GaitDAN: Cross-view Gait Recognition via Adversarial Domain Adaptation

Tianhuan Huang, Xianye Ben, *Senior Member, IEEE,* Chen Gong, *Member, IEEE,*

Wenzheng Xu, Qiang Wu, *Senior Member, IEEE*, and Hongchao Zhou

***Abstract*—View change causes significant differences in the gait appearance. Consequently, recognizing gait in cross-view scenarios is highly challenging. Most recent approaches either convert the gait from the original view to the target view before recognition is carried out or extract the gait feature irrelevant to the camera view through either brute force learning or decouple learning. However, these approaches have many constraints, such as the difficulty of handling unknown camera views. This work treats the view-change issue as a domain-change issue and proposes to tackle this problem through adversarial domain adaptation. This way, gait information from different views is regarded as the data from different sub-domains. The proposed approach focuses on adapting the gait feature differences caused by such sub-domain change and, at the same time, maintaining sufficient discriminability across the different people. For this purpose, a Hierarchical Feature Aggregation (HFA) strategy is proposed for discriminative feature extraction. By incorporating HFA, the feature extractor can well aggregate the spatial-temporal feature across the various stages of the network and thereby comprehensive gait features can be obtained. Then, an Adversarial View-change Elimination (AVE) module equipped with a set of explicit models for recognizing the different gait viewpoints is proposed. Through the adversarial learning process, AVE would not be able to identify the gait viewpoint in the end, given the gait features generated by the feature extractor. That is, the adversarial domain adaptation mitigates the view change factor, and discriminative gait features that are compatible with all sub-domains are effectively extracted. Extensive experiments on three of the most popular public datasets, CASIA-B, OULP, and OUMVLP richly demonstrate the effectiveness of our approach.**

***Index Terms*—Gait Recognition, Hierarchical Feature Aggregation, Adversarial View-change Elimination, Adversarial Domain Adaptation.**

I. Introduction

GAIT is a kind of physical and behavioral biometric feature that depicts the walking pattern of human beings. Unlike other biometrics such as the face, fingerprint, and iris, gait can be easily captured at a distance without the cooperation of subjects and is hard to disguise, which gives it high potential in various surveillance applications [1, 2].

As an identification task in vision, the essential goal of gait recognition is to learn unique and invariant representations from gait sequences. However, in real-world scenarios, gait sequences suffer from external factors like carrying, clothing conditions, and camera viewpoint switching. It brings a significant challenge to gait recognition, especially to cross-view gait recognition as dramatic appearance differences can be introduced with viewpoint variances [3–5].

To tackle the challenge above, existing appearance-based cross-view gait recognition methods primarily fall into two categories: i) transformation-based and ii) elimination-based approaches. Methods in the first category usually learn the transformation relations between different views [6–8] or project the gaits from different views onto a common view [4, 9–11]. They tend to work well in cases where the transformation between views is included in the training data. However, such transformation is typically performed between two views and cannot be well extended to handle diverse view transformations. Methods in the second category intend to eliminate the view-change interference, and can be further split into two sub-categories: 1) brute force learning [1–3, 12–15]; and 2) decouple learning [16–19]. The former is dedicated to extracting discriminative gait representations irrelevant to view changes. To this end, diverse training data under different camera views are usually first mixed. Then, regardless of view differences, models are trained based on given person IDs with the support of diverse loss functions. Decouple learning intends to split the view information from the rest of the gait features to eliminate its interference. It either deliberately arranges the training data under different views or clearly decouples the view feature from the rest of the gait features. In such a way, the model can best learn the feature irrelevant to the camera view. Compared with transformation-based methods, elimination-based methods are more flexible and can be well generalized to diverse views. However, in brute force learning, the view itself, i.e., explicit view estimation or view-specific modeling, is ignored and underrated to some extent. And in decouple learning, the decoupling process involves feature decomposition and synthesis [16, 18] using generative adversarial networks or auto-encoders, which somehow damages the spatial-temporal

Manuscript received 25 April 2023; revised 1 August 2023 and 21 February 2024. This research was funded in part by the Natural Science Foundation of China (62322111, 62271289), the Natural Science Fund for Outstanding Young Scholars of Shandong Province (ZR2022YQ60), the Research Fund for the Taishan Scholar Project of Shandong Province (tsqn202306064), Shenzhen Science and Technology Program (JCYJ20230807094000001) and Jinan ”20 Terms of New Universities” Funding Project (202333035).

T. Huang, X. Ben, W. Xu and H. Zhou are with the School of Information Science and Engineering, Shandong University, Qingdao 266237, China (e-mail: huangtianhuan@mail.sdu.edu.cn; benxianye@gmail.com; xuwenzheng@mail.sdu.edu.cn; hongchao@sdu.edu.cn).

C. Gong is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, School of Computer Science and Engineering, Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: chen.gong@njust.edu.cn).

Q. Wu is with the School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: qiang.wu@uts.edu.au).

(Corresponding authors: Xianye Ben and Hongchao Zhou.)

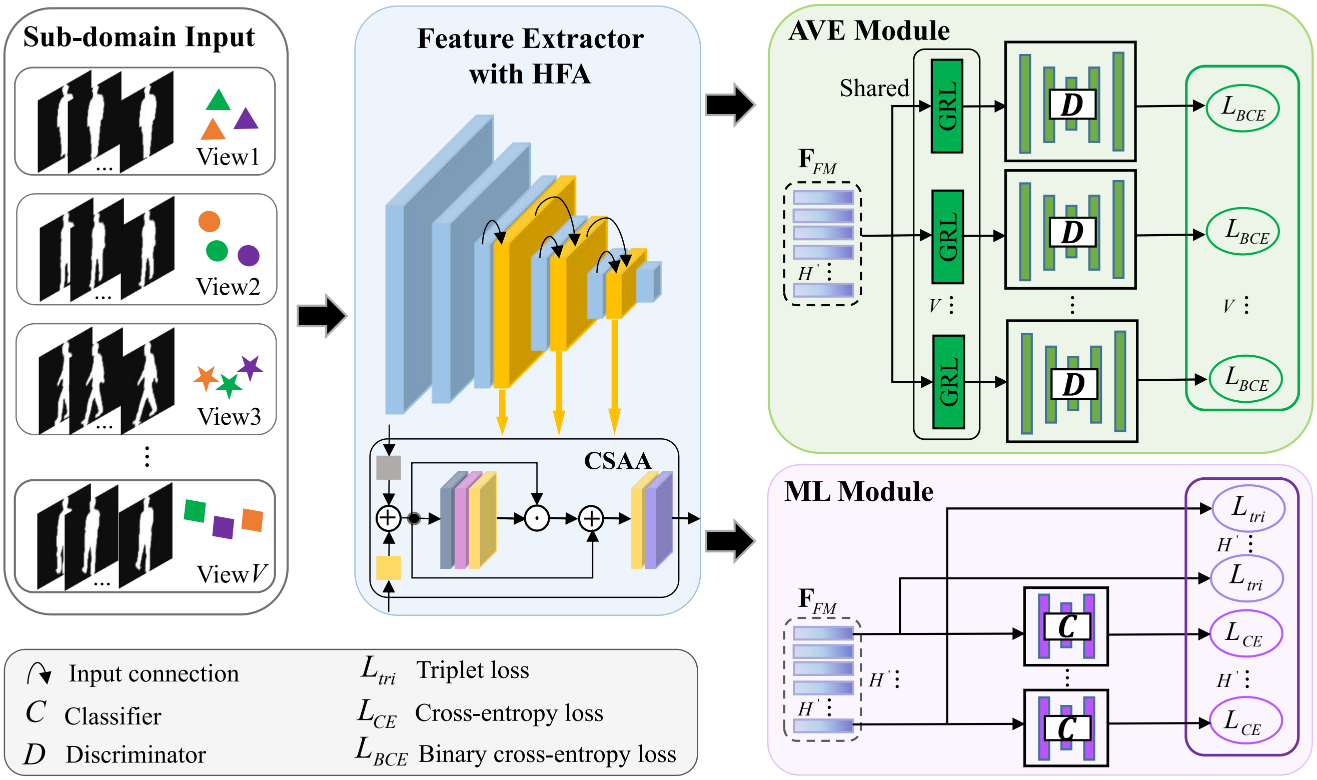


Fig. 1. The framework of the proposed GaitDAN. Our framework consists of a novel feature extractor with HFA strategy and two well-designed modules, i.e. the Adversarial View-change Elimination (AVE) module and Metric Learning (ML) module. ('▲', '♦', '●', '★' denote the samples from different sub-domains, and different colors indicate different IDs. The grey and orange squares in CSAA represent the two learnable parameters σ_1 and σ_2, respectively.)

feature in a gait sequence.

This work proposes a new approach to cross-view gait recognition that is regarded as a domain transfer problem. The gait information from different views is treated as the information from different sub-domains. The statistical distribution discrepancy caused by view changes is considered as a sub-domain shift. Thus, Domain Adaptation (DA) is adopted as the pipeline for the proposed approach. The key to successful adaptation is to learn a discriminative model that minimizes the distribution discrepancy between the source and target domains. In this work, DA does not consider one source domain against one target domain [20–22] but intends to simultaneously align gait information along multiple sub-domains. Consequently, the final feature representation of the gait for recognition is irrelevant to the view change. Inspired by unsupervised DA methods [20, 23–25], we adopt the Domain-adversarial Neural Network (DANN) [25] as the basic framework to address this challenge. The rationale for this choice stems from the fact that DANN offers several key advantages. Firstly, DANN matches the feature space distributions by modifying the feature representation itself, without considering the variation factors and complex decoupling operations behind different domains, which is more suited for our purpose of multiple sub-domain adaptation. Secondly, DANN performs feature learning and domain adaptation in a unified architecture and can be implemented using a simple back-propagation algorithm. Such a working mechanism enables the fully exploration of spatial-temporal information in gait sequences while eliminating the influence of view changes.

Therefore, we propose a novel gait domain-adversarial network (denoted as GaitDAN) for cross-view gait recognition. GaitDAN is able to learn discriminative and sub-domain-invariant gait features through end-to-end adversarial training, so that the final gait representations can be generalized well in all sub-domains. Fig. 1 illustrates the structure of GaitDAN, which consists of a novel feature extractor, an Adversarial View-change Elimination (AVE) module, and a Metric Learning (ML) module. The feature extractor is a new network with a specially designed Hierarchical Feature Aggregation (HFA) strategy, and is capable of extracting complementary spatial-temporal features of shallow-stage local detail information and high-stage semantic representation. As a result, more comprehensive spatial-temporal gait features can be obtained without losing subtle visual cues. The AVE module is the key adaptation component in GaitDAN that contains multiple view discriminators. It tries to challenge the gait feature generated by the feature extractor and distinguish them between the different sub-domains through an adversarial learning process. That is, the feature extractor intends to generate gait representation which is to fool the AVE. At the same time, the AVE feeds back to the feature extractor in the way of adversarial learning to generate a better sub-domain invariant gait feature to fool AVE. The ML module is introduced to further enhance the discriminability of gait representations in the feature space. In this way, high discriminability of the gait recognition task is guaranteed. By combining these components, the proposed

GaitDAN can produce sub-domain-invariant and discriminative gait features as the training progresses. More specifically, we make the following three major contributions.

- For the first time, we transform the view-change elimination into a domain adaptation problem, and propose a novel domain-adversarial network for cross-view gait recognition. It is in sharp contrast to current transformation-based or elimination-based methods, and makes it possible to take full advantage of spatial-temporal information while eliminating the influence of view changes. More impressively, it improves the performance of the model for cross-view gait recognition from completely unknown viewpoints.
- We propose a novel HFA strategy that can exploit comprehensive spatial-temporal information from various stages of the network and aggregate them hierarchically in a delicate attention way, which effectively enhances the discriminative capacity of the proposed method and ensures adequate mining of spatial-temporal information in gait sequences.
- We propose a simple yet effective view-change elimination method, i.e., the AVE module. By taking the advantage of sub-domain adversarial alignment, the AVE module can narrow the discrepancy across multiple view-level sub-domains in a simple way, which facilitates the end-to-end training of the whole network and further improves the robustness of gait representations.

The rest of this paper is organized as follows. Section II briefly introduces the related work. Section III explains the proposed GaitDAN in detail. In Section IV, the implementation details of GaitDAN are introduced. Meanwhile, the performance evaluation and detailed ablation study of GaitDAN are presented. Section V concludes the entire paper.

II. Related Work

In this section, we discuss related work on 1) appearance-based gait recognition, considering both transformation-based and elimination-based approaches, and 2) domain-adversarial-learning with GRL, wherein the latter inspired us to propose GaitDAN.

A. Cross-view Gait Recognition

Appearance-based cross-view gait recognition methods can be broadly classified into two categories, i.e., transformation-based and elimination-based methods.

Transformation-based methods deem the cross-view challenge as the problem of gait feature misalignment. These methods aggregate the silhouettes of a gait sequence into a template [26] or use silhouette sequences as input directly. They focus on directly learning transformations or projections between different views [4, 9–11, 27–29]. Then the gaits in one view can be transformed into another [29], or gaits in different views can be projected to a common subspace [4, 10, 11]. For instance, Makihara et al. [28] proposed a View Transformation Model to transform gait templates between views, while Ben et al. [10] proposed a Coupled Bilinear Discriminant Projection method to align gaits across different views by learning two sets of bilinear transformation matrices. To further effectively mitigate the feature misalignment between views, Xu et al. [27] proposed a Pairwise Spatial Transformer to register the gait features from different views to the target view simultaneously. However, this direct view transformation is constrained by the learned transformation models based on the current know camera views, and cannot sufficiently handle the view transformations across unknown views.

Elimination-based methods attempt to extract view invariant gait representations, and have shown state-of-the-art (SOTA) performance compared to transformation-based methods. There are two main approaches for these methods: brute force learning and decouple learning. Brute force learning typically treats the silhouettes of a gait sequence as a video. Regardless of different views, it intends to use robust spatial feature extraction and temporal modelling [1–3, 12–15, 30] to learn a strong gait representation that is irrelevant to the camera view. Various methods have been proposed. Under the constraint of triplet loss or cross-entropy loss, Chao et al. [1], Qin et al. [2] and Hou et al. [31] first extracted frame-level features from each silhouette independently and then applied temporal models such as the Max Pooling operation to encode the temporal information. Fan et al. [15] proposed a Micro-motion Capture Module to exploit the gait feature in a short-time period after spatial extraction. Analogously, Sepas-Moghaddam et al. [32] learned gait convolutional energy maps from frame-level features for temporal modeling and used an attention mechanism to focus on important recurrently learned gait representations. Chen et al. [33] conducted short-range and long-range temporal modeling to aggregate multi-features after frame-level spatial feature extraction, and utilized view assessment learning to improve the discriminability of aggregated features. Still, others [12, 13] directly extracted spatial-temporal features through a 3D convolution network. Although these methods have achieved encouraging success, the viewpoint information is still needs to be fully utilized, i.e., explicit view estimation or view-specific modeling which is ignored and underrated in these methods. Decouple learning is another approach for eliminating the influence of view changes on gait recognition. It also has corresponding advanced applications in action recognition tasks [34] and involves separating identity features and view features to obtain view-robust features. For example, Zhai et al. [17] adopted a newly designed auto-encoder to detach the identity features from the view features. Similarly, Zhang et al. [18] proposed the GaitNet to directly learn disentangled representations from gait videos. Yao et al. [16] proposed a group-supervised disentangle representation learning framework that explicitly decoupled the information in each gait sequence into pose, gait, appearance, and view features via an encoder-decoder architecture. These approaches effectively isolated the view-change information and obtained robust features that are invariable to the camera view. However, decouple learning relies on the generation and decomposition of gait sequences [16–19], which is a complex and challenging task. Current decoupling learning approaches either decompose and synthesize single-frame gait silhouette images before modeling the temporal information, or directly aggregate a gait sequence into a single image (e.g., Gait

Energy Image (GEI)) and then perform decomposition and synthesis on GEI. These operations will inevitably destroy the spatial-temporal information in gait sequences, and the errors created by the generation task are further accumulated. It significant limits the performance of gait recognition.

Therefore, the elimination-based method still presents open problems, although it demonstrates SOTA performance.

B. Domain-adversarial Learning with GRL

Domain-adversarial Learning (DAL) has emerged as a prominent technique in deep DA, and has first made the breakthrough in DA image classification [25, 35]. In DAL, a feature extractor and a domain discriminator are usually included with an adversarial objective. It is like Generative Adversarial Networks (GANs) [23–25]. The domain discriminator is trained to classifier whether the input sample is drawn from the source or target domain, while the feature extractor tries to confuse the domain discriminator to extract domain-invariant features. The optimization of the parameters of both the feature extractor and discriminator is achieved by maximizing and minimizing the domain discriminator's loss, respectively. Additionally, the label classification loss is minimized simultaneously to ensure that the extracted features possess high discriminability for original classification tasks.

The Gradient Reversal Layer (GRL) [35] was proposed for efficient adversarial training in unsupervised DA [20, 21, 25]. Except for a negative-parameter α which is not updated by the back-propagation, GRL has no parameters associated with the adversarial loss. By inserting a GRL between the feature extractor and domain discriminator, the maximization problem in the above adversarial loss can be automatically transformed into a minimized negative loss, ensuring a consistent optimization direction for the network and allowing the entire network to be routinely trained during forward and backward propagation. Specifically, during forward propagation, GRL acts as an identity transformation. During the back-propagation though, GRL takes the gradient from the subsequent network level, multiplies it by α and passes it to the preceding layer, which allows the gradient of the domain discriminant loss to be automatically inverted before back-propagating to the parameters of the feature extractor. So that, GRL implements a similar function to that performed in GANs for adversarial learning, and yields domain-invariant and discriminative features.

DAL with GRL has also been successfully applied to many other vision tasks [36–38]. For instance, He et al. [36] used GRL for adversarial learning and proposed an asymmetric tri-way Faster Region-Convolutional Neural Network (Faster-RCNN) for domain adaptive object detection, which fundamentally overcomes the source risk collapse caused by parameter sharing in general domain adaptive object detection methods and effectively ensures the adaptive safety of the detector. Niu et al. [37] proposed a novel feature fusion-and-alignment approach for remote sensing scene classification. By embedding GRLs into DAL to dynamically align the features of source and target domains, this method effectively improves the adaptive performance of features.

Motivated by such methods, the proposed GaitDAN is also based on DAL with GRL. In this paper, the gait information under different views is regarded as the data under different sub-domains, and the view-change mitigation problem is converted into a domain adaptation problem. To the best of our knowledge, GaitDAN is the first approach to cross-view gait recognition through adversarial domain adaptation.

III. The Proposed Method

In this section, we detail the proposed GaitDAN for cross-view gait recognition. We start with an overview of GaitDAN, followed by a description of key components including the feature extractor with the HFA strategy, AVE module and ML module, and end with the details of joint loss functions.

A. Overview

For supervised cross-view gait recognition, we have a labeled training set $\mathbf{X}_L$ which consists of V view-level sub-domains $\mathbf{X}^v = \{(\mathbf{x}_i^v, y_i^v)\}_{i=1}^{N_v}, v \in \{1, 2, ..., V\}$, such that each sample $\mathbf{x}_i^v$ in sub-domain $\mathbf{X}^v$ has a corresponding identity label $y_i^v \in \{1, 2, ..., P_v\}$. N_v and P_v are the numbers of samples and identities in the sub-domain $\mathbf{X}^v$, respectively. Meanwhile, the testing set $\mathbf{X}_T = \left\{\mathbf{x}_j^T\right\}_{j=1}^{N_T}$ contains N_T gait samples without identity labels from V different views. The goal of our proposed approach is to learn the discriminative gait features irrelevant to view changes through DA process.

The overall framework of the proposed GaitDAN is illustrated in Fig. 1. The gait silhouette sequences from different sub-domains are firstly input into a novel feature extractor G_F to extract fine-grained spatial-temporal features $\mathbf{F}_{FM}$. Then, to obtain view-invariant fine-grained features, the view adversarial learning procedure is incorporated into the network. It is a two-player game consisting of the feature extractor G_F and the AVE module G_{AVE}. The AVE module is trained to distinguish which sub-domain the input fine-grained gait features come from and the feature extractor G_F is fine-tuned simultaneously to confuse the AVE module. Specifically, the parameters $\mathbf{W}_F$ of feature extractor G_F are learned by maximizing the loss of AVE module, while the parameters $\mathbf{W}_{AVE}$ of AVE module are learned by minimizing the loss of AVE module. At the same time, the ML module including triplet loss and cross-entropy loss is applied to enhance the discrimination of fine-grained gait representations in the feature space. As a result, the objective of the overall framework can be formulated as:

$$\mathcal{L}(\mathbf{W}_F, \mathbf{W}_{AVE}) = \sum_{\mathbf{x}_i \epsilon \mathbf{X}_L} \mathcal{L}_{ML}(G_F(\mathbf{x}_i), y_i) - \beta \mathcal{L}_{AVE}(G_{AVE}(G_F(\mathbf{x}_i)), d_i), \tag{1}$$

where β is a trade-off parameter between two objectives that shape the gait feature learning. $\mathcal{L}_{ML}$ and $\mathcal{L}_{AVE}$ denote the loss of ML module and AVE module, respectively. d_i denotes the sub-domain labels of input samples. After the training convergence, the parameters $\hat{\mathbf{W}}_F$ and $\hat{\mathbf{W}}_{AVE}$ will deliver a saddle point of Eq. (1):

$$\begin{aligned} \hat{\mathbf{W}}_F &= \arg \min_{\mathbf{W}_F} \mathcal{L}\left(\mathbf{W}_F, \hat{\mathbf{W}}_{AVE}\right) \\ \hat{\mathbf{W}}_{AVE} &= \arg \max_{\mathbf{W}_{AVE}} \mathcal{L}\left(\hat{\mathbf{W}}_F, \mathbf{W}_{AVE}\right). \end{aligned} \tag{2}$$

Hence, the final gait representations that are discriminative and view-invariant can be obtained.

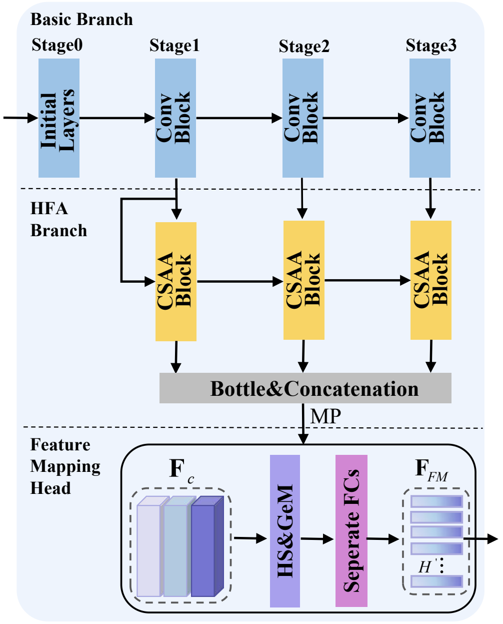


Fig. 2. The architecture of the Feature Extractor with HFA strategy.

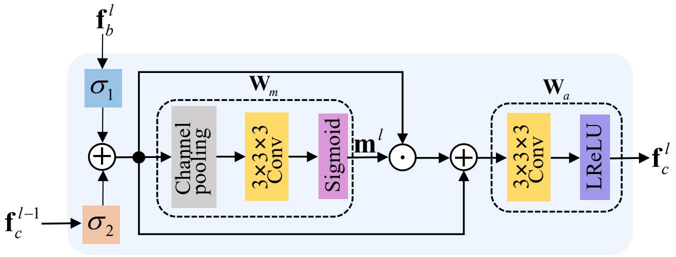


Fig. 3. The detailed architecture of the CSAA block. ⊕ and ⊙ mean the element-wise summation and multiplication, respectively.

B. Feature Extractor with the HFA Strategy

As illustrated in Fig. 2, the Feature Extractor with the HFA strategy is composed of two branches and a feature mapping head. The upper branch is the basic branch for extracting general spatial-temporal features, and can be implemented by any backbone. The following branch is the HFA branch. It takes the stage-specific features produced in basic branch as input, and is the main branch for progressively capturing comprehensive global spatial-temporal features through hierarchical feature aggregation. In the way, more integrated spatial-temporal gait features are extracted, which is more conducive to discriminative silhouette sequence-based gait recognition. Then, the extracted global features are mapped to the metric space by the feature mapping head to obtain part-based fine-grained gait representations.

In this part, the specific structure and detailed working mechanism of the basic branch and the HFA branch will be first introduced, and followed by the formulaic description of the feature mapping head. Note that in this section, gait samples from different sub-domains (views) are subjected to the same operation, so the superscript indicating the view is omitted for the convenience.

1) *Basic Branch:*

TABLE I
THE EXACT STRUCTURE OF THE BASIC BRANCH.(LEAKYRELU LAYER AFTER EACH 3D CONVOLUTION LAYER IS OMITTED.)

Stage	Layer name	In_channel	Out_channel	Kernel
Stage 0	Conv3D	1	32	(3, 3, 3)
	Conv3D	32	32	(3, 1, 1)
Stage 1	Conv3D	32	64	(3, 3, 3)
	Max Pooling	-	-	(1, 2, 2)
Stage 2	Conv3D	64	128	(3, 3, 3)
Stage 3	Conv3D	128	128	(3, 3, 3)

In this paper, a general 3D Convolutional Neural Network (CNN) is taken as the basic branch since previous SOTA works [1, 12–14] have proved that robust spatial-temporal representation is the key to the silhouette sequence-based gait recognition, and that 3D CNNs can bring great performance advantages. As shown in Fig. 2, the basic branch contains multiple network stages ('Stage 0', 'Stage 1', 'Stage 2' and 'Stage 3'), and each stage consists of initial layers or a convolution block. 'Stage 0' is the initial stage introduced to process the input gait sequences. 'Stage 1' to 'Stage 3' are different stages of the network used to extract shallow and high-level semantic information of the preprocessed input respectively. The extract structure of the network stages is listed in Tab. I.

2) *HFA Branch:*

The binarized nature of gait silhouette sequences, coupled with large appearance interference caused by view changes, results in subtle differences between subjects only at specific locations within the silhouette sequences. Therefore, it becomes crucial to utilize features extracted from the shallow stages of the network for accurate gait recognition, as they can encode the local regions in detail. Additionally, supplementing high-level features with low-level features can focus on more discriminative regions, thereby improving the feature's discriminability. Based on this, we introduce the Hierarchical Feature Aggregation (HFA) strategy to the feature extractor based on the basic branch in anticipation of obtaining more comprehensive spatial-temporal gait features. It is implemented by the HFA branch as shown in Fig. 2.

The core idea of HFA is to consider visual cues at different stages simultaneously. Nevertheless, there are distribution differences and semantic misalignment between different stage features. Direct aggregation [39, 40] like concatenation, summation or using bottle neck layers may lead to semantic confusion rather than achieving a positive complement. To this end, we introduce the attention mechanism and propose a Cross-Stage Attention Aggregation (CSAA) block in the HFA branch to incorporate cross-stage spatial-temporal features from different network stages in the basic branch. The detailed architecture of the CSAA block is shown in Fig. 3. It consists of two learnable parameters σ_1, σ_2, a cross-stage attention derivation operation $\mathbf{W}_m$, and a cross-stage attention aggregation operation $\mathbf{W}_a$.

Specifically, for input $\mathbf{x}$ from any sub-domain, general spatial-temporal features from two neighbour stages in basic branch are first combined by the learnable parameters, which can be formulated as:

$$\mathbf{f}_{ad}^{l} = \begin{cases} \mathbf{f}_{b}^{l}, l = 1 \\ \sigma_1 \mathbf{f}_{b}^{l} + \sigma_2 \mathbf{f}_{c}^{l-1}, l > 1, \end{cases} \tag{3}$$

where $\mathbf{f}_{ad}^{l}$ is the output of the element-wise weighted addition; $\mathbf{f}_{b}^{l} \in \mathbb{R}^{C_l \times T_l \times H_l \times W_l}$ $(1 < l \leq n)$ denotes the feature map extracted from basic branch at the l-th stage; C_l, T_l, H_l and W_l are the channels, frames, height and width of $\mathbf{f}_{b}^{l}$, respectively. As shown in Fig. 2, the first CSAA block in HFA branch takes $\mathbf{f}_{b}^{l}$ from "Stage 1" in basic branch as the input. For subsequent CSAA block at the l-th stage, the $\mathbf{f}_{b}^{l}$ from basic branch and the attention aggregated feature map $\mathbf{f}_{c}^{l-1}$ from the previous stage in HFA branch are taken as the input.

Then, a soft attention mask $\mathbf{m}^{l}$ indicating the importance of each position in $\mathbf{f}_{ad}^{l}$ is generated through the cross-stage attention derivation operation $\mathbf{W}_m$:

$$\mathbf{m}^{l} = \mathbf{W}_m \left(\mathbf{f}_{ad}^{l} \right), \tag{4}$$

where $\mathbf{W}_m$ is composed of a channel pooling, a $3 \times 3 \times 3$ convolution layer and a sigmoid layer. Subsequently, the generated $\mathbf{m}^{l}$ is also utilized to guide $\mathbf{W}_a$ to perform deep cross-stage attention aggregation, and the output feature map $\mathbf{f}_{c}^{l} \in \mathbb{R}^{\tilde{C}_l \times T_l \times H_l \times W_l}$ can be expressed as:

$$\mathbf{f}_{c}^{l} = \mathbf{W}_a \left(\mathbf{f}_{ad}^{l} \oplus \mathbf{f}_{ad}^{l} \odot \mathbf{m}^{l} \right), \tag{5}$$

where $\oplus$ and $\odot$ denote the element-wise summation and multiplication, respectively. $\mathbf{W}_a$ contains another $3 \times 3 \times 3$ convolution layer and a Leaky Relu layer. It is worth noting that the CSAA block takes into account the differences between different stage features in the basic branch and thus generates more accurate attention masks. With the guidance of the soft attention masks, the initial combined features can be further aggregated and more discriminative cross-stage spatial-temporal features can be extracted. By utilizing this two-step attention aggregation approach, CSAA effectively alleviate the misalignment of heterogeneous features from different stages. It is quite different from the commonly used operation of directly aggregating global general features along multi-stage. In addition, CSAA generates attention masks in the spatial-temporal domain for the initially aggregated cross-stage features via $\mathbf{W}_m$, and forgoes explicit modeling of channel interdependencies. Since the input for gait recognition is simple binary silhouette sequences lacking color and texture information, the channel weights cannot accurately reflect the importance of the channels, but may instead introduce noise and interfere with the original feature extraction, especially for shallow gait feature maps. If the channel-attention methods are introduced, it will in turn lead to performance degradation [14]. Compared to channel-attention methods, CSAA takes into account the spatial-temporal properties of gait sequences, in which case, critical spatial-temporal information can be activated by $\mathbf{W}_m$ in CSAA. Thus, more comprehensive cross-stage spatial-temporal features can be obtained.

To further encourage the semantic complement for high stage ones, a hierarchically dynamic fusion from the lower to the higher stages is employed. The bottle neck layers are utilized to adjust the channels of feature maps from different CSAA blocks. After that, resized feature maps are concatenated along the channel and a Max Pooling (MP) operation is utilized to generate the final global spatial-temporal feature $\mathbf{F}_c \in \mathbb{R}^{C' \times H' \times W'}$, which can be formulated as:

$$\mathbf{F}_c = \text{MP} \left[\mathbf{f}_c^{1}; \mathbf{f}_c^{2}; ...; \mathbf{f}_c^{n} \right], \tag{6}$$

where n denotes the number of CSAA blocks.

3) *Feature Mapping Head:*

The feature mapping head is introduced to obtain more discriminative fine-grained features. The global feature obtained from the HFA branch is firstly horizontally sliced (HS). Then the Generalized-Mean pooling (GeM) [41] is used to extract refined features from each horizontal strip as follows:

$$\mathbf{F}_{GeM} = \mathbf{W}_{GeM} \left(\mathbf{F}_c' \right), \tag{7}$$

$$\mathbf{W}_{GeM} \left(\mathbf{F}_c' \right) = \left(\mathbf{W}_{Avg} \left(\left(\mathbf{F}_c' \right)^{r} \right) \right)^{\frac{1}{r}}, \tag{8}$$

where $\mathbf{F}_{GeM} = \left\{ \mathbf{f}_{GeM}^{h} \left| h = 1, 2, ..., H' \right. \right\} \in \mathbb{R}^{C' \times H'}$ is the fine-grained part-based feature after GeM and $\mathbf{F}_c' = \left\{ \mathbf{f}_c'^{h} \left| h = 1, 2, ..., H' \right. \right\} \in \mathbb{R}^{C' \times H' \times W'}$ is the horizontal sliced feature, i.e., $H^{'}$ part-based features in total in single $\mathbf{F}_c'$ or $\mathbf{F}_{GeM}$. $\mathbf{W}_{GeM}(\cdot)$ and $\mathbf{W}_{Avg}(\cdot)$ denote the GeM pooling and Averaged pooling, respectively. The parameter r can be optimized during the training phase.

Subsequently, for each horizontal sliced feature $\mathbf{f}_{GeM}^{h}$ in $\mathbf{F}_{GeM}$, separate Fully Connected (FC) layers are employed to map the part-based gait features into a more discriminative representation space [1, 15], which can be presented as:

$$\mathbf{F}_{FM} = \mathbf{W}_{SFC} \left(\mathbf{F}_{GeM} \right), \tag{9}$$

where $\mathbf{F}_{FM} = \left\{ \mathbf{f}_{FM}^{h} \middle| h = 1, 2, ..., H' \right\} \in \mathbb{R}^{C' \times H'}$ is the output after separate feature mapping. $\mathbf{W}_{SFC}$ denotes the separate FC mapping operation.

C. *Adversarial View-change Elimination Module*

Based on the domain adaptation theory [42], a good representation of the subject for the case of cross-domains is the one by which an model cannot identify the domain origin information. The AVE module aims to reduce the distribution differences between sub-domains without specifying any particular source or target domains. Unlike the general domain adaptation problem that involves only two domains, the situation of cross-view gait recognition based on silhouette sequences is complex, involving multiple different sub-domains on one hand, and complex scene variations such as wearing and carrying situations on the other hand. Therefore, the transformation from different sub-domains to the domain invariant space is not the same. In this regard, a stepwise, refined domain adaptation method that allows samples from each sub-domain to learn their respective transformations to the domain invariant space is designed. This results in a gradual decrease in the domain offset between each sub-domain and the other sub-domains, and ultimately leads to a decrease in the difference between all sub-domains.

As shown in Fig. 4, the AVE module consists of multiple view discriminators with a shared Gradient Reverse Layer (GRL). In particular, a binary (1 v.s. others) discriminator is designed for each specific view in the AVE module. Additionally, an adversarial objective is developed to train the feature extractor and these discriminators concurrently in a min-max way. The minimum process of the discriminators' losses enables them to distinguish whether each gait input originating from this sub-domain or not, while the maximum process of their losses aims to confuse these discriminators for removing the sub-domain difference. Consequently, each sub-domain in the AVE module is treated as a temporary target domain, while others are treated as source domains. And adversarial learning is used to reduce the differences between the source and target domains. By iterative training, gait information under different camera views are finally mapped to a common embedding space where gait features cannot be discriminated between multiple sub-domains.

Specifically, for a view discriminator D_v with weight parameters $\mathbf{W}_D^v$, each part-based feature $\mathbf{f}_{FM}^h \in \mathbb{R}^{C'}$ of input sample $\mathbf{x}$ after normalization is first separately input to D_v through GRL, and the corresponding output of D_v is then fed into a softmax layer to obtain the probabilistic output $\mathbf{z} \in \mathbb{R}^2$. The process can be denoted as:

$$\mathbf{z}\left(\mathbf{f}_{FM}^h\right) = \varphi\left(\left(\mathbf{W}_D^v\right)^\top GRL\left(\frac{\mathbf{f}_{FM}^h}{\left\|\mathbf{f}_{FM}^h\right\|}\right)\right), \tag{10}$$

where φ denotes a softmax function. As mentioned in Sec. III-B3, each input sample has H' part-based features. Thus, H' probabilistic outputs of the input $\mathbf{x}$ in total. The minimum process of the view discriminator D_v is then trained by a binary cross-entropy loss defined on all the part-based features as:

$$\begin{aligned}\min_{\mathbf{W}_D^v}\mathcal{L}_{D_v}\left(\mathbf{X}_L,\mathbf{z},G_F\right) = \min_{\mathbf{W}_D^v}\Big[\tfrac{1}{H'}\textstyle\sum_{h=1}^{H'}\left\langle\mathcal{L}_{BCE}\left(\mathbf{z},\mathbf{1}\right)\right\rangle_{\mathbf{X}^v} \\ +\left\langle\mathcal{L}_{BCE}\left(\mathbf{z},\mathbf{0}\right)\right\rangle_{\bigcup\limits_{k\neq v}\mathbf{X}^k}\Big],\end{aligned} \tag{11}$$

where $\mathcal{L}_{BCE}$ denotes the binary cross-entropy loss [1] and $\langle\cdot\rangle$ denotes averaging over the set in subscript. The collective minimum optimization objective of all view discriminators can be formulated as:

$$\min_{\mathbf{W}_D^{1:V}}\mathcal{L}_{D_{1:V}}\left(\mathbf{X}_L,\mathbf{z},G_F\right) = \min_{\mathbf{W}_D^{1:V}}\left[\frac{1}{V}\sum_{\mathrm{v}=1}^{V}\mathcal{L}_{D_v}\left(\mathbf{X}_L,\mathbf{z},G_F\right)\right]. \tag{12}$$

The GRL utilized here [25, 35] is to reduce the distribution discrepancy of multi-sub-domains by maximizing the sub-domain discrimination loss (i.e., Eq. 12). As stated in Sec. II-B, it can automatically transform a maximization problem into minimizing a negative loss during back-propagation for the consistency of network optimization. Thus, the maximum objective function to optimize G_F can be formulated as follows:

$$\begin{aligned}\max_{\mathbf{W}_{G_F}}\mathcal{L}_{D_{1:V}}\left(\mathbf{X}_L,\mathbf{z},D_{1:V}\right) &\rightarrow \min_{\mathbf{W}_{G_F}}\mathcal{L}_{D_{1:V}}\left(\mathbf{X}_L,\mathbf{z},D_{1:V}\right) \\ &= \min_{\mathbf{W}_{G_F}}\left[-\mathcal{L}_{D_{1:V}}\left(\mathbf{X}_L,\mathbf{z},D_{1:V}\right)\right] \\ &= \min_{\mathbf{W}_{G_F}}\left[-\tfrac{1}{V}\textstyle\sum_{v=1}^{V}\mathcal{L}_{D_v}\left(\mathbf{X}_L,\mathbf{z},D_v\right)\right].\end{aligned} \tag{13}$$

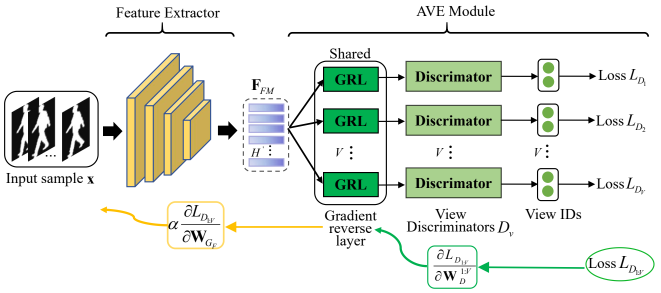


Fig. 4. The detailed architecture of the AVE module. The black arrows indicate the direction of forward propagation, while the yellow and green arrows indicate the direction of back propagation.

The forward and backward propagation of the AVE module is depicted in Fig. 4. During forward propagation, GRL is simply a common layer without any additional operation. During back-propagation, GRL inverts the gradients of the optimization objective Eq. (12) with respect to the parameters in the feature extractor and then passes backward with a negative weight α. Through GRL, the sub-domain adversarial alignment can be achieved in an end-to-end manner without fixing the generator and discriminator separately for iterative training like GANs. This greatly simplifies the overall implementation of the network and facilitates the mining of spatial-temporal features in gait sequences. Finally, robust feature representations that hard be discriminated by all view discriminators can be extracted and then the gaps between sub-domains can be effectively mitigated.

D. Metric Learning Module

As a metric learning problem, features with high discriminability are critical for cross-view gait recognition. Triplet-based loss functions are directly designed at learning discriminative features, which are more direct and suitable. Additionally, triplet-based loss functions are typically able to learn subtle features more effectively by setting a margin during training, which is particularly suitable for silhouette sequence based cross-view gait recognition. Compared to other distance-based loss function methods, e.g., the DrLim method [43], its constraints are relatively loose and more suitable. DrLim uses a contrast loss that uniformly converges the distance between all samples of the same class to 0, while the distance between samples of different classes converges to a fixed threshold. This is an extremely strict constraint for cross-view gait recognition, where there are large appearance differences between samples of the same class and the use of DrLim can confound original metric learning. Therefore, in this paper, a combined loss consisting of the separate batch-all triplet loss [44] and cross-entropy loss is adopted in the ML module

[1] $\mathcal{L}_{BCE}\left(y_i',y_i\right) = -\sum_{i=1}^{N}[y_i\log(y_i') + (1-y_i)\log(1-y_i')]$, where N denotes the number of samples, y_i and y_i' denote the true label and predict label of the sample, respectively.

to ensure the learned gait features are dispersed and highly discriminative.

Following the settings outlined in [44], P subjects and U silhouette sequences per subject are sampled to compose a mini-batch with the size of $P \times U$. For each sequence (anchor) in the mini-batch, the corresponding positive examples (pos.) and negative examples (neg.) are selected to construct sample triplets. Specifically, the anchor and positive examples have the same identity label but are different from the negative examples, and $PU\left(U-1\right)\left(PU-U\right)$ sample triplets are constituted in each mini-batch. To fully exploit fine-grained features, the batch-all triplet constraint is separately imposed on horizontal sliced features in the ML module. The complete triplet loss is defined as:

$$\mathcal{L}_{tri\_all} = \frac{1}{N_{tri}} \sum_{h=1}^{H'} \overbrace{\sum_{p=1}^{P} \sum_{u=1}^{U}}^{\text{an}chor} \overbrace{\sum_{\substack{a=1 \\ a \neq u}}^{U}}^{\text{pos.}} \overbrace{\sum_{\substack{b=1 \\ b \neq p}}^{P} \sum_{c=1}^{U}}^{neg.} \max\left\{dist + m, 0\right\}, \tag{14}$$

where N_{tri} is the number of triplets resulting in the non-zero loss terms; H' is the scale to slice the features horizontally; and m is the margin. In a sample triplet, each example has H' part-based features, and we calculate the triplet loss for each corresponding feature triplet, i.e. H' triplet losses are calculated. The $dist$ in Eq. (14) can be formulated as:

$$dist = d_{+}\left(\mathbf{f}_{FM}^{h,p,u}, \mathbf{f}_{FM}^{h,p,a}\right) - d_{-}\left(\mathbf{f}_{FM}^{h,p,u}, \mathbf{f}_{FM}^{h,b,c}\right), \tag{15}$$

where $\mathbf{f}_{FM}^{h,p,u}$ denotes the h-th horizontal feature vector in the u-th gait sequence of the p-th subject ($\mathbf{f}_{FM}^{h,p,a}$ and $\mathbf{f}_{FM}^{h,b,c}$ are similar to $\mathbf{f}_{FM}^{h,p,u}$). d_{+} and d_{-} are the euclidean distances between positive samples and negative samples, respectively.

Similarly, the cross-entropy loss is also conducted on all horizontal sliced features. As shown in Fig. 1, a classifier that predicts identity labels is trained for each horizontal sliced feature $\mathbf{f}_{FM}^{h}$, and there are H' classifiers in the ML module. The total cross-entropy loss is as follows:

$$\mathcal{L}_{CE\_all} = \frac{1}{NH'} \sum_{h=1}^{H'} \sum_{n=1}^{N} y\left(\mathbf{f}_{FM}^{h}\right) \log\left(q\left(\mathbf{f}_{FM}^{h}\right)\right), \tag{16}$$

where N is the number of samples in a mini-batch that equals to PU, $y\left(\mathbf{f}_{FM}^{h}\right)$ and $q\left(\mathbf{f}_{FM}^{h}\right)$ denote the ground truth and predict identity of $\mathbf{f}_{FM}^{h}$, respectively.

E. Joint Loss Functions

Finally, the overall objective consisting of the separate batch-all triplet loss Eq. (14) and cross-entropy loss Eq. (16) in ML module, as well as the loss introduced in Sec. III-C is conducted to optimize the proposed GaitDAN. And the function Eq. (2) of the proposed GaitDAN can be rewritten as follows:

$$\begin{gathered}\min_{\mathbf{W}_{G_F}} \mathcal{L}(\mathbf{X}_L, \mathbf{z}, D_{1:V}) = \min_{\mathbf{W}_{G_F}} (\mathcal{L}_{tri\_all} + \mathcal{L}_{CE\_all} - \beta \mathcal{L}_{D_{1:V}} (\mathbf{X}_L, \mathbf{z}, D_{1:V})) \\ \min_{\mathbf{W}_{D_{1:V}}} \mathcal{L}(\mathbf{X}_L, \mathbf{z}, G_F) = \min_{\mathbf{W}_{D_{1:V}}} \beta \mathcal{L}_{D_{1:V}} (\mathbf{X}_L, \mathbf{z}, G_F), \end{gathered} \tag{17}$$

where β denotes the weighted factor.

IV. Experiments

In this section, the datasets and implementation details are first described. Then, the performance of the proposed GaitDAN will be compared with other state-of-the-art methods on three gait databases. Finally, a detailed ablation study will be strictly performed to verify the effectiveness of each component in the proposed GaitDAN.

A. Datasets

We evaluate the proposed GaitDAN on three commonly used gait recognition datasets, i.e., CASIA-B [45], OULP [46] and OUMVLP [47].

CASIA-B is the most widely used gait dataset. It contains 124 subjects with three different variations, including viewpoint, clothing and carrying conditions. For each subject, 10 video groups under three walking conditions are collected, i.e., 6 NM (normal) (indexed as NM#01-06), 2 BG (with a bag) (indexed as BG#01-02), and 2 CL (with a cloth) (indexed as CL#01-02). In each video group, 11 videos taken under 11 different views (0°-180°with interval 18°) are included. Therefore, this dataset contains 124×(6+2+2)×11=13640 sequences. For fair comparison, the experiments in this paper are strictly following the protocol in [1, 15]. The first 74 subjects are used for training and the remaining 50 subjects are reserved for testing. During the testing phase, the first 4 sequences under NM condition (NM#01-04) are grouped into the gallery, and the rest sequences NM#05-06, BG#01-02, and CL#01-02 are used as the probe, respectively.

OULP is a gait dataset with a larger population. It consists of 4007 subjects with 2 video groups (indexed as #01-02) per subject. In each video group, 4 videos taken under view angles (55°, 65°, 75°, 85°) are available. Taking the same experimental settings as [13], samples of 3836 subjects are used for training, and five-fold cross-validation is adopted. During the test phase, the sequences with index #01 are used as the gallery, and the remaining sequences with index #02 are used as the probe.

OUMVLP is currently the largest public gait dataset which contains 10307 subjects (5153 subjects for training and 5154 subjects for testing). Similarly, there are 2 video groups (indexed as #01-02) per subject with 14 videos taken under 14 different view angles (0°, 15°, ... , 90°; 180°, 195°, ..., 270°). Consistent with the protocol in [1, 15], the sequences with index #01 of each subject are kept in the gallery and the rest sequences with index #00 are taken as the probe during the testing phase.

B. Implementation Details

Common configuration: Gait silhouette sequences are firstly pre-processed by the approach mentioned in [1] and each frame is resized to the size of 64×44. Adam optimizer [48] is utilized with the momentum of 0.9 and the initial learning rate of 10^{-4} . The margin in Eq. (14) is set to 0.2, and the frame number of each gait sequence for training is set to 30.

Network structures: The extractor in CASIA-B and OULP is shown in Fig. 2. The basic branch is listed in Tab. I. For the

HFA branch, the output channel of the CSAA block in each stage is the same as that of the next stage in the basic branch. In OUMVLP, three additional 3D convolution layers with the kernel size of (3, 3, 3) are added into the basic branch at “Stage 0”, “Stage2” and “Stage 3” to adapt the enlarged data scale. That is, there are eight 3D convolution layers, and the output channels are 64, 96, 96, 128, 192, 192, 256 and 256, respectively. Accordingly, the output channels of the CSAA blocks in HFA branch are also modified in OUMVLP case.

Parameter settings: The parameters α in GRL and β in Eq.(17) are set to -0.3 and 0.01 respectively in CASIA-B and OUMVLP, while set to -0.3 and 0.03 in OULP. The mini-batch (P, U) is set to (8, 8) for CASIA-B, (32, 4) for OULP and (32, 8) for OUMVLP. Moreover, the iteration is set to 100K, 60K, 210K for CASIA-B, OULP and OUMVLP, respectively, and the learning rate is decreased to 10^{-5} after 80K on CASIA-B while decreased to 10^{-5}, 5×10^{-6} after 150K and 200K on OUMVLP, respectively.

Testing details: During the test phase, all frames of a gait sequence are fed into the proposed GaitDAN to generate the feature representation. Then the distance between gallery and probe is defined as the average of euclidean distance of all corresponding horizontal sliced features. Finally, we calculate the recognition accuracy.

C. Performance Comparison on CASIA-B

To evaluate the performance of GaitDAN under both cross-view and cross-walking-condition cases, the comparison experiment is conducted on CASIA-B. We compare the performance of GaitDAN with several state-of-the-art methods, including CNN-LB [49], GaitNet [18], Group-supervised DRL [16], GaitSet [1], GaitPart [15], MvGAN [5], GaitSlice [50], MT3D [13], ESNet [14] and GaitGL [12]. Tab. II lists the average rank-1 accuracy for each probe view on all gallery views excluding the identical-view cases. It can be observed that GaitDAN achieves the highest accuracy with the mean recognition rates of 97.8%, 95.2% and 86.0% under NM, BG, and CL conditions respectively, which demonstrates the superiority of GaitDAN. More specifically, there are some interesting findings can be also analyzed from Tab. II:

- Coherent mining of spatial-temporal information in gait sequences by the end-to-end adversarial training contributes to superior performance. This is clearly revealed in Tab. II that compared with GaitNet and Group-supervised DRL, GaitDAN achieves an average accuracy improvement of at least 5.5%, 6.3% and 14.5% under three walking conditions, respectively. Such an improvement fully demonstrates the effectiveness of adversarial domain adaptation in view-change elimination. In addition, the design of AVE module embedded with GRLs enables the entire GaitDAN to be trained in an end-to-end adversarial manner. This approach effectively alleviates the limitations of decoupling learning, which results in the disruption of spatial-temporal information coherence in gait sequences due to decomposition and synthesis on single frame images.
- Integrated spatial-temporal feature extraction can also improve the gait recognition performance. Compared with MT3D and GaitGL, both of which are based on brute force learning and use 3D convolutions, the performance of GaitDAN under NM, BG, and CL conditions is 1.1%, 2.2%, and 4.5% higher than that of MT3D, and 0.4%, 0.7%, and 2.4% higher than that of GaitGL. This is a major improvement over the already high performance, and demonstrates the superiority of GaitDAN again. It benefits from the design of the HFA strategy and AVE module as described above. The HFA strategy allows for efficient aggregation of general local detail information and semantic representation from different network stages, thereby providing more complementary and comprehensive features, which effectively improves the discriminability of gait features.

D. Performance Comparison on OULP

To further evaluate the performance of GaitDAN, we continue to perform the evaluation on OULP, and several state-of-the-art methods are chosen for comparison, including CNNS [49], MGAN [51], and MT3D [13]. The detailed comparison results are listed in Tab. III. As can be seen in Tab. III, the proposed GaitDAN achieves the highest recognition rate in all cross-view cases with obvious performance advantages. In addition, it can be found that as the view difference between the probe and the gallery becomes larger, there is a significant degradation in performance. For example, when the view of the probe is 55°and the view of the gallery changes from 65°to 85°, the performance of MGAN decreases by 21.6% (99.4% to 77.9%). Corresponding to that, the performance degradation of GaitDAN is only 1.0% (99.6% to 98.6%) , indicating that the proposed GaitDAN can effectively eliminate the view-change interference and is more robust to view changes.

E. Performance Comparison on OUMVLP

To verify the generalization of GaitDAN on the large scale database, the evaluation of GaitDAN is completed on OUMVLP. Tab. IV lists the detailed experimental results of GaitDAN and other several state-of-the-art methods, including GEINet [52], GaitSet [1] GaitPart [15], GaitSlice [50], GLN [31], ESNet [14], GaitGL [12], and GQAN [53]. From the results in Tab. IV, we can observed that GaitDAN has the highest recognition rate except for the probe view of 195°, 210°and 225°, and achieves the competitive averaged rank-1 accuracy of 90.2%, which demonstrates the effectiveness of GaitDAN under the large-scale data scenario.

F. Ablation Study

To verify the effectiveness of each component in proposed GaitDAN, the detailed ablation studies are performed in the following parts. All experiments in this section are conducted on CASIA-B due to the richness of its data types.

*1) **Incremental evaluation of each component:*** To valid the effectiveness of the HFA strategy, AVE module and ML module, incremental evaluations are conducted on CASIA-B. The experimental results are presented in Tab. V, where the backbone consists of the basic branch and feature

TABLE II
CROSS-VIEW AVERAGE RANK-1 ACCURACIES (%) ON CASIA-B UNDER ALL DIFFERENT PROBE VIEWS EXCLUDING IDENTICAL-VIEW CASES.

Gallery NM#01-04		0°-180°											Mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#05-06	CNN-LB [49]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	84.0	94.7	91.3	78.5	89.9
	GaitNet [18]	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
	Group-supervised DRL [16]	87.9	95.2	97.0	95.1	90.5	88.0	90.9	94.8	96.5	93.7	82.7	92.0
	GaitSet [1]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [15]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitSlice [50]	95.5	**99.2**	**99.6**	**99.0**	94.4	92.5	95.0	98.1	**99.7**	98.3	92.9	96.7
	MT3D [13]	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
	ESNet [14]	95.6	98.6	99.1	97.9	96.7	94.4	96.9	98.7	99.3	98.6	95.1	97.4
	GaitGL [12]	96.0	98.3	99.0	97.9	**96.9**	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	PROPOSED	**96.6**	98.1	99.2	98.1	96.7	**95.5**	**98.0**	**99.0**	99.3	**99.1**	**96.4**	**97.8**
BG#01-02	CNN-LB [49]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitNet [18]	88.8	88.7	88.7	94.3	85.4	92.7	91.1	92.6	84.9	84.4	86.7	88.9
	Group-supervised DRL [16]	77.9	88.8	91.8	90.1	84.4	79.7	83.5	89.3	92.2	89.5	77.5	85.9
	GaitSet [1]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [15]	89.1	94.8	96.7	95.1	88.3	**94.9**	89.0	93.5	96.1	93.8	85.8	91.5
	GaitSlice [50]	90.2	96.4	96.1	94.9	89.3	85.0	90.9	94.5	96.3	95.0	88.1	92.4
	MT3D [13]	91.0	95.4	**97.5**	94.2	92.3	86.9	91.2	95.6	97.3	96.4	86.6	93.0
	ESNet [14]	92.7	95.9	96.3	94.9	93.2	87.7	90.9	96.2	97.3	**96.9**	91.7	94.0
	GaitGL [12]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	**98.2**	**96.9**	**94.5**	94.5
	PROPOSED	**93.1**	**97.2**	97.1	**96.1**	**95.0**	91.0	**93.4**	**97.0**	**98.2**	**96.9**	92.3	**95.2**
CL#01-02	CNN-LB [49]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitNet [18]	50.1	60.7	72.4	72.1	74.6	78.4	70.3	68.2	53.5	44.1	40.8	62.3
	Group-supervised DRL [16]	60.9	75.6	81.0	78.1	72.6	67.8	73.0	77.1	76.8	70.0	53.3	71.5
	GaitSet [1]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [15]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitSlice [50]	75.6	87.0	88.9	86.5	80.5	77.5	79.1	84.0	84.8	83.6	70.1	81.6
	MT3D [13]	76.0	87.6	89.8	85.0	81.2	75.7	81.0	84.5	85.4	82.2	68.1	81.5
	ESNet [14]	75.6	89.2	92.4	90.3	84.3	**80.2**	83.0	86.3	89.0	83.9	69.8	84.0
	GaitGL [12]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	PROPOSED	**78.2**	**90.9**	**93.0**	**91.4**	**86.4**	79.9	**84.7**	**89.1**	**91.2**	**87.6**	**73.2**	**86.0**

TABLE III
CROSS-VIEW AVERAGE RANK-1 ACCURACIES (%) ON OULP FOR FOUR VIEWS EXCLUDING IDENTICAL-VIEW CASES.

Probe	Method	Gallery				
		55°	65°	75°	85°	Mean
55°	CNNS [49]		98.3	96.0	80.5	91.6
	MGAN [51]		99.4	96.1	77.9	-
	MT3D [13]		**99.6**	98.1	84.7	94.2
	PROPOSED		**99.6**	**99.2**	**98.6**	**99.1**
65°	CNNS [49]	96.3		97.3	83.3	92.3
	MGAN [51]	97.7		98.5	84.4	-
	MT3D [13]	97.8		98.5	84.9	93.7
	PROPOSED	**99.6**		**99.5**	**99.0**	**99.4**
75°	CNNS [49]	94.2	97.8		85.1	92.4
	MGAN [51]	94.8	98.9		86.4	-
	MT3D [13]	96.8	99.0		86.1	94.0
	PROPOSED	**99.2**	**99.5**		**99.1**	**99.3**
85°	CNNS [49]	90.0	96.0	98.4		94.8
	MGAN [51]	86.9	97.4	99.5		-
	MT3D [13]	96.4	98.4	99.5		98.1
	PROPOSED	**99.1**	**99.5**	**99.7**		**99.4**

mapping head, and is optimized under the separate cross-entropy loss or triplet loss in the ML module, respectively. HFA∼ denotes a degenerate version of HFA strategy without CSAA blocks. In HFA∼, the output of each stage in the basic branch is simply concatenated along the channel, and the final output is directly obtained after channel adjustment with bottle neck layers.

As listed in Tab. V, the completed use of the ML module can effectively improve the performance of the model. Specifically, the average accuracy of the backbone+ML model under the three walking conditions is 0.5% higher than that of the backbone+L_{tri_all} model and 2.7% higher than that of the backbone+L_{CE} model, which also illustrates the advantages of the joint constraint of the separate triple loss and the cross-entropy loss in ML module. In addition, with the help of HFA, the performance is boosted considerably, especially under the most challenging condition (CL) and the average condition. This is because that HFA can adequately integrate the detailed visual information extracted from shallow layers and the subtle spatial-temporal clues from high layers, thereby obtaining more comprehensive and discriminative gait representations. The performance gain also suggests that mining specific local cues in shallow stages of the network is more imporatant for challenging and complex conditions (e.g. CL). Compared to simply cascading the general spatial-temporal features using bottle neck layers (Backbone+ML+HFA∼), the complete HFA provides a significant performance improvement as it considers the semantic and distribution differences between different stage features, which fully validates the rationality of the HFA's design. Moreover, the integration of the AVE module can further improve the recognition accuracy. As indicated in Tab. V, AVE brings an additional average performance improvement by 0.6% when used in conjunction with HFA. This is attributed to AVE, by which the influence of view-change can be effectively eliminated and thus improve the robustness of feature representations.

2) *Analysis of the internal structure in CSAA block:*

As described in Sec. III-B2, the CSAA block is composed of two learnable parameters σ_1, σ_2, a cross-stage attention deriva-

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3384308

TABLE IV
CROSS-VIEW AVERAGE RANK-1 ACCURACIES (%) ON OUMVLP EXCLUDING IDENTICAL-VIEW CASES.

Gallery	0°-90°, 180°-270°														Mean
Probe	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet [52]	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41.0	42.5
GaitSet [1]	79.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	86.5	89.0	89.2	87.2	87.6	86.2	87.1
GaitPart [15]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.0	88.2	88.7
GLN [31]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GaitSlice[50]	84.1	89.0	91.2	**91.6**	90.6	89.9	89.8	85.7	89.3	**90.6**	**90.7**	89.8	89.6	88.5	89.3
ESNet [14]	84.8	89.6	91.0	91.3	90.7	90.4	89.9	88.5	87.5	90.1	90.2	89.4	89.3	88.5	89.4
GaitGL [12]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
GQAN [53]	85.0	90.3	**91.3**	91.4	90.6	90.6	90.1	87.1	**89.4**	90.5	90.6	**90.0**	**89.8**	89.1	89.7
PROPOSED	**86.4**	**90.9**	**91.3**	**91.6**	**91.4**	**91.1**	**90.8**	**89.5**	89.3	90.3	90.5	**90.0**	**89.8**	**89.5**	**90.2**

TABLE V
AVERAGED RANK-1 ACCURACIES (%) OF GAITDAN FOR ABLATION STUDIES ON CASIA-B.

Methods	Accuracy			
	NM	BG	CL	Mean
Backbone+L_{CE_all}	96.0	92.3	77.2	88.5
Backbone+L_{tri_all}	97.0	93.5	81.6	90.7
Backbone+ML	96.9	93.6	83.3	91.2
Backbone+ML+HFA∼	96.9	94.1	83.7	91.6
Backbone+ML+HFA	97.5	94.5	85.3	92.4
Backbone+ML+HFA+AVE (PROPOSED)	**97.8**	**95.2**	**86.0**	**93.0**

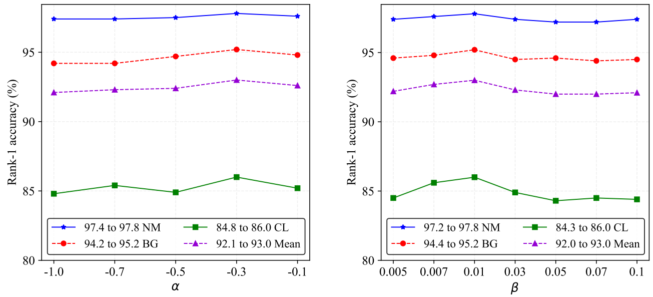
TABLE VI
AVERAGE RANK-1 ACCURACIES (%) OF THREE DEGRADATION MODELS ON CASIA-B. ('W/O' DENOTES WITHOUT.)

Methods	Accuracy			
	NM	BG	CL	Mean
PROPOSED w/o σ_1 and σ_2	97.6	94.8	85.3	92.6
PROPOSED w/o $\mathbf{W}_m$	97.5	94.6	84.8	92.3
PROPOSED w/o $\mathbf{W}_a$	97.5	94.3	84.2	92.0
PROPOSED	**97.8**	**95.2**	**86.0**	**93.0**

tion operation $\mathbf{W}_m$ and a cross-stage attention aggregation operation $\mathbf{W}_a$. To explore their individual roles, we conduct comparison experiments of GaitDAN and its three degradation models on CASIA-B, where each of these degradation models is implemented by deleting one of the above components. The experimental results are reported in Tab. VI, from which we can see that the cross-stage attention aggregation operation shows the biggest contribution among the three components. However, the absence of either the learnable parameters or the cross-stage attention derivation operation can lead to certain performance degradation. It is also well illustrated that the introduction of learnable parameters, and the utilization of cross-stage attention derivation with minimal parameters can indeed contribute to cross-stage feature aggregation. With them, the semantic and distributional differences between different stage features can be effectively mitigated. Furthermore, these three components are shown to be complementary, with the complete CSAA block, the highest results under three walking conditions can be achieved.

3) Analysis of hyper-parameters:

To evaluate how α and β affect the model learning, we conduct parameter sensitivity experiments on CASIA-B. From the results in Fig. 5, we can observe that GaitDAN performs better when α increases from -1.0 and achieves the best performance when $\alpha = -0.3$. Moreover, the performance decreases as α continues to increase. Thus, we set $\alpha = -0.3$ on CASIA-B. For β, the performance tends to increase and then decrease as β continues increasing within the range of [0.005, 0.1]. More concretely, the proposed method reaches the optimum when β attains to 0.01. Therefore, we set $\beta = 0.01$ on CASIA-B. Similarly, for OULP, we set $\alpha = -0.3$ and $\beta = 0.03$, while for the experiments on OUMVLP, the hyper-parameters are all set the same as on CASIA-B.



(a) (b)

Fig. 5. Evaluation of the hyper-parameters α and β on CASIA-B.

TABLE VII
AVERAGE RANK-1 ACCURACIES (%) ON CASIA-B UNDER THE PROBE VIEW OF 54°AND 126°. ('W/O' DENOTES 'WITHOUT', 'M_D/I' DENOTES THE DECREASE OR INCREASE IN MEAN.)

Probe	Methods	Experimental settings	Accuracy				
			NM	BG	CL	Mean	M_D/I
54°	PROPOSED	Complete training set	98.1	96.1	91.4	95.2	↓**1.8**
		w/o samples at 54°	98.1	95.5	86.6	93.4	
	GaitGL	Complete training set	97.9	95.9	87.1	93.6	↓2.2
		w/o samples at 54°	96.4	93.5	84.3	91.4	
	GaitGL+AVE	w/o samples at 54°	97.7	94.6	86.1	92.8	↑1.4
	GaitSet	Complete training set	97.8	93.4	74.6	88.6	↓4.5
		w/o samples at 54°	93.7	87.2	71.5	84.1	
	GaitSet+AVE	w/o samples at 54°	96.3	90.6	74.2	87.0	↑2.9
126°	PROPOSED	Complete training set	99.0	97.0	89.1	95.0	↓**1.0**
		w/o samples at 126°	98.6	95.4	88.0	94.0	
	GaitGL	Complete training set	98.9	96.5	87.0	94.1	↓1.5
		w/o samples at 126°	98.5	95.1	84.1	92.6	
	GaitGL+AVE	w/o samples at 126°	98.7	95.5	85.6	93.3	↑0.7
	GaitSet	Complete training set	98.3	91.7	74.1	88.0	↓2.5
		w/o samples at 126°	96.4	88.5	71.5	85.5	
	GaitSet+AVE	w/o samples at 126°	97.1	90.3	73.5	87.0	↑1.5

4) Analysis of viewpoint generalization:

To further evaluate the robustness and generalization of the proposed GaitDAN, an experiment based on CASIA-B is

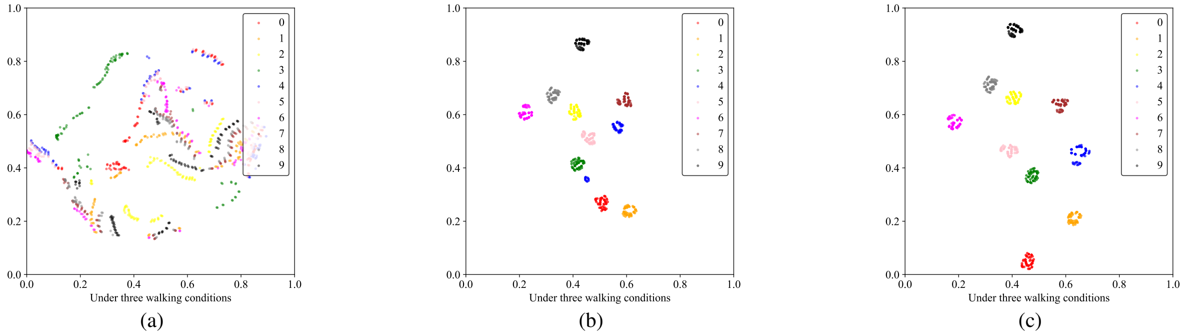


Fig. 6. T-SNE visualization of (a) the original sample features, (b) the features without view-change elimination, and (c) the features after view-change elimination under three walking conditions on CASIA-B. We visualize 10 identities. Each point represents a sample and each color defines a identity class.

designed where subjects (i.e. people) are different between the training and testing data. The view angle of the probe gait in the testing data can be included in or not in the model training. In the case that the probe gait view is not in the training data, it can best demonstrate the performance of the proposed method when dealing with an unknown gait view angle, where the existing transformation-based methods VTM [28] and CBDP [10] cannot deal with such situations. This experiment is carried out twice by using different probe gait view samples, 54 °and 126 °. For the case of unknown viewpoints, such as 54 °, the gait samples under all other known 10 viewpoints in CASIA-B dataset are used for training, so the proposed method can learn a model to transform the gait samples from different viewpoints to view-invariant space during the training phase. In this way, during the test phase, it is only necessary to input the gait samples to be recognized (e.g., samples under 54 °) into the model, and then the gait features can be directly obtained. This allows for the comparison of the distance between probe and gallery gait samples, enabling gait recognition to be carried out in various cross-view situations. For the case of an unknown viewpoint of 126 °, the test of gait samples is similar to the above. The detailed results are reported in Tab .VII.

From Tab. VII, we can find that when 54°gait samples are not included in the training data and the probe view is 54°, the average accuracy under three walking conditions decreases by about 1.8%. More specifically, the performance degradation is more slight under the NM walking condition, as NM is somewhat easier compared to BG and CL. Even so, it also elucidates the comparative advantages of GaitDAN over previous state-of-the-art approaches, such as GaitSet [1] and GaitGL [12]. Compared to them, GaitDAN has a more slight decrease in average accuracy of the three walking conditions, and still achieves the highest recognition rate under all three walking conditions when performing cross-view recognition with the unknown view angle of 54°. It demonstrates the strong generalizability of GaitDAN to different viewpoints, and further illustrates the rationality of the design of the HFA strategy and the AVE module in GaitDAN. The introduction of the HFA strategy ensures comprehensive spatial-temporal feature extraction, while the end-to-end adversarial training method of the AVE module achieves effective elimination of viewpoint variations in a simple way, facilitating the full exploitation of spatial-temporal features by the HFA strategy. In this way, GaitDAN can be tested under unknown viewpoints and obtain considerable recognition performance. As listed in Tab. VII, similar results can also be obtained when the training data does not contain 126°gait samples and the view angle of probe samples is 126°, which strongly indicates the superiority of GaitDAN.

TABLE VIII
AVERAGE RANK-1 ACCURACIES (%) ON CASIA-B OF GAITSET AND GAITGL WITH/WITHOT AVE MODULE.

Methods	Accuracy			
	NM	BG	CL	Mean
GaitSet	95.0	87.2	70.4	84.2
GaitSet+AVE	94.9	89.4	73.2	85.8
GaitGL	97.4	84.5	83.6	91.8
GaitGL+AVE	97.3	94.7	85.0	92.3

Furthermore, we test the performance of other feature extraction methods with our AVE module, i.e., GaitGL+AVE and GaitSet+AVE. As listed in Table VII, the addition of the AVE module can effectively reduce the performance degradation of GaitSet and GaitGL at 54°and 126°views, which further demonstrates the effectiveness of the AVE module.

5) *Feature Visualization:*

To illustrate the effectiveness of the AVE module to help extract highly discriminative gait features for the overall model, the t-SNE [54] technique is utilized to visualize the distributions of different features, including original gait sample features, and the features before and after view-change elimination. The visualization is shown in Fig. 6. It can be observed that the feature distribution after view-change elimination is more dispersed than that before view-change elimination. In other words, the inter-class distance between features is larger. And for some difficult samples, the intra-class distance is smaller. Therefore, these view-change-eliminated features are more separable and discriminative, making them more effective for gait recognition and increasing the likelihood of correctly identifying difficult samples. The visualization results further validate the effectiveness of the AVE module.

6) *Analysis of the AVE module:*

In addition, to quantitatively evaluate the effectiveness of

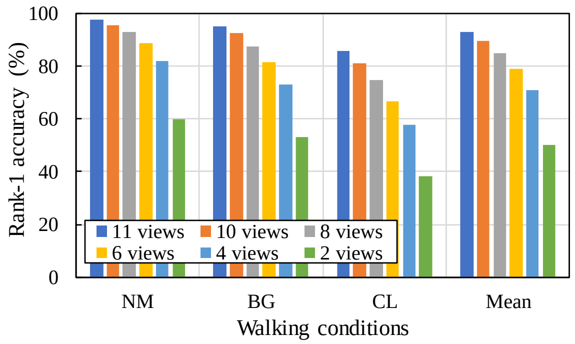


Fig. 7. Average rank-1 accuracies (%) of different training view numbers on CASIA-B.

TABLE IX
AVERAGE RANK-1 ACCURACIES (%) COMPARISON WITH SKELETON-BASED METHODS ON CASIA-B.

Methods	Accuracy			
	NM	BG	CL	Mean
GaitGraph2	80.3	71.4	63.8	71.8
GPGait	93.6	80.2	69.3	81.0
PROPOSED	97.8	95.2	86.0	93.0

AVE module, experiments of combining different feature extraction methods (GaitSet and GaitGL) with AVE module are carried out to show the performance difference with and without AVE module. The results are listed in Table VIII. It can be found that with the AVE module, although there is a slight decrease under the NM condition, there is a significant increase under both the BG and CL conditions. Compared to the original GaitSet and GaitGL, the addition of AVE results in an improvement of 1.6% and 0.5% in the mean values of the NM, BG and CL cases, respectively. This once again demonstrates the effectiveness of the AVE module.

*7) **Analysis of the impact of the number of viewpoints**:* Here we investigate the influence of the number of training viewpoints on the final performance. For the 11 viewpoints in the CASIA-B dataset (0°-180°with interval 18°), the proposed GaitDAN was trained by sequentially removing the gait samples under 1, 3, 5, 7 and 9 training viewpoints. In other words, we evaluate the cross-view gait recognition performance when the number of training viewpoints is 2, 4, 6, 8, and 10, respectively. The detailed experimental results are shown in Fig. 7. From Fig. 7, it can be obversed that the number of training viewpoints has a significant impact on the final performance, especially when the number of training viewpoints is small, the lower the test results. This is also cinsistent with the expectation because 1) the number of training samples decreases dramatically, and 2) the number of samples with unseen viewpoints becomes larger at the testing phase.

G. Comparison with skeleton-based methods

To further illustrate the validity of our proposed method, we compared our results with current SOTA human skeleton-based methods on CASIA-B, including GaitGraph2 [55] and GPGait [56]. The detailed comparison results are listed in Tab. IX. It can be found that the proposed method exhibits a significant advantage over the skeleton-based methods under the NM, BG and CL walking conditions. Notably, under the BG and CL conditions, where the walking conditions are more complex, our method is significantly improved, which once again demonstrates the superiority of the proposed method and also reflects the advantages of appearance-based cross-view gait recognition.

V. Conclusion and Future Work

In this paper, we address the task of cross-view gait recognition by casting the view-change mitigation as a domain adaptation problem of narrowing the distribution differences among view-level sub-domains. On this basis, GaitDAN is proposed to generate discriminate and sub-domain-invariant gait representations via adversarial domain adaptation. GaitDAN contains two key components, i.e., a novel feature extractor with HFA and the AVE module. The feature extractor equipped with HFA is presented to aggregate the spacial-temporal features from various stages of the network for discriminative feature extraction. The AVE module aims to match the distributions of sub-domains by making them indistinguishable for view-specific discriminators with GRL. It enables an end-to-end training of the entire framework. Therefore, the view-change information can be effectively utilized and eliminated, and at the same time, spacial-temporal information in gait sequences can be fully exploited. Experiments conducted on the three public databases, CASIA-B, OULP and OUMVLP, also demonstrate the superiority of the proposed method as well as all its components. In the future work, we intend to further investigate the differential distributions of sub-domains (viewpoints). We plan to quantitatively evaluate the differences among sub-domains and perform dynamic adversarial sub-domain adaptation for more challenging cross-view gait recognition problems. In addition, we also consider using richer input modalities such as human posture [57] to further improve the model's performance.

References

[1] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "Gaitset: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, 2021.

[2] H. Qin, Z. Chen, Q. Guo, Q. J. Wu, and M. Lu, "RPNet: Gait recognition with relationships between each body-parts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2990–3000, 2021.

[3] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1001–1015, 2019.

[4] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, 2019.

[5] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait

recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 3041–3055, 2021.

[6] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 40, no. 4, pp. 997–1008, 2009.

[7] F. Jean, A. B. Albu, and R. Bergevin, "Towards view-invariant gait modeling: Computing view-normalized body part trajectories," *Pattern Recognit.*, vol. 42, no. 11, pp. 2936–2949, 2009.

[8] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li, "A new view-invariant feature for cross-view gait recognition," *IEEE Trans. Inf. Forensic Secur.*, vol. 8, no. 10, pp. 1642–1653, 2013.

[9] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Trans. Inf. Forensic Secur.*, vol. 8, no. 12, pp. 2034–2045, 2013.

[10] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 734–747, 2020.

[11] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognit.*, vol. 90, pp. 87–98, 2019.

[12] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14 648–14 656.

[13] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *ACM Multimedia*, 2020, pp. 3054–3062.

[14] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, and Q. Wu, "Enhanced spatial-temporal salience for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6967–6980, 2022.

[15] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 14 225–14 233.

[16] L. Yao, W. Kusakunniran, P. Zhang, Q. Wu, and J. Zhang, "Improving disentangled representation learning for gait recognition using group supervision," *IEEE Trans. Multimedia*, 2022.

[17] X. Zhai, X. Ben, C. Liu, and T. Xie, "Decomposing identity and view for cross-view gait recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*. IEEE, 2022, pp. 1–6.

[18] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, 2022.

[19] T. Chai, X. Mei, A. Li, and Y. Wang, "Semantically-guided disentangled representation for robust gait recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*. IEEE, 2021, pp. 1–6.

[20] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*. IEEE, 2019, pp. 778–786.

[21] H. Ren, J. Wang, J. Dai, Z. Zhu, and J. Liu, "Dynamic balanced domain-adversarial networks for cross-domain fault diagnosis of train bearings," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[22] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. ICML*. PMLR, 2018, pp. 5423–5432.

[23] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3339–3348.

[24] P. O. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8004–8013.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[26] P. Arora, M. Hanmandlu, and S. Srivastava, "Gait based authentication using gait information image features," *Pattern Recognition Letters*, vol. 68, pp. 336–342, 2015.

[27] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, 2021.

[28] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2006, pp. 151–163.

[29] Y. Wang, C. Song, Y. Huang, Z. Wang, and L. Wang, "Learning view invariant gait features with two-stream GAN," *Neurocomputing*, vol. 339, pp. 245–254, 2019.

[30] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20 249–20 258.

[31] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 382–398.

[32] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 124–137, 2020.

[33] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, and L. Cui, "GaitAMR: Cross-view gait recognition via aggregated multi-feature representation," *Information Sciences*, vol. 636, p. 118920, 2023.

[34] Q. Men, E. S. L. Ho, H. P. H. Shum, and H. Leunga, "Focalized contrastive view-invariant learning for self-supervised skeleton-based action recognition," *Neurocomputing*, vol. 537, pp. 198–209, 2023.

[35] Y. Ganin and V. Lempitsky, "Unsupervised domain adap-

tation by backpropagation," in *Proc. ICML*, 2015, pp. 1180–1189.

[36] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 309–324.

[37] B. Niu, Z. Pan, J. Wu, Y. Hu, and B. Lei, "Multi-representation dynamic adaptation network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[38] H. Zhang, H. Cao, X. Yang, C. Deng, and D. Tao, "Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 5287–5298, 2021.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.

[41] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2018.

[42] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, pp. 151–175, 2010.

[43] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[44] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[45] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*, vol. 4. IEEE, 2006, pp. 441–444.

[46] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensic Secur.*, vol. 7, no. 5, pp. 1511–1521, 2012.

[47] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vision Appl*, vol. 10, no. 1, pp. 1–14, 2018.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[49] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, 2016.

[50] H. Li, Y. Qiu, H. Zhao, J. Zhan, R. Chen, T. Wei, and Z. Huang, "GaitSlice: A gait recognition model based on spatio-temporal slice features," *Pattern Recognit.*, vol. 124, p. 108453, 2022.

[51] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensic Secur.*, vol. 14, no. 1, pp. 102–113, 2018.

[52] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, 2016, pp. 1–8.

[53] S. Hou, X. Liu, C. Cao, and Y. Huang, "Gait quality aware network: toward the interpretability of silhouette-based gait recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[54] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[55] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 1569–1577.

[56] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang, "GPGait: Generalized pose-based gait recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 19 595–19 604.

[57] H.-M. Hsu, Y. Wang, C.-Y. Yang, J.-N. Hwang, H. L. U. Thuc, and K.-J. Kim, "GAITTAKE: Gait recognition by temporal attention and keypoint-guided embedding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2022, pp. 2546–2550.

Tianhuan Huang received the B.E. degree in Electronic Information Engineering from School of Physical Science and Technology, Nanjing normal university, Nanjing, China, in 2018. She is currently a Ph.D candidate with the School of Information Science and Engineering, Shandong University, Qingdao, China. Her current research interests include gait recognition, computer vision and machine learning.

Xianye Ben received the Ph.D. degree in pattern recognition and intelligent system from the College of Automation, Harbin Engineering University, Harbin, China, in 2010. She is currently working as a Full Professor with the School of Information Science and Engineering, Shandong University, Qingdao, China. She has authored or coauthored more than 100 papers in major journals and conferences, such as IEEE T-PAMI, IEEE T-IP, IEEE T-CSVT, IEEE T-MM, PR, CVPR, etc. Her current research interests include pattern recognition and image processing. She received the Excellent Doctorial Dissertation awarded by Harbin Engineering University. She was also enrolled by the Distinguished Young Scholars Program of Shandong University and the Shi Qingyun Female Scientists of China Society of Image Graphics.

Chen Gong received his B.E. degree from East China University of Science and Technology (ECUST) in 2010, and dual doctoral degree from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS) in 2016 and 2017, under the supervision of Prof. Jie Yang and Prof. Dacheng Tao, respectively. Currently, he is a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly include machine learning, data mining, and learning-based vision problems. He has published more than 50 technical papers at prominent journals and conferences such as IEEE T-NNLS, IEEE T-IP, IEEE T-CYB, IEEE T-CSVT, IEEE T-MM, IEEE T-ITS, CVPR, AAAI, IJCAI, ICDM, etc. He received the Excellent Doctorial Dissertation awarded by Shanghai Jiao Tong University (SJTU) and Chinese Association for Artificial Intelligence (CAAI). He was also enrolled by the Summit of the Six Top Talents Program of Jiangsu Province, China, and the Lift Program for Young Talents of China Association for Science and Technology.

Wenzheng Xu received the B.E. degree from Qingdao University, Qingdao, China, in 2021. He is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. His current research interests include gait recognition and computer vision.

Qiang Wu (M’02) received the B.Eng. and M.Eng. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree in computing science from the University of Technology Sydney, Sydney, Australia, in 2004. He is currently an Associate Professor with the School of Computing and Communications, University of Technology Sydney. His major research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. He has published more than 70 refereed papers, including those published in prestigious journals and top international conferences. Dr. Wu has been a Guest Editor of several international journals, such as the Pattern Recognition Letters (PRL) and the International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). He has served as a Chair and/or a Program Committee Member for a number of international conferences.

Hongchao Zhou received the B.Sc. degree in physics and mathematics and M.Sc. degree in control science and engineering from Tsinghua University, Beijing, China, in 2006 and 2008, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from California Institute of Technology, Pasadena, CA, USA, in 2009 and 2012, respectively. From 2012 to 2015, he was a Post-Doctoral Researcher with the Signals, Information and Algorithms Laboratory, Massachusetts Institute of Technology. He is currently a Professor with the School of Information Science and Engineering, Shandong University. His current interests include information theory, data systems, learning systems and machine learning. He was a recipient of the 2013 Charles Wilts Prize for the best doctoral thesis in electrical engineering at California Institute of Technology.