Joint Intensity Transformer Network for Gait Recognition Robust Against Clothing and Carrying Status

Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, *Member, IEEE*, and Mingwu Ren

***Abstract*— Clothing and carrying status variations are the two key factors that affect the performance of gait recognition because people usually wear various clothes and carry all kinds of objects, while walking in their daily life. These covariates substantially affect the intensities within conventional gait representations such as gait energy images. Hence, to properly compare a pair of input gait features, an appropriate metric for joint intensity is needed in addition to the conventional spatial metric. We therefore propose a unified joint intensity transformer network for gait recognition that is robust against various clothing and carrying statuses. Specifically, the joint intensity transformer network is a unified deep learning-based architecture containing three parts: a joint intensity metric estimation net, a joint intensity transformer, and a discrimination network. First, the joint intensity metric estimation net uses a well-designed encoder-decoder network to estimate a sample-dependent joint intensity metric for a pair of input gait energy images. Subsequently, a joint intensity transformer module outputs the spatial dissimilarity of two gait energy images using the metric learned by the joint intensity metric estimation net. Third, the discrimination network is a generic convolution neural network for gait recognition. In addition, the joint intensity transformer network is designed with different loss functions depending on the gait recognition task (i.e., a contrastive loss function for the verification task and a triplet loss function for the identification task). The experiments on the world's largest datasets containing various clothing and carrying statuses demonstrate the state-of-the-art performance of the proposed method.**

***Index Terms*— Joint intensity transformer network, joint intensity metric learning, gait recognition.**

I. Introduction

GAIT is an important biometric that cannot be replaced by other biometrics (e.g., fingerprints, vein patterns, irises, or the face), because it is available even at a long distance with low image resolution. Because it is an unconscious behavior, people usually do not conceal their gait intentionally. Therefore, gait-based human recognition has attracted increasingly more attention by researchers for many applications such as surveillance systems, forensics, and criminal investigations [1]–[3]. The approaches to gait recognition in the literature are divided into two main categories: model-based [4]–[7] and appearance-based [8]–[12] approaches. While the former one usually requires high-resolution videos to fit a human model, the latter is more popular for relatively low-resolution videos.

For appearance-based approaches, the gait representations mainly include motion-based features [13], [14] and silhouette-based features (such as gait energy images (GEIs) [10], frequency-domain features [15], chrono-gait images [16], and Gabor GEIs [17]), where the latter type is more popular because of its simple yet effective properties. In particular, GEIs, also known as average silhouettes [18], are widely used in many studies. However, these appearance-based gait representations are easily changed by many covariates (e.g., view, clothing, and carrying status) resulting in large intrasubject difference, which greatly affects the performance of recognition.

Although most researchers mainly investigate view angle variations [19]–[24], clothing and carrying status variations are also very common in our daily life. This is because people usually walk while carrying different kinds of bags or other items. Moreover, they often change their clothes as the temperature changes. Therefore, gait recognition techniques robust against clothing and carrying status are also of great importance.

Traditional approaches to maintain the robustness of gait recognition against covariates fall into two families: spatial metric learning-based approaches and intensity transformation-based approaches. The former one concentrates on learning more discriminant features from original spaces and containing whole-based metric learning approaches (such as linear discriminant analysis (LDA) [10], discriminant analysis with tensor representation (DATER) [25], the random subspace method (RSM) [26], [27]), and part-based approaches [28]–[30], which decompose the holistic features into multiple body part-dependent features, then enhance or attenuate parts based on how they are influenced by covariates. However, spatial positions are influenced by covariates such as clothing and carrying status quite differently depending on the instance. Hence, it is insufficient to deal with these covariates using spatial metric learning techniques alone.

Manuscript received September 20, 2018; revised February 20, 2019; accepted April 16, 2019. Date of publication April 22, 2019; date of current version July 31, 2019. This work was supported in part by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (A) JP18H04115, in part by the National R&D Program for Major Research Instruments under Grant 61727802, and in part by the National Natural Science Foundation of China under Grant 61703209. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julian Fierrez. *(Corresponding author: Xiang Li.)*

X. Li and C. Xu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0046, Japan (e-mail: lixiangmzlx@gmail.com; xuchisherry@gmail.com).

Y. Makihara and Y. Yagi are with the Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0046, Japan (e-mail: makihara@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

M. Ren is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: renmingwu@mail.njust.edu.cn).

Digital Object Identifier 10.1109/TIFS.2019.2912577

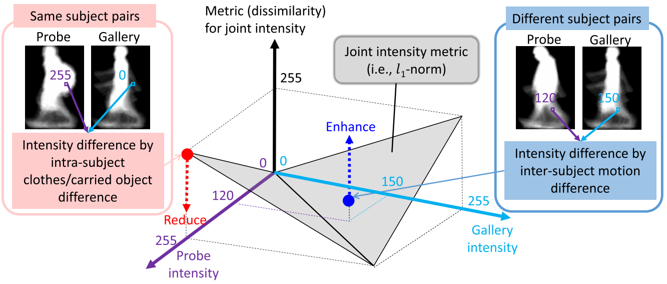

Same subject pairs
Probe
Gallery
255
0
Intensity difference by intra-subject clothes/carried object difference
Metric (dissimilarity) for joint intensity
255
Joint intensity metric (i.e., l_1-norm)
Enhance
Reduce
0
0
120
150
255
255
Probe intensity
Gallery intensity
Different subject pairs
Probe
Gallery
120
150
Intensity difference by inter-subject motion difference

Fig. 1. Meaning of joint intensity metric learning. The joint intensity metric learning is proposed to find a proper metric that can reduce the dissimilarity of the joint intensities that come from intrasubject clothes/carried object difference while enhancing the dissimilarity of the joint intensities that come from intersubject motion difference. Because a conventional joint intensity metric (i.e., l_1-norm), that returns a large dissimilarity for the intrasubject clothes/carried object difference (e.g., intensity level 255 vs. 0) and a small dissimilarity for the intersubject motion difference (e.g., intensity level 120 vs. 150), may result in a false match.

In contrast, intensity transformation-based approaches focus more on the feature representation aspect. They transform the intensity values of an original gait feature (gait energies in the case of GEI) into more discriminative values to increase the robustness against covariates. Because clothing and carrying status variations mainly affect the static components of human gait (e.g., a backpack and coat will change the torso and limb shapes) and partly affect the dynamic components of human gait (e.g., a dress can hide the leg motion and a handbag will affect the hand motion) during people's walking period, intensity transformation-based approaches are generally designed to enhance the effect of dynamic components while reducing the effect of static components. Typical approaches include handcrafted transformations like gait entropy image (GEnI) [11] and masked GEIs [12] as well as training-based transformation such as gait energy response functions [31], [32]. Recently, instead of transforming a single GEI, Makihara et al. [33] proposed a joint intensity metric learning-based method that focused on the joint intensity transformation of a pair of GEIs, which reduces the large intrasubject differences and leverages the subtle intersubject differences, as shown in Fig. 1. Through transforming gait energies, these intensity transformation-based approaches show their unique advantages dealing with the variations compared with spatial metric learning. Additionally, they can be easily combined with spatial metric learning techniques to boost performance.

In recent years, thanks to the great success of deep learning techniques, many approaches [14], [21]–[24], [34]–[41] have been proposed in the gait recognition community that significantly improve on the performance of traditional approaches. However, they all employ various types of spatial metric learning while ignoring intensity metric learning. Currently, there are no deep learning-based methods that employ intensity metric learning.

There is an existing work called the spatial transformer network [42] that regresses affine transformation parameters for spatial transformation to distorted digits. Inspired by this, we propose a new architecture to deal with the joint intensity transformation of a pair of GEIs for joint intensity metric learning. Compared with [33], which learns a fixed joint intensity metric using a framework consisting of a linear support vector machine (SVM), the proposed method utilizes deep learning networks to learn a sample-dependent joint intensity metric for intensity transformation that is more suitable for various clothing and carrying status types (appearing in different positions depending on the instance). For example, in the case of a relatively small variation, an incrementally modulated joint intensity metric is estimated, whereas for a large variation such as a large carried object, the large intrasubject difference is strongly suppressed while subtle motion differences are strongly enhanced. Finally, through the learned sample-dependent joint intensity metric, the proposed method can adaptively handle the intrasubject differences of the same subject pair caused by clothing and carrying status variations as well as the intersubject differences of different subject pairs caused by motion difference, which results in better recognition performance.

In this paper, we propose the unified joint intensity transformer network (JITN) for gait recognition that is robust against various clothing and carrying statuses and takes both spatial and intensity metric learning into consideration. To the best of our knowledge, this is the first work integrating joint intensity metric learning into a deep learning-based framework. Specifically, JITN is a unified CNN-based architecture containing three parts, i.e., a joint intensity metric estimation net (JIMEN), a joint intensity transformer, and a discrimination network (DN). More details are given in Section III. The contributions of this paper are summarized as follows:

A. A Unified CNN-Based Method Considering Both Joint Intensity and Spatial Metric Learning

The proposed JITN is a unified CNN-based method considering both joint intensity and spatial metric learning. It contains a JIMEN, a joint intensity transformer, and a DN, where the JIMEN is a well-designed encoder-decoder network to estimate the joint intensity metric, the joint intensity transformer is a transformation module, and the DN is a generic convolution neural network to learn the spatial metric. They are jointly trained from end to end.

B. Sample-Dependent Joint Intensity Metric for Intensity Transformation

Unlike the fixed joint intensity metric learned in [33], the joint intensity metric learned by the JIMEN performs a sample-dependent transformation based on the input pairs, which is more suitable for dealing with all kinds of variations in clothing and carrying status than traditional approaches, which perform common sample-independent transformations.

C. State-of-the-Art Performance

We achieve state-of-the-art performance on gait recognition under variations in clothing and carrying status on four publicly available gait databases: the OU-ISIR Large Population Gait database with real-life carried objects (OU-LP-Bag) [43], the OU-ISIR Gait database, Large Population dataset with bags β version (OU-LP-Bag β) [33], the OU-ISIR Gait Database, Treadmill Dataset B (OUTD-B) [44] and the TUM Gait from Audio, Image and Depth Database (TUM-GAID) [45].

II. Related Work

A. Spatial Metric Learning-Based Approaches

Spatial metric learning-based approaches concentrate on improving performance by learning a feature space from the original appearance-based features that is more discriminant and robust against the covariates. There are two further categories within this family: whole-based [10]–[12], [25]–[27], [46] and part-based approaches [28]–[30], [47].

For the whole-based approaches, the holistic appearance-based features are projected into a discriminative space to make them more robust against the covariate conditions. For example, Han and Bhanu [10] applied LDA to real and synthesized GEI templates to reduce intraclass variations to some extent. A RSM framework that combines multiple inductive biases also was proposed in [26], [27].

The part-based approaches decompose the holistic appearance-based features into multiple body part-dependent features and enhance the parts effective for recognition while attenuating the parts affected by the covariate conditions. This is because variations such as clothing and carrying status usually affect not the whole gait but only certain parts, and a decrease in accuracy is derived mainly from the affected parts. Thus, the part-based approaches have the potential to achieve better accuracy by appropriate treatment of the affected body parts (e.g., reducing the weights of the affected body parts for recognition). For example, in [28], the human body was divided into eight sections based on anatomical knowledge and the effect of clothing variations was mitigated by adaptively assigning larger and smaller weights to the affected and unaffected sections, respectively. Iwashita et al. [29] divided the human body into several areas equally and then estimated a comparison weight for each area. Weights were based on the similarity between the extracted features and those in the database for standard clothing.

B. Intensity Transformation-Based Approaches

Intensity transformation-based approaches transform the intensity values of an original gait feature into more discriminative values to increase the robustness against changes of the covariate conditions. For example, Bashir et al. [11] computed the GEnI using the Shannon entropy of the foreground probability at each pixel (i.e., the gait energy in the GEI). A GEnI encodes the randomness of pixel values in the silhouette images over a complete gait cycle, thereby capturing more motion information (dynamic components) rather than static information, which improves robustness against shape changes (e.g., clothing and carrying status). Masked GEI [12] is another intensity transformation-based approach that keeps the dynamic components as their original values but zero-pads the static components (i.e., almost all foreground and almost all background parts) using a certain threshold. Instead of using hand-crafted transformation, Li et al. [31] proposed a gait energy response function that transformed intensities in a data-driven way. Recently, Makihara et al. [33] proposed a joint intensity transformation-based method that focused on the joint intensity transformation of a pair images instead of a single one. Specifically, the joint intensity metric was alternately learned in conjunction with a spatial metric in a framework based on a linear SVM. However, these approaches all use traditional methods (hand-crafted design or linear optimization) to perform sample-independent transformations.

C. Deep Learning-Based Approaches

Many studies on deep learning-based gait recognition have been published recently [14], [21]–[24], [34]–[41]. For example, the work [39] is a survey on deep learning for biometrics including gait. Wolf et al. [21] designed a 3D CNN model that regarded raw silhouettes from each gait sequence as a spatiotemporal input. Battistone et al. [38] proposed a time-based graph deep learning approach to jointly exploit the temporal information and skeleton data extracted from silhouettes. Shiraga et al. [22] designed an eight-layered CNN network called GEINet using averaged silhouettes (i.e., GEI). These networks all regard gait recognition as person classification from the same gait class. In addition, Wu et al. [23] designed multiple networks with two input GEIs (i.e., a pair consisting of a probe ($query$) and gallery ($enrollment$) GEI) by considering layers to start the comparison of the input pair. Takemura et al. [24] discussed input/output architectures for CNN-based gait recognition. Their networks attempt to learn the similarity between input GEIs, then determine whether they come from the same person or not. In [36], a stacked auto-encoder was used to find invariant gait features that were robust against multiple covariates. In [37], [41], generative adversarial networks were utilized for generating feature maps without covariates. Except for silhouette-based feature GEIs, motion features (e.g., optical flow maps) were also used in some approaches [14], [34]. Additionally, apart from person authentication or identification, Liu et al. [40] also investigated deep learning-based approaches on gait-based gender recognition under clothing and carrying status variations. All of these approaches achieved significant improvements compared with traditional approaches. However, they all employ various types of spatial metric learning while ignoring intensity metric learning, which is another helpful technique for dealing with the covariates of clothing and carrying status.

III. Gait Recognition Using JITN

A. Overview

In this paper, we choose to use the GEI feature, which is the most widely employed gait representation in many works including traditional approaches [10], [26]–[28], [33] and deep learning-based approaches [22]–[24], [35]. To generate a GEI, given a raw video sequence of a subject, we first extract human silhouettes using a background subtraction-based graph-cut segmentation [48] or recent deep learning-based semantic segmentation methods such as RefineNet [49]; second, we obtain

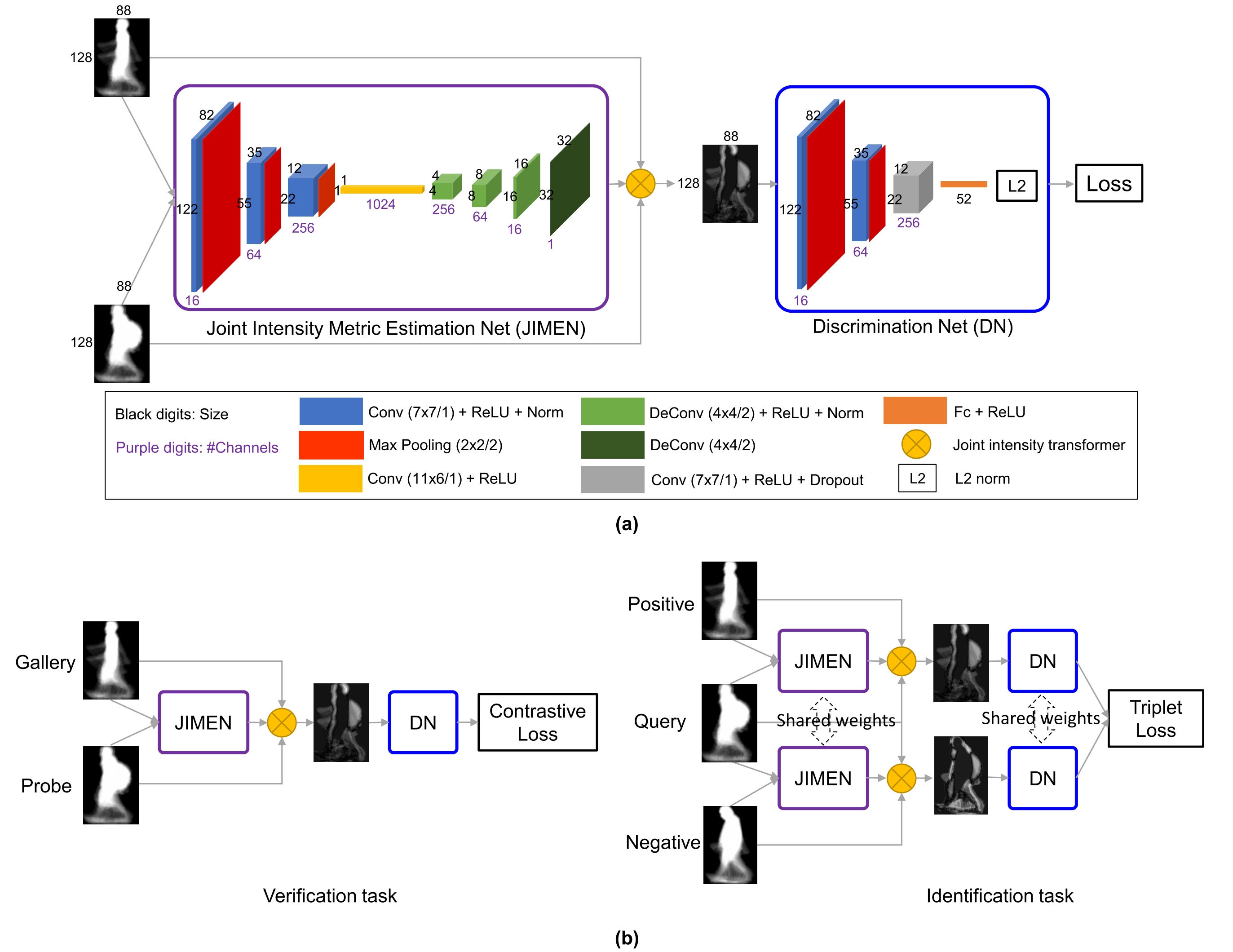


Fig. 2. Joint intensity transformer network (JITN) for gait recognition. (a) Overview of the proposed JITN framework, which consists of a joint intensity metric estimation net (JIMEN), joint intensity transformer, and a discrimination network (DN) trained from end-to-end. Conv, DeConv, ReLU, Norm, MAX-pooling, dropout, and Fc denote the convolutional layer, deconvolutional (transposed convolutional) layer, ReLU activation layer, normalization layer, max pooling layer, dropout layer, and fully connected layer respectively. The L2 norm is the L2 norm of the output feature at the previous Fc layer. The numbers in brackets written after Conv, DeConv, and MAX-pooling indicate (kernel height × kernel width / stride). (b) For different recognition tasks, we define different JITN architectures. The left architecture is for the verification task with a contrastive loss function, and the right architecture is for the identification task with a triplet loss function.

size-normalized and registered silhouettes [15] based on the extracted region's height and center; third, we detect a gait period by maximizing the auto-correlation of the size-normalized and registered silhouettes; finally, we average the silhouettes over one gait period to obtain a GEI.

The network architecture of the proposed JITN is shown in Fig. 2 (a). It consists of a JIMEN, joint intensity transformer, and a DN trained from end-to-end. Given a probe and gallery GEI pair, we first estimate the joint intensity metric using the JIMEN. Then, the joint intensity transformer generates the spatial dissimilarity feature map of the original probe and gallery using the estimated joint intensity metric. Finally, for spatial metric learning, the spatial dissimilarity feature map is fed into a DN, which outputs the final dissimilarity of the probe and gallery. More details of the modules are given in the rest of the section.

B. Joint Intensity Transformer

In this subsection, we first introduce the concept of joint intensity metric learning, which was proposed in [33]. Given a pair of gray-scale images $\boldsymbol{I}^P$ and $\boldsymbol{I}^G$ with a resolution of $H \times W$ (height by width), their dissimilarity measure, incorporating the joint intensity metric, is represented as

$$D(\boldsymbol{I}^P, \boldsymbol{I}^G; \boldsymbol{w}_I) = \sum_{i=1}^{H}\sum_{j=1}^{W} w_I(I^P_{i,j}, I^G_{i,j}), \tag{1}$$

where $w_I(I^P_{i,j}, I^G_{i,j})$ is a spatially independent dissimilarity metric for joint intensity $(I^P_{i,j}, I^G_{i,j})$ at position (i, j). In other words, $\boldsymbol{w}_I \in R^{(I_{\max}+1)\times(I_{\max}+1)}$ can be regarded as a two-dimensional look-up table from intensity pairs to dissimilarities, where $I_{\max}$ is the maximum gray value and usually is

255 for 8-bit gray images. For example, the l_1-norm is a typical joint intensity metric, i.e., $w_I(I^P_{i,j}, I^G_{i,j}) = |I^P_{i,j} - I^G_{i,j}|$.

Using this metric, we can more flexibly design joint intensities, whereas traditional linear or quadratic metrics (e.g., the l_1-norm or Mahalanobis distance) are limited to monotonically increasing metrics as the absolute difference of joint intensities increases. Although monotonically increasing properties are generally reasonable and exploited in most image matching algorithms, they are not always suitable under certain circumstances in gait recognition.

Given joint intensity metric $\boldsymbol{w}_I$, along with input probe GEI $\boldsymbol{I}^P$ and gallery GEI $\boldsymbol{I}^G$, a joint intensity transformer outputs a dissimilarity map T, whose value $T_{i,j}$ at position (i, j) is written as

$$T_{i,j} = w_{I,(p_{i,j},g_{i,j})} = \sum_{k=0}^{I_{\max}} \sum_{l=0}^{I_{\max}} \delta_{k,p_{i,j}} \delta_{l,g_{i,j}} w_{I,(k,l)}, \tag{2}$$

where $\delta_{a,b}$ is Kronecker's delta. Note that this computation is also regarded as forward propagation in the proposed deep neural network framework.

For computing a backward propagation of the loss through this joint intensity transformer, we define the partial derivative of $T_{i,j}$ with respect to $w_{I,(k,l)}$ as follows:

$$\frac{\partial T_{i,j}}{\partial w_{I,(k,l)}} = \delta_{k,p_{i,j}} \delta_{l,g_{i,j}}. \tag{3}$$

We further consider a downsampled joint intensity metric (e.g., 32 × 32) instead of a joint intensity metric with the full size of intensities (i.e., 256 × 256), because a joint intensity metric that is larger in size may need complex regression models and a relatively long time for computation. To do this, we introduce N control points distributed over intensity levels from 0 to $I_{\max}$ in both the probe and gallery at a certain interval ($\frac{I_{\max}}{N-1}$) and estimate the weights at intermediate intensities by bilinear interpolation from the adjacent control points. Note that bilinear interpolation is carried out along the diagonal and anti-diagonal directions because the joint intensity metric (e.g., the l_1-norm) is usually symmetric along the diagonal direction, as shown in Fig. 3.

Suppose $\boldsymbol{w}^d_I \in \mathbb{R}^{N \times N}$ is the downsampled joint intensity metric. Given a position (k, l) of the original joint intensity metric, we compute the weight $w_{I,(k,l)}$ by bilinear interpolation from a quadruplet of adjacent control points in $\boldsymbol{w}^d_I$ as

$$\begin{aligned} w_{I,(k,l)} = c_{k,l}&((1-a_{k,l})w^d_{I,(m_k,n_l)} + a_{k,l}w^d_{I,(m_k+1,n_l+1)}) \\ &+(1-c_{k,l})((1-b_{k,l})w^d_{I,(m_k,n_l+1)} \\ &\qquad +b_{k,l}w^d_{I,(m_k+1,n_l+2)}), \end{aligned} \tag{4}$$

where $m_k = \lfloor k/N \rfloor$, $n_l = \lfloor l/N \rfloor$, and $\lfloor . \rfloor$ is a floor function. The coefficients a, b, and c can be easily computed as $a_{k,l} = (k/N - m_k + l/N - n_l)/2$, $b_{k,l} = a_{k,l} - 1/2$, and $c_{k,l} = 1 + k/N - m_k - (l/N - n_l)$.

We can rearrange Eq. (2) by replacing $w_{I,(k,l)}$ with Eq. (4). In addition, the partial derivative of $T_{i,j}$ with respect to $w_{I,(k,l)}$ (see Eq. (3)) turns out to be four partial derivatives of $T_{i,j}$ with respect to four adjacent control points of the downsampled joint intensity metric (i.e., $w^d_{I,(m_k,n_l)}$, $w^d_{I,(m_k,n_l+1)}$, $w^d_{I,(m_k+1,n_l+1)}$, and $w^d_{I,(m_k+1,n_l+2)}$), which are easily obtained by Eq. (4) as follows:

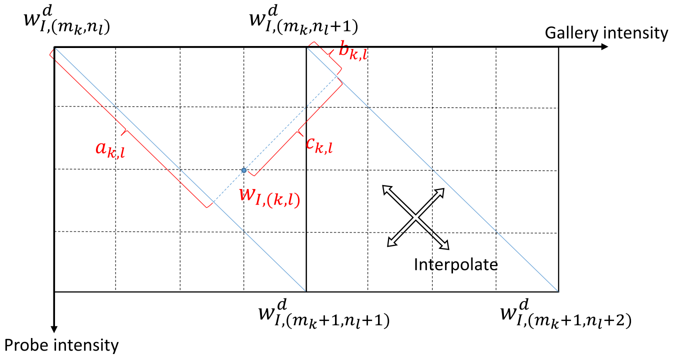


Fig. 3. Interpolation of original joint intensity metric $w_{I,(k,l)}$ using four surrounding downsampled joint intensity metrics $w^d_{I,(m_k,n_l)}$, $w^d_{I,(m_k,n_l+1)}$, $w^d_{I,(m_k+1,n_l+1)}$, and $w^d_{I,(m_k+1,n_l+2)}$ with coefficients $a_{k,l}$, $b_{k,l}$, and $c_{k,l}$ for the bilinear interpolation. Note that the interpolation is carried out along the diagonal and anti-diagonal directions.

$$\begin{aligned} \frac{\partial T_{i,j}}{\partial w^d_{I,(m_k,n_l)}} &= \delta_{k,p_{i,j}} \delta_{l,g_{i,j}} c_{k,l}(1-a_{k,l}), \\ \frac{\partial T_{i,j}}{\partial w^d_{I,(m_k,n_l+1)}} &= \delta_{k,p_{i,j}} \delta_{l,g_{i,j}} c_{k,l} a_{k,l}, \\ \frac{\partial T_{i,j}}{\partial w^d_{I,(m_k+1,n_l+1)}} &= \delta_{k,p_{i,j}} \delta_{l,g_{i,j}} (1-c_{k,l})(1-b_{k,l}), \\ \frac{\partial T_{i,j}}{\partial w^d_{I,(m_k+1,n_l+2)}} &= \delta_{k,p_{i,j}} \delta_{l,g_{i,j}} (1-c_{k,l}) b_{k,l}. \end{aligned} \tag{5}$$

C. JIMEN

We design a JIMEN to estimate the joint intensity metric, as shown on the left side of Fig. 2 (a). It takes a pair of GEIs and outputs the joint intensity metric. Specifically, the JIMEN is an encoder-decoder framework in which the encoder first takes a difference image of a probe and gallery GEI, then learns the effective subspace feature through four convolutional layers; the decoder first takes the output feature of the encoder, then generates a two-dimensional representation feature as the joint intensity metric through four deconvolutional (transposed convolutional) layers. In this network, the ReLU activation function is used for all convolutional layers and the first three deconvolutional layers, the local response normalization (LRN) [50] is used for the normalization layers, and a max pooling strategy is chosen for the pooling layers. Unlike STN [42], which adopts a fully connected layer as the final regression layer and outputs a vector of spatial deformation parameters (e.g., affine transformation parameters), we design an encoder-decoder framework because it can regress the two-dimensional joint intensity metric well.

D. DN

After the joint intensity transformation of a pair of GEIs using the joint intensity transformer, their spatial dissimilarity image is fed into a generic DN to learn the spatial metric, as shown on the right side of Fig. 2 (a). A DN shares the same architecture as the diff net in [24]. Specifically, a DN has three convolutional layers followed by a ReLU activation layer and LRN normalization layer. It also has one fully connected layer that has a 52-dimensional feature. Subsequently, the L2 layer calculates the L2 norm of the 52-dimensional feature as the final dissimilarity of the input pair of GEIs. To avoid overfitting, a dropout technique (with a ratio of 0.5) [51] is applied after the third convolutional layer. The DN is designed to reduce the final dissimilarity of a same-subject pair while increasing the final dissimilarity of a different-subject pair through the contrastive and triplet loss functions, which are introduced in the next subsection.

E. Networks for Different Gait Recognition Tasks

Gait recognition includes two kinds of tasks: gait verification and gait identification. For the verification task, a probe and gallery pair is compared and then it is determined whether they come from the same subject or different subjects. If their dissimilarity is lower than a certain acceptance threshold, they are judged to be from the same subject and vice versa. For the identification task, a probe is compared with all the galleries to find the same subject that appears in the probe. We usually calculate the dissimilarities between the probe and all the galleries, then use a nearest neighbor classifier to find the subject with the smallest dissimilarity.

Referring to [24], different gait recognition tasks have their own suitable network architectures and loss function, i.e., a Siamese network with contrastive loss for the verification task and a triplet network with triplet loss for the identification task. Therefore, we design different networks depending on different gait recognition tasks in the training phase, as shown in Fig. 2 (b). More specifically, for the verification task, we choose a contrastive loss function [52] as the loss function of the proposed JITN framework, which is defined as follows:

$$\mathcal{L}_{\text{cont}} = \frac{1}{2C}\sum_{i=1}^{C} y_i d_i^2 + (1-y_i)\max(\text{margin} - d_i, 0)^2, \tag{6}$$

where C is the number of GEI pairs for training, d_i is the dissimilarity score in the L2 norm layer of the i-th pair of GEIs, and y_i is equal to 1 if the i-th pair is the same subject pair and 0 otherwise. Using Eq. (6), the network trains its parameters such that the dissimilarity scores of the same subject pairs are always smaller than those of different subject pairs, which is suitable for the verification scenario.

For the identification task, we choose a triplet loss function [53] as the loss function of the proposed JITN framework. Triplet GEIs called $query$, $positive$, and $negative$ are fed into the network, where $positive$ is of the same subject as that of $query$ and $negative$ is of a different subject from that of $query$. Then, a triplet loss function is defined as follows:

$$\mathcal{L}_{\text{trip}} = \frac{1}{2C}\sum_{i=1}^{C} \max(\text{margin} - d_i^- + d_i^+, 0)^2, \tag{7}$$

where C is the number of triplets for training, d_i^- is the dissimilarity score in the L2 norm layer between $query$ and $negative$, and d_i^+ is the dissimilarity score in the L2 norm layer between $query$ and $positive$. Using Eq. (7), the network trains its parameters so that the dissimilarity score of $query$ and $positive$ is always smaller than the dissimilarity score of $query$ and all $negative$ GEIs, which is suitable for the identification scenario.

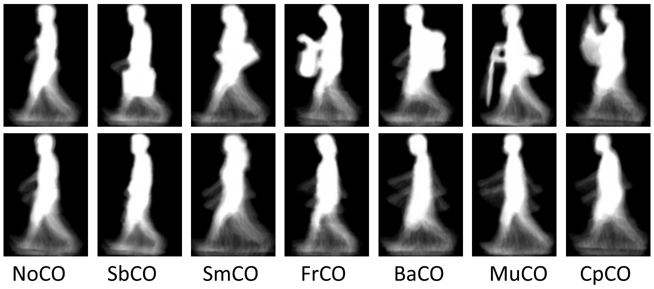


Fig. 4. Examples of GEI with seven annotated carrying status labels from OU-LP-Bag. The first row shows subjects with a carrying status and the second row shows subjects without carrying status.

IV. Experiments

A. Datasets

We use four publicly available databases,[1] OU-LP-Bag [43], OU-LP-Bag β [33], OUTD-B [44], and TUM-GAID [45], for the experiments.

OU-LP-Bag is currently the world's largest gait database with real-life carried objects. The data were collected in conjunction with an experience-based demonstration of video-based gait analysis at a science museum [54]. It includes a total of 62,528 subjects with seven annotated carrying status labels (i.e., NoCO for no carried objects, SbCO for objects carried on the side bottom, SmCO for objects carried on the side middle, FrCO for objects carried in front, BaCO for objects carried in back, MuCO for objects carried in multiple locations, and CpCO for objects carried changing from one location to another). Some typical GEI examples can be seen in Fig. 4. Each subject was captured three times to produce walking sequences. The first sequence (A_1) can be with or without carried objects (that is, some participants did not hold any objects), the other two (A_2 and A_3) are without carried objects. Among all subjects, 58,199 subjects that have a sample both in A_1 and either A_2 or A_3 were chosen for the experiments. The chosen subjects were randomly divided into two subsets: a training set (29,097 subjects) and a test set (29,102 subjects). The test set is further divided into a gallery set and a probe set. Considering the uncooperative-subject condition in real scenarios, Uddin et al. [43] introduced both cooperative and uncooperative settings in the test set. For the cooperative setting, samples from A_2 or A_3 are in the gallery set, while samples from A_1 are in the probe set. For the uncooperative setting, samples are randomly assigned into the gallery and probe sets.

[1] OU-LP-Bag, OU-LP-Bag β and OUTD-B are available at http://www.am.sanken.osaka-u.ac.jp/BiometricDB/index.html; TUM-GAID is available at https://www.mmk.ei.tum.de/en/misc/tum-gaid-database/

The OU-LP-Bag β is the beta version of OU-LP-Bag. There are 2,070 subjects in the dataset and each subject has two sequences, one with carried objects and the other without carried objects. The whole dataset is divided into three subsets: a training set, gallery set, and probe set. The training set contains 2,068 sequences of 1,034 subjects, while the remaining disjoint 1,036 subjects are included in the gallery and probe sets. The gallery set comprises sequences without carried objects, while the probe set has sequences with carried objects.

The OUTD-B has the largest number of clothing variations (up to 32). It is divided into three subsets: a training set, gallery set, and probe set. In the training set, there are 446 sequences of 20 subjects with a range of 15 to 28 different combinations of clothing. The gallery and probe sets constitute a testing set that comprises 48 subjects, which are disjoint from the 20 subjects in the training set. The gallery contains only standard clothing types (e.g., regular pants and full shirt), while the probe set includes 856 sequences of other clothing types.

The TUM-GAID simultaneously contains RGB video, depth and audio data with 305 subjects walking under four conditions: normal walking (N), carrying a backpack (B), wearing coating shoes (S), and elapsed time ($TN-TB-TS$) collected at January and April which may also exist changes in clothing or lighting condition. All subjects contain ten sequences: six normal walking ($N1-N6$), two backpack variation ($B1-B2$), and two shoes variation ($S1-S2$). 32 subjects among them contain additional ten sequences under elapsed time variation, namely $TN1-TN6$, $TB1-TB2$, $TS1-TS2$. Following the protocol of the original paper [45], the whole dataset is divided into three subsets: a training set with 100 subjects, a validation set with 50 subjects, and a test set with 155 subjects. Half of the subjects under elapsed time variation is included in the test set and another half is included in the training and validation set. In the test set, $N1-N4$ are set as the gallery set, while $N5-N6$, $B1-B2$, $S1-S2$, $TN5-TN6$, $TB1-TB2$, and $TS1-TS2$ are set as six different probe sets.

B. Implementation Details

We use Xavier's algorithm to initialize the weight parameters of all layers except for the last deconvolutional layer in JIMEN, which is separately set to initialize the joint intensity metric. The bias parameters are all set to the constant zero. The momentum for all layers is 0.9. We set the initial learning rate to 0.01 and divide it by 10 four times during the training phase. The proposed network is trained for a total of 0.1 million iterations using the stochastic gradient descent algorithm with a mini-batch size of 300. We implement the whole framework using Caffe [55] on a NVIDIA GeForce GTX TITAN X GPU with 12 G memory. The hyper-parameter margins in Eqs. (6) and (7) are experimentally set to three.

As for the joint intensity metric, which is regressed by JIMEN, we set the number N of control points distributed over the intensity levels to 32 and obtain a downsampled joint intensity metric with a size of 32×32. Regarding the initialization, we set it be the signed l_1-norm, i.e., $w_{I,(k,l)} = k-l$. To do so, we first set the weight and bias parameters of the last deconvolutional layer in JIMEN to zero, which forces the output to zero; then, we add a dummy layer initialized by the signed l_1-norm that has the same size as the joint intensity metric.

Regarding the sampling problem for training, we uncooperatively choose subjects from the training set; that is, we do not fix the gallery to be a subject with clothing or carried objects. Basically, for the verification task, we choose all the same and different subject pairs in the training set and then duplicate the same subject pairs so that their number is one-ninth that of the different subject pairs; for the identification task, we basically choose all the triplets in the training set. However, for the largest database OU-LP-Bag, there are billions of untrackable pairs and triplets. Thus, we simply randomly choose from all pairs and triplets while keeping the total number to about 10 million.

C. Evaluation Metrics

We refer to the standards defined in ISO/IEC 19795-1 on biometric performance testing and reporting [56] for evaluation metrics. For the verification task, a detection error trade-off (DET) curve, which indicates a trade-off between false non-match rate (FNMR) and false match rate (FMR) when an acceptance threshold changes, is employed. Specifically, FNMR is the proportion of genuine attempts that are falsely declared not to match a template of the same subject and FMR is the proportion of the imposter attempts that are falsely declared to match a template of another subject. In addition, we also calculate the equal error rate (EER), where FNMR is equal to FMR. For the identification task, a cumulative match characteristic (CMC) curve, which shows the identification rates of actual subjects included within each of the ranks, is employed. In addition, we calculate the rank-1 identification rate, denoted as Rank-1.

D. Learned Joint Intensity Metric

In this subsection, we analyze the learned joint intensity metrics and show some typical comparison examples using the learned metrics in Fig. 5. As mentioned before, the learned joint intensity metrics are sample-dependent trained by the proposed network (see Figs. 5 (i) and (j)), which is more suitable for dealing with all kinds of variations in clothing and carrying status than traditional approaches [31], [33], which perform common sample-independent transformations. Note that for learned joint intensity metrics, darker or brighter regions indicate an enhancement of the metrics, while grayer regions (i.e., a gray value close to 127) indicates a degradation of the metrics. We also calculated the difference between the learned joint intensity metrics and the initial one to show the changes, and then colored these changes in Fig. 5 (k) and (l); that is, the blue and yellow regions represent the degradation and enhancement of the metrics, respectively.

When compared with the initial joint intensity metric (i.e., the signed l_1-norm, see Fig. 5 (h)), for true match pairs, the learned joint intensity metrics show more degradation near the top-right and bottom-left corners, which is mainly

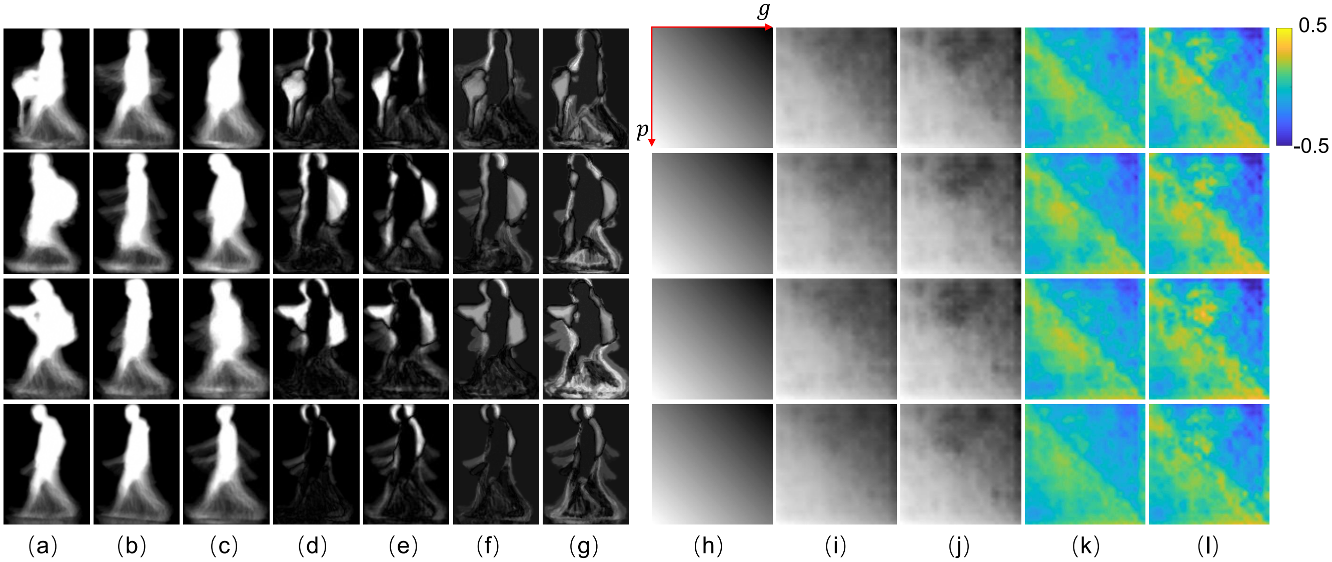


Fig. 5. Comparison examples for four probes with different types of carried objects (i.e., FrCO, BaCO, MuCO, and NoCO from top to bottom) and their corresponding learned joint intensity metric. (a) Probe. (b) Gallery (genuine). (c) Gallery (imposter). (d) and (e) Spatial dissimilarity with the absolute difference for a true match pair (probe vs. genuine) and false match pair (probe vs. imposter), respectively. (f) and (g) Spatial dissimilarity with absolute learned joint intensity metrics for a true match pair (probe vs. genuine) and false match pair (probe vs. imposter), respectively. (h) Initial joint intensity metric, i.e., signed l_1-norm. (i) and (j) Learned joint intensity metrics for a true match pair (probe vs. genuine) and false match pair (probe vs. imposter). (k) and (l) Changes in the initial and learned joint intensity metrics for a true match pair (probe vs. genuine) and false match pair (probe vs. imposter), where the blue and yellow regions represent a decline and enhancement of metrics on the joint intensity pairs, respectively.

derived from the intrasubject carrying status variations. This indicates that the learned joint intensity metrics can successfully suppress the effect of carrying status, which causes large intrasubject difference. In contrast, for false match pairs, the learned joint intensity metrics show more enhancements near the diagonal line, which is mainly derived from the motion differences (e.g., leg or hand motions). This indicates that the learned joint intensity metrics can enhance the effect of motion differences, which causes small intersubject difference. These conclusions can also be arrived at from the examples (see Figs. 5 (f) and (g)). The spatial dissimilarity of the true match pairs decreases, especially in the regions of carried objects, while the spatial dissimilarity of false match pairs increases especially in the regions of leg motion.

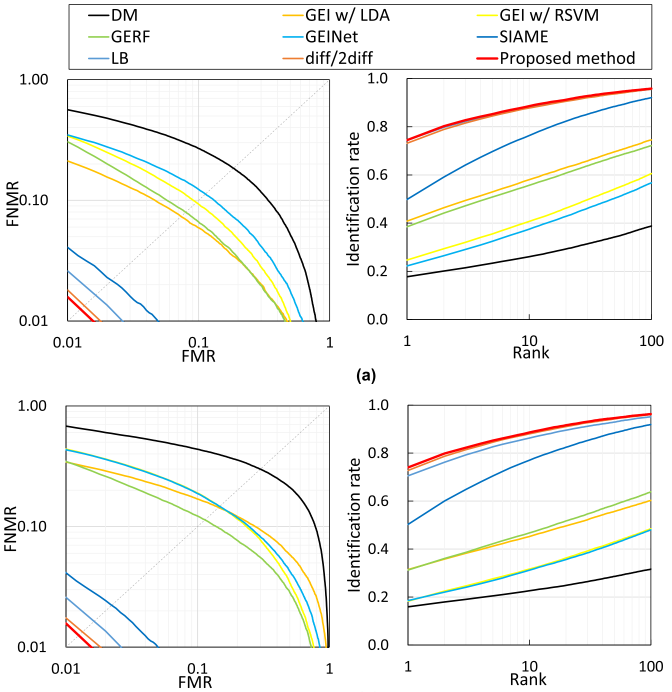
When the metrics are compared for the four different types of carried objects, for large variations such as FrCO, BaCO, and MuCO, the learned joint intensity metrics strongly suppress the large intrasubject difference and enhance the subtle motion difference. In contrast, for very small variations such as NoCO, a less drastically modulated joint intensity metric is yielded because the metric changes are relatively smaller than the other three types of large carried objects (see Figs. 5 (k) and (l), where the results in the bottom row have lighter yellow and blue colors than those in the upper three rows).

Therefore, with the learned joint intensity metrics, we get a smaller spatial dissimilarity for true match pairs regardless of carried objects in various locations (e.g., front, back, both front and back, or even with no carried objects) and a larger spatial dissimilarity for false match pairs, which successfully mitigates the effect of carrying status and leads to the correct recognition results.

E. Comparison on OU-LP-Bag

In this subsection, we evaluate the robustness of the proposed method against real-life carried objects on OU-LP-Bag. The state-of-the-art methods for comparison consist of traditional methods (direct matching (DM) of GEIs [10], spatial metric learning-based approaches such as GEI w/ LDA [57], GEI w/ RSVM [58], and intensity transformation-based approaches such as GERF [31]) and recent deep learning-based methods (GEINet [22], the Siamese GEINet (SIAME) [59], LB [23], and diff/2diff [24]). The proposed method and the compared deep learning-based benchmarks were trained from scratch on this dataset with the same protocol. The network parameters were set to the defaults given in their original papers. Of all the methods, diff/2diff, proposed by Takemura et al. [24], has the architecture that is most similar to that of the proposed method, which uses contrastive loss and triplet loss function. Takemura et al. simply take the pair-wise difference at the beginning of the network and proceed as the proposed method does with a fixed signed l_1-norm as the joint intensity metric. Thus, the comparison between the proposed method and diff/2diff reflects the effect of joint intensity metric learning well.

We show the DET and CMC curves of all methods in Fig. 6. We also show the EER and Rank-1 of each method in Table I. The results show that the proposed method achieves the best performance both in the verification and identification scenarios for both cooperative and uncooperative settings. Moreover, it outperforms traditional methods by a large margin. Although LB [23] achieves the second best Rank-1, which is only 0.05% lower than the proposed method in a cooperative setting, it faces a clear decrease of nearly 4% for Rank-1 in an uncooperative setting, which is more likely in a real scene.



(a)

(b)

Fig. 6. DET and ROC curves for the comparison experiments on OU-LP-Bag in both the (a) cooperative setting and (b) uncooperative setting. The left side shows the DET curves and the right side shows the ROC curves.

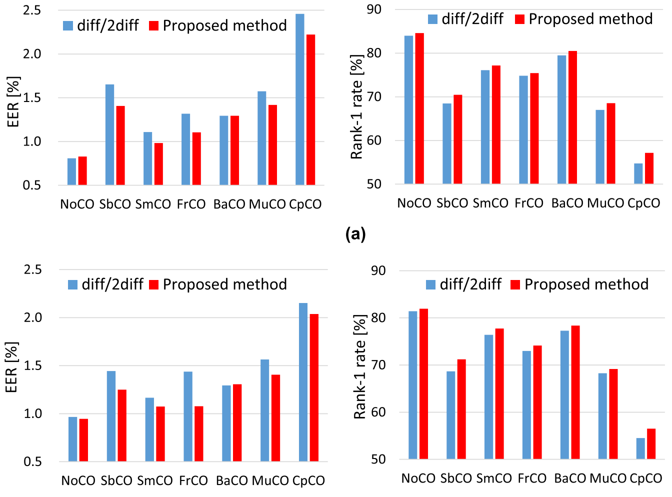
TABLE I

EER and Rank-1 [%] Results for the Comparison Experiments on OU-LP-Bag in Both Cooperative and Uncooperative Settings. For Diff /2diff, the EER Result Is From the "Diff" Method, While the Rank-1 Result Is From "2diff" Method. Bold and Italic Bold Fonts Indicate the Best and Second-Best Results, Respectively. This Convention Is Consistent Throughout This Paper

	Cooperative		Uncooperative	
Methods	EER	Rank-1	EER	Rank-1
DM [10]	18.46	17.74	29.89	15.90
GEI w/ LDA [57]	7.35	40.79	14.40	31.44
GEI w/ RSVM [58]	9.58	24.66	14.69	18.28
GERF [31]	7.97	38.48	11.35	31.24
GEINet [22]	11.29	22.26	14.68	18.52
SIAME [59]	2.17	49.80	2.22	50.27
LB [23]	1.68	***74.39***	1.66	70.53
diff/2diff [24]	***1.36***	73.14	***1.35***	***72.75***
Proposed method	**1.25**	**74.44**	**1.25**	**74.03**

In contrast, the proposed method only shows a slight decline of 0.41% for Rank-1. Besides, for the EER, the proposed method clearly outperforms LB with a 0.4% lower EER in both cooperative and uncooperative settings. Moreover, when compared with diff/2diff, the proposed method clearly outperforms it with a 0.1 % lower EER and 1.3% higher Rank-1, which demonstrates the effectiveness of the joint intensity metric learning in the proposed framework.

We also evaluated the robustness of the proposed method against different types of carried objects. We choose diff/2diff as the baseline method, which is a CNN-based method that only considers spatial metric learning. The results are shown in Fig. 7. We can see that the proposed method shows lower EERs and higher Rank-1 rates for almost all types of carried objects in both cooperative and uncooperative settings. In case of more difficult types such as FrCO, MuCO, and CpCO, the improvements of the proposed method are substantial, which is because the joint intensity metric learning can handle relatively large carried objects well regardless of their carried locations. For NoCO (no carried objects), the improvement is subtle because, without carried objects, the proposed method acts more like the baseline method and considers no joint intensity metric learning. Through this analysis, we confirm the flexibility of proposed method, which can effectively learn sample-dependent joint intensity metrics and handle all kinds of carrying statuses.



(a)

(b)

Fig. 7. EER and Rank-1 results for the proposed method and baseline method (diff/2diff) under different carried object conditions on OU-LP-Bag. (a) Cooperative setting and (b) uncooperative setting. The left side shows the EER results and the right side shows the Rank-1 results.

F. Comparison on OU-LP-Bag β

In this subsection, we evaluate the robustness of the proposed method against carrying status on OU-LP-Bag β. The comparison benchmarks are mainly from [33] and also include recent state-of-the-art deep learning-based methods (LB [23] and diff/2diff [24]). Because of the relatively small number of training samples of OU-LP-Bag β, for deep learning-based methods, we fine-tuned deep models that were pre-trained on the OU-LP-Bag dataset. To do so, we first set a smaller learning rate of 0.001 and tuned all layers of these networks. Figure 8 and Table II show the results of all methods. Here, the proposed method achieves the best performance. In contrast with JIS-ML proposed by Makihara et al. [33], which was the first to introduce joint intensity metric learning and integrated it with spatial metric learning in a traditional linear SVM framework, the proposed method successfully integrates joint intensity metric learning in a unified CNN-based framework trained from

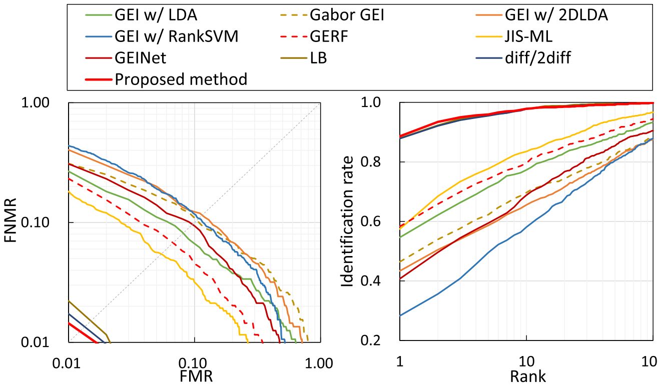


Fig. 8. DET and ROC curves for the comparison experiments on OU-LP-Bag β. The left side shows the DET curve and the right side shows the ROC curve.

TABLE II

EER AND RANK-1 [%] RESULTS FOR THE COMPARISON EXPERIMENTS ON OU-LP-BAG β

Methods	EER	Rank-1
Gabor GEI [17]	10.48	46.4
GEI w/ LDA [57]	8.10	54.6
GEI w/ 2DLDA [60]	11.47	43.3
GEI w/ RSVM [58]	10.81	28.3
GERF [31]	6.67	58.3
JIS-ML [33]	5.45	57.4
GEINet [22]	9.75	40.7
LB [23]	1.53	***87.9***
diff/2diff [24]	***1.31***	87.8
Proposed method	**1.27**	**88.1**

end to end. The proposed approach yields a much higher performance.

G. Comparison on OUTD-B

In this subsection, we evaluate the robustness of the proposed method against various clothing types on OUTD-B, which has the largest variation of labeled clothing. Similar to OU-LP-Bag β, for the deep learning-based methods, we used models pre-trained on the OU-LP-Bag dataset and applied the same fine-tuning strategy. The results of all comparison methods are shown in Fig. 9 and Table III. It can be seen that the proposed method achieves the best EER and the second best Rank-1 of all methods. Although Gabor+RSM-HDF [27] obtains the best Rank-1, we note it has several weaknesses: 1) it cannot be used for the verification task because of its majority voting scheme for all galleries and 2) because it requires multiple samples per gallery to compute the within-class scatter from the gallery set, it cannot be used in datasets with a single sample per gallery (e.g., OU-LP-Bag β). Therefore, the proposed method is promising because of its wide range of applications both in identification and verification scenarios as well as its state-of-the-art performance.

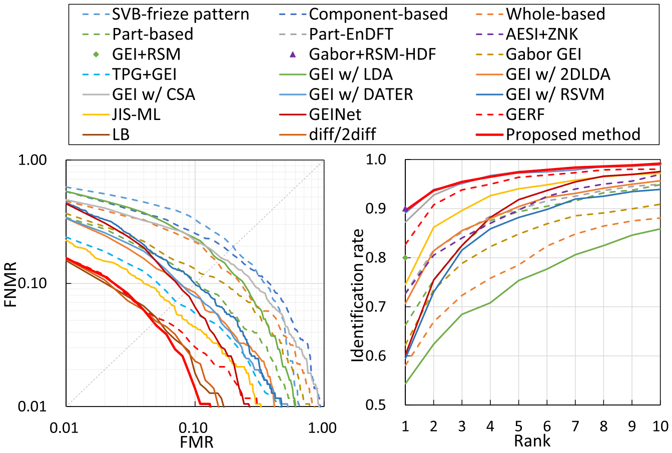


Fig. 9. DET and ROC curves for comparison experiments on OUTD-B. The left side shows the DET curve and the right side shows the ROC curve.

TABLE III

EER AND RANK-1 [%] RESULTS FOR THE COMPARISON EXPERIMENTS ON OUTD-B. N/A AND - MEAN NOT APPLICABLE AND NOT PROVIDED, RESPECTIVELY

Methods	EER	Rank-1
SVB frieze pattern [46]	19.81	-
Components-based [61]	18.25	-
Whole-based [15]	14.88	58.1
Part-based [28]	10.26	66.3
Part-EnDFT [30]	-	72.8
AESI+ZNK [62]	-	72.7
GEI+RSM [26]	N/A	80.4
Gabor+RSM-HDF [27]	N/A	**90.7**
Gabor GEI [17]	11.80	62.3
TPG+GEI [63]	7.10	-
GEI w/ LDA [57]	15.63	54.3
GEI w/ 2DLDA [60]	8.91	70.7
GEI w/ CSA [64]	16.00	-
GEI w/ DATER [25]	8.72	-
GEI w/ RSVM [58]	10.75	58.4
JIS-ML [33]	6.66	74.5
GEINet [22]	8.38	60.2
GERF [31]	5.14	82.7
LB [23]	5.11	87.3
diff/2diff [24]	***4.99***	89.1
Proposed method	**4.79**	***89.6***

H. Comparison on TUM-GAID

In this subsection, we evaluate the robustness of the proposed method on TUM-GAID dataset, which contains scenarios where both clothing and carrying status change. The method [14] shows state-of-the-art results with relatively low-resolution input features (i.e., 60×60). For fair comparison, we use the same resolution to show the performance of the proposed method under circumstances with low-resolution GEIs. To adapt the change of input size, we slightly adjust the kernel size of the fourth convolutional layer of JIMEN be 3×3. We still use fine-tuning strategy for the TUM-GAID dataset on the models, which were pre-trained on the OU-LP-Bag dataset where the features are also first resized to the low resolution. All 150 subjects from the training and validation sets are used to generate training pairs/triplets in the fine-tuning stage. The GEI features are extracted from tracked

TABLE IV

RANK-1 [%] RATE FOR THE COMPARISON EXPERIMENTS ON TUM-GAID. N, B, S, TN, TB, AND TS REPRESENT DIFFERENT PROBE SETS WITH DIFFERENT COVARIATES

Methods	N	B	S	Avg	TN	TB	TS	Avg
Baseline [45]	99.4	27.1	52.6	59.7	44.0	6.0	9.0	19.7
RSM [65]	**100**	79.0	97.0	92.0	58.0	38.0	57.0	51.3
CNN-SVM [66]	***99.7***	97.1	97.1	98.0	59.4	50.0	62.5	57.3
PFM [13]	***99.7***	**99.0**	99.0	**99.2**	**78.1**	***56.3***	46.9	60.4
DMT [14]	***99.7***	97.4	**99.7**	98.9	59.4	**62.5**	**68.8**	**63.6**
Proposed method	***99.7***	***98.1***	***99.4***	***99.1***	***62.5***	**62.5**	***65.6***	***63.5***

depth image sequences by the method described in [45], and then resized to 60×60 (the original height-to-width ratio of the subjects is kept). Because the GEIs are not as well aligned as other datasets, we employ an additional registration step (i.e., shift the probe in both horizontal and vertical directions to minimize the l_1-norm between the gallery and the shifted probe) to GEI pairs both in the training and test stages for better performance. The results of all comparison methods are shown in Table IV. Only the performance in the identification task is presented because few works reported their results in the verification task. From the results, the proposed method achieves very competitive performance compared with the best state-of-the-art method [14], which implies the proposed method can handle the change of the resolution of the input images and ensure its good performance.

I. Analysis of the Effects of Noise

Reviewing the process of GEI feature extraction, the quality of GEI features highly depends on the segmentation results of human silhouettes from raw video sequences. Although recent deep learning-based methods help to improve the segmentation results, there may still exist somewhat over-segmented or under-segmented parts (i.e., noise) near the boundaries of the human silhouettes. Therefore, robustness to such noise is very important for real world applications. Since current databases contain no noise variation, we use simulated noise data on the silhouettes for the experiment. Specifically, the OUTD-B dataset is chosen due to its relatively high-quality silhouettes. The noise is shaped into a circle with random radius from 2 to 5 pixels. The center of the noise circle is assumed to appear at the boundary pixels of the silhouettes, and we random decide it be over-segmented or under-segmented within each circle (e.g., suppose intensity values 0 and 255 belong to the background and human part, respectively; if over-segmented, all the pixels in the circle are set to be 255; if under-segmented, all the pixels are set to be 0). Figure 10 shows some examples of the noise data with three different appearance frequencies (i.e., 0.01, 0.05, and 0.1), which is defined as the ratio between the number of noise and the number of boundary pixels. Obviously, larger noise appearance frequency results in worse silhouettes and GEIs. As for the performance evaluation against noise, we simply assume a setting where both probe and gallery contain the same type of noise. We prepare two models: one is trained without any noise; another is trained together with noise. The results on both models are shown in Table V. In case of models trained without noise, the proposed method shows its robustness to moderate and small noise (i.e., 0.05 and 0.01 appearance frequencies). Moreover, even for very large noise (i.e., 0.1 appearance frequency), it still achieves better performance than most benchmarks that use the test set without noise. Additionally, if the models are trained together with noise, the proposed method could increase its robustness against noise and show much better results.

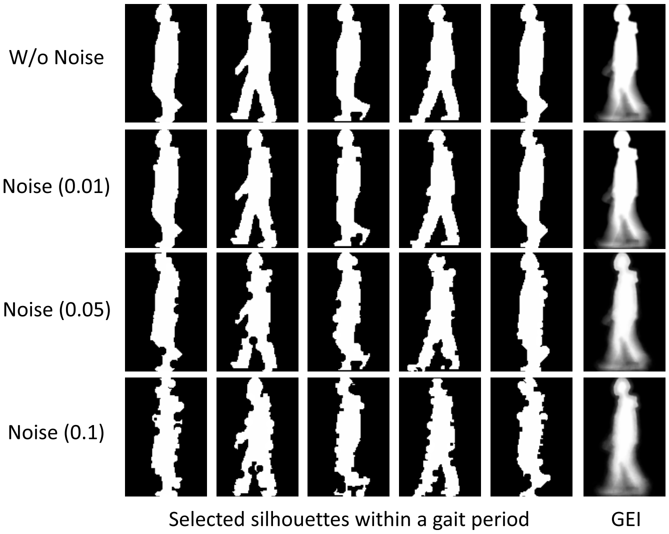


Fig. 10. Examples of some selected silhouettes and their GEIs without and with noise. There are three different appearance frequencies (i.e., 0.01, 0.05, and 0.1) of the noise, which represent different degrees that affected by the noise.

TABLE V

EER AND RANK-1 [%] RESULTS OF THE PROPOSED METHOD AGAINST NOISE UNDER TWO MODELS TRAINED W/O AND W/ NOISE ON OUTD-B

	Trained w/o noise		Trained w/ noise	
	EER	Rank-1	EER	Rank-1
W/o noise	4.79	89.6	4.89	89.7
Noise (0.01)	4.83	89.1	4.91	89.1
Noise (0.05)	6.19	85.3	5.14	87.7
Noise (0.1)	7.69	81.3	5.37	85.9

J. Stability Analysis

In this subsection, we analyze the stability of the proposed method in terms of the performance on the test set. We choose the OU-LP-Bag dataset for this experiment because it has the largest number of test samples (29,102 subjects). The whole test set is randomly divided into five equally disjoint gallery and probe sets. We use the same models as section IV-E for evaluation, which are trained on the whole training set. Table VI shows the mean value and standard deviation of the performance for the comparison methods. From the results, the proposed method is still superior to other benchmarks even if we consider the uncertainty (i.e., mean ± standard deviation).

TABLE VI

EER and Rank-1 [%] Results of the Proposed Method Compared With Two Other Benchmarks Under Cooperative Setting on OU-LP-Bag. Avg and Std Represent the Mean Value and Standard Deviation of the Results on Five Equally Disjoint Gallery and Probe Sets, Respectively

	EER		Rank-1	
	Avg	Std	Avg	Std
LB [23]	1.68	**0.04**	***82.52***	0.62
diff/2diff [24]	***1.37***	***0.05***	81.80	***0.49***
Proposed method	**1.25**	***0.05***	**82.90**	**0.46**

V. Conclusion

In this paper, we proposed a unified joint intensity transformer network for gait recognition that is robust against various clothing and carrying status. To the best of our knowledge, this is the first work integrating joint intensity metric learning into a deep learning-based framework. Specifically, JITN is a unified CNN-based architecture containing three parts: a JIMEN, a joint intensity transformer, and a DN. Additionally, it is designed with different loss functions depending on the gait recognition task. Experimental results using four publicly available datasets demonstrate the state-of-the-art performance of the proposed method compared with other state-of-the-art methods.

We use two different network structures with the contrastive/triplet losses for the verification/identification tasks, respectively, and this might be prohibited when a memory usage is limited (e.g., an embedded system). We will therefore consider to train a unified model by combining the verification/identification losses in a multi-task setting which reduces memory usage. Additionally, because the proposed method mainly focuses on the joint intensity transformation to deal with clothing and carrying status covariates, we will consider how to modify it to cope with cross-view gait recognition, which performs a spatial transformation that handles the large spatial displacement caused by view angle changes. Moreover, the combination of both joint intensity and spatial transformation for all covariates remains another direction for future work.

Acknowledgment

The authors would like to thank Md. Zasim Uddin for providing part of the results in the experiments. They also thank Kim Moravec, PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript. Moreover, they would like to thank all the reviewers for their valuable comments.

References

[1] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, 2011.

[2] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, pp. 163–175, Oct. 2013.

[3] N. Lynnerup and P. K. Larsen, "Gait as evidence," *IET Biometrics*, vol. 3, no. 2, pp. 47–54, Jun. 2014.

[4] R. Urtasun and P. Fua, "3D tracking for gait characterization and recognition," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 17–22.

[5] D. K. Wagg and M. S. Nixon, "On automated model-based extraction and analysis of gait," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 11–16.

[6] C. Yam, M. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.

[7] 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.

[8] S. Sarkar, J. Phillips, Z. Liu, I. Vega, P. G. ther, and K. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[9] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.

[10] 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.

[11] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. Int. Conf. Crime Detection Prevention*, Dec. 2009, pp. 1–6.

[12] K. Bashir, T. Xiang, and S. Gong, "Gait recognition without subject cooperation," *Pattern Recognit. Lett.*, vol. 31, no. 13, pp. 2052–2060, 2010.

[13] F. M. Castro, M. J. Marín-Jiménez, N. G. Mata, and R. Muñoz-Salinas, "Fisher motion descriptor for multiview gait recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 1, 2017, Art. no. 1756002.

[14] M. J. Marín-Jiménez, F. M. Castro, N. Guil, F. de la Torre, and R. Medina-Carnicer, "Deep multi-task learning for gait-based biometrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 106–110.

[15] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. 9th Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 151–163.

[16] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2164–2176, Nov. 2012.

[17] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[18] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 1, Aug. 2004, pp. 211–214.

[19] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 696–709, Feb. 2014.

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

[21] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4165–4169.

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

[23] 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.

[24] 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.*, to be published.

[25] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H.-J. Zhang, "Human gait recognition with matrix representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 896–903, Jul. 2006.

[26] Y. Guan, C.-T. Li, and Y. Hu, "Robust clothing-invariant gait recognition," in *Proc. 8th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP)*, Jul. 2012, pp. 321–324.

[27] Y. Guan, C. T. Li, and F. Roli, "On reducing the effect of covariate factors in gait recognition: A classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1521–1528, Jul. 2015.

[28] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, Jun. 2010.

[29] Y. Iwashita, K. Uchino, and R. Kurazume, "Gait-based person identification robust to changes in appearance," *Sensors*, vol. 13, no. 6, pp. 7884–7901, 2013.

[30] M. Rokanujjaman, M. S. Islam, M. A. Hossain, M. R. Islam, Y. Makihara, and Y. Yagi, "Effective part-based gait identification using frequency-domain gait entropy features," *Multimedia Tools Appl.*, vol. 74, no. 9, pp. 3099–3120, May 2015.

[31] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren, "Gait energy response function for clothing-invariant gait recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 257–272.

[32] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren, "Gait energy response functions for gait recognition against various clothing and carrying status," *Appl. Sci.*, vol. 8, no. 8, p. 1380, 2018.

[33] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint intensity and spatial metric learning for robust gait recognition," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5705–5715.

[34] F. M. Castro, M. J. Marín-Jiménez, N. Guil, S. López-Tapia, and N. P. de la Blanca, "Evaluation of CNN architectures for gait recognition based on optical flow maps," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2017, pp. 1–5.

[35] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2832–2836.

[36] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, May 2017.

[37] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2017, pp. 30–37.

[38] F. Battistone and A. Petrosino, "TGLSTM: A time based graph deep learning approach to gait recognition," *Pattern Recognit. Lett.*, to be published.

[39] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Comput. Surv.*, vol. 51, Jul. 2018, Art. no. 65.

[40] T. Liu, X. Ye, and B. Sun, "Clothing and carrying invariant gait-based gender recognition," *Proc. SPIE*, vol. 10836, Oct. 2018, Art. no. 108360X.

[41] 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.

[42] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. New York, NY, USA: Curran Associates, 2015, pp. 2017–2025.

[43] 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, p. 5, May 2018.

[44] Y. Makihara *et al.*, "The OU-ISIR gait database comprising the treadmill dataset," *IPSJ Trans. Comput. Vis. Appl.*, vol. 4, pp. 53–62, Apr. 2012.

[45] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 195–206, Jan. 2014.

[46] S. Lee, Y. Liu, and R. Collins, "Shape variation-based frieze pattern for robust gait recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[47] N. V. Boulgouris and Z. X. Chi, "Human gait recognition based on matching of body components," *Pattern Recognit.*, vol. 40, no. 6, pp. 1763–1770, 2007.

[48] Y. Makihara and Y. Yagi, "Silhouette extraction based on iterative spatio-temporal local color transformation and graph-cut segmentation," in *Proc. 19th Int. Conf. Pattern Recognit.*, Tampa, FL, USA, Dec. 2008, pp. 1–4.

[49] G. Lin, A. Milan, C. Shen, and I. D. Reid. (2016). "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation." [Online]. Available: https://arxiv.org/abs/1611.06612

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

[53] J. Wang *et al.* (2014). "Learning fine-grained image similarity with deep ranking." [Online]. Available: https://arxiv.org/abs/1404.4661

[54] Y. Makihara *et al.*, "Gait collector: An automatic gait data collection system in conjunction with an experience-based long-run exhibition," in *Proc. 8th IAPR Int. Conf. Biometrics (ICB)*, Halmstad, Sweden, Jun. 2016, pp. 1–8.

[55] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[56] *Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, Standard ISO/IEC 19795-1:2006(en) and ISO/IEC JTC 1/SC 37, Int. Org. Standardization, Geneva, Switzerland, 2006.

[57] N. Otsu, "Optimal linear and nonlinear solutions for least-square discriminant feature extraction," in *Proc. 6th Int. Conf. Pattern Recognit.*, 1982, pp. 557–560.

[58] R. Martín-Félez and T. Xiang, "Uncooperative gait recognition by learning to rank," *Pattern Recognit.*, vol. 47, no. 12, pp. 3793–3806, Dec. 2014.

[59] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.

[60] K. Liu, Y. Q. Cheng, and J. Y. Yang, "Algebraic feature extraction for image recognition based on an optimal discriminant criterion," *Pattern Recognit.*, vol. 26, no. 6, pp. 903–911, 2006.

[61] X. Li, S. J. Maybank, S. Yan, D. Tao, and D. Xu, "Gait components and their application to gender recognition," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 145–155, Mar. 2008.

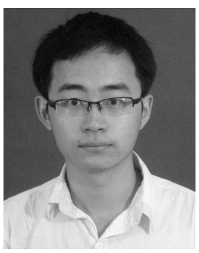
[62] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for gait recognition based upon zernike moment invariants," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 2, pp. 397–407, Jun. 2018.

[63] S. Lombardi, K. Nishino, Y. Makihara, and Y. Yagi, "Two-point gait: Decoupling gait from body shape," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1041–1048.

[64] D. Xu, S. Yan, L. Zhang, H.-J. Zhang, Z. Liu, and H.-Y. Shum, "Concurrent subspaces analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 203–208.

[65] Y. Guan, X. Wei, C.-T. Li, G. L. Marcialis, F. Roli, and M. Tistarelli, "Combining gait and face for tackling the elapsed time challenges," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep./Oct. 2013, pp. 1–8.

[66] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. P. de la Blanca, "Automatic learning of gait signatures for people identification," in *Proc. Int. Work-Conf. Artif. Neural Netw. (IWANN)*, 2017, pp. 257–270.

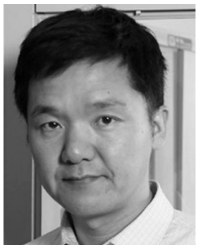


Xiang Li received the B.S. degree in computer science and technology from the Nanjing University of Science and Technology (NUST), China, in 2012, where he is currently pursuing the Ph.D. degree. Since 2016, he has been with the Institute of Scientific and Industrial Research, Osaka University, as a Visiting Researcher. His research interests are gait recognition, image processing, and machine learning.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University, in 2001, 2002, and 2005, respectively. He is currently an Assistant Professor with the Institute of Scientific and Industrial Research, Osaka University. His research interests are gait recognition, morphing, and temporal super-resolution. He is a member of the IPSJ, RJS, and JSME.

Chi Xu received the B.S. degree in computer science and technology from the Nanjing University of Science and Technology (NUST), China, in 2012, where she is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems. Since 2016, she has been with the Institute of Scientific and Industrial Research, Osaka University, Japan, as a Visiting Researcher. Her research interests are gait recognition, machine learning, and image processing.

Yasushi Yagi (M'91) received the Ph.D. degree from Osaka University, in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate, in 1990, a Lecturer, in 1993, an Associate Professor, in 1996, and a Professor, in 2003, at Osaka University. He was also the Director of the Institute of Scientific and Industrial Research, Osaka University, from 2012 to 2015, where he is the Executive Vice President. His research interests are computer vision, medical engineering, and robotics. He is a fellow of the IPSJ and a member of the IEICE and RSJ. He was a recipient of the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of T.J. Tan Best Paper in Robotics, the IEEE ICRA2008 Finalist for Best Vision Paper, the MIRU2008 Nagao Award, and the PSIVT2010 Best Paper Award. He has served as the Chair for International conferences, including the FG1998 (Financial Chair), OMINVIS2003 (Organizing Chair), ROBIO2006 (Program Co-Chair), ACCV2007 (Program Chair), PSVIT2009 (Financial Chair), ICRA2009 (Technical Visit Chair), ACCV2009 (General Chair), ACPR2011 (Program Co-Chair), and ACPR2013 (General Chair). He has also served as the Editor for the IEEE ICRA Conference Editorial Board (2007–2011). He is the Editorial Member of the IJCV and the Editor-in-Chief of the *IPSJ Transactions on Computer Vision and Applications*.



Mingwu Ren received the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2001. He is currently a Professor with the School of Computer Science and Engineering, NUST. His current research interests include computer vision, image processing, and pattern recognition.