HybridGait: A Benchmark for Spatial-Temporal Cloth-Changing Gait Recognition with Hybrid Explorations

Yilan Dong[1], Chunlin Yu[1], Ruiyang Ha[1], Ye Shi[1], Yuexin Ma[1], Lan Xu[1], Yanwei Fu[2], Jingya Wang[1]*

[1]ShanghaiTech University
[2]Fudan University
yilandong@outlook.com; yanweifu@fudan.edu.cn; {yuchl, hary2022, shiye, mayuexin, xulan1, wangjingya}@shanghaitech.edu.cn

Abstract

Existing gait recognition benchmarks mostly include minor clothing variations in the laboratory environments, but lack persistent changes in appearance over time and space. In this paper, we propose the first in-the-wild benchmark CCGait for cloth-changing gait recognition, which incorporates diverse clothing changes, indoor and outdoor scenes, and multi-modal statistics over 92 days. To further address the coupling effect of clothing and viewpoint variations, we propose a hybrid approach HybridGait that exploits both temporal dynamics and the projected 2D information of 3D human meshes. Specifically, we introduce a Canonical Alignment Spatial-Temporal Transformer (CA-STT) module to encode human joint position-aware features, and fully exploit 3D dense priors via a Silhouette-guided Deformation with 3D-2D Appearance Projection (SilD) strategy. Our contributions are twofold: we provide a challenging benchmark CCGait that captures realistic appearance changes across an expanded and space, and we propose a hybrid framework HybridGait that outperforms prior works on CCGait and Gait3D benchmarks. Our project page is available at *https://github.com/HCVLab/HybridGait*.

Introduction

Gait recognition is an intriguing biometric task that aims to identify individuals based on their unique walking patterns. Unlike traditional biometric identifiers such as faces, fingerprints, and irises, gait information is a non-cooperative identifier as well as a non-invasive differentiator, making it a effective and secure option. Consequently, gait recognition has garnered significant interest in the research community, with potential applications in security, surveillance, crime analysis and forensic search.

Extracting reliable gait features while mitigating distractions like viewpoint, and pose variations has long been a challenge in the gait community. To date, the role of clothing and accessories has also gaining increasing attention towards more practical scenarios. More recently, the long-term vision (Xu and Zhu 2021) presents a new challenge for cloth-changing person re-identification, involving non-stationary cloth-changes over a long time scale. Unfortunately, recent advances in gait recognition (Yu, Tan, and Tan 2006; Hossain et al. 2010) typically have a short-term in-the-lab assumption that only a fixed gallery of clothes is imposed to mimic such variations at a fixed location, thereby disregarding the natural changes in appearance over time and space. The very recently proposed benchmarks (Zhu et al. 2021a; Zheng et al. 2022b) break the previous laboratory assumption by introducing an in-the-wild setup, however, they not only overlook the variations in clothing and accessories but also fail to account for changes over time.

To address the aforementioned issues, we present the first in-the-wild cloth-changing gait recognition benchmark, CCGait, which captures appearance changes over time and space. Our dataset can facilitate the development of more robust and resilient gait recognition systems that can perform effectively in real-world situations especially for cloth-changing scenarios. Critically, our proposed dataset possesses unique features that set it apart from pre-existing datasets (Zheng et al. 2022b; Zhu et al. 2021b; Takemura et al. 2018; Yu, Tan, and Tan 2006). Firstly, the raw videos were captured over a 92-day period, providing a diverse range of cloth changes occurring at various frequencies without human intervention. Secondly, our dataset includes footage from both indoor and outdoor surveillance systems, capturing individuals walking along different routes and at varying speeds. Lastly, the CCGait tool provides a range of multi-modal statistics, such as 2D silhouettes, 2D/3D keypoints, and 3D human meshes. These features make our dataset a valuable resource for researchers in the field, enabling them to analyze and evaluate gait recognition algorithms under different conditions and scenarios.

We look at the problem of cloth-changing gait recognition under in-the-wild scenarios. Previous works utilize 2D silhouette information for robust extraction of gait features, while more recently 3D human meshes have been exploited to leverage the invariant properties against clothing and viewpoint variations (Zheng et al. 2022b; Han et al. 2022). However, to enable the integration of human meshes with silhouettes, existing approaches typically adopt a simple and straightforward fusion strategy, which may be inadequate in bridging the modality gap and extracting mutual information from two modalities. On this ground, a more careful design is needed to unify the two sharply different representation structures, and make full use of 3D meshes.

To implement this vision, we propose a SMPL-aided hy-

*Corresponding author.

brid network comprising three branches, where a temporal branch and a projection branch are devised to assist the appearance branch. In the temporal branch, we develop the *Canonical Alignment Spatial-Temporal Transformer* (**CA-STT**) which serves a dual purpose: firstly, it captures the cloth-irrelevant intrinsic temporal dynamics of 3D human mesh; secondly, it constructs a canonical space to bridge the gap between the 3D human mesh and the 2D rest pose by explicitly aligning the kinematic and regular grid features. In the projection branch, we first project 3D meshes to 2D silhouettes at a specific view. The projection is executed prior to training, ensuring no additional training or inference time. While the 3D SMPL projected silhouettes can eliminate large viewpoint variations, they may lack crucial appearance details. Hence, we propose a deformable alignment strategy named *Silhouette-guided Deformation* (**SilD**), which enriches the semantic content of projection appearance by leveraging original silhouettes as intermediary guidance.

We summarize our **contributions** as follows:

- Firstly, we introduce a benchmark CCGait tailored for in-the-wild clothing-change gait recognition. This dataset is designed to facilitate pragmatic research in gait recognition, specifically addressing challenges associated with appearance changes over time and space.
- Secondly, we propose a novel hybrid framework HybridGait that leverages both the projected 2D appearance and temporal dynamics of 3D human mesh. Our framework includes a Canonical Alignment Spatial-Temporal Transformer to capture the clothes-irrelevant dynamics of gait in the temporal branch, and a Silhouette-guided Deformation to address variations in viewpoint in the projection branch.
- Finally, our experimental results demonstrate that HybridGait obtains consistent improvements for gait recognition on both CCGait and Gait3D benchmarks.

Related Work

Gait Recognition. Currently, the research lines of gait recognition can be roughly grouped into two categories: appearance-based methods and model-based methods. *Appearance-based methods* (Shiraga et al. 2016; Chao et al. 2019; Hou et al. 2020; Fan et al. 2020; Huang et al. 2021b; Lin, Zhang, and Yu 2021; Lin, Zhang, and Bao 2020; Zheng et al. 2022a; Huang et al. 2021a) learn gait representations directly from binary visual cues of silhouette sequences. While Chao *et al.* (Chao et al. 2019) and Hou *et al.* (Hou et al. 2020) utilize unordered silhouette sequences for gait recognition, recent works focus on the temporal dynamics with ordered silhouettes as input, either using 1D CNN to capture the temporal information (Fan et al. 2020) or extracting spatial-temporal information via 3D convolutions (Lin, Zhang, and Bao 2020; Huang et al. 2021a,b). Although appearance-based methods are concise and effective, they hold a strong assumption that appearance statistics of individuals only have moderate changes, which is not suitable for cloth-changing gait recognition. *Model-based methods* (Yam, Nixon, and Carter 2004; Yamauchi, Bhanu, and Saito 2009; Ariyanto and Nixon 2011; Teepe et al. 2021; Liao et al. 2020) aim at modeling the structure of the human body from pose information, where Liao *et al.* (Liao et al. 2020) use 3D keypoints as human prior knowledge, while Teep *et al.* (Teepe et al. 2021) employ 2D skeletons to represent walking patterns. More recently, Zheng *et al.* (Zheng et al. 2022b) make the first attempt to explore 3D human mesh in gait recognition, which shows great potential in reducing the effect of long-term perturbations. However, they ignore the temporal dynamics and the projected appearance information within the 3D human meshes and the naive fusion approach does not maximize the utilization of 3D mesh information. Zhu *et al.*(Zhu, Zheng, and Nevatia 2023) directly extract 3D mesh information from silhouette to mitigate view and clothing disruptions. However, the accuracy of the 3D mesh information heavily rely on the original silhouette's quality, potentially causing challenges in in-the-wild scenarios. In stark contrast, our approach aims to characterize the temporal aand representation of the SMPL model and explicitly handle extreme viewpoint variations.

Gait Recognition dataset. Current public datasets can be classified into two categories: in-the-lab and in-the-wild. Previous works mainly conduct experiments on two types of popular laboratory benchmarks such as CASIA series (Wang et al. 2003; Yu, Tan, and Tan 2006; Tan et al. 2006)) and OU-ISIR series (Hossain et al. 2010; Makihara, Mannami, and Yagi 2011; Iwama et al. 2012; Tsuji, Makihara, and Yagi 2010; Takemura et al. 2018; Uddin et al. 2018)). While a few existing benchmark efforts such as CASIA-B (Yu, Tan, and Tan 2006)) and OU-ISIR Cloth (Hossain et al. 2010) have initiated the research on cloth-changing gait recognition, they focus more on laboratory moderate cloth variations in constrained environments. Recently, several in-the-wild datasets have been built to fulfill the challenging real-world demands. GREW (Zhu et al. 2021b) is collected by hundreds of cameras set in public places only within a single day, resulting in restricted clothing variations. Gait3D (Zheng et al. 2022b) builds the first in-the-wild gait recognition dataset with 3D mesh modalities, enabling the community to explore dense 3D representations for the gait. However, Gait3D is collected in a location where individuals are less likely to repeat their presence frequently. As a result, it doesn't fulfill the criteria for real-world clothing change scenarios. To address this gap, we have constructed a new dataset that specifically focuses on clothing changes for gait recognition in the wild.

The CCGait Dataset

In order to facilitate research on gait recognition, we introduce a new Cloth-Changing gait recognition dataset called CCGait. This dataset offers several distinct features compared to existing datasets, which are listed in Table 1. Specifically, the CCGait dataset was collected over a period of 92 days, during which time we captured data on individuals exhibiting a wide range of attire and appearances. This diverse collection of data is particularly useful for developing and testing gait recognition algorithms that are robust to variations in clothing, footwear, and other physical attributes. In addition to its temporal diversity, the CCGait dataset offers

Dataset	IDs	Tracklets	Views	Data Tpyes	Environment	Time(Day)	Cloth Change	No Human Intervention
CCVID	226	2,856	1	RGB	Outdoor	-	✓	✗
CCPG	200	16,566	10	RGB, Sil.	Indoor & Outdoor	-	✓	✗
CASIA-B	124	13,640	11	RGB, Sil.	Indoor	-	✓	✗
OU-ISIR Cloth	68	2,764	1	Sil.	Indoor	-	✓	✗
OU-LP Bag	62,528	187,584	1	Sil.	Indoor	-	✓	✗
OU-MVLP	10,307	288,596	14	Sil.	Indoor	-	✗	✗
FVG	226	2,856	1	RGB	Outdoor	-	✓	✗
GREW	26,345	128,671	882	Sil., Pose, Flow	Outdoor	1	✓	✓
Gait3D	4,000	25,309	39	Sil., Pose, Mesh	Indoor	7	✗	✓
CCGait	1,495	4,824	9	Sil., Pose, Mesh	Indoor & Outdoor	92	✓	✓

Table 1: Comparison of publicly available datasets for video-based cloth-changing person re-identification and gait recognition.

spatial diversity as well. Specifically, the dataset includes footage captured from both indoor and outdoor surveillance systems, providing a variety of environments in which people walk at varying speeds and along different routes. This variability in walking conditions is particularly valuable for researchers seeking to develop gait recognition algorithms that can operate effectively in real-world settings. Finally, the CCGait dataset includes a range of multi-modal statistics that researchers can use to develop and test their algorithms. Specifically, the dataset provides 2D silhouettes, 2D/3D keypoints, and 3D human meshes, allowing researchers to evaluate the performance of their algorithms across multiple modalities. This multi-modal approach can help to improve the robustness and reliability of gait recognition algorithms in a variety of settings.

Data Collection. We collected raw video footage from nine cameras located both outside and inside the building. As the locations we collected data from are places frequently visited by pedestrians on a daily basis, this greatly increases the probability of capturing the same individuals across a long period of time, even when they wear different clothing. The footage was captured at a resolution of 1,920 × 1,080 and at a frame rate of 25 FPS. For data collection, we chose to collect data during the peak pedestrian hours for 6 hours per day, 4 days per week, and collected data for over three-month period, totaling 2208 hours. Except for clothes change conditions, the CCGait dataset also involves occlusions by humans and objects. Still, it contains both indoor and outdoor sequences, as well as the sequences collected under various light conditions. All the above attributes contribute to the challenge of the CCGait dataset. Some examples of challenge cases can be seen in Figure 1. The recording has been authorized by the building administration and the collected data is solely for research purposes. Furthermore, in order to protect privacy, we will not release any RGB images.

Data Preprocessing and Annotation. Prior to annotation, we utilized FairMOT (Zhang et al. 2021), which had been fine-tuned to our dataset, to perform person tracking. We then engaged annotators to address any potential mistracking issues in the algorithm and to ensure that each sequence corresponded to only one person. Ultimately, we manually annotated a total of 4,824 sequences, resulting in 1,495 unique identities.

Generation of Gait Representations. The CCGait Dataset provides a comprehensive set of gait representations that include 3D SMPL parameters, 3D pose, 2D pose, and silhouette. Thanks to the rapid advancement of 3D pose and shape methods (Sun et al. 2021; Zhang et al. 2023; Li et al. 2023), we utilized ROMP (Sun et al. 2021), which allowed us to obtain accurate 3D models of the human body. For accurate estimation of body joints, we used PoseNet (Moon, Chang, and Lee 2019a) for 2D and 3D poses. This allowed us to obtain accurate estimates of body joint locations, which can be used to analyze human motion and detect abnormalities. To obtain a 2D silhouette, we employed HRNet-segmentation (Wang et al. 2020), which is a state-of-the-art method for semantic segmentation. To maintain accuracy, we kept the original resolution and aspect ratio of the frame during the generation. This helped to avoid any distortion or loss of information during the generation process. Some examples of multi-modal gait representations in our dataset can be found in Supplementary Material.

Data Statistics . The CCGait dataset is a collection of 4,824 sequences from 1,495 different individuals. The CCGait dataset is divided into train/test subsets with 1148/347 IDs, respectively. For the test set, we further randomly select one or two sequences from each ID to build the query set with 356 sequences, while the rest of the sequences become the gallery set with 727 sequences. More statistical details can be found in Supplementary Material.

Method

Method Overview. Taking advantage of both the model-based methods and appearance-based methods, we consider a hybrid approach that jointly takes the 3D SMPL models and silhouettes as our input. Concretely, our framework is equipped with three branches. (1) The basic *appearance branch* simply extracts body representations from silhouettes, which are susceptible to context perturbations. (2) The *temporal branch* takes the pose statistics of SMPL models as input and aims to obtain temporal representations. (3)

(a) Clothes change

(b) Occlusions

(c) Indoor & Outdoor

(d) Variation of light

Figure 1: Exemplary challenge frames extracted from the CCGait dataset are presented in this paper. Figure (a) displays the long-ter cloth changes observed for the same individual. In Figure (b), we showcase occlusions caused by both persons and objects. Figure (c) showcases diverse indoor and outdoor scenarios, while Figure (d) demonstrates variations in lighting conditions.

The *projection branch* takes the projected silhouettes from SMPL models as input to further help characterize the body representations. The overall architecture of our method is shown in Figure 2.

Appearance Branch. As the baseline of our approach, the appearance branch is adopted from (Zheng et al. 2022b), which directly takes silhouette sequences as input, denoted as $\{\boldsymbol{x}_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathbb{R}^{H\times W}$ represents the silhouette inputs, N is the sequence length. We denote the extracted shape embeddings of i-th frame from the appearance branch as $\boldsymbol{F} = \{\boldsymbol{f}_i\}_{i=1}^{N}$. In the following sections, we first introduce our elaborately-designed components: the Canonical Alignment Spatial-Temporal Transformer (CA-STT) in the temporal branch , and the Silhouette-guided Deformation Alignment in the projection branch Then, we introduce our training objectives and fusion strategy .

Canonical Alignment Spatial-Temporal Transformer

Inspired by (Zheng et al. 2021), we establish a spatial-temporal transformer framework to extract pose and motion characteristics. On top of that, a rest pose canonical space is introduced to align the model features with the appearance features in a semantically consistent way.

Spatial-Temporal Transformer. The goal of the spatial-temporal transformer is to learn the pose information across the frame scale and the motion statistics across the time scale. In the spatial transformer, the joint embeddings are first encoded into high-dimensional features, which are then fed into stacked spatial transformer layers. The output of the spatial transformer is denoted as: $\boldsymbol{F}_{sp}^{t} = \{S_i^t\}_{i=1}^{N}$. For the i-th frame, the embedding of each joint is represented as: $\boldsymbol{S}_i^t = \{\boldsymbol{f}_{ij}^t\}_{j=1}^{N_j}$, where $\boldsymbol{f}_{ij}^t \in \mathbb{R}^C$ refers to the j-th joint feature in the i-th frame, N_j is the SMPL joint number and C is the embedding dimension. Then, the joint features in each frame are averaged to obtain the frame representation. After that, frame embeddings are fed into temporal transformer layers to obtain the motion embedding $\boldsymbol{F}^t$.

Rest Pose Canonical Alignment. However, the model and shape embeddings is not semantic-consistent within the feature space. Inspired by previous works on 3D reconstruction, which utilize pixel alignment in a canonical space (He et al. 2021), we first introduce a Rest Pose Canonical space, where each human joint is assigned a coordinate within a fixed-sized 2D image. Subsequently, we present a novel approach termed Rest Pose Canonical Alignment, which involves scaling these coordinates to align with the target size image.

Given the i-th frame feature $\boldsymbol{S}_i^t$ generated by the spatial transformer, our goal is to transform $\boldsymbol{S}_i^t \in \mathbb{R}^{d\times N_j}$ into a target feature map $\boldsymbol{S}_i^{t'} \in \mathbb{R}^{d\times(h\times w)}$, with (h, w) denoting the predetermined resolution of the feature map. In our approach, we set $(h \times w)$ to the feature size of the appearance branch. However, a direct mapping lacks effective supervision. Observing that the regular grid structure retains the relative positioning of human joints, therefore, we convert this challenge into a task of finding the most pertinent human joint points for each pixel in the target space.

We first formulate the coordinate of the target patch regions as $\mathcal{R} = \{(h_r, w_r)\}_{r=1}^{N_r}$, where $N_r = h \times w$ is the number of target patch regions. Then the original rest pose canonical coordinate is defined as $\mathcal{C} = \{(h_j, w_j)\}_{j=1}^{N_j}$, where $h_j \in \{0, \cdots, H\}$ and $w_j \in \{0, \cdots, W\}$, H and W refers to the maximum values for the ordering of keypoints in the vertical and horizontal directions, respectively. After that, we can generate the proportionally scaled target canon-

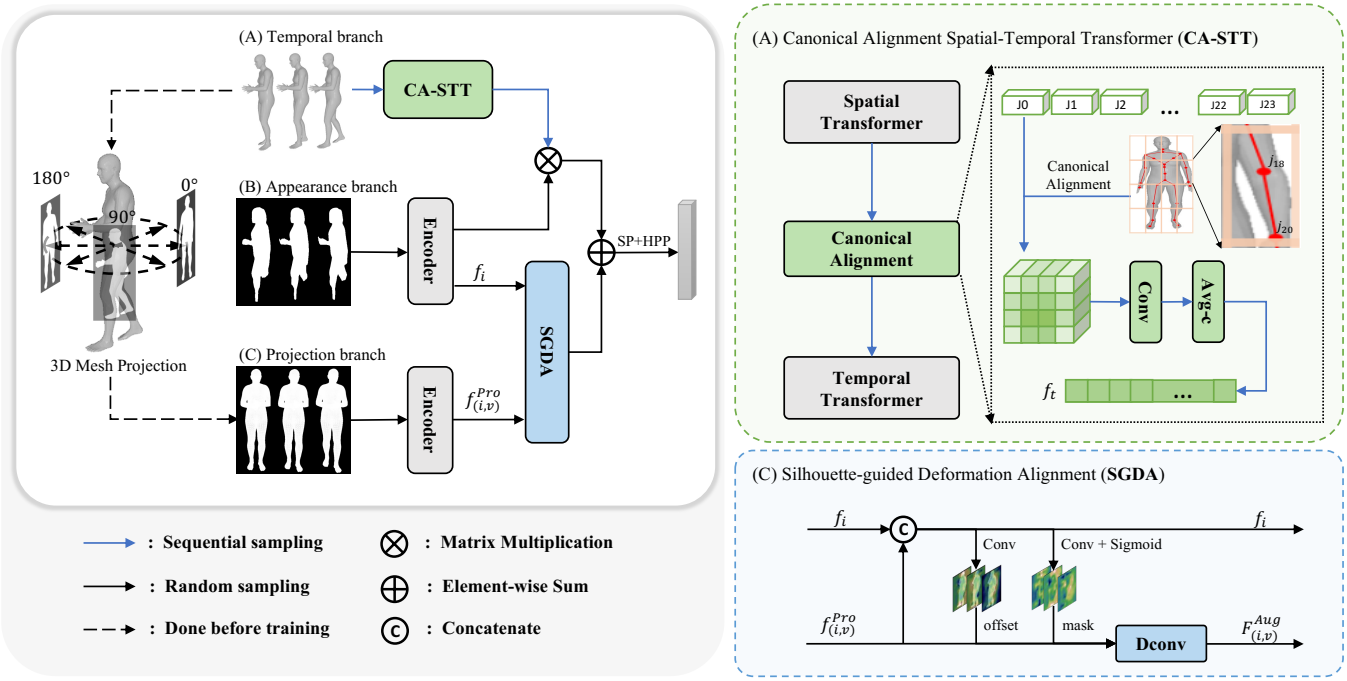

Figure 2: Our proposed framework comprises three key components. Component (B) is the basic appearance branch, responsible for extracting body representations from silhouettes, albeit susceptible to contextual perturbations. Component (A) is the temporal branch, which includes a 3D dynamic component, specifically the CA-STT model, to capture temporal information. Finally, component (C) is the Projection branch, which utilizes projected silhouettes from SMPL models as 3D-2D projection silhouettes to enhance the characterization of body representations.

ical coordinate, $\mathcal{C}' = \{(\frac{h_j \times h}{H}, \frac{w_j \times w}{W})\}_{j=1}^{N_j}$. Then, the rest pose canonical alignment can be formulated as:

$$\mathcal{A} = \left\{ (r, \{\omega_{j,r}\}_{j=1}^{N_J}) : r \in \mathcal{R},\ \omega_{j,r} \in \{0,1\} \right\}, \quad (1)$$

where $\{\omega_{j,r}\}_{j=1}^{N_J}$ can be interpreted as: for each patch region r, $\{\omega_{j,r}\}_{j=1}^{N_J}$ filters the irrelevant joints by enforcing the corresponding $\omega_{j,s} = 0$ and retaining semantic-relevant joints.

We employ k-nearest neighbors to retrieve the k closest joints to the coordinates of the region r:

$$\{\omega_{j,r}\}_{j=1}^{N_j} = \text{KNN}(\mathcal{R}_r, \mathcal{C}', k), \quad (2)$$

where KNN returns 1 if keypoint j is within the k-th closest joints, otherwise KNN returns 0.

To further obtain the semantic content of patch region r and subsequently achieve canonical alignments, we aggregate the related joint features using the obtained $\omega_{j,r}$. Formally, for each patch region $r \in \mathcal{R}$, we have:

$$(\boldsymbol{S}_i^{t'})^{(r)} = \text{Avg}(\sum_{j=1}^{N_j} \omega_{j,r} \boldsymbol{f}_{ij}^t). \quad (3)$$

Then, the transformed feature $\boldsymbol{F}_{sp}^{t'} = \{\boldsymbol{S}_i^{t'}\}_{i=1}^{N_j}$ is passed through a modulate block:

$$\boldsymbol{F}^{t'} = \text{IF}(\text{Avgpool\_c}(\text{Conv}(\boldsymbol{F}_{sp}^{t'}))), \quad (4)$$

where Avgpool_c refers to average pooling function along channel dimension, IF $(\cdot)$ refers to inverse flatten operation.

Notably, although the joints position is explicitly fixed, pose information is implicitly encoded by the spatial transformer, and then aligned to the 2D canonical space. Therefore, the temporal transformer can still learn the dynamic changes across time scales.

Silhouette-guided Deformable Alignment

To further enhance the shape information against viewpoint variations, we project the 3D SMPL model to uniform-viewed 2D silhouettes for all identities by (Moon, Chang, and Lee 2019b). The projected i-th silhouette is denoted as $x^{Pro}_{(i,v)}$, v is a specific viewpoint. Then, the projected i-th silhouettes features, $\boldsymbol{f}^{Pro}_{(i,v)}$, are encoded with a feature extractor.

However, the direct pixel-wise fusion can lead to the misalignment of body parts between distorted silhouettes and viewpoint-consistent projected silhouettes. To tackle the aforementioned concerns, we propose utilizing the original silhouette embedding f_i as a reference to guide the projected silhouette embedding $\boldsymbol{f}^{Pro}_{(i,v)}$ through the deformable convolution(Dai et al. 2017). This involves formulating the corresponding deformation offset and modulating mask as follows:

$$\boldsymbol{\Phi} = \text{Conv\_o}([\boldsymbol{f}_i, \boldsymbol{f}^{Pro}_{(i,v)}]), \quad (5)$$

$$\boldsymbol{m} = \text{Sigmoid}(\text{Conv\_m}([\boldsymbol{f}_i, \boldsymbol{f}^{Pro}_{(i,v)}])), \quad (6)$$

where $\mathbf{\Phi} = \{\Delta p_k | k = 1, ..., K\}$, K is the size of convolution kernel, $[\cdot]$ is the concatenation operation. Then a semantic-enriched projected silhouette can be aligned to the reference image:

$$\boldsymbol{f}_{(i,v)}^{Pro'}(p) = \sum_{p_k \in K} \omega(p_k) \cdot \boldsymbol{f}_{(i,v)}^{Aug}(p + p_k + \Delta p_k) \cdot \Delta m_k, \quad (7)$$

where $\boldsymbol{f}_{(i,v)}^{Pro'}$ is the output feature for the i-th frame, p is a single pixel in the feature map, p_k is the k-th regular offset of convolution kernal, Δp_k and Δm_k is the k-th learned offset and modulation scalar at location $p + p_k$. In this way, the adaptively learned offset will capture semantic correlation cues with the assistance of the original silhouette in a pixel-level alignment, and the output feature is $\boldsymbol{F}^{Pro'} = \{\boldsymbol{f}_{(i,v)}^{Pro'}\}_{i=1}^{N}$.

Training Objective

So far, with $\boldsymbol{F}$, $\boldsymbol{F}^{Pro'}$ and $\boldsymbol{F}^{t'}$, we produce the final embedding as following:

$$\hat{\boldsymbol{F}} = (\boldsymbol{F} \times \boldsymbol{F}^{t'}) + \boldsymbol{F}^{Pro'}, \quad (8)$$

where $\times$ is matrix multiplication. Finally, Set Pooling(SP) and Horizontal Pyramid Pooling(HPP)(Chao et al. 2019) are adopted to obtain the final feature vector.

In the training stage, our three-branch framework is trained in an end-to-end manner, and a combined loss is adopted. The combined loss is defined as

$$\mathcal{L} = \alpha \mathcal{L}_{tri} + \beta \mathcal{L}_{ce}, \quad (9)$$

where $\mathcal{L}_{tri}$ is the triplet loss, $\mathcal{L}_{ce}$ is the cross-entropy loss, α and β are weighting hyperparameters.

Experiments

Datasets and Evaluation Protocol. In addition to our proposed CCGait benchmark, we conducted a comparison with the Gait3D dataset (Zheng et al. 2022b), which is currently the largest gait recognition dataset with 3D representations. To ensure consistency with previous gait recognition datasets (Iwama et al. 2012; Hossain et al. 2010; Makihara, Mannami, and Yagi 2011; Tsuji, Makihara, and Yagi 2010; Uddin et al. 2018; Takemura et al. 2018; An et al. 2020; Li et al. 2022; Wang et al. 2003; Yu, Tan, and Tan 2006; Tan et al. 2006), we used the same evaluation protocol, which involves open-set instance retrieval. Specifically, we measured the similarity between a given query sequence and all sequences in the gallery set and reported the accuracy using the average Rank-1 and Rank-5 identification rates across all query sequences. To account for the retrieval of multiple instances and difficult samples, we adopted two additional evaluation metrics: mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP). These metrics provide a comprehensive evaluation of the performance of our proposed approach in comparison to the state-of-the-art methods.

Implementation Details. During the training process, we used the same configuration to train all models. The batch size is set as $32 \times 4 \times 30$, where 32 represents the number of IDs, 4 for the training sequences per ID, and 30 denotes the sequence length. The models were trained for 1200 epochs using an initial learning rate (LR)=1e-3 and the LR was reduced by a factor of 0.1 at the 200-th and 600-th epochs. We use the loss in Equ. 9 for training. The hyper-parameters in Equ. 9 are set as α=1.0 and β=0.1. And the hyper-parameters k in Equ. 2 is set to 7. We used the Adam optimizer (Kingma and Ba 2014) and set the weight decay as 5e-4. The spatial-temporal transformer, encoder structure and more details can be found in Supplementary Material.

Results and Analysis

We present a comprehensive comparison of nine state-of-the-art (SOTA) models for gait recognition. Specifically, we evaluate the performance of these models in three categories. In the model-based method category, we analyze GaitGraph (Teepe et al. 2021), which utilizes a 2D skeleton as a graph and inputs it into a Graph Convolution Network. In the appearance-based method category, we examine GEINet (Shiraga et al. 2016), GaitSet (Chao et al. 2019), GaitGL (Lin, Zhang, and Yu 2021), GaitPart (Fan et al. 2020), CSTL (Huang et al. 2021a), MTSGait (Zheng et al. 2022a) and GaitGCI(Dou et al. 2023). These models use Convolutional Neural Networks (CNNs) to learn features from various sources, including GEIs, silhouettes, and gait sequences. GaitPart divides a silhouette image into fixed parts to learn micro-motion features. CSTL learns both long-term and short-term motion for gait recognition, and MTSGait learns spatial features and multi-scale temporal dynamics. Furthermore, we examine SMPLGait (Zheng et al. 2022b), which belongs to neither the model-based nor the appearance-based method category. This model utilizes SMPL (Loper et al. 2023) in conjunction with a 2D silhouette to learn gait information.

The evaluation results of our CCGait Dataset are presented in Table 2. The findings indicate that: **(1)** the model-based method, GaitGraph, achieves only a 5.46% Rank-1 accuracy and 5.97% mAP, which is an unquestionably poor performance. **(2)** In contrast, appearance-based methods such as GaitSet, GaitPart, and GaitGL, which use 2D silhouettes as input, achieve significantly higher Rank-1 accuracies of 42.31%, 33.28%, and 33.78%, respectively, outperforming GaitGraph by a considerable margin. This demonstrates the potential of 2D silhouettes in learning temporal information. **(3)** Furthermore, SMPLGait, which utilizes 3D Mesh along with 2D silhouettes, achieves further improvement as the 3D Mesh helps to learn some Cloth-Changing information. Our HybridGait method also outperforms the other state-of-the-art methods, with a 2.52% boost in Rank-1 accuracy, 2.81% boost in Rank-5 accuracy, and higher mAP and mINP by 2.16% and 2.40%, respectively. These statistics indicate that our HybridGait method exhibits high efficiency and robustness.

Table 3 displays the evaluation results on the Gait3D dataset. We observe that the model-based method GaitGraph achieves a low accuracy, as it inevitably loses some useful gait information. The appearance-based methods, such as GaitSet, GaitPart, and GaitGL, which use 2D silhouettes as input, obtain better results. The recent SMPLGait, MTSGait

Methods	Rank-1	Rank-5	mAP	mINP
GaitGraph	7.46	15.82	8.97	5.61
GEINet	12.31	27.09	13.71	9.81
GaitSet	29.61	62.64	41.41	33.28
GaitPart	36.52	54.78	36.50	28.64
GaitGL	37.92	53.37	37.10	29.20
SMPLGait	48.60	68.54	49.14	40.72
HybridGait(Ours)	**51.12**	**71.35**	**51.30**	**43.12**

Table 2: Performance comparison of the state-of-the-art methods on CCGait Dataset.

and GaitGCI achieved further improvement due to their specific designed model. Our HybridGait method outperforms the base model SMPLGait, improving Rank-1 by 7.0% and mAP by 6.13%. This demonstrates that our method has an overall performance advantage in extracting and combining multi-modality (appearance and model information) for gait recognition. Note that we have not evaluated the MTSGait and GaitGCI methods on the CCGait dataset since their code is not publicly available.

In conclusion, our study provides a comprehensive comparison of SOTA gait recognition models. Our findings can serve as a guide for future research aimed at developing more robust models capable of handling variations in view angle and clothing changes over time.

Methods	Rank-1	Rank-5	mAP	mINP
GaitGraph	6.25	16.23	5.18	2.42
GEINet	5.40	14.20	5.06	3.14
GaitSet	36.70	58.30	30.01	17.30
GaitPart	28.20	47.60	21.58	12.36
GaitGL	29.70	48.50	22.29	13.26
CSTL	11.70	19.20	5.59	2.59
SMPLGait	46.30	64.50	37.16	22.23
MTSGait	48.70	67.10	37.63	21.92
GaitGCI	50.30	68.50	39.50	24.30
HybridGait(Ours)	**53.30**	**72.00**	**43.29**	**26.65**

Table 3: Performance comparison of the state-of-the-art methods on Gait3D Dataset.

Ablation Study

To further evaluate the different components in the model, we conduct ablation studies on both Gait3D and our CCGait dataset. The findings, as shown in Table 5, confirm that the HybridGait approach is more effective in overcoming the challenges associated with gait recognition in real-world settings.

Effect of Temporal Branch. Table 4 illustrates that using only the basic appearance branch (Appr) resulted in a Rank-1 accuracy of 42.90% and mAP of 35.19% on the Gait3D dataset. However, fusing the basic appearance branch with the temporal branch, which utilizes a Spatial-Temporal Transformer (STT), improved the Rank-1 accuracy to 48.90% and increased the mAP to 39.90%. Subsequently, replacing the STT with our proposed Canonical Alignment Spatial-Temporal Transformer (CA-STT) resulted in further accuracy improvements as a 50.26% Rank-1 accuracy and 41.36% mAP. Furthermore, when combined with the appearance branch, the CA-STT achieved a Rank-1 accuracy of 50.28% with 50.41% mAP on our CCGait dataset, demonstrating the Temporal branch's ability to effectively capture temporal information from gait data.

Methods	Gait3D		CCGait	
	R-1	mAP	R-1	mAP
Appr	42.90	35.19	44.31	36.73
Appr + STT	48.90	39.90	47.36	50.08
Appr + CA-STT	**50.26**	**41.36**	**50.28**	**50.41**

Table 4: Effect of the Temporal Branch with 3D Dynamics.

Effect of Projection Branch. As presented in Table 5, the inclusion of Projection Branch (SilD) into the basic appearance branch (Appr) leads to better performance, thus validating the effectiveness of SilD in learning more informative appearance features. Furthermore, the combination of SilD with our temporal branch results in a noticeable improvement in performance. Specifically, the projection branch with front view (0°) achieves a 3.41% boost in Rank-1 accuracy on Gait3D dataset and a 2.35% boost on CCGait dataset. Notably, when SilD is added to our proposed CA-STT, our method achieves state-of-the-art performance, highlighting the advantage of our method in gait recognition under challenging conditions.

Methods	Gait3D		CCGait	
	R-1	mAP	R-1	mAP
Appr	42.90	35.19	44.31	36.73
+ SilD	46.31	37.25	46.66	39.18
+ STT + SilD	52.10	41.98	50.35	50.84
+ CA-STT + SilD (Full)	**53.30**	**43.29**	**51.12**	**51.30**

Table 5: Effect of the Projection Branch.

Conclusion

In this paper, we propose the first in-the-wild cloth-changing gait recognition dataset, named CCGait, to facilitate the research community including diverse appearance changes over space and temporal, and other in-the-wild challenges, such as occlusion, in door & out door, lighting variations. To address the challenges posed by contextual perturbations, we further propose a hybrid approach that fully utilizes a 3D human mesh in both the Temporal branch and the Projection branch. Extensive experiments validate the efficacy of our method on a wide range of benchmarks.

Acknowledgments

This work was supported by Shanghai Sailing Program (21YF1429400, 22YF1428800), Shanghai Local College Capacity Building Program (23010503100), NSFC (No.62303319), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Shanghai Clinical Research and Trial Center and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

An, W.; Yu, S.; Makihara, Y.; Wu, X.; Xu, C.; Yu, Y.; Liao, R.; and Yagi, Y. 2020. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE transactions on biometrics, behavior, and identity science*, 2(4): 421–430.

Ariyanto, G.; and Nixon, M. S. 2011. Model-based 3D gait biometrics. In *2011 international joint conference on biometrics (IJCB)*, 1–7. IEEE.

Chao, H.; He, Y.; Zhang, J.; and Feng, J. 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8126–8133.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

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

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

Han, X.; Cong, P.; Xu, L.; Wang, J.; Yu, J.; and Ma, Y. 2022. LiCamGait: Gait Recognition in the Wild by Using LiDAR and Camera Multi-modal Visual Sensors. arXiv:2211.12371.

He, T.; Xu, Y.; Saito, S.; Soatto, S.; and Tung, T. 2021. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11046–11056.

Hossain, M. A.; Makihara, Y.; Wang, J.; and Yagi, Y. 2010. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6): 2281–2291.

Hou, S.; Cao, C.; Liu, X.; and Huang, Y. 2020. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, 382–398. Springer.

Huang, X.; Zhu, D.; Wang, H.; Wang, X.; Yang, B.; He, B.; Liu, W.; and Feng, B. 2021a. Context-sensitive temporal feature learning for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12909–12918.

Huang, Z.; Xue, D.; Shen, X.; Tian, X.; Li, H.; Huang, J.; and Hua, X.-S. 2021b. 3D local convolutional neural networks for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14920–14929.

Iwama, H.; Okumura, M.; Makihara, Y.; and Yagi, Y. 2012. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5): 1511–1521.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, J.; Bian, S.; Liu, Q.; Tang, J.; Wang, F.; and Lu, C. 2023. NIKI: Neural Inverse Kinematics with Invertible Neural Networks for 3D Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, X.; Makihara, Y.; Xu, C.; and Yagi, Y. 2022. Multi-View Large Population Gait Database With Human Meshes and Its Performance Evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2): 234–248.

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

Lin, B.; Zhang, S.; and Bao, F. 2020. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *Proceedings of the 28th ACM international conference on multimedia*, 3054–3062.

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

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Makihara, Y.; Mannami, H.; and Yagi, Y. 2011. Gait analysis of gender and age using a large-scale multi-view gait database. In *Computer Vision–ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part II 10*, 440–451. Springer.

Moon, G.; Chang, J. Y.; and Lee, K. M. 2019a. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10133–10142.

Moon, G.; Chang, J. Y.; and Lee, K. M. 2019b. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10133–10142.

Shiraga, K.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2016. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, 1–8. IEEE.

Sun, Y.; Bao, Q.; Liu, W.; Fu, Y.; Black, M. J.; and Mei, T. 2021. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11179–11188.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications*, 10: 1–14.

Tan, D.; Huang, K.; Yu, S.; and Tan, T. 2006. Efficient night gait recognition based on template matching. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, 1000–1003. IEEE.

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

Tsuji, A.; Makihara, Y.; and Yagi, Y. 2010. Silhouette transformation based on walking speed for gait identification. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 717–722. IEEE.

Uddin, M. Z.; Ngo, T. T.; Makihara, Y.; Takemura, N.; Li, X.; Muramatsu, D.; and Yagi, Y. 2018. The ou-isir large population gait database with real-life carried object and its performance evaluation. *IPSJ Transactions on Computer Vision and Applications*, 10(1): 1–11.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, L.; Tan, T.; Ning, H.; and Hu, W. 2003. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12): 1505–1518.

Xu, P.; and Zhu, X. 2021. Deepchange: A large long-term person re-identification benchmark with clothes change. *arXiv preprint arXiv:2105.14685*.

Yam, C.; Nixon, M. S.; and Carter, J. N. 2004. Automated person recognition by walking and running via model-based approaches. *Pattern recognition*, 37(5): 1057–1072.

Yamauchi, K.; Bhanu, B.; and Saito, H. 2009. Recognition of walking humans in 3D: Initial results. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 45–52. IEEE.

Yu, S.; Tan, D.; and Tan, T. 2006. 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)*, volume 4, 441–444. IEEE.

Zhang, J.; Shi, Y.; Ma, Y.; Xu, L.; Yu, J.; and Wang, J. 2023. IKOL: Inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129: 3069–3087.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11656–11665.

Zheng, J.; Liu, X.; Gu, X.; Sun, Y.; Gan, C.; Zhang, J.; Liu, W.; and Yan, C. 2022a. Gait recognition in the wild with multi-hop temporal switch. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6136–6145.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022b. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20228–20237.

Zhu, H.; Zheng, Z.; and Nevatia, R. 2023. Gait recognition using 3-d human body shape inference. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 909–918.

Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021a. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14789–14799.

Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021b. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14789–14799.