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

Xiang Li\textsuperscript{1,2} Yasushi Makihara\textsuperscript{2} Chi Xu\textsuperscript{1,2} Yasushi Yagi\textsuperscript{2} Mingwu Ren\textsuperscript{1}
\textsuperscript{1}Nanjing University of Science and Technology, \textsuperscript{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.

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 covariates 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. 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 task.

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 \( \hat{f}_{id} \) and covariate feature \( \hat{f}_{cov} \), which can be expressed as

\[
[f_{id}, f_{cov}] = E(X) \tag{1}
\]

Meanwhile, the decoder D has two decoding streams. One receives concatenated identity and covariate features \( [f_{id}, f_{cov}] \) with which to reconstruct the input GEI \( \hat{X} \) itself. The other receives concatenated identity and zero-padded covariate features \( [f_{id}, f_{0}] \) 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 identity factor. The two reconstructed GEIs can be expressed as

\[
\hat{X} = D([f_{id}, f_{cov}]) \tag{2}
\]

\[
\hat{X}_0 = D([f_{id}, f_{0}]),
\]

where \( f_0 = \theta \) is a zero-padded feature with the same dimensions as \( f_{cov} \). 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

\[
L_{\text{reconst}}(E, D) = \| X - \hat{X} \|^2_2 + \| X_0 - \hat{X}_0 \|^2_2 \tag{3}
\]

By minimizing \( L_{\text{reconst}} \), we ensure that the disentangled \( f_{id} \) and \( f_{cov} \) only contain the identity and covariate information of the input GEI, respectively, and that the predefined zero-padded \( f_0 \) 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 \( \{M_i | M_i = \{f_i^{id}, f_i^{G_i}\}, i = 1, 2, ..., N\} \) and their corresponding labels \( \{y_i | y_i = \{0, 1\}, i = 1, 2, ..., N\} \). We define a contrastive loss function as

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

where \( d_i = \|f_i^{P} - f_i^{G_i}\|_2 \) is the dissimilarity score of the given pair \((P_i, G_i)\) and \( m \) is a margin. We force the identity features of \( P_i \) and \( G_i \) closer together if they are from the same subject pair and further away if they are from different subject pairs by minimizing \( L_{\text{cont}} \), 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 \((Q_i, G_i, I_i)\), 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 \( \{T_i | T_i = \{f_i^{Q_i}, f_i^{G_i}, f_i^{I_i}\}, i = 1, 2, ..., N\} \). We define a triplet loss function as

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

where \( d_i^+ = \|f_i^{Q_i} - f_i^{G_i}\|_2 \) is the dissimilarity score of the same subject pair \((Q_i, G_i)\) and \( d_i^- = \|f_i^{Q_i} - f_i^{I_i}\|_2 \) is the dissimilarity score of the different subject pair \((Q_i, I_i)\), 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 \( L_{\text{trip}} \), which is more suitable for the identification task than for the verification task.

\( L_{\text{trip}} \) 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 \( L_{\text{sim}} \) to force the identity features of the same subject pair \((Q \) and \( G)\) to be close to each other:

\[
L_{\text{sim}}(E) = \frac{1}{N} \sum_{i=1}^{N} d_i^+.
\]

**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

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

### 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

\[
L(E, D) = \lambda_{\text{reconst}} L_{\text{reconst}}(E, D) + \lambda_{\text{cont}} L_{\text{cont}}(E),
\]

where \( \lambda_{\text{reconst}} \) and \( \lambda_{\text{cont}} \) are two hyper-parameters.

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

\[
L(E, D) = \lambda_{\text{reconst}} L_{\text{reconst}}(E, D) + \lambda_{\text{trip}} L_{\text{trip}}(E) + \lambda_{\text{sim}} L_{\text{sim}}(E),
\]

where \( \lambda_{\text{reconst}} \), \( \lambda_{\text{trip}} \), and \( \lambda_{\text{sim}} \) 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 \( \beta \) [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 \( \beta \) 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.
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 using deconvolutional (transposed convolutional) layers. The other is pre-training a model on the largest data set (i.e., translation from -5 to 5 pixels with step of 2 for both vertical and horizontal axes) on OU-LP-Bag $\beta$ and CASIA-B considering their relatively small number of samples. 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 $\beta$ 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 $(\beta_1, \beta_2)$ 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 to $(300, 2), (100, 2),$ and $(8, 16)$ for OU-LP-Bag, OU-LP-Bag $\beta$, and CASIA-B, respectively. The margin $m$ in Eqs. (5) is set to 3. The weight parameters for the joint loss functions are set as $\lambda_{\text{reconst}} = 100$ and $\lambda_{\text{cont}} = 1$ in Eq. (8) and $\lambda_{\text{reconst}} = 1000$, $\lambda_{\text{trip}} = 1$, and $\lambda_{\text{sim}} = 0.1$ in Eq. (9) for all data sets, except that $\lambda_{\text{sim}}$ 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
clear that we can successfully reconstruct the input GEIs from the encoded feature $[f_{id}, f_{cov}]$ 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 $f_{id}$ 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.

**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-
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>EER (%)</th>
<th>Rank-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor GEI [44]</td>
<td>10.48</td>
<td>46.4</td>
</tr>
<tr>
<td>GEI w/ LDA [39]</td>
<td>8.10</td>
<td>54.6</td>
</tr>
<tr>
<td>GEI w/ 2DLDA [28]</td>
<td>11.47</td>
<td>43.3</td>
</tr>
<tr>
<td>GEI w/ R SVM [36]</td>
<td>10.81</td>
<td>28.3</td>
</tr>
<tr>
<td>GEF  [25]</td>
<td>6.67</td>
<td>58.3</td>
</tr>
<tr>
<td>JIS-ML [33]</td>
<td>5.45</td>
<td>57.4</td>
</tr>
<tr>
<td>GEINet [41]</td>
<td>9.75</td>
<td>40.7</td>
</tr>
<tr>
<td>JUCNet [58]</td>
<td>-</td>
<td>79.3</td>
</tr>
<tr>
<td>Proposed</td>
<td>2.03</td>
<td>86.6</td>
</tr>
<tr>
<td>LB* [51]</td>
<td>1.53</td>
<td>87.9</td>
</tr>
<tr>
<td>diff/2diff* [43]</td>
<td>1.31</td>
<td>87.8</td>
</tr>
<tr>
<td>JITN* [26]</td>
<td>1.27</td>
<td>88.1</td>
</tr>
<tr>
<td>Proposed*</td>
<td>0.77</td>
<td>97.2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Set-NM</th>
<th>Set-BG</th>
<th>Set-CL</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEI [15]</td>
<td>99</td>
<td>60</td>
<td>30</td>
<td>63.0</td>
</tr>
<tr>
<td>Genl [3]</td>
<td>98.3</td>
<td>80.1</td>
<td>33.5</td>
<td>70.6</td>
</tr>
<tr>
<td>STIP+NN [22]</td>
<td>95.4</td>
<td>60.9</td>
<td>52</td>
<td>69.4</td>
</tr>
<tr>
<td>GEINet [41]</td>
<td>97.5</td>
<td>84.5</td>
<td>71.8</td>
<td>84.6</td>
</tr>
<tr>
<td>L-CRF [5]</td>
<td>98.6</td>
<td>90.2</td>
<td>85.8</td>
<td>91.5</td>
</tr>
<tr>
<td>Proposed</td>
<td>100</td>
<td>82.0</td>
<td>73.0</td>
<td>85.0</td>
</tr>
<tr>
<td>Proposed*</td>
<td>100</td>
<td>100</td>
<td>93.0</td>
<td>97.7</td>
</tr>
</tbody>
</table>

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 ($L_{reconstr}$ and $L_{aim}$) to evaluate their respective performance. Table 5 shows that adding the disentanglement loss improves performance, which indicates the effectiveness of our disentanglement method.

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>(90°, 72°)</th>
<th>(90°, 108°)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>CL</td>
<td>BG</td>
<td>CL</td>
</tr>
<tr>
<td>RLITDA [18]</td>
<td>75.3</td>
<td>63.2</td>
<td>76.5</td>
</tr>
<tr>
<td>LB [51]</td>
<td>93.3</td>
<td>78.3</td>
<td>89.9</td>
</tr>
<tr>
<td>L-CRF [8]</td>
<td>94.4</td>
<td>88.5</td>
<td>92.2</td>
</tr>
<tr>
<td>JUCNet [58]</td>
<td>95.9</td>
<td>-</td>
<td>95.9</td>
</tr>
<tr>
<td>GaitNet [60]</td>
<td>95.6</td>
<td>94.2</td>
<td>87.4</td>
</tr>
<tr>
<td>Proposed</td>
<td>90.0</td>
<td>76.7</td>
<td>66.7</td>
</tr>
<tr>
<td>Proposed*</td>
<td>100</td>
<td>95.6</td>
<td>100</td>
</tr>
<tr>
<td>N/A</td>
<td>93.3</td>
<td>89.2</td>
<td>89.9</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Recognition loss</th>
<th>Disentanglement loss</th>
<th>EER (%)</th>
<th>Rank-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{cont}$</td>
<td>-</td>
<td>0.98</td>
<td>N/A</td>
</tr>
<tr>
<td>$L_{cont}$</td>
<td>$L_{reconstr}$</td>
<td>0.89</td>
<td>N/A</td>
</tr>
<tr>
<td>$L_{trip}$</td>
<td>-</td>
<td>N/A</td>
<td>84.02</td>
</tr>
<tr>
<td>$L_{trip}$</td>
<td>$L_{reconstr}$</td>
<td>N/A</td>
<td>86.37</td>
</tr>
<tr>
<td>$L_{trip}$</td>
<td>$L_{reconstr}$ + $L_{aim}$</td>
<td>N/A</td>
<td>87.04</td>
</tr>
</tbody>
</table>

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 autoencoder 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


13318
