OpenGait: A Comprehensive Benchmark Study for Gait Recognition towards Better Practicality

Chao Fan, Saihui Hou, Junhao Liang, Chuanfu Shen, Jingzhe Ma, Dongyang Jin, Yongzhen Huang, and Shiqi Yu

***Abstract*—Gait recognition, a rapidly advancing vision technology for person identification from a distance, has made significant strides in indoor settings. However, evidence suggests that existing methods often yield unsatisfactory results when applied to newly released real-world gait datasets. Furthermore, conclusions drawn from indoor gait datasets may not easily generalize to outdoor ones. Therefore, the primary goal of this work is to present a comprehensive benchmark study aimed at improving practicality rather than solely focusing on enhancing performance. To this end, we first develop OpenGait, a flexible and efficient gait recognition platform. Using OpenGait as a foundation, we conduct in-depth ablation experiments to revisit recent developments in gait recognition. Surprisingly, we detect some imperfect parts of certain prior methods thereby resulting in several critical yet undiscovered insights. Inspired by these findings, we develop three structurally simple yet empirically powerful and practically robust baseline models, *i.e.*, DeepGaitV2, SkeletonGait, and SkeletonGait++, respectively representing the appearance-based, model-based, and multi-modal methodology for gait pattern description. Beyond achieving SoTA performances, more importantly, our careful exploration sheds new light on the modeling experience of deep gait models, the representational capacity of typical gait modalities, and so on. We hope this work can inspire further research and application of gait recognition towards better practicality. The code is available at https://github.com/ShiqiYu/OpenGait.**

***Index Terms*—OpenGait, Gait Recognition, Benchmark Study.**

I. Introduction

GAIT recognition has garnered increasing interest within the vision research community. It leverages physiological and behavioral characteristics observed in walking videos to authenticate individuals' identities [1]. Compared with other biometrics, such as the face, fingerprint, and iris, gait patterns can be captured from a distance in uncontrolled settings, without necessitating physical contact. Additionally, as a natural walking behavior, gait is inherently difficult to disguise, rendering it theoretically robust against common subject-related variables, such as clothing, carrying items, and posture variations. These advantages position gait recognition as a promising solution for security applications, including identity authentication and suspect tracking [2].

The widely acknowledged effectiveness and robustness of deep learning architectures [3], [4], [5], [6] have greatly propelled various vision techniques forward. Gait recognition employing deep models has likewise achieved remarkable success [2], [7]. However, emerging evidence [8], [9] suggests that many SoTA gait recognition methods perform not optimally in practical scenarios. For instance, as depicted in Fig. 1 (b), several representative gait models exhibit a significant accuracy degradation of over 40% when transitioning from laboratory testing to in-the-wild evaluation. This performance gap is likely attributed to emerging real-world noisy factors, such as complex occlusions, background variations, and illumination changes. However, through our extensive ablation study, we identify additional sensitive aspects. We discover that conclusions drawn by prior methods may vary across gait datasets. Therefore, beyond proposing an enhanced model for improved performance, the primary objective of this work is to present a comprehensive benchmark study to enhance the practicality of gait recognition. To this end, we make threefold efforts as illustrated by Fig. 1 (a): laying the groundwork with a solid codebase, exposing gaps by SoTA methods revisiting, and inspiring future research through reconstructing gait baseline models.

In previous studies, gait methods were primarily developed using private code repositories and heavily reliant on in-the-lab gait datasets, particularly CASIA-B [10] and OU-MVLP [11]. However, these datasets are partially limited, with CASIA-B comprising data from 124 subjects collected nearly 20 years ago, and OU-MVLP involving only cross-view changes. To accelerate real-world applications, this work considers building a unified evaluation platform, that covers various SoTA gait methods and emerging gait datasets captured in realistic environments, as highly desired. In line with this vision, we introduce a flexible and efficient gait recognition platform termed OpenGait. Sec. III-A will provide the highlight features of OpenGait. Thanks to its extensibility and reusability, OpenGait has been extended to new potentially influential repositories, such as Gait3D-Benchmark [9][1], FastPoseGait [12][2], and All-in-One-Gait[3]. Moreover, OpenGait has also been widely adopted in two major international gait recog-

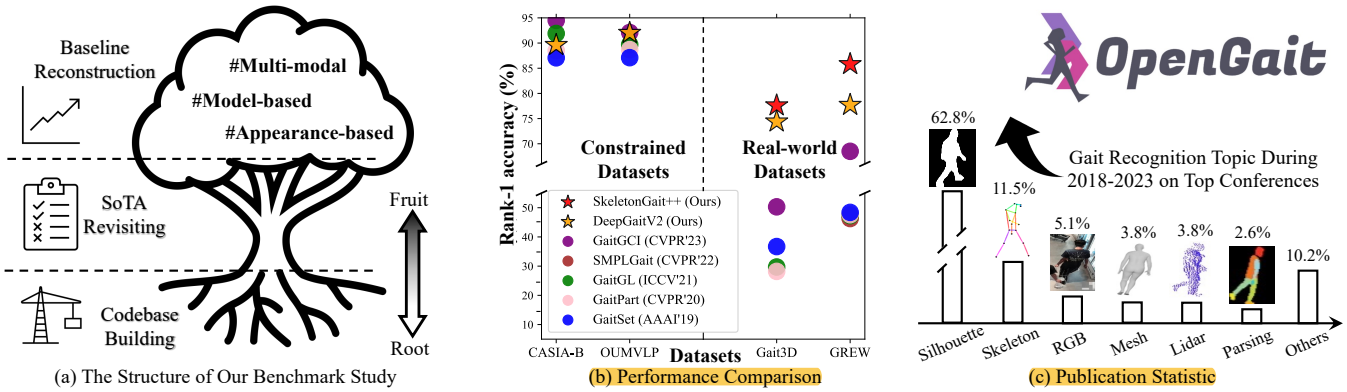
Chao Fan, Junhao Liang, Chuanfu Shen, Jingzhe Ma, Dongyang Jin, and Shiqi Yu are with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. E-mail: {12131100, 12132342, 11950016, 12031127, 12332451}@mail.sustech.edu.cn and yusq@sustech.edu.cn. Chuanfu Shen is also with the Department of Industrial and Manufacturing Systems Engineering, University of Hong Kong, Hong Kong, China.

Saihui Hou and Yongzhen Huang are with the School of Artificial Intelligence, Beijing Normal University, and also with Watrix Technology Limited Co. Ltd., Beijing, China. E-mail: {housaihui, huangyongzhen}@bnu.edu.cn.

Corresponding author: Shiqi Yu.

Manuscript received Jan 28, 2024.

[1] https://github.com/Gait3D/Gait3D-Benchmark
[2] https://github.com/BNU-IVC/FastPoseGait
[3] https://github.com/jdyjjj/All-in-One-Gait



(a) The Structure of Our Benchmark Study (b) Performance Comparison (c) Publication Statistic

Fig. 1. This benchmark study involves the codebase building, previous SoTA revisiting, and strong baseline reconstruction for gait recognition.

nition competitions, *i.e.*, HID [13], and GREW [8]. Notably, all top-10 winning teams at HID2022 [13] and 2023 [14] have utilized OpenGait as the codebase to develop new solutions.

Using OpenGait as the foundation, we reproduce several SoTA methods, with some results presented in Fig. 1 (b). Furthermore, we reassess some commonly accepted conclusions by re-implementing the ablation study on recently curated outdoor gait datasets. Surprisingly, we find that the MGP branch proposed by GaitSet [7], the FConv proposed by GaitPart [15], the local feature extraction branch proposed by GaitGL [16], and the SMPL branch proposed by SMPLGait [9], do not demonstrate superiority on real-world gait datasets. With exhaustive analysis, we reveal several critical yet overlooked limitations of existing gait research, including insufficient ablation study, lack of outdoor evaluation, and the absence of a strong backbone, *etc*.

Inspired by the above discoveries, we develop three structurally simple, experimentally powerful, and empirically robust baseline models for real-world gait recognition. Specifically, we introduce DeepGaitV2[4], SkeletonGait, and SkeletonGait++, each utilizing the silhouette image, skeletal coordinates, and the combination of these two as input, accordingly standing for the appearance-based, model-based, and multi-modal methodology for gait pattern description. As illustrated in Fig. 1 (c), the silhouette and skeleton present two of the most popular gait inputs[5] underscoring their adoption. In terms of network design, this benchmark study prioritizes the use of widely accepted blocks and gait frameworks to enhance the practicality and applicability of our findings. Through extensive experiments, several critical insights regarding the modeling capacity of shallow *vs.* deep gait models and the representational ability of gait silhouette *vs.* skeleton has been provided. By addressing these key challenges straightly, our pragmatic solutions yield significant advancements as depicted in Fig. 1 (b). More details will be provided in Sec. V.

In summary, this benchmark study contributes to gait research in three significant ways: a) OpenGait, a unified and extensible platform, is developed to facilitate the systematic analysis of gait recognition. b) We conduct a comprehensive experimental review of recent advancements in gait recognition, yielding insightful findings. c) We introduce three robust baseline models, while maintaining practicality and commonality, demonstrating significant performance enhancement.

This work stems from our CVPR2023 highlight paper, OpenGait [17], and makes three significant advancements: a) We upgrade the GaitBase to DeepGaitV2 by thoroughly exploring the capability of shallow *vs.* deep models for gait modeling. b) We combine findings from our AAAI2024 publication [19], specifically SkeletonGait and SkeletonGait++, with DeepGaitV2, resulting in a more comprehensive baseline reconstruction study. c) We broaden the experimental scope to include recently released popular gait datasets, such as CCPG [20] and SUSTech1K [21]. By aiming to enhance the practicality of gait recognition, this paper offers a more systematic analysis compared to its conference counterpart.

II. Related Work

This section starts with a literature review, involving the aspects of the gait dataset, input modality, and popular method, providing a snapshot of recent developments in the field. Following this, we offer a brief review of other related works.

A. Gait Recognition Datasets

The large-scale data collection presents an essential premise for contemporary gait recognition research. As exhibited in Table I, several popular gait datasets can be roughly categorized into two groups: the constrained and in-the-wild gait datasets, where the former ones often require subjects to repeatedly walk along fixed routes with simulating real scenarios by introducing dressing and carrying changes, while the latter ones are captured from real-world scenarios naturally covering a wide range of real-world covariates. Even though the constrained gait datasets have imported more and more human-crafted complexities, formulating gait recognition through in-the-wild settings remains a relatively practical and challenging choice according to mainstream experimental results [22]. Before the release of this paper's conference version [17], most existing

[4]We regard GaitBase, proposed by the conference version of this work [17], as an initial exploration of deep ResNet [18] for gait modeling and denote it as DeepGaitV1. Consequently, its subsequent iteration is labeled as DeepGaitV2.

[5]We count 78 gait papers released over the past 5 years, across top computer vision conferences such as CVPR, ICCV, ECCV, BMVC, WACV, ACCV, and related top conferences like AAAI, ACM MM, ICASSP, ICIP, and IJCB.

TABLE I
THE COMPARISON BETWEEN SIX POPULAR GAIT DATASETS.

Environment	Dataset	Train Set		Test Set		Cameras	Variations	Modalities	Venue
		#ID	#Seq	#ID	#Seq				
Constrained	CASIA-B [10]	74	8,140	50	5,500	11	#CV, #CL, #BG	Sil., RGB	ICPR'06
	OU-MVLP [11]	5,153	144,284	5,154	144,412	14	#CV	Sil., Ske.	IPSJ'18
	CCPG [20]	100	8,187	100	8,095	10	#CV, #CL, #BG	Sil., RGB	CVPR'23
	SUSTech1K [21]	200	5,988	850	19,228	12	#CV, #CL, #BG, *etc*	Sil., RGB, Lidar	CVPR'23
In-the-wild	GREW [8]	20,000	102,887	6,000	24,000	882	Real-world	Sil., Ske., Point	ICCV'21
	Gait3D [9]	3,000	18,940	1,000	6,369	39	Real-world	Sil., Ske., Mesh	CVPR'22

#ID and #Seq present the number of identities and sequences. #CV, #CL, and #BG are for camera viewpoints, clothing changes, and bag carrying. Sil. and Ske. respectively refer to the silhouette and skeleton.

works only verify their effectiveness on in-the-lab datasets, *e.g.*, CASIA-B [10] and OU-MVLP [11], thus posing a high risk of vulnerability for practical usage. In this benchmark study, we evaluate several representative SoTA methods on emerging real-world gait datasets thus uncovering overlooked issues. In addition, the effectiveness of our three kinds of baseline models is verified on both constrained and in-the-wild gait datasets, convincingly demonstrating the generality and significance of our findings for gait recognition.

B. Typical Gait Modalities

As shown in Fig. 1 (c), the mainstream gait modalities, such as the binary silhouette, skeleton coordinates, RGB image, human mesh, and body parsing image, are primarily derived from cameras and represented by various data formats. During this process, diverse pretreatment operations [7], [23], [24], [21], [25] and end-to-end learning manners [26], [27] have been utilized to mitigate the influence of gait-irrelevant noises like color, texture, and background cues. Out of this trend, some studies propose novel gait modalities by incorporating emerging sensors such as LiDAR [21] and event cameras [28]. However, these sensors are currently less commonly found in CCTVs, making them temporarily unsuitable for large-scale surveillance applications. Even so, OpenGait supports all of the above gait modalities regardless of the sensor type.

In addition, Fig. 1 (c) further reveals that the binary silhouette and skeleton coordinates act as two of the most prevailing gait modalities in the latest literature. They both explicitly present the human body structural characteristics, *e.g.*, the length, ratio, and movement of human limbs. Silhouettes, especially, have a more discriminative capacity by explicitly presenting the body shape information. When the gait recognition challenge evolves from constrained evaluations to in-the-wild applications, the gait models based on these two modalities have suffered different degrees of performance degradations [17], [29]. Except for recognition accuracy renaissances, this work further explores the cooperativeness and complementarity natures of these two gait modalities for 'comprehensive' gait pattern descriptions, thus proposing the multi-modal SkeletonGait++.

C. Gait Recognition Methods

For summary convenience, we roughly group gait recognition methods into three categories, *i.e.*, the model-based, appearance-based, and multi-modal methods.

Model-based Gait Recognition methods [23], [30], [27] tend to take the estimated underlying structure of the body as input, such as 2D/3D pose and SMPL [31] model. With extremely excluding visual clues, these gait modalities, typically parameterized as coordinates of body joints or customized vectors, are ideally 'clean' against factors like carrying and dressing items. In recent literature, PoseGait [23] used 3D body pose and hand-crafted structural features to overcome the changes in clothing. GaitGraph [30] introduced the graph convolutional network for 2D skeleton-based gait description. GPGait [29] developed a human-oriented transformation and a series of human-oriented descriptors to generate the unified pose representation with multiple features. Generally, the skeleton-based methods often struggle with non-shape-information situations and, as a result, usually perform unsatisfactorily on a variety of gait datasets. On the other hand, several SMPL-based methods have achieved significant advances on indoor OU-MVLP [11], *e.g.*, Li *et al.* [27] fine-tuned a pre-trained human mesh recovery network to construct the end-to-end SMPL-based model and Xu *et al.* [32] proposed an occlusion-aware human mesh model-based method to alleviate the partial occlusion problem within practical gait recognition. However, the competitiveness of this line of work on real-world gait datasets [8], [9] has not yet been verified exactly.

Appearance-based Gait Recognition methods mostly learn gait features from binary silhouettes or RGB images and benefit from informative shape characteristics. With the boom of deep learning, most current appearance-based works focus on spatial feature extraction and temporal modeling. Specifically, GaitSet [7] innovatively regarded the gait sequence as a set and utilized a maximum function to compress frame-level spatial features. Thanks to its simplicity and effectiveness, GaitSet has become one of the most influential gait methods in recent years. GaitPart [15] carefully explored the local details of the input silhouette and modeled the temporal dependencies [15]. GaitGL [16] argued that the spatially global gait representations often neglect the details, and the local region-based descriptors cannot capture the relations among neighboring parts [16]. 3DLocal [33] wanted to extract limb features through 3D local operations at adaptive scales. DyGait [34] proposed to establish spatial-temporal representations of dynamic body parts. In a statistical sense, the appearance-based methods present the focus of the recent research.

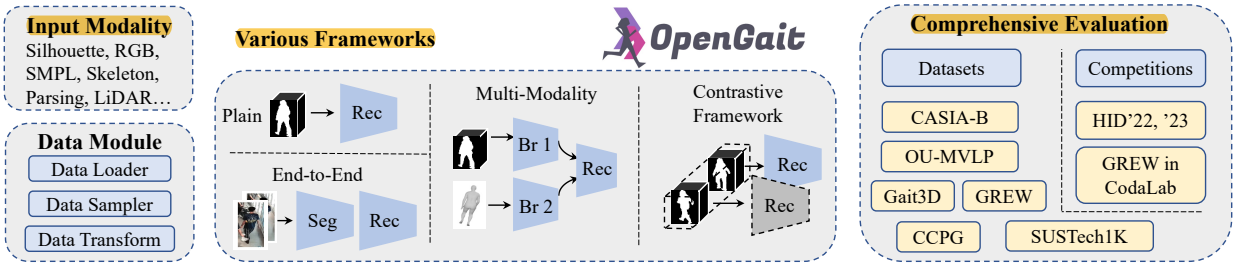


Fig. 2. The design principles of proposed codebase **OpenGait**. Seg is for *segmentation*, Rec is for *recognition*, and Br is for *branch*.

Multi-modal Gait Recognition methods typically incorporate multiple gait modalities as input, reflecting an emerging trend in gait pattern description. In recent literature, SMPLGait [9] exploited 3D information from the SMPL model to enhance the learning of gait appearance features. BiFusion [35] integrated body skeletons and silhouettes to capture the rich spatio-temporal features. ParsingGait [25] regarded the human parsing image as a form of fine-grained segmentation to improve the representational ability of the silhouette branch.

In addition to the above, many successful model-agnostic gait frameworks promote gait research as well, *e.g.*, GaitEdge [26] designed an edge-trainable silhouette to build the end-to-end gait recognition framework, GaitGCI [36] introduced the generative counterfactual intervention to alleviate the over-fitting problem within gait recognition, GaitSSB [37] collected millions of unlabelled gait sequences to pre-train the gait model by contrastive learning, LidarGait [21] built the first large-scale in-the-wild lidar-based gait method to explore precise 3D gait features from point clouds, and UDA [38] developed a novel fine-grained unsupervised domain adaptation method for robust gait recognition.

To act as a comprehensive and extensible gait recognition platform, the proposed OpenGait is compatible with almost all of the aforementioned methods.

D. Other Related Works

Survey on Gait Recognition. While survey papers [39], [40], [22] have offered comprehensive overviews of gait recognition publications, a detailed experimental analysis is still lacking. In contrast, several impactful works in other fields, such as recommendation system [41], metric learning [42], and unsupervised domain adaptation [43], have revisited SoTA experimentally. Recognizing this gap, OpenGait aims to fill this void by not only re-evaluating recent works but also uncovering new insights in the field of gait recognition.

Codebase for Computer Vision. In the computer vision community, having a robust codebase is crucial for advancing research in specific domains. For example, Amos *et al.* [44] proposed OpenFace, a face recognition library that bridges the gap between public face recognition systems and industry-leading private systems. In the field of object detection, a PyTorch toolbox called MMDetection [45] supports almost all popular detection methods, providing a convenient platform for systematic comparison. As gait recognition continues to evolve rapidly, the need for an infrastructure code platform has become increasingly evident.

III. A Glance of OpenGait Platform

Over the past few years, numerous new methods and datasets have emerged for gait recognition. However, the lack of a unified and fair evaluation platform cannot be overlooked. To facilitate academic research, this paper presents a PyTorch-based [46] toolbox, termed OpenGait, as a reasonable and dependable solution to address this issue.

A. Design Principles of OpenGait

As shown in Fig. 2, OpenGait covers the highlights below.

Compatibility with Diverse Gait Modalities. Typical gait modalities include the silhouette image, 2D/3D skeleton, and emerging ones like the SMPL model [27], [9], point cloud [21], and RGB image [24], [26]. While existing open-source repositories mostly support one specific input type, OpenGait is designed to support all of these representations.

Compatibility with Various Frameworks. The landscape of gait recognition is evolving with the emergence of novel frameworks, such as multi-modal [9], end-to-end [24], [27], [26], and contrastive learning [37] based methods. Unlike existing repositories that are specifically tailored to certain models, OpenGait is designed to be flexible and extensible, accommodating all of the above frameworks seamlessly.

Support for Various Evaluation Datasets. OpenGait offers compatibility with commonly used gait datasets, covering a wide range of scenarios and sensors. It seamlessly integrates with constrained datasets like CASIA-B [10] and OU-MVLP [11], as well as in-the-wild datasets GREW [8] and Gait3D [9]. Moreover, OpenGait also supports the latest in-the-lab gait datasets such as CCPG [20] and SUSTech1K [21]. Notably, OpenGait's compatibility has been extended to major competitions like HID [13] and GREW [8], where it has been instrumental in the development of winning solutions.

Support for SoTA. OpenGait encompasses the reproduction of most SoTA methods mentioned in the related work section. The reproduced performances match or surpass the results reported by the original papers. These rich official examples help beginners get started with the code better and faster. Moreover, this also provides a solid platform for fair and thorough comparisons in this benchmark study.

B. Main Modules

We follow the design of most PyTorch deep learning projects and divide OpenGait into three modules, *i.e.*, *data*, *modeling*, and *evaluation*, as shown in Fig. 2.

Data module contains the data loader, sampler, and transform scripts, which are responsible for loading, sampling, and pre-processing the input data flow respectively.

Modeling module is built on top of a base class (`BaseModel`) that pre-defines typical behaviors of the deep model during the training and testing phases, including optimization and inference. The four essential components of current gait recognition methods, namely, *backbone*, *neck*, *head*, and *loss*, can be customized in this class.

Evaluation module is used to assess the performance of the obtained model. Given that different datasets often come with various evaluation protocols, we integrate them into OpenGait to relieve researchers from dealing with these tedious details.

IV. State-of-the-art Methods Revisiting

With the help of OpenGait, we conduct a comprehensive re-evaluation of several SoTA gait methods. Through meticulous ablation studies, we uncover some interesting insights that differ from those presented in the original papers.

A. Experimental Recheck on Representative Methods

A lot of prior works only perform experiments on indoor gait datasets, notably CASIA-B [10] and OU-MVLP [11], with further ablation studies typically limited to CASIA-B [10]. This subsection aims to extend the analysis by conducting additional ablation studies on an in-the-wild dataset, Gait3D [9], to assess their robustness to real-world gait data.

TABLE II
The effect of MGP and Multi-scale HPP in GaitSet [7].

MGP	Multi-scale HPP	CASIA-B			Gait3D	
		NM	BG	CL	R-1	R-5
✔	✘	**95.9**	90.3	**74.2**	44.3	64.7
	✔	95.8	90.4	73.2	44.3	64.4
✘	✘	95.3	**90.5**	74.0	**45.8**	**65.1**
	✔	94.5	89.1	72.3	43.7	63.8

Re-conduct Ablation Study on GaitSet. With taking the silhouettes as input, GaitSet [7] treats the gait sequence as an unordered set and uses a simple maximum pooling function along the temporal dimension, called Set Pooling (SP), to generate the set-level understanding of the entire input video. GaitSet [7] provides insights for many subsequent works thanks to its simplicity and effectiveness. However, we find that the other two important components in GaitSet [7], namely the parallel Multi-layer Global Pipeline (MGP) and pyramid-like Horizontal Pyramid Pooling (HPP) [47], do not work well enough on both the indoor CASIA-B and outdoor Gait3D datasets. Specifically, as shown in Fig. 3 (a), MGP can be regarded as an additional branch that aggregates the hierarchical set-level characteristics. HPP [47] follows the fashion feature pyramid structure aiming to extract multi-scale part-based features. As shown in Table II, if we strip MGP or remove the multi-scale mechanism in HPP from the official GaitSet [7], the obtained model could reach the same or even better performance on both CASIA-B and Gait3D with saving over 80% training weights. This result indicates that the set-level characteristics extracted by bottom convolution blocks may be hard to benefit the final gait representation[6]. Besides, the multi-scale mechanism in HPP also provides no extra discriminative features. The cause may be that the employed statistical pooling functions are too weak to learn extra knowledge from various-scale human body parts.

TABLE III
The effect of FConv in GaitPart [15].

FConv	CASIA-B			Gait3D	
	NM	BG	CL	R-1	R-5
✔	**96.2**	**91.5**	**78.7**	29.2	48.6
✘	95.6	88.4	76.1	**36.2**	**57.0**

Re-conduct Ablation Study on GaitPart. One of the core contributions of GaitPart [15] is to point out the importance of local details with the proposed Focal Convolution (FConv) layer. Fig. 3 (b) shows the receptive field's expansion of the top-layer neuron in the network established by regular *vs.* focal convolution layers. Technically, FConv splits the input feature map into several parts horizontally and then performs a regular convolution over each part separately. As shown in Table III, we get a much higher performance (Rank-1: +7.0%) on Gait3D by degenerating the FConv to the regular convolution layer. This phenomenon exhibits that the extraction of gait features may be seriously affected by merely splitting the feature map in a fixed-window manner, due to the low-quality segmentation of wild data. The shifted window mechanism [48] may offer a solution to this issue.

TABLE IV
The effect of local branch in GaitGL [16].

Local Branch	CASIA-B			Gait3D	
	NM	BG	CL	R-1	R-5
✔	**97.4**	**94.5**	**83.6**	31.4	50.0
✘	97.1	93.7	81.9	**32.2**	**52.5**

Re-conduct Ablation Study on GaitGL. As shown in Fig 3 (c), GaitGL [16] develops the global and local convolution layer, where the local branch can be regarded as the FConv [15] employing 3D convolution while the global branch presents a standard 3D convolution layer. Similar to GaitPart [15], as shown in Table IV, removing the local branch can achieve better performance on the Gait3D dataset.

Re-conduct Ablation Study on SMPLGait. As shown in Fig. 3 (d), SMPLGait [9] consists of two elaborately-designed branches, *i.e.*, silhouette (SLN) and SMPL (3D-STN) branch. They are respectively used for 2D appearance extraction and 3D knowledge learning. SMPLGait [9] takes advantage of the 3D mesh data available in Gait3D [9] and achieves performance gains on the top of the silhouette branch. However,

[6]In the original GaitSet [7], all the obtained partial vectors will be concatenated into single feature vector used for final evaluation. However, the followup works [15], [16], [9] find this concatenation unnecessary and take the average of partial distance as a final metric. Therefore, OpenGait follows this manner and reproduces a better performance than official results.

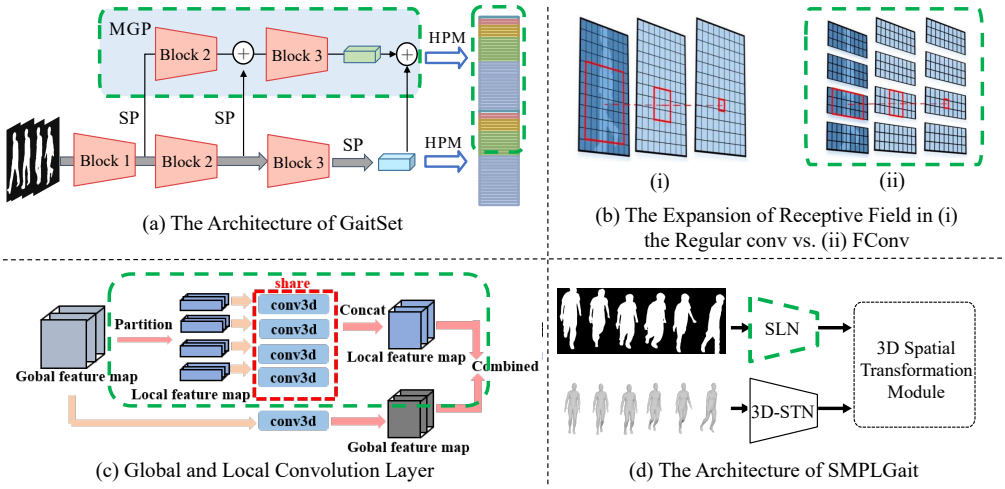


Fig. 3. The key modules of four previous SoTA methods. The modules enclosed by dotted green lines are the ones we remove or replace for comparison. (a): MGP and multi-scale mechanism in HPP are removed [7]. (b): FConvs are replaced with the regular convolution layers [15]. (c): The upper local feature branch is replaced with an independent conv3d that is identical to the lower global branch [16]. (d): SLN branch is replaced with our stronger backbone [17].

TABLE V
THE EFFECT OF SMPL BRANCH IN SMPLGAIT [9].

SMPL branch	Silhouette branch			
	SLN		ResNet9	
	R-1	R-5	R-1	R-5
✔	**46.3**	**64.5**	55.2	**75.7**
✘	42.9	63.9	**56.5**	75.2

SLN and ResNet9 respectively denote the backbone used by SMPLGait [9] and GaitBase [17].

Table V demonstrates that the proposed SMPL branch does not provide obvious benefits when we give the silhouette branch a strong backbone network like GaitBase [17].

In our view, there are three potential reasons causing the failure of the SMPL branch: a) Though the SMPL model is usually visualized as a dense mesh, its feature vector only possesses tens of dimensions that present the relatively sparse characterization of body shape and posture, making it challenging to enhance the fine-grained description of gait patterns. b) Since the SMPL model is not recognition-oriented, purposefully fine-tuning it may be more optimal than directly utilizing it to depict the subtle individual characteristics [27]. c) In the wild, estimating an accurate SMPL model that finely captures body shape and posture from a single RGB camera is still challenging. In a nutshell, introducing 3D geometrical information from the SMPL model to enhance gait representation learning is well worth further exploration.

B. Rethinking and Discussion

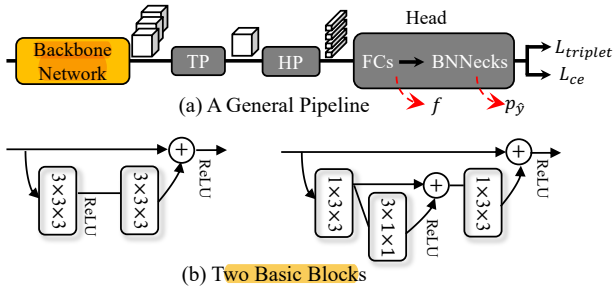
Drawing from the aforementioned findings, we propose that the model structures of previous gait recognition methods may lack robustness. This limitation stems, largely, from the heavy reliance on constrained datasets that struggled with simulating the complexity of outdoor environments. To provide a comprehensive understanding, we elaborate on this point below.

Necessity of Outdoor Evaluation. Previous methods are primarily evaluated on the indoor CASIA-B [10] and OU-MVLP [11]. We argue that this practice suffers from three significant drawbacks: a) Indoor settings. The walking videos are captured by a camera array and subjects are requested to follow a particular route. It makes the data source significantly different from real-world scenarios. b) Simple background. The simple laboratory background cannot reflect the complex changes of wild scenes. c) Outdated processing methods. The RGB videos are processed by the background subtraction algorithm significantly distinct from the deep-learning-based segmentation algorithms used for current applications.

Recently, there has been a noticeable emergence of large-scale real-world gait datasets [8], [9] aimed at advancing gait recognition from controlled laboratory settings to real scenarios. However, despite these advancements, many of the lately-published works [49], [50], [51] continue to heavily rely on constrained setups. To advance the development of gait recognition systems, we advocate for increased attention and utilization of outdoor gait datasets.

Necessity of Sufficient Ablative Experiments. The ablation study is recognized as the primary means of evaluating the effectiveness of individual components within a proposed method. However, we consider that the conclusions drawn from ablative experiments performed merely on CASIA-B [10] may lead to limited practical applicability. The primary reasons are two-fold: a) CASIA-B contains only 50 subjects for evaluation. Such few testing participants make the results vulnerable to noisy factors. b) In contemporary applications, pedestrian segmentation predominantly relies on deep learning models, diverging from the outdated background subtraction algorithm employed by CASIA-B. In principle, performing a comprehensive ablation study across diverse large-scale datasets can yield a more robust configuration of hyperparameters and the design of model architectures.

Necessity of A Strong Backbone. The quality of a model largely hinges on the capability of its backbone, and a sub-



(a) A General Pipeline

(b) Two Basic Blocks

Model	DeepGaitV2	SkeletonGait	SkeletonGait++
Category	Appearance	Model	Multi-modal
Input	Silhouette	Skeleton Map	Combination
Shape Feature	√	×	√
Structural Feature	√	√	√

(c) Three Strong Baselines

Fig. 4. (a) Overall architecture. The TP and HP respectively denote Temporal and Horizontal Pooling. The Head comprises the separate fully-connected and BNNeck [55] layers. (b-d) Employed basic blocks, including the (b) 2D, (c) 3D and (d) Pseudo 3D residual convolution blocks [18], [56], [57].

optimal backbone may inflate the effectiveness of additional modules. The evolution of CNN architectures has progressed from shallow to deep, giving rise to the emergence of excellent backbone networks such as AlexNet [52], VGG-16 [53], and ResNet [18]. However, previous works [7], [54], [9] in gait recognition have predominantly relied on plain convolutional neural networks, composed of several basic convolution layers. As gait recognition research progresses towards applications and large-scale real-world datasets [8], [9] become more accessible, the necessity for a robust and powerful backbone network becomes apparent and essential for achieving accurate and dependable gait recognition results.

V. Gait Recognition Baselines Reconstruction

Building on the insights outlined earlier, this section establishes a series of baseline models aimed at uncovering the practical principles for contemporary in-the-wild gait recognition. Our investigation encompasses appearance-based, model-based, and multi-modal methodologies, enriching the comprehensiveness of this benchmark study. To ensure broader applicability, we prioritize the development of structurally simple yet experimentally powerful and empirically robust architectures to lay the foundations for follow-ups. To this end, we undertake three-fold efforts:

- **A General Pipeline**. To ensure the baseline models' typicality and universality, we adopt a widely recognized pipeline as shown in Fig. 4 (a).
- **Two Basic Blocks**. To make it simple and scaleable, we employ only two blocks illustrated in Fig. 4 (b), *i.e.*, the 3D residual unit and its pseudo counterpart [56], [57] to build the network backbone.
- **Three Strong Baselines**. As depicted in Fig. 4 (c), the resulting DeepGaitV2, SkeletonGait, and SkeletonGait++ share an identical pipeline, each respectively taking the silhouette image, skeletal coordinates, and their combination as input, accordingly representing the appearance-based, model-based, and multi-modal baseline models in this benchmark study. Notably, we align the data format of silhouette and skeleton by drawing the latter's coordinates into an imagery heatmap, facilitating flexible comparison and fusion. Additional details will be provided in Sec. V-C.

The subsequent content of this section begins by introducing the used pipeline. Following this, we present DeepGaitV2 based on a large-scale experimental investigation on both in-the-lab and in-the-wild gait datasets. For brevity, implementation details are temporarily omitted, and they will be comprehensively presented in Sec. VI. We then proceed to introduce SkeletonGait and SkeletonGait++ in detail.

A. Pipeline

As shown in Fig. 4 (a), the backbone transforms each input frame into a 3D feature map with the height, width, and channel dimensions. Then, a temporal pooling (TP) module aggregates the obtained feature map sequence by performing the maximization along the temporal dimension, outputting a global understanding. Next, the obtained feature map is horizontally divided into several parts, and each part is pooled into a feature vector, according to the popular horizontal pooling (HP) [47]. As a result, we get several feature vectors and use separate fully connected layers (FCs) to map them into the metric space, *i.e.*, generating f. Since these part vectors will be processed independently, the following formulation loosely treats f as a single feature vector for brevity.

Given a mini-batch $\mathbb{B}$, we can get the collection of positive and negative pairs over this batch, namely $\mathbb{S}^+$ and $\mathbb{S}^-$. For each sample pair $(\overrightarrow{f}, \overleftarrow{f})$, we utilize the Euclidean norm to measure their distance, *i.e.*, $d = \left\| \overrightarrow{f} - \overleftarrow{f} \right\|$. Then the triplet loss is employed to drive the identity learning process:

$$L_{triplet} = \frac{1}{\mathcal{R}(\mathbb{S}^+)} \frac{1}{\mathcal{R}(\mathbb{S}^-)} \sum_{i \in \mathbb{S}^+} \sum_{j \in \mathbb{S}^-} \lfloor d_i - d_j + m \rfloor_+ \quad (1)$$

where $\mathcal{R}(\cdot)$ measures the collection size and m presents the loss margin. Furthermore, we use BNNecks [55] to optimize the identity distribution by estimating the probability $p_{\hat{y}}$ for class $\hat{y}$. Let the ground-truth label be y and class number be Y, then the cross-entropy loss can be formulated as:

$$\begin{aligned} p_y &= p_{\hat{y}} \quad \text{if} \quad \hat{y} = y \quad \text{else} \quad 1 - p_{\hat{y}} \\ L_{ce} &= -\frac{1}{\mathcal{R}(\mathbb{B})} \sum^{\mathbb{B}} \frac{1}{\mathrm{Y}} \sum_{y}^{\mathrm{Y}} \log(p_y) \end{aligned} \quad (2)$$

In the end, the overall loss is $L = L_{triplet} + L_{ce}$.

B. Appearance-based DeepGaitV2

Gait data is typically presented as sparse input, such the silhouette image with a size of 64×44, which might create the

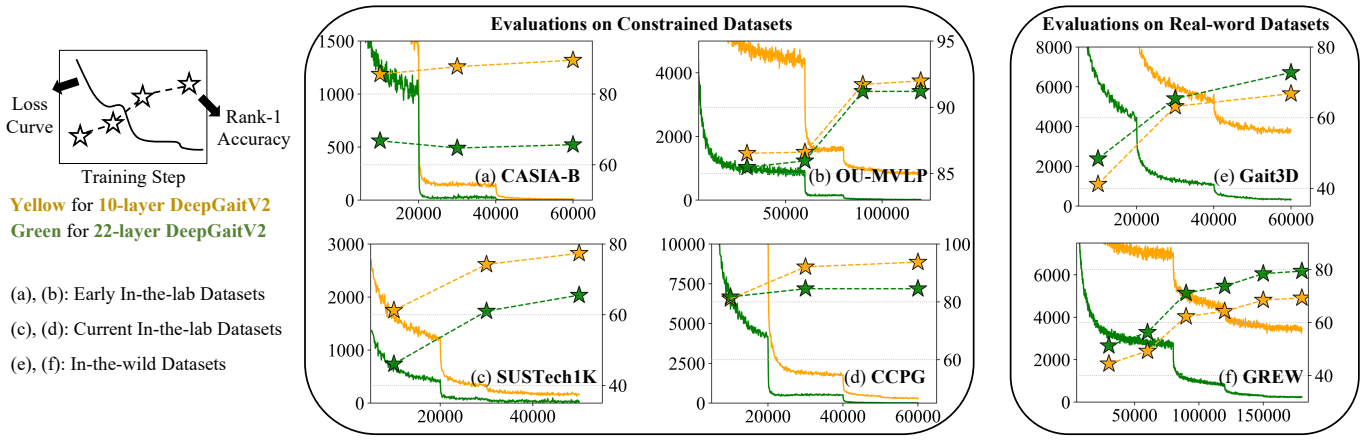


Fig. 5. The DeepGaitV2 series meets the over-fitting cases on (a) CASIA-B and (b) OU-MVLP, with the network depth increasing. The loss number presents the count of triplets that cause non-zero loss in the training batch, directly reflecting the network's convergence state.

TABLE VI
ARCHITECTURES OF DEEPGAITV2 SERIES.

Layer	Output Size	DeepGaitV2 Block	Network Depth D = 10	14	22	30
Conv 0	(T,C,64,44)	3×3, stride 1				
Stage 1	(T,C,64,44)	$\begin{bmatrix} 3\times3, C \\ 3\times3, C \end{bmatrix}$	×1	×1	×1	×1
Stage 2	(T,2C,32,22)	$\begin{bmatrix} 3\times3\times3, 2C \\ 3\times3\times3, 2C \end{bmatrix}$	×1	×2	×4	×4
Stage 3	(T,4C,16,11)	$\begin{bmatrix} 3\times3\times3, 4C \\ 3\times3\times3, 4C \end{bmatrix}$	×1	×2	×4	×8
Stage 4	(T,8C,16,11)	$\begin{bmatrix} 3\times3\times3, 8C \\ 3\times3\times3, 8C \end{bmatrix}$	×1	×1	×1	×1
TP	(1,8C,16,11)	Temporal Pooling				
HP	(1,8C,16,1)	Horizontal Pooling				
Head	(1,8C,16,1)	Separate FCs and BNNecks				

Down-sampling is performed by Stage 2 and 3 with a stride of 2. T and C respectively denote the input sequence length and arbitrary channel number.

TABLE VII
THE RANK-1 ACCURACY OF DEEPGAITV2 WITH VARIOUS DEPTHS.

Method	Depth	In-the-lab CASIA-B	In-the-lab OU-MVLP	In-the-wild Gait3D	In-the-wild GREW
GaitSet [7]	<10	87.1	87.1	36.7	48.4
GaitPart [15]	<10	88.5	88.7	28.2	47.6
GaitGL [16]	<10	**91.9**	**92.0**	29.7	47.3
GaitBase [17]	10	89.4	90.8	60.1	60.1
DeepGaitV2 (Shallow to Deep)	10	89.6	**92.0**	66.8	69.3
	14	85.5↓	**92.0**	70.8↑	75.7↑
	22	75.3↓	91.9↓	**72.8**↑	79.4↑
	30	65.7↓	91.2↓	71.7↓	**79.5**↑

Here the rank-1 accuracy on CASIA-B is the average over its three conditions.

impression that gait recognition is an 'easy' task that shallow networks can handle, as researchers used to do. In the context of developing baseline models, we agree that the gait recognition defined by constrained gait datasets, such as the CASIA-B [10], OU-MVLP [11], and the latest SUSTech1K [21] and CCPG [20] datasets, can indeed be considered as relatively simple cases that shallow gait models can deal with. This is because they merely contain limited covariates manipulated by researchers, whereas gait models may circumvent human-made challenges by focusing on evidently unchanged cues rather than subtle walking patterns [20]. Therefore, we assume that deep gait models may easily over-fit on in-the-lab datasets. To validate this hypothesis, we build DeepGaitV2 series.

As shown in Table VI, the backbone of DeepGaitV2 is initialized as a stack of five layers, *i.e.*, the initial Conv 0 and the following Stage 1 to 4. The number of residual blocks within each stage, denoted as D, can scale the network depth. For instance, if D is set to [1,1,1,1], the network depth should be 2×(1+1+1+1)+2=10, where the last +2 implies the initial Conv 0 and final head layers. As a result, Table VI presents a serial of DeepGaitV2 with a depth ranging from 10 to 30. Note we instantiate Stage 1 as pure 2D residual block [18] for computational efficiency.

With deepening the backbone, as illustrated in Table VII, there exists a noticeable distinction in performance trends between the constrained and real-world evaluations. Specifically, applications on Gait3D [9] and GREW [8] exhibit substantial benefits from properly deep networks, while contrasting outcomes are observed on CASIA-B [10] and OU-MVLP [11]. The analysis shown in Fig. 5 (a) and (b) uncover the over-fitting issues on CASIA-B and OU-MVLP, *i.e.*, the 22-layer DeepGaitV2 converges better but performs worse.

Moreover, we find similar over-fitting phenomena on the latest constrained gait datasets, namely SUSTech1K [21] and CCPG [20]. Despite the introduction of new human-controlled challenges such as poor illumination, complex occlusion, and clothing changes, as shown in Fig. 5 (c) and (d), the 22-layer DeepGaitV2 continues to exhibit the over-fitting behavior.

Combining the above observations, this paper advocates for the adoption of deep models as the fundamental architecture for analyzing in-the-wild gait datasets in future research endeavors. According to Table VII, we propose utilizing the 10-layer and 22-layer DeepGaitV2 models as baseline models for constrained and real-world evaluations, respectively. Beyond deepening DeepGaitV2, we also develop a light version for better accuracy-speed balance. Specifically, we employ pseudo 3D block [57] shown in Fig. 4 (b) to reconstruct DeepGaitV2

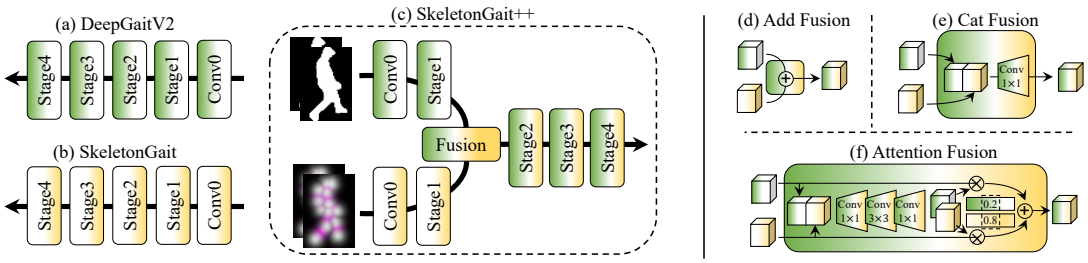


Fig. 6. The network architectures of DeepGaitV2 *v.s.* SkeletonGait *v.s.* SkeletonGait++. The 'head' part is ignored for brevity.

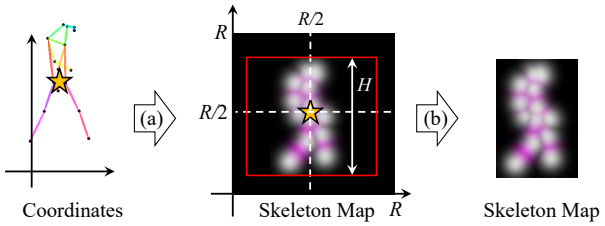


Fig. 7. The pipeline of skeleton map generation. (a) Center-normalization, scale-normalization, and skeleton rendering. (b) Subject-centered cropping.

shown in Table VI. This slight change can save about 59% training weights and 57% computation costs[7] with remaining competitive performances on both Gait3D [9] (+1.6%) and GREW [8] (-1.7%). In the following, we use the pseudo-3D DeepGaitV2 as the default unless otherwise stated.

C. Model-based SkeletonGait

Previous skeleton-based methods [23], [30], [27] typically model the coordinates of body joints as non-grid graphs using graph neural networks. In this benchmark study, we introduce a novel skeleton-based gait representation called the skeleton map, drawing inspirations from related works [58], [59], [60]. As illustrated in Fig. 4 (c), our skeleton map represents the coordinates of joints as a heatmap, *i.e.*, transforming the skeleton into a silhouette-like image without exact body shapes. This manner offers three significant advantages:

- Aligning the skeleton and silhouette data formats enables a seamless extension of DeepGaitV2 to skeleton-based methodology, eliminating the need for additional efforts to develop graph-specific gait models. This alignment ensures consistency in the network architecture between silhouette-based and skeleton-based gait recognition, making further comparisons intuitive and fair.
- The skeleton map, resembling a silhouette without body shapes, allows for an intuitive controlled experiment to observe the distinct roles of body structure *vs.* shape features used in describing gait patterns for recognition.
- Serving as an imagery input, the skeleton map can be seamlessly integrated into image-based gait models, particularly at the bottom convolution stages, facilitating the further exploration of multi-modal baseline models in this benchmark study.

As a result, SkeletonGait is developed by replacing the input of DeepGaitV2 from the silhouette to the skeleton map, as shown in Fig. 6 (b). The only architectural modification involves adjusting the input channel, as the silhouette is single-channel while the skeleton map is double-channel. Specifically, we generate the skeleton map through the following steps.

Given the coordinates of human joints (x_k, y_k, c_k), where (x_k, y_k) and c_k respectively present the location and confidence score of the k-th joint with $k \in \{1, ..., K\}$ (K denotes the number of body joints). Firstly, considering the absolute coordinates of joints relative to the original image contain much gait-unrelated information like the walking trajectory and filming distance, we introduce the pre-treatments of center- and scale-normalization to align raw coordinates:

$$\begin{aligned} x_k &= x_k - x_{\text{core}} + R/2 \\ y_k &= y_k - y_{\text{core}} + R/2 \\ x_k &= \frac{x_k - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \times H \\ y_k &= \frac{y_k - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \times H \end{aligned} \tag{3}$$

where $(x_{\text{core}}, y_{\text{core}}) = (\frac{x_{11}+x_{12}}{2}, \frac{y_{11}+y_{12}}{2})$ presents the center point of two hips (11-th and 12-th human joints, their center can be regarded as the barycenter of the human body), and $(y_{\text{max}}, y_{\text{min}})$ denotes the maximum and minimum heights of human joints. In this way, we move the barycenter of the human body to $(R/2, R/2)$ and normalize the body height to H, as shown in Fig. 7 (a).

Typically, the height of the human body is expected to exceed its width. As a result, the normalized coordinates of human joints, as defined in Eq. (3), should fall within the range of $H \times H$. But in practice, the pose estimator is imperfect and may produce some outlier joints outside the $H \times H$ scope. To address these out-of-range cases, the resolution of the skeleton map, denoted as R, should be larger than H, ensuring coverage of all the coordinates. In our experiments, let R be $2H$ is enough for all the employed datasets.

As illustrated in Fig. 7 (a), the skeleton map is initialized as a blank image with a size of $R \times R$. Then we draw it based on the normalized coordinates of human joints. Inspired by [58], we generate the joint map $\boldsymbol{J}$ by composing K Gaussian maps, where each Gaussian map is centered at a specific joint

[7]# param: 27.5 *vs.* 11.1 MB, and FLOPs: 6.8 *vs.* 2.9 GFLOPs per frame. Here only consider the backbone, and the same below unless otherwise stated.

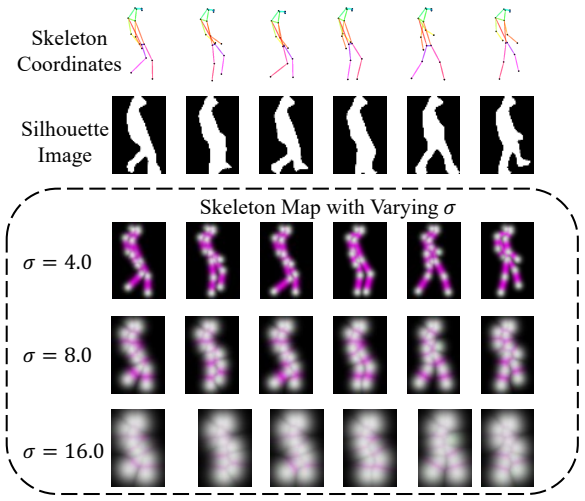


Fig. 8. More examples of the skeleton coordinates *v.s.* silhouette images *v.s.* skeleton maps.

position and contributes to all the $R \times R$ pixels:

$$\boldsymbol{J}_{(i,j)} = \sum_{k}^{K} e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2\sigma^2}} \times c_k \tag{4}$$

where $\boldsymbol{J}_{(i,j)}$ presents the value of a certain point from $\{(i,j)|i,j \in \{1,...,R\}\}$, and σ is a hyper-parameter controlling the variance of Gaussian maps.

Similarly, we can also create a limb map $\boldsymbol{L}$:

$$\boldsymbol{L}_{(i,j)} = \sum_{n}^{N} e^{-\frac{\mathcal{D}((i,j),\mathcal{S}[n^-,n^+])^2}{2\sigma^2}} \times \min(c_{n^-}, c_{n^+}) \tag{5}$$

where $\mathcal{S}[n^-, n^+]$ presents the n-th limb determined by n^--th and n^+-th joints with $n^-, n^+ \in \{1,...,K\}$. The function $\mathcal{D}((i,j),\mathcal{S}[n^-,n^+])$ measures the Euclidean distance from the point (i,j) to n-th limb, where $n \in \{1,...,N\}$ and N denotes the count of limbs.

Next, the skeleton map is obtained by stacking $\boldsymbol{J}$ and $\boldsymbol{L}$ and thus has a size of $2 \times R \times R$. Notably, for the convenience of visualization, we repeat the last channel of all the skeleton maps shown in this paper to display the visual three-channel images with the size of $3 \times R \times R$.

As shown in Fig. 7 (b), we employ subject-centered cropping to remove the blank regions, thus reducing the redundancy in skeleton maps. In practice, the vertical range is determined by the minimum and maximum heights of pixels which possess non-zero values. Meanwhile, the horizontal cropping range spans from $\frac{R-H}{2}$ to $\frac{R+H}{2}$. In this way, we remove extraneous areas outside the desired gait region, ensuring a more concise and compact skeleton map. Lastly, to align with the input size required by downstream gait models, the cropped skeleton maps are resized to $2 \times 64 \times 64$ and further cropped by the widely-used double-side cutting strategy.

As a result, Fig. 8 exhibits some examples of the used skeleton maps with varying σ. As we can see, a smaller σ produces a visually thinner skeleton map, whereas excessively large σ may lead to visual ambiguity.

Compared with other skeleton-to-image approaches [58], [59], [60], our skeleton map introduces the following gait-oriented enhancements:

- **Cleanness**. The implementation of center-normalization effectively eliminates identity-unrelated noise present in raw skeleton coordinates, *i.e.*, the walking trajectory and camera distance information.
- **Discriminability**. Preceding methods tend to directly resize the obtained images of varying sizes into a predetermined fixed size, inevitably resulting in the loss of body ratio information. Conversely, the scale-normalization and subject-centered cropping techniques outlined in this paper ensure that the skeleton map preserves the authenticity of the length and ratio of human limbs.
- **Compactness**. All the joints and limbs are drawn within a single map, optimizing the efficiency of the modeling process, as opposed to a stack of separate maps.

D. Multi-modal SkeletonGait++

To integrate the superiority of silhouette and skeleton map, as shown in Fig. 6 (c), SkeletonGait++ provides a fusion-based two-stream architecture involving the silhouette and skeleton branch. These two branches respectively share the same network architectures with DeepGaitV2 and SkeletonGait at early stages, such as Conv 0 and Stage 1. Then, a fusion module is responsible for aggregating these two feature sequences in a frame-by-frame manner. This benchmark study considers three kinds of fusion mechanisms:

- **Add Fusion.** The feature maps from the silhouette and skeleton branch are combined using an element-wise addition operation, as demonstrated in Fig. 6 (d).
- **Concatenate Fusion.** The feature maps from the silhouette and skeleton branch are first concatenated along the channel dimension, and then transformed by a plain 1×1 convolution layer, as demonstrated in Fig. 6 (e).
- **Attention Fusion.** The feature maps from the silhouette and skeleton branch are first concatenated along the channel dimension, and then transformed by a small network to form a cross-branch understanding. Here the small network is composed of a squeezing 1×1, a plain 3×3, and an expansion 1×1 convolution layer. As shown in Fig. 6 (e), a softmax layer is next employed to assign element-wise attention scores respectively for the silhouette and skeleton branch. Lastly, an element-wise weighted-sum operation is used to generate the output.

Next, the Stage 3 and 4 possess the same network architectures as the SkeletonGait. Moreover, we also consider the fusion location. Fig. 6 (c) exhibits the low-level fusion case. Another high-level fusion model aggregates the features before Stage 4, with additional Stage 2 and 3 respectively being inserted into the silhouette and skeleton branch.

VI. More Experiments

For a convincing elaboration, the key statistics of employed datasets have been presented in Table I, and some experiments have been briefly described in the previous section along

TABLE VIII
RECOGNITION RESULTS ON THE OUMVLP [11], GAIT3D [9], AND GREW [8] DATASETS.

Category	Method	Venue	Testing Datasets								
			OU-MVLP [11]	Gait3D [9]				GREW [8]			
			rank-1	rank-1	rank-5	mAP	mINP	rank-1	rank-5	rank-10	rank-20
Silhouette-based	GaitSet [7]	AAAI'19	87.1	36.7	58.3	30.0	17.3	46.3	63.6	70.3	-
	GaitPart [15]	CVPR'20	88.5	28.2	47.6	21.6	12.4	44.0	60.7	67.3	-
	GaitGL [16]	ICCV'21	89.7	29.7	48.5	22.3	13.6	47.3		-	
	DANet [61]	CVPR'23	90.7	48.0	69.7	-	-			-	
	GaitBase [17]	CVPR'23	90.8	60.1		-		64.6		-	
	GaitSSB [62]	T-PAMI'23	91.8	63.6		-		61.7		-	
	QAGait [63]	AAAI'24	-	67.0	81.5	56.5	-	59.1	74.0	79.2	83.2
	DeepGaitV2	Ours	**91.9**	**74.4**	**88.0**	**65.8**	**39.2**	**77.7**	**88.9**	**91.8**	**93.0**
Skeleton-based	GaitGraph2 [30]	CVPRW'22	62.1	11.1		-		33.5		-	
	Gait-TR [64]	ES'23	56.2	6.6		-		54.5		-	
	GPGait [29]	ICCV'23	60.5	22.5		-		53.6		-	
	SkeletonGait	Ours	67.4	38.1	56.7	28.9	16.1	77.4	87.9	91.0	93.2
Multi-modal	SMPLGait [9]	CVPR'22	-	46.3	64.5	37.2	22.2			-	
	GaitRef [65]	IJCB'23	90.2	49.0	49.3	40.7	25.3	53.0	67.9	73.0	77.5
	ParsingGait [25]	MM'23	-	76.2	-	68.2	-			-	
	HybridGait [66]	AAAI'24	-	53.3	72.0	43.3	26.7			-	
	SkeletonGait++	Ours	-	**77.6**	**89.4**	**70.3**	**42.6**	**85.8**	**92.6**	**94.3**	**95.5**

The highest results are in **bold** *for silhouette-based methods and* underlined *for skeleton-based methods. Regarding multi-modal methods, the highest results are in both* **bold and underlined**. *The same annotation is applied in the following tables.*

TABLE IX
SOME IMPLEMENTATION DETAILS.

DataSet	Batch Size	Milestones	Total Steps
CASIA-B [10]	(8, 16)	(20k, 40k, 50k)	60k
OU-MVLP [11]	(32, 8)	(60k, 80k, 100k)	120k
CCPG [20]	(8, 16)	(20k, 40K, 50k)	60k
SUSTech1K [21]	(8, 8)	(20k, 30k, 40k)	50k
Gait3D [9]	(32, 4)	(20k, 40K, 50k)	60k
GREW [8]	(32, 4)	(80k, 120k, 150k)	180k

The batch size $(8, 16)$ *indicates* 8 *subjects and* 16 *sequences per subject.*

with exploring DeepGaitV2. In this section, we provide the implementation details for completeness, and then exhibit the performance comparison and ablation study.

Implementation Details. Table IX displays the main hyper-parameters of our experiments. Unless otherwise specified, a) The silhouettes are aligned by the normalization strategy used in [11] and resized to 64×44. The spatial augmentation strategy outlined by [17] is adopted. b) Different datasets often employ various skeleton data formats, such as COCO 18 for OU-MVLP and BODY 25 for CCPG. To enhance flexibility, our implementation standardized these various formats to COCO 17 uniformly. c) At the test phase, the entire gait sequence is fed into the model directly. As for the training stage, the data sampler collects a fixed-length segment of 30 frames as input. d) The margin m in Eq. (1) is set to 0.2. For the cross-entropy loss shown in Eq. (2), a soft classification trick [67] is employed to encourage the model to be less confident on the training set. e) The SGD optimizer with an initial learning rate of 0.1 and weight decay of 0.0005 is utilized. f) For DeepGaitV2, the C controlling the channel number in Table VI is set to 64 as default. For SkeletonGait, the σ controlling the variance in Eq. (4) and (5) is set to 8.0 as default. For SkeletonGait++, the fusion strategy is instantiated as the attention fusion at the low-level stage as depicted in Fig. 6 (c) and (f). g) All the experiments strictly follow the official evaluation protocols. h) All the implementations are based on our OpenGait. All the codes are publicly available at https://github.com/ShiqiYu/OpenGait.

A. Comparison Around DeepGaitV2

Except for the degradation observed on CASIA-B [10] (Table VII), DeepGaitV2 outperforms other silhouette-based SoTA methods across all other datasets, as demonstrated in Table VIII, X, and XI. Particularly noteworthy is its performance on real-world large-scale gait datasets such as Gait3D [9] and GREW [8], where DeepGaitV2 exhibits more substantial improvements compared to its performance on constrained datasets like OU-MVLP [11], SUSTech1K [21], and CCPG [20]. This alignment with the primary objective of our benchmark study underscores its role in advancing gait recognition toward enhanced practicality.

Moreover, Table VII illustrates that the performance advantage of DeepGaitV2 primarily arises from its depth, *i.e.*, the larger size, and the higher rank-1 accuracy, before reaching the marginal effect. While this observation may seem intuitive for vision tasks, it has been often overlooked in previous gait studies, as discussed in Sec. V-B. Therefore, this work emphasizes the importance of adopting powerful network architectures as the backbone for future deep gait model designs.

B. Comparison Around SkeletonGait

As shown in Table VIII, X, and XI, SkeletonGait outperforms the latest skeleton-based methods by breakthrough improvements in most cases. It gains $+5.3\%$, $+15.6\%$, $+22.9\%$, 29.2%, and $+17.4\%$ rank-1 accuracy on the OU-MVLP [11],

TABLE X
EVALUATION WITH VARIOUS CONDITIONS ON SUSTECH1K [21].

Input	Method	Venue	Probe Sequence (R-1)								Overall	
			Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	R-1	R-5
Silhouette-based	GaitSet [7]	AAAI'19	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
	GaitPart [15]	CVPR'19	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
	GaitGL [16]	ICCV'21	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
	GaitBase [17]	CVPR'23	81.5	77.5	**49.6**	75.8	75.5	76.7	**81.4**	25.9	76.1	89.4
	DeepGaitV2	Ours	**87.4**	**84.1**	53.4	**81.3**	**86.1**	**84.8**	**88.5**	**28.8**	**82.3**	**92.5**
Skeleton-based	GaitGraph2 [30]	CVPRW'22	22.2	18.2	6.8	18.6	13.4	19.2	27.3	16.4	18.6	40.2
	Gait-TR [64]	ES'23	33.3	31.5	21.0	30.4	22.7	34.6	44.9	23.5	30.8	56.0
	MSGG [35]	MTA'23	67.11	66.16	35.92	63.31	61.58	58.07	66.59	17.88	33.8	-
	SkeletonGait	Ours	67.9	63.5	36.5	61.6	58.1	67.2	79.1	50.1	63.0	83.5
Multi-modal	BiFusion [35]	MTA'23	69.8	62.3	45.4	60.9	54.3	63.5	77.8	33.7	62.1	83.4
	ParsingGait [25]	MM'23	82.7	73.3	30.4	76.4	70.8	74.6	84.3	**50.5**	75.0	90.7
	SkeletonGait++	Ours	**89.1**	**87.2**	**55.3**	**85.3**	**87.3**	**87.6**	**91.3**	47.9	**85.6**	**95.0**

TABLE XI
EVALUATION WITH VARIOUS CONDITIONS ON CCPG [20].

Input	Model	Venue	Gait Evaluation Protocol					ReID Evaluation Protocol				
			CL	UP	DN	BG	Mean	CL	UP	DN	BG	Mean
Silhouette-based	GaitSet [7]	AAAI'19	60.2	65.2	65.1	68.5	64.8	77.5	85.0	82.9	87.5	83.2
	GaitPart [15]	CVPR'20	64.3	67.8	68.6	71.7	68.1	79.2	85.3	86.5	88.0	84.8
	AUG-OGBase [20]	CVPR'23	52.1	57.3	60.1	63.3	58.2	70.2	76.9	80.4	83.4	77.7
	GaitBase [17]	CVPR'23	71.6	75.0	76.8	78.6	75.5	88.5	92.7	93.4	93.2	92.0
	DeepGaitV2	Ours	**78.6**	**84.8**	**80.7**	**89.2**	**83.3**	**90.5**	**96.3**	**91.4**	**96.7**	**93.7**
Skeleton-based	GaitGraph2 [30]	CVPRW'22	5.0	5.3	5.8	6.2	5.1	5.0	5.7	7.3	8.8	6.7
	Gait-TR [64]	ES'23	15.7	18.3	18.5	17.5	17.5	24.3	28.7	31.1	28.1	28.1
	MSGG [35]	MTA'23	29.0	34.5	37.1	33.3	33.5	43.1	52.9	57.4	49.9	50.8
	SkeletonGait	Ours	40.4	48.5	53.0	61.7	50.9	52.4	65.4	72.8	80.9	67.9
Multi-modal	BiFusion [35]	MTA'23	62.6	67.6	66.3	66.0	65.6	77.5	84.8	84.8	82.9	82.5
	ParsingGait [25]	MM'23	55.3	58.9	64.0	66.7	61.2	73.5	78.4	85.2	87.0	81.0
	SkeletonGait++	Ours	**79.1**	**83.9**	**81.7**	**89.9**	**83.7**	**90.2**	**95.0**	**92.9**	**96.9**	**93.8**

Gait3D [9], GREW [8], SUSTech1K [21], and CCPG [20] datasets, respectively.

To exclude the potential positive influence brought by the model size of SkeletonGait, we reduce its channels by half, thus making its model size nearly identical to that of GPGait [29], *i.e.*, 2.85 v.s. 2.78M. After that, SkeletonGait reached the rank-1 accuracy of 33.2% and 70.9% on Gait3D [9] and GREW [8], still maintaining a higher result than prior skeleton-based methods shown in Fig. VIII.

To demonstrate the robustness of SkeletonGait to different upstream pose estimators, we conduct experiments using both the officially provided AlphaPose and OpenPose data on OU-MVLP [11], resulting in a rank-1 accuracy of 67.4% and 65.9%, respectively. These two results consistently surpass other SoTA skeleton-based methods, revealing the robustness of SkeletonGait. Table VIII presents the higher one.

As discussed in Sec. V-C, the skeleton map can be perceived as a silhouette devoid of body shape information. Through a detailed comparison of the performance between SkeletonGait and DeepGaitV2, we obtain the following thoughts:

- **Importance of Structural Features**. Structural features play a more important role than previously shown. They contribute to over 50% of the overall performance, as indicated by the rank-1 accuracy ratios between SkeletonGait and DeepGaitV2.
- **Superiority of Skeleton Data**. In situations where silhouette data might be less reliable, such as the night

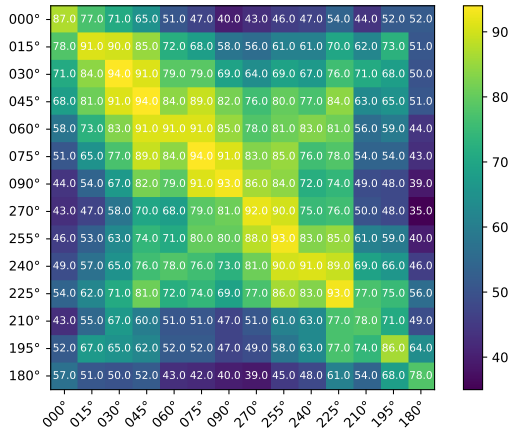


Fig. 9. The rank-1 accuracy of SkeletonGait on OU-MVLP [11] over the probe-gallery view pairs.

TABLE XII
ABLATION STUDY ON GAIT3D [9].

(a) DeepGaitV2

Condition C in Table VI	Rank-1	Param.	GFlops
C=32	67.9	2.8M	0.7G
C=64	74.4	11.1M	2.9G
C=128	**75.0**	44.4M	11.4G

(b) SkeletonGait

Condition σ in Eq. (4)	Performance Rank-1	mAP
$\sigma = 4.0$	37.5	28.5
$\sigma = 8.0$	**38.1**	**28.9**
$\sigma = 16.0$	36.0	26.9

(c) SkeletonGait++

Condition Fusion in Fig. 6	Low-Level Rank-1	Low-Level mAP	High-Level Rank-1	High-Level mAP
Add	76.5	69.6	76.2	69.5
Cat	76.7	69.7	42.2	69.4
Attention	77.6	**70.3**	**78.2**	70.2

The highest performance is in **bold**.

case of SUSTech1K in Table X, SkeletonGait exhibits a considerable performance advantage over DeepGaitV2, highlighting the potential of skeleton data in such cases.

- **Cross-view Challenge**. The results on the diagonal and anti-diagonal of Fig. 9 respectively present the performance of SkeletonGait in identical- and nearly symmetric-view cases, where other regions reflect cross-view conditions. As we can see, the cross-view scenarios remain a major challenge for SkeletonGait.
- **Concerns on GREW**. The GREW dataset is widely acknowledged as the most challenging gait dataset due to its largest scale and real-world settings. However, SkeletonGait achieves a comparable performance compared to DeepGaitV2 on GREW, rather than on other datasets, as shown in Table VIII. Considering SkeletonGait works well on the cross-limited-view cases as shown in Fig. 9, we assume that GREW may contain not enough cross-view covariates. Further investigations are warranted.

C. Comparison Around SkeletonGait++

As demonstrated in Table VIII, X, and XI, SkeletonGait++ achieves a new SoTA with notable improvements compared to other multi-modal methods. Specifically, it gains +1.4%, +21.2%, +9.5%, and 23.2% rank-1 accuracy on the Gait3D [9], GREW [8], SUSTech1K [21], and CCPG [20] datasets, respectively. Here the result missing on OU-MVLP [11] is due to the lack of frame-by-frame alignment of skeleton and silhouette inputs for this dataset.

Compared to DeepGaitV2, the additional skeleton branch of SkeletonGait++ notably enhances the recognition accuracy, particularly when the body shape becomes less reliable. This augmentation is explicitly evident in challenging scenarios involving object carrying, occlusion, and poor illumination conditions, as observed on SUSTech1K (Table X).

To better understand the differences among DeepGaitV2, SkeletonGait, and SkeletonGait++, Fig. 10 visualizes their activation maps through the CAM technique [68]. As a result, DeepGaitV2 pays more attention to regions that exhibit distinct and discriminative body shapes. On the other hand, SkeletonGait can only concentrate on 'clean' structural features over the body joints and limbs. In comparison, SkeletonGait++ strikes a balance between these approaches, effectively capturing the 'comprehensive' gait patterns that are rich in both body shape and structural characteristics. Especially for the challenging cases shown in Fig. 10 (b), SkeletonGait++ adaptively leverages the skeleton branch to support the robust gait representation learning. This is an urgent need for practical

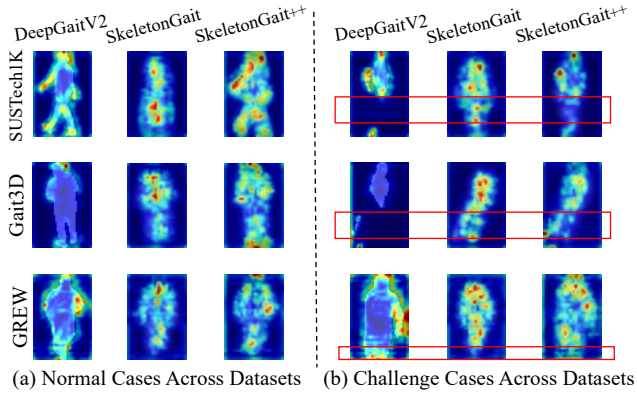


Fig. 10. The activation visualization of DeepGaitV2 *vs.* SkeletonGait and SkeletonGait++ on SUSTech1K [21], Gait3D [9], and GREW [8].

applications, and we also think this is the main reason causing the SkeletonGait++'s performance gains on Gait3D [9] and GREW [8] datasets.

D. Ablation Study

The preceding paragraphs have thoroughly explored the relationship and distinction among DeepGaitV2, SkeletonGait, and SkeletonGait++, through comprehensive experiments across all the utilized datasets. As a supplementary analysis, here we further investigate their robustness to the hyper-parameters as outlined below.

To determine the suitable network capacity for DeepGaitV2, Table VII has deepened the backbone depth from 10 to 30 layers. We further widen the network, and the results in Table XII (a) exhibit limited performance improvements but introduce significant training parameters and computation costs. Therefore, we set the network width C to 64 as default for a better accuracy-computation trade-off.

To determine the empirical value of σ in generating the skeleton map as described in Eq. (4) and (5), we range it from 4 to 16, and Table XII (b) shows that: a) SkeletonGait is robust to the value of σ. b) $\sigma = 8.0$ is an experimentally optimal choice. These findings also underscore the robustness and discriminative capacity of the proposed skeleton map as a gait modality used for human identification.

To determine the fusion location and mode within SkeletonGait++ shown in Fig. 6, we try all the combinations, and Table XII (c) reveals that: a) SkeletonGait++ is robust to both fusion location and mode. b) The low-level attention fusion is an experimentally optimal choice. Moreover, these

observations also highlight the promising potential of multi-modal gait recognition methods.

VII. Conclusion and Future Work

This work embarks on a journey to fulfill the urgent need for practical gait recognition by basic codebase building, previous SoTA revisiting, and new baseline construction, exhibiting a comprehensive benchmark study aimed at catalyzing further advancements in this field. As a result, OpenGait serves as an easy-to-use platform, while DeepGaitV2, SkeletonGait, and SkeletonGait++ are proposed as representatives of the appearance-based, model-based, and multi-modal gait recognition baseline models, respectively.

Beyond the above contributions, this work also makes it clear that the dataset scale, model size, and modality designs stand out as three of the bottlenecks in gait recognition. While these aspects are commonly acknowledged in computer vision, our experiments offer concrete insights into why the gait community continues to rely predominantly on shallow gait models and binary silhouette images. This work presents a good attempt to expose these gait-specific issues, recognizing that there remains considerable ground to cover for real-world gait recognition applications.

VIII. Acknowledgement

We would like to thank ...

References

[1] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 12, pp. 1505–1518, 2003. 1

[2] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 209–226, 2016. 1

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015. 1

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. 1

[5] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36. 1

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299. 1

[7] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 8126–8133. 1, 2, 3, 5, 6, 7, 8, 11, 12

[8] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou, "Gait recognition in the wild: A benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 789–14 799. 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13

[9] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3d representations and a benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13

[10] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 441–444. 1, 3, 4, 5, 6, 8, 11

[11] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Transactions on Computer Vision and Applications*, vol. 10, no. 1, pp. 1–14, 2018. 1, 3, 4, 5, 6, 8, 11, 12, 13

[12] S. Meng, Y. Fu, S. Hou, C. Cao, X. Liu, and Y. Huang, "Fastposegait: A toolbox and benchmark for efficient pose-based gait recognition," *arXiv preprint arXiv:2309.00794*, 2023. 1

[13] S. Yu, Y. Huang, L. Wang, Y. Makihara, S. Wang, M. A. R. Ahad, and M. Nixon, "Hid 2022: The 3rd international competition on human identification at a distance," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022, pp. 1–9. 2, 4

[14] S. Yu, W. Chenye, L. Wang, Z. Yuwei, W. Li, W. Ming, L. Qing, L. Wenlong, W. Runsheng, H. Yongzhen, W. Liang, M. Yasushi, and R. A. Md Atiqur, "Human identification at a distance: Challenges, methods and results on hid 2023," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 2023, pp. 1–7. 2

[15] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 225–14 233. 2, 3, 5, 6, 8, 11, 12

[16] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 648–14 656. 2, 3, 5, 6, 8, 11, 12

[17] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "Opengait: Revisiting gait recognition towards better practicality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9707–9716. 2, 3, 6, 8, 11, 12

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 2, 7, 8

[19] C. Fan, J. Ma, D. Jin, C. Shen, and S. Yu, "Skeletongait: Gait recognition using skeleton maps," 2023. 2

[20] W. Li, S. Hou, C. Zhang, C. Cao, X. Liu, Y. Huang, and Y. Zhao, "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 (CVPR)*, June 2023, pp. 13 824–13 833. 2, 3, 4, 8, 11, 12, 13

[21] C. Shen, C. Fan, W. Wu, R. Wang, G. Q. Huang, and S. Yu, "Lidargait: Benchmarking 3d gait recognition with point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 1054–1063. 2, 3, 4, 8, 11, 12, 13

[22] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges," *arXiv preprint arXiv:2206.13732*, 2022. 2, 4

[23] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020. 3, 9

[24] C. Song, Y. Huang, Y. Huang, N. Jia, and L. Wang, "Gaitnet: An end-to-end network for gait based human identification," *Pattern recognition*, vol. 96, p. 106988, 2019. 3, 4

[25] J. Zheng, X. Liu, S. Wang, L. Wang, C. Yan, and W. Liu, "Parsing is all you need for accurate gait recognition in the wild," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 116–124. 3, 4, 11, 12

[26] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, "Gaitedge: Beyond plain end-to-end gait recognition for better practicality," *arXiv preprint arXiv:2203.03972*, 2022. 3, 4

[27] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proceedings of the Asian conference on computer vision*, 2020. 3, 4, 6, 9

[28] Y. Wang, X. Zhang, Y. Shen, B. Du, G. Zhao, L. Cui, and H. Wen, "Event-stream representation for human gaits identification using deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3436–3449, 2022. 3

[29] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang, "Gpgait: Generalized pose-based gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 595–19 604. 3, 11, 12

[30] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: graph convolutional network for skeleton-based gait recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2314–2318. 3, 9, 11, 12

[31] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015. 3

[32] C. Xu, Y. Makihara, X. Li, and Y. Yagi, "Occlusion-aware human mesh model-based gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1309–1321, 2023. 3

[33] Z. Huang, D. Xue, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua, "3d local convolutional neural networks for gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 920–14 929. 3

[34] M. Wang, X. Guo, B. Lin, T. Yang, Z. Zhu, L. Li, S. Zhang, and X. Yu, "Dygait: Exploiting dynamic representations for high-performance gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 424–13 433. 3

[35] Y. Peng, S. Hou, K. Ma, Y. Zhang, Y. Huang, and Z. He, "Learning rich features for gait recognition by integrating skeletons and silhouettes," *arXiv preprint arXiv:2110.13408*, 2021. 4, 12

[36] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li, "Gaitgci: Generative counterfactual intervention for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5578–5588. 4

[37] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu, "Learning gait representation from massive unlabelled walking videos: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 920–14 937, 2023. 4

[38] K. Ma, Y. Fu, D. Zheng, Y. Peng, C. Cao, and Y. Huang, "Fine-grained unsupervised domain adaptation for gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 313–11 322. 4

[39] C. Filipi Gonçalves dos Santos, D. d. S. Oliveira, L. A. Passos, R. Gonçalves Pires, D. Felipe Silva Santos, L. Pascotti Valem, T. P. Moreira, M. Cleison S. Santana, M. Roder, J. Paulo Papa *et al.*, "Gait recognition based on deep learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–34, 2022. 4

[40] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[41] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? a worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM conference on recommender systems*, 2019, pp. 101–109. 4

[42] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *European Conference on Computer Vision*. Springer, 2020, pp. 681–699. 4

[43] ——, "Unsupervised domain adaptation: A reality check," *arXiv preprint arXiv:2111.15672*, 2021. 4

[44] B. Amos, B. Ludwiczuk, M. Satyanarayanan *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, no. 2, p. 20, 2016. 4

[45] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019. 4

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019. 4

[47] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33018295 5, 7

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 5

[49] P. Zhang, H. Dou, Y. Yu, and X. Li, "Adaptive cross-domain learning for generalizable person re-identification," in *European Conference on Computer Vision*. Springer, 2022, pp. 215–232. 6

[50] W. Yu, H. Yu, Y. Huang, and L. Wang, "Generalized inter-class loss for gait recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 141–150. 6

[51] W. Xiang, H. Yang, D. Huang, and Y. Wang, "Multi-view gait video synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6783–6791. 6

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. 7

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 7

[54] X. Huang, D. Zhu, H. Wang, X. Wang, B. Yang, B. He, W. Liu, and B. Feng, "Context-sensitive temporal feature learning for gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 909–12 918. 7

[55] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "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*, 2019, pp. 0–0. 7

[56] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160. 7

[57] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541. 7, 8

[58] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978. 9, 10

[59] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1159–1168. 9, 10

[60] R. Liao, Z. Li, S. S. Bhattacharyya, and G. York, "Posemapgait: A model-based gait recognition method with pose estimation maps and graph convolutional networks," *Neurocomputing*, vol. 501, pp. 514–528, 2022. 9, 10

[61] K. Ma, Y. Fu, D. Zheng, C. Cao, X. Hu, and Y. Huang, "Dynamic aggregated network for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 076–22 085. 11

[62] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu, "Learning gait representation from massive unlabelled walking videos: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2023. 11

[63] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao, Y. Huang, P. Li, and S. Xu, "Qagait: Revisit gait recognition from a quality perspective," *arXiv preprint arXiv:2401.13531*, 2024. 11

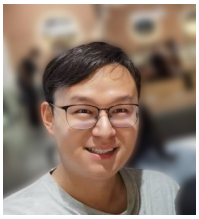
[64] C. Zhang, X.-P. Chen, G.-Q. Han, and X.-J. Liu, "Spatial transformer network on skeleton-based gait recognition," *Expert Systems*, vol. 40, no. 6, p. e13244, 2023. 11, 12

[65] H. Zhu, W. Zheng, Z. Zheng, and R. Nevatia, "Gaitref: Gait recognition with refined sequential skeletons," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, pp. 1–10. 11

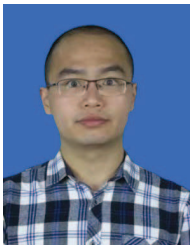
[66] Y. Dong, C. Yu, R. Ha, Y. Shi, Y. Ma, L. Xu, Y. Fu, and J. Wang, "Hybridgait: A benchmark for spatial-temporal cloth-changing gait recognition with hybrid explorations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1600–1608. 11

[67] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 382–398. 11

[68] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929. 13



Chao Fan received the B.E. and M.S. degrees from Xi'an University of Technology in 2018 and University of Science and Technology Beijing in 2021, respectively. He is currently a Ph.D. candidate with Department of Computer Science and Engineering, Southern University of Science and Technology. His research interests include gait recognition and computer vision.



Saihui Hou received the B.E. and Ph.D. degrees from University of Science and Technology of China in 2014 and 2019, respectively. He is currently an Assistant Professor with School of Artificial Intelligence, Beijing Normal University. His research interests include computer vision and machine learning. He recently focuses on gait recognition which aims to identify different people according to the walking patterns.

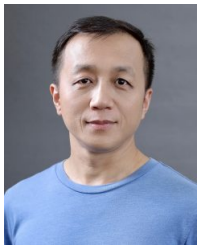
Junhao Liang received the B.E. degrees from Southern University of Science and Technology in 2021. He is currently a Master student with Department of Computer Science and Engineering, Southern University of Science and Technology. His research interests include gait recognition and image generation.

Chuanfu Shen received his B.E. degree in 2019 from the Southern University of Science and Technology in Shenzhen, China. He is currently a Ph.D. candidate jointly enrolled in Computer Science and Engineering at the Southern University of Science and Technology, and Industrial and Manufacturing Systems Engineering at The University of Hong Kong. His research interests include human retrieval, gait recognition, and point cloud classification.

Jingzhe Ma received the B.E. and M.S. degrees in computer science and technology from Zhengzhou University in 2017 and 2020, respectively. He is currently a Ph.D. candidate with Department of Computer Science and Engineering, Southern University of Science and Technology. His research interests include human video synthesis and gait recognition.

Dongyang Jin received the B.E. degrees from Southern University of Science and Technology in 2023. He is currently a Master student with Department of Computer Science and Engineering, Southern University of Science and Technology. His research interests include gait recognition and large vision models.

Yongzhen Huang received the B.E. degree from Huazhong University of Science and Technology in 2006, and the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2011. He is currently a Professor at School of Artificial Intelligence, Beijing Normal University. He has published one book and more than 80 papers at international journals and conferences such as TPAMI, IJCV, TIP, TSMCB, TMM, TCSVT, CVPR, ICCV, ECCV, NIPS, AAAI. His research interests include pattern recognition, computer vision and machine learning.



Shiqi Yu is currently an Associate Professor in the Department of Computer Science and Engineering, Southern University of Science and Technology, China. He received his B.E. degree in computer science and engineering from the Chu Kochen Honors College, Zhejiang University in 2002, and Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2007. He worked as an Assistant Professor and an Associate Professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences from 2007 to 2010, and as an Associate Professor in Shenzhen University from 2010 to 2019. His research interests include gait recognition, face detection and computer vision.