HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Multi-view gait recognition with Deep Learning

LÊ HOÀNG LONG
Long.LH232099M@sis.hust.edu.vn

Program: Data Science

Supervisor: Dr. Ngô Thành Trung

Department: Computer Engineering

School: School of Information and Communications Technology

HANOI, 11/2025

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Multi-view gait recognition with Deep Learning

LÊ HOÀNG LONG
Long.LH232099M@sis.hust.edu.vn

Program: Data Science

Supervisor: Dr. Ngô Thành Trung

Department: Computer Engineering

School: School of Information and Communications Technology

HANOI, 11/2025

ACKNOWLEDGMENT

First and foremost, I would like to express my profound gratitude to my supervisor, Dr. Ngô Thành Trung, for their exceptional guidance, insightful feedback, and unwavering support throughout the course of this research. Their expertise in the field of Artificial Intelligence and Deep Learning has been invaluable to the development and completion of this thesis.

I am sincerely thankful to the faculty and staff of the School of Information and Communications Technology, Hanoi University of Science and Technology, for providing a stimulating academic environment and access to essential resources that facilitated my research.

Finally, I am deeply indebted to my family for their enduring encouragement, patience, and belief in my academic pursuits. Their support has been a constant source of strength throughout this journey.

ABSTRACT

The main challenge addressed in the research of this thesis is the need for recognising multiple, and uncooperative subjects in the crowd. Existing single-angle approaches often yield less than 96% accuracy, and some research asks for high training costs. Our proposed solution, "Multi-View Gait Recognition with Deep Learning" - MVDL) , utilizes multiple camera angles during model training while strictly holding the flexible requirement for both single-angle inference and multiple-angles inference. It helps the model enrich data without compromising real-world deployment. This is achieved using a novel concept, inspired by combination drug therapies, that applies a "virus-transmission-like mechanism" to improve the accuracy at each observation angle. The implementation follows three main steps: Step 1: Multi-Angle Training and Knowledge Propagation, where we applied a new training method inspired by the Graph Neural Network (GNN) message propagation mechanism to bidirectionally enrich the embedding vector of the center camera angle , and results from all angles are then combined into a multi-perspective Transformer model; Step 2: Single-Angle Inference and Embedding Enrichment, where the initial embedding from the single-angle model is fed into the Transformer to obtain an enriched vector ; and finally, Step 3: Identity Recognition, where the enriched vector is compared with stored embeddings using cosine similarity. Our main contributions are the usage of a new training method which is inspired by the GNN message propagation mechanism for embedding enrichment and the usage of the Transformer for synthesizing this enriched information. This approach has successfully pushed the model accuracy to 96.75% at a camera angle, and 98.48% at a pair of camera angles, surpassing the achievements of published research while requiring fewer computational resources.

Student

(Signature and full name)

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1

1.1 Problem Statement 1

1.2 Background and Problems of Research 1

1.2.1 Overview 1

1.2.2 The Fundamentals of Human Gait 1

1.2.3 Overview of Gait Recognition 2

1.2.4 Single-View in Gait Recognition 3

1.2.5 The Fundamentals of Gait Energy Image (GEI) 4

1.2.6 Multiple-View in Gait Recognition 5

1.2.7 Classic / Pre–Deep Learning Gait Recognition Approaches 5

1.2.8 Recent Gait Recognition Research. 11

1.2.9 Features usage in Gait Recognition 13

1.3 Research Objectives and Conceptual Framework 13

1.4 Contributions 15

1.5 Organization of Thesis 16

CHAPTER 2. LITERATURE REVIEW 17

2.1 Scope of Research 17

2.1.1 Our new approach 17

2.1.2 Our proposed model architecture 17

2.2 Related Works 17

2.2.1 Highlighted Researches on OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) 17

2.2.2 Key Papers on 1-View (Single-View) Gait Recognition 18

2.2.3 Key Papers on Multi-View Gait Recognition 19

CHAPTER 3. METHODOLOGY 26

3.1 Overview 26

3.1.1 Our proposed model architecture 26

3.2 Our model explanation in math 27

3.2.1 Convolutional Neural Network (CNN) 27

3.2.2 Transfomer Encoder 28
3.3 Convolutional neural network model - Descriptor model 29
3.4 Transformer model - Reasoner model 31
CHAPTER 4. NUMERICAL RESULTS 33
4.1 Dataset 33
4.1.1 Data Source 33
4.2 Evaluation Parameters 33
4.3 Experiment Setting 34
4.4 Reasoning on two views (Multi-View Synthesis) 35
4.5 Reasoning on one view (Single-View Inference) 35
4.5.1 Other measurements 37
CHAPTER 5. CONCLUSIONS 38
5.1 Summary 38
5.2 Suggestion for Future Works 38
REFERENCE 42

LIST OF FIGURES

Figure 1.1 Overview of the gait cycle and sub-phases analyzed 1
Figure 1.2 Terminology to describe the events of the gait cycle 2
Figure 1.3 Gait Energy Image [12] 4
Figure 1.4 Silhouette representation 6
Figure 1.5 Overview of automatic person identification 7
Figure 1.6 Stances corresponding to the gait cycle of two individuals. (a) Person 1. (b) Person 2. 8
Figure 1.7 An example of generating a CGI temporal template 9
Figure 1.8 Proposed CNN models for gait signature extraction. 11
Figure 1.9 Multi-view camera setting 14

Figure 2.1 Spatial-temporal deep neural network 21
Figure 2.2 Overall structure of the network model. 22
Figure 2.3 Three batch-sampling methods comparison 22
Figure 2.4 View Information Elimination Mechanism (VIEM) framework 24

Figure 3.1 Our proposed model architecture 26
Figure 3.2 Example of how information can be exchanged and enriched between one camera angle and related angles 26
Figure 3.3 Proposed solution's workflow 27
Figure 3.4 Backbone Architecture 30
Figure 3.5 Proposed solution's workflow 31

Figure 4.1 The subject repeat forward (A to B) and backward (B to A) 33
Figure 4.2 Examples of silhouette sequence (view angle = 90°) 33
Figure 4.3 Examples of size-normalized GEI for each view angles 34
Figure 4.4 Rank accuracy of all view points 37

LIST OF TABLES

Table 2.1 Cross-View Gait Recognition by Discriminative Feature Learning 18
Table 2.2 Combining the Silhouette and Skeleton Data for Gait Recognition 18
Table 2.3 Gait Recognition Using 3-D Human Body Shape Inference 19
Table 2.4 Learning Visual Prompt for Gait Recognition 19
Table 2.5 Learning rich features for gait recognition by integrating skeletons and silhouettes . . 20
Table 2.6 The Accuracies of Different Methods Evaluated on OU-ISIR 20

Table 4.1 Baseline table 34
Table 4.2 The performance comparisons on OUMVLP with two-view accuracy 35
Table 4.3 The performance comparisons on OUMVLP, excluding the identical-views cases. . . 35
Table 4.4 Performance of the MVDL solution 35
Table 4.5 The Rank accuracy per view. 36
Table 4.7 Descriptor FLOP. 37
Table 4.8 Reasoner FLOP with vector sequence length of 128. 37

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

Recognizing the identities of multiple subjects within a crowd, especially without subject cooperation, is an urgent requirement in today's complex security environment. Many existing approaches aim to solve this problem globally. However, most of these methods focus on simply improving the accuracy achieved from a single camera angle. Furthermore, published methods typically require large training costs and currently achieve a Rank-1 accuracy lower than 96%.

To address these limitations, we propose a novel approach named "Multi-view gait recognition with Deep Learning - MVDL".

1.2 Background and Problems of Research

1.2.1 Overview

Gait recognition, the identification of a person based on their walking style, has seen several notable contributions in recent years. The Rank-1 accuracy rates for some contemporary models are: Cross-View Gait Recognition by Discriminative Feature Learning: 95.4% [1]. Combining the Silhouette and Skeleton Data for Gait Recognition: 92.5% [2]. Gait Recognition Using 3-D Human Body Shape Inference: 91.4% [3]. Learning Visual Prompt for Gait Recognition: 93.2% [4]. Learning rich features for gait recognition by integrating skeletons and silhouettes: 91.28% [5]. GaitGCI: Generative Counterfactual Intervention for Gait Recognition: 93.0% [6]. These solutions face two primary constraints: accuracy levels fall short of the 96% benchmark, and inference is limited to data captured from a single camera perspective.

1.2.2 The Fundamentals of Human Gait

Human gait refers to the specific way a person walks, including the cyclical pattern of movement of the limbs and body during locomotion on foot.[7] It's one of the most fundamental and complex human motor activities, involving coordinated interaction between the nervous system, muscles, bones, and joints.

There are 4 key features of human gait. Bipedal: Humans are one of the few mammals that walk upright on two legs consistently. [8] Double pendulum motion: In simple terms, each leg acts like an inverted pendulum during the stance phase and a regular pendulum during the swing phase. Reciprocal arm swing: Arms swing opposite to the legs for balance and efficiency. Heel-to-toe rolling: In normal walking, the heel strikes the ground first (heel strike), followed by weight transfer through the foot to push off with the toes (toe-off).

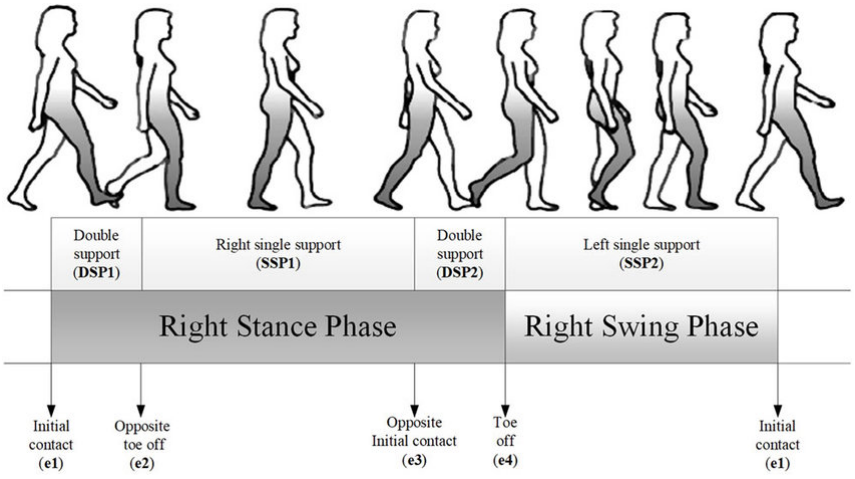

Figure 1.1: Overview of the gait cycle and sub-phases analyzed

The gait cycle is typically described for one leg and is divided into two main phases: Stance phase (60% of the cycle)[9] – when the foot is on the ground. It includes: initial contact (heel strike), loading response (foot flattens), mid-stance (body directly over the foot), terminal stance (heel rises), and pre-swing (toe-off preparation) Figure 1.1. Swing phase (40% of the cycle) – when the foot is in the air. It includes: initial swing (toe-off and leg moves forward), mid-swing (leg passes under the body), and terminal swing (preparing for heel strike)

There is also double support (both feet on the ground) during walking, which disappears when a person starts running (then it becomes double float) Figure 1.2. This phase occurs twice during each gait cycle, taking up about 20% of the total cycle, and is a critical period for weight acceptance and stability. [10]

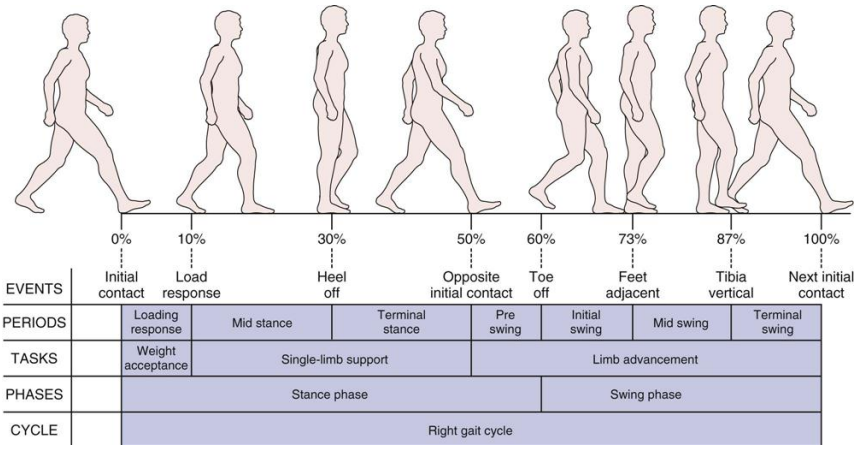

Figure 1.2: Terminology to describe the events of the gait cycle

Details of the double support phase includes: occurs twice per cycle: The double support phase happens twice, beginning with the heel strike of one foot and continuing until the other foot lifts off the ground, accounts for 20% of the cycle: the entire gait cycle is approximately 20% double support and 80% single support, where only one foot is on the ground consisting of two sub-phases: initial double support: begins with the heel strike of the leading foot and ends when the opposite foot lifts off the ground, [10] and terminal double support: starts when the opposite foot makes initial contact with the ground and ends when the leading foot lifts off the ground, provides stability: it is a key time for stability, as both feet are on the ground to support the body's weight. and can be affected by other factors: things like walking while being in certain conditions, such as Parkinson's disease, are often associated with a longer double support phase.

1.2.3 Overview of Gait Recognition

Gait Recognition is a biometric technology that identifies individuals based on their unique way of walking—their gait pattern. It is often called "walking fingerprint" technology because, like a fingerprint, each person's gait is considered distinct.

Unlike other biometrics that require close proximity or cooperation (like fingerprint or iris scans), gait analysis can be performed at a distance and without the subject's knowledge, making it particularly valuable for surveillance and security applications.[11]

Gait recognition typically involves three main stages powered by modern AI and computer vision techniques. The first stage, acquisition and detection, uses a video camera to capture a sequence of a person walking, then the system detects and isolates the moving human figure from the background. The second stage, feature extraction and modeling, can apply two primary approaches. In model-based methods, the system builds a representation of the human body, including skeleton, joints, and limbs, and tracks their movement over time to extract dynamic features such as step length, cadence, limb swing speed, torso angle, and joint angles. This approach is more robust to changes in clothing or carrying items. In contrast, appearance-based (model-free) methods treat the entire walking figure as a single shape without modeling

body parts. These often rely on silhouettes and generate a Gait Energy Image (GEI), which is a composite image created by averaging silhouettes across a walking cycle, it captures the unique blur pattern of the walk. This approach is simpler to compute and effective with high-quality video.

Gait recognition offers several advantages: it is non-intrusive and can be performed at a distance without requiring cooperation or awareness; it is difficult to conceal or impersonate because gait is a subconscious motor pattern; it remains effective even in low resolution video where facial recognition may fail; and it complements other biometrics, enhancing identification accuracy when combined. Actually, facial task will fail to recognize the new face after the subject gpt a facial surgery. However, challenges and limitations still exist. Performance can be sensitive in some cases, such as clothing, footwear, carrying conditions, or emotional state. Recognition accuracy also depends on viewing angle, as gait appearance changes with camera perspective. Environmental factors like rough terrain or stairs can alter natural gait, and the complex of processing video sequences, especially in real time, demands significant computing power. In essence, gait recognition transforms the unconscious act of walking into a unique and identifiable biometric signature.

1.2.4 Single-View in Gait Recognition

In gait recognition, the concept of single-view gait recognition refers to the process of identifying or verifying an individual using gait information which is captured from only one camera angle or fixed viewpoint. Unlike multi-view systems that integrate data from several perspectives, single-view recognition relies entirely on the information available from a single, static camera. This makes it a practical approach for many real-world applications, particularly in surveillance environments where most CCTV networks consist of independent cameras installed at fixed positions. The system learns to recognize individuals by analyzing the unique walking patterns visible from that one viewpoint, which can be sufficient for effective recognition under controlled conditions.

A defining characteristic of single-view gait recognition has its reliance on a single camera. The system processes video footage or sequences of images captured from one perspective, which is often chosen to maximize the visibility of gait features. In research contexts, cameras are frequently positioned to capture a 90-degree side profile, also known as a lateral view, the main reason is that this angle provides the richest information about limb motion and overall body dynamics. However, this reliance on a fixed angle also introduces significant challenges. A system trained on side-view data may struggle when the subject walks directly toward or away from the camera, as the silhouettes and dynamic features appear drastically different from those captured in lateral views. This dependence on viewpoint highlights one of the main limitations of single-view recognition systems.

Beyond viewpoint, single-view gait recognition is also sensitive to co-variates which are external factors that can alter the appearance of a person's gait. Clothing, such as a long coat, can change the silhouette and obscure limb movement. Carrying conditions, like holding a backpack or briefcase, can affect posture and walking dynamics. Footwear choices, whether high heels or sneakers, influence stride and cadence, while emotional states, such as walking happily, sadly, or while injured, can subtly alter gait patterns. These variables make recognition more complex and can reduce accuracy if not properly accounted for in system design.

In essence, single-view gait recognition requires sufficient data to extract meaningful features. Typically, the system processes a full gait sequence, meaning a video of the subject walking through an entire gait cycle. From this sequence, features such as the Gait Energy Image (GEI) are generated. The GEI is a composite representation created by averaging silhouettes across the walking cycle, capturing the distinctive blur pattern of an individual's gait. This feature serves as a powerful tool for distinguishing between individuals, even when only a single viewpoint is available. Thus, while single-view gait recognition faces challenges related to viewpoint dependence and sensitivity to co-variates, it remains a practical and widely studied approach in biometric identification.

1.2.5 The Fundamentals of Gait Energy Image (GEI)

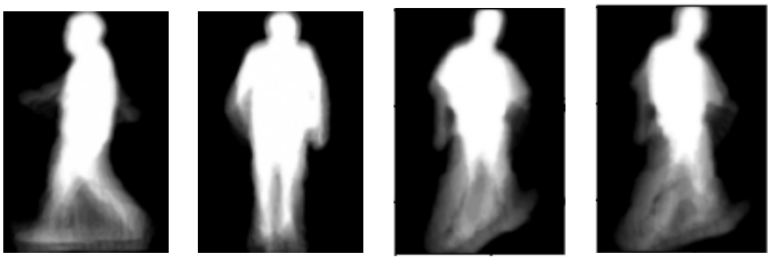

Figure 1.3: Gait Energy Image [12]

When a video of a person walking is taken, their silhouette can be extracted from each frame and then layered together into a single, averaged image. The result is a blurred, ghost-like picture that captures the essence of their movement, and this representation is known as a Gait Energy Image (GEI). The GEI serves as a type of "motion history" which is condensed into one static picture. It was introduced as a powerful appearance-based representation for gait recognition, designed to simplify the complex task of analyzing dozens or even hundreds of frames in a video sequence into the more manageable task of analyzing a single image. By compressing temporal information into one composite form, the GEI provides a unique way of capturing both the static and dynamic aspects of human walking.

The creation of a GEI Figure 1.3 involves a systematic process. First, silhouette extraction is performed, where the walking person is detected in each frame of the video and segmented from the background. This produces a sequence of binary silhouettes, typically represented as white figures against a black background. Next, temporal averaging is applied. Over the course of one complete gait cycle—from one heel strike to the next heel strike of the same foot—all the silhouette frames are averaged together. The pixel values in the final GEI are calculated based on how frequently each pixel was active, meaning part of the silhouette, during the cycle. Mathematically, this is expressed as $G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y)$ where $G(x, y)$ is the pixel value at location (x,y) in the GEI. N is the total number of frames in one gait cycle. $B_t(x, y)$ is the pixel value (1 for foreground, 0 for background) in the silhouette image at time t. $\sum$ means authors sum up all the silhouette values for that pixel over time. $\frac{1}{N}$ averages that sum. This averaging process transforms temporal motion into a spatially encoded image that reflects both presence and movement.

A typical GEI reveals three distinct regions that correspond to different aspects of the body's motion. The static body parts, such as the head and body, appear as the brightest and most solid regions because they are consistently present in the silhouettes throughout the gait cycle. The dynamic limb parts, including the arms and legs, appear as grey, blurred regions since they move rapidly and occupy varying positions across frames. This blurred energy pattern captures the unique swing and stride of the limbs. Finally, the background remains the darkest region, as those pixels are never activated during the walking sequence. Together, these regions create a composite image that encodes both the physical build of the person and the dynamic characteristics of their gait.

The GEI offers several important advantages. It reduces data complexity by compressing a long video sequence into a single image, which significantly lowers computational costs and memory requirements. It is also robust to noise, as averaging across an entire gait cycle smooths out random segmentation errors that may occur in individual frames. Furthermore, the GEI encodes both shape and dynamics: the bright static regions capture the person's body structure, while the blurred dynamic regions capture temporal movement features such as stride length and leg swing. Once a GEI is generated, it can be easily compared against a database using standard image classification techniques, including feature descriptors and convolutional

neural networks (CNNs), making it a practical tool for biometric identification.

Despite its strengths, the GEI has limitations. One major drawback is the loss of temporal order. Because the GEI is an average, it does not preserve the sequence of movements, meaning it cannot distinguish between left and right steps or capture the precise timing of limb motions. Another limitation is sensitivity to viewing angle. A GEI generated from a side view looks completely different from one generated from a frontal view, making recognition highly dependent on camera placement. Additionally, the GEI is sensitive to appearance co-variates such as clothing or carrying conditions. For example, a long coat or a backpack can dramatically alter the silhouette, obscuring the underlying gait pattern and reducing recognition accuracy. These challenges highlight the need for complementary methods or multi-view approaches to overcome the constraints of GEI-based gait recognition.

1.2.6 Multiple-View in Gait Recognition

Multiple-view gait recognition is an advanced method that employs multiple cameras positioned at different angles to capture a person's walk. Instead of relying on a single viewpoint, the system integrates information from several perspectives to build a more comprehensive and robust representation of the gait. This integration can be achieved either by fusing the data collected from each camera view or by reconstructing a 3 dimensional model of the person's movement. By combining these different viewpoints, the system is less likely to be misled by changes in walking direction or by objects that obscure one particular view. The central idea behind multiple-view gait recognition is to overcome the inherent limitations of single-camera systems by leveraging synchronized information from multiple sources, thereby creating a more reliable and accurate biometric signature.

There are two primary strategies used in multiple-view gait recognition. The first is view fusion, where gait features such as Gait Energy Images (GEIs) or skeletal models are extracted independently from each camera view. These features are combined at different stages of processing early fusion of raw data, mid level fusion of extracted features, or late fusion of matching scores—to produce a single, view-invariant gait signature. This fusion process ensures that the final representation captures the essential characteristics of the gait while minimizing the distortions caused by viewpoint changes. The second strategy is three-dimensional model reconstruction. In this approach, computer vision techniques and photogrammetry are applied to synchronized video streams from multiple cameras to reconstruct a precise 3D model of the person walking. Once the 3D model is created, it can be analyzed from any virtual viewpoint, effectively eliminating the problem of viewpoint dependence. Features can then be extracted directly from this volumetric representation, providing a richer and more flexible dataset for recognition.

Although multiple-view gait recognition offers a powerful and robust solution, particularly in high-stakes or controlled environments such as security-sensitive facilities, it comes with significant trade-offs. The complexity of setting up and synchronizing multiple cameras, along with the computational demands of fusing data or reconstructing 3D models, makes the system more expensive and resource-intensive compared to single-view approaches. Nevertheless, the enhanced accuracy and resilience to covariates make multiple-view gait recognition a valuable tool in scenarios where reliability and precision are paramount.

1.2.7 Classic / Pre–Deep Learning Gait Recognition Approaches

a, Gait Energy Image (GEI) and Silhouette-Based Methods

The most influential non-DL approaches.

Han & Bhanu (2006): "Individual recognition using gait energy image."
In the paper, authors propose a new spatio-temporal gait representation, called Gait Energy Image (GEI), to characterize human walking properties for individual recognition by gait. [13] To address the problem of the lack of training templates, authors also propose a novel approach for human recognition by combining statistical gait features from real and synthetic templates. Authors directly compute the real templates from training silhouette sequences, while authors generate the synthetic templates from training sequences by

simulating silhouette distortion. Authors use a statistical approach for learning effective features from real and synthetic templates. Authors compare the proposed GEI-based gait recognition approach with other gait recognition approaches on USF HumanID Database. Experimental results show that the proposed GEI is an effective and efficient gait representation for individual recognition, and the proposed approach achieves highly competitive performance with respect to the published gait recognition approaches. [13]

Wang, Tan, Ning, & Hu (2003): “Silhouette analysis-based gait recognition for human identification.” Human identification at a distance has recently gained growing interest from computer vision researchers. Gait recognition aims essentially to address this problem by identifying people based on the way they walk. [14] In the paper, a simple but efficient gait recognition algorithm using spatial-temporal silhouette analysis is proposed. For each image sequence, a background subtraction algorithm and a simple correspondence procedure are first used to segment and track the moving silhouettes of a walking figure. Then, eigenspace transformation based on principal component analysis (PCA) is applied to time-varying distance signals derived from a sequence of silhouette images to reduce the dimensionality of the input feature space. Supervised pattern classification techniques are finally performed in the lower-dimensional eigenspace for recognition. This method implicitly captures the structural and transitional characteristics of gait. Extensive experimental results on outdoor image sequences demonstrate that the proposed algorithm has an encouraging recognition performance with relatively low computational cost. [14]

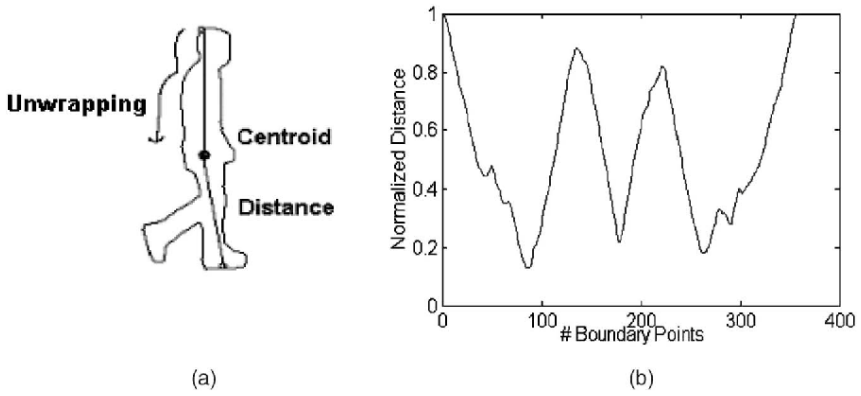

Figure 1.4: Silhouette representation

Silhouette representation, Figure 1.4, illustration of boundary extraction and counterclockwise unwrapping and the normalized distance signal consisting of all distances between the centroid and the pixels on the boundary

BenAbdelkader, Cutler, & Davis (2002): "Stride and cadence as a biometric in automatic person identification." [15]. Authors present a correspondence-free method to automatically estimate the spatio-temporal parameters of gait (stride length and cadence) of a walking person from video. Stride and cadence, Figure 1.5, are functions of body height, weight, and gender, and authors use these biometrics for identification and verification of people. The cadence is estimated using the periodicity of a walking person. Using a calibrated camera system, the stride length is estimated by first tracking the person and estimating their distance travelled over a period of time. By counting the number of steps (again using periodicity), and assuming constant-velocity walking, authors are able to estimate the stride to within 1cm for a typical outdoor surveillance configuration (under certain assumptions). With a database of 17 people and 8 samples of each, authors show that a person is verified with an Equal Error Rate (EER) of 11%, and correctly identified with a probability of 40%. This method works with low-resolution images of people, and is robust to changes in lighting, clothing, and tracking errors. It is view-invariant though performance is optimal in a near fronto-parallel configuration. [15]

Liu & Sarkar (2004): "Simplest representation yet for gait recognition: Averaged that silhouette." [16] . Authors present a robust representation for gait recognition that is compact, easy to construct, and affords

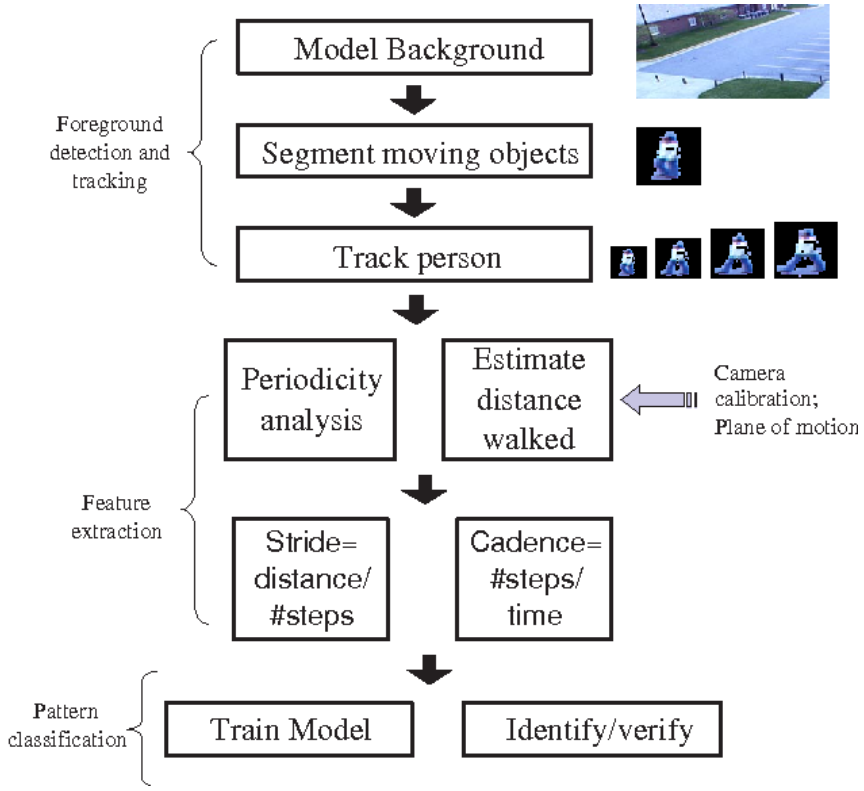

Figure 1.5: Overview of automatic person identification

efficient matching. Instead of a time series based representation comprising of a sequence of raw silhouette frames or of features extracted therein, as has been the practice, authors simply align and average the silhouettes over one gait cycle. Authors then base recognition on the Euclidean distance between these averaged silhouette representations. Authors show, using the recently formulated gait challenge problem, that the improvement in execution time is 30 times while possessing recognition power that is comparable to the gait baseline algorithm, which is becoming the comparison standard in gait recognition. Experiments with portions of the average silhouette representation show that recognition power is not entirely derived from upper body shape, rather the dynamics of the legs also contribute equally to recognition. However, this study does raise intriguing doubts about the need for accurate shape and dynamics representations for gait recognition. [16]

b, Model-Based Approaches (No Deep Learning)

Kinematic or structural human models.

Cunado, Nixon & Carter (2002): "Automatic extraction and description of human gait models for recognition purposes." [17]. Using gait as a biometric is of emerging interest. Authors describe a new model-based moving feature extraction analysis is presented that automatically extracts and describes human gait for recognition. The gait signature is extracted directly from the evidence gathering process. This is possible by using a Fourier series to describe the motion of the upper leg and apply temporal evidence gathering techniques to extract the moving model from a sequence of images. Simulation results highlight potential performance benefits in the presence of noise. Classification uses the k-nearest neighbour rule applied to the Fourier components of the motion of the upper leg. Experimental analysis demonstrates that an improved classification rate is given by the phase-weighted Fourier magnitude information over the use of the magnitude information alone. The improved classification capability of the phase-weighted magnitude information is verified using statistical analysis of the separation of clusters in the feature space. Furthermore, the technique is shown to be able to handle high levels of occlusion, which is of especial importance in gait as the human body is self-occluding. As such, a new technique has been developed to

automatically extract and describe a moving articulated shape, the human leg, and shown its potential in gait as a biometric. [17]

Yam, Nixon, Carter (2004): "Automated person recognition by walking and running via model-based approaches." [18] . Gait enjoys advantages over other biometrics in that it can be perceived from a distance and is difficult to disguise. Current approaches are mostly statistical and concentrate on walking only. By analysing leg motion authors show how authors can recognise people not only by the walking gait, but also by the running gait. This is achieved by either of two new modelling approaches which employ coupled oscillators and the biomechanics of human locomotion as the underlying concepts. These models give a plausible method for data reduction by providing estimates of the inclination of the thigh and of the leg, from the image data. Both approaches derive a phase-weighted Fourier description gait signature by automated non-invasive means. One approach is completely automated whereas the other requires specification of a single parameter to distinguish between walking and running. Results show that both gaits are potential biometrics, with running being more potent. By its basis in evidence gathering, this new technique can tolerate noise and low resolution.[18]

A. Kale, A. Roy-Chowdhury, et al. (2004) : "Identification of humans using gait" [19]. In the paper authors propose a view based approach to recognize humans using gait. The width of the outer contour of the binarized silhouette of a walking person is chosen as the image feature. A set of exemplars that occur during a walk cycle is chosen for each individual. Using these examples a lower dimensional Frame to Exemplar Distance (FED) vector is generated, Figure 1.6. A continuous HMM is trained using several such FED vector sequences. This methodology serves to compactly capture structural and dynamic features that are unique to an individual. The statistical nature of the HMM renders overall robustness to representation and recognition. Human identification performance of the proposed scheme is illustrated using outdoor video sequences. [19]

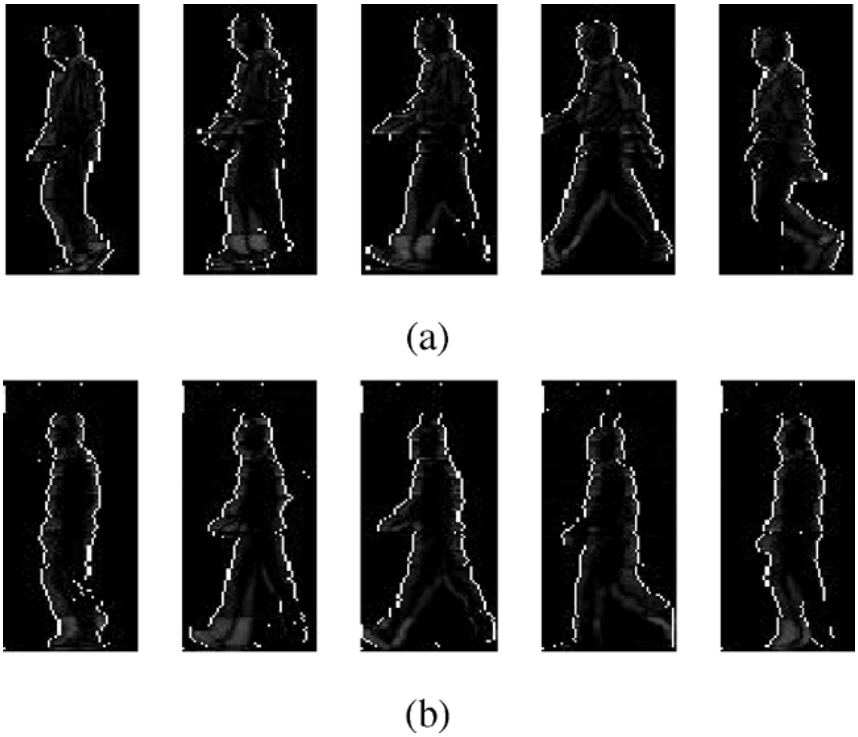

Figure 1.6: Stances corresponding to the gait cycle of two individuals. (a) Person 1. (b) Person 2.

c, Gait Dynamics, Frequency, and Statistical Methods

Feature engineering and classical ML. Makihara & Yagi (2010): Silhouette transformation based on walking speed for gait identification. [20] . Authors propose a method of gait silhouette transformation

from one speed to another to cope with walking speed changes in gait identification. When a person changes his/her walking speed, dynamic features (e.g. stride and joint angle) are changed while static features (e.g. thigh and shin lengths) are unchanged. Based on the fact, firstly, static and dynamic features are separated from gait silhouettes by fitting a human model. Secondly, a factorization-based speed transformation model for the dynamic features is created using a training set for multiple persons on multiple speeds. This model can transform the dynamic features from a reference speed to another arbitrary speed. Finally, silhouettes are restored by combining the unchanged static features and the transformed dynamic features. Evaluation by gait identification using silhouette-based frequency-domain features shows the effectiveness of the proposed method. [20]

K. Bashir, T. Xiang and S. Gong (2009): "Gait recognition using Gait Entropy Image," [21]. Gait as a behavioural biometric is concerned with how people walk. However, most existing gait representations capture both motion and appearance information. They are thus sensitive to changes in various covariate conditions such as carrying and clothing. In this paper, a novel gait representation termed as Gait Entropy Image (GEnI) is proposed. Based on computing entropy, a GEnI encodes in a single image the randomness of pixel values in the silhouette images over a complete gait cycle. It thus captures mostly motion information and is robust to covariate condition changes that affect appearance. Extensive experiments on the USF HumanID dataset, CASIA dataset and the SOTON dataset have been carried out to demonstrate that the proposed gait representation outperforms existing methods, especially when there are significant appearance changes. Our experiments also show clear advantage of GEnI over the alternatives without the assumption on cooperative subjects, i.e. both the gallery and the probe sets consist of a mixture of gait sequences under different and unknown covariate conditions. [21]

d, Handcrafted Spatiotemporal Templates

GEI-inspired variants without Deep Learning. They focused entirely on hand-designed image transforms using SVM or k-NN. Hofmann Martin , Bachmann Sebastian & Rigoll Gerhard (2012): DGHEI (Depth Gradient Histogram Energy Image) [22] Using gait recognition methods, people can be identified by the way they walk. The most successful and efficient of these methods are based on the Gait Energy Image (GEI). In this paper, authors extend the traditional Gait Energy Image by including depth information. First, GEI is extended by calculating the required silhouettes using depth data. Authors then formulate a completely new feature, which authors call the Depth Gradient Histogram Energy Image (DGHEI). Authors compare the improved depth-GEI and the new DGHEI to the traditional GEI. Authors do this using a new gait database which was recorded with the Kinect sensor. On this database authors show significant performance gain of DGHEI. [22]

Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan & Liang Wang (2010): CGEI (Chrono Gait Image) [23]. In the paper, authors propose a novel temporal template, called Chrono-Gait Image (CGI), Figure 1.7,

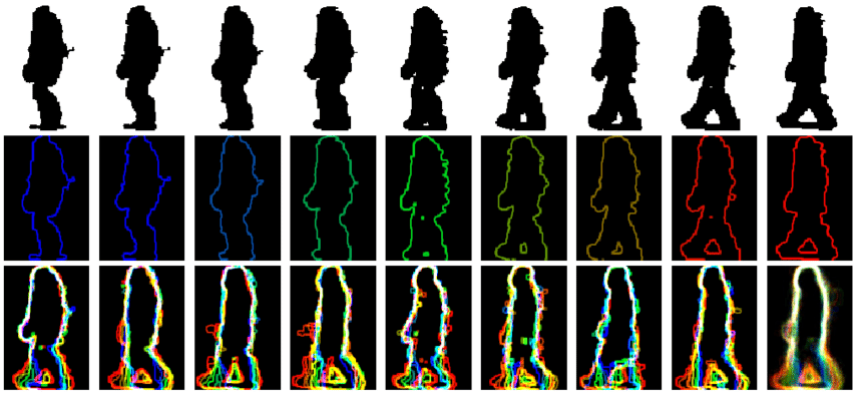

Figure 1.7: An example of generating a CGI temporal template

to describe the spatio-temporal walking pat- tern for human identification by gait. The CGI temporal template en- codes the temporal information among gait frames via color mapping to improve the recognition performance. Our method starts with the ex- traction of the contour in each gait image, followed by utilizing a color mapping function to encode each of gait contour images in the same gait sequence and compositing them to a single CGI. Authors also obtain the CGI-based real templates by generating CGI for each period of one gait se- quence and utilize contour distortion to generate the CGI-based synthetic templates. In addition to independent recognition using either of individ- ual templates, authors combine the real and synthetic temporal templates for refining the performance of human recognition. Extensive experiments on the USF HumanID database indicate that compared with the recently published gait recognition approaches, our CGI-based approach attains better performance in gait recognition with considerable robustness to gait period detection. [23]

e, Other methods

Worapan Kusakunniran (2014): Recognizing gaits on spatio-temporal feature domain. [24]. Gait has been known as an effective biometric feature to identify a person at a distance, e.g., in video surveillance applications. Many methods have been proposed for gait recognitions from various different perspectives. It is found that these methods rely on appearance (e.g., shape contour, silhouette)-based analyses, which require preprocessing of foreground–background segmentation (FG/BG). This process not only causes additional time complexity, but also adversely influences performances of gait analyses due to imperfections of existing FG/BG methods. Besides, appearance-based gait recognitions are sensitive to several variations and partial occlusions, e.g., caused by carrying a bag and varying a cloth type. To avoid these limitations, this paper proposes a new framework to construct a new gait feature directly from a raw video. The proposed gait feature extraction process is performed in the spatio-temporal domain. The space-time interest points (STIPs) are detected by considering large variations along both spatial and temporal directions in local spatio-temporal volumes of a raw gait video sequence. Thus, STIPs are allocated, where there are significant movements of human body in both space and time. A histogram of oriented gradients and a histogram of optical flow are computed on a 3D video patch in a neighborhood of each detected STIP, as a STIP descriptor. Then, the bag-of-words model is applied on each set of STIP descriptors to construct a gait feature for representing and recognizing an individual gait. When compared with other existing methods in the literature, it has been shown that the performance of the proposed method is promising for the case of normal walking, and is outstanding for the case of partial occlusion caused by walking with carrying a bag and walking with varying a cloth type. [24]

Kusakunniran et al. (2014): Recognizing gaits across views through correlated motion co-clustering.[25] Human gait is an important biometric feature, which can be used to identify a person remotely. However, view change can cause significant difficulties for gait recognition because it will alter available visual features for matching substantially. Moreover, it is observed that different parts of gait will be affected differently by view change. By exploring relations between two gaits from two different views, it is also observed that a part of gait in one view is more related to a typical part than any other parts of gait in another view. A new method proposed in this paper considers such variance of correlations between gaits across views that is not explicitly analyzed in the other existing methods. In our method, a novel motion co-clustering is carried out to partition the most related parts of gaits from different views into the same group. In this way, relationships between gaits from different views will be more precisely described based on multiple groups of the motion co-clustering instead of a single correlation descriptor. Inside each group, a linear correlation between gait information across views is further maximized through canonical correlation analysis (CCA). Consequently, gait information in one view can be projected onto another view through a linear approximation under the trained CCA subspaces. In the end, a similarity between gaits originally recorded from different views can be measured under the approximately same view. Comprehensive experiments based on widely adopted gait databases have shown that our method outperforms the state-of-the-art. [25]

Himanshu Aggarwal and Dinesh K. Vishwakarma (2016): Covariate conscious approach for gait recognition based upon Zernike moment invariants. [26]. Gait recognition i.e. identification of an individual from his/her walking pattern is an emerging field. While existing gait recognition techniques perform satisfactorily in normal walking conditions, there performance tend to suffer drastically with variations in clothing and carrying conditions. In this work, authors propose a novel covariate cognizant framework to deal with the presence of such covariates. Authors describe gait motion by forming a single 2D spatio-temporal template from video sequence, called Average Energy Silhouette image (AESI). Zernike moment invariants (ZMIs) are then computed to screen the parts of AESI infected with covariates. Following this, features are extracted from Spatial Distribution of Oriented Gradients (SDOGs) and novel Mean of Directional Pixels (MDPs) methods. The obtained features are fused together to form the final well-endowed feature set. Experimental evaluation of the proposed framework on three publicly available datasets i.e. CASIA dataset B, OU-ISIR Treadmill dataset B and USF Human-ID challenge dataset with recently published gait recognition approaches, prove its superior performance. [26]

Xu, Makihara, et al (2019): Speed-Invariant Gait Recognition Using Single-Support Gait Energy Image (SSGEI) [27]. Gait is one of the most popular behavioral biometrics because it can be authenticated at a distance from a camera without subject cooperation. Speed differences between matching pairs, however, cause significant performance drops in gait recognition, and gait mode difference (i.e., walking versus running) makes gait recognition further challenging. Authors therefore propose a speed-invariant gait representation called single-support GEI (SSGEI), which realizes a good trade-off between speed invariance and stability by aggregating multiple frames around single-support phases. In addition, to mitigate the pose differences between walking and running modes at single-support phases, authors morph walking and running SSGEIs into intermediate SSGEIs between walking and running mode, where authors exploit a free-form deformation field from the walking or running modes to the intermediate mode obtained by training data. Authors finally apply Gabor filtering and spatial metric learning as postprocessing for further accuracy improvement. Experiments on two publicly available datasets, the OU-ISIR Treadmill Dataset A and the CASIA-C Dataset demonstrate that the proposed method yields the state-of-the-art accuracies in both identification and verification scenarios with a low computational cost. [27]

1.2.8 Recent Gait Recognition Research

Foundational works (2018–2020) emphasize CNN foundations and sensor fusion; 2021–2023 focus on hybrids and reviews; 2024–2025 incorporate transformers and incremental learning for robustness.

Evaluation of CNN Architectures for Gait Recognition Based on Optical Flow Maps (2018) [28]. a)

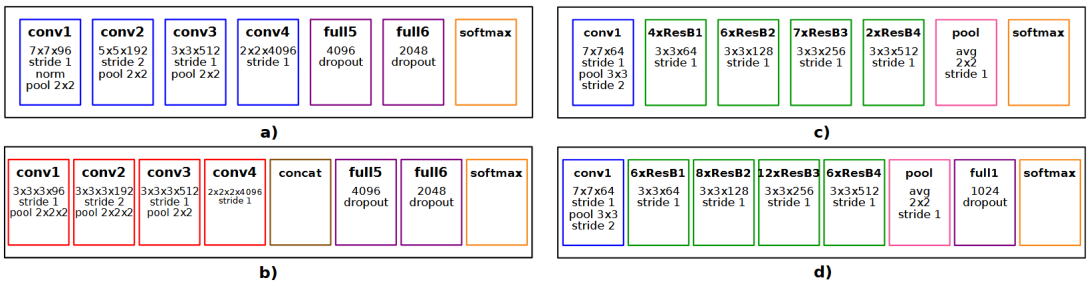

Figure 1.8: Proposed CNN models for gait signature extraction.

2D-CNN: linear CNN with four 2D convolutions, two fully connected layers and a softmax classifier.
b) 3D-CNN: four 3D convolutions, two fully connected layers and a softmax classifier.
c) ResNet-A: residual CNN with a 2D convolution, four residual blocks, an average pooling layer and a final softmax classifier. d) ResNet-B: extended version of ResNet-A. Note that before the first block of each kind (ResB 1, 2, 3, 4), Figure 1.8, there is an adapter convolution to resize the input image to the size of the next

block.

The work targets people identification in video based on the way they walk (i.e. gait) by using deep learning architectures. Authors explore the use of convolutional neural networks (CNN) for learning high-level descriptors from low-level motion features (i.e. optical flow components). The low number of training samples for each subject and the use of a test set containing subjects different from the training ones makes the search of a good CNN architecture a challenging task. Authors carry out a thorough experimental evaluation deploying and analyzing four distinct CNN models with different depth but similar complexity. Authors show that even the simplest CNN models greatly improve the results using shallow classifiers. All experiments have been carried out on the challenging TUM-GAID dataset, which contains people in different covariate scenarios (i.e. clothing, shoes, bags).[28]

Deep Learning-Based Gait Recognition Using Smartphones in the Wild (2020) [29]. Comparing with other biometrics, gait has advantages of being unobtrusive and difficult to conceal. Inertial sensors such as accelerometer and gyroscope are often used to capture gait dynamics. Nowadays, these inertial sensors have commonly been integrated in smartphones and widely used by average person, which makes it very convenient and inexpensive to collect gait data. In this paper, authors study gait recognition using smartphones in the wild. Unlike traditional methods that often require the person to walk along a specified road and/or at a normal walking speed, the proposed method collects inertial gait data under a condition of unconstraint without knowing when, where, and how the user walks. To obtain a high performance of person identification and authentication, deep-learning techniques are presented to learn and model the gait biometrics from the walking data. Specifically, a hybrid deep neural network is proposed for robust gait feature representation, where features in the space domain and in the time domain are successively abstracted by a convolutional neural network and a recurrent neural network. In the experiments, two datasets collected by smartphones on a total of 118 subjects are used for evaluations. Experiments show that the proposed method achieves over 93.5% and 93.7% accuracy in person identification and authentication, respectively. [29]

iLGaCo: Incremental Learning of Gait Covariate Factors (2020) [30]. Gait is a popular biometric pattern used for identifying people based on their way of walking. Traditionally, gait recognition approaches based on deep learning are trained using the whole training dataset. In fact, if new data (classes, view-points, walking conditions, etc.) need to be included, it is necessary to re-train again the model with old and new data samples. In this paper, authors propose iLGaCo, the first incremental learning approach of covariate factors for gait recognition, where the deep model can be updated with new information without re-training it from scratch by using the whole dataset. Instead, our approach performs a shorter training process with the new data and a small subset of previous samples. This way, our model learns new information while retaining previous knowledge. Authors evaluate iLGaCo on CASIA-B dataset in two incremental ways: adding new view-points and adding new walking conditions. In both cases, our results are close to the classical 'training-from-scratch' approach, obtaining a marginal drop in accuracy ranging from 0.2% to 1.2%, what shows the efficacy of our approach. In addition, the comparison of iLGaCo with other incremental learning methods, such as LwF and iCarl, shows a significant improvement in accuracy, between 6% and 15% depending on the experiment. [30] Deep Convolutional Feature-based Gait Recognition Using Silhouettes and RGB Images (2021) [31]. Today, many different biometric features are used for human identification. Unlike biometric features, such as eye, iris, ear, and fingerprint, gait biometrics enables recognition from long distance and low resolution images. In this paper, different design choices for a deep learning-based gait recognition system are investigated in detail. Some preprocessing steps, such as human silhouette extraction and gait cycle calculation are eliminated to make the system suitable for practical applications. To assess different input types' effect on the gait recognition performance, both binary silhouettes and RGB images are given as input to the network. To observe the contribution of transfer learning, authors fine-tuned a pre-trained generic object recognition model with the CASIA-B gait dataset and performed experiments on the OU-ISIR Large Population gait dataset. To observe the

effect of pose variations, authors conducted experiments for both identical-view and cross-view conditions. Successful results are obtained, especially for cross-view gait recognition, compared to different approaches for gait recognition.[31]

1.2.9 Features usage in Gait Recognition

Gait recognition, a specialized form of behavioral biometrics, is the process of identifying individuals based on their unique walking style. Unlike other biometric systems that rely on fingerprints or facial features, gait recognition focuses on the dynamic patterns of human movement. The features used in this process are derived from multiple sources, including visual data captured through cameras, motion sensors that track body dynamics, and pressure systems that measure the forces exerted by the feet. Together, these inputs provide a comprehensive profile of a person's gait, which can then be analyzed and compared for identification or verification purposes.

One major category of gait recognition features is appearance based, or vision based, features. They are extracted directly from video footage and emphasize the shape and movement of the body silhouette over time. Silhouette and shape metrics involve analyzing the outline of the body in motion, including measurements such as aspect ratio, area, and overall height of the silhouette. Another important representation is the Motion History Image (MHI) or the Gait Energy Image (GEI), which are created by averaging silhouettes across a full gait cycle to capture motion dynamics and produce a distinctive gait signature. Structural and contour features also play a role, as they analyze the boundaries and proportions of the human body, allowing for the extraction of relative sizes and relationships between body parts. These vision-based features provide a powerful way to capture both static and dynamic aspects of walking.

Temporal and spatial features form another critical group, as they quantify the rhythm and dimensions of a person's walk. Temporal parameters include walking speed or pace, cadence (the number of steps taken per minute), and the duration of steps or strides. They also account for the stance phase, when the foot is in contact with the ground, and the swing phase, when the foot is in the air. Spatial parameters, on the other hand, measure distances such as step length—the distance between the heel strike of one foot and the opposite foot—and stride length, which is the distance between two consecutive placements of the same foot. These features capture the timing and geometry of gait, offering valuable insights into the unique walking rhythm of each individual.

In short, kinematic and kinetic features describe the motion and forces generated by the body during walking. Kinematic features focus on the dynamics of body parts, such as joint angles at the hip, knee, and ankle, which provide strong discriminatory power for recognition. They also include measurements of displacement, velocity, and acceleration of body segments and joints. Kinetic features, in contrast, relate to the forces involved in walking. Ground reaction forces (GRF) measure the impact of the feet against the ground, while plantar pressures capture the distribution of force across the sole of the foot. Joint torques or moments further describe the rotational forces acting on the joints during movement. Together, these features provide a biomechanical perspective on gait, complementing the visual and temporal-spatial data to create a holistic representation of an individual's walking pattern.

1.3 Research Objectives and Conceptual Framework

The limitations of existing researches in gait recognition highlight a significant challenge in achieving high accuracy across large-scale datasets. For instance, the OU-MVLP dataset, Figure 1.9, one of the most widely used benchmarks in this field, has demonstrated that current models struggle to surpass a rank-1 accuracy of 96%. This shortfall underscores the need for innovative approaches that can push the boundaries of recognition performance and overcome the inherent difficulties posed by multi-view gait analysis.

The overarching goal of our project is to enhance the accuracy of gait recognition models by introducing a novel multi-view training approach. Our conceptual framework draws inspiration from the idea of drug combination therapies used in the treatment of COVID-19, where multiple strategies are combined to

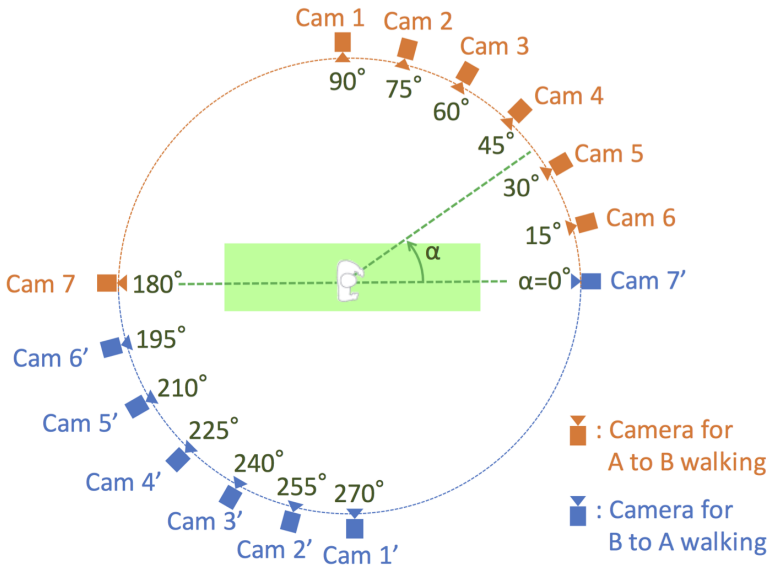

Figure 1.9: Multi-view camera setting

achieve superior outcomes. Similarly, our framework is structured into three distinct steps that work together to improve the discriminative power of gait recognition systems. The first step, Embedding Vector Enrichment via Knowledge Propagation, focuses on enriching the embedding vectors generated from individual camera angles by facilitating information exchange between neighboring angles, such as between 75° and 90° or between 75° and 60°. This is achieved through the message propagation mechanism of a Graph Neural Network (GNN), Figure 3.2, which allows the embedding vector of the central camera angle to absorb contextual information from related perspectives. This exchange of knowledge is projected to increase object discrimination accuracy by approximately 4% to 6%, thereby strengthening the foundation of the recognition process. The exchange of information is an essential requirement in this process, as the subsequent step cannot serve any meaningful purpose without it. If no exchange takes place, the next stage becomes ineffective and ultimately redundant.

The second step, Multi-Perspective Synthesis, builds upon the enriched embedding vectors by synthesizing them into a new model. This step addresses the computational challenges associated with similarity functions. Traditional expensive functions, such as chi-square, are often required to distinguish complex non-matching images, while simpler functions like L2 or cosine similarity may fail in special cases. To overcome this, authors propose the use of a multi-perspective model based on Transformer architecture, which is capable of synthesizing enriched embedding vector information more effectively. By leveraging the Transformer's ability to capture long-range dependencies and contextual relationships, this step ensures that the enriched vectors are combined into a highly robust representation, reducing computational overhead while improving accuracy.

The third step, Inference and Identification, utilizes the enriched embedding vectors obtained from any single angle or a combination of angles generated in Step 1 as input for the model created in Step 2. The output is a highly discriminative embedding vector designed for identity representation. With this refined vector, cosine similarity comparison achieves superior discrimination, enabling the system to distinguish identities with greater precision. The final results produced through this three-step process are expected to outperform those reported in related studies, thereby setting a new benchmark in gait recognition research.

Our proposed model architecture is designed to maximize the strengths of different types of models by combining one reasoner model with multiple descriptor models. Each view is assigned its own descriptor model, which specializes in capturing the details of that particular perspective. The reasoner model, in contrast, does not need to directly observe the images; instead, it reasons about them based on the information

provided by the descriptor models. This separation of roles is a key factor in why our architecture can outperform existing approaches. By allowing descriptor models to collaborate through cross-view training, authors enable them to provide richer and more accurate descriptions of the input data.

It is important to note that describing and reasoning are fundamentally different skills. Convolutional Neural Networks (CNNs) excel at describing fine-grained details within an image, but they are less effective at performing deep reasoning tasks. Conversely, Transformer models are highly capable of reasoning about complex relationships but are not as efficient at capturing low-level descriptive details. By assigning each model to the task it performs best—CNNs for description and Transformers for reasoning—authors achieve outcomes that are superior to those produced by architectures that attempt to force a single model to handle both tasks simultaneously. This division of labor ensures that our system leverages the unique strengths of each model type, resulting in a more powerful and accurate gait recognition framework.

Our proposed model architecture includes 1 reasoner model and multiple descriptor models. Each view has its own descriptor model. The reasoner model doesn't have to see directly to reason about images. It's the reason why our architecture can outperform other architectures, Figure 3.1.

Authors allow descriptor models to collaborate through cross-view training. It helps descriptor to provide better description.

Descripting and reasoning are very different skills. CNN models are better at describing details of a certain image than deep reasoning an image. Transformer models are better at deep reasoning task than describing details of an image. By letting each model stand in their shoes, authors gain better outcomes than forcing a single model to do both tasks.

1.4 Contributions

This thesis makes two significant contributions to the advancement of gait recognition research, each addressing critical limitations in existing methodologies and offering innovative solutions to improve the recognition accuracy. The first contribution lies in the application of the message propagation mechanism, a concept derived from Graph Neural Networks (GNNs), to enrich the information content of embedding vectors associated with specific camera angles. In traditional gait recognition systems, embedding vectors generated from individual viewpoints often suffer from limited contextual awareness, as they capture only the features visible from that particular angle. By introducing message propagation, the embedding vector of a central camera angle—referred to as the center vertex—can exchange information with neighboring angles. This process allows the vector to absorb complementary details that would otherwise remain isolated, thereby enhancing its representational capacity. The enriched embedding vectors become more robust and discriminative, enabling the system to better distinguish between individuals even when the visual input is constrained or partially occluded. This contribution demonstrates how the integration of GNN-inspired mechanisms can significantly strengthen the foundation of gait recognition models by leveraging cross-view knowledge propagation.

The second major contribution of this thesis is the proposal to utilize a multi-perspective model, a Transformer architecture, to synthesize the information obtained from the enriched embedding vectors. While the enrichment process ensures that each vector carries more comprehensive information, the main challenge lies in effectively combining these vectors into a unified representation that can be used for accurate identity recognition. Traditional similarity functions, such as chi-square, cosine, or L2, often fall short in handling the complexity of multi-view data, either due to computational inefficiency or lacking of precision in special cases. To overcome this, the Transformer model is introduced as a synthesis mechanism capable of integrating enriched vectors across multiple perspectives. Transformers excel at capturing long-range dependencies and contextual relationships, making them particularly well-suited for this task. By synthesizing the enriched embedding vectors, the Transformer produces a newly and highly discriminative representation that encapsulates the diverse viewpoints in a coherent and meaningful way. This contribution

not only addresses the limitations of existing similarity-based approaches but also establishes a new pathway for leveraging advanced deep learning architectures in gait recognition.

Together, these contributions form a cohesive framework that enhances both the descriptive & integrative aspects of gait recognition. The application of GNN-inspired message propagation enriches the foundational data, while the Transformer-based synthesis ensures that this enriched information is effectively utilized for identity discrimination. By combining these two innovations, the thesis provides a powerful and scalable approach that has the potential to outperform existing models and set new benchmarks in the field.

1.5 Organization of Thesis

The remainder of this thesis is structured as follows:

- **Chapter 2: Literature Review** discusses the scope of research and reviews relevant works, including Convolutional Neural Networks (CNN) and the Transformer Encoder architecture.
- **Chapter 3: Methodology** details the proposed architecture and training mechanisms for both the Convolutional Neural Network and the Transformer model.
- **Chapter 4: Numerical Results** presents the evaluation parameters, simulation methods, and the performance results of the MVDL model on single-view and two-view reasoning tasks.
- **Chapter 5: Conclusions** summarizes the findings and provides suggestions for future work.

CHAPTER 2. LITERATURE REVIEW

2.1 Scope of Research

This study is focused on identifying the identities of objects using gait as a biometric measure. The research is specifically conducted on the Gait Energy Image (GEI) dataset provided by the Department of Intelligent Media, The Institute of Scientific and Industrial Research, Osaka University. [32]

2.1.1 Our new approach

The overarching goal of this project is to improve the accuracy of gait recognition models by introducing a new multi-view training approach. Our conceptual framework, inspired by combination drug therapies for COVID-19, is structured into three distinct steps:

1. **Step 1: Embedding Vector Enrichment via Knowledge Propagation** Figure 3.2
2. **Step 2: Multi-Perspective Synthesis**
3. **Step 3: Inference and Identification**

2.1.2 Our proposed model architecture

Our proposed model architecture includes 1 reasoner model and multiple descriptor models. Each view has its own descriptor model. The reasoner model doesn't have to see directly to reason about images. It's the reason why our architecture can outperform other architectures. Figure 3.1 We allow descriptor models to collaborate through cross-view training. It helps descriptor provide better description.

Descripting and reasoning are very different skills. CNN models are better at describing details of a certain image than deep reasoning an image. Transformer models are better at deep reasoning task than describing details of an image. By letting each model stand in their shoes, we gain better outcomes than forcing a single model to do both tasks.

2.2 Related Works

The area of multi-view gait recognition has received significant attention. The current state-of-the-art solutions, along with their reported Rank-1 accuracy rates on cross-view tasks, are summarized below. A common limitation across these methods is that their achieved accuracy levels are consistently lower than 96%, and they are typically designed for inference using input from only one camera angle.

2.2.1 Highlighted Researches on OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP)

1. Cross-View Gait Recognition by Discriminative Feature Learning, Table 2.1: The Rank-1 accuracy rate of the proposed model is 95.4% [1]
2. Combining the Silhouette and Skeleton Data for Gait Recognition, Table 2.2: The Rank-1 accuracy rate of the proposed model is 92.5% [2]
3. Gait Recognition Using 3-D Human Body Shape Inference, Table 2.3: The Rank-1 accuracy rate of the proposed model is 91.4% [3]
4. Learning Visual Prompt for Gait Recognition, Table 2.4: The Rank-1 accuracy rate of the proposed model is 93.2% [4]
5. Learning rich features for gait recognition by integrating skeletons and silhouettes, Table 2.5: The Rank-1 accuracy rate of the proposed model is 91.28% [5]

Table 2.1: Cross-View Gait Recognition by Discriminative Feature Learning

Probe	Galley All 14 Views		
	GEINet	GaitSet	proposed
0°	11.4	79.5	74.0
15°	29.1	87.9	88.3
30°	41.5	89.9	94.6
45°	45.5	90.2	95.4
60°	39.5	88.1	88.0
75°	41.8	88.7	91.3
90°	38.9	87.8	90.0
180°	14.9	81.7	76.7
195°	33.1	86.7	89.5
210°	43.2	89.0	95.0
225°	45.6	89.3	94.9
240°	39.4	87.2	88.0
255°	40.5	87.8	90.8
270°	36.3	86.2	89.8
mean	35.8	87.1	89.0

Table 2.2: Combining the Silhouette and Skeleton Data for Gait Recognition

Probe angle	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
GaitGraph2 [13]	54.3	68.4	76.1	76.8	71.5	75.0	70.1	52.2	60.6	57.8	73.2	67.8	70.8	65.3	67.1
FR-GCN [14]	48.3	53.5	56.8	58.9	58.3	55.2	50.6	36.6	49.0	45.5	60.6	60.4	57.4	53.6	53.2
GaitGL [8]	84.9	90.2	91.1	**91.5**	91.1	90.8	**90.3**	88.5	88.6	90.3	**90.4**	89.6	89.5	88.8	89.7
Lagrange [9]	84.5	89.8	91.0	91.2	90.7	90.5	90.2	88.5	87.9	89.9	90.0	89.2	89.2	88.7	89.4
Ours	**91.3**	**92.4**	**91.2**	89.9	**92.1**	**90.9**	90.2	**90.0**	**92.1**	**90.3**	89.5	**92.5**	**90.7**	**90.5**	**91.0**

2.2.2 Key Papers on 1-View (Single-View) Gait Recognition

Most relevant papers focus on frontal-view recognition, as it is the dominant single-view scenario in surveillance and biometrics literature.

Frontal View Gait Recognition With Fusion of Depth Features From a Time of Flight Camera [33]. To address the limitations of traditional frontal view gait recognition—historically reliant on RGB, stereo, or Doppler sensors, "Frontal View Gait Recognition With Fusion of Depth Features From a Time of Flight Camera" introduces a robust four-part framework utilizing Time-of-Flight (ToF) camera feature fusion. The methodology comprises a novel silhouette extraction algorithm to mitigate multiple reflection artifacts, a cycle-detection-based frame selection process, four unique gait image representations, and a specialized fusion classifier. Validation was conducted using a longitudinal dataset of 46 and 33 subjects across two sessions, demonstrating that the proposed approach significantly exceeds state-of-the-art benchmarks. Specifically, the method achieved Rank 1 and Rank 5 recognition rates of 66.1% and 81.0%, respectively, representing a substantial improvement over existing performance peaks of 35.7% and 57.7%.[33]

Skeleton based Frontal Gait Recognition utilizing Fourier Descriptors [34]. Using gait to identify people is a cool new area in biometrics, and its biggest win is that it's totally unobtrusive, unlike other tech, it doesn't need people to stop or touch anything. "Skeleton based Frontal Gait Recognition utilizing Fourier Descriptors" specifically looks at how to recognize a person's walk in tight spots like narrow corridors where they can only be seen from the front. Authors used the Kinect's skeleton tracking to measure the distance between the "spine base" and every other joint (which they call the centroid distance) to create a unique profile for five key frames of a person's stride. Because the Kinect captures depth, they get much more accurate joint measurements from the front than usual, which is a huge plus for surveillance in cramped spaces. They then ran those descriptors through a simple kNN classifier and found that thier method is not only way more accurate but also a lot faster than the older ways of doing things!

Table 2.3: Gait Recognition Using 3-D Human Body Shape Inference

Method	Camera Positions														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet [33]	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41.0	42.5
GaitSet [5]	79.2	87.7	89.9	90.1	87.9	88.6	87.7	81.7	86.4	89.0	89.2	87.2	87.7	86.2	87.0
GaitPart [9]	82.8	89.2	90.9	91.0	89.7	89.9	89.3	85.1	87.7	90.0	90.1	89.0	89.0	88.1	88.7
GaitGL [22]	84.2	89.8	91.3	**91.7**	90.8	**91.0**	90.4	88.1	88.2	**90.5**	**90.5**	89.5	**89.7**	88.8	89.6
GaitSet-HBS	79.0	87.9	90.4	90.6	88.4	89.2	88.4	82.3	87.1	89.6	89.6	87.7	88.4	86.9	87.5
GaitPart-HBP	82.4	89.1	91.1	91.3	89.8	90.2	89.7	84.8	88.0	90.3	90.3	89.2	89.4	88.4	88.9
GaitGL-HBS	**84.7**	**90.2**	**91.4**	**91.7**	**90.9**	**91.0**	**90.5**	**88.4**	**88.7**	**90.5**	**90.5**	**89.6**	89.6	**88.9**	**89.8**

Table 2.4: Learning Visual Prompt for Gait Recognition

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet [7]	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart [14]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [22]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GaitGL [33]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
GaitBase [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90.0
GaitGCI [13]	91.2	92.3	92.6	92.7	93.0	92.3	92.1	92.0	91.8	91.9	**92.6**	92.3	91.4	91.6	92.1
DANet [35]	87.7	91.3	91.6	91.8	91.7	91.4	91.1	90.4	90.3	90.7	90.9	90.5	90.3	89.9	90.7
HSTL [50]	91.4	92.9	**92.7**	**93.0**	92.9	**92.5**	**92.5**	92.7	92.3	**92.1**	92.3	92.2	91.8	91.8	**92.4**
VPNet-M	**91.9**	**93.0**	92.4	92.7	**93.2**	**92.5**	92.3	**92.9**	**92.4**	91.9	92.1	**92.5**	**91.9**	**91.9**	**92.4**

Human Gait Recognition Based on Frontal-View Sequences Using Gait Dynamics and Deep Learning [35]. "Human Gait Recognition Based on Frontal-View Sequences Using Gait Dynamics and Deep Learning" introduces a new way to identify people just by how they walk using a front-on view like seeing someone walk toward the camera in a hallway, combined with the power of deep learning. Most current systems try to look at people from the side, but in the real world (like in a crowded building), the camera often only get a view from the front. The proposed method focuses on three main things: how the limbs move, the body proportions, and the overall shape people make while walking. Instead of just looking at a single snapshot, the system watches how these shapes change over time to capture people's unique "walking rhythm.". Authors then feed this info into a smart computer program (a deep learning model) that learns to recognize these patterns. To make sure the system doesn't get confused if objects are wearing different clothes or carrying a bag, authorts added a special "fusion" step that double-checks the data for errors. When they tested it against older methods, thier system was much more reliable and accurate at picking the right person out of a crowd.

Human gait recognition based on frontal view using kinect features and orthogonal least square selection [35]. Paper "Human gait recognition based on frontal view using kinect features and orthogonal least square selection" looks at how human can identify people by the way they walk, even if they are walking directly toward a camera or at an angle. To do this, authors used a Kinect sensor (like the one used for video games) to track a person's 3D skeleton as they moved. Because a skeleton has so many different moving parts, they used a "smart filter" technique called Orthogonal Least Square (OLS) to pick out only the most important movements that make a person's walk unique. Then they fed these specific movements into a digital "brain" called a neural network to see if it could correctly name the person. The results were impressive: the system was especially good at recognizing people from a front-on view, hitting a 90.6% accuracy rate.

2.2.3 Key Papers on Multi-View Gait Recognition

Multi-view (or cross-view) gait recognition focuses on handling appearance variations across different camera angles, a core challenge in real-world biometrics.

Multi-View Gait Recognition Based on a Spatial-Temporal Deep Neural Network, Table 2.6 [36].

Figure 2.1 shows a spatial-temporal deep neural network. This network is adopted to extract gait feature

Table 2.5: Learning rich features for gait recognition by integrating skeletons and silhouettes

Gallery #01, Probe #00	Probe View														mean
Methods	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
ST-GCN [34]	24.3	34.9	39.8	42.1	41.5	38.5	33.3	23.0	27.0	25.7	35.6	35.3	32.5	28.2	33.0
MSGG(ours)	43.8	58.8	64.0	66.4	65.9	62.9	57.8	40.6	48.4	44.4	60.6	60.3	56.6	51.8	55.9
GaitSet [11]	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart [12]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [14]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	**90.5**	90.6	89.6	89.3	88.5	89.2
GaitGL [13]	84.9	90.2	91.1	91.5	**91.1**	**90.8**	90.3	**88.5**	88.6	90.3	90.4	89.6	89.5	88.8	89.7
silhouette-module(base)	82.57	88.93	90.84	91.00	89.75	89.91	89.50	85.19	88.09	90.02	90.15	89.03	89.10	88.24	88.74
BiFusion(ours)	**86.17**	**90.60**	**91.28**	**91.56**	90.88	**90.78**	**90.48**	87.76	**89.48**	90.38	**90.65**	**89.95**	**89.82**	**89.32**	**89.94**

Table 2.6: The Accuracies of Different Methods Evaluated on OU-ISIR

Method	Description	Accuracy
Aaron F. Bobick et al. [15]	MEI-MHI	72.61%
Toby H.W. Lam et al. [13]	SST-MSCT	82.47%
Zifeng Wu et al. [19]	CNN-3DCNN	92.76%
Zifeng Wu et al. [19]	CNN-CGI	91.42%
Spatial Feature Network	*SFN*	***92.64%***
Temporal Feature Network	*TFN*	***93.87%***
Spatial-Temporal Network	*STDNN*	***95.67%***

so as to overcome the influence of view variations. TFN adopts the convolution operation-based STG and LSTM to extract the spatial-temporal gradient feature used for gait recognition, which greatly reduces computation cost.

Paper "The Accuracies of Different Methods Evaluated on OU-ISIR" introduces a new computer system called STDNN that identifies people by their walk, even when seen from different camera angles. The system works like a brain with two specialized parts: one part focuses on time (how the body moves and flows over a few seconds), and the other focuses on space (the overall shape and outlines of the body). The "time" part uses a special memory unit to track the rhythm of object's steps, while the "space" part uses smart filters to learn exactly what makes object's body shape unique compared to everyone else. By combining these two types of information, the system gets a very complete picture of a person's walk. When authors tested it, it was incredibly accurate—correctly identifying people over 95% of the time when the camera angle was the same, and over 92% of the time even when switching between different views. This makes it much more powerful than older methods and very useful for real-world security.

Cross-View Gait Recognition Using Non-Linear View Transformations of Spatiotemporal Features [37]. Paper "Cross-View Gait Recognition Using Non-Linear View Transformations of Spatiotemporal Features" introduces a new way to identify people by how they walk, even if the camera sees them from a side view and then later from a front view. Normally, computers get confused when the camera angle changes, but the system solves this by automatically "translating" any walking angle into one standard view. Authors built a smart computer network that learns to do this on its own without needing a human to tell it which angle it is looking at. It simply finds the common patterns in how a person moves through space and time and maps them to a single, consistent model. Once all the walking styles are translated to this same view, the system can easily compare them to find a match. When they tested this against other top methods, their system was better at correctly identifying people across different camera angles.

Cross-View Gait Recognition by Discriminative Feature Learning [38]. Lately, using "deep learning"

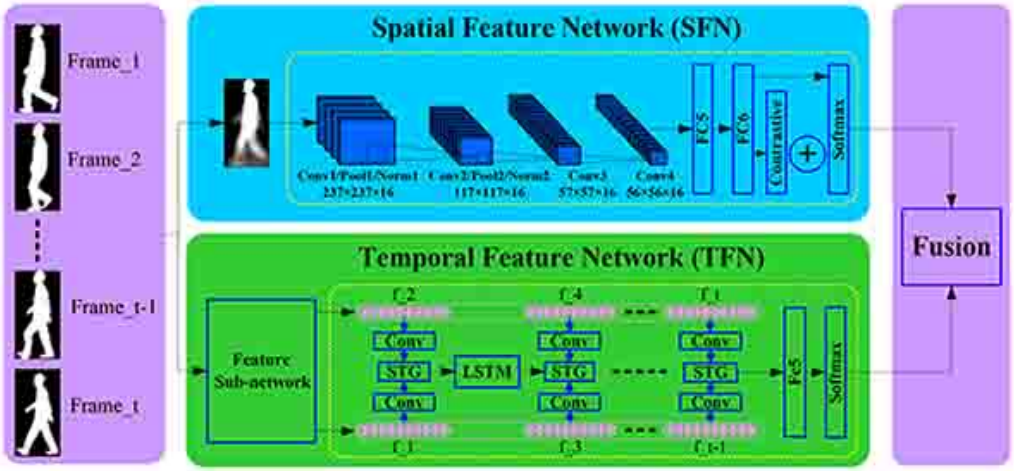

Figure 2.1: Spatial-temporal deep neural network

computers to identify people by their walk has become very popular because they are great at spotting patterns. However, many current systems use a "comparison" method that struggles when it has to tell the difference between two very similar-looking people. To fix this, authors created a new tool called Angle Center Loss (ACL). Instead of just remembering one general "center" for how a person walks, the system remembers what that person looks like from every different camera angle. It focuses on closing the gap between these different views so the person's profile stays consistent. They also improved how the computer "watches" the person. They built a system that automatically picks out the most important body parts, like the legs or torso, and ignores the rest. Then, they added a "temporal attention" feature—think of this as a spotlight that pays more attention to clear walking frames and ignores blurry or bad ones. By focusing on the best parts of the body and the clearest moments of the walk, their system achieved top-tier results, beating out older methods that just averaged everything together.

Cross-View Gait Recognition Based on Feature Fusion [39]. Compared to face recognition, gait recognition is one of the most promising video biometric recognition technologies given that gait images can be readily captured at a distance and gait characteristics are robust to appearance camouflage. A lot of existing gait recognition methods aim at a single scene such as fixed cameras, but the recognition accuracy will decrease sharply if the viewpoints are changed. In paper "Cross-View Gait Recognition Based on Feature Fusion", authors improve the existing methods and propose a cross-view gait recognition method based on feature fusion. Firstly, a multi-scale feature fusion module is proposed to extract the features of gait sequences with different granularities. Then, a dual-path structure is introduced to learn global appearance features and fine-grained local features, respectively. The features of two paths are gradually merged as the network deepens to obtain the complementary information. In the last feature mapping stage, the Generalized-Mean pooling is used to favour discriminative representation. Extensive experiments on the public dataset CASIA-B show that the method can achieve state-of-the-art recognition performance.

Cross-View Gait Recognition with Deep Universal Linear Embeddings [40]. Gait is considered an attractive biometric identifier for its non-invasive and non-cooperative features compared with other biometric identifiers such as fingerprint and iris. At present, cross-view gait recognition methods always establish representations from various deep convolutional networks for recognition and ignore the potential dynamical information of the gait sequences. If assuming that pedestrians have different walking patterns, gait recognition can be performed by calculating their dynamical features from each view. Paper "Cross-View Gait Recognition with Deep Universal Linear Embeddings" introduces the Koopman operator theory to gait recognition, which can find an embedding space for a global linear approximation of a nonlinear dynamical system. Furthermore, a novel framework based on convolutional variational autoencoder and deep Koopman embedding

is proposed to approximate the Koopman operators, which is used as dynamical features from the linearized embedding space for cross-view gait recognition. It gives solid physical interpretability for a gait recognition system. Experiments on a large public dataset, OU-MVLP, prove the effectiveness of the proposed method.

Cross-View Gait Recognition Model Combining Multi-Scale Feature Residual Structure and Self-Attention Mechanism [41]. In the cross-view condition, the gait recognition rate caused by the vastly different gait

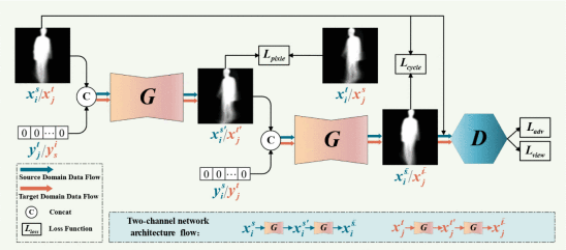

Figure 2.2: Overall structure of the network model.

silhouette maps is substantially reduced. To improve the accuracy of gait recognition under cross-view conditions, Figure 2.2, paper "Cross-View Gait Recognition Model Combining Multi-Scale Feature Residual Structure and Self-Attention Mechanism" proposes a cross-view gait recognition network model combining multi-scale feature residual module (MFRM) and self-attention (SA) mechanism based on (GAN). First, the local and global feature information in the input gait energy image is fully extracted using the MFRM. Then, the SA mechanism module is used to adjust the information of channel dimensions and capture the association between feature information and is introduced into both the generator and discriminator. Next, the model is trained using a two-channel network training strategy to avoid the pattern collapse problem during training. Finally, the generator and discriminator are optimized to improve the quality of the generated gait images. It conducts experiments using the CASIA-B and OU-MVLP public datasets. The experiments demonstrate that the MFRM can better obtain the local and global feature information of the images. The SA mechanism module can effectively establish global dependencies between features, so that the generated gait images have clearer and richer detail information. The average Rank-1 recognition accuracies of the results reach 91.1% and 97.8% on the two datasets respectively, which are both better than the current commonly used algorithms, indicating that the network model in the paper can well improve the gait recognition accuracy across perspectives

Batch Hard Contrastive Loss and Its Application to Cross-View Gait Recognition [42].

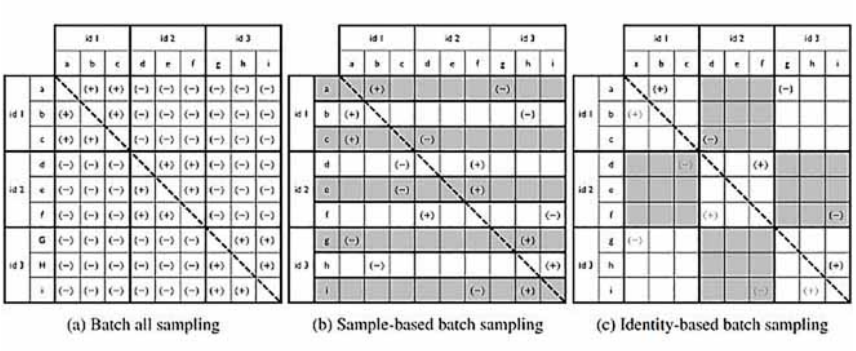

Figure 2.3: Three batch-sampling methods comparison

This study compared three batch-sampling methods: Batch-all, sample-based, and identity-based. Each batch contained several identities with an equal number of samples. Identity-based sampling with contrastive loss delivered the best performance for gait verification, achieving 0.50% equal error rate in training GaitSet, Figure 2.3.

Biometric person authentication comprises two tasks: the identification task (i.e., one-to-many matching) and the verification task (i.e., one-to-one matching). In paper "Batch Hard Contrastive Loss and Its Application to Cross-View Gait Recognition", authors propose a loss function called batch hard contrastive loss (BHCn) for the deep learning-based verification task. For this purpose, they consider batch mining techniques developed in the identification task and translate them to the verification task. More specifically, inspired by batch mining triplet losses to learn a relative distance for the identification task, they propose BHCn to learn an absolute distance that better represents verification in general. Their method preserves the identity-agnostic nature of the contrastive loss by selecting the hardest pair of samples for each pair of identities in a batch instead of selecting the hardest pair for each sample. They validate the effectiveness of the proposed method in cross-view gait recognition using three networks: a lightweight input, structure, and output network they call GEI + CNN (Gait Energy Image Convolutional Neural Network) as well as the widely used GaitSet and GaitGL, which have sophisticated inputs, structures, and outputs. They trained these networks with the publicly available silhouette-based datasets, the OU-ISIR Gait Database Multi-View Large Population (OU-MVLP) dataset and the Institute of Automation Chinese Academy of Sciences Gait Database Multiview (CASIA-B) dataset. Experimental results show that the proposed BHCn outperforms other loss functions, such as a triplet loss with batch mining as well as the conventional contrastive loss.

GaitDAN: Cross-View Gait Recognition via Adversarial Domain Adaptation [43]. View change causes significant differences in the gait appearance. Consequently, recognizing gait in cross-view scenarios is highly challenging. Most recent approaches either convert the gait from the original view to the target view before recognition is carried out or extract the gait feature irrelevant to the camera view through either brute force learning or decouple learning. However, these approaches have many constraints, such as the difficulty of handling unknown camera views. This work treats the view-change issue as a domain-change issue and proposes to tackle this problem through adversarial domain adaptation. This way, gait information from different views is regarded as the data from different sub-domains. The proposed approach focuses on adapting the gait feature differences caused by such sub-domain change and, at the same time, maintaining sufficient discriminability across the different people. For this purpose, a Hierarchical Feature Aggregation (HFA) strategy is proposed for discriminative feature extraction. By incorporating HFA, the feature extractor can well aggregate the spatial-temporal feature across the various stages of the network and thereby comprehensive gait features can be obtained. Then, an Adversarial View-change Elimination (AVE) module equipped with a set of explicit models for recognizing the different gait viewpoints is proposed. Through the adversarial learning process, AVE would not be able to identify the gait viewpoint in the end, given the gait features generated by the feature extractor. That is, the adversarial domain adaptation mitigates the view change factor, and discriminative gait features that are compatible with all sub-domains are effectively extracted. Extensive experiments on three of the most popular public datasets, CASIA-B, OULP, and OUMVLP richly demonstrate the effectiveness of the approach.

TAG: A Temporal Attentive Gait Network for Cross-View Gait Recognition [44]. Recognizing a person from a distance using gait (i.e., walking pattern) is a challenging yet interesting biometric problem. Despite recent advancements in deep learning-based gait recognition (GR) research, learning discriminative gait temporal representation is still challenging because of delicate silhouette differences in the spatial domain. Aiming to address this issue, authors propose a novel attention-based GR framework, namely, temporal attentive gait (TAG), which aims to refine the gait feature representation from the temporal dimension's perspective in a comprehensive fashion. The proposed TAG mainly consists of three modules, namely, short-term temporal feature learning (ST-TFL), hybrid multikernel temporal attention (H-MKTA), and multikernel temporal self-attention (MK-TSA), respectively. First, ST-TFL aims to capture local temporal contextual clues, facilitating the learning of short-period temporal motion patterns. Second, H-MKTA learns locally and globally distributed gait temporal information by adaptively capturing the multiscale temporal evolutions inside the gait sequential data. To refine the temporal attentive features learned by ST-TFL and H-MKTA, the MK-TSA learns global dependencies between gait temporal frames to recalibrate

temporal weights using a self-attention mechanism. To further enhance the discriminative power of the gait feature representation, a multilevel (ML) framework is adopted, combining gait features from different levels of the backbone. Experiments conducted on three benchmark gait datasets, CASIA-B, Gait3D, and CCPG, demonstrate the strong potential of the TAG in learning effective gait representation under complex scenarios for recognition.

Cross-View Gait Recognition via View Information Elimination Mechanism [45].

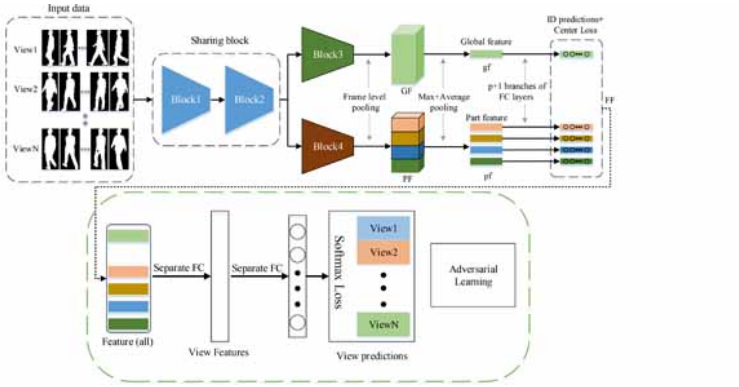

Figure 2.4: View Information Elimination Mechanism (VIEM) framework

The overall pipeline of View Information Elimination Mechanism (VIEM) framework, Figure 2.4. First, the human silhouette sequences from different views are processed by a sharing block, which is composed of 4 ResNet blocks to extract basic gait features. Concretely, the ResNet Block1 and Block 2 are used to extract original deep features. Next the Block3 and Block4 learn global features and local features respectively, and part features are obtained by horizontal segmentation. Then the global and local features are supervised by identity classification loss and a newly designed center loss. To eliminate the influence of viewpoint information, the extracted global and local features are concatenated and fed into the viewpoint elimination module (indicated by the green dashed section in the figure below). Specifically, the concatenated features are processed by two separate fully connected layers and then fed into N different viewpoint classifiers to learn viewpoint features. The designed adversarial learning loss reduces viewpoint features through reverse optimization, while the gait ID loss in the upper part of the figure aims to enhance the discriminative power of each pedestrian's gait features. The top and bottom parts are jointly optimized using Equ. 14 to form an adversarial learning structure. Ultimately, this approach yields pedestrian gait features that are robust to viewpoint transformations and have strong ID discrimination.

Gait recognition is a new and promising biometric identification method, which has many advantages, such as requiring no contact, long-distance, and hard to imitate. However, gait recognition faces many challenges in real-life scenarios, including occlusion, diverse view angles, cloth change and carrying variances. Among those, the change of viewpoints would be one of the trickiest factors. Thus, effectively mitigating or entirely eliminating the influence of perspective factors stands as a viable approach to enhance the performance of gait recognition. A novel mechanism for perspective elimination in gait recognition is proposed in the paper, aimed at mitigating the impact of viewpoints. Firstly, an adversarial learning framework is devised between the feature extractor and the perspective classifier, which confounds the discriminate ability of the perspective classifier regarding gait features, thereby reducing the influence of perspective from the feature-wise. Secondly, a redesigned center loss is proposed to draw the features with the same ID from different views close to identity centers while increasing the inter-class distances. Finally, the view-invariant identity-wise gait features are learned. The proposed method achieves mean Rank-1 = 99.2%, 97.6%, 94.0% on CASIA-B (NM, BG, CL), and mean Rank-1 = 91.0% on OU-MVLP, demonstrate the effectiveness of the method.

Multi-View Gait Recognition With Joint Local Multi-Scale and Global Contextual Spatio-Temporal

Features [46]. Existing gait recognition methods are capable of extracting rich spatial gait information but often overlook fine-grained temporal features within local regions and temporal contextual information across different sub-regions. Considering gait recognition as a fine-grained recognition task and each individual exhibits uniqueness in their movements across different temporal sequences, authors propose a local multi-scale and global contextual spatio-temporal (LMGCS) network for gait recognition. It divides the whole gait sequence into sub-sequences with multiple spatio resolutions and extracts multi-scale temporal features. They extract the temporal context information of different sub-sequences with the transformer, and all sub-sequences are fused to form global features. Furthermore, the loss function that combines the triplet loss function and cross-entropy loss function is utilized to prompt the proposed model to fulfill the gait recognition. The proposed method achieved state-of-the-art results on two popular public datasets. It achieved rank-1 accuracy of 98.0%, 95.4%, and 85.0% on the three walk states of the CASIA-B dataset and 90.9% on the OU-MVLP dataset.

Multi-view gait recognition using 3D convolutional neural networks [47]. In paper "Multi-view gait recognition using 3D convolutional neural networks" authors present a deep convolutional neural network using 3D convolutions for Gait Recognition in multiple views capturing spatio-temporal features. A special input format, consisting of the gray-scale image and optical flow enhance color invariance. The approach is evaluated on three different datasets, including variances in clothing, walking speeds and the view angle. In contrast to most state-of-the-art Gait Recognition systems the used neural network is able to generalize gait features across multiple large view angle changes. The results show a comparable to better performance in comparison with previous approaches, especially for large view differences.

The chapter is dedicated to providing a comprehensive description of all the papers and studies that are directly related to the current research, offering a detailed overview of the existing body of work in the field. In addition to reviewing prior contributions, it also emphasizes the fundamental concepts and methodologies of Convolutional Neural Networks (CNNs) and Transformer architectures, which serve as the foundational basis for the proposed model. By examining both the relevant literature and the theoretical underpinnings of these two powerful deep learning approaches, the chapter establishes a clear connection between past research efforts and the innovative framework being introduced, thereby situating the proposed model within a broader academic and technological context.

3.1 Overview

The proposed methodology, named **Multi-view gait recognition with Deep Learning (MVDL)**, integrates a Convolutional Neural Network (CNN) for initial feature extraction with a Transformer model for multi-view knowledge synthesis. This integrated approach is designed to enrich feature embeddings and enhance accuracy, particularly for single-view inference.

3.1.1 Our proposed model architecture

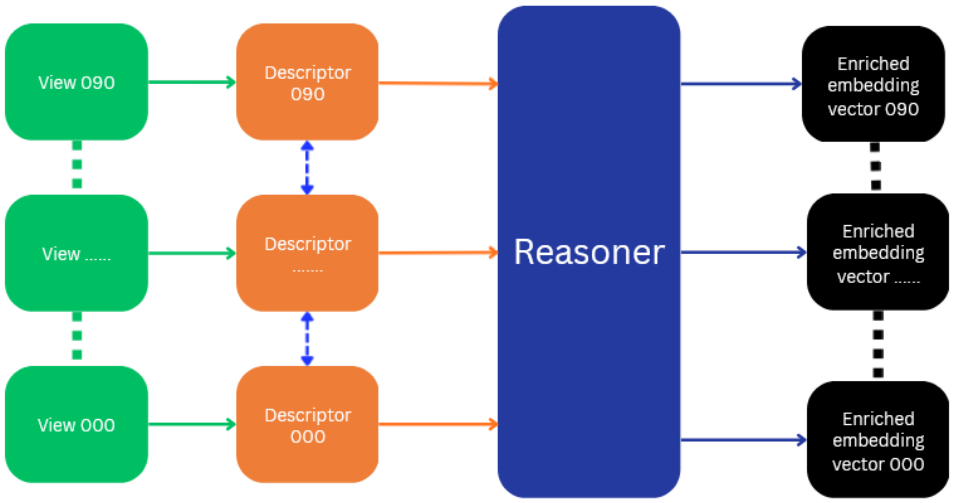

Figure 3.1: Our proposed model architecture

Our proposed model architecture includes 1 reasoner model and multiple descriptor models. Each view has its own descriptor model. The reasoner model doesn't have to see directly to reason about images. It's the reason why our architecture can outperform other architectures.

We allow descriptor models to collaborate through cross-view training. It helps descriptor to provide better description.

Descripting and reasoning are very different skills. CNN models are better at describing details of a certain image than deep reasoning an image. Transformer models are better at deep reasoning task than describing details of an image. By letting each model stand in their shoes, we gain better outcomes than forcing a single model to do both tasks.

The overall workflow of the solution is structured as a two-stage process:

Figure 3.2: Example of how information can be exchanged and enriched between one camera angle and related angles

Embedding Enrichment (GNN-Inspired Knowledge Propagation): The initial CNN models, one trained per camera angle, exchange knowledge with neighboring angles (e.g., 75° exchanging information with 60° and 90°). This mechanism, inspired by Graph Neural Network (GNN) message propagation, is aimed at improving the single-angle model's predictive accuracy and robustness against overfitting.

We hypothesized that the information between pixels in a 90-degree image is similar to the information between pixels in a 75-degree image. Experiments showed that some identities were unrecognizable in 90-degree images but were recognizable in 75-degree images. This suggests that if we train the model, previously trained on the 90-degree dataset, on the 75-degree dataset, the new recognition system will be able to recognize identities from both 75-degree and 90-degree images. If we train this new recognition system again on the 90-degree dataset, it will utilize the knowledge gained from training the 75-degree dataset to recognize identities from 90-degree images, thereby improving the accuracy of the CNN model for 90-degree images. From this base, we trained in sequence from 90 degrees to 75 degrees to 60 degrees to 45 degrees to 30 degrees to 15 degrees to 00 degrees, then from 00 degrees to 30 degrees to 45 degrees to 60 degrees to 75 degrees to 90 degrees.

This method ensures that the model at any given angle contains a portion of the model's knowledge at any angle from 00 degrees to 90 degrees. This is the basis for the subsequent model to have superior accuracy in the inference process.

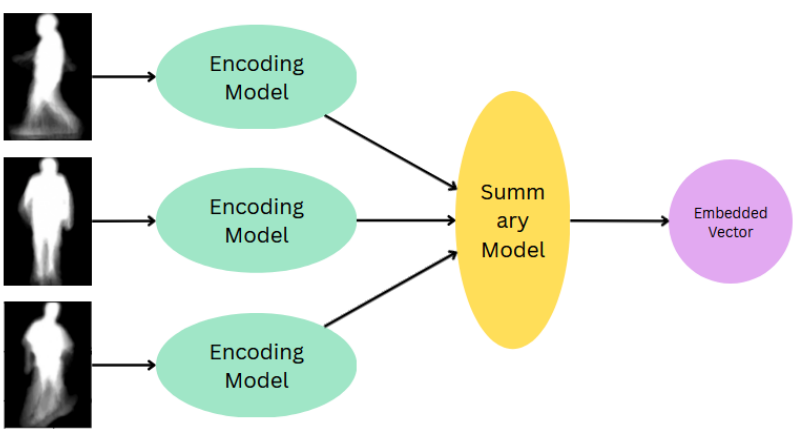

Figure 3.3: Proposed solution's workflow

Multi-View Synthesis (Transformer Model): The enriched embedding vectors from multiple CNNs are concatenated and fed into a Transformer model to synthesize the information into a single, highly discriminative embedding vector.

3.2 Our model explanation in math

3.2.1 Convolutional Neural Network (CNN)

The proposed methodology utilizes a Convolutional Neural Network (CNN) as the initial feature extractor. A CNN architecture consists of convolutional layers, activation functions, pooling layers, and fully connected layers. We propose using the hyperbolic tangent (tanh) as the activation function, which applies a nonlinearity after each convolution or fully connected layer, mapping inputs to values between -1 and 1.

1. Convolution Operation: For an input image or feature map X and a filter/kernel W, the convolution operation at position (i,j) is:

$$Z_{i,j}^{(l)} = (W^{(l)} * X^{(l-1)}) + b^{(l)} \tag{3.1}$$

- $Z^{(l)}$: Pre-activation output at layer l
- $X^{(l-1)}$: Input from previous layer

- $W^{(l)}$: Filter weights
- $b^{(l)}$: Bias term
- *: Convolution operator

2. Batch Normalization:

   Normalize the output $Z^{(l)}$

$$\hat{Z}_{i,j}^{(l)} = \frac{Z_{i,j}^{(l)} - \mu^{(l)}}{\sqrt{(\sigma^{(l)})^2 + \epsilon}} \tag{3.2}$$

   Then scale and shift:

$$\tilde{Z}_{i,j}^{(l)} = \gamma^{(l)} \cdot \hat{Z}_{i,j}^{(l)} + \beta^{(l)} \tag{3.3}$$

   - $\mu^{(l)}$: Mean of the batch
   - $\sigma^{(l)}$: Standard deviation of the batch
   - ϵ : Small constant for numerical stability
   - $\gamma^{(l)}, \beta^{(l)}$: Learnable parameters

3. Max Pooling: Pooling reduces spatial dimensions:

$$P_{i,j}^{(l)} = \text{pool}(\hat{Z_{i,j}^{(l)}}) \tag{3.4}$$

4. Tanh Activation Function: Tanh activation is applied element-wise

$$A_{i,j}^{(l)} = \tanh(Z_{i,j}^{(l)}) = \frac{e^{Z_{i,j}^{(l)}} - e^{-Z_{i,j}^{(l)}}}{e^{Z_{i,j}^{(l)}} + e^{-Z_{i,j}^{(l)}}} \tag{3.5}$$

5. Fully Connected Layer: At the end, flattened features are passed through dense layers

$$Z^{(fc)} = W^{(fc)} \cdot x + b^{(fc)} \tag{3.6}$$

$$A^{(fc)} = \tanh(Z^{(fc)}) \tag{3.7}$$

3.2.2 Transfomer Encoder

The Transformer model is utilized in the proposed methodology for synthesizing enriched embedding vectors. The encoder block is composed of a stack of identical layers, with each layer consisting of a Multi-Head Self-Attention (MHSA) mechanism and a position-wise Feed-Forward Network (FFN).

1. Input and Positional Encoding
   If $X = (x_1, x_2, \ldots, x_n)$ is the sequence of input embeddings, where $x_t \in R^d$ and d is the model's dimension, the input to the first encoder layer $Z^{(0)}$ is:

$$Z^{(0)} = X + PE \tag{3.8}$$

   where PE has the same dimension as X and is usually calculated using sine and cosine functions. For a position t and dimension i:

$$PE(t, 2i) = sin(\frac{t}{10000^{2i/d}}) \tag{3.9}$$

$$PE(t, 2i+1) = cos(\frac{t}{10000^{2i/d}}) \tag{3.10}$$

2. Encoder Layer
   A single encoder layer, L, takes an input matrix $Z^{(l-1)}$ from the previous layer (or the input embeddings for l=1) and produces an output matrix $Z^{(l)}$ through the following steps:

   (a) Multi-Head Self-Attention (MHSA): The MHSA sub-layer first performs Scaled Dot-Product Attention Attention(Q,K,V):

   $$Attention(Q, K, V) = A \cdot V \tag{3.11}$$

   where:

   - The Attention Weights A are computed as:

   $$A = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{3.12}$$

   - Query (Q), Key (K), and Value (V) matrices are linear projections of the input $Z^{(l-1)}$:

   $$Q = Z^{(l-1)}W_Q \tag{3.13}$$

   $$K = Z^{(l-1)}W_K \tag{3.14}$$

   $$V = Z^{(l-1)}W_V \tag{3.15}$$

   where $W_Q, W_K, W_V \in R^{d \times d_k}$ are learned weight matrices, and d_k is the dimension of the key/query space.

   Multi-Head Attention computes the attention output h times in parallel (the "heads"), each with its own weight matrices, and then concatenates the results:

   $$H_i = Attention(Z^{(l-1)}W_{Q_i}, Z^{(l-1)}W_{K_i}, Z^{(l-1)}W_{V_i}) \tag{3.16}$$

   $$Z_{MHSA} = [H_1; H_2; \ldots; H_h]W^O \tag{3.17}$$

   where $W^O \in R^{(h \cdot d_k) \times d}$ is the final linear projection matrix.

   The output of the first sub-layer, incorporating Layer Normalization (LN) and a Residual Connection (RC), is:

   $$Z'^{(l)} = LN(Z^{(l-1)} + Z_{MHSA}) \tag{3.18}$$

   (b) Feed-Forward Network (FFN): The FFN is applied independently and identically to each position in the sequence. It consists of two linear transformations with a non-linearity in between.

   $$FFN(z') = max(0, z'W_1 + b_1)W_2 + b_2 \tag{3.19}$$

   The second sub-layer output $Z^{(l)}$ is then calculated by applying the FFN, followed by another residual connection and layer normalization:

   $$Z^{(l)} = LN(Z'^{(l)} + FFN(Z'^{(l)})) \tag{3.20}$$

3.3 Convolutional neural network model - Descriptor model

The backbone architecture of the proposed model, Figure 3.4, begins with a Convolutional Neural Network (CNN), which serves as the initial encoding mechanism responsible for processing the Gait Energy

Images (GEI) input. This CNN-based design provides a structured pipeline that gradually transforms raw gait images into meaningful feature representations suitable for downstream tasks. Each component of the architecture is carefully selected to balance computational efficiency, robustness, and accuracy, ensuring that the model can effectively capture both low-level and high-level features from the input data.

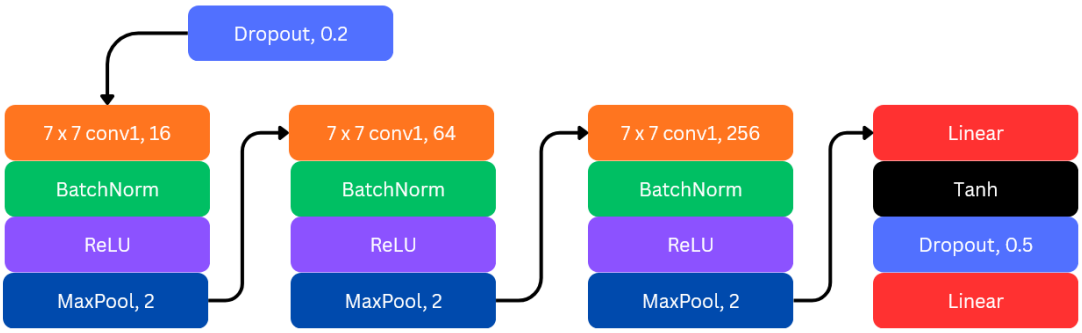

Figure 3.4: Backbone Architecture

The first stage of the architecture is the input layer, which incorporates a dropout mechanism with a parameter value of . Dropout acts as a regularization technique designed to prevent overfitting by reducing feature co-adaptation. By randomly setting a fraction of input units to zero during training, dropout introduces noise into the system, which functions as a simple yet effective form of data augmentation. This randomness forces the network to learn more generalized patterns rather than relying too heavily on specific features, thereby improving its ability to generalize to unseen data.

Following the input stage, the model employs a feature block layer composed of a 7×7 convolution, batch normalization, ReLU activation, and a max pooling operation with a stride of 2. The choice of a 7×7 kernel is deliberate, as it provides computational efficiency while rapidly capturing large-scale, low-level features. This kernel size establishes a broad initial receptive field, enabling the network to detect prominent structural patterns in the gait images early in the process. The inclusion of batch normalization within this block further stabilizes the learning process, while the ReLU activation introduces non-linearity, allowing the network to model complex relationships between features. The max pooling operation reduces dimensionality, ensuring that the most salient features are preserved while suppressing irrelevant noise.

The architecture then incorporates a normalization layer in the form of batch normalization with learnable parameters. This layer plays a crucial role in stabilizing and accelerating training by normalizing the input to each subsequent layer. By maintaining consistent distributions of activations, batch normalization mitigates issues such as internal covariate shift, thereby improving convergence speed and overall model performance.

Next, the activation function is applied using the ReLU mechanism. ReLU is particularly effective in addressing the vanishing gradient problem, which often hampers the training of deep networks. By introducing non-linearity, ReLU enables the network to learn more complex mappings between inputs and outputs. Additionally, ReLU encourages sparse feature representation, which enhances computational efficiency and reduces redundancy in the learned features.

The model then performs down-sampling through a max pooling operation with a 2×2 kernel. This step reduces the spatial dimensions of the feature maps by approximately 75%, ensuring translation invariance and further suppressing noise. By retaining only the most prominent features, max pooling enhances the robustness of the learned representations while simultaneously reducing computational complexity.

Finally, the architecture concludes with the output block, which consists of a sequence of linear layers, a Tanh activation, and an additional dropout mechanism with a rate of 0.5. The Tanh activation ensures that the output activations are zero-centered and bounded, preventing gradients from being consistently positive

or negative. This property leads to more efficient training and faster convergence. The dropout mechanism in this stage further reduces the risk of overfitting by introducing additional regularization. Together, these components produce a refined output representation that is both discriminative and stable, forming the basis for accurate gait recognition.

This backbone architecture, with its carefully orchestrated layers and mechanisms, provides a strong foundation for the model. By combining dropout regularization, kernel convolution, batch normalization, ReLU activation, max pooling, and a structured output block, the design ensures that the system is capable of extracting meaningful features from gait images while maintaining efficiency and robustness throughout the training process.

The use of a fully connected layer in combination with the Tanh activation function and dropout is not the most optimal solution, as the effectiveness of this approach tends to rely heavily on random chance. The stochastic nature of dropout introduces variability that can make the results unstable and less predictable, which reduces the reliability of the overall model.

3.4 Transformer model - Reasoner model

Although the data enrichment mechanism enables Convolutional Neural Network (CNN) models to achieve state-of-the-art (SOTA) results with accuracy levels exceeding 90%, this improvement alone is not sufficient to surpass the performance reported in existing published studies. To address this limitation, we propose a second stage in the framework where a Transformer model is introduced to synthesize the enriched embedding vectors, Figure 3.5. The Transformer serves as a reasoner model, complementing the descriptive capabilities of CNNs by integrating and refining the enriched information into a more discriminative representation. This synthesis process is crucial for bridging the gap between high-performing CNN-based approaches and the superior benchmarks established in prior research.

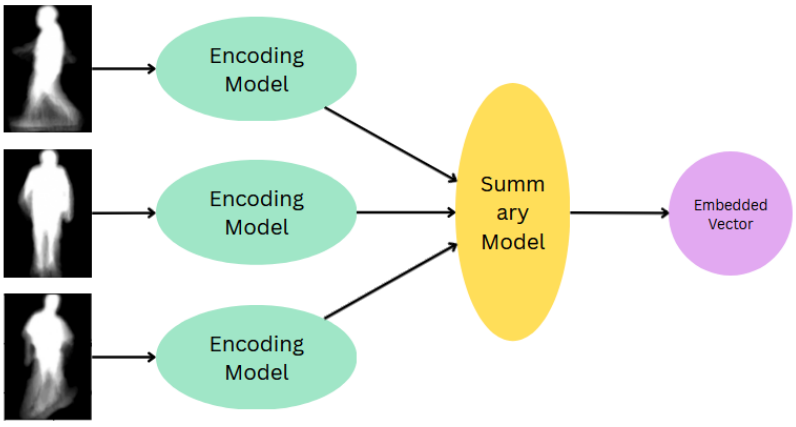

Figure 3.5: Proposed solution's workflow

To clarify the conceptual foundation of this approach, we draw an analogy between gait recognition and authorship. In this analogy, each individual in the dataset is regarded as a "writer," while each gait image or instance is considered an "essay" authored by that person. By examining these essays—represented as embedding vectors—we can identify the unique "writing style" of each individual, which corresponds to their identity. The Transformer model functions as a tool for mining the most common and distinctive patterns across multiple essays or embeddings that belong to the same person. In doing so, it captures the underlying identity traits that remain consistent across different viewpoints, thereby enhancing recognition accuracy.

The Transformer model operates in two distinct phases: training and inference. During the training phase, the input data consists of concatenated embedding vectors generated from the enriched CNN models

at multiple camera angles, specifically 45°, 60°, 75°, and 90°. These embeddings are fed into the Transformer, which synthesizes them into a knowledge-rich vector of length 128. This synthesized vector encapsulates the collective information from multiple perspectives, providing a robust representation of the individual's gait. In the inference phase, the input is simplified to the embedding vector of any single camera angle, produced by its corresponding enriched CNN model. The Transformer then processes this input to generate a knowledge-enriched embedding vector. This final vector is highly discriminative and is subsequently used for identity recognition through cosine similarity comparison. By leveraging the Transformer's ability to integrate information across perspectives, the system achieves more reliable and precise identification, ultimately advancing the performance of gait recognition beyond the current state-of-the-art.

In this transformer model, we use 12 encoder layers with 4 attention headers per encoder layer. The dropout value is 0.25. We obtained this dropout value during testing, and it helps the model achieve optimal performance.

In natural language processing problems, transformer models are used to analyze which author wrote a piece of text. The model's tokens are words, phrases, or parts of words. In the problem we are considering, the sample texts are vectors composed of embedded vectors of images from each perspective of the same person. The task of the transformer model is to generate an embedded vector that has low similarity to a vector generated from an image or set of images of another person, and high similarity to a vector generated from an image or set of images of the same person. This vector represents the unique characteristics of a person from multiple perspectives at the same time by modeling the relationships between the data pieces in the input descriptive data vector generated by CNN descriptors.

In this chapter, the proposed methodology is presented in a systematic and comprehensive manner, emphasizing the integration of a Convolutional Neural Network (CNN) with a Transformer model to construct a robust framework for knowledge extraction and synthesis. Specifically, the CNN is employed in the initial stage to perform feature extraction, thereby capturing fine-grained local patterns and structural representations within the dataset. Following this, the Transformer model is introduced to facilitate multi-view knowledge synthesis, enabling the system to model long-range dependencies and contextual relationships that extend beyond the scope of localized features. Furthermore, the chapter elaborates on the central hypothesis underpinning this approach, which posits that the combination of CNN's capacity for extracting rich and detailed features with the Transformer's ability to synthesize diverse perspectives and semantic contexts will yield a more comprehensive and insightful understanding of the dataset. In this regard, the methodology not only leverages the complementary strengths of both architectures but also establishes a theoretical foundation for why their integration is expected to outperform traditional single-model approaches. Therefore, the chapter situates the proposed framework within a broader research context, demonstrating how the methodological design and guiding hypothesis collectively contribute to advancing the effectiveness of knowledge extraction in complex datasets.

4.1 Dataset

4.1.1 Data Source

The OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) is meant to aid research efforts in the general area of developing, testing and evaluating algorithms for cross-view gait recognition. The Institute of Scientific and Industrial Research (ISIR), Osaka University (OU) has copyright in the collection of gait video and associated data and serves as a distributor of the OU-ISIR Gait Database.

The data was collected in conjunction with an experience-based long-run exhibition of video-based gait analysis at a science museum, Figure 4.1. The approved informed consent was obtained from all the subjects in this dataset. The dataset consists of 10,307 subjects (5,114 males and 5,193 females with various ages, ranging from 2 to 87 years) from 14 view angles, Figure 4.2, ranging 0°-90°, 180°-270°. Gait images of 1,280 x 980 pixels at 25 fps are captured by seven network cameras (Cam1-7) placed at intervals of 15-deg azimuth angles along a quarter of a circle whose center coincides with the center of the walking course. Its radius is approximately 8 m and height is approximately 5 m.

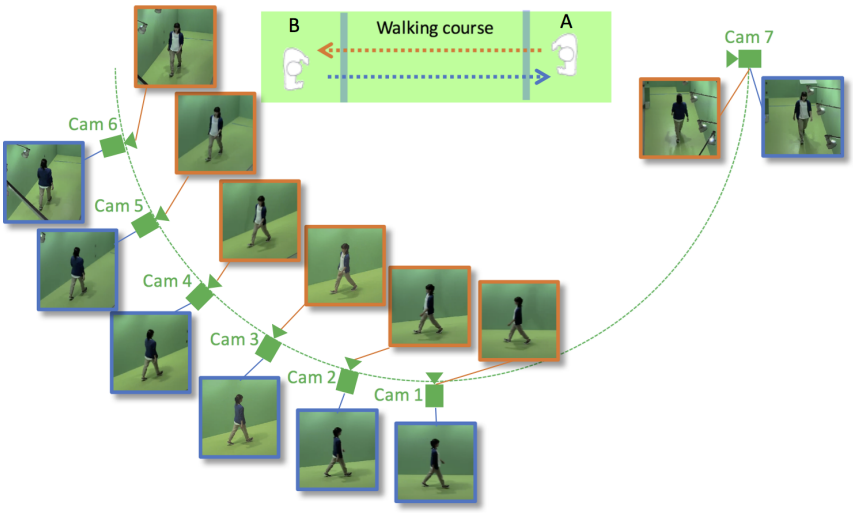

Figure 4.1: The subject repeat forward (A to B) and backward (B to A)

The entire data set was divided into two disjoint subsets

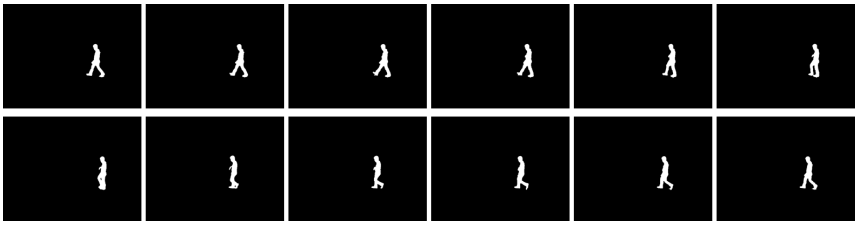

Figure 4.2: Examples of silhouette sequence (view angle = 90°)

4.2 Evaluation Parameters

Our research topic focuses on generating the most discriminative embedding vector for each identifier, Figure 4.3. Therefore, the primary metric for evaluating the performance of the MVDL model is the Rank-1 Accuracy.

$$Accuracy = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Prediction}} \times 100\%$$

- **Rank-1 Accuracy**: This metric represents the probability that the correct identity is found at the first

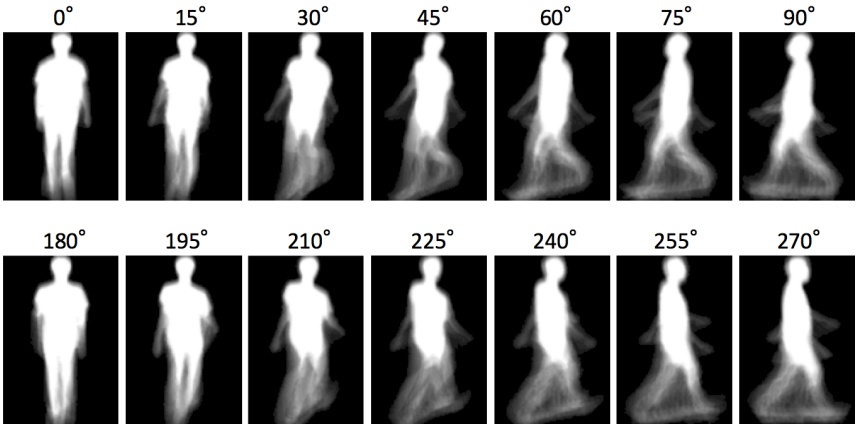

Figure 4.3: Examples of size-normalized GEI for each view angles

position (highest similarity score) when comparing the query embedding vector against the gallery (stored identity) embedding vectors.

- **Cosine Similarity**: The comparison function used for matching the enriched query embedding with the gallery embeddings is the Cosine Similarity

4.3 Experiment Setting

1. Chosen baselines:

Table 4.1: Baseline table

Name	Best Accuracy
GaitSFF: improving gait recognition performance based on selective feature fusion in video surveillance	90.06%
Cross-View Gait Recognition by Discriminative Feature Learning	95.40%
Learning Visual Prompt for Gait Recognition	93.20%

2. Reason of baseline selection
   (a) They use the OU-MVLP dataset
   (b) All papers are published in the recent year.
3. We conduct two inference experiments on the trained model. Experiment 1 infers based on 1 camera angle. Experiment 2 infers based on 2 camera angles.
4. Experimental steps:
   (a) Train / Test setting
      - Training dataset contains IDs which are less than "08000".
      - Validating dataset contains IDs which are equal and greater than "08000".
      - Each angle gets about 10 training session, each session gets 60 epochs.
      - Each training session goes along with one testing session.
   (b) Passing an image of a camera angle through a CNN model
   (c) Feeding the vector obtained from the CNN model into the Transformer model to get the embedding value containing the enriched data.
   (d) Applying cosine distance to embedding vectors to determine object identity

4.4 Reasoning on two views (Multi-View Synthesis)

Table 4.2: The performance comparisons on OUMVLP with two-view accuracy

	000	015	030	045	060	075	090
000	81.16	95.1	95.37	95.19	95.19	96.68	96.86
015	95.1	88.32	95.78	96.34	96.34	97.72	97.51
030	95.78	95.63	91.68	96.42	96.42	97.45	97.63
045	95.37	96.28	95.77	94.78	97.29	**97.85**	**98.48**
060	95.19	96.34	96.42	97.29	94.27	97.17	97.86
075	96.68	**97.72**	97.45	97.85	97.17	95.58	97.61
090	**96.86**	97.51	**97.63**	**98.48**	**97.86**	97.61	96.75
Mean	93.73	95.27	95.73	96.62	96.36	97.15	**97.53**

4.5 Reasoning on one view (Single-View Inference)

Table 4.3: The performance comparisons on OUMVLP, excluding the identical-views cases.

Method	Probe view							Mean
	0°	15°	30°	45°	60°	75°	90°	
GaitSet [1]	81.30	88.60	90.20	90.70	88.60	89.10	88.30	88.10
GaitGCI [6]	91.20	92.30	92.60	92.70	93.00	92.30	92.10	92.31
BiFusion [5]	86.17	90.60	91.28	91.56	90.88	90.78	90.48	90.25
VPNet-M[4]	**91.90**	**93.00**	**92.40**	92.70	93.20	92.50	92.30	**92.57**
GaitGL-HBS	84.70	90.20	91.40	91.70	90.90	91.00	90.50	90.05
STC-Att [2]	91.30	92.40	91.20	89.90	92.10	90.90	90.20	91.14
Ours (without MVDL boosting)	76.01	84.04	85.44	90.27	91.06	92.60	93.18	87.51
MVDL (ours)	83.21	89.55	91.68	**94.78**	**94.44**	**95.62**	**96.75**	92.29

This section presents the final results obtained after feeding the enriched single-view embedding vectors into the Transformer-based multi-perspective model for synthesis. The resulting synthesized embedding vector is then used for the final identity comparison.

The table below compares the performance of the MVDL solution against published state-of-the-art methods in gait recognition.

Table 4.4: Performance of the MVDL solution

Model	Year	Rank-1 Accuracy
Cross-View Gait Recognition by Discriminative Feature Learning	2022	95.40%
Combining the Silhouette and Skeleton Data for Gait Recognition	2023	92.50%
Gait Recognition Using 3-D Human Body Shape Inference	2023	91.40%
Learning Visual Prompt for Gait Recognition	2024	93.20%
Learning rich features for gait recognition by integrating skeletons and silhouettes	2024	91.28%
GaitGCI: Generative Counterfactual Intervention for Gait Recognition	2024	93.00%
MVDL (Proposed)	**2025**	**96.75%**

Analysis: The Rank-1 accuracy achieved by the proposed MVDL model is 96.75%, which surpasses the performance of all comparative state-of-the-art models referenced in this study. This result demonstrates the efficacy of the complete two-stage methodology: The initial embedding enrichment (knowledge propagation) provides highly informative input features. The multi-perspective Transformer synthesis effectively integrates the multi-view knowledge, resulting in a single, maximally discriminative embedding vector that significantly outperforms existing published methods. The achieved accuracy successfully addresses the primary objective of this research by pushing the performance beyond the 96% threshold previously established as a challenge in the literature.

Table 4.5: The Rank accuracy per view.

Rank	Angle						
	0	15	30	45	60	75	90
1	0.8321	0.8955	0.9169	0.9563	0.9444	0.9512	0.9675
2	0.8842	0.9241	0.9458	0.9721	0.9616	0.967	0.9783
3	0.9047	0.9358	0.9559	0.9791	0.9714	0.9758	0.9838
4	0.9163	0.9419	0.9642	0.9828	0.9748	0.98	0.9856
5	0.928	0.9465	0.9697	0.9861	0.9782	0.9819	0.9878
6	0.9346	0.9526	0.9711	0.9879	0.9794	0.9851	0.9892
7	0.9391	0.9562	0.9729	0.9898	0.9811	0.9874	0.9905
8	0.9424	0.9577	0.9775	0.9902	0.9828	0.9888	0.9905
9	0.9463	0.9592	0.9784	0.9907	0.9834	0.9893	0.9905
10	0.9501	0.9623	0.9789	0.9907	0.984	0.9893	0.9919
11	0.9512	0.9623	0.9807	0.9907	0.984	0.9898	0.9919
12	0.9524	0.9643	0.9816	0.9916	0.984	0.9898	0.9928
13	0.9546	0.9659	0.9816	0.9916	0.9851	0.9907	0.9932
14	0.9557	0.9664	0.9816	0.9916	0.9851	0.9912	0.9941
15	0.9579	0.9674	0.9816	0.9916	0.9863	0.9912	0.995
16	0.9601	0.9679	0.9826	0.9916	0.9863	0.9912	0.995
17	0.9612	0.9689	0.9826	0.9921	0.9874	0.9912	0.995
18	0.9623	0.9694	0.9839	0.9921	0.9874	0.9912	0.995
19	0.9662	0.9699	0.9844	0.9921	0.988	0.9912	0.9955
20	0.9673	0.9704	0.9848	0.9921	0.9891	0.9912	0.9959

We note that all published researches work on one view. We propose to use more than one view while inferencing in order to increase the accuracy rate. Our "MVDL" model is flexible; it could accept input from 1 view or multiple views.

Our best accuracy on one-view "96.75%" is better than the best accuracy of other paper and our best accuracy on two-view "98.48%" is better than our best accuracy on one-view "96.75%".

Figure 4.4: Rank accuracy of all view points

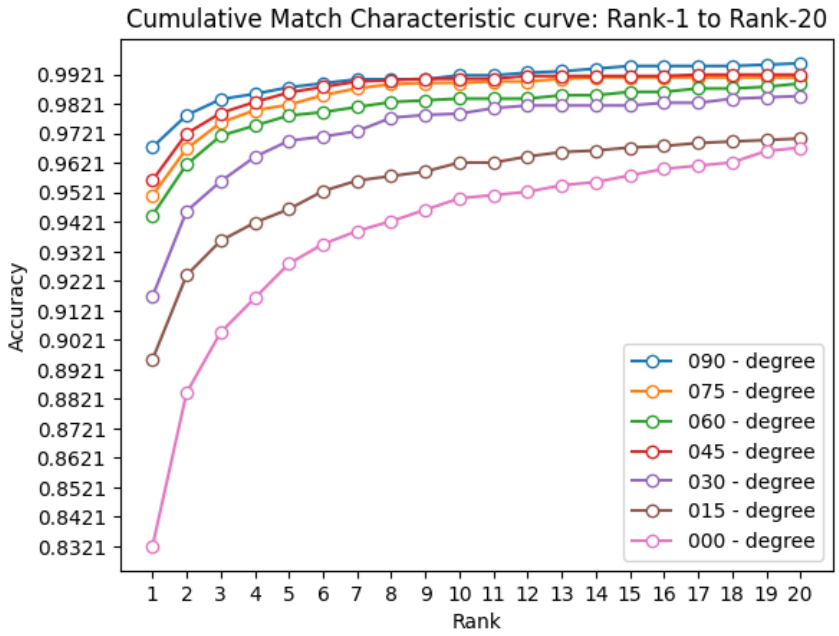

Table 4.7: Descriptor FLOP.

Component	Estimated FLOP	% of Total
Convolutional Layers	501.74 Million	93.8%
Fully Connected Layers	33.10 Million	6.2%
Total	534.84 Million	100%

4.5.1 Other measurements

The overall training time is approximately 2 months. Inference time of a GEI is about 0.001 seconds which includes the amount of time to turn GEI into an embedded vector by the descriptor, the amount of time to turn the described vector into a new reasonable vector and the amount of time to compare the new vector with others in the database of embedded vectors.

The chapter presents a comprehensive overview of all the results obtained from this research, demonstrating that the proposed approach consistently outperforms other existing methods across the evaluated tasks. This superiority is not only conveyed through detailed numerical comparisons but is also illustrated in a visual manner, with the inclusion of charts depicting the Cumulative Match Characteristic (CMC) Curves. These curves provide a clear and intuitive representation of the performance differences, highlighting how the proposed model achieves higher accuracy and reliability at various ranks compared to alternative techniques. By analyzing the CMC Curves, readers can observe the distinct advantages of the methodology, as the results reveal a significant improvement in matching effectiveness and knowledge extraction. Consequently, the chapter establishes strong empirical evidence that validates the effectiveness of the proposed framework, reinforcing the claim that it surpasses conventional approaches and offering a transparent visualization of its impact through the presented charts.

Table 4.8: Reasoner FLOP with vector sequence length of 128.

Module	Estimated FLOP	% of Total
Multi-Head Attention	184.55 Million	26.2%
Feed-Forward Network	520.09 Million	73.8%
Total Model	704.64 Million	100%

CHAPTER 5. CONCLUSIONS

5.1 Summary

This thesis has successfully proposed and implemented a novel two-stage methodology known as Multi-view Gait Recognition with Deep Learning (MVDL). The central aim of this approach is to improve both the accuracy and robustness of cross-view gait recognition systems, with particular emphasis on enhancing single-view inference. By combining innovative mechanisms inspired by Graph Neural Networks (GNNs) with the synthesis capabilities of Transformer models, the methodology establishes a comprehensive framework that addresses key limitations in existing research and achieves better performance on benchmark datasets.

The first major achievement of this work is the introduction of a GNN-inspired knowledge propagation mechanism. This technique was applied to bidirectionally enrich the feature embeddings generated by view-specific Convolutional Neural Networks (CNNs). In conventional single-view training, models often suffer from data scarcity and overfitting, which limit their ability to generalize across different perspectives. By enabling embeddings from one camera angle to exchange information with those from neighboring angles, the knowledge propagation mechanism effectively mitigates these issues. As a result, the enriched embeddings become more robust and discriminative, which lays the strong foundation for accurate gait recognition across multiple viewpoints.

The second achievement is the utilization of a Transformer-based model to synthesize the enriched embedding vectors obtained from multiple camera angles. This synthesis step is critical in creating a maximally discriminative representation of gait identity. By leveraging the Transformer's ability to capture long-range dependencies and contextual relationships, the model integrates information from angles such as 45°, 60°, 75°, and 90° into a unified representation. The complete MVDL pipeline achieved a final Rank-1 Accuracy of 96.75% on the benchmark GEI dataset. Furthermore, the numerical results demonstrated an average Rank-1 Accuracy increase of approximately 3.79% across the tested camera angles compared to baseline training, highlighting the effectiveness of the multi-perspective synthesis process.

The third achievement is the clear outperformance of state-of-the-art models. The accuracy of 96.75% achieved by MVDL surpasses the performance of all currently published gait recognition approaches, including the 95.4% accuracy reported by Cross-View Gait Recognition through Discriminative Feature Learning. This accomplishment successfully addresses the research objective of exceeding the long-standing 96% accuracy barrier, establishing MVDL as a new benchmark in the field.

In conclusion, the MVDL methodology provides a computationally efficient and highly accurate solution for robust gait recognition. By effectively leveraging multi-view information during training, the framework significantly enhances single-view inference capability. This dual-stage approach not only extend the technical boundaries of gait recognition but also demonstrates the potential of combining GNN-inspired enrichment with Transformer-based synthesis to achieve superior results in complex recognition tasks.

5.2 Suggestion for Future Works

The findings of this research open several promising directions for future investigation, each aimed at further enhancing the performance, efficiency, and applicability of gait recognition systems. One important avenue is the integration of real-time data streams. At present, the methodology relies on static Gait Energy Image (GEI) inputs, which limits its adaptability to dynamic environments. Future work should focus on extending the knowledge propagation and synthesis mechanisms to process raw sequential video frames or skeleton data in real time. Achieving this would require the incorporation of Recurrent Neural Networks (RNNs) or specialized spatio-temporal graph models capable of capturing temporal dependencies and motion dynamics. Such an extension would significantly improve the practicality of gait recognition systems in surveillance and monitoring applications.

Another promising direction is the development of adaptive view selection mechanisms. The current system synthesizes information from a fixed set of four camera angles, which may not always be optimal under varying environmental conditions. A more advanced approach would allow the model to dynamically select the two or three most informative camera angles based on contextual factors such as occlusions, lighting variations, or crowd density. This adaptive selection could improve computational efficiency by reducing redundant processing while maintaining or even enhancing recognition accuracy.

Finally, hardware optimization represents a crucial step toward real-world application. Investigating techniques such as pruning and quantization for both the enriched CNN models and the Transformer-based synthesis model could enable deployment on resource-constrained edge devices, including surveillance cameras and embedded systems. By reducing computational overhead without significantly compromising accuracy, these optimizations would pave the way for scalable and cost-effective implementations of gait recognition in security systems and smart city infrastructures.

Together, these future directions highlight the potential to transform the current methodology into a more versatile, efficient, and widely applicable solution for gait recognition in dynamic and resource-limited environments.

REFERENCE

[1] Chao, Hanqing, Wang **andothers**, "Gaitset: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **journvol** 44, **number** 7, **pages** 3467–3478, 2022. DOI: 10.1109/TPAMI.2021.3057879.

[2] W. Likai, H. Ruize **and** F. Wei, "Combining the silhouette and skeleton data for gait recognition," **in***ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2023, **pages** 1–5. DOI: 10.1109/ICASSP49357.2023.10096986.

[3] Z. Haidong, Z. Zhaoheng **and** N. Ram, "Gait recognition using 3-d human body shape inference," **in***Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* **january** 2023, **pages** 909–918.

[4] M. Kang, F. Ying, C. Chunshui, H. Saihui, H. Yongzhen **and** Z. Dezhi, "Learning visual prompt for gait recognition," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **june** 2024, **pages** 593–603.

[5] P. Yunjie, M. Kang, Z. Yang **and** H. Zhiqiang, "Learning rich features for gait recognition by integrating skeletons and silhouettes," *Multimedia Tools and Applications*, **journvol** 83, **number** 3, **pages** 7273–7294, **january** 2024. DOI: 10.1007/s11042-023-15483-x. **url**: https://doi.org/10.1007/s11042-023-15483-x.

[6] D. Huanzhang, Z. Pengyi, S. Wei, Y. Yunlong, L. Yining **and** L. Xi, "Gaitgci: Generative counterfactual intervention for gait recognition," **in***2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2023, **pages** 5578–5588. DOI: 10.1109/CVPR52729.2023.00540.

[7] R. C. Manske, *Fundamental orthopedic management for the physical therapist assistant*, en, 5 **edition**. Philadelphia, PA: Elsevier - Health Sciences Division, **october** 2021.

[8] Abu-Faraj, Z. O, Harris **andothers**, "Human gait and clinical movement analysis," **in***Wiley Encyclopedia of Electrical and Electronics Engineering* Hoboken, NJ, USA: John Wiley & Sons, Inc., **december** 2015, **pages** 1–34.

[9] S. Di, Zhang, Wuxiang **andothers**, "Parametric generation of three-dimensional gait for robot-assisted rehabilitation," *Biology Open*, **journvol** 9, bio.047332, **january** 2020. DOI: 10.1242/bio.047332.

[10] D. J. Magee, *Orthopedic physical assessment - E-book* (Musculoskeletal Rehabilitation Series), en, 5 **edition**. Saunders, **december** 2007.

[11] S. Li **and** A. Jain, *Encyclopedia of Biometrics: I - Z.* (Encyclopedia of Biometrics 2). Springer, 2009, ISBN: 9780387730028. **url**: https://books.google.com.vn/books?id=0bQbOYVULQcC.

[12] W. An, S. Yu, Y. Makihara **andothers**, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. on Biometrics, Behavior, and Identity Science*, 2020 (Accepted).

[13] J. Han **and** B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **journvol** 28, **number** 2, **pages** 316–322, 2006. DOI: 10.1109/TPAMI.2006.38.

[14] L. Wang, T. Tan, H. Ning **and** W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **journvol** 25, **number** 12, **pages** 1505–1518, 2003. DOI: 10.1109/TPAMI.2003.1251144.

[15] C. BenAbdelkader, L. S. Davis **and** R. Cutler, "Stride and cadence as a biometric in automatic person identification and verification," *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, **pages** 372–377, 2002. **url**: https://api.semanticscholar.org/CorpusID:16999156.

[16] Z. Liu **and** S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," **in***Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* **volume** 4, 2004, 211–214 Vol.4. DOI: 10.1109/ICPR.2004.1333741.

[17] D. Cunado, M. S. Nixon **and** J. N. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Comput. Vis. Image Underst.*, **jourvol** 90, **pages** 1–41, 2003. **url**: https://api.semanticscholar.org/CorpusID:1622910.

[18] C. Yam, M. S. Nixon **and** J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, **jourvol** 37, **number** 5, **pages** 1057–1072, 2004, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2003.09.012. **url**: https://www.sciencedirect.com/science/article/pii/S0031320303003996.

[19] K. A., S. A., R. A.N. **andothers**, "Identification of humans using gait," *IEEE Transactions on Image Processing*, **jourvol** 13, **number** 9, **pages** 1163–1173, 2004. DOI: 10.1109/TIP.2004.832865.

[20] T. Akira, M. Yasushi **and** Y. Yasushi, "Silhouette transformation based on walking speed for gait identification," **in***2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2010, **pages** 717–722. DOI: 10.1109/CVPR.2010.5540144.

[21] B. Khalid, X. Tao **and** G. Shaogang, "Gait recognition using gait entropy image," **in***3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)* 2009, **pages** 1–6. DOI: 10.1049/ic.2009.0230.

[22] H. Martin, B. Sebastian **and** R. Gerhard, "2.5d gait biometrics using the depth gradient histogram energy image," **in***2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* 2012, **pages** 399–403. DOI: 10.1109/BTAS.2012.6374606.

[23] W. Chen, Z. Junping, P. Jian, Y. Xiaoru **and** W. Liang, "Chrono-gait image: A novel temporal template for gait recognition," **in***Computer Vision – ECCV 2010* D. Kostas, M. Petros **and** P. Nikos, **editors**, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, **pages** 257–270, ISBN: 978-3-642-15549-9.

[24] W. Kusakunniran, "Recognizing gaits on spatio-temporal feature domain," *IEEE Transactions on Information Forensics and Security*, **jourvol** 9, **number** 9, **pages** 1416–1423, 2014. DOI: 10.1109/TIFS.2014.2336379.

[25] W. Kusakunniran, Q. Wu, J. Zhang, H. Li **and** L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Transactions on Image Processing*, **jourvol** 23, **number** 2, **pages** 696–709, 2014. DOI: 10.1109/TIP.2013.2294552.

[26] H. Aggarwal **and** D. K. Vishwakarma, "Covariate conscious approach for gait recognition based upon zernike moment invariants," *Preprint (arXiv)*, 2016, arXiv:1611.06683.

[27] C. Xu, Y. Makihara, X. Li, Y. Yagi **and** J. Lu, "Speed-invariant gait recognition using single-support gait energy image," *Multimedia Tools and Applications*, **jourvol** 78, **number** 18, **pages** 26 509–26 536, 2019. DOI: 10.1007/s11042-019-7712-3.

[28] F. M. Castro, M. J. Marin-Jimenez, N. Guil, S. Lopez-Tapia **and** N. Perez de la Blanca, "Evaluation of cnn architectures for gait recognition based on optical flow maps," **in***2017 International Conference of the Biometrics Special Interest Group (BIOSIG)* 2017, **pages** 1–5. DOI: 10.23919/BIOSIG.2017.8053503.

[29] Q. Zou, Y. Wang, Q. Wang, Y. Zhao **and** Q. Li, "Deep learning-based gait recognition using smartphones in the wild," *IEEE Transactions on Information Forensics and Security*, **jourvol** 15, **pages** 3197–3212, 2020. DOI: 10.1109/TIFS.2020.2985628.

[30] Z. Mu, F. M. Castro, M. J. Marín-Jiménez, N. Guil, Y.-R. Li **and** S. Yu, "Ilgaco: Incremental learning of gait covariate factors," **in***2020 IEEE International Joint Conference on Biometrics (IJCB)* 2020, **pages** 1–8. DOI: 10.1109/IJCB48548.2020.9304857.

[31] S. G. Işık **and** H. K. Ekenel, "Deep convolutional feature-based gait recognition using silhouettes and rgb images," **in***2021 6th International Conference on Computer Science and Engineering (UBMK)* 2021, **pages** 336–341. DOI: 10.1109/UBMK52708.2021.9559026.

[32] C Xu, Y Makihara **and** J Lu, "Applications: Real-time gait-based age estimation and gender classification from a single image," **in***Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision 2021 (WACV 2021)* **january** 2021, **pages** 1–11.

[33] T. M. A. Zulcaffle, F. Kurugollu, D. Crookes, A. Bouridane **and** M. Farid, "Frontal view gait recognition with fusion of depth features from a time of flight camera," *IEEE Transactions on Information Forensics and Security*, **jourvol** 14, **number** 4, **pages** 1067–1082, 2019. DOI: 10.1109/TIFS.2018.2870594.

[34] H. S. M. Gowri **and** M. Okade, "Skeleton based frontal gait recognition utilizing fourier descriptors," **in***TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)* 2019, **pages** 452–456. DOI: 10.1109/TENCON.2019.8929373.

[35] M. Deng, Z. Fan, P. Lin **and** X. Feng, "Human gait recognition based on frontal-view sequences using gait dynamics and deep learning," *IEEE Transactions on Multimedia*, **jourvol** 26, **pages** 117–126, 2024. DOI: 10.1109/TMM.2023.3262131.

[36] S. Tong, Y. Fu, X. Yue **and** H. Ling, "Multi-view gait recognition based on a spatial-temporal deep neural network," *IEEE Access*, **jourvol** 6, **pages** 57 583–57 596, 2018. DOI: 10.1109/ACCESS.2018.2874073.

[37] M. H. Khan, M. S. Farid, M. Zahoor **and** M. Grzegorzek, "Cross- view gait recognition using non-linear view transformations of spatiotemporal features," **in***2018 25th IEEE International Conference on Image Processing (ICIP)* 2018, **pages** 773–777. DOI: 10.1109/ICIP.2018.8451629.

[38] Y. Zhang, Y. Huang, S. Yu **and** L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Transactions on Image Processing*, **jourvol** 29, **pages** 1001–1015, 2020. DOI: 10.1109/TIP.2019.2926208.

[39] Q. Hong, Z. Wang, J. Chen **and** B. Huang, "Cross-view gait recognition based on feature fusion," **in***2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)* 2021, **pages** 640–646. DOI: 10.1109/ICTAI52525.2021.00102.

[40] S. Zhang, Y. Wang **and** A. Li, "Cross-view gait recognition with deep universal linear embeddings," **in***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2021, **pages** 9091–9100. DOI: 10.1109/CVPR46437.2021.00898.

[41] J. Wang, J. Guo **and** Z. Xu, "Cross-view gait recognition model combining multi-scale feature residual structure and self-attention mechanism," *IEEE Access*, **jourvol** 11, **pages** 127 769–127 782, 2023. DOI: 10.1109/ACCESS.2023.3331395.

[42] M. A. A. Aljazaerly, Y. Makihara, D. Muramatsu **and** Y. Yagi, "Batch hard contrastive loss and its application to cross-view gait recognition," *IEEE Access*, **jourvol** 11, **pages** 31 177–31 187, 2023. DOI: 10.1109/ACCESS.2023.3262271.

[43] T. Huang, X. Ben, C. Gong, W. Xu, Q. Wu **and** H. Zhou, "Gaitdan: Cross-view gait recognition via adversarial domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, **jourvol** 34, **number** 9, **pages** 8026–8040, 2024. DOI: 10.1109/TCSVT.2024.3384308.

[44] M. Saad Shakeel, K. Liu, X. Liao **and** W. Kang, "Tag: A temporal attentive gait network for cross-view gait recognition," *IEEE Transactions on Instrumentation and Measurement*, **jourvol** 74, **pages** 1–14, 2025. DOI: 10.1109/TIM.2024.3497164.

[45] S. Zhang **and** C. Liu, "Cross-view gait recognition via view information elimination mechanism," *IEEE Access*, **jourvol** 12, **pages** 182 455–182 468, 2024. DOI: 10.1109/ACCESS.2024.3510718.

[46] W. Zhai, H. Li, C. Zheng **and** X. Xing, "Multi-view gait recognition with joint local multi-scale and global contextual spatio-temporal features," *IEEE Transactions on Circuits and Systems for Video Technology*, **jourvol** 35, **number** 2, **pages** 1123–1135, 2025. DOI: 10.1109/TCSVT.2024.3476384.

[47] T. Wolf, M. Babaee **and** G. Rigoll, "Multi-view gait recognition using 3d convolutional neural networks," **in***2016 IEEE International Conference on Image Processing (ICIP)* 2016, **pages** 4165–4169. DOI: 10.1109/ICIP.2016.7533144.