BigGait: Learning Gait Representation You Want by Large Vision Models

Dingqiang Ye[1,2*], Chao Fan[1,2*], Jingzhe Ma[1,2], Xiaoming Liu[3], and Shiqi Yu[1,2†]

[1] Research Institute of Trustworthy Autonomous System, Southern University of Science and Technology, Shenzhen, China

[2] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

[3] Michigan State University, Michigan, United States

{11810121, 12131100, 12031127}@mail.sustech.edu.cn, liuxm@cse.msu.edu, yusq@sustech.edu.cn

Abstract

Gait recognition stands as one of the most pivotal remote identification technologies and progressively expands across research and industry communities. However, existing gait recognition methods heavily rely on task-specific upstream driven by supervised learning to provide explicit gait representations like silhouette sequences, which inevitably introduce expensive annotation costs and potential error accumulation. Escaping from this trend, this work explores effective gait representations based on the all-purpose knowledge produced by task-agnostic Large Vision Models (LVMs) and proposes a simple yet efficient gait framework, termed ***BigGait****. Specifically, the Gait Representation Extractor (GRE) within BigGait draws upon design principles from established gait representations, effectively transforming all-purpose knowledge into implicit gait representations without requiring third-party supervision signals. Experiments on CCPG, CAISA-B* and SUSTech1K indicate that BigGait significantly outperforms the previous methods in both within-domain and cross-domain tasks in most cases, and provides a more practical paradigm for learning the next-generation gait representation. Finally, we delve into prospective challenges and promising directions in LVMs-based gait recognition, aiming to inspire future work in this emerging topic. The source code is available at* `https://github.com/ShiqiYu/OpenGait`.

1. Introduction

Vision-based gait recognition aims to identify individuals based on their unique walking patterns. Compared to other biometric modalities like face, fingerprint, and iris, gait stands out for its non-intrusive nature and the ability to identify individuals at long distances without the need for active cooperation. These distinctive advantages make gait recognition exceptionally suitable for a range of security applications, like suspect tracking and identity investigation [34].

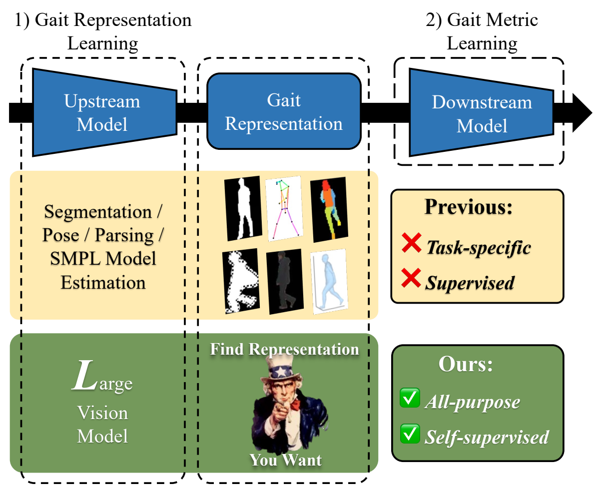

Figure 1. The upstream and downstream parts of existing gait recognition methods are responsible for gait representation and metric learning, respectively.

Existing gait recognition methods [3, 27–29, 31, 50, 51, 53] heavily depend on upstream tasks driven by supervised learning, as illustrated by Fig. 1. These tasks encompass a range of objectives, including, but not limited to, pedestrian segmentation [45], human parsing [32], body posture [2], and SMPL [33] model estimation. In general, the upstream model serves the purpose of providing task-specific priors to filter out the gait-irrelevant cues within walking videos, especially for RGB-encoded background and texture characteristics. Next, some inductive biases may be utilized to refine the intermediate gait representation, ranging from basic operations like size alignment [42], coordinate normalization [15] to sophisticated ones like making the silhouette edges differentiable [28] and enforcing the appearance reconstruction [27, 50]. Briefly, existing gait representations are largely predetermined by upstream supervised tasks.

*Equal contribution.
†Corresponding author.

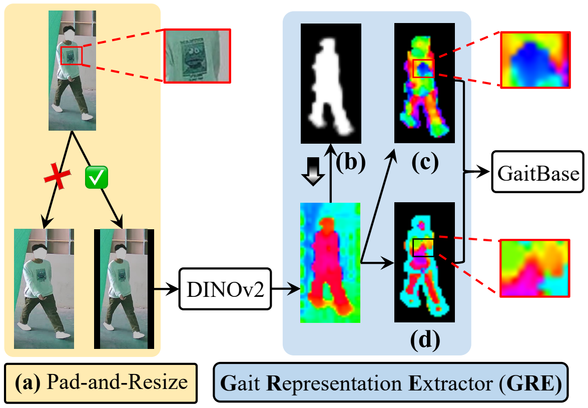

Figure 2. The illustration of (a) body proportion preservation trick is provided in the Supplementary Material. (b)-(d) respectively present the visualization of intermediate representation generated by (b) mask, (c) appearance, and (d) denoising branch.

Escaping from the reliance on upstream task-specific models, this paper makes a pioneering effort to acquire desired gait features from *task-agnostic Large Vision Models* (LVMs). We are strongly driven by the following insights:

- The recent breakthroughs [17, 23, 35, 37] have confirmed the discriminability and generalization of all-purpose features produced by LVMs, showcasing an attractive opportunity to enhance gait metric learning.
- Self-supervised pre-training of LVMs obviates the need for labeled datasets required for training the state-of-the-art gait upstream models, thus eliminating the substantial costs necessary for annotating elements such as the silhouette, skeleton, and more, on a large scale.
- The task-agnostic knowledge embedded in LVMs is learned from web-scale datasets without task-specific supervision, thus ideally avoiding error accumulation imposed by specific upstream tasks to a large extent.

To bring the above visions to life, we propose the first LVMs-based gait framework termed *BigGait*[1]. Structurally, we configure the upstream and downstream models as DINOv2 [35] and GaitBase [13] with minor adjustments, thanks to their representativeness in the domains of LVMs and gait research, respectively. In this paper, we regard LVMs-based gait recognition as a challenging task focusing on transforming all-purpose features into effective gait representations. To achieve this goal, we propose the *Gait Representation Extractor* (GRE) bridging the upstream and downstream models, as illustrated in Fig. 2:

- **Mask Branch**. Borrowing the design of the silhouette, as shown in Fig. 2 (b), GRE develops a mask branch to autonomously infer the foreground without supervision, thus excluding the background interference on the whole.
- **Appearance Branch**. As observed in Fig. 2 (c), background removal makes foreground features diverse and discriminative over body parts, resulting in a parsing-like representation. However, this branch may potentially introduce texture-related noise.
- **Denoising Branch** in Fig. 2 (d) introduces smoothness constraints along the spatial dimension to reduce the high-frequency textural characteristics. Moreover, a diversity constraint is introduced to prevent trivial solutions.

Overall, our GRE module adopts the design principles of some established gait representations and distinguishes itself by leveraging all-purpose knowledge and removing upstream task-specific supervision. Technically, GRE relies solely on soft constraints to extract effective gait representations, *i.e.*, capturing truthful body structural characteristics while effectively excluding gait-irrelevant noise. Moreover, this work also finds that the utilization of soft constraints may introduce two major challenges: (1) **Interpretability**. In contrast to the more visually intuitive modality such as the silhouette and skeleton, which directly convey physical meanings, the learned representation derived from soft geometrical constraints, presented as the multi-channel feature map, may not offer the same level of intuitiveness. (2) **Purity**. Completely suppressing the texture characteristics of RGB videos through soft constraints remains a challenge for gait recognition [28]. Our comprehensive cross-clothing and cross-domain evaluations provide strong evidence of the dominant role of gait features within the acquired gait representations. Nevertheless, we find that data distribution may still influence outcomes.

In summary, this paper presents a pioneering attempt to open up the promising LVMs-based gait recognition research. The main contributions can be outlined as follows:

- BigGait charts a groundbreaking and practical learning paradigm for the next-generation gait representation, with gait guidance shifting from task-specific priors to LVMs-based all-purpose knowledge.
- We establish the effectiveness of all-purpose knowledge for gait description, as evidenced by its discriminability and generalization across various gait datasets, including CCPG [25], CASIA-B* [47] and SUSTech1K [39]. BigGait achieves the SoTA performance in most cases.

2. Related Works

Upstream Models for Gait Representation Learning. Walking videos often contain numerous extraneous elements, including background distractions and variations in color, texture, clothing, and carried items within the foreground. Prior methods have introduced a range of task-specific models to address these challenges. These models are designed to provide binary silhouettes [3, 10, 13, 30], body skeletons [29, 43], human parsing [53], SMPL models [27], or a combination of these representations [52], to serve as the input for downstream gait models. Moreover,

[1] Both 'large' and 'big' can refer to LVMs' size. We use the latter since it is relatively common in daily conversations.

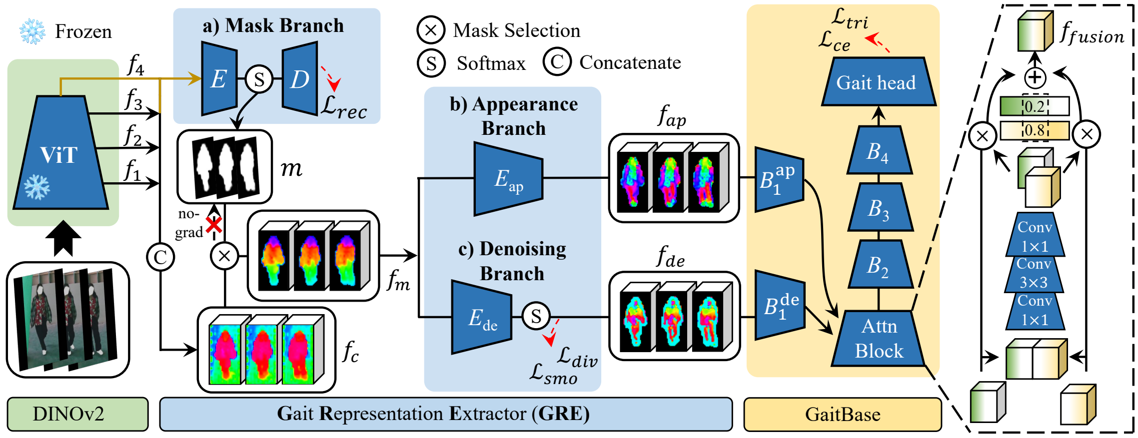

Figure 3. The workflow of BigGait. Specifically, the upstream model is instantiated as DINOv2 aiming to produce all-purpose features. The central gait representation extractor (GRE) owns three branches respectively responsible for the background removal, feature transformation, and feature refining. In the end, the modified GaitBase is employed for gait metric learning.

several end-to-end methods [28, 50] aim to optimize the intermediate gait representation globally. However, it is important to recognize that these upstream tasks heavily rely on supervised learning, giving rise to three primary challenges: a) considerable annotation costs, b) restricted form of gait representation, and c) potential accumulation of estimation errors. In this paper, BigGait instantiates the upstream model as a self-supervised LVM, constructing the desired gait representation by leveraging the all-purpose knowledge in a novel manner.

Downstream Models for Gait Metric Learning. Depending on the gait representations used, downstream gait models are designed accordingly. For instance, Convolutional Neural Networks (CNNs) are extensively employed for modeling image-based gait representations such as silhouettes and human parsing, while Graph Convolutional Networks (GCNs) are commonly employed for learning skeleton-based representations. In the majority of cases, downstream gait models place their emphasis on local and global temporal modeling [9, 21, 22], along with the creation of spatially hierarchical descriptions [10, 15, 30].

Large Vision Models. Inspired by the success of LLMs [1, 7], researchers have ventured into exploring foundation models in the field of computer vision [23, 35, 37], *i.e.*, learning generalizable features from web-scale data sources by scalable vision models. For example, CLIP [37] used a form of textual supervision to guide the training of visual representation. SAM [23] developed a promotable model and pre-trained it on a broad dataset using a task that enables powerful generalization for image segmentation. This paper focuses on another alternative termed self-supervised LVMs [5, 6, 19], where features are learned from images alone without supervision. Specifically, the employed DINOv2 [35] demonstrated that self-supervised learning has the potential to learn all-purposed visual features if pre-trained on a large quantity of curated data. Here, all-purpose features aim to work out of the box on any task, *e.g.*, image classification, and dense recognition like semantic segmentation and depth estimation. The main goal of this paper is to learn gait representation from these all-purpose features, thereby benefiting from the advantages of LVMs.

Other Related Works. Recently, GaitSSB [12] introduced a self-supervised pre-training benchmark tailored for silhouette-based gait recognition. Notably, this pre-training approach is gait-specific and directly influences the downstream model. It still heavily relies on an upstream supervised segmentation model, whereas in the case of DINOv2 within our BigGait, the pre-training is task-agnostic and serves as the upstream model without the need for supervision. Another distinctive advantage of BigGait lies in the multi-branch architecture of the GRE module, which offers a gait representation resembling the capabilities of multi-modality, even though it derives from a single source. While typical multi-modal methods [36, 52] require multiple supervised upstream models, we can create diverse gait representations or their combinations relying solely on all-purpose features obtained from self-supervised LVMs.

3. Method

As shown in Fig. 3, BigGait consists of three parts, *i.e.*, the upstream frozen DINOv2 [35], the central GRE module, and the downstream adjusted GaitBase [13], respectively responsible for all-purpose feature extraction, gait representation construction, and gait metric learning.

3.1. Overview

BigGait takes an RGB video as input and processes each frame in parallel. For simplicity, we focus on a single frame in the following description. Before feeding the image into the upstream model, this work adopts a Pad-and-Resize

Table 1. Parameter and GFLOPs of different upstream/overall methods. The GFLOPs are calculated for the input size 448×224.

Upstream Obj.	Methods	Backbone	#Params (M)	GFLOPs
Segmentation	UNet3+ [20]	UNet3+	27.0	308.2
	DeepLabV3+ [4]	ResNet-50	26.8	43.7
Human Parsing	SCHP [24]	A-CE2P	66.6	33.5
	HRNet [45]	HRNet-W48	70.1	61.9
Pose Estimation	HRNet [41]	HRNet-W32	28.5	15.8
	DEKR [16]	HRNet-W32	29.5	17.1
All-purpose (**Ours**)	DINOv2 [35]	ViT-S/14	21.5	11.0
		ViT-L/14	302.9	155.3
Overall	**Upstream**	**Downstream**	**#Params (M)**	**GFLOPs**
Sil.-based	DeepLabV3+	GaitBase	34.2	45.4
Parsing-based	SCHP	GaitBase	74.0	35.2
Skeleton-based	HRNet-W32	Gait-TR	29.0	31.2
BigGait (**Ours**)	ViT-S/14 + GRE	GaitBase	30.8	12.7

trick to resize it into a fixed resolution of 448×224 while keeping the truthful body proportions. The illustration of this detail is provided in the Supplementary Material.

Upstream Model. The upstream DINOv2 owns a scalable ViT [8] as the backbone, with parameter counts ranging from 21M to $1,100$M. In this study, we opt for the smallest and second largest counterparts, namely, ViT-S/14 (21M) and ViT-L/14 (302M), to construct BigGait-S and BigGait-L, respectively. Given the resized RGB image, the ViT first divides it into non-overlapping patches with a size of 14×14, thus taking 32×16 tokenized vectors as input. Next, the positional encodings are added, and subsequent stacked ViT blocks transform the features. We gather tokens from low to high layers with uniform intervals, and all these 32×16 tokens are concatenated according to the spatial correspondence to form a feature map. These details follow the official implementation [35]. Finally, we upsample the feature map to 64×32 as output.

We leverage the officially provided model checkpoint pre-trained on LVD-142M [35] and freeze the imported parameters during gait metric learning. To reduce carbon emissions, we establish BigGait-S as the reference version, even though BigGait-L exhibits slightly higher performance. Thanks to this choice, the upstream part of BigGait is even smaller than many typical segmentation, human parsing, and pose estimation networks widely used by existing gait methods, as shown in Tab. 1.

Central Module. Directly using the learned all-purpose features can result in inferior performances. In light of this, this paper pays primary attention to transforming all-purpose features into useful gait representations, thereby proposing the GRE module. Specifically, the GRE module includes three primary components, *i.e.*, the mask, appearance, and denoising branches respectively responsible for the background removal, feature transformation, and feature refining. This module outputs two types of gait representations, and Sec. 3.2 will provide details.

Downstream Model. We make a minor modification to GaitBase [13] so that it can process the two-stream input. As shown in Fig. 3, B_1, B_2, B_3 and B_4 refer to the standard blocks within GaitBase, where B_1^{ap} and B_1^{de} own the different parameters but share the identical structure with B_1. Then, an attention block possessing the widely used cross-and-select architecture [26] targets aggregating the generated two-branch features. Excluding the above modifications, our implementation of GaitBase makes no architectural difference from the official one [13]. The training is driven by the triplet L_{tri} and cross-entropy $L_{L_{ce}}$ losses.

Visualization. To understand the content of each intermediate feature map, namely f_c, f_m, f_{ap}, and f_{de} depicted in Fig. 3, we conduct a PCA using the pixel-wise features from the entire dataset. Projecting pixel-wise features onto the first three PCA bases allow us to visualize a high-dim feature map as an RGB color image. This visualization approach is inherited from DINOv2 [35].

3.2. Gait Representation Extractor

As shown in Fig. 3, f_1, f_2, f_3, f_4 denote the feature map generated by various stages of the ViT backbone with the corresponding semantic hierarchy spanning from low to high levels, which is a common practice suggested by DINOv2. Each of them is $2\times$ upsampled by bilinear interpolation to improve the resolution, *i.e.*, exhibited as a 3-D tensor with a size of $384 \times 64 \times 32$ while the first dimension denotes the output channel of the upstream DINOv2. Additionally, we concatenate these feature maps along the channel dimension to form f_c, *i.e.*, with a size of $1,536 \times 64 \times 32$, which is rich in both low-level details and high-level semantics. As shown in Fig. 3, f_c is dominated by the foreground and background regions accompanied by noisy spots.

Mask Branch. Inspired by the design of the silhouette, the GRE module develops a mask branch to remove the background noise on the whole. Specifically, an auto-encoder is proposed to generate the foreground mask based on f_4:

$$\begin{aligned} m &= \text{softmax}(E(f_4)) \\ \bar{f}_4 &= D(m) \\ L_{rec} &= \left\| f_4 - \bar{f}_4 \right\|_2 , \end{aligned} \tag{1}$$

where E and D denote the linear convolution layer with a kernel size of 1×1 and the output channel of 2 and 384, respectively. The softmax function is conducted along the channel dimension, and L_{rec} presents the reconstruction loss. Here, we use f_4 rather than f_c due to its better empirical results. The cause may be that f_4 contains higher-level semantically separable features.

Intuitively, the mask branch works like a PCA by searching two mutually exclusive first components. Thanks to the segmentable nature of f_4, as shown in Fig. 4, the mask m can be easily learned in this way. Notably, the m is a two-channel tensor, and which channel represents the foreground is uncertain. So we select the channel whose activations lay more in the image center as the foreground mask. Moreover, we perform the binarization and closing operations to reduce the potential cavities and breakpoints. These

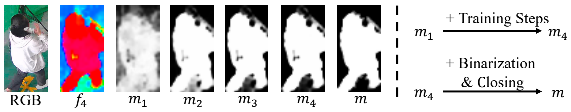

Figure 4. Visualization of unsupervised mask learning.

common tricks do not appear in Fig. 3 for brevity, *i.e.*, the symbol of m in Fig. 3 denotes the final foreground mask.

After masking the background regions in f_c with m, we obtain f_m:

$$f_m = m \cdot f_c, \tag{2}$$

where $\cdot$ denote the multiplication. Fig. 3 shows that this process not only largely filters out the background noise but also makes foreground features diverse and discriminative over body parts, resulting in a parsing-like effect.

Appearance Branch. To effectively capture characteristics within f_m, GRE develops an appearance branch:

$$f_{ap} = E_{\text{ap}}(f_m), \tag{3}$$

where E_{ap} denotes a linear convolution layer with the kernel size of 1×1 and output channel of C. Here C is a hyper-parameter determining the channel of intermediate gait representation f_{ap}. Note the functionality of E_{ap} can be easily approximated by the early layers of GaitBase. Here we introduce this layer mainly for GPU memory saving by ensuring $C \ll 1,536$ (the channel number of f_m). Overall, E_{ap} linearly adjusts features along the channel dimension and then directly feeds its output f_{ap} into the gait model.

By comparing the input f_m with the output f_{ap}, as shown in Fig. 3, the emergence of noisy elements within the latter should be primarily driven by the downstream gait model rather than the linear E_{ap}. This observation is consistent with findings in related studies [25, 28], suggesting that the gait model tends to emphasize the clearly unchanged gait-irrelevant cues rather than the subtle gait patterns. Further visualizations in Sec. 3.3 reveal the correlation between these noises and the clothing texture on the body.

To alleviate this issue, the GRE module develops a denoising branch parallel to this appearance branch.

Denoising Branch. This branch regards texture noises as high-frequency signals along spatial dimensions, thus introducing a smoothness loss L_{smo}, formulated as:

$$\begin{aligned} f_{de} &= \text{softmax}(E_{\text{de}}(f_m)) \\ L_{smo} &= |\text{sobel}_{\text{x}} * f_{de}| + |\text{sobel}_{\text{y}} * f_{de}|, \end{aligned} \tag{4}$$

where E_{de} denotes non-linear layers composed by a 1×1 convolution, a batch normalization, a GELU, and an additional 1×1 convolution. Its output channel is set to C which is the same as that of the appearance branch for brevity. The softmax function is performed along the channel dimension to normalize the features. The smoothness loss is defined by the absolute value of channel-wise image gradients along both the width and height axes. In Eq. 4, sobel$_x$ and sobel$_y$ refer to the x-axis and y-axis Sobel operator [40], and the symbol of $*$ represents the convolution operation.

From a definitional standpoint, L_{smo} can not remove the gradual textural noise effectively. But for gait recognition, we consider that the evident textures on body images, *e.g.*, the clothing logo and pattern, are generally high-frequency characteristics that can be easily eliminated by L_{smo}.

Relying solely on the smoothness loss L_{smo} can potentially lead to a trivial solution, where each pixel of f_{de} essentially presents nearly identical feature vectors, resulting in a significant loss of representational diversity. To tackle this issue, we introduce an additional diversity loss based on information entropy. The softmax normalization ensures a certain probability distribution of activations across the channel dimension, and the uniform one corresponds to the maximum information entropy. Thus, the diversity loss is:

$$\begin{aligned} p_i &= \text{sum}(f_{de}^i) / \sum_{i=1}^{C} \text{sum}(f_{de}^i) \\ L_{div} &= \text{log}C + \sum_{i=1}^{C} p_i \text{log} p_i, \end{aligned} \tag{5}$$

where f_{de}^i and p_i respectively represents the activation map of the i-th channel and the corresponding frequency to entire activations. The constant term log$C$ denotes the maximum entropy and is imported to prevent the negative loss.

Overall, the learning of the denoising branch is entirely driven by the above two soft constraints. These losses only perform in the foreground produced by the mask branch. Finally, we fuse f_{ap} and f_{de} using attention weights:

$$f_{fusion} = Attn(\text{B}_1^{ap}(f_{ap}), \text{B}_1^{de}(f_{de})), \tag{6}$$

where $Attn$ is an attention block, following [14].

3.3. Visualization of Intermediate Representation

To provide an intuitive understanding, we visualize each intermediate representation in BigGait, and compare it with traditional gait representations, as shown in Fig. 5. All-purpose features f_m created by the upstream model (DINOv2) provide diverse and discriminative body parts after masking the background regions, *i.e.*, purple head, red abdomen, green limbs and blue shoe. Both f_{ap} in the appearance branch and f_{de} in the denoising branch are derived from f_m, with the former inheriting features by a linear transformation and showing body shape representation with high-frequency texture noises, and the latter embodying highly consistent skeleton-like representation since deploying smoothness and diversity constraints removes most texture noises. However, compared with traditional gait representations with direct physical meanings, the gait representation based on soft geometrical constraints may need more interpretability on its physical meaning in future.

3.4. Implementation Details

Loss. The overall loss can be formulated as:

$$L = L_{tri} + L_{ce} + \gamma_{rec} L_{rec} + \gamma_{smo} L_{smo} + \gamma_{div} L_{div}, \tag{7}$$

where the first two terms are recognition losses of the downstream gait model, L_{rec} is the mask reconstruction loss, and

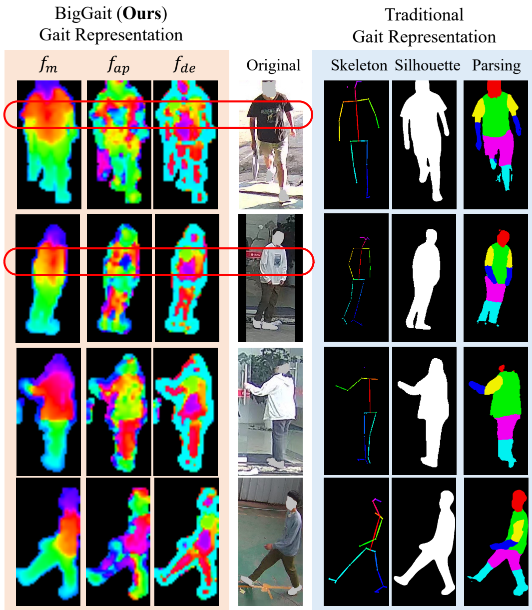

Figure 5. The visualization of intermediate representations generated by BigGait *v.s.* three traditional gait representations. The red boxes indicate regions with strong texture patterns.

Table 2. The amount of the identities (#ID) and sequences (#Seq) covered by the employed datasets.

Data Set	Train Set #ID	Train Set #Seq	Test Set #ID	Test Set #Seq	Setting
CCPG	100	$8,388$	100	$8,178$	CL, UP, DN, BG
CASIA-B*	74	$8,140$	50	$5,500$	NM, BG, CL
SUSTech1K	250	$6,011$	750	$19,228$	Various

the last two are the smoothness and diversity losses of the denoising branch. In our implementations, $\gamma_{rec}, \gamma_{smo}$ and γ_{div} are respectively set to 1.0, 0.01, and 5.0.

Training. All images are processed with Pad-and-Resize to the size of 448×224. We use the SGD optimizer with a weight decay of 0.0005, a momentum of 0.9 and an initial learning rate of 0.1. The learning rate is reduced by a factor of 0.1 at the 15k, 25k, 30k and 35k iterations, and the total number of iterations is 40k. The batch size (p, k, l) is set to $(8, 8, 30)$, where p is the number of IDs, k the number of sequences per ID, and l the number of frames per sequence. The sampling strategy of frames from sequence follows GaitBase [13]. The data augmentation operation only contains randomly flipping all images in a sequence.

4. Experiments

4.1. Datasets

In our experiments, three popular cross-clothing and multi-view gait datasets are employed, *i.e.*, CCPG [25], CASIA-B* [28, 47] and SUSTech1K [39]. Among them, CCPG acts as the primary benchmark since it prioritizes challenging cloth-changing scenarios, featuring a diverse collection of coats, pants, and bags in various colors and styles. The key statistics of these gait datasets are listed in Tab. 2. Our implementation follows official protocols, including the training and gallery/probe set partition strategies. During the testing, all datasets adopt comprehensive gait evaluation protocols for multi-view scenes, and rank-1 accuracy is considered the primary metric.

Table 3. Rank-1 accuracy on CCPG: BigGait *v.s.* recent SoTA models using different inputs. GaitBase^p: human parsing inputs. GaitBase^{p+s}: parsing and silhouette inputs.

Input	Model	Venue	CL	UP	DN	BG	Mean
Skeleton	GaitGraph2 [44]	CVPRW'22	5.0	5.3	5.8	6.2	5.6
	Gait-TR [49]	ES'23	15.7	18.3	18.5	17.5	17.5
	MSGG [36]	MTA'23	29.0	34.5	37.1	33.3	33.5
Sils	GaitSet [3]	TPAMI'22	60.2	65.2	65.1	68.5	64.8
	GaitPart [10]	CVPR'20	64.3	67.8	68.6	71.7	68.1
	AUG-OGBase [25]	CVPR'23	52.1	57.3	60.1	63.3	58.2
	GaitBase [13]	CVPR'23	71.6	75.0	76.8	78.6	75.5
	DeepGaitV2 [11]	Arxiv	78.6	84.8	80.7	89.2	83.3
Parsing	GaitBase^p	CVPR'23	59.1	62.1	66.8	68.1	64.0
Parsing+Sils	GaitBase^{p+s}	CVPR'23	73.6	76.2	79.1	79.2	77.0
Skeleton+Sils	SkeletonGait++ [14]	AAAI'24	79.1	83.9	81.7	89.9	83.7
RGB+Sils	GaitEdge [28]	ECCV'22	66.9	74.0	70.6	77.1	72.2
RGB	AP3D [18]	ECCV'20	53.4	57.3	69.7	91.4	67.8
	PSTA [46]	ICCV'21	42.2	52.2	60.3	84.5	59.8
	PiT [48]	TII'22	41.0	47.6	64.3	91.0	61.0
	BigGait	CVPR'24	**82.6**	**85.9**	**87.1**	**93.1**	**87.2**

4.2. Main Results

Note BigGait stands for BigGait-S, as mentioned in Sec. 3.1, where BigGait-L refers to its large version.

Within-domain Evaluation. We compare BigGait with various SoTA methods, such as the skeleton-based [36, 44, 49], silhouette-based [3, 10, 13, 25], and end-to-end gait recognition methods [28], and video-based ReID methods [18, 46, 48], on the challenging CCPG. The Supplementary Material provides more within-domain results on ReID evaluation protocols and on other datasets [39, 47, 54].

As shown in Tab. 3, BigGait achieves significantly better performance than other SoTA methods on CCPG. In the following, we measure BigGait's effectiveness in filtering out gait-irrelevant noises from two aspects.

First, compared to video-based ReID methods [18, 46, 48], BigGait outperforms them considerably, *e.g.*, $+29.2\%$ for full-changing (CL), $+28.6\%$ for ups-changing (UP), $+17.4\%$ for pants-changing (DN) and $+1.7\%$ for bag-changing (BG) scenarios. Accordingly, we consider that BigGait can efficiently extract robust identity representations from raw RGB videos. Fig. 6 provides intuitive evidence by visualizing the activation difference between BigGait and video-based ReID methods.

Second, compared to silhouette-based gait methods [3, 10, 13, 25], which perform well due to immunity to color/texture noises, BigGait still outperforms them by $+4.0\%$, $+1.1\%$, $+6.4\%$, and $+3.9\%$ in CL, UP, DN, and BG settings. This indicates that BigGait's immunity to gait-

Table 4. Rank-1 accuracy on three popular benchmark datasets: BigGait *v.s.* recent SoTA methods. This is a cross-domain evaluation and comparison, in which all methods are trained on one dataset and tested on the remaining two datasets.

(a) Trained on **CCPG**

Model	Test Set					
	CASIA-B*			SUSTech1K		
	NM	BG	CL	Clothing	Night	Overall
GaitSet [3]	47.4	40.9	25.8	8.2	11.0	12.8
GaitBase [13]	59.1	52.7	30.4	9.5	13.1	16.8
AP3D [18]	53.7	46.2	11.9	36.2	**51.6**	55.3
PSTA [46]	49.7	42.3	8.8	25.7	29.4	40.6
BigGait	**77.4**	**71.5**	**33.6**	**43.7**	44.8	**56.4**

(b) Trained on **CASIA-B***

Model	Test Set						
	CCPG				SUSTech1K		
	CL	UP	DN	BG	Clothing	Night	Overall
GaitSet	**10.6**	16.4	17.2	24.9	7.2	12.2	12.8
GaitBase	**10.6**	18.1	**21.4**	28.7	8.1	11.8	15.6
AP3D	2.1	2.9	3.9	6.1	29.3	47.4	48.3
PSTA	1.7	1.9	3.4	5.0	19.9	37.5	34.6
BigGait	7.5	**19.5**	14.2	**43.0**	**36.9**	**60.2**	**64.8**

(c) Trained on **SUSTech1K**

Model	Test Set						
	CCPG				CASIA-B*		
	CL	UP	DN	BG	NM	BG	CL
GaitSet	14.0	**23.7**	20.3	43.2	63.3	50.8	26.4
GaitBase	**16.8**	21.7	**26.0**	42.7	73.1	61.2	**28.2**
AP3D	5.5	7.9	13.9	35.1	56.7	48.1	15.3
PSTA	3.7	5.7	9.5	26.5	31.2	27.7	10.6
BigGait	4.5	11.5	11.9	**45.5**	**91.1**	**85.8**	18.7

irrelevant noises is comparable to silhouette-based ones.

Based on these analyses, we consider that gait features are the primary characteristics extracted by BigGait, even if gait-irrelevant noises cannot be absolutely eliminated from RGB inputs. We further validate this conclusion through cross-domain and visualization experiments.

Cross-domain Evaluation. Tab. 4 shows cross-domain performance comparisons. BigGait's cross-domain performance varies depending on the training set used. When trained on CCPG, BigGait exhibits strong adaptability to unseen datasets, *i.e.*, outperforming both video-based ReID methods [18, 46] and silhouette-based methods [3, 13]. When the training set is CASIA-B*, BigGait exhibits less impressive performances in some settings, *i.e.*, performing well in the ups-changing (UP), but poorly in the full-changing (CL) and pants-changing (DN). When crossing domains from SUSTech1K, BigGait struggles in more cross-dressing settings (CL, UP and DN).

We consider that BigGait's power in filtering gait-irrelevant noises is impacted by the training data biases. Specifically, the probability of sampling cross-dressing data pairs from SUSTech1K is lower than 5.0%, likely making BigGait focus on unchanged clothing cues rather than subtle gait patterns. Fortunately, with the enrichment of training data, BigGait can self-correct this mistake. CASIA-B* includes over 35.0% cross-dressing sample pairs, all of which are ups-changing, which improve BigGait's resistance to ups-changing (UP), as shown in Tab. 4 (b). CCPG has more diverse outfits than others, containing over 95.0% cross-dressing pairs, with 78.5% for ups-changing, 80.0% for pants-changing, and 62.0% for full-changing. As shown in Tab. 4 (a), BigGait can learn robust gait representations by training on CCPG. These analyses indicate that the distribution of training data can significantly influence BigGait's outcomes, *i.e*, the more cross-dressing changes given, the more cross-dressing capacity obtained.

Visualization of Activation Maps. To delve deeper into the gait representation learned by BigGait, Fig. 6 shows visualizations of activation maps from both BigGait and video-based ReID methods [18, 46, 48], driven by the popular Grad-CAM [38] algorithm. These maps are from the first layer in BigGait's downstream model, second layer in AP3D [18] and PSTA [46], and fourth block in

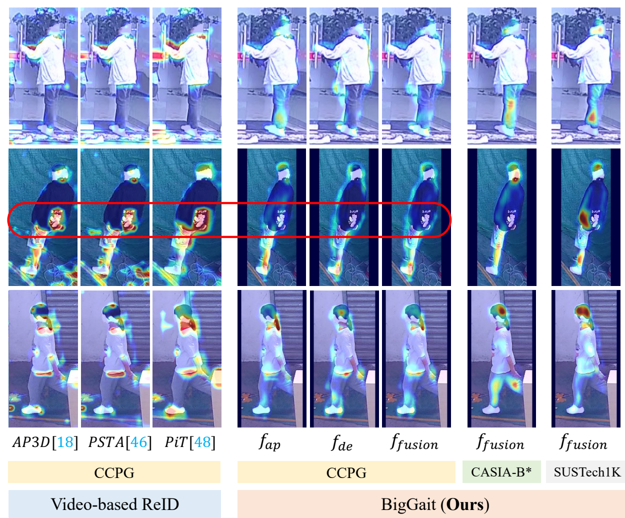

Figure 6. Visualization of activation maps. By background removal and texture denoising, BigGait pay more attention on robust gait pattern than video-based ReID methods.

PiT [48]. Four key insights emerge: a) Unlike video-based ReID methods distracted by cluttered backgrounds, BigGait solely attends to the body. b) The video-based ReID methods are sensitive to high-frequency clothing texture, as shown in the red box of Fig. 6. c) f_{ap} and f_{de} target different body regions, with f_{ap} emphasizing limbs and head via parsing-like priors, and f_{de} highlighting human contours by suppressing texture noises on body. Fusing f_{ap} and f_{de} using attention weights to obtain f_{fusion}. d) As mentioned before, CASIA-B* only contains ups-changes, and SUSTech1K has few cloth-changes. Thus BigGait incorrectly highlights unchanged pants and clothing when trained on CAISA-B* and SUSTech1K, respectively. More visualizations are provided in Supplementary Material.

4.3. Ablation Study

All experiments are conducted on CCPG. Due to limited GPU resources, we adopt 8 frames per sequence in the ablation study to train BigGait for more efficient iterations. Despite slight performance drop, BigGait's performance in Tab. 5 (a) remains comparable with SoTA methods in Tab. 3.

Pad-and-Resize. Compared (a) and (b) in Tab. 5, the Pad-

Table 5. Ablation of each branch and Pad-and-Resize: w/ and w/o mask branch, denoising branch, appearance branch and Pad-and-Resize strategy. ✓* denotes direct feeding to the downstream.

Index	Mask	Denoising	Appearance	Pad-and-Resize	CL	UP	DN	BG
(a)	✓	✓	✓	✓	**76.0**	**79.1**	**84.2**	93.0
(b)	✓	✓	✓	×	70.5	72.9	82.8	93.0
(c)	×	×	×	✓	55.6	62.7	73.7	87.8
(d)	×	✓	✓	✓	70.5	76.1	80.9	92.2
(e)	✓	×	✓	✓	66.1	71.6	80.7	**93.2**
(f)	✓	✓	×	✓	65.1	68.4	79.2	90.1
(g)	✓*	×	×	✓	36.7	46.8	52.4	46.4

and-Resize strategy is helpful for BigGait by faithfully preserving body aspect ratio.

Various Branch. In Tab. 5 (c), without the GRE module, the performance of BigGait drops to be similar to video-based ReID methods [18] in Tab. 3. This highlights the effectiveness of GRE module. Compared (a) to (d-f) in Tab. 5, each branch contributes substantially. This implies a mutually beneficial relationship among them. In Tab. 5 (g), BigGait's foreground mask is less discriminable for recognition compared to the silhouette in Tab. 3. We consider that BigGait's foreground masks is approximate regions rather than precise segmentations. SoTA results in Tab. 5 (a) are achieved even without finely segmented foregrounds.

Denoising Branch. Comparing (a) to (b-d) in Tab. 6, smoothness loss alone leads to a trivial solution, while combining it with diversity constraint improves results. Compared (a) with (e) and (f) in Tab. 6, we explore an appropriate channel number for the denoising branch.

Upstream and Downstream Models. Compared (c) with (d) in Tab. 7, BigGait-L exhibits slightly higher performance than BigGait-S. Compared (c) with (e) in Tab. 7, we verify that the gains primarily come from the knowledge embedded in DINOv2 rather than its model architecture. Compared (c) and (f) in Tab. 7, the performance suffers when fine-tuning the upstream DINOv2. Further, we explore the influence of different upstream/downstream models for BigGait. Compared (c) and (g) in Tab. 7, although the performance drops when changing BigGait's downstream model, it still surpasses its silhouette-based counterpart in (a). As shown in Tab. 7 (h), BigGait remains effective even when changing BigGait's upstream model, *i.e.*, replacing DINOv2 [35] with SAM [23]. In most cases, it still outperforms its silhouette-based counterpart in (b). This highlights BigGait's flexibility for upstream/downstream models.

5. Challenges and Limitations

Challenges for LVMs-based Gait Recognition. Beyond attractive profits brought by BigGait, this paper also reveals two primary challenges for LVMs-based gait recognition: a) *Interpretability.* While we have introduced several explicit human gait priors as soft constraints, the learned gait representations, in comparison to conventional ones defined by clear and intuitive physical attributes, remain partially understandable. b) *Purity.* In gait methods that directly utilize RGB videos as input, a recurring challenge involves effectively reducing gait-irrelevant noise within walking sequences. This task becomes even more demanding when attempting to preserve the purity of gait characteristics in LVMs-based gait recognition without explicit supervision.

Table 6. Ablation of denoising branch: exploring channel number, and w/ and w/o smoothness and diversity loss.

Index	Dimension	Smoothness	Diversity	CL	UP	DN	BG
(a)	16	✓	✓	**76.0**	79.1	**84.2**	93.0
(b)	16	×	×	69.1	73.7	81.3	91.4
(c)	16	✓	×	68.8	73.4	80.7	92.6
(d)	16	×	✓	74.5	**80.1**	83.1	93.2
(e)	8	✓	✓	72.8	77.3	83.8	**93.5**
(f)	32	✓	✓	68.5	73.4	81.2	92.0

Table 7. Ablation of upstream and downstream models.

Index		Upstream	Downstream	CL	UP	DN	BG
-	(a)	Segmentation	GaitSet	60.2	65.2	65.1	68.5
	(b)		GaitBase	71.6	75.0	76.8	78.6
BigGait (Ours)	(c)	DINOv2^{Frozen} (ViT-S/14)	GaitBase	76.0	79.1	84.2	93.0
	(d)	DINOv2^{Frozen} (ViT-L/14)	GaitBase	**79.0**	**82.3**	**86.7**	**94.5**
	(e)	DINOv2^{Scratch} (ViT-S/14)	GaitBase	36.6	44.0	62.4	79.3
	(f)	DINOv2$^{Fine\text{-}tune}$ (ViT-S/14)	GaitBase	74.8	78.1	82.6	92.7
	(g)	DINOv2^{Frozen} (ViT-S/14)	GaitSet	63.5	66.5	71.7	79.0
	(h)	SAM^{Frozen} (ViT-B/16)	GaitBase	71.2	76.0	84.8	93.5

Limitations of this paper. a) The influence of different upstream LVMs for BigGait is only preliminarily explored. This issue is worth further study. b) The GRE module lacks of spatial-temporal designs. c) This paper focus on high-frequency texture noises instead of low-frequency color noises, since it is the high-frequency that still heavily impacts existing RGB-based methods in the red box of Fig. 6. Further improvements to color noises are expected.

6. Conclusion

In this paper, we present an innovative and efficient methodology for the next-generation gait representation construction, termed BigGait, with gait guidance shifting from task-specific priors to LVMs-based all-purpose knowledge. Results on CCPG, CASIA-B* and SUSTech1K show that BigGait significantly outperforms the previous methods in both within-domain and cross-domain tasks in most cases, indicating that BigGait is a more practical paradigm for learning general gait representation. Moreover, this work discusses the challenge in LVMs-based gait recognition, which provides some possible directions on future research. The work may also provide inspiration to employ all-purpose knowledge produced by LVMs for other vision tasks.

Acknowledgements

This work was supported in part by Shenzhen International Research Cooperation Project (Grant No. GJHZ20220913142611021) and National Natural Science Foundation of China (Grant No. 61976144).

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020. 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 1

[3] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Cross-View Gait Recognition Through Utilizing Gait As a Deep Set. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (7):3467–3478, 2022. 1, 2, 6, 7

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 4

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 3

[6] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[9] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li. MetaGait: Learning to Learn an Omni Sample Adaptive Representation for Gait Recognition. In *European Conference on Computer Vision (ECCV)*, pages 357–374, 2022. 3

[10] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. GaitPart: Temporal Part-Based Model for Gait Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14233, 2020. 2, 3, 6

[11] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring Deep Models for Practical Gait Recognition. *arXiv preprint arXiv:2303.03301*, 2023. 6

[12] Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu. Learning Gait Representation from Massive Unlabelled Walking Videos: A Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45 (12):14920–14937, 2023. 3

[13] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. OpenGait: Revisiting Gait Recognition Towards Better Practicality. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 2, 3, 4, 6, 7

[14] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. SkeletonGait: Gait Recognition Using Skeleton Maps. In *Association for the Advancement of Artificial Intelligence (AAAI)*, page [in press], 2024. 5, 6

[15] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. GPGait: Generalized Pose-based Gait Recognition. In *International Conference on Computer Vision (ICCV)*, pages 19595–19604, 2023. 1, 3

[16] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14676–14686, 2021. 4

[17] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised Pretraining of Visual Features in the Wild. *arXiv preprint arXiv:2103.01988*, 2021. 2

[18] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In *European Conference on Computer Vision (ECCV)*, pages 228–243, 2020. 6, 7, 8

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders are Scalable Vision Learners. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3

[20] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1055–1059, 2020. 4

[21] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-Sensitive Temporal Feature Learning for Gait Recognition. In *International Conference on Computer Vision (ICCV)*, pages 12909–12918, 2021. 3

[22] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3D Local Convolutional Neural Networks for Gait Recognition. In *International Conference on Computer Vision (ICCV)*, pages 14920–14929, 2021. 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 3, 8

[24] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-Correction for Human Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6): 3260–3271, 2020. 4

[25] Weijia Li, Saihui ‘Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An In-Depth Exploration of Person Re-Identification and Gait Recognition in

Cloth-Changing Conditions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13824–13833, 2023. 2, 5, 6

[26] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective Kernel Networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019. 4

[27] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end Model-based Gait Recognition. In *Asian Conference on Computer Vision (ACCV)*, page [no pagination], 2020. 1, 2

[28] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. GaitEdge: Beyond Plain End-to-End Gait Recognition for Better Practicality. In *European Conference on Computer Vision (ECCV)*, pages 375–390, 2022. 1, 2, 3, 5, 6

[29] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-Based Temporal-Spatial Network (PTSN) for Gait Recognition with Carrying and Clothing Variations. In *Chinese Conference on Biometric Recognition (CCBR)*, pages 474–483, 2017. 1, 2

[30] Beibei Lin, Shunli Zhang, and Xin Yu. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. In *International Conference on Computer Vision (ICCV)*, pages 14648–14656, 2021. 2, 3

[31] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, Yiyang Su, Pegah Varghaei, Kai Wang, Xingguang Zhang, Stanley Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, and Xiaoming Liu. FarSight: A Physics-Driven Whole-Body Biometric System at Large Distance and Altitude. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 6227–6236, 2024. 1

[32] Kunliang Liu, Ouk Choi, Jianming Wang, and Wonjun Hwang. CDGNet: Class Distribution Guided Network for Human Parsing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4473–4482, 2022. 1

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34 (6):[no pagination], 2015. 1

[34] Mark S Nixon and John N Carter. Automatic Recognition by Gait. *Proceedings of the IEEE*, 94(11):2013–2024, 2006. 1

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4, 8

[36] Yunjie Peng, Kang Ma, Yang Zhang, and Zhiqiang He. Learning Rich Features for Gait Recognition by Integrating Skeletons and Silhouettes. *Multimedia Tools and Applications*, 83(3):7273–7294, 2024. 3, 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2, 3

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[39] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. LidarGait: Benchmarking 3D Gait Recognition With Point Clouds. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1063, 2023. 2, 6

[40] Irwin Sobel, Gary Feldman, et al. A 3x3 Isotropic Gradient Operator for Image Processing. *a talk at the Stanford Artificial Project*, pages 271–272, 1968. 5

[41] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 4

[42] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, 2018. 1

[43] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph Convolutional Network for Skeleton-Based Gait Recognition. In *International Conference on Image Processing (ICIP)*, pages 2314–2318, 2021. 2

[44] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a Deeper Understanding of Skeleton-based Gait Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1569–1577, 2022. 6

[45] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(10):3349–3364, 2020. 1, 4

[46] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In *International Conference on Computer Vision (ICCV)*, pages 12026–12035, 2021. 6, 7

[47] Shiqi Yu, Daoliang Tan, and Tieniu 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)*, pages 441–444, 2006. 2, 6

[48] Xianghao Zang, Ge Li, and Wei Gao. Multi-direction and Multi-scale Pyramid in Transformer for Video-based Pedestrian Retrieval. *IEEE Transactions on Industrial Informatics*, 18(12):8776–8785, 2022. 6, 7

[49] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial Transformer Network on Skeleton-based Gait Recognition. *Expert Systems*, 40(6):e13244, 2023. 6

[50] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait Recognition via Disentangled Representation Learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4710–4719, 2019. 1, 3

[51] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On Learning Disentangled Representations for Gait Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(1):345–360, 2022. 1

[52] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait Recognition in the Wild with Dense 3D Representations and A Benchmark. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20228–20237, 2022. 2, 3

[53] Jinkai Zheng, Xinchen Liu, Shuai Wang, Lihao Wang, Chenggang Yan, and Wu Liu. Parsing is All You Need for Accurate Gait Recognition in the Wild. In *The 31st ACM International Conference on Multimedia (ACM MM)*, page 116–124, 2023. 1, 2

[54] Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-Covariate Gait Recognition: A Benchmark. In *Association for the Advancement of Artificial Intelligence (AAAI)*, page [in press], 2024. 6