Cross-View Gait Recognition with Deep Universal Linear Embeddings

Shaoxiong Zhang, Yunhong Wang, Annan Li *
State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing, China
{zhangsx, yhwang, liannan}@buaa.edu.cn

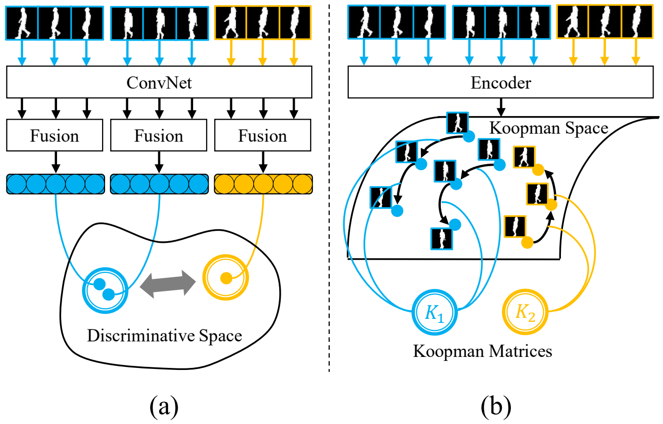
Abstract

Gait is considered an attractive biometric identifier for its non-invasive and non-cooperative features compared with other biometric identifiers such as fingerprint and iris. At present, cross-view gait recognition methods always establish representations from various deep convolutional networks for recognition and ignore the potential dynamical information of the gait sequences. If assuming that pedestrians have different walking patterns, gait recognition can be performed by calculating their dynamical features from each view. This paper introduces the Koopman operator theory to gait recognition, which can find an embedding space for a global linear approximation of a nonlinear dynamical system. Furthermore, a novel framework based on convolutional variational autoencoder and deep Koopman embedding is proposed to approximate the Koopman operators, which is used as dynamical features from the linearized embedding space for cross-view gait recognition. It gives solid physical interpretability for a gait recognition system. Experiments on a large public dataset, OU-MVLP, prove the effectiveness of the proposed method.

1. Introduction

Gait recognition aims to identify people by recognizing their body shape and walking patterns. Compared to other biometrics such as fingerprint or iris, gait requires relatively low cooperation and can be performed at a longer distance. Besides, it is also difficult to camouflage. Therefore, gait recognition can be applied in some special senses such as criminal investigation [28].

Although the progress is encouraging, gait recognition still suffers from many external factors such as carrying condition, varying pace, clothing, and camera viewpoints, which degrade the performance of gait recognition systems. Among all these unfavorable factors, camera viewpoints could be the most tricky one [43]. Prior arts proved that the performance of a single-view gait recognition system would drop drastically if the viewpoint is changed [43, 34].



(a)

(b)

Figure 1. Previous works focus on feature fusion of gait silhouette sequences and search for discriminative space where distance is small for the feature pair with the same identity (a). We calculate dynamical features in the *Koopman* space where gait images evolve linearly (b), and then we recognize them from their dynamical features.

To solve these problems, many deep learning models have been proposed for cross-view gait recognition and achieved great performances. In general, such approaches can be grouped into two categories, *i.e.* appearance-based approaches and model-based approaches, respectively. The former [43, 5, 33, 50, 45] is mainstream for gait recognition in the past few years. These methods extract features from gait silhouette images and optimize the intra-individual distance in the feature space by metric learning loss functions without gait cycle modeling. In addition, temporal fusion units [6, 9, 44, 5, 8, 23] and part-division units [17, 5, 8, 50] are proposed to combine features of silhouette sequences and local parts.

The model-based gait recognition [2, 22, 24] focus on reconstructing body structures from gait sequences in a mathematical manner. A three-dimensional model conveys more information than a two-dimensional one and can be con-

[*] Corresponding author.

structed to represent the gait pattern. Therefore, it can achieve acceptable performance against viewpoints variation in theory. This point is also supported by some biomechanical gait analysis [14, 15]. However, the performances of such approaches are vulnerable to the accuracy of pose estimation and the quality of silhouette sequences, which limited their development.

In general, appearance-based methods are good at feature representation but suffer from insufficient data, while model-based approaches are more robust to view differences but challenging to construct. Although deep convolutional neural networks (ConvNets) can provide a robust feature extractor and achieve excellent performance in controlled scenarios, exiting models still cannot deal with large view differences or variations of clothing and article-carrying very well. Because, in essence, ConvNet is still a two-dimensional template, and the human body is a three-dimensional object. It is not surprising that even having the exact person's data, the model still cannot deal with his/her 2D projections that are not included in the training set. This issue is also known as the *ill-posed* problem of computer vision.

Inspired by works on inertial sensor-based gait analysis [29, 4], biomechanical gait analysis [14] and dynamical analysis of human gait [3], we realize that dynamical features are competitive in gait recognition since it models the essence of human gait, the motion process, rather than pure human shapes. Therefore, different from most existing deep learning methods, we explore cross-view gait recognition from a dynamical system perspective. More specifically, we introduce the *Koopman* theory, which is a popular tool for analyzing nonlinear systems in the literature of fluid mechanics [31, 42]. In fact, the *Koopman* theory has been already applied to computer vision as video background separations [10], image spoofing [37], and motion detection [7]. As for gait recognition, the most relevant work is by Wang et al. [39], in which windowed-dynamic mode decomposition is applied for generating gait energy images. However, only the static gait feature is investigated in their work.

As shown in Figure 1, the *Koopman* theory focuses on the systematic linear representation of nonlinear systems, which provides a new way of representing the walk cycles of gait. We propose a novel framework for cross-view gait recognition by approximating *Koopman* operator (see Figure 2). First, aligned silhouettes are fed into a convolutional variational auto-encoder (VAE) for image-level encoding. Then, we enforce additional constraints and loss functions [26] to identify *Koopman* operators where the dynamics evolve linearly following. Finally, a fully-connected network is trained for final gait representation from the *Koopman* matrix.

In summary, we make the following three major contributions.

- We introduce the *Koopman* theory to dynamic feature extraction from gait silhouettes. To our knowledge, this is the first study to apply *Koopman* analysis.
- We propose a novel framework for cross-view gait recognition by integrating convolutional variational autoencoder and deep *Koopman* embedding.
- We conduct experiments on a widely used large gait database, OU-MVLP [36]. The results prove the effectiveness of our method, which makes an essential contribution to understanding the connections between gait recognition and human walking dynamics.

2. Related Work

In this section, we will give a brief introduction to recent works on gait recognition. Before the deep learning era, time series analysis methods are applied in some works, such as Auto-regressive Modeling [38] and Hidden Markov Mode [25], for dynamics modeling. These models usually have strong assumptions, while they also cannot fit nonlinear systems well. In the deep learning era, ConvNets [21] has been proved successful in numerous computer vision tasks, and it has also been adopted for gait recognition and achieved admirable performance. In general, the ConvNet based approaches can be grouped into two categories, appearance-based approaches and model-based approaches. Meanwhile, according to the type of input data, the proposed works can also be grouped into template-based approaches and sequence-based approaches.

Most template-based methods applied ConvNets to extract gait features from a single gait image, such as the Gait Energy Images (GEI) [11] or other GEI-like template images [1]. Wu et al. [43] proposed three ConvNets with different architectures and conducted a series of experiments that significantly improved cross-view gait recognition performance. Similar ConvNets can be found in [32, 48, 49]. Meanwhile, some generative models are also proposed to transform gait images from one view to another, such as auto-encoder [47] and generative adversarial network [46, 12, 41].

Some works establish the model directly from the gait silhouette sequence rather than the GEI. They apply temporal models to encode information across time, such as Feature Map Pooling [6], Long Short Term Memory [9], and three-dimensional ConvNet [44]. Some latest works on large gait databases present competitive performance. Chao et al.[5] presented a novel perspective regarding gait as a set of silhouettes rather than continuous sequences. They believe that the silhouette's appearance contains position information, which is a replacement of temporal information. Thus, they apply a simple ConvNet to extract frame-level gait features from the silhouette and then use a pooling operation to aggregate frame-level features into a single setlevel feature. Zhang et al. [50] proposed a model combined with ConvNets for single image feature extraction and LSTM attention models for attention scores on the frame-level ConvNet. Fan et al. [8] proposed a novel part-based model with a micro-motion capture module, which also provides an approach of temporal modeling.

Recently, some works focus on model-based gait recognition methods [2, 22, 24]. They reconstruct mathematical structures of the human body from gait image sequences. Three-dimensional data of human walking can be constructed, and it conveys more information than two-dimensional data. Therefore, it can solve the cross-view problem with a three-dimensional model rotation. However, it suffers excessive details, which degrades the performance of cross-view gait recognition.

3. The Koopman Operators

In this section, we present the basics of *Koopman* operators [31] and extending dynamic mode decomposition [42]. The *Koopman* operator is a linear but infinite-dimensional operator, approximate by a data-driven method. For a nonlinear dynamical system, the *Koopman* observation functions map the original state space to an embedding space where the dynamics would evolve universally linearly. Extended dynamic mode decomposition (EDMD) is a method that approximates the *Koopman* eigenvalue, eigenfunction, and mode tuples. The EDMD procedure requires two prerequisites: a data set of snapshots and a dictionary of observation functions.

Given a discrete-time dynamical system, $x_t \in \mathcal{M}$ at the time step t, described by:

$$x_{t+1} = F(x_t) \tag{1}$$

where F represents the dynamics that map the state of the system forward in time. *Koopman* theory provides an alternative description of dynamical systems in terms of the evolution of functions, that is *Koopman* operator $\mathcal{K}$, which is an infinite-dimensional linear operator. Denote eigenfunctions as $\varphi_p : \mathcal{M} \to \mathcal{F}$ and eigenvalues λ_p of the *Koopman* operator $\mathcal{K}$, we have

$$\mathcal{K}\varphi_p(x_t) = \lambda_p \varphi_p(x_t), \quad p = 1, 2, \dots \tag{2}$$

Consider a vector-valued function $g : \mathcal{M} \to \mathcal{F}$. $\mathcal{K}$ maps g into a new function $\mathcal{K}g$, satisfying:

$$\mathcal{K}g(x_t) = g(F(x_t)) \tag{3}$$

If g lies within the span of the eigenfunctions φ_p, g can be expanded in terms of eigenfunctions as

$$g(x_t) = \sum_{p=1}^{\infty} \varphi_p(x_t) v_p \tag{4}$$

Then we have

$$\begin{aligned} g(F(x_t)) &= \mathcal{K}g(x_t) \\ &= \sum_{p=1}^{\infty} \mathcal{K}\varphi_p(x_t) v_p \\ &= \sum_{p=1}^{\infty} \lambda_p \varphi_p(x_t) v_p \end{aligned} \tag{5}$$

Therefore, the dynamic of the system is linear if we regard λ_p as coefficients:

$$g(F(x_t)) = Kg(x_t) \tag{6}$$

where the *Koopman* operator $\mathcal{K}$ will yield a matrix K to the subspace spanned by φ_p. Conventionally, observation functions g can be determined by hand-designed methods from the knowledge of underlying physics. Then, the system identification problem can be transformed into finding the *Koopman* matrix K, which can be solved by linear regression given collected numerical data. In summary, the *Koopman* operator theory focuses on the linear representation of a nonlinear system, capturing the full information of the original nonlinear system.

4. Proposed Approach

4.1. Problem Formulation

Given a sequence of gait silhouettes, it can be regarded as time sequence data $\{x_t\}$, where $t \in [1, 2, ..., M]$, and M is the number of frames in this gait sequence. The *Koopman* theory indicates that by representing a nonlinear dynamics system in linear space with *Koopman* operator, prediction for linear systems can be used for system state analysis. Suppose we regard human walking as dynamic systems by assuming pedestrians have unique walking patterns. In that case, we can calculate different *Koopman* matrix K_i for subject i from their gait silhouette sequences $\{x_{i,t}\}$:

$$g(x_{i,t+1}) = K_i g(x_{i,t}) \tag{7}$$

Once K_i can be estimated from $\{x_{i,t}\}$ via the least-squares solution, we can recognize the identities of pedestrians after comparing the similarity of estimated walking patterns $\hat{K}_i$ as:

$$\hat{K}_i^{\mathrm{T}} = LS(\Phi(\{x_{i,t}\})) \tag{8}$$

where Φ is a convolutional neural network representing the observation functions g, and LS stands for the least-squares solution. Therefore, we formulate the gait recognition task as follows:

$$feat_i = \Psi(LS(\Phi(\{x_{i,t}\}))) \tag{9}$$

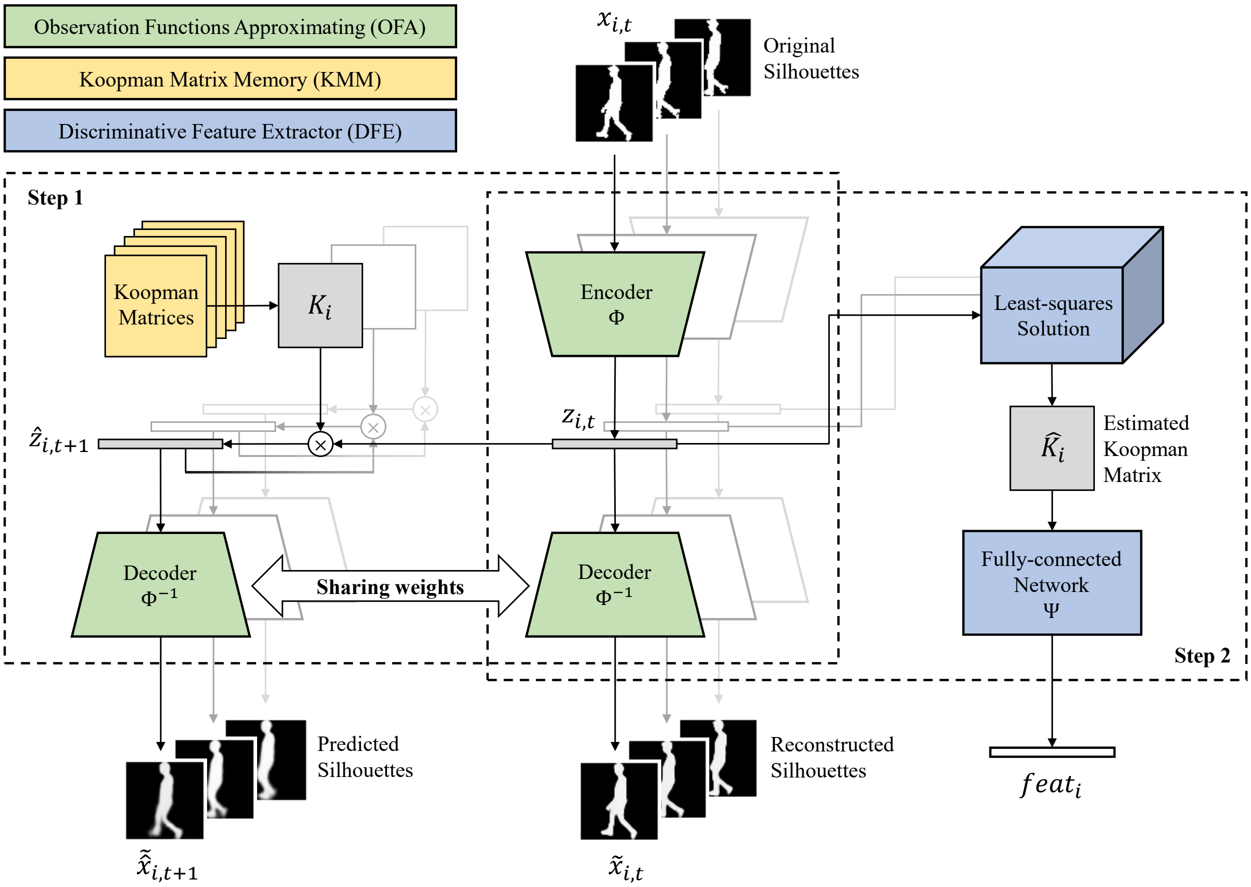


Figure 2. The framework of our proposed method. In training step one, the OFA and the KMM module are trained. In training step two, only parameters in the DFE module are trained with parameters in OFA frozen.

where Ψ is a fully connected network that transforms the estimated *Koopman* matrix $\hat{K}_i$ into a new feature in a discriminative space for individual identification.

Besides, the estimated *Koopman* matrix $\hat{K}_i$ can also be used to predict the future since it contains the original system's information. In our model, the future images of a pedestrian $\{\tilde{\hat{x}}_{i,t}\}$ can be predicted as:

$$\tilde{\hat{x}}_{i,t+T} = \hat{K}_i^T \Phi(x_{i,t}) \tag{10}$$

4.2. Model Architecture

Our model consists of three key components, including the Observation Function Approximating module (OFA), the *Koopman* Matrix Memory (KMM), and the Discriminative Feature Extractor module (DFE).

In the OFA module, a variational auto-encoder (VAE) [20] with convolutional layers is applied to leverage the power of deep learning to represent eigenfunctions of the *Koopman* operator. The KMM module contains learnable parameters K_i for each individual in the training set, which can be trained via backpropagation. In the DFE module, a simple fully-connected network is used to transform the estimated Koopman matrices into a discriminative space for cross-view recognition. The model is illustrated in Figure 2.

Observation Functions Approximating

In the Observation Functions Approximating module, the input data is an aligned gait silhouette sequence. An image $x_{i,t}$ with identity i at time step t is fed into an encoder Φ,

$$z_{i,t} = \Phi(x_{i,t}) \tag{11}$$

The encoder Φ contains six convolutional layers and two fully-connected layers, shown in Table 1. After fully-connected layers, the network gives a mean vector μ and a standard deviation vectors σ. Both of them are D-dimensional. Then we get the code of $x_{i,t}$ sampled from the distribution $N(\mu, \sigma^2)$. The encoder Φ aims to transform original input data $x_{i,t} \in \mathbb{R}^n$ into *Koopman* space $z_{i,t} \in \mathcal{F}$ with the help of non-linear transformation of deep network instead of original observation functions. The output size

Table 1. The architecture of the encoder Φ, the decoder Φ^{-1} and the fully connected network Ψ. Activation function ReLUs are skipped after each convolutional and fully-connected layer except FC 2 and FC 7. Batch normalization layers after FC 5 and FC 6 are also skipped. The strings following each convolutional layer are formatted as the filters' size, the dimensions of the feature maps. Conv stands for convolution operator, and Deconv stands for 2D transposed convolution operator, while outpadding stands for the additional size added to one side of the output shape.

Layers	Architecture
Conv 1	5×5, 8, padding 2
Conv 2	3×3, 8, stride 2, padding 1
Conv 3	3×3, 16, padding 1
Conv 4	3×3, 16, stride 2, padding 1
Conv 5	3×3, 32, padding 1
Conv 6	3×3, 32, stride 2, padding 1
FC 1	$32 \times 8 \times 8$ to 1024
FC 2	1024 to $2D$
FC 3	D to 1024
FC 4	1024 to $32 \times 8 \times 8$
Deconv 1	3×3, 32, stride 2, padding 1, outpadding 1
Deconv 2	3×3, 32, padding 1
Deconv 3	3×3, 16, stride 2, padding 1, outpadding 1
Deconv 4	3×3, 16, padding 1
Deconv 5	3×3, 8, stride 2, padding 1, outpadding 1
Deconv 6	5×5, 8, padding 2
FC 5	2*D to 4096
FC 6	4096 to 2048
FC 7	2048 to 512

for Φ is D, which means that in this work, we set $\mathcal{F} = \mathbb{R}^D$. D is a hyperparameter of this model, which can be decided experimentally. In this work, we set $D = 128$. We hold the idea that a 128-dimensional space is enough to approximate the *Koopman* matrix for recognition because the gait cycle is a relatively simple dynamical system.

A decoder Φ^{-1} is applied to ensure that the code $z_{i,t}$ in the *Koopman* space keeps most of the useful information in original images, rather than converging on outliers such as zeros, while the architecture is also shown in Table 1,

$$\tilde{x}_{i,t} = \Phi^{-1}(z_{i,t}) \tag{12}$$

This encoder-decoder module extracts human walking patterns by transforming original gait sequences into the *Koopman* space. Instead of hand-designed observation functions, an encoder-decoder structure is capable of representing any arbitrary function, including desired *Koopman* eigenfunctions [26]. Therefore, our model can accurately fit human walking dynamics without hand-designed functions.

Koopman Matrix Memory

We assume that individuals have unique walking patterns. Therefore, their *Koopman* matrices should be the same while walking at an even pace. To achieve this assumption in the training phase, a *Koopman* Matrix Memory K is constructed for the training dataset. $K = [K_i]$ is a learnable parameter matrix, trained via backpropagation from predicting the next frame in the same gait sequence. Each individual in the training set has one unique K_i, which is a $D \times D$ matrix and initialized randomly. After an input gait sequence $X = [x_{i,t}]$ is encoded into $Z = [z_{i,t}]$, K_i is used to predict the state for the next snapshot in the *Koopman* space by

$$[\hat{z}_{i,2}, \hat{z}_{i,3}, ..., \hat{z}_{i,t+1}] = K_i[z_{i,1}, z_{i,2}, ..., z_{i,t}] \tag{13}$$

K_i is directly loaded from *Koopman* Matrix Memory K. In this way, it is ensured that input gait sequences with different view angles can be encoded into the same space. It should be noticed that the *Koopman* Matrix Memory K is only used for training the OFA module. After parameters in the OFA module are frozen, this KMM module will be removed from the model and not be used.

Discriminative Feature Extractor

Finally, a simple fully connected network Ψ transforms the estimated *Koopman* matrix $\hat{K}_i$ into a new feature in a discriminative space:

$$feat_i = \Psi(\hat{K}_i) \tag{14}$$

where Euclidean distance can be applied to measure the similarity of two features. It should also be noticed that the input data $\hat{K}_i$ is calculated via the least-squares estimation shown in Equation 8, rather than K_i from the *Koopman* Matrix Memory. The architecture of Ψ is listed in Table 1.

4.3. Loss Functions

The reconstruction accuracy of the autoencoder in the OFA module is achieved to reduce the spatial information lost. $\mathcal{L}_\Phi$ refers to the difference between the original gait silhouette sequences and the recovered gait silhouette sequences from linear space, following:

$$\mathcal{L}_\Phi = \|x_{i,t} - \Phi^{-1}(\Phi(x_{i,t}))\| \tag{15}$$

Meanwhile, an additional loss $\mathcal{L}_{\mu,\sigma}$ is applied to enhance the model generation ability [20], which tries to push the distributions as close as possible to unit Gaussian,

$$\mathcal{L}_{\mu,\sigma^2} = KL(N(\mu, \sigma^2)\|N(0,1)) \tag{16}$$

where KL stands for Kullback-Leibler divergence.

According to the *Koopman* theory [31, 42], we learn the linear dynamics K_i to ensure linear dynamics: $\Phi(x_{i,t+1}) = K_i \Phi(x_{i,t})$. More generally, we enforce linear prediction over S time steps with the loss:

$$\mathcal{L}_{linear} = \|\Phi(x_{i,t+S}) - K_i^S \Phi(x_{i,t})\| \tag{17}$$

In addition, future gait images are also need to be predicted with $\mathcal{L}_{furure}$ over S time steps:

$$\mathcal{L}_{furure} = \|x_{i,t+S} - \Phi^{-1}(K_i^S \Phi(x_t))\| \tag{18}$$

In these losses, norm $\| \cdot \|$ is a mean-squared error, and they are all averaged in a training batch.

As for the DFE module, a triplet loss with hard mining [13] $\mathcal{L}_{triplet}$ and a Softmax loss $\mathcal{L}_{softmax}$ are both employed for identity recognition. In a training batch, $p \times k$ gait silhouette sequences randomly selected from the training set, where p denotes the number of selected subjects and k for the number of different views. For each data in a training batch as an anchor, the hardest positive data and the hardest negative data are selected for $\mathcal{L}_{triplet}$. Meanwhile, an additional classifier is applied for $\mathcal{L}_{softmax}$, which takes y_i as input and is omitted in Figure 2.

4.4. Implementation Details

During the training, the whole model is trained within two steps. We train the OFA and the KMM module together, without the DFE module in training step one. Then we freeze the OFA module's parameters and train the DFE module alone without the KMM module in training step t-wo. The reason is that the least-squares solution is applied to calculated *Koopman* matrices $\hat{K}_i$. If we train this model entirety, it requires back-propagating the loss through the Eigendecomposition step in the least-squares solution, which is unstable [40]. Besides, if the *Koopman* space cannot be established correctly, the *Koopman* matrices $\hat{K}_i$ are all irrational. Therefore, a two-step training strategy is employed.

In training step one, the OFA and the KMM module are trained together with loss function:

$$\mathcal{L}_{step1} = \alpha \mathcal{L}_{\Phi} + \beta \mathcal{L}_{\mu,\sigma^2} + \gamma \mathcal{L}_{linear} + \lambda \mathcal{L}_{furure} \tag{19}$$

In training step two, the DFE module is trained with loss function:

$$\mathcal{L}_{step2} = \xi \mathcal{L}_{triplet} + \mathcal{L}_{softmax} \tag{20}$$

We randomly select 32 continuous frames in both training steps in a gait silhouette sequence as one training sample for training. In the testing phase, we calculated one *Koopman* matrix for every 32 frames. Therefore, we get more than one *Koopman* matrix because there are always more than 32 frames in a gait sequence. The final distance between a probe sample and the gallery samples will be the average distance of all the calculated *Koopman* matrices of this probe sequence.

5. Experiments

5.1. Dataset

The OU-ISIR Gait Dataset, Multi-View Large Population Dataset (OU-MVLP) [36] is the largest public gait database. It contains 10,307 subjects with 14 views ($0°, 15°, ..., 90°, 180°, 195°, ..., 270°$) per subject and 2 sequences per view. The gait sequences have been divided into two sets, including 5,153 subjects in the training set and 5,154 subjects in the testing set. For evaluation, one sequence is kept in the gallery, and the other one acts as a probe. Silhouette sequences are released with the original image size and background removed. We conduct size-normalization following [5] with 64×64 pixels in each gait silhouette.

5.2. Hyper-Parameters Details

We initialize the weights of ConvNet, the weights and bias of fully connected layers by normal initialization with standard deviation equaling to 0.01. In optimization, the Adam algorithm [19] was implemented. In triplet loss, we set the margin as 2.0. The batch size is set to $p = 32, k = 16$ for the whole training phase. In training step one, the learning rate is set to $1e^{-3}$ for the first 20,000 iterations with $S = 8$ in Equation 17 and 18. Then the learning rate is changed to $1e^{-4}$ for the rest 80,000 iterations with $S = 16$. This setting is used to stabilize the training process. Other hyper-parameters in Equation 19 and 20 are set as: $\alpha = 0.001, \beta = 0.002, \gamma = 1, \lambda = 0.01$, and $\xi = 10$. In training step two, we train the model with learning rate $1e^{-4}$ for 50,000 iterations, and $1e^{-5}$ for the next 20,000 iterations.

5.3. Visualization

We visualize the original gait silhouette sequences $\{x_{i,t}\}$ and the predicted ones $\{\tilde{\hat{x}}_{i,t+1}\}$ to prove the effectiveness of the *Koopman* theory on gait, as shown in Figure 4. The predicted gait silhouette sequences $\{\tilde{\hat{x}}_{i,t+1}\}$ are calculated as Equation 10. The predicted gait silhouette sequences hold the same walking characteristics as original ones, including crookback and arm swing.

Furthermore, we exchange the calculated *Koopman* matrices of the first two identities in Figure 4, and predict the gait silhouette sequences again, shown in Figure 3. We find that the gait phases are the same in both figures, while their walking features are exchanged. For example, from identity *A* in Figure 3, the amplitude of arm swing get larger and the person becomes crookback, which is similar to identity *B* in Figure 4. This visualization proves that the estimated *Koopman* matrix $\hat{K}_i$ can represent the whole gait sequence. Given an initial gait phase $x_{i,t=1}$, we can predict the next

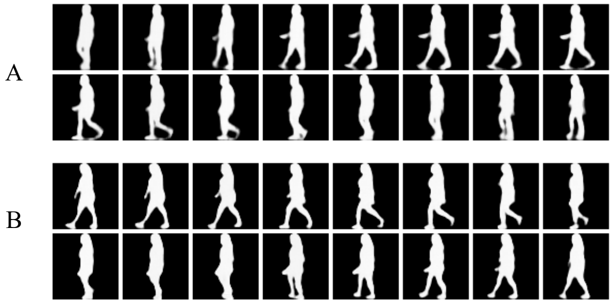


Figure 3. Visualization of predicted silhouette sequences of the first two identities *A* and *B* in Figure 4 with their *Koopman* matrices exchanged. The predicted silhouette sequence of the *A* is calculated with the *Koopman* matrices of *B*, and vice versa.

several gait silhouettes with $\hat{K}_i$ easily by matrix multiplication. These findings may help us understand the strong physical interpretability of the proposed gait dynamical feature in a gait recognition system.

5.4. Comparison and Discussion

Table 2, 3 and 4 show the comparison with previous works on OU-MVLP dataset. In Table 2 and 3, we compare our performance with some non-deep learning methods and template-based methods. Our model presents better performance than all of the previous template-based methods. These results indicate that the dynamical gait feature is discriminative and competitive on the cross-view gait recognition task.

In Table 4, we compare our model with the state-of-the-art sequence-based approaches, including GaitSet [33], GaitPart [8], and ACL+local+temporal (ACL+) [50]. Unlike template-based methods, these sequence-based approaches extract deep features for each gait silhouette and concatenate them together for recognition. Therefore, these approaches outperform the template-based methods greatly. It is rather disappointing that there are still gaps between our performance and theirs.

A possible explanation might be that the two tasks, dynamical analysis, and identity recognition, are not integrated perfectly. In other words, the whole framework is not an end-to-end model. Since it is the first work to introduce the *Koopman* theory with the deep learning model into the gait recognition system, firstly, we construct the *Koopman* space to check its effectiveness on gait images. After that, we optimize the remaining parameters for recognition with the *Koopman* space fixed. That is the reason that our final performance is degraded compared with the deep end-to-end models.

Another possible explanation is that we do not apply the part-based modules in our model. Part-based modules, *e.g.* Horizontal Pyramid Mapping [33], Horizontal Pooling [8], and Local Feature Extraction Module [50], employ partial features of the human body and have been proved to be beneficial to recognition. We focus on linear representation on a full-body moving system in this work and do not design a local features extraction module. Therefore, we can not achieve the state-of-the-art performance as theirs.

Notwithstanding these limitations, the study suggests that dynamical features can provide solid physical interpretability for a gait recognition system and achieve acceptable performance on a cross-view gait recognition task in such a large gait database. Considering the two explanations above, we believe that there are still opportunities for further enhancements.

Table 2. Averaged rank-1 identification rates in four key view pairs on OU-MVLP dataset, compared with GEINet [32], Siamese [48], CNN-LB [43], MGANs [12], Attention [18], and DiGGAN [16]. The data of the first four methods are from [16].

Gallery	90°			
Probe	0°	30°	60°	90°
GEINet	3.4	21.5	50.2	90.7
Siamese	7.9	26.5	36.5	82.1
CNN-LB	2.2	14.0	41.2	91.7
MGANs	2.1	12.0	22.0	85.9
Attention	23.7	36.1	57.2	89.1
DiGGAN	44.6	58.9	66.0	90.0
Ours	44.9	80.5	88.3	95.4

Table 3. Averaged rank-1 identification rates in four angular difference on OU-MVLP dataset, compared with LDA[30], VTM[27], GEINet [32], CNN-LB [43], 3in+2diff[35], Attention [18], PST-2LB*+PST-4in (PST+) [45], ACL+local+temporal (ACL+) [50]. The data of the first four methods are from [35].

Method	**Angular Difference**				**Mean**
	0°	30°	60°	90°	
LDA	81.6	10.1	0.8	0.1	24.4
VTM	77.4	2.7	0.6	0.2	20.5
GEINET	85.7	40.3	13.8	5.4	40.7
CNN-LB	89.9	42.2	15.2	4.5	42.6
3in+2diff	89.5	55.0	30.0	17.3	52.7
Attention	89.1	57.2	36.1	23.7	55.7
PST+	93.9	69.2	41.9	25.9	63.1
ACL+	99.5	95.8	77.1	66.3	88.3
Ours	92.4	83.5	65.1	46.0	71.8

6. Conclusion

In this paper, we formulate a cross-view gait recognition task from a dynamic system perspective. We assume that people have different walking patterns, which can be

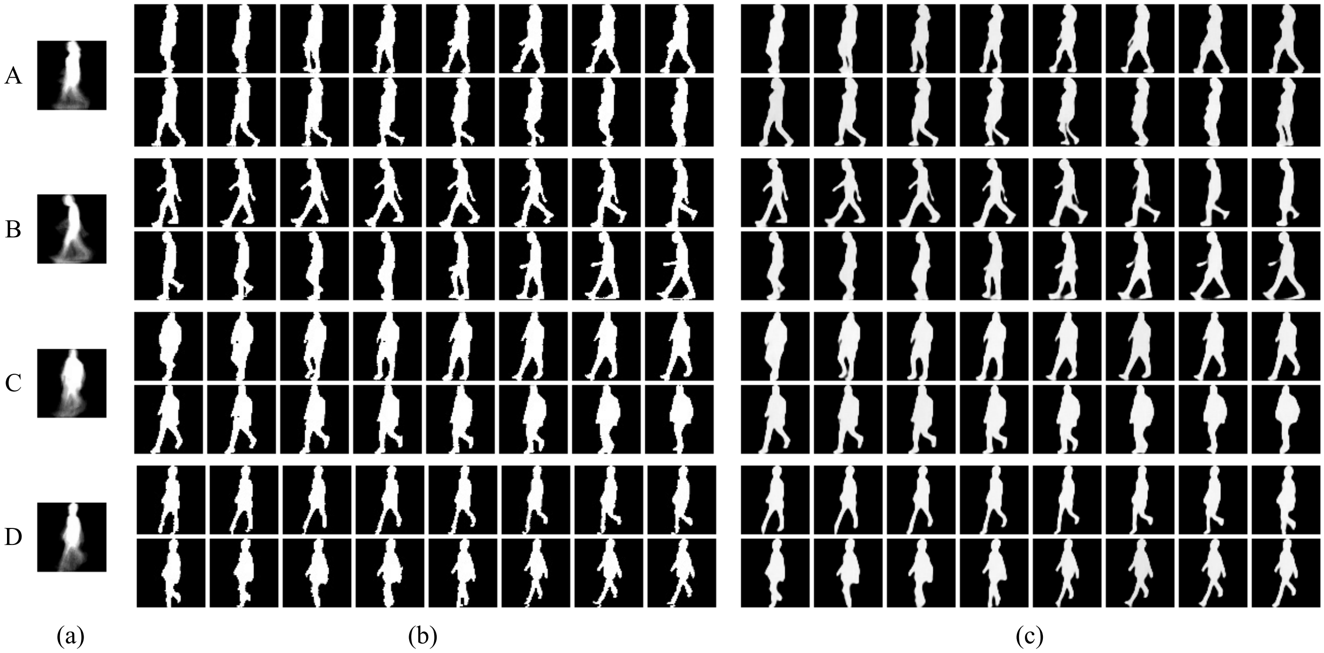


Figure 4. Visualization of the original gait images and the predicted images of four identities with two views: *A* and *B* are 90°; *C* and *D* are 45°. Each row represents the same identity with different gait images: (a) gait energy images, (b) original silhouette sequences $\{x_{i,t}|t \in [1,16]\}$, (c) predicted silhouette sequences $\{\tilde{\hat{x}}_{i,t+1}|t \in [1,16]\}$from the first frames of their original silhouette sequences $x_{i,t=1}$ and the estimated *Koopman* matrix $\hat{K}_i$. The four identities are from the first four samples from gallery testing set of the OU-MVLP dataset.

Table 4. Averaged rank-1 identification rates on OU-MVLP dataset, excluding identical-view cases, compared with GEINet [32], GaitSet [33], GaitPart [8], and ACL+local+temporal (ACL+) [50].

Probe	GEINet	GaitSet	GaitPart	ACL+	**Ours**
0°	11.4	79.5	82.6	74.0	56.2
15°	29.1	87.9	88.9	88.3	73.7
30°	41.5	89.9	90.8	94.6	81.4
45°	45.5	90.2	91.0	95.4	82.0
60°	39.5	88.1	89.7	88.0	78.4
75°	41.8	88.7	89.9	91.3	78.0
90°	38.9	87.8	89.5	90.0	76.5
180°	14.9	81.7	85.2	76.7	60.2
195°	33.1	86.7	88.1	89.5	72.0
210°	43.2	89.0	90.0	95.0	79.8
225°	45.6	89.3	90.1	94.9	80.2
240°	39.4	87.2	89.0	88.0	76.7
255°	40.5	87.8	89.1	90.8	76.3
270°	36.3	86.2	88.2	89.8	73.9
Mean	35.8	87.1	88.7	89.0	74.7

used for recognition against view changing. Then, to extract walking patterns from human silhouette sequences, the *Koopman* theory is applied for system linearization. We propose a framework based on convolutional variational auto-encoder and deep *Koopman* embedding, where gait systems are linearized. Therefore, the coefficient can be easily calculated to approximate the *Koopman* operator as walking patterns for recognition.

Finally, we conduct some experiments on a widely used gait database, OU-MVLP. The visualization and identification results prove that extracted gait dynamical features can represent human walking well, and it can also achieve an acceptable identification performance on cross-view gait recognition. Meanwhile, it provides physical interpretability for a gait recognition system. Overall, this study strengthens the idea that dynamical information contributes to gait recognition, which provides a new perspective and approach to gait recognition.

Acknowledgement

This work was supported by the Key Program of National Natural Science Foundation of China (Grant No. U20B2069) and CCF-Tencent Rhino-Bird Research Fund.

References

[1] Himanshu Aggarwal and Dinesh Kumar Vishwakarma. Covariate conscious approach for gait recognition based upon

zernike moment invariants. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):397–407, 2017. 2

[2] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020. 1, 3

[3] Gary P Austin. Motor control of human gait: A dynamic systems perspective. *Orthopaedic Physical Therapy Clinics of North America*, 10(1):17–34, 2001. 2

[4] Rafael Caldas, Marion Mundt, Wolfgang Potthast, Fernando Buarque de Lima Neto, and Bernd Markert. A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait & posture*, 57:204–210, 2017. 2

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8126–8133, 2019. 1, 2, 6

[6] Qiang Chen, Yunhong Wang, Zheng Liu, Qingjie Liu, and Di Huang. Feature map pooling for cross-view gait recognition based on silhouette sequence images. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 54–61. IEEE, 2017. 1, 2

[7] N Benjamin Erichson and Carl Donovan. Randomized low-rank dynamic mode decomposition for motion detection. *Computer Vision and Image Understanding*, 146:40–50, 2016. 2

[8] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14233, 2020. 1, 3, 7, 8

[9] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using lstm. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 325–330. IEEE, 2016. 1, 2

[10] Jacob Grosek and J Nathan Kutz. Dynamic mode decomposition for real-time background/foreground separation in video, 2014. 2

[11] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005. 2

[12] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2018. 2, 7

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 6

[14] Fabian Horst, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Müller, and Wolfgang I Schöllhorn. Explaining the unique nature of individual gait patterns with deep learning. *Scientific reports*, 9(1):1–13, 2019. 2

[15] F Horst, M Mildner, and WI Schöllhorn. One-year persistence of individual gait patterns identified in a follow-up study–a call for individualised diagnose and therapy. *Gait & posture*, 58:476–480, 2017. 2

[16] BingZhang Hu, Yan Gao, Yu Guan, Yang Long, Nicholas Lane, and Thomas Ploetz. Robust cross-view gait identification with evidence: A discriminant gait gan (diggan) approach on 10000 people, 2018. 7

[17] Yijun Huang, Yaling Liang, Zhisong Han, and Minghui Du. Two-stream convolutional network extracting effective spatiotemporal information for gait recognition. In *2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 43–48. IEEE, 2019. 1

[18] Yuanyuan Huang, Jianfu Zhang, Haohua Zhao, and Liqing Zhang. Attention-based network for cross-view gait recognition. In *International Conference on Neural Information Processing*, pages 489–498. Springer, 2018. 7

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4, 5

[21] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010. 2

[22] Na Li, Xinbo Zhao, and Chong Ma. A model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping, 2020. 1, 3

[23] Shuangqun Li, Wu Liu, and Huadong Ma. Attentive spatial–temporal summary networks for feature learning in irregular gait recognition. *IEEE Transactions on Multimedia*, 21(9):2361–2375, 2019. 1

[24] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 1, 3

[25] Zongyi Liu and Sudeep Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006. 2

[26] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):1–10, 2018. 2, 5

[27] Yasushi Makihara, Ryusuke Sagawa, Yasuhiro Mukaigawa, Tomio Echigo, and Yasushi Yagi. Gait recognition using a view transformation model in the frequency domain. In *European conference on computer vision*, pages 151–163. Springer, 2006. 7

[28] Daigo Muramatsu, Yasushi Makihara, Haruyuki Iwama, Takuya Tanoue, and Yasushi Yagi. Gait verification system for supporting criminal investigation. In *2013 2nd IAPR Asian Conference on Pattern Recognition*, pages 747–748. IEEE, 2013. 1

[29] Thanh Trung Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. *Pattern Recognition*, 47(1):228–237, 2014. 2

[30] Nobuyuki Otsu. Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In *Proceedings of the 6th International Conference on Pattern Recognition, 1982*, pages 557–560, 1982. 7

[31] Clarence W Rowley, IGOR MEZI?, Shervin Bagheri, Philipp Schlatter, DANS HENNINGSON, et al. Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, 641(1):115–127, 2009. 2, 3, 6

[32] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016. 2, 7, 8

[33] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern Recognition*, 96:106988, 2019. 1, 7, 8

[34] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 1

[35] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 7

[36] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, 2018. 2, 6

[37] Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony TS Ho. Detection of face spoofing using visual dynamics. *IEEE transactions on information forensics and security*, 10(4):762–777, 2015. 2

[38] Ashok Veeraraghavan, Amit K Roy-Chowdhury, and Rama Chellappa. Matching shape sequences in video with applications in human movement analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1896–1909, 2005. 2

[39] Jiawei Wang, Edel B Garcia, Shiqi Yu, and Dexin Zhang. Windowed dmd for gait recognition under clothing and carrying condition variations. In *Chinese Conference on Biometric Recognition*, pages 484–492. 2017. 2

[40] Wei Wang, Zheng Dang, Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Backpropagation-friendly eigendecomposition. In *Advances in Neural Information Processing Systems*, pages 3162–3170, 2019. 6

[41] Yanyun Wang, Chunfeng Song, Yan Huang, Zhenyu Wang, and Liang Wang. Learning view invariant gait features with two-stream gan. *Neurocomputing*, 339:245–254, 2019. 2

[42] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. 2, 3, 6

[43] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):209–226, 2016. 1, 2, 7

[44] Weiwei Xing, Ying Li, and Shunli Zhang. View-invariant gait recognition method by three-dimensional convolutional neural network. *Journal of Electronic Imaging*, 27(1):013010, 2018. 1, 2

[45] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Cross-view gait recognition using pairwise spatial transformer networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1, 7

[46] Shiqi Yu, Haifeng Chen, Edel B Garcia Reyes, and Norman Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–37. 2017. 2

[47] Shiqi Yu, Haifeng Chen, Qing Wang, Linlin Shen, and Yongzhen Huang. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, 239:81–93, 2017. 2

[48] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu. Siamese neural network based gait recognition for human identification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2832–2836. IEEE, 2016. 2, 7

[49] Xianfu Zhang, Shouqian Sun, Chao Li, Xiangyu Zhao, and Yuping Hu. Deepgait: A learning deep convolutional representation for gait recognition. In *Chinese Conference on Biometric Recognition*, pages 447–456. 2017. 2

[50] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *IEEE Transactions on Image Processing*, 29:1001–1015, 2019. 1, 3, 7, 8