Exploring More from Multiple Gait Modalities for Human Identification

Dongyang Jin[1*], **Chao Fan**[2,1*], **Weihua Chen**[3], **Shiqi Yu**[1†]

[1]Department of Computer Science and Engineering, Southern University of Science and Technology, China
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China
[3]Alibaba Group
12332451@mail.sustech.edu.cn, chaofan996@szu.edu.cn, kugang.cwh@alibaba-inc.com, yusq@sustech.edu.cn

Abstract

The gait, as a kind of soft biometric characteristic, can reflect the distinct walking patterns of individuals at a distance, exhibiting a promising technique for unrestrained human identification. With largely excluding gait-unrelated cues hidden in RGB videos, the silhouette and skeleton, though visually compact, have acted as two of the most prevailing gait modalities for a long time. Recently, several attempts have been made to introduce more informative data forms like human parsing and optical flow images to capture gait characteristics, along with multi-branch architectures. However, due to the inconsistency within model designs and experiment settings, we argue that a comprehensive and fair comparative study among these popular gait modalities, involving the representational capacity and fusion strategy exploration, is still lacking. From the perspectives of fine vs. coarse-grained shape and whole vs. pixel-wise motion modeling, this work presents an in-depth investigation of three popular gait representations, i.e., silhouette, human parsing, and optical flow, with various fusion evaluations, and experimentally exposes their similarities and differences. Based on the obtained insights, we further develop a C^2Fusion strategy, consequently building our new framework MultiGait++. C^2Fusion preserves commonalities while highlighting differences to enrich the learning of gait features. To verify our findings and conclusions, extensive experiments on Gait3D, GREW, CCPG, and SUSTech1K are conducted.

Code — https://github.com/ShiqiYu/OpenGait

Introduction

The pedestrian gait presented in walking videos typically involves the visual characteristics of body shape and limb movements. This unique application, known as gait recognition, distinguishes itself from other biometrics techniques such as face, fingerprint, and iris recognition, by offering the flexibility of non-intrusive and long-distance usage without necessitating the subject's cooperation. Furthermore, gait is inherently hard to disguise and conceal. These attributes render gait recognition particularly well-suited for unconstrained security applications such as suspect tracking and retrieval (Nixon and Carter 2006).

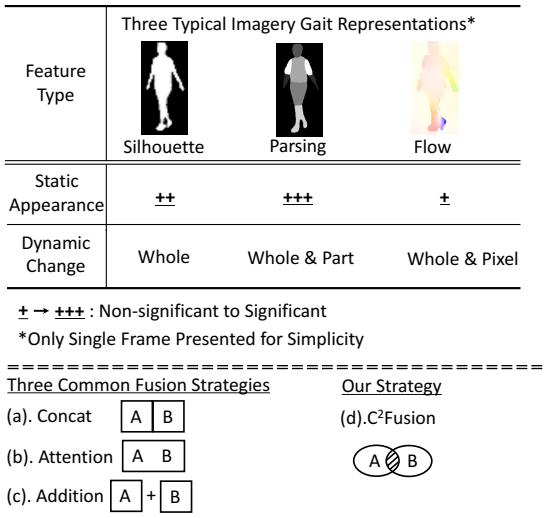

Figure 1: Top: comparing three typical gait modalities, i.e., the binary silhouette, body parsing, and optical flow images. Bottom: the comparison between three common fusion strategies with our C^2Fusion.

To mitigate the influence of irrelevant visual cues such as texture and background, gait recognition methods often rely on derived gait representations extracted from RGB videos rather than the videos themselves as inputs. The widely used ones include binary silhouettes (Chao et al. 2022; Fan et al. 2020; Lin, Zhang, and Yu 2021; Shen et al. 2023b; Wang et al. 2024b; Fan et al. 2023a,b), skeleton coordinates (Liao et al. 2017; Teepe et al. 2021; Fan et al. 2024b), SMPL model (Li et al. 2020; Zheng et al. 2022a), body parsing (Zheng et al. 2023), and optical flow images (Castro et al. 2024; Feng, Yuan, and Fan 2023). Among these, the silhouette is favored for its clarity in color and richness in shape, making it the most popular choice for gait modeling. Recent studies have pointed out the limitations in silhouette images, citing a lack of fine-grained part-level shape (Zheng et al. 2023) and explicit body structural characteristics (Peng et al. 2024), and introduce extra body parsing images and human joints coordinate, often driven by multi-branch networks, to reach the new state-of-the-art recognition performance. In this paper, we agree that the multimodal approach represents

*These authors contributed equally.
†Corresponding author.

a pivotal direction for advancing gait recognition research.

Despite ongoing efforts, a fair and comprehensive comparative study among some representative gait modalities remains necessary, due to the usual misalignments in model and experimental settings. To advance this understanding, this work investigates three typical gait modalities: silhouette, human parsing, and optical flow images, as depicted in Figure 1. These modalities, each with distinct physical meanings, emphasize different aspects of gait: whole-body shapes, body parts, and pixel-level motion dynamics, respectively, offering varied levels of feature granularity. Among them, optical flow images are particularly notable for capturing pixel-level motion in each frame, setting them apart from silhouettes and human parsing images that primarily focus on body shapes. While these modalities have their differences, they also share similarities. For instance, the overall contours of the human body can be represented by both binary silhouettes and human parsing images. All three modalities convey dynamic changes in human limbs. Leveraging three common fusion strategies depicted in Figure 1 (a-c), this work introduces the MultiGait series, i.e., a set of uni- and multimodal baselines, that provide a comprehensive study on gait modality fusion. This approach allows for a fair and thorough examination of the efforts of each modality for multimodal gait modeling.

Building on the exploration of similarities and differences among silhouette, body parsing, and optical flow images, this work goes a step further by proposing a novel gait modality fusion strategy termed C^2Fusion, leading to the development of a new multimodal method named MultiGait++. As illustrated in Figure 1(d), the core idea of C^2Fusion is to extract shared characteristics across different modalities while simultaneously encouraging each modality to emphasize its unique attributes beyond these commonalities. This design forces MultiGait++ to fully explore the diverse features offered by given modalities, thereby enhancing its discriminative capacity. Experiments show that MultiGait++ achieves a new SoTA on the Gait3D (Zheng et al. 2022a), GREW (Zhu et al. 2021), CCPG (Li et al. 2023) and SUSTech1K (Shen et al. 2023a) datasets.

Overall, this work contributes to the field in two-fold:

- **An In-Depth Gait Modality Fusion Study under Fair Conditions:** MultiGait marks one of the first comparative studies on uni- vs. multimodal gait recognition under fair conditions. Through extensive experiments, we comprehensively examine the specific contributions and limitations of modalities such as silhouette, human parsing, and optical flow images for gait description.
- **A Novel Multimodal Gait Recognition Method:** We introduce MultiGait++, featured by its core component, C^2Fusion. This approach maximizes the extraction of diverse features from given modalities, thereby enhancing the overall representation of gait patterns.

Related Work

Gait Modalities

Gait modalities are captured by a range of sensors, including traditional RGB cameras as well as emerging technologies like event cameras (Wang et al. 2022), LiDAR (Shen et al. 2023a; Wang et al. 2023a, 2024a; Guo et al. 2025), and fisheye cameras (Xu et al. 2023). Among these, RGB-based cameras continue to dominate gait recognition due to their cost-effectiveness and seamless integration with existing CCTV systems. Consequently, the following literature review focuses on gait modalities extracted from RGB images, which include binary silhouettes, optical flow images, human parsing, and 2D/3D poses. Each modality offers distinct advantages, yet they all aim to minimize the influence of gait-irrelevant factors such as clothing color, texture, and background. We categorize gait recognition methods into two groups based on the number of modalities used: unimodal and multimodal.

Unimodal Gait Recognition Methods

This kind of method usually extracts gait features from the sequence of binary silhouettes, 2D/3D coordinates, human parsing, or optical flow images.

With the rapid advancement of deep learning, silhouette-based methods primarily focus on extracting both spatial and temporal features of gait. For instance, GaitSet (Chao et al. 2022) treats the gait sequence as a set and uses a maximum function to compress frame-level spatial features. GaitPart (Fan et al. 2020) pays more attention to spatial and temporal details, showcasing their significance. GaitGL (Lin et al. 2022) introduces a GLConv block to capture global and local spatial features simultaneously. The latest OpenGait (Fan et al. 2024a) offers a comprehensive exploration of deep model design for outdoor gait recognition, achieving strong performance across various datasets.

In pose-based methods, PoseGait (Liao et al. 2020) combines 3D skeleton data with hand-crafted features to address the challenges posed by changes in viewpoint and clothing. GaitGraph (Teepe et al. 2021) uses a graph convolution network to learn gait modalities based on 2D skeletons. SkeletonGait (Fan et al. 2024b) proposes an image-like skeleton modality to enhance gait feature learning, offering novel insights into the role of body structure characteristics.

Human parsing is visually similar to silhouettes that provide fine-grained body part annotations. Recent studies on parsing-based gait recognition (Zou et al. 2024a) typically focus on modeling both the entire body and individual body parts. For example, GaitParsing (Wang et al. 2023b) addresses the self-occlusion problem through human parsing. ParsingGait (Zheng et al. 2023) extracts holistic body features while incorporating a GCN branch to capture structural relationships among various body parts.

Individuals have unique movements and speeds, making movement behavior a crucial aspect of human identity recognition. While silhouettes and human parsing focus on body shape, optical flow is gaining attention for its representation of instantaneous motion. Ye (Ye, Sun, and Xu 2023) and Xu (Xu, Li, and Hou 2023) proposed that the motion information in optical flow can better match individuals with similar body shapes. AttenGait (Castro et al. 2024) demonstrates that the modality information in optical flow is richer than that in silhouettes, showing great potential for gait recognition.

Multimodal Gait Recognition Methods

An increasing number of methods are now focused on extracting rich features from multiple gait modalities. For instance, SMPLGait (Zheng et al. 2022b) utilizes the 3D SMPL model to refine the learning from gait silhouettes. BiFusion (Peng et al. 2024) integrates skeletons and silhouettes to capture the comprehensive spatiotemporal characteristics of human gait. Feng (Feng, Yuan, and Fan 2023) explored the complementary nature of movement and shape information by combining silhouettes and optical flow images. XGait (Zheng et al. 2024) proposes a novel cross-granularity alignment method to unleash the power of gait modalities of different granularity. Additionally, SkeletonGait++ (Fan et al. 2024b) combines silhouettes and the proposed skeleton maps in a frame-by-frame manner, fully leveraging the strengths of CNNs. In this work, we propose a new idea that extracts common characteristics across different modalities while simultaneously encouraging each modality to express its unique attributes, thereby enhancing the learning of rich multimodal gait features.

Method

Silhouette, Parsing, and Optical Flow for Gait

Here we study three mainstream gait modalities: **silhouette, human parsing, and optical flow**. In previous research, silhouette has been the primary focus, while human parsing presented an emerging topic showing significant potential. Although optical flow has been less shiny in recent leading venues for gait recognition, we include it due to its fine-grained nature for motion description.

To ensure an intuitive and fair investigation, we adopt a uniform framework to model these three modalities and their combinations. Structurally, we utilize the architecture of DeepGaitV2 (Fan et al. 2024a) thanks to its straightforward design and strong performance. As a result, three unimodal baselines, *i.e.*, MultiGait^{s} (totally identical to DeepGaitV2), MultiGait^{p}, MultiGait^{f}, as well as two multimodal ones, *i.e.*, MultiGait^{s+p} and MultiGait^{s+f}, are shown in Figure. 2, where s, p and f denote the input of silhouette, human parsing, and optical flow images, respectively.

These baselines share identical blocks, with the multimodal ones employing additional naive multi-branch structures. Additionally, we introduce MultiGait^{2s}, which doubles the channels of MultiGait^{s}, to eliminate the effects of increased parameters brought by multi-branch designs. To enrich the research scope, we consider various feature fusion locations (input, middle, and high level) and mechanisms (element-wise addition, channel-wise concatenation, and cross-attention fusion[1]), as shown in Figure. 2 (b). For now, we defer the implementation details to focus on key insights that advance multimodal gait recognition and inspire the design principles behind our MultiGait++.

Human Parsing: Intuitively, human parsing images offer more detailed cues of body part labels, shapes, and structures compared to binary silhouettes. This should ideally allow for a more fine-grained interpretation of human gait. However, as shown in Table 1, MultiGait^{p} does not meet these expectations. We attribute this discrepancy to the inherently complex nature of human parsing extraction, which, while providing additional characteristics, can also introduce potential noise due to its fine-grained segmentation granularity compared to silhouettes. This also explains why existing methods tend to combine the silhouette and human parsing rather than replacing the former with the latter.

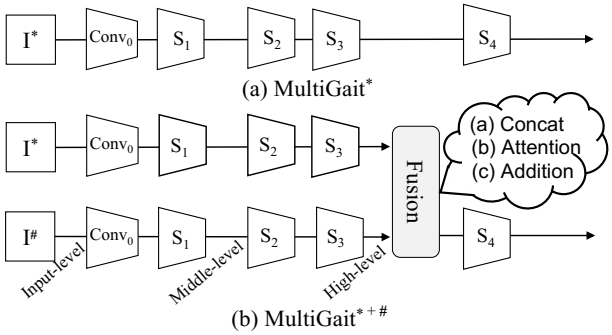

Figure 2: The architecture of the MultiGait series. Here the symbols * and # can be instantiated with any of the employed gait modalities, such as the silhouette, human parsing, and optical flow, in theory.

Optical Flow: Optical flow images excel at capturing pixel-level motion but struggle to represent body shapes (the major advantage of gait silhouettes). Surprisingly, Table 1 shows that MultiGait^{f} achieves highly competitive performance compared with the silhouette-based MultiGait^{s}. This highlights the untapped potential of optical flow in enhancing gait description.

Silhouette+Human Parsing: Table 1 indicates that directly combining silhouettes and human parsing images, *i.e.*, MultiGait^{s+p}, can enhance the performance significantly. Moreover, we observe that these gains are nearly uniform over various fusion locations and mechanisms. This suggests a high degree of homogeneity in how these two modalities describe human gait, despite their differences in detail. Therefore, this work plans to merge them at the input level, saving the model's complexity and computational demands while preserving SoTA performance.

Silhouette+Optical Flow: We observe that integrating silhouettes and optical flow images, *i.e.*, MultiGait^{s+f}, at a higher level will result in more performance improvements. This suggests that silhouettes, which represent body shape, and optical flow, which capture pixel-level motion, complement each other but are not homogeneous in data forms. By fusing them at a high level, the combined utility is enhanced.

Based on the above comprehensive study, we present solid evidence showing that:

- The silhouette provides accurate whole-body shapes, while human parsing adds detailed body part features. Thanks to homogeneity in data forms, they can be effectively fused at the input level. On the other hand, optical flow presents distinct pixel-wise motion characteristics

[1]Here we utilize the popular cross-attention mechanism from SkeletonGait++ (Fan et al. 2024b).

Method	Fusion Location	Fusion Mechanism	Probe Sequence (Rank-1 acc) NM	BG	CL	CR	UB	UN	OC	NT	Overall R-1	R-5
MultiGait^{s}	N/A		86.5	82.8	49.2	80.4	83.3	81.9	86.0	28.0	80.9	91.9
MultiGait^{p}	N/A		84.7	77.3	29.4	78.2	75.8	80.2	87.9	43.3	77.3	91.3
MultiGait^{f}	N/A		84.4	82.8	54.9	79.3	82.4	79.8	89.7	35.0	80.5	92.5
MultiGait^{2s}	N/A		87.6	83.9	50.1	81.3	85.1	83.6	86.7	29.1	81.9	92.3
MultiGait^{s+p}	Input	Concatenation	89.6	87.1	45.1	85.4	87.4	86.7	90.3	43.2	85.1	94.9
	Middle	Concatenation	90.6	86.5	44.6	85.7	86.7	86.6	91.2	44.5	85.2	94.9
		Attention	90.2	86.7	43.5	85.6	87.0	87.1	91.5	43.0	85.1	94.8
		Addition	89.5	87.0	41.5	85.7	87.2	86.6	91.4	45.1	85.2	94.9
	High	Concatenation	90.9	86.7	42.4	85.8	87.4	87.3	92.3	44.5	85.4	94.9
		Attention	90.4	85.9	38.4	84.9	86.8	86.4	91.3	43.6	84.5	94.5
		Addition	91.1	87.3	43.5	86.2	88.0	87.4	92.3	45.5	85.8	95.0
MultiGait^{s+f}	Input	Concatenation	86.3	83.2	51.5	80.8	84.1	82.5	86.5	27.8	81.3	92.1
	Middle	Concatenation	87.0	83.7	51.8	81.2	84.1	82.3	87.0	29.3	81.7	92.4
		Attention	87.4	84.2	52.0	81.5	84.9	82.6	87.5	28.6	82.1	92.6
		Addition	87.4	84.1	52.8	81.6	85.2	83.3	87.7	28.4	82.3	92.5
	High	Concatenation	88.8	86.1	55.2	83.2	87.0	84.6	88.5	31.0	83.9	93.4
		Attention	89.1	85.4	54.1	82.7	86.3	84.7	88.9	31.2	83.5	93.2
		Addition	88.6	85.5	55.2	83.1	86.8	84.2	88.6	32.0	83.7	93.5

Table 1: Recognition results of MultiGait on SUSTech1K.

and should be appended at a higher level to capture its unique contributions.

- Exploring the shared discriminative features across these three gait modalities, while also identifying and leveraging their unique characteristics, is essential for enriching the learning of multimodal gait representations.

MultiGait++

Motivation: Silhouette, human parsing, and optical flow images are all image-based gait modalities that, despite their differences, naturally share a substantial amount of common features. Recognizing this point, we propose to encourage each branch to highlight its unique discriminative characteristics beyond these commonalities, thereby enriching the features for multimodal gait description.

Overall: In line with practices established by MultiGait series, we develop a new multimodal method named MultiGait++. Specifically, MultiGait++ processes each frame equally and fuses them in a frame-by-frame manner, thus the following formulation considers a single frame for brevity.

As shown in Figure 3, MultiGait++ employs a three-stream architecture. The appearance branch takes the concatenation of silhouette and human parsing image as input, while the parallel motion branch processes the optical flow image. It is important to note that both branches capture different granularities of body shape and dynamic features, with the terms 'appearance' and 'motion' used here only for clarification in presentation.

The first component of C^{2}Fusion, i.e., the C^2 module, then extracts shared features of the above two branches to form an additional common branch. Meanwhile, the C^2 module also refines the output features of these two branches to emphasize their differences.

After Stage2 and 3, another component of C^{2}Fusion, a straightforward concatenation fusion operation, then aggregates the features from all three branches.

The final Stage4 and gait head project the rich features extracted from multiple gait modalities into the identity metric space. Following the general practices summarized by OpenGait (Fan et al. 2023c), the gait head comprises several widely used components, including temporal pooling, horizontal pooling, separate fully connected layers, and BNNecks (Luo et al. 2019). The overall training process is driven by the triplet and softmax losses.

In the following section, we focus on the key component of C^{2}Fusion, specifically the C^2 module.

C^{2}Fusion: Through the Conv0 and Stage1, the appearance and motion branch respectively output feature f_{ap} and f_{mo} with a shape of $C \times H \times W$ (channel$\times$ height $\times$ width).

As shown in Figure 3 (a), the C^2 module initiates a global understanding through a cross-attention operation:

$$m_{ap}, m_{mo} = \text{Softmax}(E_{ap}(f_{ap}), E_{mo}(f_{mo})), \quad (1)$$

where E_{ap} and E_{mo} represent two simple squeeze-and-excitation networks with a squeeze rate of 16. With the help of an element-wise softmax function, these networks project the f_{ap} and f_{mo} to an identical-size attention map m_{ap} and m_{mo}, respectively.

Intuitively, the smaller value within $|m_{ap} - m_{mo}|$ reveals the more commonalities between branches since they result in similar attention activations. Next, we conduct an element-wise minimize operation (Min) between m_{ap} and

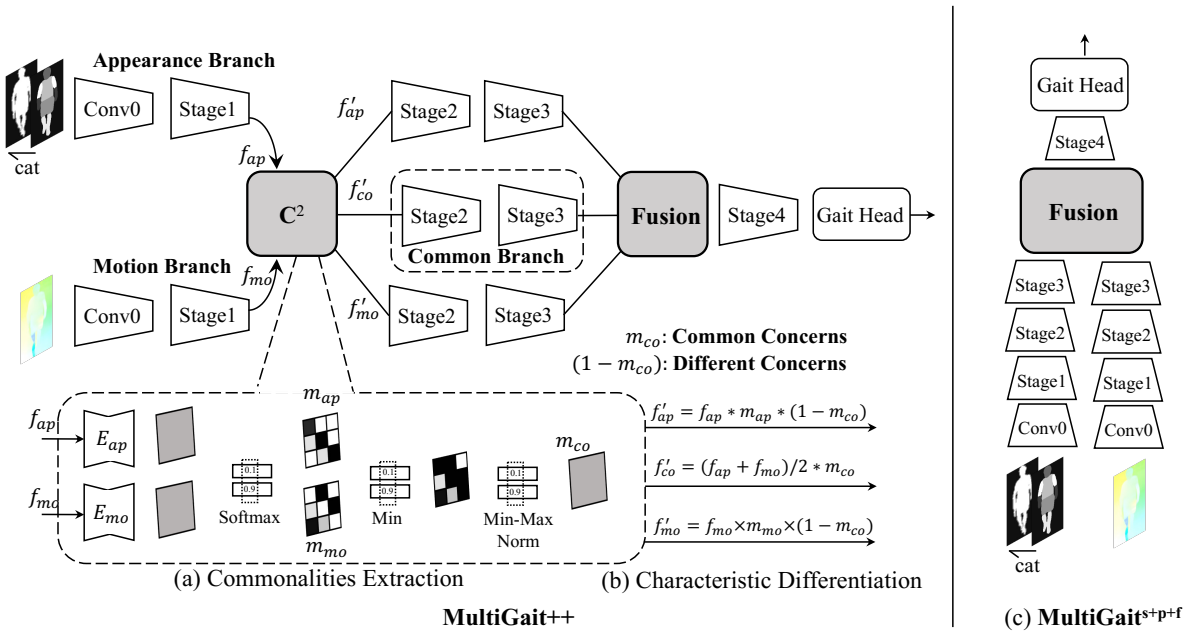

Figure 3: Left: Our pipeline of MultiGait++. Right: The architecture of MultiGait^{s+p+f}

DataSet	Batch Size	Milestones	Steps
SUSTech1K	(8, 8, 10)	(20k, 30k, 40k)	50k
CCPG	(8, 16, 30)	(20k, 40k, 50k)	60k

Table 2: Implementation details. The batch size (q, p, k) indicates q IDs, p sequences per ID, and k frames per sequences.

m_{mo} and perform a further Min-Max normalization (Norm) over the spatial dimension:

$$m_{co} = \text{Norm}(\text{Min}(m_{ap}, m_{mo})), \tag{2}$$

where the output attention map m_{co} can represent the common concerns of different branches since the larger value within m_{co} means the smaller value within $|m_{ap} - m_{mo}|$.

Similarly, we define the different concerns of branches by m_{di}, e.g., $1 - m_{co}$. Finally, the initial features of the common branch (the commonalities between branches) can be formulated as:

$$f'_{co} = \frac{f_{ap} + f_{mo}}{2} * m_{co}, \tag{3}$$

meanwhile, we refine the appearance and motion branch (highlight differences of each branch) by:

$$\begin{aligned} f'_{ap} &= f_{ap} * m_{ap} * m_{di}, \\ f'_{mo} &= f_{mo} * m_{mo} * m_{di} \end{aligned} \tag{4}$$

as shown in Figure 3 (b).

Experiment

Dataset

We conduct main experiments on the popular SUSTech1K and CCPG datasets, which offer publicly available RGB videos, allowing for flexible extraction of multiple gait modalities. Specifically, we utilize the officially provided silhouettes and personally perform the tasks of human parsing and optical flow extraction. Detailed pretreatment procedures are provided in the **Supplementary Materials**[2]. The SUSTech1K offers many scenarios, including the normal(NM), bags(BG), clothing(CL), carrying(CR), umbrella(UB), uniform (UM), occlusion(OC), and nighttime(NT) conditions. Alternatively, the CCPG is designed around the challenges of clothing factors, featuring a diverse collection of full-(CL), up-(UP), down-clothing(DN), and bag(BG) changes. Our experiments strictly follow official evaluation protocols and take rank-1 accuracy(R-1) and mean average precision(mAP) as the primary metric.

GREW (Zhu et al. 2021) and Gait3D (Zheng et al. 2022a) are two large real-world datasets. In recent years, more and more research (Fan et al. 2023c, 2024a; Shen et al. 2024) suggest paying more attention to real-world datasets. Therefore, we conduct additional experiments on these two challenging real-world datasets, despite the absence of parsing and flow images in each dataset.

Implementation Details

1) Silhouette, human parsing, and optical flow images are aligned and resized to 1×64×44, 1×64×44, and 3×64×44 (Takemura et al. 2018). 2) Table 2 shows the main hyper-parameters of our experiments; 3) The spatial augmentation strategy suggested by OpenGait (Fan et al. 2023c) is adopted; 4) The SGD optimizer with an initial learning rate of 0.1 and weight decay of 0.0005 is utilized. 5) A naive combination of MultiGait^{s+p}(input-level and cat

[2]Available at https://arxiv.org/pdf/2412.11495

Modality	Method	Probe Sequence (R-1)								Overall	
		NM	BG	CL	CR	UB	UN	OC	NT	R-1	R-5
Skeleton	GaitGraph2 (Teepe et al. 2022)	22.2	18.2	6.8	18.6	13.4	19.2	27.3	16.4	18.6	40.2
	SkeletonGait (Fan et al. 2024b)	55.0	51.0	24.7	49.9	42.3	52.0	62.8	43.9	50.1	72.6
Sils	GaitSet (Chao et al. 2022)	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
	GaitPart (Fan et al. 2020)	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
	GaitGL (Lin, Zhang, and Yu 2021)	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
	GaitBase (Fan et al. 2023c)	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	89.4
	DeepGaitV2 (Fan et al. 2023a)	86.5	82.8	49.2	80.4	83.3	81.9	86.0	28.0	80.9	91.9
Parsing	GaitBase^{p} (Fan et al. 2023c)	80.3	71.1	32.5	72.6	66.3	73.5	81.7	40.7	71.7	88.4
	DeepGaitV2^{p} (Fan et al. 2023a)	84.7	77.3	29.4	78.2	75.8	80.2	87.9	43.3	77.3	91.3
Flow	DeepGaitV2^{f} (Fan et al. 2023a)	84.4	82.8	**54.9**	79.3	82.4	79.8	89.7	35.0	80.5	92.5
Sils+Skeleton	BiFusion (Peng et al. 2024)	69.8	62.3	45.4	60.9	54.3	63.5	77.8	33.7	62.1	83.4
	SkeletonGait++ (Fan et al. 2024b)	85.1	82.9	46.6	81.9	80.8	82.5	86.2	**47.5**	81.3	95.5
Sils+Parsing+Flow	MultiGait^{s+p+f} (Ours)	90.0	87.8	44.2	86.2	88.3	87.7	91.8	44.8	86.0	95.1
	MultiGait++ (Ours)	**92.0**	**89.4**	50.4	**87.6**	**89.7**	**89.1**	**93.4**	45.1	**87.4**	**95.6**

Table 3: Evaluation with different attributes on SUSTech1K.

Modality	Method	CL		UP		DN		BG		Mean
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Skeleton	GaitGrapgh2 (Teepe et al. 2022)	5.0	2.5	5.3	4.0	5.8	4.2	6.2	4.6	5.6
	GaitTR (Zhang et al. 2023)	15.7	9.7	18.3	16.1	18.5	16.4	17.5	15.3	17.5
	SkeletonGait (Fan et al. 2024b)	40.4	20.8	48.5	35.8	53.0	40.3	61.7	48.1	50.9
Sils	GaitSet (Chao et al. 2022)	60.2	47.4	65.2	63.2	65,1	61.9	68.5	65.8	64.8
	GaitPart (Fan et al. 2020)	64.3	48.0	67.8	63.2	68.6	63.8	71.7	66.4	68.1
	GaitBase (Fan et al. 2023c)	71.6	58.5	75.0	71.6	76.8	73.3	78.6	75.3	75.5
	DeepGaitV2 (Fan et al. 2023a)	78.6	62.1	84.8	81.5	80.7	78.1	89.2	85.8	83.3
Parsing	GaitBase^{p} (Fan et al. 2023c)	59.1	44.1	62.1	57.5	66.8	61.9	68.1	64.2	64.0
	DeepGaitV2^{p} (Fan et al. 2023a)	69.6	51.7	75.8	71.4	75.8	71.6	83.3	79.3	76.1
Flow	DeepGaitV2^{f} (Fan et al. 2023a)	68.8	49.5	73.3	67.5	75.0	68.0	76.8	71.1	73.5
RGB	GaitEdge (Liang et al. 2022)	66.9	-	74.0	-	70.6	-	77.1	-	72.2
	BigGait (Ye et al. 2024)	82.6	-	85.9	-	**87.1**	-	**93.1**	-	87.2
Sils+Skeleton	BiFusion (Peng et al. 2024)	62.6	46.7	67.6	64.1	66.3	61.9	66.0	61.9	65.6
	SkeletonGait++ (Fan et al. 2024b)	79.1	63.6	83.9	81.2	81.7	79.3	89.9	87.0	83.7
Sils+Parsing	XGait (Zheng et al. 2024)	72.8	59.5	77.0	74.3	79.1	75.7	80.5	78.2	77.4
Sils+Parsing+Flow	MultiGait^{s+p+f} (Ours)	81.3	65.2	87.2	83.6	82.9	80.3	90.6	87.4	85.5
	MultiGait++ (Ours)	**83.9**	**68.5**	**89.0**	**85.8**	86.0	**82.5**	91.5	**88.9**	**87.6**

Table 4: Evaluation with different attributes on CCPG.

fusion) and MultiGait^{s+f}(high-level and attention fusion), termed MultiGait^{s+p+f} as shown in Figure 3 (c), is introduced to act as a strong baseline.

Comparison Around MultiGait++

Results on SUSTech1K: As shown in Table 3, both MultiGait++ and its baseline, MultiGait^{s+p+f}, achieve rank-1 accuracies that significantly surpass other state-of-the-art methods in most cases. This demonstrates the substantial benefits of multimodal modeling for gait recognition. More importantly, MultiGait++ outperforms its baseline, demonstrating the effectiveness of our C^{2}Fusion strategy.

Results on CCPG: On the more challenging CCPG dataset, MultiGait++ outperforms all other SoTA methods across all conditions, as shown in Table 4. Excluding its baseline, MultiGait++ raises the SoTA standard by +4.8%, +4.2%, +4.3%, and +1.6% in rank-1 accuracy on the CL, UP, DN, and BG subsets, respectively. It also achieves considerable

Modality	Method	GREW			Gait3D			
		R-1	R-5	R-10	R-1	R-5	mAP	mINP
Sils	GaitSet (Chao et al. 2022)	46.3	63.6	70.3	36.7	58.3	30.0	17.3
	GaitPart (Fan et al. 2020)	44.0	60.7	67.3	28.2	47.6	21.6	12.4
	GaitGL (Lin, Zhang, and Yu 2021)	47.3	-		29.7	48.5	22.3	13.6
	GaitBase (Fan et al. 2023c)	60.1	75.5	80.4	64.6		-	
	QAGait (Wang et al. 2024b)	59.1	74.0	79.2	67.0	81.5	56.5	-
	DeepGaitV2 (Fan et al. 2023a)	77.7	88.9	91.8	74.4	88.0	65.8	39.2
Skeleton	GaitGraph2 (Teepe et al. 2022)	33.5	-		11.1		-	
	GaitTR (Zhang et al. 2023)	54.5	-		6.6		-	
	SkeletonGait (Fan et al. 2024b)	77.4	87.9	91.0	38.1	56.7	28.9	16.1
Sils+Skeleton	GaiRef (Zhu et al. 2023)	53.0	67.9	73.0	49.0	49.3	40.7	25.3
	MSAFF (Zou et al. 2024b)	57.4	73.0	78.3	48.1	66.6	38.5	23.5
	SkeletonGait++ (Fan et al. 2024b)	85.8	92.6	94.3	77.6	89.4	70.3	42.6
Sils+Flow	GaitFusion (Feng, Yuan, and Fan 2023)	83.1	91.3	93.6			-	
	MultiGait^{s+f} (Ours)	91.4	96.4	97.5			-	
	MultiGait++$^{s+f}$ (Ours)	**93.4**	**97.3**	**98.3**			-	
Sils+Parsing	XGait (Zheng et al. 2024)		-		80.5	91.9	73.3	55.4
	MultiGait^{s+p} (Ours)		-		83.0	94.5	78.6	62.7
	MultiGait++$^{s+p}$ (Ours)		-		**85.4**	**94.9**	**80.5**	**65.2**

Table 5: Recognition results on two real-world gait datasets, involving GREW, and Gait3D.

idx	m_{co} in Eq.3	m_{di} in Eq.4	NM	UB	UN	OC	NT	Overall
(a)	×	×	90.0	87.8	87.7	91.8	44.8	86.0
(b)	×	✓	90.1	88.7	87.9	93.0	**45.4**	86.5
(c)	✓	×	90.6	89.0	88.1	92.5	43.5	86.4
(d)	✓	✓	**92.0**	**89.7**	**89.1**	**93.4**	45.1	**87.4**

Table 6: Ablation study on common concerns m_{co} and different concerns m_{di}.

improvements compared to its baseline, i.e., +2.1% in rank-1 accuracy averaged over CCPG. These results underscore the exceptional capability and practicality of MultiGait++ in handling complex clothing variations.

More Results on Other Real-world Datasets: To further validate the effectiveness of MultiGait++, we conduct additional experiments on two challenging real-world datasets, GREW and Gait3D, despite the absence of parsing and flow images in each dataset. In this phase, we modify MultiGait++ (and MultiGait) into a two-branch input design, with a combination of silhouette and parsing for Gait3D, and silhouette and flow for GREW.

The results shown in Table 5 demonstrate the notable superiority of both MultiGait and MultiGait++. Excluding its baseline MultiGait, MultiGait++ raises the SoTA standard by +7.6% and +4.9% in rank-1 accuracy on GREW and Gait3D. It is also worth mentioning that MultiGait++ continues to outperform its high-performance baseline, MultiGait, underscoring the robustness and effectiveness of the proposed C^{2}Fusion.

Ablation Study

To validate the C^2 module's effectiveness, we conduct ablation experiments on m_{co} and m_{di} in Table 6. Specifically, Table 6 (b) means that removing m_{co} but remaining m_{di} in Figure 3 (b); Table 6 (c) means that removing the term of m_{di} but remaining the term of m_{co} in Figure 3 (b); Table 6 (a) and 6 (d) means the MultiGait^{s+p+f} and MultiGait++.

Fusing these three modalities in various ways consistently yields high performance on SUSTech1K (compared to other SoTA methods in Table 3), highlighting the benefits of multimodal methods. Furthermore, the C^2 module in MultiGait++, which preserves shared characteristics (by m_{co}) while emphasizing unique features (by m_{di}), further enhances recognition accuracy.

Conclusion

This work studies three typical gait modalities: silhouettes, human parsing, and optical flow. It emphasizes the importance of multimodal methods in gait recognition. Furthermore, the proposed C^{2}Fusion strategy encourages each modality to highlight its unique features while preserving shared characteristics, effectively enriching the features for multimodal gait description. The integrated approach achieves superior accuracy and demonstrates that multimodal gait recognition has much to explore in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62476120, and the Shenzhen International Research Cooperation Project under Grant GJHZ20220913142611021.

References

Castro, F. M.; Delgado-Escaño, R.; Hernández-García, R.; Marín-Jiménez, M. J.; and Guil, N. 2024. AttenGait: Gait recognition with attention and rich modalities. *Pattern Recognition*, 148: 110171.

Chao, H.; Wang, K.; He, Y.; Zhang, J.; and Feng, J. 2022. GaitSet: Cross-View Gait Recognition Through Utilizing Gait As a Deep Set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3467–3478.

Fan, C.; Hou, S.; Huang, Y.; and Yu, S. 2023a. Exploring Deep Models for Practical Gait Recognition. *arXiv preprint arXiv:2303.03301*.

Fan, C.; Hou, S.; Liang, J.; Shen, C.; Ma, J.; Jin, D.; Huang, Y.; and Yu, S. 2024a. OpenGait: A Comprehensive Benchmark Study for Gait Recognition towards Better Practicality. *arXiv preprint arXiv:2405.09138*.

Fan, C.; Hou, S.; Wang, J.; Huang, Y.; and Yu, S. 2023b. Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; and Yu, S. 2023c. OpenGait: Revisiting Gait Recognition Towards Better Practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9707–9716.

Fan, C.; Ma, J.; Jin, D.; Shen, C.; and Yu, S. 2024b. SkeletonGait: Gait Recognition Using Skeleton Maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1662–1669.

Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; and He, Z. 2020. GaitPart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14225–14233.

Feng, Y.; Yuan, J.; and Fan, L. 2023. GaitFusion: Exploring the Fusion of Silhouettes and Optical Flow for Gait Recognition. In Iliadis, L.; Papaleonidas, A.; Angelov, P.; and Jayne, C., eds., *Artificial Neural Networks and Machine Learning – ICANN 2023*, 88–99. Cham: Springer Nature Switzerland. ISBN 978-3-031-44195-0.

Guo, W.; Liang, Y.; Pan, Z.; Xi, Z.; Feng, J.; and Zhou, J. 2025. Camera-LiDAR Cross-modality Gait Recognition. In *European Conference on Computer Vision*, 439–455. Springer.

Li, W.; Hou, S.; Zhang, C.; Cao, C.; Liu, X.; Huang, Y.; and Zhao, Y. 2023. 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*, 13824–13833.

Li, X.; Makihara, Y.; Xu, C.; Yagi, Y.; Yu, S.; and Ren, M. 2020. End-to-end model-based gait recognition. In *Proceedings of the Asian Conference on Computer Vision*.

Liang, J.; Fan, C.; Hou, S.; Shen, C.; Huang, Y.; and Yu, S. 2022. GaitEdge: Beyond Plain End-to-End Gait Recognition for Better Practicality. In *Computer Vision – ECCV 2022*.

Liao, R.; Cao, C.; Garcia, E. B.; Yu, S.; and Huang, Y. 2017. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In *Chinese conference on biometric recognition*, 474–483. Springer.

Liao, R.; Yu, S.; An, W.; and Huang, Y. 2020. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98: 107069.

Lin, B.; Zhang, S.; Wang, M.; Li, L.; and Yu, X. 2022. Gaitgl: Learning discriminative global-local feature representations for gait recognition. *arXiv preprint arXiv:2208.01380*.

Lin, B.; Zhang, S.; and Yu, X. 2021. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14648–14656.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Nixon, M. S.; and Carter, J. N. 2006. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11): 2013–2024.

Peng, Y.; Ma, K.; Zhang, Y.; and He, Z. 2024. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia Tools and Applications*, 83(3): 7273–7294.

Shen, C.; Fan, C.; Wu, W.; Wang, R.; Huang, G. Q.; and Yu, S. 2023a. LidarGait: Benchmarking 3D Gait Recognition With Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1054–1063.

Shen, C.; Lin, B.; Zhang, S.; Yu, X.; Huang, G. Q.; and Yu, S. 2023b. Gait recognition with mask-based regularization. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.

Shen, C.; Yu, S.; Wang, J.; Huang, G. Q.; and Wang, L. 2024. A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets, and Challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10.

Teepe, T.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2022. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1569–1577.

Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2021. GaitGraph: graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2314–2318. IEEE.

Wang, R.; Shen, C.; Fan, C.; Huang, G. Q.; and Yu, S. 2023a. PointGait: Boosting End-to-End 3D Gait Recognition with Point Clouds via Spatiotemporal Modeling. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.

Wang, R.; Shen, C.; Marin-Jimenez, M. J.; Huang, G. Q.; and Yu, S. 2024a. Cross-Modality Gait Recognition: Bridging LiDAR and Camera Modalities for Human Identification. *arXiv preprint arXiv:2404.04120*.

Wang, Y.; Zhang, X.; Shen, Y.; Du, B.; Zhao, G.; Cui, L.; and Wen, H. 2022. Event-Stream Representation for Human Gaits Identification Using Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3436–3449.

Wang, Z.; Hou, S.; Zhang, M.; Liu, X.; Cao, C.; and Huang, Y. 2023b. GaitParsing: Human Semantic Parsing for Gait Recognition. *IEEE Transactions on Multimedia*.

Wang, Z.; Hou, S.; Zhang, M.; Liu, X.; Cao, C.; Huang, Y.; Li, P.; and Xu, S. 2024b. QAGait: Revisit Gait Recognition from a Quality Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5785–5793.

Xu, C.; Makihara, Y.; Li, X.; and Yagi, Y. 2023. Gait recognition from fisheye images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1030–1040.

Xu, J.; Li, H.; and Hou, S. 2023. Attention-based gait recognition network with novel partial representation PGOFI based on prior motion information. *Digital Signal Processing*, 133: 103845.

Ye, D.; Fan, C.; Ma, J.; Liu, X.; and Yu, S. 2024. BigGait: Learning Gait Representation You Want by Large Vision Models. *arXiv preprint arXiv:2402.19122*.

Ye, H.; Sun, T.; and Xu, K. 2023. Gait Recognition Based on Gait Optical Flow Network with Inherent Feature Pyramid. *Applied Sciences*, 13(19): 10975.

Zhang, C.; Chen, X.-P.; Han, G.-Q.; and Liu, X.-J. 2023. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, e13244.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022a. Gait Recognition in the Wild with Dense 3D Representations and A Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022b. 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*, 20228–20237.

Zheng, J.; Liu, X.; Wang, S.; Wang, L.; Yan, C.; and Liu, W. 2023. Parsing is All You Need for Accurate Gait Recognition in the Wild. In *Proceedings of the 31st ACM International Conference on Multimedia*, 116–124.

Zheng, J.; Liu, X.; Zhang, B.; Yan, C.; Zhang, J.; Liu, W.; and Zhang, Y. 2024. It Takes Two: Accurate Gait Recognition in the Wild via Cross-granularity Alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8786–8794.

Zhu, H.; Zheng, W.; Zheng, Z.; and Nevatia, R. 2023. Gaitref: Gait recognition with refined sequential skeletons. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.

Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021. Gait Recognition in the Wild: A Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14789–14799.

Zou, S.; Fan, C.; Xiong, J.; Shen, C.; Yu, S.; and Tang, J. 2024a. Cross-Covariate Gait Recognition: A Benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7855–7863.

Zou, S.; Xiong, J.; Fan, C.; Shen, C.; Yu, S.; and Tang, J. 2024b. A multi-stage adaptive feature fusion neural network for multimodal gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.