Cross-View Gait Recognition Using Pairwise Spatial Transformer Networks

Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, *Member, IEEE*, and Jianfeng Lu

***Abstract*— In this paper, we propose a pairwise spatial transformer network (PSTN) for cross-view gait recognition, which reduces unwanted feature mis-alignment due to view differences before a recognition step for better performance. The proposed PSTN is a unified CNN architecture that consists of a pairwise spatial transformer (PST) and subsequent recognition network (RN). More specifically, given a matching pair of gait features from different source and target views, the PST estimates a non-rigid deformation field to register the features in the matching pair into their intermediate view, which mitigates distortion by registration compared with the case of direct deformation from the source view to target view. The registered matching pair is then fed into the RN to output a dissimilarity score. Although registration may reduce not only intra-subject variations but also inter-subject variations, we can still achieve a good trade-off between them using a loss function designed to optimize recognition accuracy. Experiments on three publicly available gait datasets demonstrate that the proposed method yields superior performance for both verification and identification scenarios by combining any gait recognition network benchmarks with the PST.**

***Index Terms*— Pairwise spatial transformer, convolutional neural network, gait recognition, cross-view.**

I. Introduction

GAIT is one popular behavioral biometric modality that can be used to authenticate a person from his/her walking style. Compared with other physiological biometrics (e.g., DNA, fingerprints, irises, and faces), it exhibits unique advantages in applications such as surveillance and criminal investigation using cameras (e.g., closed-circuit television (CCTV)) because it works even for a subject captured at a long distance from a camera without his/her cooperation (i.e., a person at a low image resolution) and also is difficult to keep on concealing and impersonating in daily life. Gait recognition has therefore been of great importance for many applications in surveillance, forensics, and criminal investgation [1]–[3].

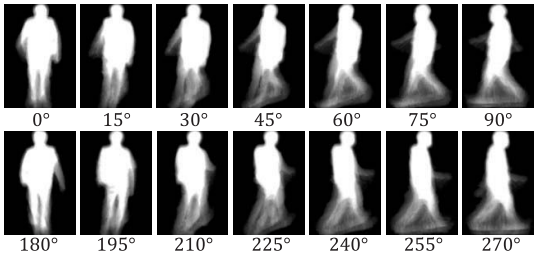


Fig. 1. GEI examples from the same subject with different view angles. Obvious intra-subject variations exist among the GEI features from different observation view angles.

Involving uncooperative subjects, however, makes gait recognition more easily affected by various covariates, including view [4], [5], clothing [6], [7], and walking speed [8], [9]. Among these covariates, view variation is one of the most common challenging factors and often exists in real applications (e.g., CCTV footage captured from different observation view angles). As shown in Fig. 1, the view changes raise large intra-subject variations in widely used appearance-based gait features, such as the gait energy image (GEI) [10], which may drastically degrade the performance of gait recognition.

Extensive studies [4], [5], [11]–[16] on cross-view gait recognition mainly fall into two categories: the generative approaches and discriminative approaches. The generative approaches generally transform gait features from one view (e.g., gallery view) to a different view (e.g., probe view) [4], [13], or transform features from different views into a common canonical view (e.g., side view) [11], [17]. However, these approaches do not guarantee optimal recognition accuracy [14], [15] because they essentially consider not recognition accuracy but the quality of transformed gait features. By contrast, discriminative approaches mainly aim at learning view-invariant subspaces or metrics to directly optimize the discrimination capability without undertaking registration among features from various views, such as traditional linear discriminant analysis (LDA) [10] and rank support vector machine (rank SVM) [18]. It is, however, difficult to find a robust subspace or metric for non-aligned features, particularly for the case of large view differences.

Manuscript received July 9, 2019; revised December 7, 2019; accepted February 15, 2020. Date of publication February 21, 2020; date of current version January 7, 2021. This work was supported in part by the JSPS Grants-in-Aid for Scientific Research (A) under Grant JP18H04115, in part by the Jiangsu Provincial Science and Technology Support Program under Grant BE2014714, in part by the Programme of Introducing Talents of Discipline to Universities under Grant B13022, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions. This article was recommended by Associate Editor S. Gao. *(Corresponding author: Chi Xu.)*

Chi Xu and Xiang Li 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 567-0047, Japan (e-mail: xuchisherry@gmail.com; lixiangmzlx@gmail.com)

Yasushi Makihara and Yasushi Yagi are with the Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan (e-mail: makihara@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

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

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2020.2975671

Recently, by introducing convolutional neural network (CNN) frameworks, the performance of cross-view gait recognition has been promoted increasingly further [5], [14]–[16], [19]–[23]. The great success of CNN-based methods is partly because of the max-pooling layers, which allow the networks to be somewhat spatially invariant to the intra-subject variations in the body parts that result from view differences [15]. On the other hand, the max pooling layer may wash out subtle inter-subject variation (e.g., a slight back contour difference caused by body shape variation), and hence we need to consider the trade-off between spatial displacement invariance caused by view variation and maintaining subtle inter-subject variation. In fact, [15] suggested maintaining the subtle inter-subject variation by taking the difference at the low level before going through the pooling layer, when the view angle difference is small. This implies that CNN-based gait recognition accuracy may further improve by incorporating a preceding registration process for a matching pair of gait features, in a similar manner to face recognition accuracy improving as a result of image registration in advance for pose variation [24], [25] and expression variation [26].

The spatial transformer network (STN) [27] is one such registration technique, and includes a spatial transformer (ST) module that explicitly performs spatial transformation on input features. Because the ST is typically used as a sub-network of the entire STN designed for main tasks such as object classification [27], [28] and face recognition [29]–[31], the ST is trained to provide registration parameters that are suitable for the main task. The conventional ST often takes a single input and regresses affine transformation parameters to transform the input features to a canonical view or pose for the main task. Whereas such an architecture is suitable for classification tasks, such as digit recognition [27], it is not necessarily suitable for matching tasks, such as cross-view gait recognition. For example, assume that the canonical view is set to side view 90°, whereas a matching pair is observed from 0° and 30°. In this case, it is infeasible to transform gait features from 0° and 30° into those from the canonical view, that is, 90°, and hence, a direct application of the conventional ST is an unsuitable choice for cross-view gait recognition.

We therefore propose a unified pairwise spatial transformer network (PSTN) that contains a pairwise spatial transformer (PST) module for cross-view gait recognition, which takes a pair of probe and gallery gait features as the input in the network architecture. Instead of transforming a single input feature into a canonical view using affine transformation, the features in the input pair from different views are both registered into their intermediate view via the learned appropriate non-rigid transformation by the proposed PSTN, where the intra-subject differences caused by the view variations are well suppressed while maintaining the inter-subject differences simultaneously. The contributions of this work are four-fold.

A. PST for Cross-View Matching

Rather than transforming the single input into a general canonical view in the conventional ST [27], the proposed PST transforms a pair of inputs (i.e., probe and gallery) into their intermediate view between the probe view and gallery view, which avoids unnecessary large distortion. To the best of our knowledge, this is the first time that geometric feature registration has been introduced into a CNN-based gait recognition framework.

B. Managing Arbitrary View Combination Without Knowing the View Information in Advance With a Single Unified CNN Model

Unlike existing generative approaches that require view information in advance, the proposed method does not require view information throughout the training and testing processes, because the PST is trained to output a suitable transformation field for each input pair without view information.

C. A Unified CNN Framework That Involves Both Generative and Discriminative Models

The proposed PSTN is composed of a PST and subsequent recognition network (RN), which is a unified CNN framework that involves both generative and discriminative models, and hence the optimal transformation for the main recognition task that achieves a trade-off between intra-subject variations and inter-subject variations is predicted by the proposed PSTN, unlike traditional generative methods, which only aim to reduce the intra-subject differences in feature generation rather than optimizing recognition accuracy.

D. State-of-the-Art Performance Among GEI-Based Methods on Three Publicly Available Datasets

We evaluated the proposed method on three publicly available gait datasets: the OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) [32], OU-ISIR Gait Database, Large Population Dataset (OULP) [33], and CASIA Gait Database, Dataset B (CASIA-B) [34]. OU-MVLP is the world's largest gait database with wide view variation, which enables a more statistically reliable performance evaluation, whereas OULP and CASIA-B are datasets widely used for existing cross-view gait recognition studies. By combining the proposed PST with any gait recognition network benchmarks as the following RN, the proposed PSTN achieves performance improvement on the three datasets, which yields state-of-the-art accuracy in the field of GEI-based methods for both verification and identification scenarios.

II. Related Work

A. Generative Approaches to Cross-View Gait Recognition

Generative approaches to cross-view gait recognition can be divided into two categories: geometry-based and example-based. Some geometry-based approaches construct a three-dimensional (3D) human model for gallery subjects from two-dimensional (2D) images for multiple views using a model-fitting [35], [36] or visual intersection method [37], whereas some methods [11], [38] project gait templates into a canonical view (i.e., side view) based on the assumption that the human body is well approximated as a planar object on a sagittal plane. These methods are, however, only applicable for

cooperative scenarios with multiple calibrated cameras, or only work for the case in which view differences from the side view are small, where the aforementioned assumption holds [39].

Most generative approaches belong to the example-based category, which learns the transformation between different views based on the training set. Makihara *et al.* [4] proposed a view transformation model (VTM) that applies singular value decomposition (SVD) on frequency-domain features. Subsequently, a variety of VTM-based methods were proposed to improve performance using different algorithms, such as support vector regression (SVR) [12], [13] and multi-layer perceptron [40], and using 3D training gait models [41]. The aforementioned methods may corrupt the geometric continuity of the human body that results in non-humanoid generated gait features; hence, El-Alfy *et al.* [39] proposed a geometric view transformation model (GVTM), which is the only model that considers both geometric deformation and example-based learning to avoid possible corruption in appearance-based gait features.

However, generative methods all require view information to construct a transformation for each view combination in advance. Additionally, they only ensures the optimal generation of gait features rather than recognition accuracy. Furthermore, the learned transformation is generic across the population, which also fails to represent subject individuality in the transformation.

B. Discriminative Approaches to Cross-View Gait Recognition

Different from generative approaches, discriminative approaches use machine learning techniques to directly optimize the discrimination capability without feature registration. One typical method is to apply traditional LDA after dimension reduction using principal component analysis [10]. Rather than LDA, Lu and Tan [42] used uncorrelated discriminant simplex analysis; Mansur *et al.* [43] used multi-view discriminant analysis (MvDA); and Martin-Felez and Xiang [18] exploited rank SVM. Zhang *et al.* [44] proposed discriminative projection with list-wise constraints and rectification (DPLCR) using a new gait representation called the gait individuality image (GII).

Instead of learning a common relatively view-invariant subspace or metric, canonical correlation analysis (CCA) [45] was introduced to project features on two latent subspaces with maximal correlation. Kusakunniran *et al.* [46] further applied correlated motion co-clustering (CMCC) to solve the weakly correlated problem in global gait features, and Xing *et al.* [47] proposed complete canonical correlation analysis (C3A) to improve the performance of CCA for high-dimensional features. Unlike the aforementioned CCA-based methods that need view information, unitary linear view-invariant discriminative projection (ViDP) [48] transfers features into latent space without knowing the view angles.

Although discriminative approaches generally achieve better results than generative approaches, most of them still work poorly for the case of large view variations because it is quite challenging to find robust view-invariant subspaces or metrics with high generalizations for such non-aligned gait features.

C. CNN-Based Approaches to Cross-View Gait Recognition

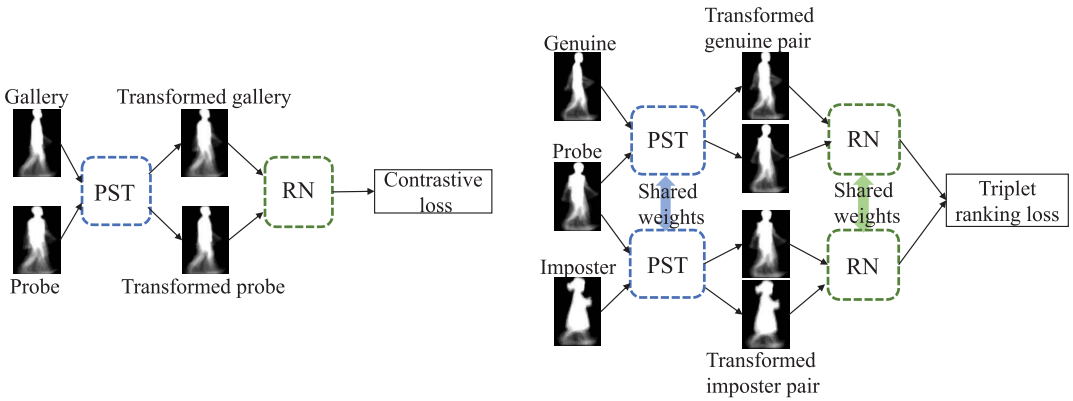
To date, CNN-based approaches have achieved state-of-the-art performance in cross-view gait recognition, where various input features and network structures have been discussed. Wu *et al.* [19] used raw silhouette images and Wolf *et al.* [49] exploited spatio-temporal features as inputs. Reference [23] directly inputted RGB images to disentangle appearance and pose features, and used the latter one for subsequent recognition. However, most CNN-based approaches [5], [14], [15], [20], [50] directly feed the GEI into the networks. GEINet, designed by Shiraga *et al.* [14], is a typical network and has a similar structure to AlexNet [51]. Some approaches [5], [20] have demonstrated that CNN models with two inputs, where the similarities between the two inputs are learned to discriminate whether they are from the same subject or different subjects, achieve better performance than the aforementioned one-input networks. In [15], different CNN architectures were explored to consider the type of recognition task (i.e., verification and identification) and degree of view differences. Zhang *et al.*. [22] proposed a joint network to combine the advantages of using a single input and a pair of inputs with the quintuplet loss function. A recent work named GaitSet [21] regarded gait as a set of independent frames, without considering the order information of silhouette frames in the gait sequence, which achieved prominent gait recognition performance.

Although these approaches have achieved promising results, the networks that only consider discrimination learning without feature registration are still limited in their ability to be invariant to large spatial displacement on the inputs by only using convolutional layers and local pooling layers [27], which may also wash out subtle personal gait characteristics (e.g., body shape and walking style) simultaneously.

Some approaches have used generative adversarial networks (GAN), which generate a gait feature of the common canonical view (i.e., side view) [50] or generate a probe feature of the same view as that of a gallery feature after detecting the view angles [16]. However, GAN-based methods encounter the following two limitations: First, GAN models need to optimize both the feature generation quality and recognition accuracy simultaneously, and hence they suffer from hyperparameter tuning. Second, the generated gait feature may be corrupted because GAN is not a geometric but an example-based approach, and hence the geometric continuity of the human body is never considered.

D. STN

Recently, Jaderberg *et al.* [27] proposed a differentiable module, the ST, to explicitly perform parameterized spatial transformation inside the network by inserting it into any existing CNN architecture, which improves the invariance to significantly large spatial displacement for the network. The optimal transformation for the main task (e.g., classification) is adaptively learned based only on a loss function for the main task in an end-to-end manner, without any extra supervision or modification to the objective function, unlike the GANs' optimization of a loss function, which considers both feature generation and the performance of the main task alternately.



(a) Verification scenario (b) Identification scenario

Fig. 2. Overview of the proposed PSTN framework, which contains a PST and RN. Two networks are designed for two types of gait recognition scenarios according to [15]. (a) is for the gait verification task using contrastive loss with a pair of input GEIs, and (b) is for the gait identification task using triplet loss with triplet input GEIs.

With the illustration of state-of-the-art performance for digit recognition in the original paper of STN [27], STN has been extended to many applications, such as face recognition. Whereas [29] adopted affine transformation, [52] compared the performance of ST based on three types of homogeneous transformations for face alignment. Shi and Jain [31] designed an ST-based attention network to extract both global and local features, and Wu *et al.* [30] proposed a recursive ST to hierarchically model a more complex transformation by performing affine transformation in divided local regions. To render non-rigid transformation, [28] introduced diffeomorphisms into ST, which results in a 'squarification' effect and better performance for face verification.

The aforementioned ST modules mainly aim at transforming a single input image into a common canonical view, which cannot manage gait feature transformation well when the canonical view is not in-between the probe and gallery views or the view difference between the probe and gallery views is considerably large, where the gait behavior in one view cannot been seen in another view.

III. Gait Recognition Using PSTN

A. Overview

An overview of the proposed PSTN framework is shown in Fig. 2. Following most CNN-based approaches [5], [14], [15], [20], [50], GEI is adopted as the input gait feature, which is most widely used for gait recognition studies because it is effective despite its simplicity. Given raw gait videos, silhouette sequences can be first extracted using segmentation methods (e.g., background subtraction-based graph-cut segmentation [53], recent state-of-the-art deep learning-based semantic segmentation methods, such as RefineNet [54], followed by a boundary refinement method such as DenseCRF [55] since the semantic segmentation methods extracts often larger segments than the actual object).[1] The height is then normalized and the region center of the silhouettes is registered, and the gait period is detected based on the autocorrelation of the size-normalized and registered silhouette sequences [4]. Thus, GEI is obtained by averaging the silhouettes over one gait period, which reflects both the static and dynamic (e.g., arm swing and leg motion) gait information.

The proposed CNN framework is composed of two parts: PST and RN. Given an input pair of probe and gallery GEIs, instead of a common transformation used by existing generative approaches, an appropriate sample-dependent geometric transformation is regressed by the PST to transform both the probe and gallery GEIs from different views into their intermediate view, which are further fed into the subsequent RN to obtain the final dissimilarity between this pair. Similar to [15], we design two networks for two types of gait recognition scenarios: verification (i.e., one-to-one matching) and identification (i.e., one-to-many matching). Whereas the verification network uses the contrastive loss, the network for identification adopts the triplet loss using triplet GEIs, which is similar to the Siamese network [20], where the parameters for PST and RN are shared, respectively.

Details for the PST and RN are given in the following sections.

B. PST

Similar to the conventional ST [27], the proposed PST also consists of three components, that is, the localization network, grid generator, and sampler, which is shown in Fig. 3. Instead of affine transformation used in many studies, a non-rigid transformation based on free-form deformation (FFD) [56] is used for the PST because the FFD is suitable for reflecting the transformation of non-rigid objects (e.g., human body) because of its high flexibility. Moreover, the FFD also retains geometric continuity in adjacent regions [56], and hence never corrupts the personalized gait characteristics, whereas such corruption in feature generation may easily occur in existing example-based generative approaches [4], [13].

For example, in case where a test subject walks in an extraordinary manner (e.g., extremely large arm swing or heavy stoop), or owns an extraordinary body shape

[1]We used extracted silhouette sequences released by each dataset provider in our experiments.

(e.g., extremely fat or thin), which are extraordinary gait features never included in the training samples, example-based generative approaches may easily make these extraordinary (or distinctive) gait characteristics disappear and may generate more common gait features instead. Moreover, assuming that a training sample unfortunately contains noise in the background region (e.g., over-segmented isolated foreground regions in background area in GEI) and that a test sample resembles to the training sample with noise by chance, the noise may pop-out for the generated gait feature for the test sample. By contrast, using the geometric transformation with a spatially smooth warping field, the isolated noisy foreground regions never pop-out newly and extraordinary (or distinctive) gait features are likely to be kept to some extent thanks to the property of the geometric continuity, which is more beneficial for the subsequent recognition task.

Therefore, we adopt FFD-based geometric transformation for the PST module, and consider transforming a pair of probe and gallery GEIs from different views into their intermediate view to avoid unnecessary distortion. During the training of the entire PSTN, the PST is supervised by the loss of the following RN, which aims to learn a transformation that achieves a trade-off between intra-subject and inter-subject variations, and further leads to optimal recognition performance. We describe an overview of the FFD framework and details of the three components in the following.

1) Overview of FFD: We first introduce the general representation of FFD in this section. To represent FFD, we first allocate a set of grid-type control points on the GEI [56], [57], as shown in Fig. 4. More specifically, given source GEI $G^s \in \mathbb{R}^{H\times W}$, where H and W are the height and width of the image, respectively, we set n_W control points with interval $\Delta x = (W-1)/(n_W-1)$ and n_H control points with interval $\Delta y = (H-1)/(n_H-1)$ for the horizontal and vertical directions, respectively. The spatial position of the (i, j)-th control point located in the i-th column and j-th row is denoted by $\boldsymbol{p}_{i,j} = [i\Delta x, j\Delta y]^T$ $(i = 0, \ldots, n_W-1, j = 0, \ldots, n_H-1)$.

We then define a set of 2D displacement vectors that indicate a deformation on the control points as $\boldsymbol{u}$, where the displacement vector on the (i, j)-th control point is denoted by $\boldsymbol{u}_{i,j} = [u_{i,j}, v_{i,j}]^T \in \mathbb{R}^2$. Therefore, the position of the (i, j)-th control point is transformed to $\boldsymbol{p}_{i,j} + \boldsymbol{u}_{i,j}$. The entire warping field throughout image $F(\boldsymbol{u})$ is obtained using interpolation from the displacement vectors on the control points, which represents the coordinate correspondence between the source GEI and transformed GEI for each pixel. Finally, the transformed GEI is obtained as $G^t = G^s \circ F(\boldsymbol{u})$, where $\circ$ denotes a transformation operator. In practice, we usually implement a backward (or inverse) warping field (i.e., that from the target image to the source image) instead of a forward warping field (i.e., that from the source image to the target image) when we try to transform a source image to a target image [58].[2]

2) Localization Network: Given an input pair of probe and gallery GEIs, the localization network regresses a transformation parameter vector, that is, a set of displacement vectors on the control points $\boldsymbol{u}$. The transformation parameter vector $\boldsymbol{u}$ is used to define an inverse warping field from the intermediate view to a probe view and gallery view, as described later.

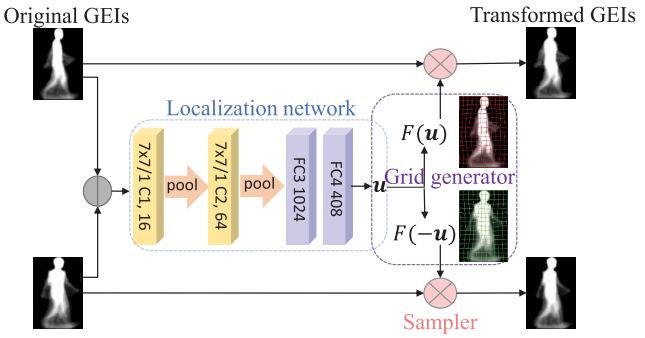


Fig. 3. Structure of the proposed PST, which is composed of the localization network, grid generator, and sampler. In the localization network, C, FC, and pool denote the convolution layer, fully connected layer, and pooling layer, respectively. The digits written before C represent the filter size with the stride, whereas those after C represent the number of filters. The digits after FC represent the number of output neurons. This notation is used to illustrate the network structure throughout this paper.

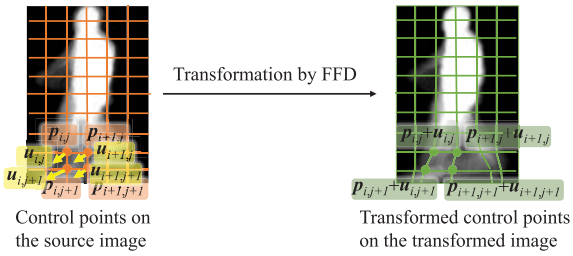


Fig. 4. Illustration of the FFD framework.

The localization network is then designed as a simple CNN whose input is a subtraction image of the original GEI pair and output is transformation parameter vector $\boldsymbol{u}$. Note that both the original probe and gellery GEI can be the minuend for the input of subtraction image. More specifically, the localization network constitutes two convolutional layers, two pooling layers, and two fully connected layers (see Fig. 3). A max pooling with 2×2 pixels with stride 2 is used for the pooling layers, and the rectified linear unit (ReLU) activation function [59] is used for all convolution layers and the first fully connected layer. Additionally, local response normalization (LRN) [60] is applied before the max-pooling layers. Transformation parameter vector $\boldsymbol{u}$ is regressed by the last fully connected layer.

3) Grid Generator: After obtaining transformation parameter vector $\boldsymbol{u}$ from the localization network, the grid generator produces a warping field, which describes the deformation on each pixel. To avoid unnecessary large distortion, we consider transforming both probe and gallery GEIs from the original views to their intermediate view, where the transformation between the intermediate view to each probe and gallery view is symmetrical to each other. Note that the intermediate view does not indicate the physically exact medial view, but an apparent intermediate view derived from the symmetric

[2]Readers may refer to page 25—26 and page 31—35 in Chapter 3 of [58] for more details.

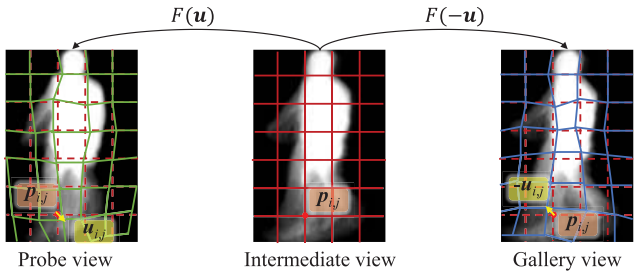


Fig. 5. Illustration of the opposite pair of warping fields $F(\boldsymbol{u})$ and $F(-\boldsymbol{u})$.

transformation for the following two reasons. First, because the training data could be generally collected under arbitrary views, it is infeasible to prepare enough training samples from the physically exact medial view, which are necessary for constructing the transformation from the probe/gallery view to the physically exact medial view. Second, because the warping fields from a probe and a gallery to the physically exact medial view is asymmetric, two different warping fields are required for estimation, which increases the number of parameters for a STN. On the other hand, if the apparent intermediate view is adopted for the proposed PST, only a pair of GEIs without any view constraint (or information) is required for the training process, and only a single warping field is needed for estimation thanks to the property of the symmetric transformation, which saves the number of parameters.

We therefore define warping field $F(\boldsymbol{u})$ from the GEI of the intermediate view (referred to as the intermediate GEI) to the input probe GEI using piecewise linear interpolation from the displacement vectors on the control points, and define a reverse version, $F(-\boldsymbol{u})$, as the warping field from the intermediate GEI to the input gallery GEI. Please also note that we define the intermediate view so as that deformation vectors from the intermediate view to the probe and gallery are just opposite each other, i.e., $\boldsymbol{u}$ and $-\boldsymbol{u}$.

Assume there is a GEI at a virtual intermediate view with the originally regular red grid, as shown in Fig. 5. The probe GEI is obtained by applying the warping field $F(\boldsymbol{u})$, where the spatial position of the (i, j)-th control point $\boldsymbol{p}_{i,j}$ is transformed to the corresponding position in the warped green grid by the displacement vector $\boldsymbol{u}_{i,j}$ (i.e., the transformed position of the control point is $\boldsymbol{p}_{i,j} + \boldsymbol{u}_{i,j}$). Similarly, the gallery GEI is obtained by applying the warping field $F(-\boldsymbol{u})$, where the position of the same control point is transformed to that in the warped blue grid by the displacement vector $-\boldsymbol{u}_{i,j}$ (i.e., the transformed position of the control point is $\boldsymbol{p}_{i,j} - \boldsymbol{u}_{i,j}$), which is just the opposite of $\boldsymbol{u}_{i,j}$. Thus, an entirely symmetric transformation between the intermediate view to each probe and gallery view is guaranteed, which is an important aspect for constructing deformation between two different states.

More specifically, the displacement at position (x, y) on the entire warping field $F(\boldsymbol{u})$ is denoted by $\boldsymbol{f}_{x,y} = [f^u_{x,y}, f^v_{x,y}]^T \in \mathbb{R}^2$ $(x = 0, \ldots, W-1, y = 0, \ldots, H-1)$. First, two general weighting functions are defined as $w_0(x) = 1 - (x - \lfloor x \rfloor)$ and $w_1(x) = x - \lfloor x \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function, and two variables are defined as $\bar{x} = x/\Delta x$ and $\bar{y} = y/\Delta y$. Displacement $\boldsymbol{f}_{x,y}$ is then computed using piecewise linear interpolation from its four neighboring control points as

$$\boldsymbol{f}_{x,y} = \sum_{k=0}^{1}\sum_{l=0}^{1} w_k(\bar{x}) w_l(\bar{y}) \boldsymbol{u}_{\lfloor \bar{x} \rfloor + k, \lfloor \bar{y} \rfloor + l}. \tag{1}$$

The displacement for each pixel on warping field $F(-\boldsymbol{u})$ is obtained in the same manner.

Once the grid generator is defined as above, the gradient of $f^d_{x,y}$ $(d \in \{u, v\})$ is also computed with respect to each $d_{\lfloor \bar{x} \rfloor + k, \lfloor \bar{y} \rfloor + l}$ used during the back-propagation process as

$$\frac{\partial f^d_{x,y}}{\partial d_{\lfloor \bar{x} \rfloor + k, \lfloor \bar{y} \rfloor + l}} = w_k(\bar{x}) w_l(\bar{y}), \tag{2}$$

where $k, l \in \{0, 1\}$.

4) Sampler: As the final procedure in the PST, the sampler generates the output image pairs by transforming each of input probe and gallery GEIs, respectively. Let the input probe and gallery GEIs be $G^i_p, G^i_g \in \mathbb{R}^{H \times W}$, and transformed probe and gallery GEIs be $G^o_p, G^o_g \in \mathbb{R}^{H \times W}$, where $G^i_p = G^o_p \circ F(\boldsymbol{u})$, $G^i_g = G^o_g \circ F(-\boldsymbol{u})$. For pixel (x^o_p, y^o_p) of the output G^o_p, its corresponding source coordinates in the input are obtained as $(x^i_p, y^i_p) = (x^o_p + f^u_{x^o_p, y^o_p}, y^o_p + f^v_{x^o_p, y^o_p})$, where $f^u_{x^o_p, y^o_p}$ and $f^v_{x^o_p, y^o_p}$ are the horizontal and vertical displacements, respectively, for this pixel according to warping field $F(\boldsymbol{u})$. The intensity value $I^o_{x^o_p, y^o_p}$ at position (x^o_p, y^o_p) is sampled using bilinear interpolation as

$$I^o_{x^o_p, y^o_p} = \sum_{k=0}^{1}\sum_{l=0}^{1} w_k(x^i_p) w_l(y^i_p) I^i_{\lfloor x^i_p \rfloor + k, \lfloor y^i_p \rfloor + l}, \tag{3}$$

where $I^i_{\lfloor x^i_p \rfloor + k, \lfloor y^i_p \rfloor + l}$ is the intensity at position $(\lfloor x^i_p \rfloor + k, \lfloor y^i_p \rfloor + l)$ on original probe G^i_p, which indicates the four nearest pixels to spatial location (x^i_p, y^i_p).

The partial derivatives of $I^o_{x^o_p, y^o_p}$ with respect to its related displacements $f^u_{x^o_p, y^o_p}$ and $f^v_{x^o_p, y^o_p}$ are given as

$$\begin{aligned}
\frac{\partial I^o_{x^o_p, y^o_p}}{\partial f^u_{x^o_p, y^o_p}} &= \frac{\partial I^o_{x^o_p, y^o_p}}{\partial x^i_p} = \sum_{k=0}^{1}\sum_{l=0}^{1} c_k w_l(y^i_p) I^i_{\lfloor x^i_p \rfloor + k, \lfloor y^i_p \rfloor + l}, \\
\frac{\partial I^o_{x^o_p, y^o_p}}{\partial f^v_{x^o_p, y^o_p}} &= \frac{\partial I^o_{x^o_p, y^o_p}}{\partial y^i_p} = \sum_{k=0}^{1}\sum_{l=0}^{1} c_l w_k(x^i_p) I^i_{\lfloor x^i_p \rfloor + k, \lfloor y^i_p \rfloor + l},
\end{aligned} \tag{4}$$

where the coefficients c_k and c_l are defined as

$$c_k = \begin{cases} 1 & k = 1 \\ -1 & k = 0, \end{cases} \quad c_l = \begin{cases} 1 & l = 1 \\ -1 & l = 0. \end{cases} \tag{5}$$

The forward and back-propagation are executed similarly for the gallery. Consequently, the loss gradients are enabled to flow back throughout the entire network, from the RN to the localization network in PST.

C. RNs Considering the Recognition Scenarios and View Variation Degree

The transformed probe and gallery GEI pairs from the PST are subsequently fed into the RN for discriminative feature learning. Because the proposed PST could be freely combined with any CNN model, we choose four state-of-the-art network

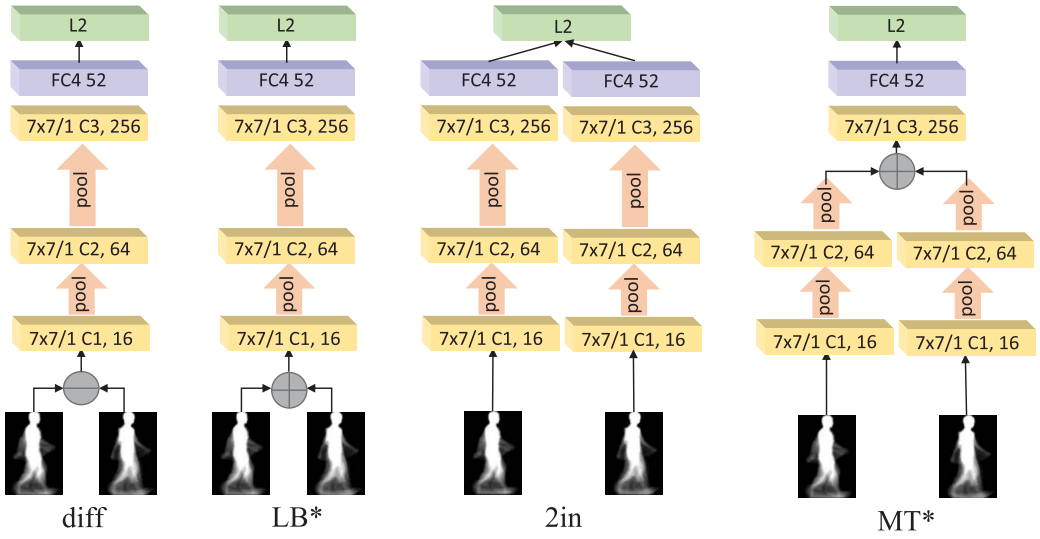


Fig. 6. RNs with four different architectures. The transformed probe and gallery GEI pairs are used as the inputs. L2 indicates the L2 norm/distance of the one/two output from the FC4 layer. diff/2in directly takes the difference between the input GEI pair/output pair of the FC4 layer, whereas LB*/MT* uses pair filters to compute the pixel-wise weighted sum of the two inputs/features at the C3 layer. 2in is a Siamese network, and MT* also shares parameters in the first two convolutional layers.

architectures for gait recognition as the RNs to investigate the performance of PSTN.

The structures of the four RNs, diff, 2in [15], LB*, and MT*, are shown in Fig. 6. As introduced in Section III-C2, contrastive loss and triplet loss are more suitable for gait verification and identification, respectively [15]; therefore, we modify the structures of the original LB and MT [5] by replacing cross-entropy loss with contrastive loss and triplet loss, respectively, and changing the output dimension of the last fully connected layer accordingly (denoted by LB* and MT*). As a result, the four RNs share the same basic architecture, which is composed of three convolutional layers and two max-pooling layers with 2×2 pixels with stride 2, in addition to a fully connected layer. The ReLU activation function [59] is applied to all convolution layers and fully connected layers. The LRN [60] is used before the max-pooling layers and the dropout technique [61] is used for the FC4 layer. Finally, the L2 norm/distance of the one/two output from the FC4 layer is computed, which is also considered as the dissimilarity score between the input pair of probe and gallery GEIs during the test stage. The differences among these four RNs will be discussed in detail in Section III-C1.

1) RNs Considering the View Variation Degree: To successfully recognize that the input GEI pairs are from the same subject or different subjects, it is necessary to consider a trade-off between the inter-subject differences and intra-subject differences that result from the view variations; hence, different network architectures are required that depend on the degree of appearance change caused by different view angles [15].

The structures diff and LB* match (i.e., determines the difference) the input transformed GEI pair at the initial stage (i.e., first layer) of the networks, which makes the models more sensitive to local differences that arise from both intra-subject and inter-subject variations, and hence they are more suitable for the case of relatively smaller view variations, where the inter-subject variations are larger than the intra-subject differences. Whereas diff directly takes the difference between the input pair before feeding it into the network, LB* simulates the subtraction by calculating the pixel-wise weighted sum using the pair filters [5].

By contrast, 2in and MT* compare the two inputs in the higher level of the networks, and hence allow more spatially invariant features to be extracted before the matching process. Therefore, 2in and MT* are more favorable for the case of relatively larger view variations because the intra-subject spatial differences are larger than the inter-subject variations because of the large view differences. Similar to the difference between diff and LB*, 2in computes the subtraction between the outputs of the last FC4 layer with a Siamese network [20] (i.e., weights are shared between two columns), whereas MT* takes the weighted differences between the features learned at the C3 layer, which is also based on the Siamese network at the bottom two convolutional layers.

2) RNs Considering the Recognition Scenario: Two scenarios are considered in gait recognition: verification and identification. In the verification scenario, a pair of probe and gallery GEIs is provided to assess whether the GEIs are from the same subject by comparing their dissimilarity score with an acceptance threshold. Regarding gait identification, a probe is compared with all the galleries to locate the genuine GEI derived from the same subject as the probe, which is determined by calculating the smallest dissimilarity score using the nearest neighbor classifier.

To successfully discriminate the input GEI pairs in the verification scenario, it is necessary for the absolute dissimilarity scores of the same subject pairs to be smaller than those of the different subject pairs, which coincides with the definition of the contrastive loss function as [62]

$$L_{\text{con}} = \frac{1}{2N} \sum_{n=1}^{N} (\alpha_n d_n^2 + (1 - \alpha_n) \max(\text{margin} - d_n, 0)^2), \quad (6)$$

where N is the number of training GEI pairs and d_n is the dissimilarity score (i.e., L2 norm/distance of the one/two outputs from the FC4 layer) of the n-th GEI pair. α_n is set to one when the GEIs in the n-th pair originate from the same subject, and zero otherwise. Consequently, the contrastive loss function is suitable to be used for the proposed PSTN in the gait verification scenario.

In the identification scenario, the dissimilarity score between the probe and genuine (i.e., true match in the gallery) GEI is required to be relatively smaller than that of the probe and imposters (i.e., false match in the gallery) to obtain a correct match. To achieve this, a triplet that characterizes a relative dissimilarity ranking order for the three GEI images, that is, probe, genuine, and imposter, is adopted as the input of the entire proposed framework, which is similar to [15]. Correspondingly, two parallel PSTs and RNs with respectively shared weights are designed for the input genuine pair and imposter pair, as shown on the right of Fig. 2. A triplet loss function is used for the proposed PSTN, which is defined as [63]

$$L_{\text{tri}} = \frac{1}{2N}\sum_{n=1}^{N}\max(\text{margin} - d_{\text{imp}}^{n} + d_{\text{gen}}^{n}, 0)^2, \qquad (7)$$

where N is the number of training GEI triplets, d_{imp}^{n} is the dissimilarity score of the imposter pair, and d_{gen}^{n} is that of the genuine pair for the n-th input triplet. As a result, $d_{\text{gen}}^{n}(n = 1, \ldots, N)$ is trained to be relatively smaller than d_{imp}^{n}, which satisfies the requirement of an accurate match for the identification task.

D. Training Process

To boost the performance of the proposed PSTN, we first pre-trained the PST part using a subset of the whole training set without any additional training data, and then used the pre-trained model as the initialization for the PST part to fine-tune the whole PSTN.

More specifically, the PST part was first pre-trained using only the same subject pairs from two randomly selected view angles, which aimed to optimize the deformation effects applied by the PST. Concretely speaking, we minimized the differences between the transformed training GEI pairs using the Euclidean loss. To smooth the output images, we additionally defined a regularizer loss with coefficient λ to ensure the spatial consistency between displacements at adjacent control points [57]. Thus, the Euclidean loss and regularizer loss with its coefficient constituted the total loss for pre-training the PST.

Once the PST is pre-trained, we first used the pre-trained PST model to initialize the weights of the PST part in the PSTN, and then fine-tuned the entire PSTN only with the contrastive/triplet loss introduced in Section III-C2 in an end-to-end manner, where the deformations $F(\boldsymbol{u})$ and $F(-\boldsymbol{u})$ were again simultaneously modified to obtain optimal recognition accuracy.

IV. Experiments

A. Datasets

We evaluated the proposed method on three publicly available datasets: OU-MVLP [32], OULP [33], and CASIA-B [34].

OU-MVLP is currently the world's largest gait database with wide view variation, and was collected in conjunction with an experience-based long-run exhibition at a science museum (i.e., Miraikan). The dataset contains 10,307 subjects captured from 14 view angles, ranging from 0° to 90° and 180° to 270° in 15° intervals. Examples of GEIs from each view angle can be found in Fig. 1. Two sequences (i.e., probe and gallery sequences) are provided for each subject from each view angle. This dataset was used for all our experiments to make the performance evaluation more statistically reliable. Following the protocol of the dataset [32], 5,153 subjects were used for training, and the other disjoint 5,154 subjects were used for testing. Based on the perspective projection assumption [64], GEIs with view angles over 180° were flipped right-to-left to roughly align the walking direction in the GEIs, which is easier to estimate (i.e., leftward or rightward) compared with the exact view angle estimation [15]. In the training phase, GEIs from all view angles were fed into the PSTN simultaneously, and in the test stage, following [15], performance was evaluated for each combination of four typical views: 0°, 30°, 60°, and 90°.

The OULP dataset is the second largest gait dataset and consists of over 4,000 subjects with four different views: 55°, 65°, 75°, and 85°. Similar to OU-MVLP, both probe and gallery sequences are provided for each view angle. To compare our results with the benchmarks [14], [39], [43], [65], [66] in Section IV-F, the same subset that they used was chosen, which comprised 1,912 subjects. The subset was further divided into two disjoint sets of equal size, which were used for training and testing separately.

The CASIA-B dataset includes a relatively small number of subjects, that is, 124, but also has a wide view variation. Eleven view angles ranging from 0° to 180° in 18° intervals with six normal walking sequences (NM #01-06) per view are provided for each subject. This dataset allows us to evaluate performance for low-quality gait silhouettes with segmentation errors. The same protocol protocol as [5], [16] was used for the experiment in Section IV-G. Specifically, the first 74 subjects were used for training, and performance was evaluated using the remaining 50 subjects, where four sequences (NM #01-04) were chosen as galleries, whereas the other two sequences (NM #05-06) were used as probes.

B. Implementation Details

We initialized the weight parameters of all layers using Xavier's algorithm [67], except for the last fully connected layer in the PST, which was initialized to zero. The bias terms were all set to zero initially. The momentum for weights and bias terms was 0.9, and the weight decay was zero. The network parameters were trained using the stochastic gradient descent algorithm [68] with a min-batch size of 600. Basically, the initial learning rate was set to 0.001 for the PST and 0.01 for the RN, which were both divided by 10 four times during the training stage. The proportion of dropping neurons was set to 0.5 for the dropout technique applied in the last layer of the RN. The hyperparameters of the margin in Eqs. (6) and (7) were both set to 3 empirically.

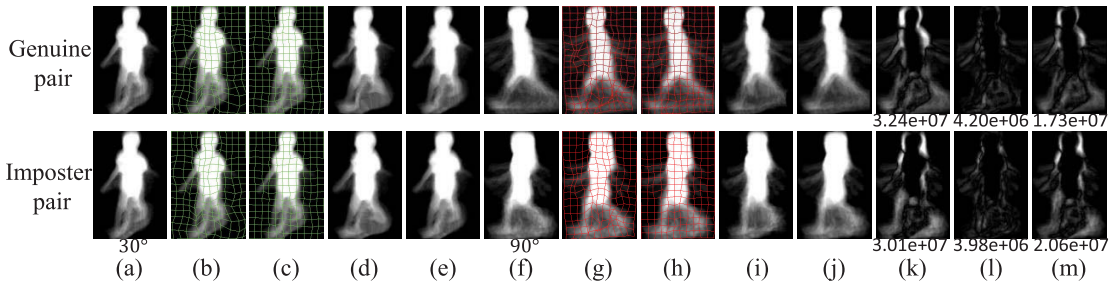


Fig. 7. Examples of transformed GEIs from the fine-tuned PSTN (PST-LB*) and pre-trained PST. The first and second row show the learned transformations for a genuine pair and imposter pair, respectively. (a) Original probe GEI from 30°. (b) Probe warping field learned by pre-trained PST. (c) Probe warping field learned by PSTN. (d) Transformed probe using pre-trained PST. (e) Transformed probe using PSTN. (f) Original gallery GEI from 90°. (g) Gallery warping field learned by pre-trained PST. (h) Gallery warping field learned by PSTN. (i) Transformed gallery using pre-trained PST. (j) Transformed gallery using PSTN. (k) Absolute difference image and corresponding Euclidean distance between original GEIs (a) and (f). (l) Absolute difference image and Euclidean distance between transformed GEIs (d) and (i) using pre-trained PST. (m) Absolute difference image and Euclidean distance between transformed GEIs (e) and (j) using PSTN.

Regarding the constitution of training sample pairs for the gait verification task, we used all the same subject pairs by randomly selecting two view angles, and randomly chose parts from all different subject pairs to maintain the ratio of the number of same subject pairs to that of different subject pairs equal to 1 : 9. To make up the training triplets for the identification scenario, we randomly selected approximately 30 million triplets from all possible sample combinations. Considering the quite limited training subjects in the CASIA-B dataset, we used all the pairs[3] and triplets to increase the amount of training data.

To evaluate the recognition performance, we computed the equal error rate (EER) of the false acceptance rate and false rejection rate to measure accuracy in the verification scenario, and used the rank-1 identification rate as the evaluation criterion for the identification scenario.

C. Visualization of PST

We first visualize the transformation learned by the proposed PST module using the examples of a genuine pair and imposter pair with the same probe to illustrate the sample-dependence of the transformation. We choose a case of relatively larger view variation, where the probe and two galleries are from 30° and 90°, respectively. To better understand the trade-off between the intra-subject and inter-subject differences, we show both the transformation learned by the pre-trained PST and the entire fine-tuned PSTN (we used LB* as the RN as an example, denoted by PST-LB*) in Fig. 7.

Comparing the two corresponding learned warping fields for the genuine pair and imposter pair (Fig. 7(b)(c) and (g)(h)), it is obvious that the learned transformations vary for different sample pairs, which makes it possible to generate a transformation that represents individual gait characteristics using a single CNN model rather than a common deformation typically applied by traditional methods [4], [39]. Because of the view variation, both the original GEI pairs show large image differences (Fig. 7(k)). By transforming the original GEIs with the learned deformation fields into the intermediate view (Fig. 7(d)(e)(i)(j)), such differences are reduced by both the pre-trained PST (Fig. 7(l)) and fine-tuned PSTN (Fig. 7(m)).

As a result, the pre-trained PST significantly reduces the intra-subject difference of the genuine pair by transforming the probe and gallery into almost the same intermediate view. The imposter pair, however, is still even more similar than the genuine pair after the transformation, which degrades the discrimination capability (e.g., body shape change between Figs. 7(a) and (d) or (f) and (i) in the second row), and may further result in a false match in the subsequent recognition task (larger difference for the genuine pair than the imposter pair in Fig. 7(l)). This means that the pre-trained PST alone may risk overly registering the GEI pairs regardless of the difference derived from the intra-subject view variation or the inter-subject variation.

By contrast, the proposed PSTN fine-tuned by the recognition loss (i.e., contrastive/triplet loss) somewhat weakens the registration effect and hence makes dissimilarity measures both for the genuine pair and imposter pair larger than those by the pre-trained PST. Importantly, we notice that the dissimilarity for the imposter pair increases more than that for the genuine pair by changing the pre-trained PST to the fine-tuned PSTN, and consequently the dissimilarity for the imposter pair becomes larger than that for the genuine pair (Fig. 7(m)). This implies that the PSTN fine-tuned by the main recognition task works to reduce the intra-subject difference caused by view variation while not unnecessarily reducing the inter-subject difference (e.g., body shape change observed between Figs. 7(a) and (d) or (f) and (i)), which are two conflicting aspects; that is, the fine-tuned PSTN achieves a good trade-off between reducing the intra-subject difference caused by view variations and maintaining the inter-subject difference, unlike the pre-trained PST, which tends to overly register, and the baseline without PST, which does nothing regarding registration.

D. Effect of PST

To confirm the effectiveness of the proposed PST module, we compared the recognition performance between the

[3]The same subject pairs were duplicated to keep the ratio of the number of same subject pairs to that of different subject pairs still equal to 1 : 9.

TABLE I

COMPARISON BETWEEN THE RNS W/O AND W/ A PST MODULE FOR BOTH VERIFICATION AND IDENTIFICATION SCENARIOS USING OU-MVLP. EACH ROW IN THE TABLE SHOWS THE RESULTS OF THE PURE RN (BEFORE SLASH) AND THAT COMBINED WITH THE PST (AFTER SLASH). BOLD AND ITALIC BOLD INDICATE THE BEST AND SECOND-BEST RESULTS FOR EACH ANGULAR DIFFERENCE, RESPECTIVELY. THIS FONT CONVENTION IS USED TO INDICATE PERFORMANCE THROUGHOUT THIS PAPER

(a) Mean EER (%) for each angular difference and the total mean EER.

Methods	Angular difference				
	0°	30°	60°	90°	Mean
diff [15]/ PST-diff	1.1/0.9	3.0/2.3	5.7/4.5	7.2/5.8	3.7/2.9
2in [15]/ PST-2in	1.3/1.2	2.4/2.1	3.5/3.1	4.4/***3.9***	2.6/2.4
LB*/ PST-LB*	0.9/**0.6**	2.9/1.9	5.8/4.3	7.7/5.4	3.8/2.6
MT*/ PST-MT*	1.1/***0.7***	2.9/1.8	5.5/3.7	7.1/4.9	3.6/2.4
diff+2in [15]/ PST-diff+PST-2in	1.0/***0.7***	2.0/***1.6***	3.4/***2.9***	4.2/**3.7**	2.4/***2.0***
LB*+MT*/ PST-LB*+PST-MT*	0.8/**0.6**	2.5/***1.6***	5.1/3.6	6.7/4.6	3.3/2.2
LB*+2in/ PST-LB*+PST-2in	***0.7***/**0.6**	1.9/**1.5**	3.6/**2.8**	4.6/**3.7**	2.4/**1.9**
diff+MT*/ PST-diff+PST-MT*	0.9/**0.6**	2.4/1.7	5.0/3.5	6.3/4.7	3.2/2.3

(b) Mean rank-1 identification rate (%) for each angular difference and the total mean rank-1 rate.

Methods	Angular difference				
	0°	30°	60°	90°	Mean
2diff [15]/ PST-2diff	91.1/92.3	46.0/52.6	20.2/25.3	9.5/12.9	46.2/50.8
3in [15]/ PST-4in	88.8/90.0	54.7/60.3	31.3/36.6	19.1/23.4	52.9/57.2
2LB*/ PST-2LB*	90.2/92.4	49.1/58.1	23.5/30.2	11.0/15.5	48.2/54.4
2MT*/ PST-2MT*	88.9/91.1	53.1/60.7	27.9/35.4	14.8/20.7	51.0/57.0
2diff+3in [15]/ PST-2diff+PST-4in	92.1/**94.4**	62.0/***68.6***	34.9/***40.6***	20.3/***25.4***	57.5/***62.7***
2LB*+2MT*/ PST-2LB*+PST-2MT*	92.1/94.1	57.9/66.4	30.2/38.3	15.7/22.1	54.2/60.7
2LB*+3in/ PST-2LB*+PST-4in	92.7/93.9	62.6/**69.2**	35.4/**41.9**	20.3/**25.9**	58.0/**63.1**
2diff+2MT*/ PST-2diff+PST-2MT*	93.1/***94.3***	58.3/65.2	30.0/36.8	15.5/21.4	54.6/59.9

RNs without and with the PST module (i.e., PSTN) for the same parameter settings. More specifically, for the verification scenario, we used the EER to evaluate all the four RNs introduced in Section III-C1, that is, diff, 2in, LB*, and MT*, and those combined with the PST, which are denoted by PST-diff, PST-2in, PST-LB*, and PST-MT*, respectively. For the identification scenario, we computed the rank-1 identification rate to compare the performance of parallel RNs using the triplet loss, that is, 2diff, 3in [15], 2LB* and 2MT*, with each including the PST module as PST-2diff, PST-4in, PST-2LB*, and PST-2MT*, respectively. Similar to [5], [15], we additionally created a score-level fusion between a pair of networks suitable for small and large view variation (e.g., diff and 2in) to further improve the recognition accuracy, which was performed by simply averaging the L2 distances output from the individual CNN models.

We conducted the experiments on OU-MVLP because of its statistical reliability in terms of both the number of subjects and view variations. Following [15], we also report the mean results for each angular difference based on the full combinations of four typical view angles, 0°, 30°, 60°, and 90°, in addition to the total mean results for each method, as shown in Table I.

Comparing the results of the RNs with the corresponding PSTNs, it is clear that the recognition performance improves when the proposed PST module is combined with any RN for both the verification and identification scenarios, which demonstrates the effectiveness of integrating the PST into a CNN framework for cross-view gait recognition. The proposed PSTN gains more significant improvement for the case of larger view variation (e.g., 60° and 90° view difference), for which it is more difficult for the pure RNs to extract spatially invariant features. By registering both the input GEIs from two views into an appropriate intermediate view, the entire PSTNs can be more invariant to large spatial displacements raised by considerable view variations. Additionally, it is interesting to find that the proposed PST also slightly improves the results for the same view case, where the original GEI pairs with some posture change (e.g., looking down in the probe but walking normally in the gallery sequence) are also effectively aligned by the PST for better matching.

TABLE II

EER (%) OF PST-LB*+PST-2in (BEFORE SLASH) AND RANK-1 IDENTIFICATION RATE (%) OF PST-2LB*+PST-4in (AFTER SLASH) FOR EACH INDIVIDUAL COMBINATION OF FOUR TYPICAL VIEW ANGLES. PROBE AND GALLERY ARE DENOTED BY P AND G, RESPECTIVELY

G \ P	0°	30°	60°	90°	Mean
0°	0.9/89.5	1.8/61.8	3.9/27.3	3.5/27.5	2.5/51.5
30°	2.1/57.0	0.5/95.7	1.3/72.0	1.4/58.3	1.3/70.8
60°	4.4/23.9	1.3/73.0	0.6/94.1	1.2/75.7	1.9/66.7
90°	4.0/24.3	1.6/58.1	1.2/75.8	0.4/96.1	1.8/63.6
Mean	2.9/48.7	1.3/72.1	1.7/67.3	1.6/64.4	1.9/63.1

On the other hand, the models with the high-level matching structure (e.g., PST-2in) achieve better performance than the low-level matching structure (e.g., PST-diff) for the case of larger view differences, whereas the latter models are more effective than the former in the scenario of smaller view variations, which is consistent with the analysis in Section III-C1, in addition to insight from [15]. Additionally,

TABLE III

COMPARISON OF THE PROPOSED METHOD WITH OTHER STATE-OF-THE-ART METHODS ON OU-MVLP

(a) EER (%)

Methods	Angular difference				
	0°	30°	60°	90°	Mean
DM	6.5	25.2	41.4	46.2	27.2
LDA [65]	6.2	22.7	35.7	40.1	24.0
VTM [4]	6.5	26.8	34.2	38.5	25.0
GEINet [14]	2.4	5.9	12.7	17.2	8.1
Original LB [5]	1.0	3.3	6.7	9.3	4.3
Original MT [5]	***0.9***	2.5	5.2	7.0	3.3
diff [15]	1.1	3.0	5.7	7.2	3.7
2in [15]	1.3	2.4	3.5	4.4	2.6
diff+2in [15]	1.0	2.0	3.4	4.2	***2.4***
PST-LB* (proposed)	**0.6**	***1.9***	4.3	5.4	2.6
PST-2in (proposed)	1.2	2.1	***3.1***	***3.9***	***2.4***
PST-LB*+PST-2in (proposed)	**0.6**	**1.5**	**2.8**	**3.7**	**1.9**

(b) Rank-1 identification rate (%)

Methods	Angular difference				
	0°	30°	60°	90°	Mean
DM	77.4	2.4	0.2	0.0	20.3
LDA [65]	81.6	10.1	0.8	0.1	24.4
VTM [4]	77.4	2.7	0.6	0.2	20.5
GEINet [14]	85.7	40.3	13.8	5.4	40.7
Original LB [5]	89.9	42.2	15.2	4.5	42.6
Original MT [5]	89.3	49.0	20.9	8.2	46.9
2diff [15]	89.1	40.8	17.6	7.8	42.9
3in [15]	85.7	47.8	26.3	15.9	47.9
2diff+3in [15]	89.5	55.0	30.0	17.3	52.7
PST-2LB* (proposed)	***92.4***	58.1	30.2	15.5	54.4
PST-4in (proposed)	90.0	***60.3***	***36.6***	***23.4***	***57.2***
PST-2LB*+PST-4in (proposed)	**93.9**	**69.2**	**41.9**	**25.9**	**63.1**

it is understandable that the improvement for PST-diff is relatively larger than that for PST-2in because the former is more sensitive to spatial variations and hence easier to improve by involving spatial transformation before feature learning. Finally, the fusions of high-level and low-level matching networks all obtain better results than those of using a single CNN model. Given that the fusions of PST-LB* + PST-2in and ST-2LB* + PST-4in yield the best performance for verification and identification, respectively, we also provide their EER/rank-1 identification rates for each individual combination of four view angles in Table II, and only report the results based on this four related networks for the following experiments.

E. Comparison on OU-MVLP

In this section, the proposed method is compared with the state-of-the-art methods on OU-MVLP. In addition to the benchmark of the generative approach, that is, VTM [4][4] and one typical discriminative approach, that is, LDA [65], we also provide the results of the baseline, that is, direct matching (DM) between the original GEI image pairs, and

[4]The exact view angle information is required for VTM to generate a transformation model for each view pair.

TABLE IV

COMPARISON OF THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS ON OULP. GALLERY VIEW IS FIXED TO 85°

(a) EER (%)

Methods	Probe view				
	55°	65°	75°	85°	Mean
DM	30	14	4	4	13
LDA [65]	8	5	4	-	-
GMLDA [66]	12	9	5	-	-
MvDA [43]	7	5	4	-	-
GVTM [39]	4	3	2	-	-
GEINet [14]	2.7	1.8	1.0	-	-
Original LB [5]	0.84	0.55	0.42	0.42	0.56
Original MT [5]	0.90	***0.40***	0.42	0.35	0.52
diff [15]	0.72	0.42	0.42	0.37	0.48
2in [15]	0.94	0.46	0.33	***0.31***	0.51
diff+2in [15]	***0.64***	**0.31**	***0.25***	***0.31***	***0.38***
PST-LB* (proposed)	0.73	0.42	**0.21**	***0.31***	0.42
PST-2in (proposed)	0.80	0.42	0.31	***0.31***	0.46
PST-LB*+PST-2in (proposed)	**0.50**	**0.31**	**0.21**	**0.21**	**0.31**

(b) Rank-1 identification rate (%)

Methods	Probe view				
	55°	65°	75°	85°	Mean
DM	2	16	81	92	48
LDA [65]	56	91	96	-	-
GMLDA [66]	68	82	95	-	-
MvDA [43]	88	96	97	-	-
GVTM [39]	92	96	98	-	-
GEINet [14]	80.4	91.5	94.8	-	-
Original LB [5]	91.95	98.22	99.27	98.95	97.10
Original MT [5]	89.33	97.38	98.12	98.74	95.89
2diff [15]	95.29	98.64	99.16	99.37	98.12
3in [15]	89.12	96.13	98.01	98.74	95.50
2diff+3in [15]	95.71	98.54	99.16	***99.69***	98.27
PST-2LB* (proposed)	***96.34***	***99.06***	**99.79**	**99.90**	***98.77***
PST-4in (proposed)	91.21	96.65	97.91	98.95	96.18
PST-2LB*+PST-4in (proposed)	**97.18**	**99.27**	***99.37***	**99.90**	**98.93**

those of the state-of-the art CNN-based methods, that is, GEINet [14], original LB and MT [5], in addition to the results of diff/2diff, 2in/3in, and the fusion of them,[5] which are all originally from [15]. The mean EERs and rank-1 identifications for each angular difference in addition to the total mean over all the combinations of four view angles are shown in Table III.

The CNN-based methods clearly outperform the traditional generative and discriminative approaches both in terms of verification and identification scenarios. The networks that apply contrastive loss (e.g., diff) and triplet loss (e.g., 2diff) obtain better results than those using cross-entropy loss (e.g., original LB), particularly for the case of larger view differences. Among all methods using a single CNN model, the proposed PST-LB*/2LB* and PST-2in/4in yield the best performance for recognition under small and large view variations,

[5]We did not compare with [21] because it takes a set of silhouette images as the inputs, where the PST module cannot be directly applied since both view registration and phase registration for each frame need to be considered; therefore, we mainly focus on the comparison with GEI-based methods.

TABLE V

COMPARISON OF THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS ON CASIA-B. THE MEAN RESULT OVER ALL 10 GALLERY VIEWS FOR EACH PROBE VIEW IS SHOWN, WHERE THE IDENTICAL VIEW IS EXCLUDED FOR THE GALLERY

(a) EER (%)

Methods	Probe view											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
DM	36.1	31.9	28.9	29.0	33.1	32.7	30.8	29.2	31.0	32.3	35.2	31.8
diff [15]	6.4	4.8	4.4	4.2	4.7	6.2	5.6	4.5	3.8	5.1	7.3	5.2
2in [15]	6.8	5.6	4.7	3.8	4.7	4.6	5.3	4.4	4.6	5.3	7.8	5.2
diff+2in [15]	5.5	4.4	3.7	3.4	3.8	4.6	4.3	3.4	3.1	4.3	6.0	4.2
PST-LB* (proposed)	**4.6**	**2.8**	***3.2***	***2.7***	***3.4***	***3.8***	***3.4***	***3.1***	**2.4**	***3.0***	**4.6**	***3.4***
PST-2in (proposed)	6.6	5.4	3.7	3.8	4.0	4.6	4.3	4.2	4.0	3.5	***5.2***	4.5
PST-LB*+PST-2in (proposed)	***5.3***	***3.7***	**2.8**	**2.6**	**2.9**	**3.0**	**3.0**	**2.7**	***2.6***	**2.5**	**4.6**	**3.2**

(b) Rank-1 identification rate (%). GEI Ensemble and GEI-temporal Ensemble indicate the Ensemble of networks with GEI and Ensemble of networks with GEI and temporal information in [5], respectively.

Methods	Probe view											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
DM	17.0	22.0	22.0	22.4	24.0	26.4	27.2	24.6	22.6	25.0	15.6	22.6
ViDP [48]	-	-	-	64.2	-	60.4	-	65.0	-	-	-	-
MGAN [16]	-	-	-	84.2	-	72.3	-	83.0	-	-	-	-
Original LB [5]	79.1	88.4	95.7	92.8	89.1	87.0	89.3	92.1	94.4	89.4	75.4	88.4
Original MT [5]	***87.7***	92.0	95.3	94.2	89.9	86.5	90.2	95.0	96.5	92.9	82.9	91.2
GEI Ensemble [5]	***87.7***	93.3	***97.3***	***95.6***	***93.4***	90.5	***92.9***	***96.2***	***97.5***	93.8	***85.1***	***93.0***
GEI-temporal Ensemble [5]	**88.7**	**95.1**	**98.2**	**96.4**	**94.1**	***91.5***	**93.9**	**97.5**	**98.4**	***95.8***	**85.6**	**94.1**
2diff [15]	78.6	89.6	93.0	91.4	86.8	85.0	86.8	91.0	93.6	91.0	77.8	87.7
3in [15]	81.4	89.8	93.2	90.4	89.0	85.4	87.0	90.0	92.0	92.6	78.8	88.1
2diff+3in [15]	83.2	91.2	95.8	93.4	91.2	87.8	89.4	93.6	96.0	***95.8***	81.6	90.8
PST-2LB* (proposed)	83.6	91.4	93.6	91.6	91.0	89.2	89.0	94.6	95.8	91.6	79.6	90.1
PST-4in (proposed)	83.2	91.0	93.2	92.2	89.6	86.2	89.2	91.8	90.8	93.2	81.4	89.3
PST-2LB*+PST-4in (proposed)	87.0	***93.8***	96.2	94.4	92.2	**91.8**	92.0	95.0	96.0	**96.4**	84.8	92.7

respectively, whereas their fusion achieves the best results for all degrees of view angle difference.

F. Comparison on OULP

In addition to the DM, LDA, and CNN-based methods mentioned in the previous section, the proposed method was also compared with several additional benchmarks of traditional methods: generalized multiview LDA (GMLDA) [66], MvDA [43], and GVTM [39] using OULP. The performance was evaluated using the same settings as [39], [43], where the gallery view was fixed to 85°, whereas the probe included all four view angles. We implemented the original LB/MT [5] and diff/2in/2diff/3in [15] for the performance evaluation using this protocol.

As shown in Table IV, the proposed PSTN achieves much better results than the GVTM, which is the generative approach using a common non-rigid deformation between each pair of different view angles. Therefore, an automatically learned sample-dependent transformation is demonstrated to be more effective than a common one for achieving high recognition accuracy. Because OULP contains relatively small view variations (i.e., maximum of 30°), which is more suitable for using networks with a low-level matching structure, those matching the sample pairs at a high level achieve relatively worse performance. Generally, most of the CNN-based methods obtain almost saturated performance (less than 1% EER and almost 100% rank-1 rate), and the proposed PST-LB*+PST-2in and PST-2LB*+PST-4in yield the smallest EER and highest rank-1 identification rate, respectively.

G. Comparison on CASIA-B

We finally compared the methods on CASIA-B, which is another dataset that includes wide view variation, but a small number of subjects with low-quality silhouette images. Considering the large amount of network parameters in our models, we used the protocol containing relatively more training data, i.e., 74 training subjects, which was also utilized in [5] and [16]. To compare with the proposed method, we implemented diff/2in and 2diff/3in using the same protocol. Additionally, we selected the state-of-the-art traditional method, which was also evaluated using the same dataset settings, that is, ViDP [48], and several typical networks in [5], in addition to a GAN-based method, that is, Multi-task GAN (MGAN) [16], for comparison in the identification scenario only because of the lack of EER results in their original papers.

Similar to the other methods, we evaluated all 11 probe views, whereas the gallery view set that corresponded to each probe view contained the other 10 views, that is, we only excluded the identical view (i.e., the same view) from the total 11 views. The mean results over all 10 gallery views for each probe view are shown in Table V. Because of the small size of the training set, the improvements of the proposed methods are not as large as that on OU-MVLP.

Particularly, the networks with a high-level matching structure sometimes cannot outperform those with a low-level matching structure (e.g., PST-4in vs. PST-2LB*) because they may fail to extract sufficiently invariant features using a small training set. As a result, the proposed PST-LB* + PST-2in achieves the best performance for the verification scenario. Considering the small number of test subjects (i.e., 50), the proposed GEI-based PST-2LB* + PST-4in still obtains competitive results for identification scenario compared with GEI Ensemble and GEI-temporal Ensemble [5], which is a fusion of five networks using GEI, and a fusion of eight networks using GEI and/or temporal information, respectively.

V. Conclusion

In this paper, a PSTN for cross-view gait recognition was presented, which combines a PST for spatial transformation and an RN for discrimination learning in a unified framework. To the best of our knowledge, this is the first time that geometric feature registration has been integrated into a CNN architecture in the gait recognition community. To avoid unnecessary large distortion, an input matching pair of GEIs from different views were both registered into their intermediate view with a non-rigid deformation field predicted by the PST, and further fed into the subsequent RN to obtain the dissimilarity score. A loss function for optimizing the main recognition task was designed for the PSTN, which made the learned transformation realize a good trade-off between maintaining inter-subject variations and suppressing intra-subject variations that resulted from the view differences to achieve optimal recognition accuracy. Experiments on three publicly available datasets demonstrated the effectiveness of the proposed method, which achieved state-of-the-art performance among GEI-based methods in both the verification and identification scenarios.

One important future research avenue is the extension of the PST module considering both view registration and phase registration for each frame, in conjunction with the recent silhouette-based recognition networks, such as [21]. Because geometric feature registration is also applicable to other covariates, such as posture change caused by the walking/running speed, a future research direction is to evaluate the performance of the proposed PSTN under other covariates in gait recognition. Rather than using a simple score-level fusion of networks with low-level and high-level matching structures, a unified model that combines these two structures to accommodate different degrees of view variation is also worth investigating in the future. Additionally, considering that the sample-dependent transformation predicted by the PST may contain discriminative individualities, including the transformation parameters in subsequent feature learning may improve the final recognition performance, and this remains future work.

Acknowledgment

The authors would like to thank Dr. Noriko Takemura for providing help and valuable suggestions for the implementations and experiments and thank Maxine Garcia, PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

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, May 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] 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.

[5] 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.

[6] M. Altab 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.

[7] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren, "Gait energy response function for clothing-invariant gait recognition," in *Proc. 13th Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan, Nov. 2016, pp. 257–272.

[8] Y. Guan and C.-T. Li, "A robust speed-invariant gait recognition system for walker and runner identification," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–8.

[9] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Speed invariance vs. Stability: Cross-speed gait recognition using single-support gait energy image," in *Proc. 13th Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan, Nov. 2016, pp. 52–67.

[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] A. Kale, A. K. R. Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Jul. 2003, pp. 143–150.

[12] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1–8.

[13] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012.

[14] 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)*, Jun. 2016, pp. 1–8.

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

[16] 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.

[17] F. Jean, R. Bergevin, and A. B. Albu, "Computing and evaluating view-normalized body part trajectories," *Image Vis. Comput.*, vol. 27, no. 9, pp. 1272–1284, Aug. 2009, doi: 10.1016/j.imavis.2008.11.009.

[18] R. Martín-Félez and T. Xiang, "Uncooperative gait recognition by learning to rank," *Pattern Recognit.*, vol. 47, no. 12, pp. 3793–3806, Dec. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320314002325

[19] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1960–1968, Nov. 2015.

[20] 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.

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

[22] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4695–4704.

[23] Z. Zhang *et al.*, "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4705–4714.

[24] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.

[25] A. Jourabloo and X. Liu, "Pose-invariant face alignment via CNN-based dense 3D model fitting," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 187–203, Apr. 2017.

[26] V. Naresh Boddeti, M.-C. Roh, J. Shin, T. Oguri, and T. Kanade, "Face alignment robust to pose, expressions and occlusions," 2017, *arXiv:1707.05938*. [Online]. Available: http://arxiv.org/abs/1707.05938

[27] M. Jaderberg, K. Simonyan, A. Zisserman, and K. kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. New York, NY, USA: Curran Associates, 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

[28] N. S. Detlefsen, O. Freifeld, and S. Hauberg, "Deep diffeomorphic transformer networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4403–4412.

[29] X. Peng, N. Ratha, and S. Pankanti, "Learning face recognition from limited training data using deep neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1442–1447.

[30] W. Wu, M. Kan, X. Liu, Y. Yang, S. Shan, and X. Chen, "Recursive spatial transformer (ReST) for alignment-free face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3792–3800.

[31] Y. Shi and A. Jain, "Improving face recognition by exploring local features with visual attention," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 247–254.

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

[33] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

[34] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Hong Kong, vol. 4, Aug. 2006, pp. 441–444.

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

[36] H.-D. Yang and S.-W. Lee, "Reconstruction of 3D human body pose for gait recognition," in *Proc. IAPR Int. Conf. Biometrics 2006*, Jan. 2006, pp. 619–625.

[37] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Proc. the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. CVPR*, vol. 1, Dec. 2001, pp. I–I.

[38] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 40, no. 4, pp. 997–1008, Aug. 2010.

[39] H. El-Alfy, C. Xu, Y. Makihara, D. Muramatsu, and Y. Yagi, "A geometric view transformation model using free-form deformation for cross-view gait recognition," in *Proc. 4th IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2017.

[40] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Cross-view and multi-view gait recognitions based on view transformation model using multi-layer perceptron," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 882–889, May 2012.

[41] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 140–154, Jan. 2015.

[42] J. Lu and Y.-P. Tan, "Uncorrelated discriminant simplex analysis for view-invariant gait signal computing," *Pattern Recognit. Lett.*, vol. 31, no. 5, pp. 382–393, Apr. 2010.

[43] A. Mansur, Y. Makihara, D. Muramatsu, and Y. Yagi, "Cross-view gait recognition using view-dependent discriminative analysis," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–8.

[44] Z. Zhang, J. Chen, Q. Wu, and L. Shao, "GII representation-based cross-view gait recognition by discriminative projection with list-wise constraints," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2935–2947, Oct. 2018.

[45] K. Bashir, T. Xiang, and S. Gong, "Cross view gait recognition using correlation strength," in *Proc. Procedings Brit. Mach. Vis. Conf.*, 2010, pp. 1–11.

[46] 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.

[47] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognit.*, vol. 50, pp. 107–117, Feb. 2016.

[48] 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.

[49] 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.

[50] 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. Workshops (CVPRW)*, Jul. 2017, pp. 532–539.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[52] Y. Zhong, J. Chen, and B. Huang, "Toward end-to-end face recognition through alignment learning," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1213–1217, Aug. 2017.

[53] 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.

[54] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[55] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, 2011, pp. 109–117.

[56] T. W. Sederberg and S. R. Parry, "Free-form deformation of solid geometric models," *ACM SIGGRAPH Comput. Graph.*, vol. 20, no. 4, pp. 151–160, Aug. 1986, doi: 10.1145/15886.15903.

[57] Y. Makihara, D. Adachi, C. Xu, and Y. Yagi, "Gait recognition by deformable registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 561–571.

[58] J. Chaki and N. Dey, *A Beginner's Guide to Image Preprocessing Techniques* (Intelligent Signal Processing and Data Analysis). Boca Raton, FL, USA: CRC Press, 2018. [Online]. Available: https://books.google.co.jp/books?id=Dfp0DwAAQBAJ

[59] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn. (ICML)*. Madison, WI, USA: Omnipress, 2010, pp. 807–814. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104425

[60] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

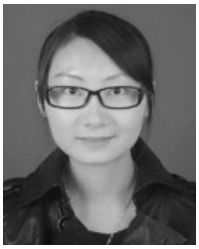
[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

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

[63] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Washington, DC, USA: IEEE Computer Society, Jun. 2014, pp. 1386–1393, doi: 10.1109/CVPR.2014.180.

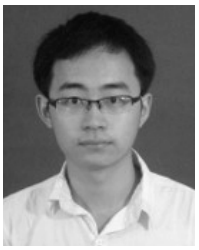
[64] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Which reference view is effective for gait identification using a view transformation model?" in *Proc. IEEE Comput. Soc. Workshop Biometrics*, New York, NY, USA, Jun. 2006.

[65] N. Otsu, "Optimal linear and nonlinear solutions for least-square discriminant feature extraction," in *Proc. 6th Int. Conf. Pattern Recognit.*, 1982, pp. 557–560.
[66] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.
[67] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 9, May 2010, pp. 249–256.
[68] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2007, pp. 161–168.



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 system. Since January 2016, she has been a Visiting Researcher with the Institute of Scientific and Industrial Research, Osaka University, Japan. Her research interests are gait recognition, machine learning, and image processing.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University in 2001, 2002, and 2005, respectively. He was appointed specially as an Assistant Professor (full-time), an Assistant Professor, and an Associate Professor from the Institute of Scientific and Industrial Research, Osaka University, in 2005, 2006, and 2014, respectively, where he is currently a Professor with the Institute for Advanced Co-Creation Studies. His research interests are computer vision, pattern recognition, and image processing, including gait recognition, pedestrian detection, morphing, and temporal super resolution. He is a member of IPSJ, IEICE, RSJ, and JSME. He has obtained several honors and awards, including the Second International Workshop on Biometrics and Forensics (IWBF 2014), the IAPR Best Paper Award, the 9th IAPR International Conference on Biometrics (ICB 2016), the Honorable Mention Paper Award, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology, Research Category, in 2014. He has served as an Associate Editor-in-Chief for the *IEICE Transactions on Information and Systems*, an Associate Editor for the *IPSJ Transactions on Computer Vision and Applications* (CVA), a Program Co-Chair for the 4th Asian Conference on Pattern Recognition (ACPR 2017), and an Area Chair for ICCV 2019, CVPR 2020, and ECCV 2020.



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 January 2016, he has been a Visiting Researcher with the Institute of Scientific and Industrial Research, Osaka University, Japan. His research interests are gait recognition, image processing, and machine learning.

Yasushi Yagi (Member, IEEE) 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 was the Executive Vice President from 2015 to 2019. He is currently a Professor with the Institute of Scientific and Industrial Research, Osaka University. His research interests are computer vision, medical engineering, and robotics. He is a fellow of IPSJ and a member of IEICE and RSJ. He was awarded 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. International conferences for which he has served as the Chair include: FG1998 as the Financial Chair, OMINVIS2003 as the Organizing Chair ROBIO2006 as the Program Co-Chair, ACCV2007 as the Program Chair, PSVIT2009 as the Financial Chair, ICRA2009 as the Technical Visit Chair, ACCV2009 as the General Chair, ACPR2011 as the Program Co-Chair, and ACPR2013 as the General Chair. He has also served as an Editor for the IEEE ICRA Conference Editorial Board from 2007 to 2011. He is also an Editorial Member of IJCV and the Editor-in-Chief of *IPSJ Transactions on Computer Vision and Applications*.

Jianfeng Lu received the B.S. degree in computer software and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, China. He is currently a Professor with the Nanjing University of Science and Technology. His research interests include image processing, pattern recognition, and data mining.