Integral Pose Learning via Appearance Transfer for Gait Recognition

Panjian Huang, Saihui Hou, *Member, IEEE*, Chunshui Cao, Xu Liu, Xuecai Hu, and Yongzhen Huang, *Senior Member, IEEE*

***Abstract*— Gait recognition plays an important role in video surveillance and security by identifying humans based on their unique walking patterns. The existing gait recognition methods have achieved competitive accuracy with shape and motion patterns under limited-covariate conditions. However, when extreme appearance changes distort discriminative features, gait recognition yields unsatisfactory results under cross-covariate conditions. In this work, we first indicate that the integral pose in each silhouette maintains an appearance-unrelated discriminative identity. However, the monotonous appearance variables in a gait database cause gait models to have difficulty extracting integral poses. Therefore, we propose an Appearance-transferable Disentangling and Generative Network (GaitApp) to generate gait silhouettes with rich appearances and invariant poses. Specifically, GaitApp leverages multi-branch cooperation to disentangle pose features and appearance features, and transfers the appearance information from one subject to another. By simulating a person constantly changing appearances under limited-covariate conditions, downstream models enable to extract discriminative integral pose features. Extensive experiments demonstrate that our method allows representative gait models to stand at a new altitude, further promoting the exploration to cross-covariate gait recognition. All the code is available at https://github.com/Hpjhpjhs/GaitApp.git**

***Index Terms*— Integral pose, appearance transfer, gait recognition, disentangling representation learning.**

I. Introduction

GAIT recognition is one of the most valuable biometric recognition at a long distance and has great applicability in the fields of surveillance and forensics [1], [2]. However, gait recognition has become one of the most difficult fine-grained action classification problems due to complex covariates (*e.g.*, cross-view, cross-carrying, cross-clothes, and occlusion conditions) [3], [4], [5]. Despite the significant progress achieved in the cross-view gait recognition [6], [7], [8], [9], [10], [11], [12], the distortion of human appearance with clothes, carrying, and other covariates has become the current biggest bottleneck in gait recognition tasks.

It is common sense that gait recognition mainly depends on two types of discriminative patterns: human shapes and body part motions extracted from gait silhouettes [7], [13], [14]. For shape-based methods [6], [8], [9], when fragile human shape patterns are distorted by appearance covariates and discriminative features are hidden, model degradation occurs. For methods [7], [13], [15] emphasizing motion patterns, a large accuracy gap is also observed between conditions with and without clothes changes. Ideally, motion patterns are essential descriptors of human gaits. However, some keypoint-based methods [16], [17] that only use motion patterns still obtain unsatisfactory results. Therefore, we naturally ask: *How can shape patterns and motion patterns be effectively integrated despite their distinctive properties, and how can their respective defects be filtered out*? We attempt to decouple this question into two perspectives: 1) Keypoint-based methods oversimplify predefined human poses (*e.g.*, 17 keypoints with human biases); 2) Silhouettes contain rich pose information, but include a large amount of interference information.

Therefore, we claim that the integral pose in each gait silhouette contains compact motion patterns. Here, the integral pose consists of the angle of the head lift, the angle between two thighs, the length of the two thighs and the height of the hands off the ground. As Fig. 1(a) shows, under various clothes-changing scenarios, the deformation of the human shape changes drastically. However, importantly, the integral poses of different people exhibit obvious differences at the same phase of a gait cycle, while the integral pose of the same person remains consistent. For example, at the phase with the maximum gait stride, the integral poses (*e.g.*, the angle of the head lift, the angle between the thighs, and the height of the hands off the ground) are still distinct between different persons and are consistent for the same person. Theoretically, appearance-unrelated integral pose features can be obtained by simulating a person constantly changing appearance to constrain the representation learning process. However, since the diversity limitations of gait databases (*e.g.*, the widely used CASIA-B [18] changes only upper clothes) and gait collection

Manuscript received 26 June 2023; revised 15 January 2024 and 6 March 2024; accepted 20 March 2024. Date of publication 27 March 2024; date of current version 7 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62276025 and Grant 62206022, in part by Beijing Municipal Science and Technology Commission under Grant Z231100007423015, and in part by Shenzhen Technology Plan Program under Grant KQTD20170331093217368. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fernando Alonso-Fernandez. *(Panjian Huang and Saihui Hou contributed equally to this work.) (Corresponding authors: Yongzhen Huang; Xuecai Hu.)*

Panjian Huang and Xuecai Hu are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China (e-mail: 202231081025@mail.bnu.edu.cn; huxc1208@bnu.edu.cn).

Saihui Hou and Yongzhen Huang are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China, and also with Watrix Technology Limited Company Ltd., Beijing 100088, China (e-mail: housaihui@bnu.edu.cn; huangyongzhen@bnu.edu.cn).

Chunshui Cao and Xu Liu are with Watrix Technology Limited Company Ltd., Beijing 100088, China (e-mail: chunshui.cao@watrix.ai; xu.liu@watrix.ai).

Digital Object Identifier 10.1109/TIFS.2024.3382606

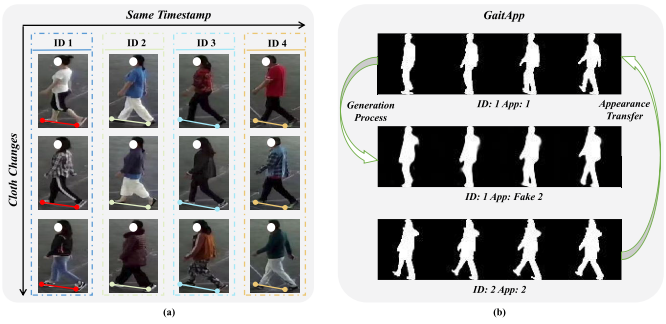


Fig. 1. (a) Integral poses. At the maximum gait strides of persons (viewed in coloured lines), although extreme appearance changes occur due to clothes changes, the same person maintains a consistent integral pose (*e.g.*, the angle of the head lift, the angle between thighs, and the height of hands from the ground) at the same phase of different gait cycles (better viewed in each column), but different persons have obviously different integral poses (better viewed in each row). (b) Appearance transfer. GaitApp transfers the appearance of ID 2 to ID 1, generating a new gait sequence for ID 1 with an invariant pose and a new appearance (denoted as *Fake 2*).

with rich covariates requires considerable labour and complex environmental conditions [3], the key challenge is how to extract integral pose features based on gait silhouettes under insufficient appearance changes.

Driven by the above analysis, we design an Appearance-transferable Disentangling and Generative Network (**GaitApp**) to transfer appearance information from one person to another while maintaining pose consistency. The generated gait silhouettes with rich appearances can be flexibly fed into downstream tasks. As Fig. 1(b) shows, given a pair of gait silhouette sequences (*e.g.*, ID 1 with App I and ID 2 with App II), GaitApp transfers the appearance information from ID 2 to ID 1, generating a new gait silhouette sequence containing the same pose as that of ID 1 and the same appearance as that of ID 2. Therefore, GaitApp enables each gait silhouette to maintain discriminative pose information with a number of appearance changes, which forces gait recognition models to extract appearance-invariant integral pose features. In summary, this work introduces three aspects of integral poses:

(1) Concept. Integral poses embody adaptive human regions, preventing visual perturbations, which are derived by contrasting the same person with different appearances but the same pose.

(2) Insight. Integral poses represent a new perspective that endeavours to combine the benefits of silhouettes and keypoints, as they contain rich shapes and possible motion information. This represents the pivotal essence of gait recognition.

(3) Inspiration. Integral poses serve as the inspiration for GaitApp, which is a pioneering solution. We aspire for this innovation to inspire further valuable explorations towards cross-covariate gait recognition.

Experiments conducted on CASIA-B [18] and FVG [19] show that our proposed method significantly improves the cross-covariate recognition performances of representative models without extra parameters. For example, GaitApp improves the CL accuracy of GaitSet [6] by 5.06% on CASIA-B and by 7.6% on FVG.

The contributions of this work are summarized as follows:

(1) Theoretically, we indicate the discriminative property of integral poses, which provides a new potential direction towards cross-covariate gait recognition and interpretability for gait research.

(2) Methodologically, we propose an Appearance transferable Disentangling and Generative Network (GaitApp) to generate new gait silhouette sequences with greater appearance diversity and invariant poses. To the best of our knowledge, this is one of the first attempts to synthesize frame-level gait silhouettes.

(3) Practically, extensive experiments demonstrate the effectiveness, compatibility, and generalizability of our method, which results in significant cross-covariate performance improvements for the representative gait models. In addition, GaitApp is applicable and valuable for potential research and exhibits practicality due to its low-cost and effective appearance transfer.

II. Related Work

A. Gait Recognition

Gait recognition models can generally be categorized into two main types:

1) Model-Based Gait Recognition: Zhao et al. [20] propose a 3D gait recognition framework to extract 3D human features by reconstructing a 3D human model with multi-view conditions. GaitGraph [16] employs a human pose estimator to extract 2D skeletons, and a GCN to extract the inherent structure of the skeleton. PoseGait [17] utilizes 3D human pose coordinates to design handcrafted features, which are then used to extract temporal-spatial features via CNNs. LUGAN [21] employs adversarial training to learn full-rank transformation matrices from the source pose and target views, and generate multi-view pose sequences for each single-view sample to reduce the cross-view variance.

2) Appearance-Based Gait Recognition: Gait silhouettes have proven the effectiveness and efficiency in gait recognition [6], [7], [13], [15]. GaitSet [19] first indicates that the appearance of a silhouette has contained its position information and extracts set-level features from an unordered gait sequence. GaitPart [7] presents that the different parts of the human body contain different visual and motion patterns, which are independent to extract spatial-temporal features. GaitGL [10] employs global and local feature extraction based on the 3DCNNs to extract fine-grained temporal-spatial features. In addition, some works use other input modalities, *e.g.*, GaitEdge [22] based on RGB frames, GaitParsing [23] based on human parsing, LidarGait [24] based on point clouds, Cross-modal Transfer Model [25] based on data captured by RGB-D, UGaitNet [26] with multimodal inputs.

B. Disentangling Representation Learning

Disentangling Representation Learning (DRL) has been leveraged in human identification to address occlusion, clothes-changing, and cross-view problems. Zhang et al. [19] introduce an encoder-decoder framework called GaitNet to disentangle pose information from visual RGB images based

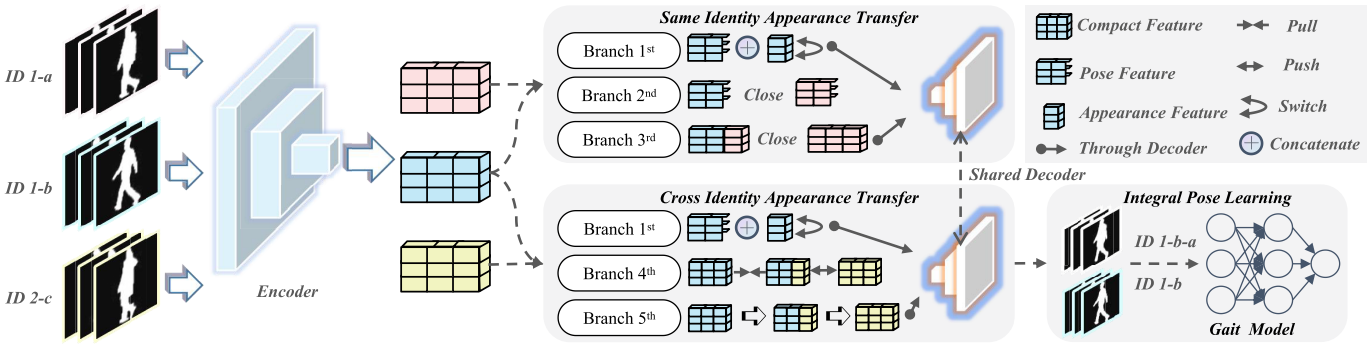


Fig. 2. The overview of GaitApp. ID 1-a and ID 1-b denote the same person with different appearances. ID 2-c represents another person with a different appearance. (1) First, encoder extracts compact features from gait silhouettes. (2) Then, two compact features from ID 1-a and ID 1-b flow into same identity appearance transfer (SIAT) with Branch 1st, Branch 2nd, and Branch 3rd for disentangling compact features into pose features and appearance features. (3) meanwhile, two compact features from ID 1-b and ID 2-c flow into cross identity appearance Transfer (CIAT) with Branch 1st, Branch 4th, and Branch 5th for the disentangling process and identity constraints. Finally, GaitApp generates gait silhouettes with new appearances (*i.e.*, ID 1-b-a) but the same pose silhouettes (*i.e.*, ID 1-b) for integral pose learning through the downstream gait model.

on two assumptions. First, the cross-frame appearance features in a sequence should be consistent, while the mean pose features of two sequences for the same person should be similar. Li et al. [27] propose ICDNet with semi-supervised DRL to disentangle identity and covariate features based on GEI inputs. ICDNet manually splits the compact feature of a GEI input into an identity feature and a covariate feature, and then two types of reconstruction are performed: self-reconstruction and the reconstruction of a canonical version without covariate features. Yao et al. [28] present group supervision DRL to disentangle poses, gaits, appearances, and view angles, which are aggregated to the final recognition process. Group supervision leverages multiple swapping operations between the same attributes to complete the disentangling process. GaitApp is different from the above mentioned methods in terms of four aspects:

(1) Frame-level Appearance Transfer. GaitApp conducts appearance transfer silhouette generation at the **frame level**, distinguishing from the previous methods that are primarily focused on GEIs [27], RGB [19], or decoupled processes with recognition objectives [19], [27], [28].

(2) Challenges. First, the appearance transfer process in gait silhouettes lacks paired samples with pose invariance and appearance variation. Second, the appearance transfer between frames within the same sequence is not consistent. Finally, identity information may be compromised by appearance transfer.

(3) Solution. GaitApp employs multi-branch cooperation to disentangle pose and appearance features, maintain the consistency of appearance transfer, and implicitly preserve the consistency of identities.

(4) Plug-and-Play Module. The data generated by GaitApp can be flexibly provided to downstream gait models, without imposing constraints on the architectures of the models, requiring extra parameters, or reducing the inference speed.

Person Re-identification (PR) has also demonstrated the effectiveness of disentangling and generative learning [29], [30], [31], [32], [33] in recent years. However, the task of appearance-transferable gait silhouette synthesis has basic differences:

(1) Challenges. Appearance transfer with silhouettes in GR primarily tackles the ambiguous human body shapes and edges. On the other hand, appearance transfer in PR usually involves different modalities, *e.g.*, RGB images, videos, and thermal images. As a result, the appearance transfer in PR mainly concerns the preservation of invariant backgrounds (*e.g.*, trees and obstacles) while transferring a large amount of complex information in the foreground (*e.g.*, colors or logos on clothes).

(2) Information. GR leverages a pair of gait sequences with different appearances to constrain the constraining appearance transfer with temporal consistency (*e.g.*, GEIs), and imposes implicit constraints or pose consistency to maintain identity consistency. On the other hand, PR has rarely been studied in video-based appearance transfer, and relies mainly on image-based appearance transfer. In contrast, PR employs single frame reconstruction to achieve background invariant and conduct human appearance transfer. Additionally, PR incorporates extra information (*e.g.*, faces) [29], [32] to maintain identity consistency.

Therefore, our framework is designed to minimize the difficulty of implementing the disentangling and generative processes. Additionally, we do not require fully decoupled features, and our work aims to generate realistic silhouettes with ***obvious appearance transfer*** and ***invariant poses***.

III. Proposed Method

This work aims to extract robust integral pose information by simulating a person constantly changing appearances under limited-covariate conditions. Therefore, we propose an Appearance-transferable Disentangling and Generative Network (**GaitApp**) to generate gait silhouettes with rich appearances and invariant poses. As illustrated in Fig. 2, GaitApp, an encoder-decoder architecture, leverages Same Identity Appearance Transfer (SIAT) and Cross Identity Appearance Transfer (CIAT) to disentangle the pose and appearance features by multi-branch cooperation. Finally, the generated silhouettes with new appearances and invariant poses, are fed into the downstream gait models for integral pose learning. For a better understanding, here we provide a brief overview

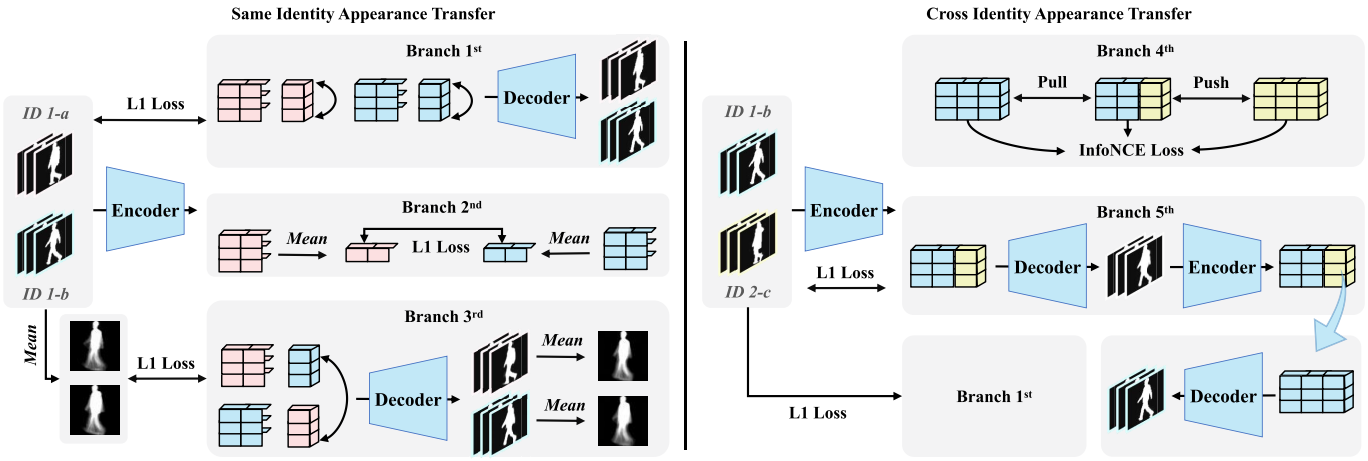


Fig. 3. The process of SIAT and CIAT. SIAT (Left.): Branch 1st, Branch 2nd and Branch 3rd are cooperated to disentangle the pose and appearance features. CIAT (Right.): Branch 4th and Branch 5 th are used to weakly supervise the identity consistency meanwhile maintaining the appearance diversity. In this case, Branch 1st used in Same Identity Appearance Transfer is also reserved.

of the full process. Two silhouette sequences (*i.e.*, ID 1-a and ID 1-b) share the same ID 1 but possess different appearances a and b, while the third sequence belongs to ID 2 with appearance c. Encoder extracts the compact feature of each silhouette in sequence. In SIAT, Branch 1st adopts Cross-Frame Reconstruction for each compact feature, aiming at maintaining pose consistency and avoiding over-represented appearance features. Branch 2nd employs Pose Feature Similarity to disentangle the pose features where a pair of pose features from the same person with different appearances (*i.e.*, ID-1-a and ID-1-b) are used as inputs. Branch 3rd leverages GEI Similarity to disentangle appearance features where two compact feature sequences for the same person transfer their appearance features to explicitly impose the appearance transfer constraint. In CIAT, Branch 4th employs Weak Identity Constraint (*i.e.*. InfoNCE) to maintain identity consistency during appearance transfer where triple compact features (*e.g.*, ID 1-b, ID 2-c, and ID 1-b-c for generated samples) serve as positive samples, negative samples, and anchor samples. Branch 5th adopts Gait Recycle Consistency to disentangle the pose features and appearance features through implicit appearance transfer and cycle reconstruction. Branch 1st is also applied in CIAT. Notably, Branch 1st, Branch 3rd and Branch 5th apply image-level constraints via Decoder, while Branch 2nd and Branch 4th apply feature-level constraints. It is worth mentioning that we adopt the same-view sampling strategy since our method aims at transferring appearances meanwhile maintaining invariant poses, which mitigates the learning difficulty faced by GaitApp. Next, we provide more details about GaitApp.

A. Feature Definition and Split

In this work, we empirically define pose and appearance features for two main reasons: 1) Since the predefined features contain human biases and we cannot distinguish which features are essential, we select pose and appearance features for simplicity. 2) GaitApp aims to generate frame-level gait silhouettes with obvious appearance changes and invariant poses, without focusing on disentangling learning for recognition tasks. Specifically, the pose features contain identity information, and the appearance features include identity-unrelated silhouette shapes (*e.g.*, carrying bags and clothes). We first leverage Encoder to extract a compact feature from a gait silhouette, and then, we split the compact feature into two parts: a pose feature and an appearance feature.

Formulation. Given an input gait silhouette sequence $\mathcal{X}_{in} \in \mathbb{R}^{\mathcal{S}\times\mathcal{C}_{in}\times\mathcal{H}\times\mathcal{W}}$, where $\mathcal{S}$ is the length of the gait sequence (e.g., $\mathcal{S}=30$), $\mathcal{C}_{in}$ is the number of channels (e.g., $\mathcal{C}_{in}=1$), and $(\mathcal{H},\mathcal{W})$ is the image size of each silhouette. Encoder aims to extract compact features $\mathcal{X}_{out}$ from each frame in $\mathcal{X}_{in}$. Then, the compact features $\mathcal{X}_{out}$ are manually split into pose features $\mathcal{F}_p$ and appearance features $\mathcal{F}_a$, respectively. This process can be formulated as follows:

$$\mathcal{X}_{out} = \mathcal{E}(\mathcal{X}_{in}) \tag{1}$$

$$\mathcal{F}_p, \mathcal{F}_a = Split(\mathcal{X}_{out}) \tag{2}$$

where $\mathcal{E}$ denotes Encoder, $\mathcal{X}_{out} \in \mathbb{R}^{\mathcal{S}\times\mathcal{C}_{out}\times1\times1}$, $\mathcal{F}_p \in \mathbb{R}^{\mathcal{S}\times\mathcal{C}_{part1}\times1\times1}$, $\mathcal{F}_a \in \mathbb{R}^{\mathcal{S}\times\mathcal{C}_{part2}\times1\times1}$, *i.e.*, $\mathcal{X}_{out}$ is split into two parts along the channels. To achieve feature disentanglement and silhouette generation, we divide the appearance transfer process into two categories based on gait representation properties: Same Identity Appearance Transfer (SIAT) and Cross Identity Appearance Transfer (CIAT).

B. Same Identity Appearance Transfer

The key to transferring the appearance information of a target silhouette to a source silhouette is to establish the relation between the appearance changes before and after. Fig. 3(Left.) shows that SIAT leverages, Branch 1st, Branch 2nd and Branch 3rd to separate the pose and appearance features.

Branch 1st: Cross-Frame Reconstruction. *"One person should be able to take off his or her coat and wear the coat again while maintaining status consistency."* The pose feature of the current silhouette is paired with the appearance feature of another silhouette in the same gait sequence to reconstruct the current silhouette.

Formulation. Given an input feature sequence $\mathcal{X}_{out} = \{(f_p^{s_n}, f_a^{s_n})|n \in 1, 2, 3, \ldots, \mathcal{S}\}$, the objective is as follows:

$$\mathcal{L}_{Branch-1} = \|\mathcal{D}(f_p^{s_i}, f_a^{s_j}) - \mathcal{I}_{s_i}\|_1 \tag{3}$$

where $\mathcal{D}$ denotes Decoder. $\mathcal{D}(f_p^{s_i}, f_a^{s_j})$ is the reconstructed silhouette $\tilde{\mathcal{I}}_{s_i}$ corresponding to the original input $\mathcal{I}_{s_i}$ at the s_i frame of the sequence. $f_a^{s_j}$ is the appearance feature of a random frame in the sequence.

Effect. Cross-frame reconstruction not only maintains the invariance of the poses but also prevents the over-representation of appearance features.

Branch 2nd: Pose Feature Similarity. *"One person takes off his or her shirt or coat, and the statuses maintain consistency."* Only Branch 1st may cause the pose feature to encode the full information of the corresponding silhouette, whereas the appearance feature degrades to a constant vector. To enable the appearance feature to contain rich covariate information of a silhouette, the pose feature needs to be as pure as possible. The main idea is to maintain the consistency between the pose features of two gait sequences produced by the same person with different appearances. However, it is unlikely to depict frame-level pose-alignment silhouettes from two gait sequences. Therefore, the second branch leverages the similarity between the means of the pose features in the two gait sequences.

Formulation. Given two gait sequences from the same person with different appearances, the pose feature similarity is defined as follows:

$$\mathcal{L}_{Branch-2} = \|\frac{1}{\mathcal{S}}\sum_{n=1}^{\mathcal{S}} f_p^{s_n} - \frac{1}{\mathcal{S}}\sum_{n=1}^{\mathcal{S}} f_p^{t_n}\|_1 \tag{4}$$

where $f_p^{s_n}$ and $f_p^{t_n}$ represent the pose features of one person in two gait sequences.

Effect. Pose feature similarity forces the pose feature to address appearance-unrelated information.

Branch 3rd: GEI Similarity. *"One person wears his or her coat and walks twice, and the statuses maintain consistency."* Our work aims to perform appearance transfer in a frame-level space for flexible applications. Usually, we need to model a relation between two images with only appearance differences. However, it is unlikely to achieve frame-level pose alignment from two gait sequences. We adopt Gait Energy Image [34] (GEI) similarity to approach this objective. Since GEIs are robust to pose information but sensitive to appearances, we enforce the GEI similarity between the generated gait sequence and the gait sequence with the target appearance.

Formulation. Given a source gait feature sequence $\mathcal{F}^s = \{(f_p^{s_n}, f_a^{s_n})|n \in 1, 2, 3, \ldots, \mathcal{S}\}$, GEI^s, and a target gait feature sequence $\mathcal{F}^t = \{(f_p^{t_n}, f_a^{t_n})|n \in 1, 2, 3, \ldots, \mathcal{S}\}$, GEI^t, the process is as follows:

$$\begin{aligned}\widetilde{\mathcal{F}}^s, \widetilde{\mathcal{F}}^t = Swap(\mathcal{F}^s, \mathcal{F}^t) = \{(f_p^{s_n}, f_a^{t_n})|n \in 1, 2, \ldots, \mathcal{S}\} \\ \&\{(f_p^{t_n}, f_a^{s_n})|n \in 1, 2, \ldots, \mathcal{S}\}\end{aligned} \tag{5}$$

$$\widetilde{GEI}^s, \widetilde{GEI}^t = \frac{1}{\mathcal{S}}\sum_{i=1}^{\mathcal{S}} \mathcal{D}(\widetilde{\mathcal{F}}^s), \frac{1}{\mathcal{S}}\sum_{i=1}^{\mathcal{S}} \mathcal{D}(\widetilde{\mathcal{F}}_t) \tag{6}$$

$$\mathcal{L}_{Branch-3} = \|GEI^s - \widetilde{GEI}^t\|_1 + \|GEI^t - \widetilde{GEI}^s\|_1 \tag{7}$$

Effect. GEI similarity enhances the visual appearance transfer in the image-level space. Notably, since gait recognition models usually take sequences as inputs, we treat the appearance as a sequence-level status representation.

C. Cross Identity Appearance Transfer

Although pose and appearance features can be disentangled through SIAT, visual failures and feature confusion (*e.g.*, identity information leakage) may occur during the appearance transfers in different identities. Different from SIAT where the image-level reconstruction enables to supervise identity consistency, the ground truth in the cross-identity condition is unavailable. Therefore, as Fig. 3(Right.) shows, GaitApp introduces Branch 4th and Branch 5th to implicitly supervise the identity consistency and realize a better visual transfer under the cross-identity condition. Similarly, Branch 1st is also used in Cross Identity Appearance Transfer, here is not described repeatedly.

Branch 4th: Weak Identity Constraint. *"One person wears the coats of others, but the identity should maintain consistency."* Usually, downstream recognition models leverage identity supervision to the upstream generation task in an end-to-end manner. However, we do not leverage the prior knowledge of recognition models for the following reasons: 1) The sequence-level information extracted by appearance-based gait models is difficult to distil into the frame-level appearance-transferable task. 2) The triplet loss [35] is widely used to narrow intra-class distances in the gait recognition task, but it also damages the appearance diversity of the generated gait sequences. Motivated by contrastive learning [32], [36], [37], [38], we leverage the InfoNCE [36] paradigm to implicitly and weakly supervise identity consistency in the feature space.

Formulation. Given a pair of input feature sequences $\mathcal{F}_c^s, \mathcal{F}_c^t = \{[f_c^{s_n}, f_c^{t_n}], n \in 1, 2, 3, \ldots, \mathcal{S}\}$, the principle is as follows:

$$(q, k_+, k_-) = \{(f_c^{s_n}, \widetilde{f}_c^{s_n}, f_c^{t_n}), n \in 1, 2, 3, \ldots, \mathcal{S}\} \tag{8}$$

$$\mathcal{L}_{Branch-4} = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{\mathcal{M}} \exp(q \cdot k_i/\tau)\Big)} \tag{9}$$

where $f_c^{s_n}$, $\widetilde{f}_c^{s_n}$ and $f_c^{t_n}$ denote the compact feature of the source silhouette, generated gait silhouette and target silhouette, respectively. q, k_+, k_-, τ, and $\mathcal{M}$ denote Query, Positive Keys, Negative Keys, Temperature and the number of keys for a query, respectively. Specifically, to maintain identity consistency and rich appearance diversity, we choose only one negative sample for a query.

Branch 5th: Gait Recycle Consistency. *"One person takes off his or her coat and wears another's coat and then takes it off and wear the original coat again. The identity should be consistent during this process."* The ground truth of an appearance-changing gait sequence is not available in the cross-identity problem. Therefore, we leverage Gait Recycle Consistency (**GRC**) to implicitly complete the appearance-transferable process. GRC enables the generated gait sequence to maintain the same pose as that of the source subject and the same appearance as that of the target subject.

TABLE I

THE RANK-1 ACCURACY (%) ON CASIA-B FOR DIFFERENT PROBE VIEWS EXCLUDING THE IDENTICAL-VIEW CASES.FOR EVALUATION, THE SEQUENCES OF NM-1,2,3,4 FOR EACH SUBJECT ARE TAKEN AS THE GALLERY

	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM	Base	**93.00**	**99.10**	99.80	97.80	**96.10**	92.80	96.60	97.80	99.20	98.20	**91.10**	96.50
	GaitApp-Base	92.70	98.30	99.80	**98.10**	95.00	**93.90**	**97.20**	**99.20**	**99.50**	**99.10**	89.50	**96.57**
	GaitSet [6]	92.80	98.80	99.30	98.00	94.60	91.60	95.40	97.90	98.40	97.50	**90.40**	95.88
	GaitApp-GaitSet	**93.20**	**98.90**	99.30	98.00	**95.10**	**93.10**	**96.20**	**99.00**	**99.00**	**98.20**	87.20	**96.11**
	GaitPart [7]	92.80	98.00	99.20	**98.80**	95.20	92.70	95.50	98.10	**99.30**	**97.80**	**90.50**	96.17
	GaitApp-GaitPart	**93.50**	**98.30**	**99.50**	98.20	**95.50**	**92.80**	**96.60**	**98.80**	99.10	97.30	88.60	**96.20**
	GaitGL [10]	95.10	**98.50**	**99.10**	**97.90**	**96.60**	**95.00**	**97.80**	**98.90**	**99.00**	**98.80**	93.70	**97.31**
	GaitApp-GaitGL	**95.20**	97.90	98.80	97.70	95.60	94.30	97.20	98.70	98.90	98.60	**93.90**	96.98
BG	Base	88.30	95.40	95.00	94.14	90.30	84.40	88.30	92.80	96.80	96.67	85.00	91.56
	GaitApp-Base	**88.80**	**96.60**	**96.60**	**94.45**	**91.50**	**86.30**	**91.80**	**95.20**	**96.90**	**96.47**	**87.20**	**92.89**
	GaitSet [6]	**87.20**	93.70	94.40	92.02	86.90	81.00	85.90	92.10	96.40	**95.66**	**84.60**	89.99
	GaitApp-GaitSet	86.80	**94.60**	**95.80**	**92.93**	**88.80**	**84.30**	**88.80**	**93.60**	**97.20**	95.45	83.40	**91.06**
	GaitPart [7]	88.40	93.50	**96.30**	92.83	87.10	82.90	88.10	93.60	95.20	93.13	84.10	90.47
	GaitApp-GaitPart	**88.70**	**94.40**	95.70	**93.33**	**90.30**	**86.10**	**90.40**	**95.60**	**96.40**	**94.04**	**85.10**	**91.83**
	GaitGL [10]	91.90	**96.00**	97.00	95.25	94.30	89.30	92.60	96.20	97.50	**97.07**	**92.20**	94.48
	GaitApp-GaitGL	**92.00**	95.20	97.00	**95.66**	**94.60**	**90.90**	**92.80**	**96.70**	**98.00**	96.87	89.70	**94.49**
CL	Base	73.10	82.80	86.00	81.60	75.90	74.00	75.10	78.30	80.10	78.90	66.70	77.50
	GaitApp-Base	**77.10**	**88.80**	**89.00**	**86.40**	**81.20**	**77.70**	**82.00**	**84.70**	**85.80**	**82.80**	**69.00**	**82.23**
	GaitSet [6]	67.90	79.80	82.60	78.30	71.70	68.10	71.60	75.80	77.20	75.10	62.80	73.72
	GaitApp-GaitSet	**74.00**	**84.90**	**86.50**	**83.00**	**76.40**	**74.10**	**76.10**	**80.30**	**84.70**	**81.20**	**65.40**	**78.78**
	GaitPart [7]	71.60	83.20	85.80	81.40	76.30	71.00	77.70	79.80	80.80	79.20	66.00	77.53
	GaitApp-GaitPart	**75.80**	**87.60**	**89.80**	**84.10**	**78.60**	**77.70**	**80.80**	**84.60**	**86.50**	**83.80**	**70.70**	**81.82**
	GaitGL [10]	76.30	90.60	90.40	89.70	84.40	78.20	84.10	86.80	87.70	85.10	70.20	83.96
	GaitApp-GaitGL	**77.90**	**91.60**	**93.80**	**91.10**	**86.90**	**81.60**	**88.10**	**90.80**	**92.00**	**87.40**	**71.30**	**86.59**

Formulation. Given a source gait sequence $\mathcal{X}^s$ and a target gait sequence $\mathcal{X}^t$, this case is shown as follows:

$$(\mathcal{F}_p^s, \mathcal{F}_a^s), (\mathcal{F}_p^t, \mathcal{F}_a^t) = \mathcal{E}(\mathcal{X}^s), \mathcal{E}(\mathcal{X}^t) \tag{10}$$

$$\widetilde{\mathcal{X}}^s, \widetilde{\mathcal{X}}^t = \mathcal{D}(\mathcal{F}_p^s, \mathcal{F}_a^t), \mathcal{D}(\mathcal{F}_p^t, \mathcal{F}_a^s) \tag{11}$$

$$(\widetilde{\mathcal{F}}_p^s, \widetilde{\mathcal{F}}_a^t), (\widetilde{\mathcal{F}}_p^t, \widetilde{\mathcal{F}}_a^s) = \mathcal{E}(\widetilde{\mathcal{X}}^s), \mathcal{E}(\widetilde{\mathcal{X}}^t) \tag{12}$$

$$\mathcal{X}^{s-recycle}, \mathcal{X}^{t-recycle} = \mathcal{D}(\widetilde{\mathcal{F}}_p^s, \mathcal{F}_a^s), \mathcal{D}(\widetilde{\mathcal{F}}_p^t, \mathcal{F}_a^t) \tag{13}$$

$$\mathcal{L}_{Branch-5} = \|\mathcal{X}^s - \mathcal{X}^{s-recycle}\|_1 + \|\mathcal{X}^t - \mathcal{X}^{t-recycle}\|_1 \tag{14}$$

where s and t denote the source and target samples, respectively, and $\mathcal{F}_p$ and $\mathcal{F}_a$ denote the pose and appearance features, respectively.

Finally, the generated gait sequences with new appearances and invariant poses are fed into the following gait recognition model for integral pose learning, further improving the recognition accuracy.

D. Training Strategy

Our work has two trainable networks: GaitApp and downstream gait recognition models. We train GaitApp with the following objectives:

$$\mathcal{L}_{GaitApp} = \mathcal{L}_{Branch-1} + \mathcal{L}_{Branch-2} + \mathcal{L}_{Branch-3} + \mathcal{L}_{Branch-4} + \mathcal{L}_{Branch-5} \tag{15}$$

After the training of GaitApp, we freeze the parameters of GaitApp to generate gait sequences with new appearances and invariant poses for training the downstream gait recognition models. In addition,, the vanilla training database is also added to the training process for cross-view learning. In this work, we select OpenGait-Baseline[1] (denoted as Base), GaitSet [6], GaitPart [7], and GaitGL [10] to evaluate the effectiveness, compatibility, and generalizability of our method. Here, we take Base as an example, and the loss function is defined as follows:

$$\mathcal{L}_{Base} = \lambda_1^o \mathcal{L}_{tp-original} + \lambda_2^o \mathcal{L}_{ce-original} + \lambda_1^g \mathcal{L}_{tp-generate} + \lambda_2^g \mathcal{L}_{ce-generate} \tag{16}$$

where $\mathcal{L}_{tp}$ and $\mathcal{L}_{ce}$ denote Triplet Loss and Cross-entropy Loss, respectively; λ_1^o and λ_2^o represent the weight of original branch, which is consistently set to downstream gait models in their manuscripts; λ_1^g and λ_2^g represent the weight of generation branch.

[1] https://github.com/ShiqiYu/OpenGait

TABLE II

THE RANK-1 AND RANK-5 ACCURACY (%) ON FVG UNDER 5 PROTOCOLS WALKING SPEED (WS), CARRYING BAG WHILE WEARING A HAT(BGHT), CHANGING CLOTHES (CL), MULTIPLE PERSONS (MP), AND ALL VARIATIONS (ALL). THE FIRST AND SECOND COLUMNS IN EACH PROTOCOL DENOTE THE GALLERY AND PROBE ON EVALUATION

Protocol	WS		BGHT		CL		MP		ALL	
Index of Gallery & Probe										
Session 1	2	4-9	2	10-12	-	-	-	-	2	1, 3-12
Session 2	2	4-6	-	-	2	7-9	2	10-12	2	1, 3-12
Session 3	-	-	-	-	-	-	-	-	-	1 - 12
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Base	**98.35**	99.67	100.00	100.00	56.12	72.57	95.78	99.58	75.84	81.13
GaitApp-Base	98.02	99.67	100.00	100.00	**62.45**	**83.97**	**97.05**	**100.00**	**77.87**	**83.42**
GaitSet [6]	**98.68**	99.67	100.00	100.00	55.27	70.04	96.20	98.31	75.57	80.60
GaitApp-GaitSet	98.02	99.67	100.00	100.00	**62.87**	**82.28**	**96.62**	**98.73**	**77.34**	**82.80**
GaitPart [7]	98.02	99.34	100.00	100.00	77.22	**94.94**	96.20	99.16	81.04	85.71
GaitApp-GaitPart	**98.68**	**99.67**	100.00	100.00	**79.32**	94.09	**96.62**	99.16	**81.39**	85.71
GaitGL [10]	**98.68**	99.67	100.00	100.00	83.97	96.20	97.89	**99.58**	82.80	86.33
GaitApp-GaitGL	98.35	99.67	100.00	100.00	**86.92**	**97.89**	97.89	99.16	**83.60**	**86.51**

TABLE III

CROSS-DOMAIN SCENARIO FROM CASIA-B TO FVG

Method	CASIA-B → FVG				
	WS	BGHT	CL	MP	ALL
Base	89.11	78.79	67.51	89.87	73.37
GaitApp-Base	**91.09**	**78.79**	**70.89**	**91.98**	**75.49**
GaitSet	88.78	72.73	62.03	85.65	71.08
GaitApp-GaitSet	**91.09**	**72.73**	**64.98**	**86.08**	**72.57**
GaitPart	89.77	75.76	73.42	86.50	73.99
GaitApp-GaitPart	**91.09**	**84.85**	**75.53**	**89.45**	**75.75**
GaitGL	93.73	93.94	73.84	91.98	77.34
GaitApp-GaitGL	**95.05**	**96.97**	**76.37**	**92.41**	**78.48**

IV. EXPERIMENTS

A. Database

Cross-covariate gait recognition usually involves cross-view, cross-carrying, and cross-clothing conditions. Significant progress has been achieved with respect to the cross-view problem in recent studies [6], [7], [10] while the cross-carrying and cross-clothes problems still require further exploration and refinement. In particular, cross-clothing gait recognition remains the most challenging problem since it extremely degrades shape and motion patterns. Therefore, we take the cross-clothes protocol to facilitate our study. Many previous public gait databases do not collect cross-clothing sequences, e.g., OUMVLP [39] or contain only one type of covariate and view variation, e.g., OU-LP-Bag [40]. In addition, OU-Treadmill-B [41] contains various clothes, but only one sequence of clothing types for each subject, which is unable to facilitate our study. The recent gait databases collected in the wild, GREW [42] and Gait3D [43] also cannot satisfy our requirements since whether a subject possesses cross-clothing variations is inaccessible. Specifically, GREW captures videos for one day and Gait3D randomly samples two segments of continuous two-hour video clips from seven-day raw videos in a large supermarket, which makes it difficult to collect cross-clothing sequences for a subject. Finally, we evaluate our method on two public gait databases. **CASIA-B [18].** There are 124 subjects on CASIA-B. Each subject possesses three variations, normal walking (NM), carrying bags (BG) and cloth-changing condition (CL), with 11 views (0°, 18°, ..., 180°). Due to the low segmentation quality of the 5th subject, we utilize the first 73 subjects as the training set and the last 50 subjects as the test set according to protocols [18]. For all gait models, the image size of each silhouette is set to 64×64, and the number of sequences in a batch is set to $(8, 16)$ where 8 subjects are sampled, and each subject contains 16 sequences. For evaluation, NM 5-6, BG 1-2 and CL 1-2 of each subject are taken as the probe, and NM 1-4 are as the gallery.

1) Front-View Gait Database [19]: As Tab. II shows, FVG collected Session 1 (135 subjects) and Session 2 (79 subjects) in 2017 and 2018, respectively, and Session 3 (12 subjects) in both years [44], [45]. Each subject includes 12 sequences with 4 variations captured by three front view angles (−45°, 0°, 45°). We define 5 protocols according to [44], i.e., Walking Speed (WS), Carrying Bag while Wearing a Hat (BGHT), Changing Clothes (CL), Multiple Persons (MP), and all variations (ALL). The first 136 subjects are used as the training data, and the remaining 90 subjects are employed as the test data. The image size of each silhouette and batch size are set to 64×64 and (8, 16), respectively.

B. Implementation Details

GaitApp consists of an encoder and a decoder. Encoder consists of several Conv Blocks, LeakyReLU and Max Pooling, and Decoder contains several ConvTrans Blocks, Batch Normalization, LeakyReLU and Sigmoid. More details are presented in Tab. IV. GaitApp can be a compatible module for

TABLE IV
THE ARCHITECTURE OF GAITAPP. CONV BLOCK CONSISTS OF TWO CONVOLUTION LAYERS WITH LEAKYRELU. CONVTRANS BLOCK CONSISTS OF A CONVOLUTION TRANSPOSE LAYER, BATCH NORMALIZATION AND LEAKYRELU. K, S, AND P ARE FOR THE KERNEL SIZE, STRIDE, AND PADDING, RESPECTIVELY

Encoder		
Layers	Output size	(K), (S), (P)
Conv Block1	(T, 32, 64, 64)	(5, 5, 32), (1), (2) (3, 3, 32), (1), (1)
Max Pooling	(T, 32, 32, 32)	(2, 2), (0), (0)
Conv Block2	(T, 64, 32, 32)	(3, 3, 64), (1), (1) (3, 3, 64), (1), (1)
Max Pooling	(T, 64, 16, 16)	(2, 2), (0), (0)
Conv Block3	(T, 128, 16, 16)	(3, 3, 128), (1), (1) (3, 3, 128), (1), (1)
Adaptive AvePooling	(T, 128, 1, 1)	-
Pose Appearance	(T, 96, 1, 1) (T, 32, 1, 1)	-
Decoder		
ConvTrans Block1	(T, 64, 4, 4)	(4, 4, 64), (1), (0)
ConvTrans Block2	(T, 32, 8, 8)	(4, 4, 32), (2), (1)
ConvTrans Block3	(T, 16, 16, 16)	(4, 4, 16), (2), (1)
ConvTrans Block4	(T, 8, 32, 32)	(4, 4, 8), (2), (1)
ConvTrans Sigmoid	(T, 1, 64, 64)	(4, 4, 1), (2), (1)

gait models with silhouette inputs. Therefore, we reproduce several representative gait models with different properties, OpenGait-Baseline (denoted as Base), GaitSet [6], GaitPart [7] and GaitGL [10] referred to as Opengait[1]. GaitApp uses Adam as the optimizer with an initial learning rate of 1e-4 and the weight decay of 5e-4. The number of training epochs is set to 20K, and the learning rate is reset to 1e-5 after 10K. All the models are trained on NVIDIA 3090 GPUs.

C. Performance Comparison

GaitApp enables the generation of frame-level gait silhouettes that downstream gait models can use to learn integral pose features, which is robust to cross-covariate problems.

In this section, we show the results of our method with the representative models on CASIA-B [18] and FVG [19]. Notably, since cross-clothing condition severely degrades shape and motion patterns, we use the CL protocol on CASIA-B and FVG as the main criteria.

1) CASIA-B: Tab. I shows that increasing the CL diversity for each subject by our method on CASIA-B with our method further improves the CL performance of all the downstream models (+4.73% for Base, +5.06% for GaitSet, +4.29% for GaitPart, +2.63% for GaitGL) without damaging the NM condition. In addition, the BG performances of Base, GaitSet and GaitPart increase by +1.33%, +1.07% and +1.36%, respectively. These results demonstrate the effectiveness of integral pose information for cross-covariant gait recognition.

2) FVG: Tab. II shows that our method also significantly improves the CL protocol with Rank-1 and Rank-5 over all downstream models (+6.33%, +11.4% for Base, +7.60%, +12.24% for GaitSet, +2.1%, −0.85% for GaitPart, +2.95%, +1.69% for GaitGL). It is noted that CL performance improvement attained on GaitPart is not evident since the cross-clothing variables for the front view angle cause GaitPart to have difficulty extracting partial motion patterns. Notably, since the other benchmarks reach saturation on FVG, we consider these saturated metrics as references to ensure that GaitApp provides gait data with rich appearances, enabling gait models to learn integral poses without introducing adverse effects.

3) Cross-Domain Scenarios: We conduct cross-domain experiments to demonstrate the robust gains achieved by integral pose learning with GaitApp. Tab. III presents the results obtained after training on CASIA-B and testing on FVG where GaitApp yields significant gains in the cross-domain benchmark, which is consistent with the source-domain evaluation. As a result, the integral pose can provide robust identity information, even under the most challenging CL scenario.

D. Ablation Study

GaitApp aims to generate gait silhouettes with new appearances and invariant poses. Therefore, we provide more cross-covariate examples. In addition, we carefully design a multi-branch cooperation to disentangle pose features and appearance features. To better understand the principles, we visualize the results of each branch. We also explore the impact of appearance transfer on downstream models and the interpretability of the integral poses.

1) Visualization for Appearance Transfer: GaitApp aims to generate frame-level gait silhouettes with rich appearances and invariant poses, forcing gait recognition models to learn integral poses. Fig. 4(a) shows that the appearance of Target 44 with CL is transferred to Source 57 with NM, and then a new gait sequence with the same identity that of Source 57 and the corresponding poses is generated (viewed in the second row) meanwhile a new appearance is obtained from Target 44. More examples are visualized in Fig. 4(b, c, d).

2) Branch Analysis: To better understand and validate multi-branch principles from the visualization perspective, Fig. 7 shows that Branch 1st, Branch 3rd and Branch 5th cooperate to achieve visual appearance reconstruction in SIAT and CIAT. From the design perspective, Branch 1st disentangles pose features through Cross-Frame Reconstruction within a sequence, and thus is applied in both SIAT and CIAT. Branch 2nd and Branch 3rd, which require consistent poses and appearances within the same ID samples, are exclusively applied in SIAT. Branch 4th implicitly maintains identity consistency in the cross-ID appearance transfer and is only applied in CIAT. As for Branch 5th, since the ground truth for appearance-transfer gait sequences is unavailable in CIAT, we employ Branch 5th to achieve implicit appearance transfer through the self-supervised mechanism. However, theoretically, Branch 5th can be also applied in both SIAT and CIAT, as a combination of supervised and self-supervised learning to further enhance the appearance transfer ability. To better validate this principle, we conduct ablation experiments utilizing Branch 5th in both SIAT and CIAT. Tab. VI indicates

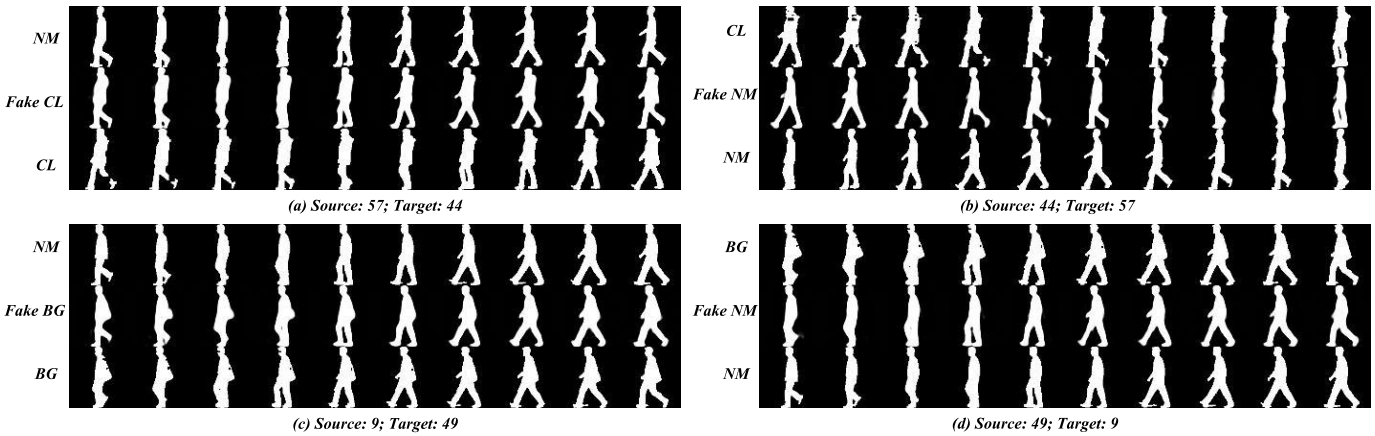


Fig. 4. The visualization of appearance transfer. Source (viewed in the first row) and target (viewed in the third row) denote the numbers of the subject on CASIA-B. GaitApp enables to transfer the appearances of the target silhouettes to the source silhouettes meanwhile maintaining the pose consistency of source silhouettes. Notably, here, the covariate type is not strict (*i.e.*, the BG transfer process also includes clothes transfer, and even hairstyle transfer).

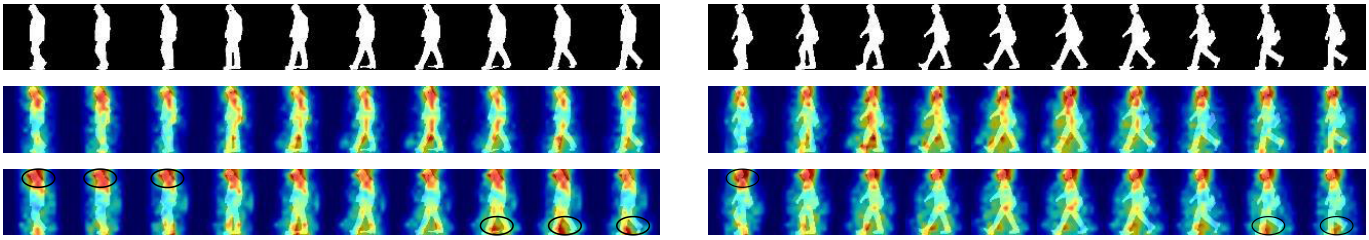


Fig. 5. Class activation map [46]. The first, second, and third rows denote the gait silhouettes, the CAM extracted by GaitSet, and the CAM extracted by GaitApp-GaitSet, respectively. The black circles represent key regions of the integral poses, including the angle of the head lift, the regions of the shoulder and neck, and the length of the two thighs.

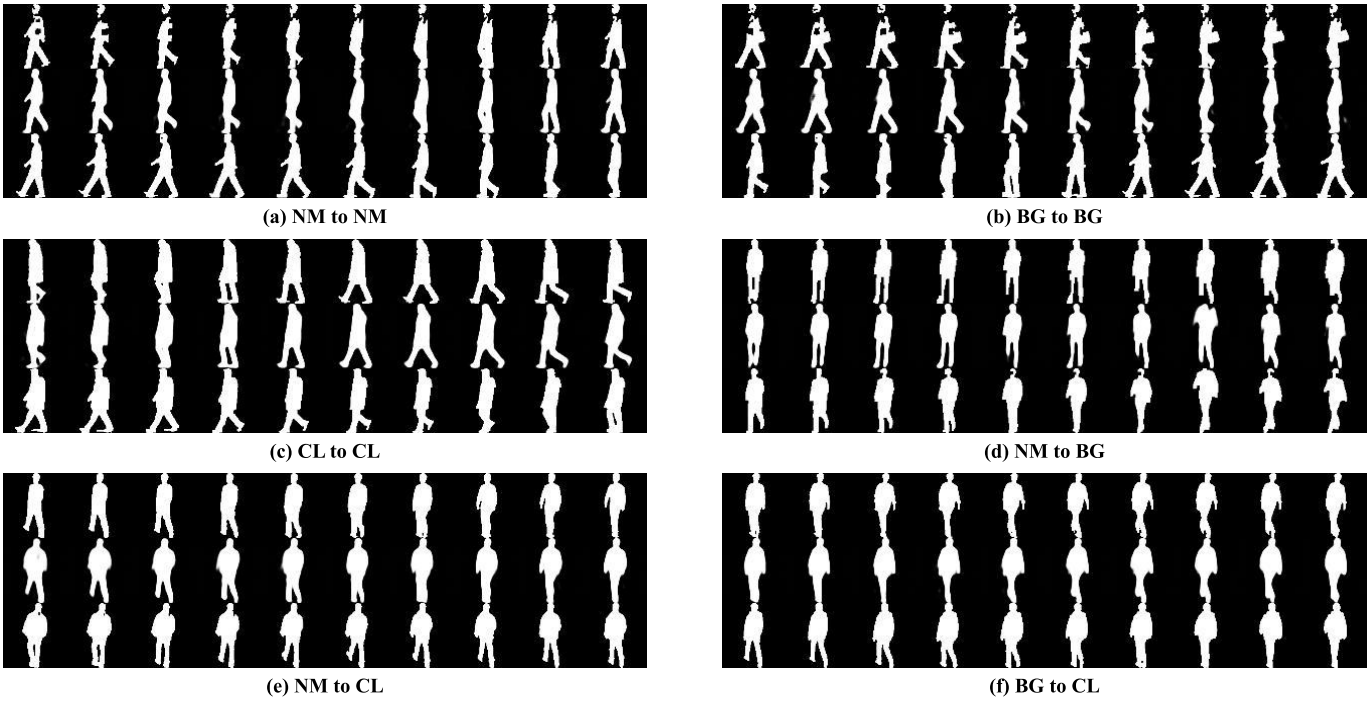


Fig. 6. The visualization of appearance transfer. Source (viewed in the first row) and target (viewed in the third row) denote the subject on CASIA-B. GaitApp enables to transfer the appearances of the target silhouettes into the source silhouettes meanwhile maintaining the pose consistency of source silhouettes. Notable, here covariate type is not strict (i.e., The process of BG transfer also contains clothes transfer, even hairstyle transfer).

that the additional GRC in SIAT, while incurring extra training overhead, does not result in significant recognition gains.

3) Impact for Appearance Transfer: As shown in Tab. V, appearance transfer improves CL performance with increasing

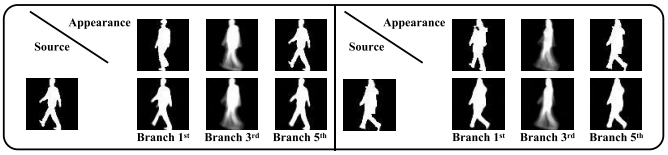


Fig. 7. Branch visualization. Branch 1st, Branch 3rd and Branch 5th are for cross-frame Reconstruction, GEI similarity and gait recycle consistency, respectively.

TABLE V

THE IMPACT OF APPEARANCE TRANSFER ON CASIA-B. THE λ_1^g AND λ_2^g DENOTE THE LOSS WEIGHT OF THE GENERATION BRANCH. WE SET A SCALE FACTOR OF 10 BETWEEN λ_1^g AND λ_2^g ACCORDING TO OPENGAIT[1]

Method	λ_1^g	λ_2^g	NM	BG	CL
GaitApp-Base	0.0	0.0	96.50	91.56	77.50
	0.1	0.01	96.80	92.38	80.21
	0.5	0.05	**96.82**	92.49	81.25
	1.0	0.1	96.57	**92.89**	82.23
	2.0	0.2	96.05	91.88	**82.45**

TABLE VI

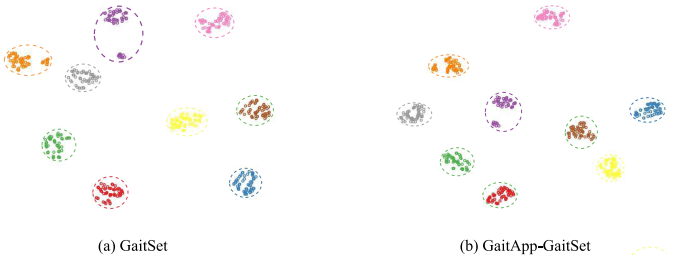
THE IMPACT ON SIAT W/ GAIT RECYCLE CONSISTENCY (GRC) IN GAITAPP. $^\diamond$ DENOTES THE ORIGINAL RESULTS (*i.e.*, SIAT W/O GRC)

Method	NM	BG	CL	Mean
Base w/ GRC	96.54	93.09	81.76	90.46
Base$^\diamond$	**96.57**	**92.89**	**82.23**	**90.56**
GaitSet w/ GRC	96.04	91.16	77.66	88.29
GaitSet$^\diamond$	**96.11**	**91.06**	**78.78**	**88.65**
GaitPart w/ GRC	96.30	91.63	80.24	89.39
GaitPart$^\diamond$	**96.20**	**91.83**	**81.82**	**89.95**
GaitGL w/ GRC	96.01	93.42	86.16	91.86
GaitGL$^\diamond$	**96.98**	**94.49**	**86.59**	**92.69**

λ_1^g and λ_2^g. However, the over-learning in the cross-covariate condition may damage the discriminative features (*e.g.*, hand and body motion) in NM condition. Therefore, the magnitude of appearance transfer for the downstream recognition task is a trade-off.

4) Interpretability for Integral Poses: To further understand integral poses, we visualize the Class Activation Map (CAM) [46] of GaitSet [6] and GaitApp-GaitSet in Fig. 5. Compared with GaitSet, GaitApp further complements the integral pose information, including the angle of the head lift, the shoulder and neck regions, and the length of the two thighs (viewed in black circles).

5) Feature Distribution: We randomly select nine IDs from the CASIA-B test set, extract one-half of the sequences for each ID, and use each colour to represent a distinct ID (*i.e.*, class) extracted from GaitSet or GaitApp-GaitSet for t-SNE visualization. The clear contrast between (a) and (b) in Fig. 8 reveals that, for each class, sub-clusters are contained within their clusters. GaitApp-GaitSet yields more compact intra-class representations.



(a) GaitSet (b) GaitApp-GaitSet

Fig. 8. T-SNE of feature distribution on CASIA-B test database. (a) Gait features of nine identities (marked by various colours) extracted from GaitSet. (b) Gait features of the same nine identities extracted from GaitApp-GaitSet.

V. DISCUSSION

1) Appearance Transfer. GaitApp can generate gait silhouettes with rich appearances and invariant poses. Fig. 6 shows the cross-covariate appearance transfer on CASIA-B [18]. It is worth noting that since appearance information consists of clothes, bags, and other covariates, these cross-type transfers on CASIA-B are not completely strict (e.g., BG transfer still includes clothes information). Additionally, CASIA-B may contain segmentation errors and low-quality silhouettes, but GaitApp enables to complement missing regions through statistics information from the training data.

2) Accuracy Performance. Since GaitApp is a plug-and-play module, we select different types of gait models, such as GaitSet [6], GaitPart [7], and GaitGL [10] for validating the effectiveness and generalizability of GaitApp. We expect that a kind of new gait feature can provide novel insights instead of incremental performance improvements.

3) Ethical Statement. We strictly follow ethical research laws. CASIA-B and FVG are two public gait databases where the subjects have provided their consent on such data collection. Additionally, our research is based on silhouettes, which also protect the privacy of human information.

VI. CONCLUSION

In this work, we first introduce integral pose learning, aiming to adaptively learn robust regions that are derived by contrasting the same person with different appearances but the same pose, marking a new direction in cross-covariate gait recognition. Towards this goal, we propose GaitApp with multi-branch cooperation to generate frame-level gait silhouettes with rich appearances and invariant poses. To the best of our knowledge, this is one of the first attempts to generate frame-level silhouettes with diverse appearances. Extensive experiments validate the effectiveness of integral pose learning with samples generated by GaitApp in cross-covariate and cross-domain scenarios. In addition, since GaitApp is a low-cost yet effective appearance transfer, it can be a potential application in gait research, *e.g.*, hard sample mining, or reducing the labour costs for gait data collection in the industry.

Limitations and Future Works. Although GaitApp does not require frame alignment, it still relies on pairs of samples exhibiting diverse appearances, which poses challenges in terms of acquiring and annotating large-scale gait data in

the wild. Recently, generative models (*e.g.*, Diffusion Models) [47], which incrementally engender target data through continual noise diffusion without paired inputs, have achieved remarkable milestones. Consequently, we will continue to explore unsupervised or self-supervised generation paradigms. Additionally, appearance transfer at the image level may induce visual blurring and confusion in identity information, potentially affecting the downstream recognition task. We will try to explore appearance transfer at the feature level with the recognition process in a unified framework in the future. Finally, the samples generated by GaitApp with pose invariance and rich appearances can serve as ordinary gait data, which may provide a potential direction for the interpretability of cross-covariate gait recognition via quantitative and qualitative analyses (*e.g.*, causal analysis).

REFERENCES

[1] 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, Feb. 2017.

[2] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 102–113, Jan. 2019.

[3] C. Song, Y. Huang, W. Wang, and L. Wang, "CASIA-E: A large comprehensive dataset for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2801–2815, Mar. 2023.

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

[5] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3102–3115, Dec. 2019.

[6] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8126–8133.

[7] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14225–14233.

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

[9] S. Hou, X. Liu, C. Cao, and Y. Huang, "Set residual network for silhouette-based gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 384–393, Jul. 2021.

[10] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14648–14656.

[11] 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.*, vol. 34, no. 11, pp. 8978–8988, Nov. 2023.

[12] J. Wang et al., "Causal intervention for sparse-view gait recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 77–85.

[13] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3054–3062.

[14] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023.

[15] X. Huang et al., "Context-sensitive temporal feature learning for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12909–12918.

[16] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hormann, and G. Rigoll, "GaitGraph: Graph convolutional network for skeleton-based gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2314–2318.

[17] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[18] 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. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, Aug. 2006, pp. 441–444.

[19] Z. Zhang et al., "Gait recognition via disentangled representation learning," in *Proc. CVPR*, Jun. 2019, pp. 4710–4719.

[20] G. Zhao, G. Liu, H. Li, and M. Pietikainen, "3D gait recognition using multiple cameras," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, Apr. 2006, pp. 529–534.

[21] H. Pan, Y. Chen, T. Xu, Y. He, and Z. He, "Toward complete-view and high-level pose-based gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2104–2118, 2023.

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

[23] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao, and Y. Huang, "GaitParsing: Human semantic parsing for gait recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 4736–4748, 2024.

[24] C. Shen, F. Chao, W. Wu, R. Wang, G. Q. Huang, and S. Yu, "Lidar-Gait: Benchmarking 3D gait recognition with point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1054–1063.

[25] X. Gu, Y. Guo, F. Deligianni, B. Lo, and G.-Z. Yang, "Cross-subject and cross-modal transfer for generalized abnormal gait pattern recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 546–560, Feb. 2021.

[26] M. J. Marin-Jimenez, F. M. Castro, R. Delgado-Escano, V. Kalogeiton, and N. Guil, "UGaitNet: Multimodal gait recognition with missing input modalities," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5452–5462, 2021.

[27] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13306–13316.

[28] L. Yao, W. Kusakunniran, P. Zhang, Q. Wu, and J. Zhang, "Improving disentangled representation learning for gait recognition using group supervision," *IEEE Trans. Multimedia*, vol. 25, pp. 4187–4198, May 2022.

[29] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

[30] W. Xu, H. Liu, W. Shi, Z. Miao, Z. Lu, and F. Chen, "Adversarial feature disentanglement for long-term person re-identification," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1201–1207.

[31] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.

[32] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2004–2013.

[33] Y. Yan et al., "Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1523–1529.

[34] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

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

[36] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[37] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14960–14969.

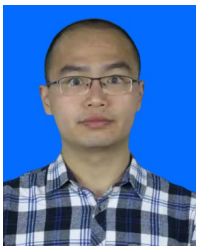
[38] T. Teepe, J. Gilg, F. Herzog, S. Hormann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1569–1577.

[39] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

[40] M. Z. Uddin et al., "The OU-ISIR large population gait database with real-life carried object and its performance evaluation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–11, Dec. 2018.

[41] Y. Makihara et al., "The OU-ISIR gait database comprising the treadmill dataset," *IPSJ Trans. Comput. Vis. Appl.*, vol. 4, pp. 53–62, 2012.
[42] Z. Zhu et al., "Gait recognition in the wild: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Jun. 2021, pp. 14789–14799.
[43] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3D representations and a benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 20228–20237.
[44] 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, Jan. 2022.
[45] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen, "Clothes-changing person re-identification with RGB modality only," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1060–1069.
[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[47] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33. Vancouver, BC, Canada: Curran Associates, 2020, pp. 6840–6851.

Panjian Huang received the B.E. degree from Qingdao University in 2019 and the M.S. degree from The University of Manchester in 2020. He is currently pursuing the Ph.D. degree with Beijing Normal University. His research interests include computer vision, pattern recognition, and gait recognition.



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

Chunshui Cao received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2013 and 2018, respectively. During the Ph.D. study, he joined the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. From 2018 to 2020, he was a Post-Doctoral Fellow with the PBC School of Finance, Tsinghua University. He is currently a Research Scientist with Watrix Technology Limited Company Ltd. His research interests include pattern recognition, computer vision, and machine learning.

Xu Liu received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2013 and 2018, respectively. He is currently a Research Scientist with Watrix Technology Limited Company Ltd. His research interests include gait recognition, object detection, and image segmentation.

Xuecai Hu received the B.Sc. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2015 and 2021, respectively. He is currently with the Postdoctoral Scientific Research Station, School of Artificial Intelligence, Beijing Normal University. His research interests include computer vision, image enhancement, semantic segmentation, and pose estimation.

Yongzhen Huang (Senior Member, IEEE) received the B.E. degree from Huazhong University of Science and Technology in 2006 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2011. He is currently a Professor with the School of Artificial Intelligence, Beijing Normal University. He has published one book and more than 100 papers at international journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, CVPR, ICCV, ECCV, NIPS, and AAAI. His research interests include pattern recognition, computer vision, and machine learning.