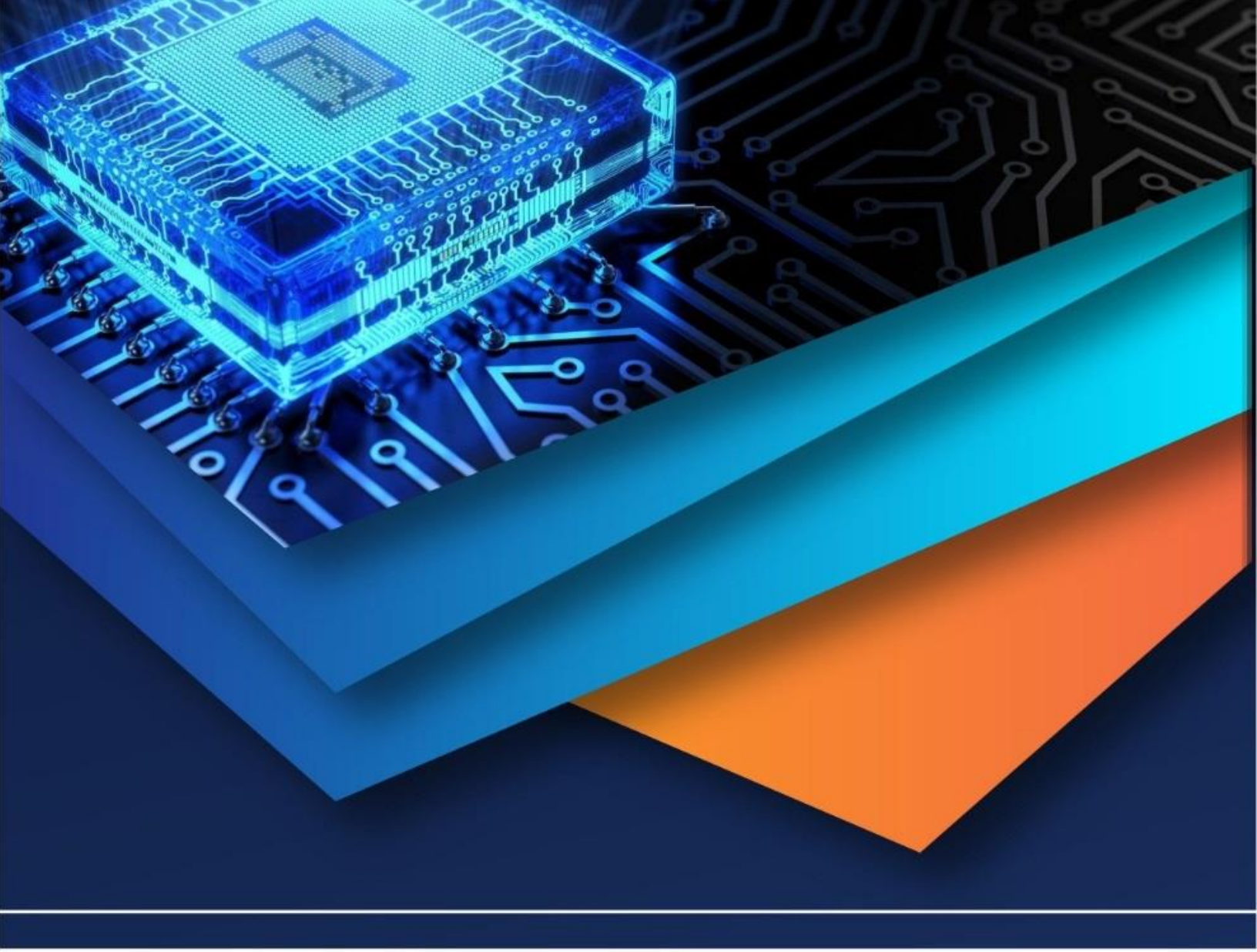
Several directions for future work can be identified:

1) Cross-dataset evaluation on datasets such as OU-MVLP and GREW to assess generalisability.
2) Multi-view fusion to improve recognition in multi-camera setups.
3) Skeleton-based integration combining appearance and model-based features.
4) Lightweight architectures through knowledge distillation or pruning for edge deployment.
5) Real-time optimisation for surveillance applications.
6) Adversarial robustness analysis for security-critical sce-narios.
7) Self-supervised pre-training using unlabelled walking video data.

REFERENCES

[1] M. S. Nixon, T. N. Tan, and R. Chellappa, Human Identification Based on Gait. Springer, 2006.
[2] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," ACM Computing Surveys, vol. 51, no. 5, pp. 1–35, 2018.
[3] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 1,
[4] pp. 4–20, 2004.
[5] J. E. Boyd and J. J. Little, "Biometric gait recognition," in Advanced Studies in Biometrics, Springer, 2005, pp. 19–42.
[6] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," Pattern Recogni-tion, vol. 98, p. 107069, 2020.
[7] J. Han and B. Bhanu, "Individual recognition using gait energy image," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 2, pp. 316–322, 2006, doi: 10.1109/TPAMI.2006.38.
[8] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 2, pp. 209–226, 2017.
[9] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in Proc. ICB, 2016, pp. 1–8.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017,
[12] pp. 5998–6008.
[13] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait recognition using temporal-spatial feature learning for treadmill and overground walking," IEEE Access, vol. 10, pp. 90280–90291, 2022.
[14] Q. Wu, R. Xiao, K. Xu, J. Ni, B. Li, and Z. Xu, "GaitFormer: Revisiting intrinsic periodicity for gait recognition," arXiv:2307.13259, 2023. [Online]. Available: https://arxiv.org/abs/2307.13259
[15] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in Proc. AAAI, vol. 33, 2019, pp. 8126–

8133.
[16] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in Proc. CVPR, 2020, pp. 14225–14233.
[17] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in Proc. ICCV, 2021, pp. 14648–14656.
[18] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in Proc. ICPR, vol. 4, 2006, pp. 441–444, doi: 10.1109/ICPR.2006.67.
[19] Institute of Automation, Chinese Academy of Sciences, "CASIA Gait Database." [Online]. Available: https://english.ia.cas.cn/db/201610/t20161026_169403.html
[20] National Institute of Standards and Technology, "CASIA Gait Database," Biometric and Forensic Research Database Catalog. [Online]. Available: https://tsapps.nist.gov/BDbC/Search/Details/574



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY