Gait Recognition via Semi-supervised Disentangled Representation Learning to Identity and Covariate Features

Xiang Li[1,2] Yasushi Makihara[2] Chi Xu[1,2] Yasushi Yagi[2] Mingwu Ren[1]
[1] Nanjing University of Science and Technology, [2] Osaka University
{lixiangmzlx, xuchisherry}@gmail.com {makihara, yagi}@am.sanken.osaka-u.ac.jp
renmingwu@mail.njust.edu.cn

Abstract

Existing gait recognition approaches typically focus on learning identity features that are invariant to covariates (e.g., the carrying status, clothing, walking speed, and viewing angle) and seldom involve learning features from the covariate aspect, which may lead to failure modes when variations due to the covariate overwhelm those due to the identity. We therefore propose a method of gait recognition via disentangled representation learning that considers both identity and covariate features. Specifically, we first encode an input gait template to get the disentangled identity and covariate features, and then decode the features to simultaneously reconstruct the input gait template and the canonical version of the same subject with no covariates in a semi-supervised manner to ensure successful disentanglement. We finally feed the disentangled identity features into a contrastive/triplet loss function for a verification/identification task. Moreover, we find that new gait templates can be synthesized by transferring the covariate feature from one subject to another. Experimental results on three publicly available gait data sets demonstrate the effectiveness of the proposed method compared with other state-of-the-art methods.

1. Introduction

The gait is an important biometric feature used in human identity recognition at a distance because it can be recorded at a long distance without subject cooperation in contrast with the case for other biometric features (e.g., faces, fingerprints, and irises). Additionally, the gait is an unconscious characteristic and is generally not disguised by people. Gait-based recognition thus has many potential applications, such as surveillance systems, forensics, and criminal investigation [6, 20, 31].

Previous gait recognition studies can be largely categorized according to the extracted features into model-based approaches [47, 53, 27, 5, 10, 54, 2] and appearance-based approaches [15, 32, 44, 24, 48, 4, 37, 29, 59]. The appearance-based approaches are more widely used in the gait recognition community owing to their effectiveness and efficiency. However, they suffer from large intra-subject differences owing to there being many covariates, such as the carrying status, clothing, posture change, and viewing angle.

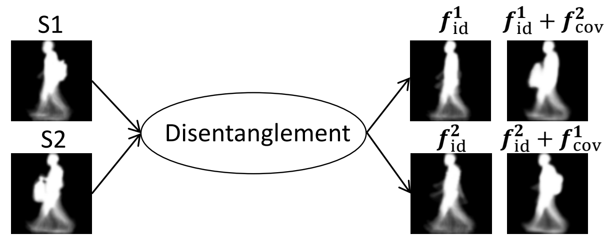

Figure 1. Given two gait templates with different carrying statuses for two subjects ($S1$ and $S2$), our method implicitly disentangles their identity and covariate features ($\boldsymbol{f}_{\mathrm{id}}$ and $\boldsymbol{f}_{\mathrm{cov}}$) and reconstructs gait templates without any carrying statuses using only the identity feature. Moreover, we can swap the subjects' covariate features and generate new gait templates with the carrying statuses of each other.

Appearance-based gait recognition requires the extraction of identity features that are invariant to the covariates. Such invariant approaches fall into two families: discriminative approaches [15, 52, 30, 12, 13, 33, 41, 50, 51, 43, 58, 7, 26] and generative approaches [23, 32, 34, 38, 35, 1, 11, 55, 56, 16]. The former aims at directly extracting an invariant identity feature subspace from the original gait representations, while the latter aims at generating gait representations from different covariate conditions into those under a same covariate condition. However, they all focus on learning identity feature subspaces or image spaces that are invariant to covariates and seldom involve learning features from the covariate aspect, which may lead to failure modes when variations due to a covariate overwhelm those due to the identity.

To remedy the above problem, we propose a method of appearance-based gait recognition using disentangled representation learning (DRL) to consider both identity and covariate features. The idea is inspired by prior work [60] in which pose and appearance features were disentangled from RGB imagery and LSTM-based integration of pose features over time were employed for gait recognition. Although the effect of appearance features such as the color and texture of clothing, which are useless if a subject changes clothes, were successfully eliminated, we argue that the use of RGB imagery still has shortcomings. First, Zhang et al. [60] assumed two conditions to disentangle pose and appearance features; one is that the appearance features are consistent within a sequence while the other is that each training subject should involve at least two sequences with totally different appearance features. These conditions may, however, not always be satisfied, resulting in contamination of the appearance factor into the pose features. As an example, the first condition may be unsatisfied if the illumination condition suddenly changes during a sequence (including the case that the body surface normal changes relative to an incident light direction through limb movement); the second condition may be unsatisfied if the training subjects only change clothes partially or change into clothes of a different color but similar texture. Second, the color and texture information from the RGB inputs were regarded as a type of covariate by [60], which can be easily handled simply using silhouette-based representations as many gait recognition works have done.

We therefore extend the disentanglement idea of [60] to directly disentangle identity and covariate features from silhouette-based gait representations. We divide covariates into two categories that have different effects on the gait representations and may require different disentanglement strategies. Specifically, the first category includes the carrying status and clothing that physically change the body shape of a subject, and this category has a type of clear canonical condition; i.e., a gait template with no covariates. As an example, we regard a gait template without carried objects (COs) and with sufficiently tight clothes (e.g., a subject in tights) as the canonical condition for the carrying status and clothing. The second category includes viewing angles that introduce changes common among all subjects, and this category does not have a clear and suitable canonical condition for all viewing angles. The present paper focuses on the first category. Also for gait representation, we choose the gait energy image (GEI) [15], which is the most widely used gait representation in the gait recognition community.

Specifically, we first use an encoder to disentangle the input GEI into low-dimensional identity and covariate features. We then use a decoder to perform two reconstructions; one is the self-reconstruction of the input GEI from the disentangled identity and covariate features while the other is the reconstruction of another GEI of the same identity as the input GEI but with no covariates (the canonical condition) from the disentangled identity and zero-padded covariate features, where we give the ground-truth GEI with no covariates. With this design, we can successfully disentangle the input GEI into identity and covariate features. Finally, we feed a pair or triplet of identity features into a contrastive/triplet loss for a verification/identification task.

To this end, (1) we use silhouette-based gait representations to avoid unnecessary color and texture covariates; (2) we explore DRL to disentangle the identity and more common but difficult covariates, such as the carrying status and clothing; and (3) we overcome the contamination problem by simultaneously reconstructing the input GEI and its canonical version in a semi-supervised manner (i.e., we give covariate labels to the GEIs with no covariates (e.g., "no covariate") but not to other GEIs with a covariate in the training stage).

Moreover, we find that given a disentangled identity feature from one subject $S1$ and a disentangled covariate feature from another subject $S2$, the decoder can reconstruct a new GEI sample that has the same identity characterized from $S1$ and the same covariate characterized from $S2$, as shown in Fig. 1. We can thus transfer covariate characteristics freely from one subject to another to generate new GEIs, which we call GEI editing.

We summarize our contributions as follows.

1) An identity and covariate feature-based disentanglement network (ICDNet) for gait recognition.

We introduce semi-supervised DRL to disentangle identity and covariate features for gait recognition for the first time. After disentanglement, the identity features are pure and discriminative for gait recognition.

2) GEI editing: covariate transfer from one subject to another.

We can generate new GEIs by transferring the disentangled covariate feature from one subject to another. This might be beneficial for future research on data augmentation in gait recognition.

3) State-of-the-art performance.

We achieved state-of-the-art performance on three publicly available gait databases: the OU-ISIR Large Population Gait database with real-life COs (OU-LP-Bag) [46], the OU-ISIR Gait database, Large Population data set with bag β version (OU-LP-Bag β) [33], and CASIA-B gait database [57].

2. Related work

Appearance-based gait recognition approaches

Appearance-based gait recognition approaches are mainly divided into discriminative and generative approaches. The first category aims at extracting a discriminative subspace against covariates using traditional metric

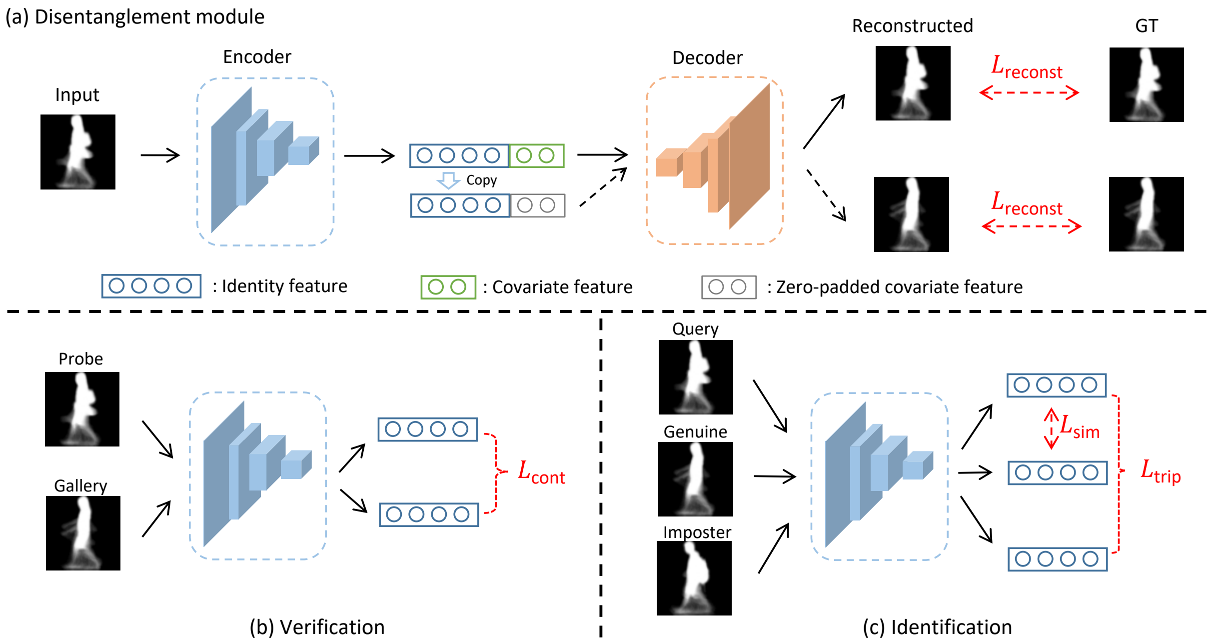

Figure 2. Overview of the proposed **ICDNet**. (a) The disentanglement module uses an encoder to disentangle latent identity and covariate features from an input GEI and a decoder to perform two reconstructions; one is self-reconstruction of the original input GEI (represented by the solid line) while the other is the reconstruction of the input GEI without any covariates (represented by the dotted line). (b) Verification scenario: a pair (probe and gallery) first passes through the encoder of the disentanglement module (a), and the disentangled identity features are then used for the verification loss. (c) Identification scenario: a triplet (query, genuine, and imposter) first passes through the encoder of the disentanglement module (a), and the disentangled identity features are then used for the identification loss.

learning techniques (e.g., linear discriminant analysis [15], discriminant analysis with tensor representation [52], the random subspace method [13], and joint intensity and spatial metric learning [33]) or current deep neural networks. Particularly, deep learning-based approaches are more popular because of their much higher performance. For example, Shiraga et al. [41] proposed a light convolutional neural network (CNN) for classifying single-input GEIs with a cross-entropy loss. Thereafter, several studies [51, 43, 26] conducted similarity learning for a pair or triplet of input GEIs with contrastive or triplet losses. Instead of using GEIs as network inputs, Wolf et al. [50] and Chao et al. [7] directly designed CNNs for silhouette frames.

The second category aims at generating gait representations from different covariate conditions into those under a same covariate condition using subspace analysis techniques [23, 32, 34, 38, 35, 1] or generative adversarial networks (GANs) [55, 56, 16]. For example, Makihara et al. [32] proposed a view transformation model to transform gait features from a gallery view condition to a probe view condition. Yu et al. [55, 56] proposed GAN-based generation networks named GaitGAN and GaitGANv2, which generate invariant GEIs of the side view in the normal status from any input GEIs with covariates.

However, the above approaches focus on learning identity feature subspaces or image spaces invariant to covariates and seldom consider learning features from the covariate aspect, which may lead to failure modes when variations due to a covariate overwhelm those due to the identity. In contrast, our method considers both identity and covariate feature learning and achieves obvious gains for the identity feature with the ablation of the covariate feature.

Disentangled representation learning

DRL is expected to provide gains by separating the underlying structure of data into disjoint meaningful variables, which helps clarify the deep models and determine what types of hidden features are actually learned. Zhang et al. [60] introduced DRL to the field of gait recognition for the first time, where pose and appearance features of subjects were disentangled from RGB imagery. Although DRL is new in studies on gait, it has been well explored in studies on other biometrics (i.e., face recognition). For example, Tran et al. [45] and Peng et al. [40] disentangled pose variation from face images for pose-invariant face recognition, while Zhao et al. [61] generated age-invariant face features through the disentanglement of age variation.

In contrast with [60], our method avoids the difficulty of the disentanglement of RGB information and gives new meaningful disentangled variables (i.e., identity and covariate features) for silhouette-based gait representation. Compared with DRL in face recognition [45, 40, 61], which requires additional covariate labels (e.g., pose or age labels), our method is designed for no explicit labels (except for "no covariate" labels) because the covariates (i.e., the carrying status and clothing) that we target in our paper do not have clear labels. For example, even for the same carry bags, the carrying status (e.g., shape and location) largely depend on the subject. Only the label of the canonical condition (i.e., "no covariate") for partial training subjects is needed in our semi-supervised DRL.

3. Proposed method

3.1. Overview

We propose a method of gait recognition that applies DRL to disentangle identity and covariate features from GEIs[1]. In our problem setting, we assume that each training subject has a gait template without covariates (e.g., a GEI without COs), while we do not have labels of the covariate conditions for the other gait templates (e.g., a subject may carry a backpack, briefcase, suitcase or even nothing, but we never know it in advance, which also provides a test case). We therefore try making the most of the partial labels "no covariate" for better disentanglement.

An overview of the proposed **ICDNet** is shown in Fig. 2. A basic disentanglement module, which has one encoding and two decoding streams, processes all GEIs in the training set. A pair (probe P and gallery G) or triplet (query Q, genuine G, and imposter I) of identity features is then fed into verification or identification loss functions for verification or identification training, separately. In a test case, only the encoder of the disentanglement module is used to disentangle the identity and covariate features for each input GEI. The Euclidean distance between two subjects' identity features is computed as the dissimilarity score. We finally judge whether subjects are the same or different by comparing the score with an acceptance threshold for the verification scenario (one-to-one matching) or find the smallest dissimilarity score among galleries for the identification scenario (one-to-many matching).

3.2. Disentanglement module

The disentanglement module is a vital component of **ICDNet**. As shown in Fig. 2 (a), the module has an encoder E and a decoder D. The encoder E has one encoding stream that receives an input GEI X and outputs the latent identity feature $\boldsymbol{f}_{\mathrm{id}}$ and covariate feature $\boldsymbol{f}_{\mathrm{cov}}$, which can be expressed as

$$[\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_{\mathrm{cov}}] = E(X). \tag{1}$$

Meanwhile, the decoder D has two decoding streams. One receives concatenated identity and covariate features $[\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_{\mathrm{cov}}]$ with which to reconstruct the input GEI $\hat{X}$ itself. The other receives concatenated identity and zero-padded covariate features $[\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_0]$ with which to reconstruct a GEI $\hat{X}_0$ of the same training subject as in the input GEI and also without covariates. Intuitively speaking, through zero-padding of the covariate feature vector, we can knock out the covariate feature or make it invalid to ensure that the covariate factor never contaminates the identity factor when reconstructing the GEI $\hat{X}_0$ without covariates; i.e., a sort of purified GEI that only contains the identify factor. The two reconstructed GEIs can be expressed as

$$\begin{aligned} \hat{X} &= D([\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_{\mathrm{cov}}]) \\ \hat{X}_0 &= D([\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_0]), \end{aligned} \tag{2}$$

where $\boldsymbol{f}_0 = \boldsymbol{0}$ is a zero-padded feature with the same dimensions as $\boldsymbol{f}_{\mathrm{cov}}$. Note that it does not matter whether the input GEI X actually involves a covariate; i.e., the network just tries outputting the GEI without covariates for both outputs if the input GEI does not involve a covariate.

The two reconstructed GEIs $\hat{X}$ and $\hat{X}_0$ are supposed to be similar to their corresponding ground truth GEIs X (the input GEI) and X_0 (the GEI of the same identity as X but without covariates). To achieve this, we define the reconstruction loss as

$$L_{\mathrm{reconst}}(E, D) = \|X - \hat{X}\|_2^2 + \|X_0 - \hat{X}_0\|_2^2. \tag{3}$$

By minimizing L_{reconst}, we ensure that the disentangled $\boldsymbol{f}_{\mathrm{id}}$ and $\boldsymbol{f}_{\mathrm{cov}}$ only contain the identity and covariate information of the input GEI, respectively, and that the predefined zero-padded $\boldsymbol{f}_0$ indicates there is no covariates. In this semi-supervised manner, we ensure the disentanglement property of the proposed method.

3.3. Gait recognition using an identity feature

We use disentangled identity features for gait recognition. There are generally two types of biometric recognition task: verification and identification. The verification task (i.e., one-to-one matching) aims at judging whether a given pair of probe and gallery are from the same subject. The identification task (i.e., one-to-many matching) aims at finding a correct match from multiple enrolled galleries given a probe (i.e., a query). A previous study [43] presented a detailed discussion on suitable network architectures and loss functions for the two different biometric recognition tasks. We therefore design two different networks and loss functions for the two tasks as follows.

Verification task. (see Fig. 2 (b)) We first prepare disentangled identity features from a pair of GEIs (P, G) and its corresponding binary label y (where values of 1 and 0 mean that the pair is from the same and different subjects, respectively); we then feed the features into a contrastive loss function [14].

[1]The reader may refer to [15] for details of how to extract a GEI from a silhouette sequence.

Suppose there are N pairs of identity features $\{M_i|M_i = (\boldsymbol{f}_{\text{id}}^{P_i}, \boldsymbol{f}_{\text{id}}^{G_i}), i = 1, 2, ..., N\}$ and their corresponding labels $\{y_i|y_i = \{0, 1\}, i = 1, 2, ..., N\}$. We define a contrastive loss function as

$$L_{\text{cont}}(E) = \frac{1}{N}\sum_{i=1}^{N}\{y_i d_i + (1 - y_i)\max(m - d_i, 0)\}, \tag{4}$$

where $d_i = \|\boldsymbol{f}_{\text{id}}^{P_i} - \boldsymbol{f}_{\text{id}}^{G_i}\|_2^2$ is the dissimilarity score of the given pair (P_i and G_i) and m is a margin. We force the identity features of P and G closer together if they are from the same subject pair and further away if they are from different subject pairs by minimizing L_{cont}, which is more suitable for the verification task than for the identification task.

Identification task. (see Fig. 2 (c)) We first prepare disentangled identity features from a triplet of GEIs (Q, G, I), where Q and G are from the same subject while Q and I are from two different subjects. We then feed the features into a triplet loss function [49].

Suppose there are N triplets of identity features $\{T_i|T_i = (\boldsymbol{f}_{\text{id}}^{Q_i}, \boldsymbol{f}_{\text{id}}^{G_i}, \boldsymbol{f}_{\text{id}}^{I_i}), i = 1, 2, ..., N\}$. We define a triplet loss function as

$$L_{\text{trip}}(E) = \frac{1}{N}\sum_{i=1}^{N}\max(m - d_i^- + d_i^+, 0), \tag{5}$$

where $d_i^+ = \|\boldsymbol{f}_{\text{id}}^{Q_i} - \boldsymbol{f}_{\text{id}}^{G_i}\|_2^2$ is the dissimilarity score of the same subject pair (Q_i and G_i) and $d_i^- = \|\boldsymbol{f}_{\text{id}}^{Q_i} - \boldsymbol{f}_{\text{id}}^{I_i}\|_2^2$ is the dissimilarity score of the different subject pair (Q_i and I_i), and m is a margin. We force the identity features of Q and G closer than those of the same Q and other I by minimizing L_{trip}, which is more suitable for the identification task than for the verification task.

L_{trip} only restricts the relative distance between the same subject and different subject pairs and thus does not force the distances of the same subject pairs to be absolutely close to each other. Considering the disentanglement property that the disentangled identity features from the same subject should be similar, we define another loss function referred to as the identity similarity loss L_{sim} to force the identity features of the same subject pair (Q and G) to be close to each other:

$$L_{\text{sim}}(E) = \frac{1}{N}\sum_{i=1}^{N} d_i^+. \tag{6}$$

Sampling of pairs and triplets. We employ batch all sampling [17] for both contrastive and triplet losses. For each batch, we first randomly choose P subjects and K samples per subject. There are thus a total of PK samples in a batch. We then select all combinations of pairs and triplets in this batch, resulting in $PK(PK - 1)$ pairs and $PK(PK - K)(K - 1)$ triplets. Considering the severe imbalance between the number of same subject pairs ($N_s = PK(K - 1)$) and the number of different subject pairs ($N_d = PK(PK - K)$), we modify Eq. (4) so as to normalize the losses for the same and different subject pairs as

$$L_{\text{cont}}(E) = \frac{1}{N_s}\sum_{i=1}^{N} y_i d_i + \frac{1}{N_d}\sum_{i=1}^{N}(1 - y_i)\max(m - d_i, 0). \tag{7}$$

3.4. Joint loss functions

Considering both disentanglement and recognition aspects, we define joint loss functions by the weighted summation of the aforementioned loss functions and train them in an end-to-end manner.

Specifically, for the verification task, the joint loss function is defined as

$$L(E, D) = \lambda_{\text{reconst}} L_{\text{reconst}}(E, D) + \lambda_{\text{cont}} L_{\text{cont}}(E), \tag{8}$$

where λ_{reconst} and λ_{cont} are two hyper-parameters.

For the identification task, the joint loss function is defined as

$$\begin{aligned} L(E, D) = \lambda_{\text{reconst}} L_{\text{reconst}}(E, D) + \lambda_{\text{trip}} L_{\text{trip}}(E) \\ + \lambda_{\text{sim}} L_{\text{sim}}(E), \end{aligned} \tag{9}$$

where λ_{reconst}, λ_{trip}, and λ_{sim} are three hyper-parameters.

Finally, the parameters of E and D are optimized by minimizing the joint loss function $L(E, D)$.

4. Experiment

4.1. Data sets

We evaluate the proposed method on three publicly available gait databases: OU-LP-Bag [46], OU-LP-Bag β [33], and CASIA-B [57].

OU-LP-Bag has the largest number of subjects (62,528 subjects) of any gait database available worldwide and contains the covariate of real-life COs. Following the same protocol as [46], the training set contains 29,097 subjects with two sequences with and without COs, and the test set contains other 29,102 disjoint subjects. There are two versions of probe and gallery sets prepared in the test set under cooperative and uncooperative settings. For the cooperative setting, the gallery set only contains sequences without COs whereas the probe set contains sequences with seven types of COs annotated by carrying locations; for the uncooperative setting, gallery and probe sets randomly swap sequences and both have sequences with and without COs.

OU-LP-Bag β contains 4,140 sequences of 2,070 subjects with and without COs. The training set contains 1,034 subjects while the test set contains the other 1,036 disjoint subjects. The gallery set contains sequences without COs whereas the probe set contains sequences with COs.

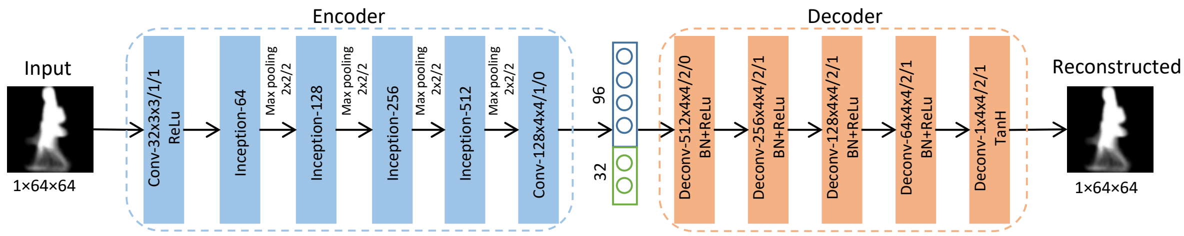

Figure 3. Detailed architectures of the encoder and decoder. The digits after "Conv" and "Deconv" indicate $\#channels \times \#kernel_h \times \#kernel_w / \#stride / \#padding$; "Inception-$N$" indicates an inception module [42] with the output channel number N.

CASIA-B contains 124 subjects from 11 views. There are 10 sequences per subject and view. Among them, six are of normal walking (NM), two are of carrying a bag (BG), and the remaining two are of wearing a coat (CL). Following [51], the first four sequences under NM are chosen as the gallery (NM #1–4). The other six sequences are used as three probe sets: (1) Set-NM contains two NM sequences (NM #5–6), (2) Set-BG contains two BG sequences (BG #1–2), and (3) Set-CL contains two CL sequences (CL #1–2).

4.2. Implementation details

Network architectures. The detailed architectures of the encoder and decoder are shown in Fig. 3. The encoder takes an input GEI of size $1 \times 64 \times 64$ and outputs latent identity and covariate features, which are experimentally set as 96 and 32 dimensional vectors, respectively. The backbone of the encoder is designed on the basis of the Inception module in GoogLeNet [42] to extract features at multiple scales. The decoder takes latent identity and covariate features as input and outputs the reconstructed GEI, which is designed using deconvolutional (transposed convolutional) layers.

Training strategies. We employ two training strategies: one is training a model from scratch for each data set while the other is pre-training a model on the largest data set (i.e., OU-LP-Bag) and then fine-tuning the pre-trained model on the other two smaller data sets, which takes advantage of better generalization capabilities on the largest data set. For the first strategy, we apply an additional data augmentation (i.e., translation from -5 to 5 pixels with step of 2 for both vertical and horizontal axes) on OU-LP-Bag β and CASIA-B considering their relatively small number of samples. For the second strategy, we only use the original data sets.

Parameter settings. We train the proposed **ICDNet** in an end-to-end manner using the Adam optimizer [21]. For the train-from-scratch strategy, the initial learning rate is set to 0.0002 and the momentum term (β_1, β_2) is set to $(0.5, 0.999)$. After 100,000 iterations, we decrease the learning rate to 0.00002 and run for 50,000 further iterations. For the fine-tuning strategy, the models at 150,000 iterations on OU-LP-Bag are used for initialization. We set the initial learning rate as 0.00002 and only run 10,000 iterations. The batch all sampling parameters (P, K) are set

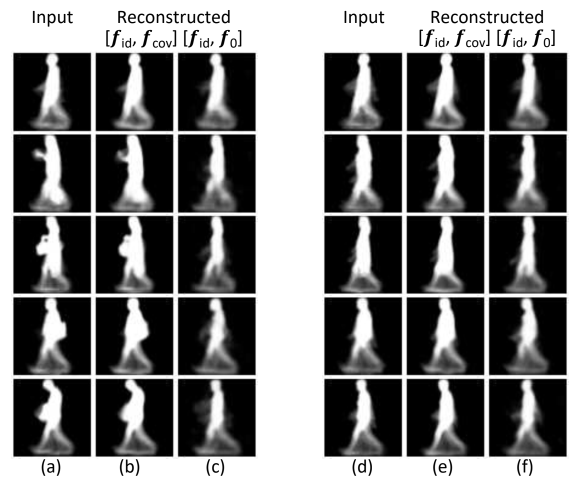

Figure 4. Self-reconstruction examples. Each row is for the same subject. Left and right sides show the results of the input GEIs with and without COs, respectively. Columns (a) and (d) are the input GEIs; columns (b) and (e) are reconstructed by $[\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_{\mathrm{cov}}]$; columns (c) and (f) are reconstructed by $[\boldsymbol{f}_{\mathrm{id}}, \boldsymbol{f}_0]$.

to $(300, 2)$, $(100, 2)$, and $(8, 16)$ for OU-LP-Bag, OU-LP-Bag β, and CASIA-B, respectively. The margin m in Eqs. (5)(7) is set to 3. The weight parameters for the joint loss functions are set as $\lambda_{\mathrm{reconst}} = 100$ and $\lambda_{\mathrm{cont}} = 1$ in Eq. (8) and $\lambda_{\mathrm{reconst}} = 1000$, $\lambda_{\mathrm{trip}} = 1$, and $\lambda_{\mathrm{sim}} = 0.1$ in Eq. (9) for all data sets, except that λ_{sim} is set to 0.0001 for CASIA-B.

4.3. Evaluation metrics

According to the biometrics performance standard [19], for the verification task, we report the equal error rate (EER) of a false match rate (FMR) and a false non-match rate (FNMR), and a detection error trade-off (DET) curve that describes the trade-off between the FNMR and FMR when an acceptance threshold changes. For the identification task, we report the rank-1 identification rate (denoted by Rank-1) and a cumulative match characteristic (CMC) curve that describes identification rates within each of the ranks.

4.4. Visualization of reconstructed GEIs

We qualitatively evaluate the proposed method by visualizing the reconstruction results on OU-LP-Bag. We first show several self-reconstruction examples in Fig. 4. It is

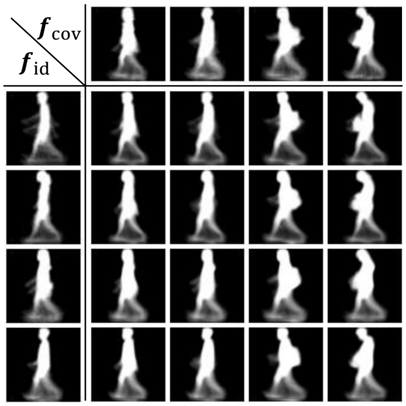

Figure 5. Cross-reconstruction examples for two different subjects (GEI editing). In each reconstruction example, the identity feature $\boldsymbol{f}_{\text{id}}$ is from the left-most subject and the covariate feature $\boldsymbol{f}_{\text{cov}}$ is from the top-most subject.

clear that we can successfully reconstruct the input GEIs from the encoded feature $[\boldsymbol{f}_{\text{id}}, \boldsymbol{f}_{\text{cov}}]$ no matter whether the input GEIs have COs (see Fig. 4 (b) and (e)). We also successfully reconstruct GEIs of the same subject without COs (see Fig. 4 (c) and (f)), which are similar to the input GEIs without COs (see Fig. 4 (d)), and the reconstruction results in Fig. 4 (c) and (f) are similar to each other, which implies similar identity features $\boldsymbol{f}_{\text{id}}$ are obtained for the same subject no matter whether the input GEIs have COs. Moreover, we find that not only the COs themselves but also some posture changes (e.g., hand raising and body bending) induced by the carrying status can be regarded as covariate features, which is evident from the fact that they are also eliminated in the reconstructed GEIs without COs (see Fig. 4 (c)).

We next combine the identity and covariate features from two different subjects and see if the covariate feature from one subject can be transferred to the other subject. Results are shown in Fig. 5. In line with our expectations, the reconstructed GEI samples share similar identity features for each row and similar covariate features for each column. Moreover, we further confirm that the transferred covariate features contain many covariates, including COs, posture, and clothes.

Through the evaluation, we verify that the proposed method can disentangle identity and covariate features.

4.5. Comparison with state-of-the-art approaches

OU-LP-Bag. We evaluate benchmarks reported in the original database study [46] and a current state-of-the-art method [26] as well as the proposed method by following the original experimental protocols in [46]. All results are presented in Fig. 6 and Table 1. For each of the cooperative/uncooperative settings and recognition tasks, we match the state-of-the-art performance and outperform the second-best benchmark [26] by a large margin (e.g., an EER that is more than 0.3 % lower and a Rank-1 rate that is more than 12 % higher), which shows the advantages of our method in terms of recognition performance.

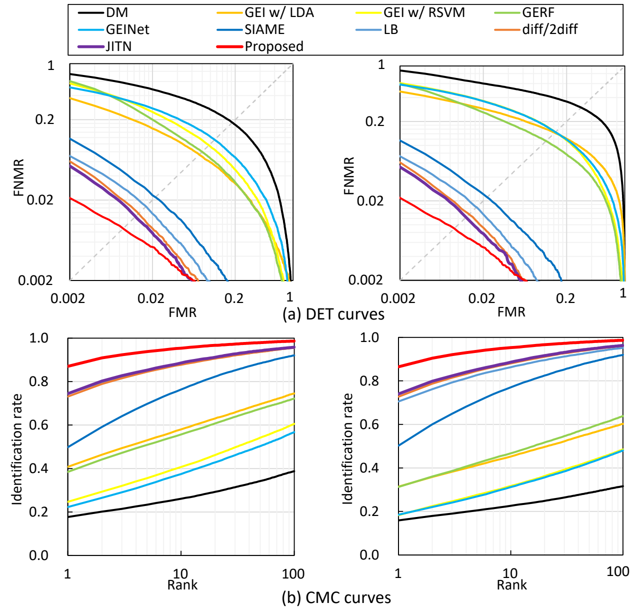

Figure 6. DET and CMC curves on OU-LP-Bag. The left/right side shows results for the cooperative/uncooperative setting.

Table 1. EERs and Rank-1 [%] on OU-LP-Bag under cooperative and uncooperative settings. Models are trained from scratch. Bold and italic bold fonts respectively indicate the best and second-best results. This convention is used consistently throughout the paper.

Methods	Cooperative		Uncooperative	
	EER	Rank-1	EER	Rank-1
DM [15]	18.46	17.74	29.89	15.90
GEI w/ LDA [39]	7.35	40.79	14.40	31.44
GEI w/ RSVM [36]	9.58	24.66	14.69	18.28
GERF [25]	7.97	38.48	11.35	31.24
GEINet [41]	11.29	22.26	14.68	18.52
SIAME [9]	2.17	49.80	2.22	50.27
LB [51]	1.68	74.39	1.66	70.53
diff/2diff [43]	1.36	73.14	1.35	72.75
JITN [26]	***1.25***	***74.44***	***1.25***	***74.03***
Proposed	**0.89**	**87.04**	**0.90**	**86.49**

OU-LP-Bag β. Existing methods adopt two different training strategies for the data set. As mentioned in section 4.2, one is training from scratch as adopted in [33, 58] while the other is fine-tuning on a pre-trained model on a larger data set (i.e., OU-LP-Bag), which was first adopted in [26] considering the relatively small number of subjects in OU-LP-Bag β. For fair comparison, we compare our method with other benchmarks using each strategy accordingly. All results are presented in Fig. 7 and Table 2. The results show that our method performs better for both strategies.

CASIA-B. We focus the experiments of CASIA-B on the side view (or nearly side view) because our method currently does not aim at the viewing angle covariate. We use two protocols for these experiments. Protocol 1 is taken from [8], whereby the first 24 subjects are taken as the train-

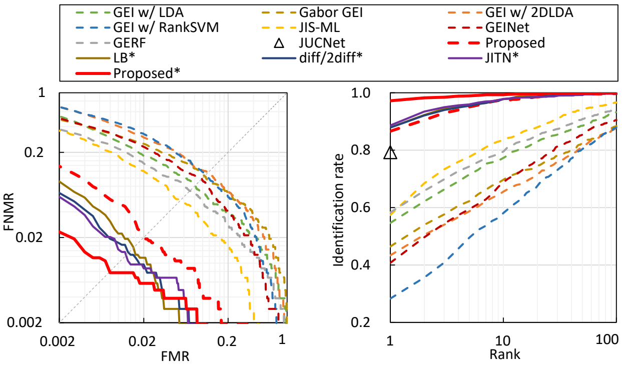

Figure 7. DET and CMC curves on OU-LP-Bag β.

Table 2. EERs and Rank-1 [%] on OU-LP-Bag β. "-" indicates not provided. "*" indicates models are fine-tuned from the pre-trained models on OU-LP-Bag. This convention is used consistently throughout the paper.

Methods	EER	Rank-1
Gabor GEI [44]	10.48	46.4
GEI w/ LDA [39]	8.10	54.6
GEI w/ 2DLDA [28]	11.47	43.3
GEI w/ RSVM [36]	10.81	28.3
GERF [25]	6.67	58.3
JIS-ML [33]	***5.45***	57.4
GEINet [41]	9.75	40.7
JUCNet [58]	-	***79.3***
Proposed	**2.03**	**86.6**
LB* [51]	1.53	87.9
diff/2diff* [43]	1.31	87.8
JITN* [26]	***1.27***	***88.1***
Proposed*	**0.77**	**97.2**

Table 3. Rank-1 [%] on CASIA-B for protocol 1.

Methods	Set-NM	Set-BG	Set-CL	Mean
GEI [15]	***99***	60	30	63.0
GEnI [3]	98.3	80.1	33.5	70.6
STIP+NN [22]	95.4	60.9	52	69.4
GEINet [41]	97.5	84.5	71.8	84.6
L-CRF [8]	98.6	***90.2***	***85.8***	***91.5***
Proposed	**100**	82.0	73.0	85.0
Proposed *	**100**	**100**	**93.0**	**97.7**

ing set while the remaining 100 subjects are taken as the test set. Only the side-view angle is used for this protocol. Protocol 2 is taken from [18, 51, 8, 58, 60] considering both walking conditions (BG or CL) and viewing angle changes. The first 34 subjects are used as the training set and the remaining 90 subjects as the test set. In our case, we limit the viewing angle changes to the nearly side view angles (probe vs. gallery: 90° vs. 72° and 90° vs. 108°).

Tables 3 and 4 give results for protocol 1 and 2, respectively. Only Rank-1 rate is reported because EERs are reported for hardly any benchmarks. The results show that our method performs worse under the train-from-scratch strategy. We argue that this may be because the number of subjects in the training set for protocols 1 and 2 is limited for CASIA-B compared with the other two data sets, resulting

Table 4. Rank-1 [%] on CASIA-B for protocol 2.

Methods	(90°, 72°)		(90°, 108°)		Mean	
	BG	CL	BG	CL	BG	CL
RLTDA [18]	75.3	63.2	76.5	72.1	75.9	67.7
LB [51]	93.3	78.3	88.9	75.6	91.1	77.0
L-CRF [8]	94.4	88.5	89.2	85.7	91.8	87.1
JUCNet [58]	***95.9***	-	***95.9***	-	***95.9***	-
GaitNet [60]	95.6	***94.2***	87.4	***86.5***	91.5	***90.4***
Proposed	90.0	76.7	87.8	66.7	88.9	71.7
Proposed*	**100**	**95.6**	**100**	**93.3**	**100**	**94.5**

Table 5. Ablation study of loss functions in terms of EERs and Rank-1 [%] on OU-LP-Bag under a cooperative setting. "N/A" indicates not applicable.

Recognition loss	Disentanglement loss	EER	Rank-1
L_{cont}	-	***0.98***	N/A
L_{cont}	L_{reconst}	**0.89**	N/A
L_{trip}	-	N/A	84.02
L_{trip}	L_{reconst}	N/A	***86.37***
L_{trip}	$L_{\text{reconst}} + L_{\text{sim}}$	N/A	**87.04**

in a lack of generalization capability for our method. However, we achieve the best performance once we apply the fine-tuning strategy.

4.6. Ablation study of loss functions

We analyze how each loss function affects the performance of our method on OU-LP-Bag. While keeping the recognition loss (L_{cont} or L_{trip}), we add or remove the disentanglement loss (L_{reconst} and L_{sim}) to evaluate their respective performance. Table 5 shows that adding the disentanglement loss improves performance, which indicates the effectiveness of our disentanglement method.

5. Conclusion

We proposed a method of gait recognition named **ICDNet**, which applies semi-supervised DRL to disentangle identity and covariate features. We designed an auto-encoder that encodes an input GEI into identity and covariate features and reconstructs the input GEI and that of the same subject without covariates using partial labels on the covariate. We presented qualitative and quantitative evaluations to show the successful disentanglement of identity and covariate features and the improvement in performance with disentanglement. We also confirmed the proposed method makes cross-reconstruction possible, which shows the potential of gait data augmentation in future work. Moreover, because we currently excluded the viewing angle, it will be another future work to design a more comprehensive network that handles all covariates.

Acknowledgment This work was supported by JSPS Grants-in-Aid for Scientific Research (A) JP18H04115, JSPS Grant-in-Aid for Scientific Research on Innovative Areas 19H05692, and the National Natural Science Foundation of China (Grant No. 61727802).

References

[1] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi. Video from nearly still: an application to low frame-rate gait recognition. In *Proc. of the 25th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, pages 1537–1543, Providence, RI, USA, Jun. 2012. 1, 3

[2] G. Ariyanto and M. Nixon. Marionette mass-spring model for 3d gait biometrics. In *Proc. of the 5th IAPR International Conference on Biometrics*, pages 354–359, March 2012. 1

[3] K. Bashir, T. Xiang, and S. Gong. Gait recognition using gait entropy image. In *Proc. of the 3rd Int. Conf. on Imaging for Crime Detection and Prevention*, pages 1–6, Dec. 2009. 8

[4] K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, Oct. 2010. 1

[5] A. Bobick and A. Johnson. Gait recognition using static activity-specific parameters. In *Proc. of the 14th IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 423–430, 2001. 1

[6] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011. 1

[7] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI Conference on Artificial Intelligence*, 2019. 1, 3

[8] X. Chen, J. Weng, W. Lu, and J. Xu. Multi-gait recognition based on attribute discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2018. 7, 8

[9] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546, June 2005. 7

[10] D. Cunado, M. Nixon, and J. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003. 1

[11] B. Decann and A. Ross. Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In *Proc. of the SPIE, Biometric Technology for Human Identification VII*, volume 7667, pages 1–13, 2010. 1

[12] B. DeCann, A. Ross, and M. Culp. On clustering human gait patterns. In *2014 22nd International Conference on Pattern Recognition*, pages 1794–1799, Aug 2014. 1

[13] Y. Guan, C. T. Li, and F. Roli. On reducing the effect of covariate factors in gait recognition: A classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1521–1528, July 2015. 1, 3

[14] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *The IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. 4

[15] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, 2006. 1, 2, 3, 4, 7, 8

[16] Y. He, J. Zhang, H. Shan, and L. Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, Jan 2019. 1, 3

[17] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 5

[18] H. Hu. Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7):1274–1286, July 2013. 8

[19] ISO/IEC 19795-1:2006(en). Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. Technical report, International Organization for Standardization, ISO/IEC JTC 1/SC 37, Geneva, Switzerland, 2006. 6

[20] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi. Gait verification system for criminal investigation. *IPSJ Transactions on Computer Vision and Applications*, 5:163–175, Oct. 2013. 1

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv: 1412.6980 (2014), 2014. 6

[22] W. Kusakunniran. Recognizing gaits on spatio-temporal feature domain. *IEEE Transactions on Information Forensics and Security*, 9(9):1416–1423, Sept 2014. 8

[23] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Support vector regression for multi-view gait recognition based on local motion feature selection. In *Proc. of IEEE computer society conference on Computer Vision and Pattern Recognition 2010*, pages 1–8, San Francisco, CA, USA, Jun. 2010. 1, 3

[24] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognition*, 44:973–987, April 2011. 1

[25] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren. Gait energy response function for clothing-invariant gait recognition. In *Asian Conference on Computer Vision*, pages 257–272, 2016. 7, 8

[26] 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 Transactions on Information Forensics and Security*, pages 1–1, 2019. 1, 3, 7, 8

[27] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Proceedings of the 12th Chinese Conference on Biometric Recognition*, pages 474–483, 2017. 1

[28] K. Liu, Y. Cheng, and J. Yang. Algebraic feature extraction. *IEEE Trans. Circuits Syst. Video Technol*, 26(6):903–911, 2006. 8

[29] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006. 1

[30] J. Lu and Y.-P. Tan. Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. *Pattern Recognition Letters*, 31(5):382–393, 2010. 1

[31] N. Lynnerup and P. Larsen. Gait as evidence. *IET Biometrics*, 3(2):47–54, 6 2014. 1

[32] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *Proc. of the 9th European Conference on Computer Vision*, pages 151–163, Graz, Austria, May 2006. 1, 3

[33] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *Proc. of the 30th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 5705–5715, Jul. 2017. 1, 2, 3, 5, 7, 8

[34] Y. Makihara, A. Tsuji, and Y. Yagi. Silhouette transformation based on walking speed for gait identification. In *Proc. of the 23rd IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun 2010. 1, 3

[35] A. Mansur, Y. Makihara, R. Aqmar, and Y. Yagi. Gait recognition under speed transition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2521–2528, June 2014. 1, 3

[36] R. Martin-Felez and T. Xiang. Uncooperative gait recognition by learning to rank. *Pattern Recognition*, 47(12):3793 – 3806, 2014. 7, 8

[37] S. Mowbray and M. Nixon. Automatic gait recognition via fourier descriptors of deformable objects. In *Proc. of the 1st IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 566–573, 2003. 1

[38] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, and Y. Yagi. Gait-based person recognition using arbitrary view transformation model. *IEEE Trans. on Image Processing*, 24(1):140–154, Jan 2015. 1, 3

[39] N. Otsu. Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In *Proc. of the 6th Int. Conf. on Pattern Recognition*, pages 557–560, 1982. 7, 8

[40] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[41] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *Proc. of the 8th IAPR Int. Conf. on Biometrics (ICB 2016)*, number O19, pages 1–8, Halmstad, Sweden, Jun. 2016. 1, 3, 7, 8

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 6

[43] 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 Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2018. 1, 3, 4, 7, 8

[44] D. Tao, X. Li, X. Wu, and S. J. Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1700–1715, Oct 2007. 1, 8

[45] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017. 3

[46] M. Z. Uddin, T. T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, and Y. Yagi. The ou-isir large population gait database with real-life carried object and its performance evaluation. *IPSJ Transactions on Computer Vision and Applications*, 10(1):5, May 2018. 2, 5, 7

[47] D. Wagg and M. Nixon. On automated model-based extraction and analysis of gait. In *Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 11–16, 2004. 1

[48] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan. Human identification using temporal information preserving gait template. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2164 –2176, nov. 2012. 1

[49] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 5

[50] T. Wolf, M. Babaee, and G. Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 4165–4169, Sept 2016. 1, 3

[51] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, Feb 2017. 1, 3, 6, 7, 8

[52] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H. jiang Zhang. Human gait recognition with matrix representation. *IEEE Trans. Circuits Syst. Video Technol*, 16(7):896–903, 2006. 1, 3

[53] C. Yam, M. Nixon, and J. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, 2004. 1

[54] K. Yamauchi, B. Bhanu, and H. Saito. 3d human body modeling using range data. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3476–3479, Aug 2010. 1

[55] S. Yu, H. Chen, E. B. Garcia Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 1, 3

[56] S. Yu, R. Liao, W. An, H. Chen, E. B. Garcia, Y. Huang, and N. Poh. Gaitganv2: Invariant gait feature extraction using generative adversarial networks. *Pattern Recognition*, 87:179 – 189, 2019. 1, 3

[57] 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. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006. 2, 5

[58] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li. Learning joint gait representation via quintuplet loss minimization. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2019. 1, 7, 8

[59] Y. Zhang, Y. Huang, L. Wang, and S. Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 93:228 – 236, 2019. 1

[60] Z. Zhang, L. Tran, X. Yin, Y. Atoum, J. Wan, N. Wang, and X. Liu. Gait recognition via disentangled representation learning. In *Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019. 2, 3, 8

[61] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, H. Lan, F. Zhao, L. Xiong, Y. Xu, J. Li, S. Pranata, et al. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *AAAI Conference on Artificial Intelligence*, 2019. 3