Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.075704

ARTICLE

GaitMAFF: Adaptive Multi-Modal Fusion of Skeleton Maps and Silhouettes for Robust Gait Recognition in Complex Scenarios

Zhongbin Luo[1,2], **Zhaoyang Guan**[3], **Wenxing You**[2], **Yunteng Wang**[2] **and Yanqiu Bi**[4,5,*]

[1]College of Computer Science, Chongqing University, Chongqing, 400044, China
[2]School of Traffic and Transportation, Chongqing Jiaotong University, Chongqing, 400074, China
[3]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208, USA
[4]School of Civil Engineering, Chongqing Jiaotong University, Chongqing, 400074, China
[5]National & Local Joint Engineering Research Center of Transportation Civil Engineering Materials, Chongqing Jiaotong University, Chongqing, 400074, China

*Corresponding Author: Yanqiu Bi. Email: biyanqiu@cqjtu.edu.cn

Received: 06 November 2025; Accepted: 15 December 2025; Published: 12 March 2026

ABSTRACT: Gait recognition is a key biometric for long-distance identification, yet its performance is severely degraded by real-world challenges such as varying clothing, carrying conditions, and changing viewpoints. While combining silhouette and skeleton data is a promising direction, effectively fusing these heterogeneous modalities and adaptively weighting their contributions in response to diverse conditions remains a central problem. This paper introduces GaitMAFF, a novel Multi-modal Adaptive Feature Fusion Network, to address this challenge. Our approach first transforms discrete skeleton joints into a dense Skeleton Map representation to align with silhouettes, then employs an attention-based module to dynamically learn the fusion weights between the two modalities. These fused features are processed by a powerful spatio-temporal backbone with Weighted Global-Local Feature Fusion Modules (WFFM) to learn a discriminative representation. Extensive experiments on the challenging CCPG and Gait3D datasets show that GaitMAFF achieves state-of-the-art performance, with an average Rank-1 accuracy of 84.6% on CCPG and 58.7% on Gait3D. These results demonstrate that our adaptive fusion strategy effectively integrates complementary multi-modal information, significantly enhancing gait recognition robustness and accuracy in complex scenes and providing a practical solution for real-world applications.

KEYWORDS: Gait recognition; multi-modal fusion; adaptive feature fusion; skeleton map; silhouette

1 Introduction

Gait, the manner of human walking, is a unique biometric characteristic that can be captured at a distance without subject cooperation, offering significant advantages in applications such as public security, intelligent surveillance, and healthcare [1]. Unlike other biometrics like fingerprints or iris, gait is difficult to disguise and can be recognized even from low-resolution video footage, making it highly suitable for a wide array of real-world scenarios.

Despite its potential, gait recognition in complex, unconstrained environments remains a significant challenge. The performance of existing methods often degrades drastically when faced with real-world complexities such as variations in clothing, carrying conditions, diverse walking directions, partial occlusions, and fluctuating illumination. Silhouette-based gait recognition methods, which have demonstrated considerable success in controlled laboratory settings by extracting features from 2D body shapes [2,3],

are particularly vulnerable to these appearance-related interferences, leading to a loss of robustness in practical applications.

To mitigate the impact of appearance variations, skeleton-based gait recognition has emerged as a promising alternative. By representing gait through the spatio-temporal patterns of body keypoints, these methods can effectively handle variations in clothing and some forms of occlusion [4,5]. However, skeleton data inherently lacks fine-grained appearance information, limiting its discriminative power, and its performance is often contingent on the accuracy of the underlying pose estimation algorithms, which can also be compromised in complex scenes.

Recognizing the complementary nature of silhouette and skeleton modalities, multi-modal gait recognition has garnered increasing attention. The rich appearance details from silhouettes and the structural motion cues from skeletons, if fused effectively, can lead to a more comprehensive and robust gait representation [6–9]. Nevertheless, existing multi-modal fusion techniques often face challenges such as semantic misalignment between different data types (e.g., directly fusing 2D skeleton coordinates with silhouette images), inadequate exploration of inter-modal correlations, and suboptimal fusion strategies that may not adaptively weigh the contributions of each modality under varying conditions.

To address these limitations, this paper introduces GaitMAFF (Multi-modal Adaptive Feature Fusion Network), a novel framework designed to enhance gait recognition accuracy and robustness in complex scenarios by effectively integrating silhouette and skeleton information. The core contributions of GaitMAFF are threefold:

1. **A novel Skeleton Map representation:** We transform discrete 2D skeleton coordinates into a dense, image-like heatmap representation. This Skeleton Map preserves the structural and motion patterns of gait while facilitating better spatial and semantic alignment with silhouette images, thereby reducing noise and improving the efficacy of subsequent feature fusion.
2. **An adaptive modality feature fusion network:** We propose an attention-driven fusion module that dynamically learns to weigh the importance of silhouette and skeleton features. This adaptive mechanism allows GaitMAFF to effectively exploit the complementary strengths of each modality and mitigate their respective weaknesses, leading to a more discriminative fused representation.
3. **A robust spatio-temporal feature learning backbone:** The fused multi-modal features are processed by a powerful backbone network that incorporates effective techniques for global-local feature integration (such as the Weighted Global-Local Feature Fusion Module, WFFM) and potentially multi-scale temporal aggregation, enabling the extraction of fine-grained and discriminative spatio-temporal gait patterns.

Extensive experiments conducted on challenging real-world gait datasets, CCPG [10] and Gait3D [11], demonstrate that GaitMAFF significantly outperforms state-of-the-art single-modal and multi-modal methods, particularly under conditions with clothing variations, carrying items, and viewpoint changes.

The remainder of this paper is organized as follows: Section 2 reviews related work in gait recognition. Section 3 details the proposed GaitMAFF framework. Section 4 presents the experimental setup, results, and analysis. Finally, Section 5 concludes the paper and discusses future research directions.

2 Related Work

Gait recognition has been an active research area for decades, with numerous approaches proposed. These methods can be broadly categorized based on the input data representation: silhouette-based, skeleton-based, and more recently, multi-modal approaches that combine the strengths of different representations.

2.1 Silhouette-Based Gait Recognition

Silhouette-based methods have historically been the most prevalent due to the intuitive representation of human body shape during locomotion. Early works often relied on holistic representations of silhouettes, such as the Gait Energy Image (GEI) proposed by Han and Bhanu [12], which averages silhouettes over a gait cycle to create a compact template. Subsequent research focused on extracting more discriminative features from silhouette sequences. For instance, GaitSet [2] treated a gait sequence as an unordered set of silhouettes and employed a neural network to learn set-level features, demonstrating robustness to sequence length variations. GaitPart [3] introduced a micro-motion capture module and a focal convolution layer to learn fine-grained part-level features from horizontally divided silhouette parts. To address view variations, some methods explored view transformation models or view-invariant feature learning.

Later advancements, such as GLN [13] and GaitGL [14], focused on leveraging feature pyramids or combining global and local information to enhance feature representation. Our previous explorations also led to the development of modules like the Weighted Global-Local Feature Fusion Module (WFFM) and Attention-based Multiscale Temporal Aggregation (AMTA) [15], which aimed at refining silhouette-based feature extraction by adaptively weighting global and local cues and optimizing temporal modeling. While these methods achieve impressive performance in controlled or less challenging environments, their reliance on clean silhouette quality makes them susceptible to noise, occlusion, and significant appearance changes (e.g., varying clothes, carrying bags) prevalent in real-world scenarios like those found in the Gait3D [11] and CCPG [10] datasets. Furthermore, OpenGait [16] has provided a comprehensive benchmark for evaluating various gait recognition methods, promoting standardization in the field.

2.2 Skeleton-Based Gait Recognition

Skeleton-based gait recognition utilizes human pose information, typically a set of 2D or 3D joint coordinates, to model gait dynamics. This representation is inherently more robust to variations in clothing and texture compared to silhouettes. Early methods often involved handcrafted features derived from joint trajectories, angles, or limb lengths. With the advent of deep learning and reliable human pose estimation algorithms [17], learning-based approaches have become dominant.

Graph Convolutional Networks (GCNs) are a natural fit for modeling the non-Euclidean structure of skeleton data. GaitGraph [4] and subsequent works like GaitGCN++ [5] modeled the human skeleton as a graph and applied GCNs to learn spatio-temporal features. Other approaches have focused on designing specific network architectures to capture temporal dependencies and structural information from skeleton sequences. However, a primary limitation of skeleton-based methods is the lack of detailed appearance information, which can be crucial for distinguishing individuals, especially when motion patterns are similar. Furthermore, the accuracy of skeleton-based recognition is highly dependent on the precision of the upstream pose estimation, which can be unreliable in crowded or occluded scenes.

Moreover, recent studies have shown promising results in related fields by employing advanced techniques. For instance, Naik Bukht et al. [18] proposed a robust human interaction recognition method using Extended Kalman Filter, demonstrating the efficacy of temporal filtering in handling dynamic actions. Similarly, Chen et al. [19] introduced a two-stream spatio-temporal GCN-Transformer network for skeleton-based action recognition, effectively combining local and global features. Jiang et al. [20] also explored deep learning approaches for dynamic human activity recognition, highlighting the importance of capturing temporal dependencies. These works provide valuable insights into robust feature extraction and temporal modeling, which inspire our multi-modal fusion strategy.

2.3 Multi-Modal Gait Recognition

To leverage the complementary strengths of silhouettes and skeletons, multi-modal gait recognition has emerged as a promising research direction. Silhouettes provide rich appearance details, while skeletons offer robust structural and motion cues. The main challenge lies in effectively fusing these heterogeneous data sources.

Early fusion attempts often involved simple concatenation or element-wise addition of features extracted by separate branches. More recent works have explored sophisticated fusion mechanisms. For example, Fan et al. [21] proposed converting skeleton coordinates into heatmap-like skeleton maps to facilitate fusion with silhouette features at the image level and introduced a two-branch network. Li et al. [7] and Cui and Kang [8] utilized Transformer architectures to model cross-modal interactions between silhouette and skeleton features. MSGG [6] proposed a bi-modal network to mine discriminative gait patterns from skeletons and combine them with silhouette representations. GaitMA [9] introduced pose-guided feature fusion.

Despite these advancements, several challenges persist in multi-modal fusion. Firstly, achieving effective alignment and bridging the semantic gap between raw skeleton coordinates (or basic graph representations) and pixel-based silhouette images can be difficult, potentially leading to suboptimal fusion or noise introduction. Secondly, many existing fusion strategies are static or do not adequately adapt to the varying importance of each modality under different conditions (e.g., heavy occlusion might reduce the reliability of silhouettes, while poor pose estimation might affect skeleton quality). Recent works such as BigGait [22] have explored large-scale pre-training to learn more robust features, indicating a trend towards leveraging massive data for better generalization. Our proposed GaitMAFF addresses these issues by first transforming skeletons into an image-like Skeleton Map for better alignment with silhouettes, and then employing an adaptive attention-based fusion network to dynamically weigh and integrate features from both modalities, thereby capturing more comprehensive and robust gait representations for complex scenes.

Compared to existing methods such as SkeletonGait++ [21] and GaitRef [23], which primarily rely on static fusion or concatenation strategies, our GaitMAFF introduces a novel dynamic fusion mechanism. Specifically, while SkeletonGait++ utilizes a dual-branch structure with heatmap-like skeleton maps, it lacks an adaptive mechanism to weigh the importance of modalities based on their reliability in complex scenarios. Similarly, GaitRef focuses on refining skeleton sequences but does not fully exploit the complementary pixel-level details from silhouettes through an attention-driven approach. In contrast, GaitMAFF innovatively incorporates an attention-based module that learns to dynamically adjust the fusion weights, ensuring that the model can selectively emphasize the more reliable modality. Furthermore, unlike methods that use standard convolutional backbones, we integrate the Weighted Global-Local Feature Fusion Module (WFFM) to extract multi-granular spatio-temporal features, enhancing the discriminative power of the fused representation.

3 Methods

In this section, we present our proposed Multi-modal Adaptive Feature Fusion Network (GaitMAFF) for robust gait recognition in complex scenarios. GaitMAFF is designed to effectively fuse complementary information from silhouette and skeleton modalities, leveraging a novel skeleton map representation, an adaptive fusion mechanism, and a powerful spatio-temporal feature extraction backbone.

3.1 Overall Architecture

The overall architecture of GaitMAFF is illustrated in Fig. 1. The network takes a sequence of silhouette images $S_{sil} \in \mathbb{R}^{1\times T\times H\times W}$ and a corresponding sequence of generated skeleton maps $S_{ske} \in \mathbb{R}^{2\times T\times H\times W}$ as input, where T is the number of frames, and H and W are the height and width of each frame, respectively.

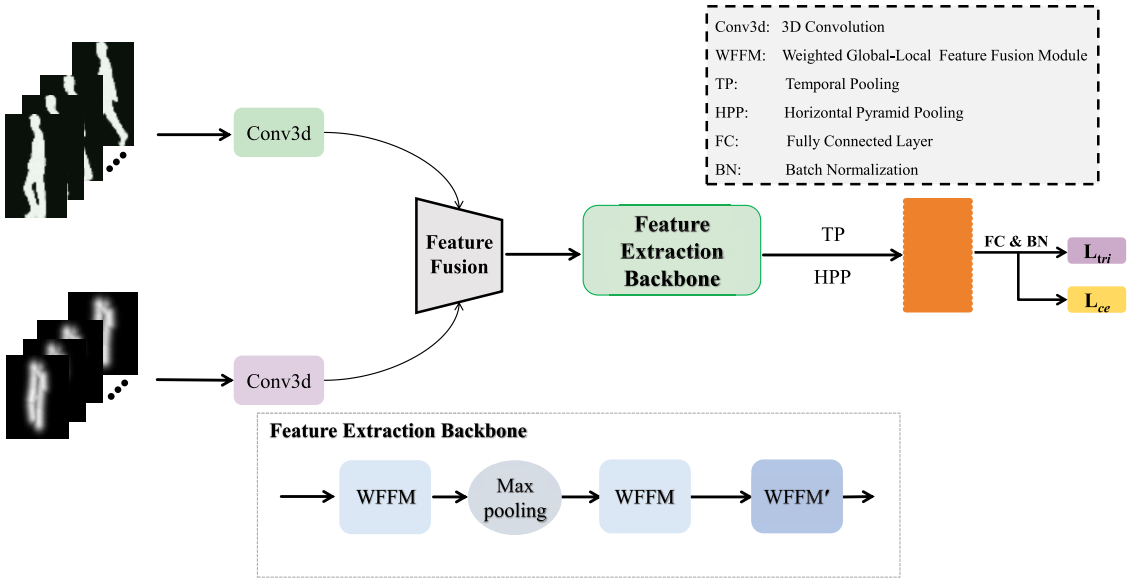

Figure 1: The overall architecture of the proposed GaitMAFF network. It comprises skeleton map generation, shallow feature extraction for each modality, an adaptive modality fusion network, a deep spatio-temporal feature extraction backbone featuring WFFM modules, and final feature mapping for recognition

The pipeline consists of the following main stages:

1. **Gait Skeleton Map Generation:** Raw 2D keypoint coordinates are transformed into a 2-channel image-like representation (Section 3.2).
2. **Shallow Feature Extraction:** Initial spatio-temporal features are extracted independently from S_{sil} and S_{ske} using separate 3D Convolutional (Conv3D) layers, yielding F_{sil} and F_{ske}.
3. **Adaptive Modality Feature Fusion:** F_{sil} and F_{ske} are fed into an adaptive fusion network to produce a unified multi-modal feature representation F_f (Section 3.3).
4. **Deep Spatio-temporal Feature Extraction Backbone:** The fused features F_f are processed by a deep backbone network, incorporating Weighted Global-Local Feature Fusion Modules (WFFM), to learn highly discriminative gait features (Section 3.4).
5. **Feature Mapping and Loss Function:** The extracted deep features are mapped to an embedding space using Temporal Pooling (TP) and Horizontal Pyramid Pooling (HPP), followed by a fully connected (FC) layer. The network is trained with a combined triplet loss and cross-entropy loss (Section 3.5).

3.2 Gait Skeleton Map Generation

To effectively fuse skeleton data with silhouette images and mitigate the semantic gap, we transform the raw 2D joint coordinates into a dense, image-like representation termed Skeleton Map, inspired by [21]. This process enhances the spatial representation of skeletal structures and facilitates pixel-level alignment with silhouettes.

As shown in Fig. 2, given a sequence of 2D human keypoint coordinates (x_k, y_k, c_k) for K joints (e.g., $K = 17$ from COCO format via HRNet [17]), where (x_k, y_k) is the spatial location of the k-th joint and c_k is its

confidence score, the Skeleton Map is generated as follows: First, the raw joint coordinates in each frame are normalized. The center of the body (e.g., midpoint of the hip joints) is translated to the center of a predefined canvas of size $R \times R$. The skeleton is then scaled so that its height is normalized to h.

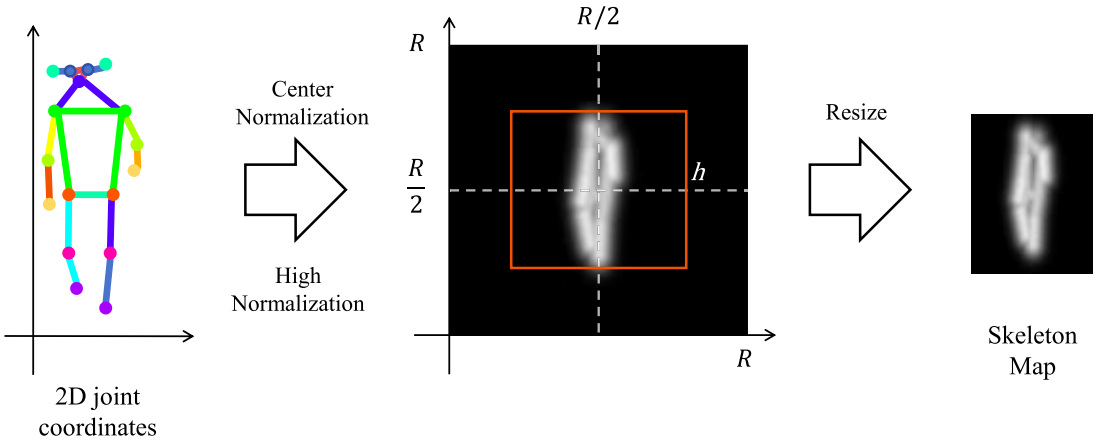

Figure 2: Gait skeleton map generation

The Skeleton Map consists of two channels: a Joint Map (J) and a Limb Map (L). The Joint Map J is generated by placing a 2D Gaussian kernel centered at each joint location and summing them up, weighted by their confidence scores:

$$J_{(u,v)} = \sum_{k=0}^{K} c_k \cdot \exp\left(-\frac{(u - x'_k)^2 + (v - y'_k)^2}{2\sigma^2}\right) \tag{1}$$

where (u, v) are pixel coordinates on the canvas, (x'_k, y'_k) are the normalized coordinates of the k-th joint, and σ is a hyperparameter controlling the variance of the Gaussian kernels.

The Limb Map L represents the connections between joints. For each predefined limb n connecting joints n_1 and n_2, a Gaussian distribution is rendered along the line segment connecting n_1 and n_2:

$$L_{(u,v)} = \sum_{n=1}^{N_{limbs}} \min(c_{n_1}, c_{n_2}) \cdot \exp\left(-\frac{D((u,v), \text{Limb}_n)^2}{2\sigma^2}\right) \tag{2}$$

where N_{limbs} is the total number of limbs, $D((u, v), \text{Limb}_n)$ is the minimum Euclidean distance from pixel (u, v) to the line segment representing Limb_n, and $\min(c_{n_1}, c_{n_2})$ uses the minimum confidence of the two endpoint joints to weight the limb's presence.

The resulting 2-channel (J and L) Skeleton Map of size $R \times R$ is then center-cropped and resized (e.g., to 64×44) to match the input dimensions of the silhouette stream for the subsequent network stages.

3.3 Adaptive Modality Feature Fusion Network

After extracting shallow features $F_{sil} \in \mathbb{R}^{C \times T \times H \times W}$ from silhouettes and $F_{ske} \in \mathbb{R}^{C \times T \times H \times W}$ from skeleton maps using initial Conv3D layers, these features are fed into our adaptive modality fusion network. This network (illustrated in Fig. 3) aims to effectively integrate the complementary information from both modalities while allowing the model to learn the relative importance of each.

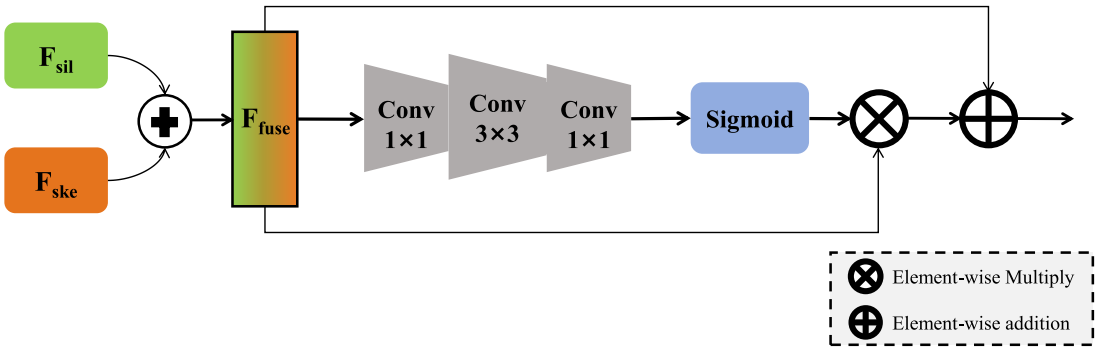

Figure 3: Detailed structure of the adaptive modality feature fusion network

First, the features from the two modalities are combined through element-wise addition to create an initial fused feature F_{add}:

$$F_{add} = F_{sil} + F_{ske} \tag{3}$$

This initial fusion preserves spatial correspondence and reduces dimensionality compared to concatenation.

To adaptively weigh the contributions of different spatial locations and channels in the fused features, we generate a spatial attention weight matrix w_f. F_{add} is passed through a three-stage 2D convolutional block: a 1×1 convolution for channel-wise feature recalibration and dimensionality reduction, followed by a 3×3 convolution to capture local contextual information, and another 1×1 convolution to restore the channel dimension. The output is then passed through a Sigmoid activation function:

$$w_f = \sigma_{sigmoid}(\text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(F_{add})))) \tag{4}$$

The Sigmoid function maps the weights to the range $(0,1)$, indicating the importance of each feature element.

The final fused feature F_f is obtained by applying these learned weights to F_{add} and adding a residual connection to facilitate gradient flow and enhance feature richness:

$$F_f = F_{add} + F_{add} \odot w_f \tag{5}$$

where $\odot$ denotes element-wise multiplication. This F_f serves as the input to the deep feature extraction backbone.

3.4 Deep Spatio-Temporal Feature Extraction Backbone

The adaptively fused multi-modal features F_f are then fed into a deep spatio-temporal feature extraction backbone to learn highly discriminative gait representations. This backbone typically consists of several Conv3D layers, interspersed with Weighted Global-Local Feature Fusion Modules (WFFM) and max-pooling layers for spatial downsampling. The specific architecture (e.g., number of layers, channels) is detailed in Section 4.2 based on the target datasets (e.g., CCPG, Gait3D).

A key component of our backbone is the Weighted Global-Local Feature Fusion Module (WFFM), adapted from our prior work [15] to effectively process the fused multi-modal features. WFFM aims to extract both coarse-grained global motion patterns and fine-grained discriminative details from local body parts. The WFFM module (Fig. 4) takes an input feature map $X_{in} \in \mathbb{R}^{C_{in}\times t\times h\times w}$.

1. **Global Feature Path:** A Conv3D layer processes X_{in} to capture holistic spatio-temporal features $Y_{global} \in \mathbb{R}^{C_{out} \times t \times h \times w}$.
2. **Local Feature Path:** X_{in} is first partitioned along the height dimension into n anatomically-inspired local regions (e.g., head, upper body, legs, feet). Each local region X_{local}^{j} is then processed by an independent Conv3D layer $Conv_{local}^{j}$ to learn part-specific features. The resulting local features are concatenated along the height dimension to form $Y_{local} \in \mathbb{R}^{C_{out} \times t \times h \times w}$. This structured partitioning, as opposed to uniform slicing, helps preserve the integrity of motion features from distinct body parts.
3. **Weighted Fusion:** The global and local features are then adaptively fused using learnable channel-wise weights p_g and p_l:

$$Y_{WFFM} = Y_{global} \odot p_g + Y_{local} \odot p_l \tag{6}$$

In the final WFFM module of the backbone (denoted WFFM'), before the HPP layer, the global and local features are typically concatenated channel-wise instead of weighted addition, i.e., $Y_{WFFM'} = \text{cat}(Y_{global}, Y_{local})$, to provide a richer set of features for the subsequent pooling stage.

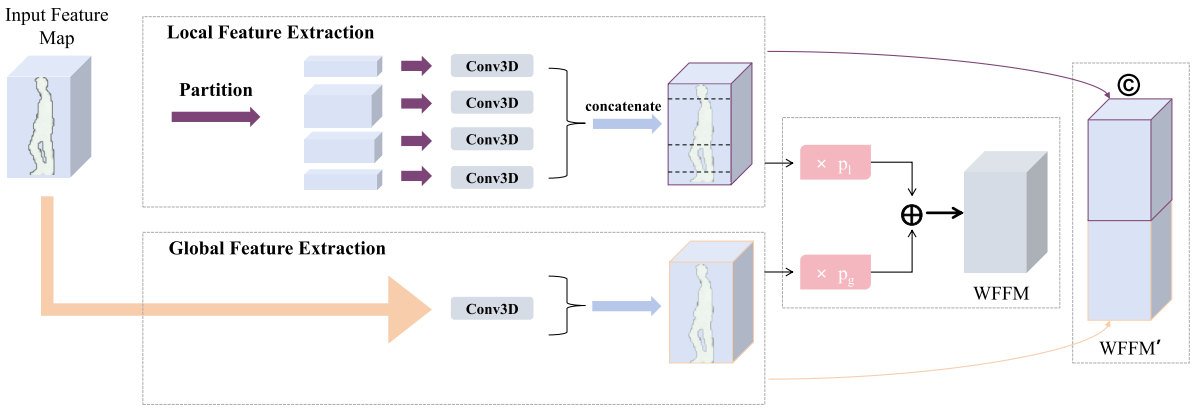

Figure 4: Structure of the weighted global-local feature fusion module (WFFM)

This backbone structure allows GaitMAFF to learn a hierarchical and comprehensive representation from the fused silhouette and skeleton data.

3.5 Feature Mapping and Loss Function

After the deep feature extraction backbone, a Temporal Pooling (TP) layer, typically max pooling along the time axis, is applied to aggregate frame-level features into a sequence-level representation. Subsequently, Horizontal Pyramid Pooling (HPP) [2,14] is employed. HPP divides the feature maps into multiple horizontal strips at different scales. Features within each strip are aggregated using Generalized-Mean (GeM) pooling [14], and the resulting vectors are concatenated. This HPP feature is then passed through a fully connected (FC) layer to obtain the final gait embedding.

The network is trained using a combination of Triplet Loss (L_{tri}) for metric learning and Cross-Entropy Loss (L_{ce}) for classification:

$$L_{total} = \alpha L_{tri} + \beta L_{ce} \tag{7}$$

where α and β are balancing hyperparameters (typically set to 1). The Triplet Loss aims to pull features from the same identity (anchor and positive) closer while pushing features from different identities (anchor and negative) further apart in the embedding space:

$$L_{tri} = \max(0, D(f_a, f_p) - D(f_a, f_n) + margin) \tag{8}$$

where f_a, f_p, f_n are the embeddings for an anchor, a positive sample, and a negative sample, respectively, $D(\cdot, \cdot)$ is the Euclidean distance, and $margin$ is a predefined margin. The Cross-Entropy Loss supervises the classification of gait embeddings into identity labels.

4 Experiments

In this section, we conduct extensive experiments to evaluate the performance of our proposed GaitMAFF network. We first describe the datasets and evaluation protocols, followed by implementation details. Then, we compare GaitMAFF with state-of-the-art (SOTA) methods on challenging datasets. Finally, we perform ablation studies to validate the effectiveness of key components in GaitMAFF.

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

We evaluate GaitMAFF on two large-scale public datasets captured in complex, real-world scenarios:

- **CCPG Dataset [10]:** This dataset is specifically designed for evaluating gait recognition under clothing variations. It contains over 16,000 sequences from 200 subjects, each wearing seven different outfits and walking in both indoor and outdoor settings. We follow the standard protocol: data from the first 100 subjects (IDs 0-99) is used for training, and data from the remaining 100 subjects (IDs 100-199) is used for testing. The test set includes different covariate conditions: full clothing change (CL-Full), upper-body clothing change (CL-UP), lower-body clothing change (CL-DN), and carrying a bag (BG).
- **Gait3D Dataset [11]:** This is a large-scale dataset captured by 39 cameras in a real supermarket surveillance environment. It comprises 25,309 sequences from 4000 subjects. We use the standard split: 18,940 sequences from the first 3000 subjects for training and 6369 sequences from the remaining 1000 subjects for testing. During testing, one sequence per subject is randomly selected as the probe, and the rest form the gallery.

4.1.2 Evaluation Metrics

We use standard evaluation metrics for gait recognition:

- **Rank-N Accuracy:** The percentage of probe samples for which the correct identity is found within the top-N retrieved gallery samples. We report Rank-1, and sometimes Rank-5 accuracy.
- **Mean Average Precision (mAP):** Provides a comprehensive measure of retrieval quality across all recall levels.
- **Mean Inverse Negative Penalty (mINP) [24]:** Measures the model's performance on the most difficult positive matches, reflecting its robustness.

All evaluations exclude identical-view cases unless otherwise specified.

4.2 Implementation Details

4.2.1 Input Processing

For all datasets, input silhouette frames and generated skeleton maps were resized to 64 × 44 pixels. For each sequence, $T = 30$ frames were sampled following the strategy in [3]. For skeleton map generation, 2D keypoints were first extracted using HRNet [17] pre-trained on COCO. The Gaussian variance hyperparameter σ in Eqs. (1) and (2) was set to 8.0. Specifically, the preprocessing pipeline involves: (1) Normalizing skeleton coordinates to a centered canonical view; (2) Generating Joint and Limb Maps; (3) Resizing both silhouette and skeleton maps to 64 × 44; and (4) Applying a uniform temporal sampling strategy to obtain fixed-length sequences ($T = 30$).

4.2.2 Network Architectures

The detailed backbone architectures for GaitMAFF on CCPG and Gait3D are presented in Tables 1 and 2, respectively. These tables specify the input/output channels, kernel sizes, and partitioning strategies for WFFM modules. The initial Conv3D layer for shallow feature extraction (before fusion) uses a 3 × 3 × 3 kernel with stride 1. The WFFM modules also use 3 × 3 × 3 kernels.

Table 1: GaitMAFF backbone architecture on CCPG Dataset (after modality fusion)

Module type	Input channels	Output channels	Kernel size	WFFM partition
Initial Conv3D (Sil/Ske)	1/2	32	(3, 3, 3)	–
		— Adaptive Modality Fusion Output (Input to Backbone) —		
Fused Input	32			
WFFM	32	64	(3, 3, 3)	(8, 32, 16, 8)
WFFM	64	64	(3, 3, 3)	(8, 32, 16, 8)
Max Pooling	–	–	(1, 2, 2)	–
WFFM	64	128	(3, 3, 3)	(4, 16, 8, 4)
WFFM	128	128	(3, 3, 3)	(4, 16, 8, 4)
WFFM’	128	256	(3, 3, 3)	(4, 16, 8, 4)

Table 2: GaitMAFF backbone architecture on Gait3D dataset (after modality fusion)

Module type	Input channels	Output channels	Kernel Size	WFFM partition
Initial Conv3D (Sil/Ske)	1/2	16	(3, 3, 3)	–
		— Adaptive Modality Fusion Output (Input to Backbone) —		
Fused Input	16			
Conv3D	16	32	(3, 3, 3)	–
WFFM	32	64	(3, 3, 3)	(8, 32, 16, 8)
WFFM	64	128	(3, 3, 3)	(8, 32, 16, 8)
Max Pooling	–	–	(1, 2, 2)	–
WFFM’	128	256	(3, 3, 3)	(4, 16, 8, 4)

4.2.3 Training Details

The margin for triplet loss was set to 0.2. We used the stochastic gradient descent (SGD) optimizer with a weight decay of 5×10^{-4}. For the CCPG dataset, the batch size was (8 identities, 16 sequences per identity). The initial learning rate was 0.01, decayed by a factor of 0.1 at 20, 40, and 70 K iterations, with a total of 80 K iterations. For the Gait3D dataset, the batch size was (32 identities, 4 sequences per identity). The initial learning rate was 0.01, decayed by 0.1 at 30, 60, and 80 K iterations, for a total of 90 K iterations. Label smoothing was applied to the cross-entropy loss for improved generalization.

4.3 Comparison with State-of-the-Art Methods

4.3.1 Results on CCPG Dataset

Table 3 shows the Rank-1 accuracy of GaitMAFF compared to SOTA silhouette-based, skeleton-based, and multi-modal methods on the CCPG dataset under various covariate conditions. GaitMAFF achieves an average Rank-1 accuracy of 84.6%, outperforming most existing methods. Specifically, under the challenging CL-Full condition, GaitMAFF achieves 80.5%, demonstrating its robustness to significant appearance changes. Compared to the best performing silhouette-only method (DeepGaitV2 [25] with 83.3% average) and skeleton-only method (MSGG (Ske) [6] with 34.6% average), GaitMAFF shows a clear advantage, highlighting the benefits of our adaptive multi-modal fusion. While SkeletonGait++ [21] shows strong performance, particularly in the BG condition, GaitMAFF exhibits competitive or superior results across most clothing variation scenarios (CL-Full, CL-DN). This underscores GaitMAFF's ability to effectively integrate complementary cues and adapt to appearance variations.

Table 3: Rank-1 accuracy (%) comparison with SOTA methods on the CCPG dataset

Modality	Method	CL-Full	CL-UP	CL-DN	BG	Average
Skeleton-based	GaitGraph2 [26]	5.0	5.3	5.8	6.2	5.6
	Gait-TR [27]	15.9	19.3	18.5	19.6	18.3
	MSGG (Ske) [6]	29.1	35.2	37.1	36.9	34.6
Silhouette-based	GaitSet [2]	60.6	65.1	64.6	69.0	64.8
	GaitPart [3]	64.1	67.5	67.9	72.4	68.0
	GaitBase [16]	72.1	75.5	76.3	78.9	75.7
	DeepGaitV2 [25]	78.6	84.8	80.7	89.2	83.3
Multi-modal	BiFusion (MSGG) [6]	62.6	67.6	66.3	66.0	65.6
	SkeletonGait++ [21]	79.1	83.9	81.7	89.9	84.1
	GaitMAFF (Ours)	80.5	84.0	85.0	88.9	84.6

4.3.2 Results on Gait3D Dataset

Table 4 presents the performance of GaitMAFF on the Gait3D dataset against other leading methods. GaitMAFF achieves a Rank-1 accuracy of 58.7%, mAP of 47.9%, and mINP of 29.6%. This performance is significantly better than most single-modal methods. For instance, it surpasses the silhouette-based GaitSet [2] (36.7% Rank-1) and the skeleton-based SkeletonGait [21] (36.1% Rank-1) by a large margin. Compared to our silhouette-only backbone (GaitATGL, which was 51.0% Rank-1), GaitMAFF's multi-modal approach yields a substantial improvement of 7.7 percentage points in Rank-1 accuracy. This highlights the critical role of fusing skeleton information in highly complex and unconstrained surveillance scenarios.

Among multi-modal methods, GaitMAFF also demonstrates SOTA performance, outperforming recent methods like HybridGait [28] (53.3% Rank-1) and GaitRef [23] (49.0% Rank-1). These results strongly validate the effectiveness of GaitMAFF's adaptive fusion and robust feature representation in handling the diverse challenges present in Gait3D.

Table 4: Performance comparison with SOTA methods on the Gait3D dataset (%)

Modality	Method	Rank-1	Rank-5	mAP	mINP
Skeleton-based	GaitMGL [29]	22.7	42.7	18.7	9.5
	SkeletonGait [21]	36.1	57.9	28.4	17.2
	GaitSet [2]	36.7	58.3	30.0	17.3
Silhouette-based	GaitPart [3]	28.2	47.6	21.6	12.4
	GaitGL [14]	29.7	48.5	22.3	13.3
	GLN [13]	31.4	52.9	24.7	13.6
	GaitGCI [30]	50.3	68.5	39.5	24.3
	GaitATGL [15]	51.0	68.9	39.1	23.2
Multi-modal	GaitRef [23]	49.0	69.3	40.7	25.3
	HybridGait [28]	53.3	72.0	43.3	26.7
	SMPLGait [11]	46.3	64.5	37.2	22.2
	MSAFF [31]	48.1	66.6	38.5	23.5
	GaitMAFF (Ours)	58.7	76.5	47.9	29.6

4.4 Visualization of Feature Space

To provide an intuitive demonstration of our model's discriminative capability, we employ t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the feature space distributions. We compare GaitMAFF with several state-of-the-art methods on the test sets of both the CCPG and Gait3D datasets, as depicted in Fig. 5.

On the CCPG dataset, the feature clusters generated by GaitMAFF are visibly more compact and distinctly separated compared to those from the strong silhouette-based method DeepGaitV2 [25] and the multi-modal SkeletonGait++ [21]. While the baseline methods show considerable overlap between clusters of different identities, GaitMAFF effectively pulls features of the same identity together while pushing others apart. A similar trend is observed on the more challenging Gait3D dataset, where GaitMAFF produces significantly more defined and well-separated clusters than the silhouette-based GaitATGL [15] and the multi-modal HybridGait [28].

Collectively, these visualizations provide direct qualitative evidence that our adaptive multi-modal fusion strategy leads to a more discriminative feature representation, achieving superior intra-class compactness and inter-class separability. This intuitively explains the quantitative performance gains reported in our experiments.

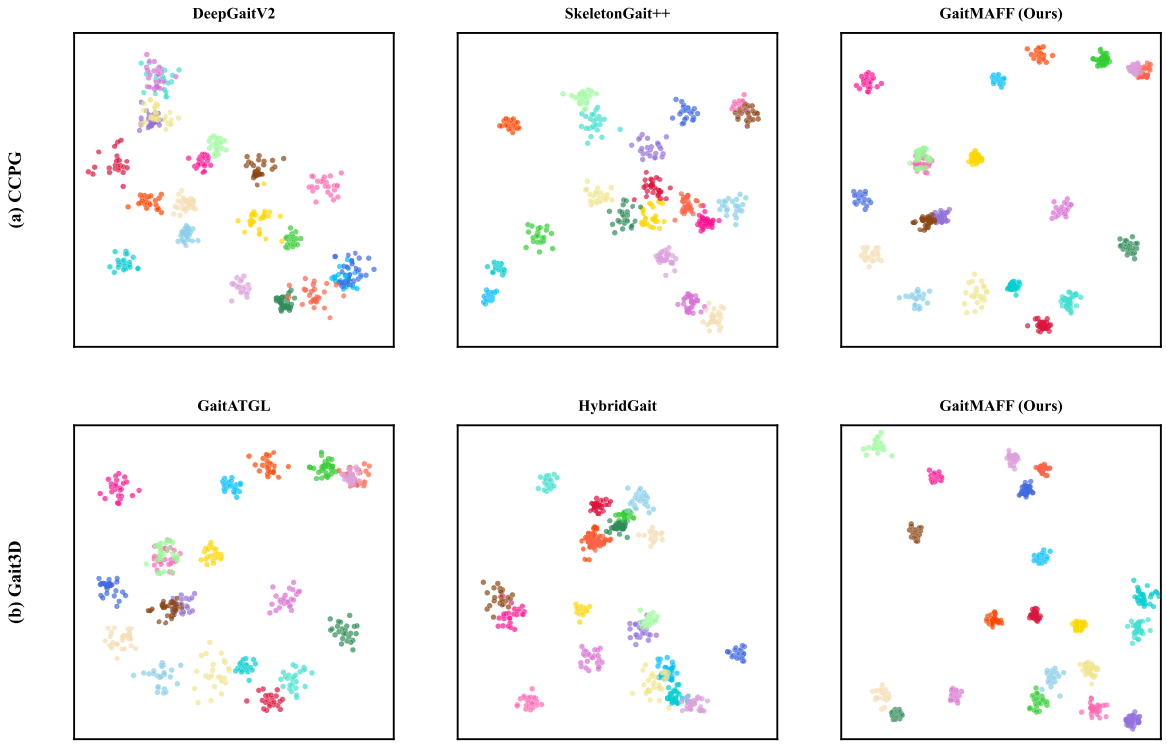

Figure 5: t-SNE visualization of feature distributions on the (**a**) CCPG and (**b**) Gait3D test sets. Each color represents a different subject identity. Our GaitMAFF model produces more compact and well-separated clusters compared to other SOTA methods, demonstrating its superior discriminative power

4.5 Ablation Studies

We conduct ablation studies on the CCPG dataset to analyze the contribution of key components in GaitMAFF. We report average Rank-1 accuracy across all four test conditions (CL-Full, CL-UP, CL-DN, BG).

4.5.1 Effectiveness of Multi-Modal Input

Table 5 investigates the impact of using different input modalities.

Table 5: Ablation study on the impact of input modalities on CCPG (Average Rank-1 %)

Method	CL-Full	CL-UP	CL-DN	BG	Average
Skeleton-map only	42.2	46.7	52.3	54.3	48.9
Silhouette only	77.6	83.3	81.1	86.8	82.2
GaitMAFF (Multi-Modal)	80.5	84.0	85.0	88.9	84.6

- **Skeleton-Map Only:** Using only the generated Skeleton Maps as input to the GaitMAFF backbone.
- **Silhouette Only:** Using only silhouettes as input to the GaitMAFF backbone.
- **GaitMAFF (Silhouette + Skeleton-Map):** Our full proposed multi-modal approach.

The results clearly demonstrate the superiority of the multi-modal approach. GaitMAFF (84.6%) significantly outperforms both Skeleton-Map Only (48.9%) and Silhouette Only (82.2%). The 2.4 percentage point

improvement over the strong Silhouette Only baseline highlights the effective complementary information provided by the skeleton maps and the efficacy of our fusion strategy. The substantial gain over Skeleton-Map Only (35.7 percentage points) confirms that appearance information from silhouettes is crucial.

4.5.2 Effectiveness of Adaptive Fusion Strategy

Table 6 compares our proposed adaptive fusion network with simpler fusion strategies:

Table 6: Ablation study on different modality fusion strategies on CCPG (Average Rank-1 %)

Fusion strategy	CL-Full	CL-UP	CL-DN	BG	Average
Addition	79.8	83.3	83.6	88.5	83.8
Concatenation	80.3	84.3	84.1	88.6	84.3
GaitMAFF (Adaptive)	80.5	84.0	85.0	88.9	84.6

- **Addition:** Element-wise addition of F_{sil} and F_{ske}, then fed to the backbone.
- **Concatenation:** Channel-wise concatenation of F_{sil} and F_{ske}, followed by a 1×1 Conv3D to adjust channels, then fed to the backbone.
- **GaitMAFF (Adaptive Fusion):** Our proposed fusion method.

Our adaptive fusion (84.6%) outperforms both direct addition (83.8%) and concatenation (84.3%). This indicates that adaptively learning the weights for combining features, as done in our fusion network, is more effective than static fusion operations, allowing the model to better emphasize relevant information from each modality.

4.5.3 Effectiveness of WFFM in the Backbone

To validate the contribution of the Weighted Global-Local Feature Fusion Modules (WFFM) within the GaitMAFF framework, we replace all WFFM modules in the backbone (after the modality fusion stage) with standard Conv3D blocks (maintaining similar parameter counts by adjusting channel numbers or block depth). As shown in Table 7, using WFFM (84.6%) leads to a notable improvement over using standard Conv3D blocks (Avg. 82.9%). This confirms that WFFM's strategy of structured local feature partitioning and adaptive global-local fusion is beneficial for processing the fused multi-modal features as well, capturing more discriminative details.

Table 7: Ablation study on the WFFM module in GaitMAFF's backbone on CCPG (Average Rank-1 %)

Backbone feature extractor	Average rank-1
Standard Conv3D blocks	82.9
WFFM (Ours)	84.6

4.5.4 Ablation on WFFM Partition Ratios

The partition strategy in WFFM (e.g., 8/32/16/8 for head/upper-body/legs/feet) is designed based on human anatomical proportions. To validate this choice, we compared it with other partition strategies on the CCPG dataset.

Table 8 shows that the anatomically aligned partition (8/32/16/8) outperforms the uniform partition (83.1%) and a coarser 3-part strategy (82.8%). This confirms that aligning the feature splitting with the semantic structure of the human body (allocating more channels to the complex upper body and legs, and fewer to the head and feet) is crucial for effective feature extraction.

Table 8: Comparison of different WFFM partition strategies on CCPG (Average Rank-1 %)

Partition strategy (Head/Upper/Legs/Feet)	Average rank-1
Uniform (16/16/16/16)	83.1
Coarse (16/32/16)	82.8
Anatomical (8/32/16/8)	84.6

4.5.5 Effectiveness of Skeleton Map Components

To further validate our Skeleton Map representation, we analyze the individual contributions of its two components: the Joint Map and the Limb Map. We conduct experiments on the CCPG dataset using variants of GaitMAFF where the skeleton modality is represented by only the Joint Map, only the Limb Map, or both, as proposed.

The results in Table 9 show that both components are crucial for achieving the best performance. While using only the Joint Map (Avg. 83.1%) or only the Limb Map (Avg. 82.5%) already provides a significant boost compared to a skeleton-less baseline, combining them yields the highest accuracy (Avg. 84.6%). This confirms that the joint locations and their connections (limbs) provide complementary information that, when fused, creates a more comprehensive and robust gait representation.

Table 9: Ablation study on skeleton map components on CCPG (Average Rank-1 %)

Skeleton map input	CL-Full	CL-UP	CL-DN	BG	Average
Joint map only	79.1	83.0	82.4	87.9	83.1
Limb map only	78.5	82.2	81.7	87.5	82.5
Joint + Limb maps (Ours)	80.5	84.0	85.0	88.9	84.6

4.5.6 Sensitivity to Pose Estimation Noise

To evaluate the robustness of GaitMAFF against pose estimation errors, we conducted an experiment on the CCPG dataset by introducing synthetic noise to the skeleton keypoints. We added Gaussian noise with zero mean and varying standard deviations ($\sigma_{noise} \in \{0, 2, 4, 6, 8\}$ pixels) to the coordinates of the input joints before generating the Skeleton Maps. The results are summarized in Table 10.

As shown in Table 10, GaitMAFF maintains relatively stable performance under low to moderate noise levels ($\sigma_{noise} \leq 4$). Even with significant noise ($\sigma_{noise} = 8$), the accuracy only drops by 4.2%, remaining above 80%. This resilience can be attributed to our adaptive fusion mechanism, which can dynamically down-weight the skeleton modality when its reliability is compromised, relying more on the silhouette branch. This experiment confirms GaitMAFF's robustness to pose estimation errors likely to be encountered in real-world scenarios.

Table 10: Performance of GaitMAFF under varying levels of synthetic pose noise on the CCPG dataset (Average Rank-1 %)

Noise level (σ_{noise})	0	2	4	6	8
Rank-1 accuracy	84.6	84.2	83.5	82.1	80.4
Drop	–	−0.4	−1.1	−2.5	−4.2

4.6 Computational Complexity

To evaluate the practical applicability of our proposed method, we analyze the computational complexity of GaitMAFF in terms of the number of parameters and Floating Point Operations (FLOPs). We compare our model with several other state-of-the-art methods.

As shown in Table 11, GaitMAFF demonstrates a competitive balance between performance and computational cost. While achieving superior accuracy, its parameter count (5.5 M) and computational load (2.8 GFLOPs) are comparable to other recent multi-modal approaches like SkeletonGait++ and HybridGait. This indicates that the performance gains from our adaptive fusion and WFFM modules are achieved without introducing excessive computational overhead, making GaitMAFF a practical solution for real-world deployment.

Table 11: Comparison of computational complexity. FLOPs are calculated for a single sequence of 30 frames

Method	Parameters (M)	FLOPs (G)
GaitSet [2]	2.5	1.5
GaitPart [3]	3.1	2.1
SkeletonGait++ [21]	5.2	2.5
HybridGait [28]	5.8	3.2
GaitMAFF (Ours)	5.5	2.8

5 Conclusion

This paper introduced GaitMAFF, a Multi-modal Adaptive Feature Fusion Network, designed to enhance gait recognition robustness in complex real-world scenarios. GaitMAFF effectively integrates silhouette and skeleton modalities through three key innovations: a novel Skeleton Map representation for improved cross-modal alignment, an adaptive attention-based fusion network to dynamically weigh modal contributions, and a powerful spatio-temporal backbone incorporating Weighted Global-Local Feature Fusion Modules (WFFM) for discriminative feature learning.

Experimental results on the challenging CCPG and Gait3D datasets demonstrate GaitMAFF's superior performance over state-of-the-art single-modal and multi-modal methods. Notably, GaitMAFF achieved an average Rank-1 accuracy of 84.6% on CCPG and 58.7% on Gait3D, showcasing its efficacy in handling significant appearance variations and complex surveillance conditions. Ablation studies confirmed the significant contributions of each proposed component.

In essence, GaitMAFF offers a significant step towards practical and reliable gait-based identification in diverse and challenging settings. However, several limitations remain. First, the reliance on high-quality 2D pose estimation means that severe occlusions or failures in the upstream pose estimator can still degrade performance, despite our adaptive fusion. Second, although our computational complexity is comparable

to existing multi-modal methods, it may still be high for resource-constrained edge devices. Third, the current framework primarily focuses on standard RGB scenarios, which may raise privacy concerns in certain applications. Future work will focus on three main directions: (1) Designing lightweight versions of the fusion module and backbone to facilitate deployment on edge devices; (2) Exploring cross-dataset transfer learning techniques to enhance the model's generalization ability across different domains without extensive retraining; and (3) Extending the framework to incorporate privacy-preserving modalities, such as thermal imaging or depth sensors, to address growing privacy concerns in biometric applications.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Natural Science Foundation of Chongqing Municipality, grant number CSTB2022NSCQ-MSX0503.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Zhongbin Luo and Yanqiu Bi; methodology, Zhongbin Luo and Zhaoyang Guan; software, Zhongbin Luo; validation, Zhongbin Luo and Wenxing You; formal analysis, Zhongbin Luo; data curation, Wenxing You; writing—original draft preparation, Zhongbin Luo; writing—review and editing, Yanqiu Bi; visualization, Wenxing You and Yunteng Wang; supervision, Yanqiu Bi; project administration, Yanqiu Bi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in CCPG at https://github.com/BNU-IVC/CCPG and Gait3D at https://gait3d.github.io/.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Sepas-Moghaddam A, Etemad A. Deep gait recognition: a survey. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):264–84. doi:10.1109/tpami.2022.3151865.
2. Chao H, He Y, Zhang J, Feng J. Gaitset: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2019. Vol. 33. p. 8126–33.
3. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, et al. Gaitpart: temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2020. p. 14225–33.
4. Teepe T, Khan A, Gilg J, Herzog F, Hörmann S, Rigoll G. Gaitgraph: graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). Piscataway, NJ, USA: IEEE; 2021. p. 2314–8.
5. Hasan MB, Ahmed T, Ahmed S, Kabir MH. GaitGCN++: improving GCN-based gait recognition with part-wise attention and DropGraph. J King Saud Univ-Comput Inf Sci. 2023;35(7):101641. doi:10.1016/j.jksuci.2023.101641.
6. Peng Y, Ma K, Zhang Y, He Z. Learning rich features for gait recognition by integrating skeletons and silhouettes. Multimedia Tools Appl. 2024;83(3):7273–94. doi:10.1007/s11042-023-15483-x.
7. Li G, Guo L, Zhang R, Qian J, Gao S. TransGait: multimodal-based gait recognition with set transformer. Appl Intell. 2023;53(2):1535–47. doi:10.1007/s10489-022-03543-y.
8. Cui Y, Kang Y. Multi-modal gait recognition via effective spatial-temporal feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 17949–57.
9. Min F, Guo S, Fan H, Dong J. GaitMA: pose-guided multi-modal feature fusion for gait recognition. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ, USA: IEEE; 2024. p. 1–6.

10. Li W, Hou S, Zhang C, Cao C, Liu X, Huang Y, et al. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 13824–33.
11. Zheng J, Liu X, Liu W, He L, Yan C, Mei T. Gait recognition in the wild with dense 3D representations and a benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 20228–37.
12. Han J, Bhanu B. Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell. 2005;28(2):316–22. doi:10.1109/tpami.2006.38.
13. Hou S, Cao C, Liu X, Huang Y. Gait lateral network: learning discriminative and compact representations for gait recognition. In: European Conference on Computer Vision. Glasgow, UK: Springer; 2020. p. 382–98.
14. Lin B, Zhang S, Yu X. Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2021. p. 14648–56.
15. Xu Y, Xi H, Ren K, Zhu Q, Hu C. Gait recognition via weighted global-local feature fusion and attention-based multiscale temporal aggregation. J Electron Imaging. 2025;34(1):013002.
16. Fan C, Liang J, Shen C, Hou S, Huang Y, Yu S. Opengait: revisiting gait recognition towards better practicality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 9707–16.
17. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2019. p. 5693–703.
18. Bukht TFN, Alazeb A, Mudawi NA, Alabdullah B, Alnowaiser K, Jalal A, et al. Robust human interaction recognition using extended kalman filter. Comput Mater Contin. 2024;81(2):2987–3002.
19. Chen D, Chen M, Wu P, Wu M, Zhang T, Li C. Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. Sci Rep. 2025;15(1):4982. doi:10.1038/s41598-025-87752-8.
20. Jiang H, Ye L, Hu J, Chen X, Chen S, Zhang W, et al. WarmGait: thermal array-based gait recognition for privacy-preserving person Re-ID. IEEE Trans Mob Comput. 2025:1–14. doi:10.1109/tmc.2025.3608447.
21. Fan C, Ma J, Jin D, Shen C, Yu S. SkeletonGait: gait recognition using skeleton maps. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2024. Vol. 38. p. 1662–9.
22. Ye D, Fan C, Ma J, Liu X, Yu S. Biggait: learning gait representation you want by large vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 200–10.
23. Zhu H, Zheng W, Zheng Z, Nevatia R. Gaitref: gait recognition with refined sequential skeletons. In: 2023 IEEE International Joint Conference on Biometrics (IJCB). Piscataway, NJ, USA: IEEE; 2023. p. 1–10.
24. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC. Deep learning for person re-identification: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2021;44(6):2872–93. doi:10.1109/tpami.2021.3054775.
25. Fan C, Hou S, Huang Y, Yu S. Exploring deep models for practical gait recognition. arXiv:2303.03301. 2023.
26. Teepe T, Gilg J, Herzog F, Hörmann S, Rigoll G. Towards a deeper understanding of skeleton-based gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 1569–77.
27. Zhang C, Chen XP, Han GQ, Liu XJ. Spatial transformer network on skeleton-based gait recognition. Expert Syst. 2023;40(6):e13244. doi:10.1111/exsy.13244.
28. Dong Y, Yu C, Ha R, Shi Y, Ma Y, Xu L, et al. HybridGait: a benchmark for spatial-temporal cloth-changing gait recognition with hybrid explorations. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2024. Vol. 38. p. 1600–8.
29. Zhang Z, Wei S, Xi L, Wang C. GaitMGL: multi-scale temporal dimension and global-local feature fusion for gait recognition. Electronics. 2024;13(2):257. doi:10.3390/electronics13020257.

30. Dou H, Zhang P, Su W, Yu Y, Lin Y, Li X. Gaitgci: generative counterfactual intervention for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 5578–88.
31. Zou S, Xiong J, Fan C, Shen C, Yu S, Tang J. A multi-stage adaptive feature fusion neural network for multimodal gait recognition. IEEE Trans Biom Behav Identity Sci. 2024;6(4):539–49. doi:10.1109/tbiom.2024.3384704.