Learning Multi-view Generator Network for Shared Representation

Tian Han†
Department of Statistics
University of California, Los Angeles
Los Angeles, California 90095
Email: hantian@ucla.edu

Xianglei Xing† ⋆
College of Automation
Harbin Engineering University
Harbin 150001, China
Email: xingxl@hrbeu.edu.cn

Ying Nian Wu
Department of Statistics
University of California, Los Angeles
Los Angeles, California 90095
Email: ywu@stat.ucla.edu

***Abstract*—Multi-view representation learning is challenging because different views contain both the common structure and the complex view specific information. The traditional generative models may not be effective in such situation, since view-specific and common information cannot be well separated, which may cause problems for downstream vision tasks. In this paper, we introduce a multi-view generator model to solve the problem of multi-view generation and recognition in a unified framework. We propose a multi-view alternating back-propagation algorithm to learn multi-view generator networks by allowing them to share common latent factors. Our experiments show that the proposed method is effective for both image generation and recognition. Specifically, we first qualitatively demonstrate that our model can rotate and complete faces accurately. Then we show that our model can achieve state-of-art or competitive recognition performances through quantitative comparisons.**

Keywords: Multi-view learning, Generator networks, Gait recognition

I. Introduction

Multi-view data have become increasingly accessible in many areas of scientific analysis including video surveillance, social computing, and environmental science, where data are collected from diverse domains, or described by different feature sets, or different "views." For example, a document can be described using both images and audios, or be described in multiple languages. Human's identity can be represented by multiple biometric features, such as face, gait, fingerprint, iris etc. Human faces or gait sequences captured by multiple cameras from different viewpoints can be utilized together for person identification in uncontrolled environment. Therefore, multi-view representation learning has wide applicability.

Consensus and complementarity are two main principles behind the multi-view representation learning [1], [2]. Consensus principle aims to maximize the agreement on the representations learned from multiple distinct views. For example, among unconstrained face recognition and gait recognition, pose variations or viewpoint variations are the bottleneck for real-world applications. In such applications, differences between intra-subjects under two views are usually much larger than the differences between inter-subjects under the same view, which makes the multi-view learning problem challenging. Consensus principle is commonly employed by various multi-view representation learning methods to obtain a common representation for multiple heterogeneous spaces. This usually comes in two types: (1) finding common subspace across different views, and (2) transforming non-normal views to normal view. For the first type, Canonical Correlation Analysis (CCA) based algorithms [3]–[6] and coupled projections (CP) based algorithms [7]–[9] are two representative methods which tend to find a common subspace such that the heterogeneous attributes can be eliminated in this consensus subspace. Specifically, CCA based algorithms aim to maximize the correlations (or the principal angles) of variables among different views, while CP based algorithms aim to minimize the distances between the projecting point pairs that have similar relations in the original heterogeneous sets. For the second type of consensus principle enforcement, there are several notable models which transform the data under different views into a normal view. Specifically, view transformation based model (VTM) [10]–[12] performs a linear transformation from non-normal views to normal view. Stacked progressive auto-encoders (SPAE) [13] models further extend VTM by using deep neural network to model complex non-linear transformations from the non-frontal face images to frontal ones in a progressive way. Recently, generative model based methods, such as generative adversarial networks (GAN) [14], have been employed to rotate images from non-normal views to the normal view. TP-GAN [15] is proposed to recover a frontal face in a data-driven way, while GaitGAN [16] is proposed to generate the side view gait images as invariant gait features for multi-view recognition task.

Different views of data usually contain complementary information, therefore the complementarity principle states that multiple views can be employed to comprehensively and accurately represent the data. The complementarity principle has also been employed by many multi-view representation learning methods in different ways. For the task of human identification at a distance, face and gait features are combined to obtain better performance than the methods that only use the face or gait feature alone. For example, Zhou and Bhanu [17] conducts feature concatenation after normalization and dimension reduction to fuse face and gait information. Xing

† These two authors contributed equally.
⋆ This work was done when the author was working as a visiting scholar at UCLA.

and Wang [18] fuses the gait and face features by computing the weighted mean of the two projecting features in the coupled subspace. For the task of human pose inference, the image features and different pose information are connected with one another in a latent space, which integrates the complementary information underlying different views [19].

In contrast to the existing multi-view representation learning methods, which usually utilize only the consensus principle or the complementarity principle alone, we introduce a novel multi-view generator network through shared latent representation (named MvGSR), which can essentially take advantage of both the consensus and complementarity principles.

The contributions of our paper are as follows:

- We propose the multi-view generator network for shared representation. The common knowledge for different views can be easily learned by shared latent factors, and view-specific characteristics can be effectively encoded through view-specific generator net.
- We propose to improve the multi-view recognition by generating data under arbitrary views through view-specific generator nets, thus complementary knowledge can be utilized through view-specific generator nets. Combined with the alternating back-propagation algorithm, we can efficiently and effectively transform the data from different views into any common view, therefore further boosting the performance on recognition.
- We conduct experiments both qualitatively and quantitatively. We show qualitatively that our method can be naturally used in image generation, including face rotation and face completion. We then quantitatively show that our method is competitive with or outperforms the state-of-art baselines in terms of multi-view gait recognition.

II. Model and Algorithm

A. Single-view generator model

Let $\mathbf{Y}$ be the observed signal of dimension D, such as an image, an audio and a video sequence. We assume that $\mathbf{Y}$ can be generated by a generator network [14]. At the top of the network is a layer of latent factors $\mathbf{Z} = (Z_k, k = 1, ..., d)$. The model can be treated as a non-linear generalization of factor analysis:

$$\begin{aligned} &\mathbf{Y} = G(\mathbf{Z}; W) + \epsilon, \\ &\mathbf{Z} \sim \mathrm{N}(0, \mathbf{I}_d),\ \epsilon \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_D), \end{aligned} \tag{1}$$

where $\mathbf{Z}$ and ϵ are independent. $G(\mathbf{Z}; W)$ is a non-linear transformation parametrized by a top-down convolutional neural network that consists of multiple layers of deconvolution, ReLU non-linearity, and up-sampling. W consists of all the weight and bias parameters of the network.

B. Multi-view generator model

The traditional generator model has been shown to be effective in image generation [14], [20]–[22]. However, if $\mathbf{Y}$ is heterogeneous or inhomogeneous, the original generator model can be less effective. The model tends to encode all the variations in the data in the latent factors in a highly non-linear and non-interpretable way. Although such $\mathbf{Z}$ can be used to generate sharp images, they hardly contain any useful information for downstream vision tasks. Therefore, we introduce a multi-view generator model to solve this problem.

Suppose $\mathbf{Y}$ contains signals that come from m different source domains, i.e., $\mathbf{Y} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, ..., \mathbf{Y}^{(m)}\}$. The number of domains can be determined by the specific application and we assume the signals are obtained across m domains. To effectively model the shared representations, we consider the following model (2):

$$\begin{cases} \mathbf{Y}^{(1)} = G_1(\mathbf{Z}; W_1) + \epsilon_1 \\ \mathbf{Y}^{(2)} = G_2(\mathbf{Z}; W_2) + \epsilon_2 \\ ... \\ \mathbf{Y}^{(m)} = G_m(\mathbf{Z}; W_m) + \epsilon_m \\ \mathbf{Z} \sim \mathrm{N}(0, \mathbf{I}_d)\ \epsilon_v \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_D) \end{cases} \tag{2}$$

where the vector of the latent factors $\mathbf{Z}$ is shared across signals from different domains, and G_v denotes the generator sub-network corresponding the v-th view, $v \in \{1, \ldots, m\}$. In this way, the domain or view specific variation is encoded through its corresponding view-specific generator sub-network, while the latent factors are forced to represent the common features among all the observations.

C. Multi-view alternating back-propagation

To learn from this multi-view generator model, we introduce a multi-view alternating back-propagation algorithm based on [22]. The generator model can be learned from training examples $\{\mathbf{Y}_i, i = 1, ..., n\}$ by the maximum likelihood estimation (MLE). Since the error term ϵ is normally distributed, the MLE is equivalent to L_2 reconstruction error $\sum_{i=1}^{n} \|\mathbf{Y}_i - G(\mathbf{Z}_i; W)\|^2$. The basic idea of learning is to iteratively optimize $\{\mathbf{Z}_i\}$ and W until convergence. More rigorously, since the latent factors are random variables, sampling method is employed to account for the uncertainty in $\mathbf{Z}_i$.

Specifically, for the training examples from m domains, i.e., $\mathbf{Y} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, ..., \mathbf{Y}^{(m)}\}$, the model can be written as $\mathbf{Z} \sim p(\mathbf{Z})$ and $[\mathbf{Y}^{(v)}|\mathbf{Z}, W_v] \sim p(\mathbf{Y}^{(v)}|\mathbf{Z}, W_v)$, where $v \in \{1, ...m\}$. The complete data log-likelihood is thus

$$\begin{aligned} \log p(\mathbf{Y}, \mathbf{Z}; W) &= \log \left[p(\mathbf{Z}) \prod_{v=1}^{m} p(\mathbf{Y}^{(v)}|\mathbf{Z}, W_v) \right] \\ &= -\sum_{v=1}^{m} \frac{1}{2\sigma^2} \|\mathbf{Y}^{(v)} - G_v(\mathbf{Z}; W_v)\|^2 - \frac{1}{2}\|\mathbf{Z}\|^2 + \mathrm{C}, \end{aligned} \tag{3}$$

where C denotes the constant w.r.t $\mathbf{Z}$ and W. The network parameter $W = \{W_1, W_2, ..., W_m\}$ is learned by maximizing the observed-data log-likelihood $L(W)$ which integrates out the unknown latent factors $\mathbf{Z}$. More precisely, the gradient of $L(W)$ can be obtained from:

$$\begin{aligned} \frac{\partial}{\partial W_v} L(W) &= \frac{\partial}{\partial W_v} \log p(\mathbf{Y}; W) \\ &= \mathbf{E}_{p(\mathbf{Z}|\mathbf{Y}, W)} \left[\frac{\partial}{\partial W_v} \log p(\mathbf{Y}^{(v)} \mid \mathbf{Z}, W_v) \right]. \end{aligned} \tag{4}$$

The expectation can be approximated by Monte Carlo samples drawn from the posterior distribution $p(\mathbf{Z}|\mathbf{Y}, W) \propto p(\mathbf{Y}, \mathbf{Z}; W) = p(\mathbf{Z}) \prod_{v=1}^{m} p(Y^{(v)}|\mathbf{Z}; W_v)$. Specifically, the latent factors are inferred using Langevin sampling method and they are updated as follows:

$$\mathbf{Z}_{t+1} = \mathbf{Z}_t + \frac{\delta^2}{2} \frac{\partial}{\partial \mathbf{Z}} \log p(\mathbf{Y}, \mathbf{Z}_t; W) + \delta \mathcal{E}_t \quad (5)$$

where the Gaussian standard random noise term $\mathcal{E}$ is added to prevent the stochastic gradient step to be trapped by the local modes, and δ denotes the step size of the Langevin dynamics. It can be shown that given sufficient transition steps, the obtained $\mathbf{Z}$ follows the posterior distribution. In this paper, the transition of $\mathbf{Z}$ starts from the updated latent factors from the previous learning iteration, so that the persistent updating results in a sufficiently long chain to sample from the posterior distribution while greatly reducing the computational burden in that we only need to use a small number of transition steps in each learning iteration.

For each training example $\mathbf{Y}_i$, we run Langevin dynamics Eq.(5) to get the corresponding posterior sample $\mathbf{Z}_i$, then this sample is used for gradient computation in Eq.(4). More precisely, the parameter W is learned through Monte Carlo approximation:

$$\frac{\partial}{\partial W_v} L(W) \approx \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial}{\partial W_v} \log p(\mathbf{Y}_i^{(v)} \mid \mathbf{Z}_i; W_v) \right]. \quad (6)$$

The whole algorithm iterates through two steps: (1) inferential step that infers the latent factors through the Langevin dynamics, and (2) learning step that updates the network parameter W by stochastic gradient descent. Gradient computations in both steps are powered by the efficient back-propagation. The left part of Figure 1 illustrates the structure and the training process of the proposed model.

D. Shared representation and recognition by generation

The proposed model can be used to learn the shared representation (the latent factors $\mathbf{Z}$) across multiple views. On the one hand, our proposed multi-view generator learning scheme is able to rotate data to different views and recover the incomplete data from multiple views in the training stage. On the other hand, the shared representation obtained from our proposed scheme contains some identity information and can be employed in recognition problem.

For multi-view recognition problem, suppose we have the gallery dataset $\mathbf{Y}^g$ and the probe dataset $\mathbf{Y}^p$ coming from the normal view and the test view, respectively. We can infer their latent factors $\mathbf{Z}^g$ and $\mathbf{Z}^p$, which are in the shared distribution, and perform classification in this common subspace. Specifically, in the testing phase, when given data from any view v, we can infer the corresponding latent vector $\mathbf{Z}^v$ using Eq.(5) and we change $\log p(\mathbf{Y}, \mathbf{Z}; W)$ by using only the given view v as:

$$\begin{aligned} \log p(\mathbf{Y}, \mathbf{Z}; W) &= \log \left[p(\mathbf{Z}) p(\mathbf{Y}^{(v)} | \mathbf{Z}, W_v) \right] \\ &= -\frac{1}{2\sigma^2} \|\mathbf{Y}^{(v)} - G_v(\mathbf{Z}; W_v)\|^2 - \frac{1}{2} \|\mathbf{Z}\|^2 + \mathrm{C}. \end{aligned} \quad (7)$$

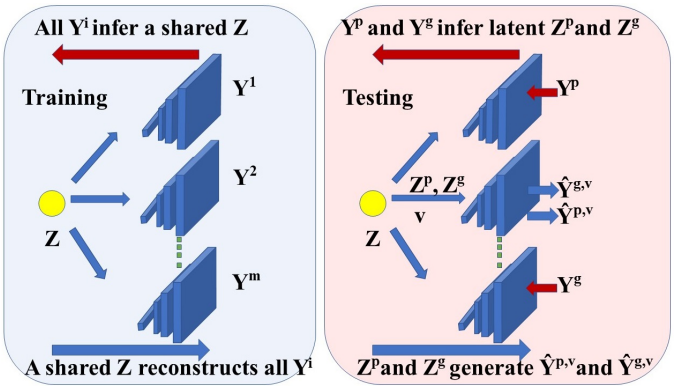


Fig. 1. Overview of our method. The left panel illustrates the training process of the proposed method. The right panel illustrates our method for multi-view recognition task. Suppose the probe dataset $\mathbf{Y}^p$ and the gallery dataset $\mathbf{Y}^g$ come from two different views p and g. Our method first infers their latent factors $\mathbf{Z}^g$ and $\mathbf{Z}^p$, and then generates $\hat{\mathbf{Y}}^{p,v}$ and $\hat{\mathbf{Y}}^{g,v}$ by feeding the inferred latent factors $\mathbf{Z}^p$ and $\mathbf{Z}^g$ to a sub-network under some common view. The complementary information for generating $\hat{\mathbf{Y}}^v$ under all views ($v \in \{1, \ldots, m\}$) and the latent factors $\mathbf{Z}$ are employed by our method.

Moreover, we can improve the multi-view recognition by generating data $\hat{\mathbf{Y}}^{g,v}$ and $\hat{\mathbf{Y}}^{p,v}$ by feeding the inferred latent factors $\mathbf{Z}^g$ and $\mathbf{Z}^p$ to the sub-generators for a common view $v \in \{1, \ldots, m\}$. Figure 1 illustrates this multi-view recognition by generation scheme. It is worth noting that although $\mathbf{Y}^g$ and $\mathbf{Y}^p$ are from different domains or views, the generated $\hat{\mathbf{Y}}^{g,v}$ and $\hat{\mathbf{Y}}^{p,v}$ obey the same distribution, and can be effective in recognition problem. Furthermore, we can perform the match-score fusion [23] which combines both the latent factors $\mathbf{Z}$ and the generated data $\hat{\mathbf{Y}}^v$ in all common views $v \in \{1, \ldots, m\}$. This scheme has the advantage of encoding both the common attributes (in $\mathbf{Z}$) and view-specific discriminating information (in $\hat{\mathbf{Y}}^v$) through a unified way. Let $\mathbf{D}(\cdot, \cdot)$ denote the L_2 distances between two matrices, and define $\mathbf{D}_\mathbf{Z} \equiv \mathbf{D}(\mathbf{z}^p, \mathbf{Z}^g)$, and $\mathbf{D}_\mathbf{Y}^v \equiv \mathbf{D}(\hat{\mathbf{y}}^{p,v}, \hat{\mathbf{Y}}^{g,v})$, where $\mathbf{D}_\mathbf{Z}$ and $\mathbf{D}_\mathbf{Y}^v$ are both vectors which calculate the L_2 distances between a probe identity and the gallery set containing N_g identities. The normalized match-score of latent factors $\mathbf{z}^p$ and $\mathbf{Z}^g$ can be computed as:

$$S_\mathbf{Z}(\mathbf{z}^p, \mathbf{Z}^g) = \frac{\exp\{-\mathbf{D}_\mathbf{Z}\}}{\sum_{i=1}^{N_g} \exp\{-\mathbf{D}_\mathbf{Z}\}(i)}, \quad (8)$$

and the normalized match-score of the generated $\hat{\mathbf{y}}^{p,v}$ and $\hat{\mathbf{Y}}^{g,v}$ can be computed as:

$$S_\mathbf{Y}^v(\hat{\mathbf{y}}^{p,v}, \hat{\mathbf{Y}}^{g,v}) = \frac{\exp\{-\mathbf{D}_\mathbf{Y}^v\}}{\sum_{i=1}^{N_g} \exp\{-\mathbf{D}_\mathbf{Y}^v\}(i)}. \quad (9)$$

Finally, we can fuse the match score of the latent factors $\mathbf{Z}$ and the generated data $\hat{\mathbf{Y}}^v$ in all common views $v \in \{1, \ldots, m\}$ to obtain the fusing match score S_F, defined as:

$$S_F = \sum_{v=1}^{m} w_i S_\mathbf{Y}^v + w_{m+1} S_\mathbf{Z}, \quad (10)$$

Fig. 2. Face rotation results for different subjects. First column: face image under standard pose (0°). Second to fifth column pairs: each pair shows the rotated face by our method (left) and the ground truth target (right).

where $w_i, i \in \{1, \cdots, m+1\}$ are the fusing coefficients which can be optimized on a validation set. The probe identity will be classified to the class for which the fusing match score is the largest. Utilizing the information from all the views helps the recognition problem, since different identities may be difficult to recognize in a particular view but may be easier in some other views. We shall elaborate on these points in the following experiment section. In this paper, we only focus on the image domain, but note that the proposed framework can also be used for other data domains.

III. Experiments

A. Experiment setup and network structure

For the first qualitative experiment on faces, we scale the face image so that the pixel intensities are within $[-1, 1]$, and we build up two generator nets, one for standard pose 0° (G_1) and one for another rotated pose (G_2). For each generator net, we adopt the structure similar to [21]. Specifically, we learn a 5-layer convNet which has deconvolutional kernel of size 4×4 and of stride 2, and has 512, 256, 128, 64, 3 filters from top to bottom. Each deconvolution layer is followed by ReLU non-linearity and batch normalization [24] except the last layer where tanh is used. For the quantitative experiment, we build up eleven generator sub-nets, where for each sub-net, we use the similar network structure except that we do not use batch normalization and we use the sigmoid in the last layer. All the input data for our experiments are normalized and cropped to 64×64. We use Adam optimizer [25] with initial learning rate 0.0002 and we run Langevin dynamics for 20 steps in each iteration with step size 0.1.

B. Qualitative results on the Multi-PIE face database

Multi-PIE database [26] is one of the most widely used face dataset which has a wide range of pose and illusion conditions. Specifically, it consists of 3 sessions of images of 249 identities under 15 poses and 20 different illumination conditions. We train our model on a selected subset which covers 5 poses, i.e., $\{-60^\circ, -30^\circ, 0^\circ, 30^\circ, 60^\circ\}$, and 50 randomly selected subjects under all illuminations. The subjects under odd numbers of illumination conditions are used for training and those under even numbers of illumination conditions are used for testing. We show qualitatively that our model cannot only rotating faces, but can also learn to recover the incomplete faces under different poses.

Fig. 3. Face completion results for different noise patterns. Each row represents the completion results under $B20$, $R0.5$ and $R0.9$ patterns respectively. In each row, the first and fourth columns represent the masked images. The second and fifth columns represent the images after filling in the missing parts using our method. The third and sixth columns represent the ground truth images.

1) Rotation: We first consider generating faces with the same identity and illumination condition but with the desired pose. The input data to our model are the image pairs which consist of the same subjects under the standard pose and the desired pose. In the testing stage, we have image pairs of 10 testing subjects under these two poses, and we use images under the standard pose as our gallery set and the other images as our testing set. After training, given the gallery set, we first run the Langevin dynamics for 200 steps based on G_1 to get the inferred latent factors. These factors are then fed into the other learned generator net G_2 to get the rotated images under the desired pose.

Figure 2 shows the qualitative results of rotating face images to -60°, -30°, 30°, and 60°. It can be seen that our shared generator model can accurately generate face images under different poses and they are visually similar to the ground truth testing images.

2) Completion: The proposed shared generator model can be adapted to the task of image completion in which we only consider the observed or visible pixels for inference.

We experiment with two types of noise patterns: one is the random pixel occlusion where we randomly block 50% or 90% pixels of the training images. The other is random patch occlusion where we randomly place a 20×20 block

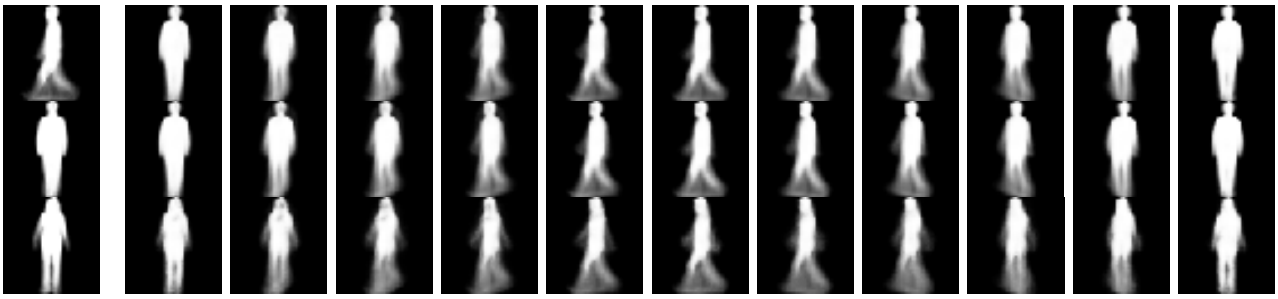


Fig. 4. Multi-view gait generation results for different subjects. The first column shows the original GEIs, and the second to twelfth columns show the generated GEIs under all the 11 views by our multi-view generative model. Specifically, the first row consists of one subject's original GEI under 90° view, and 11 generated GEIs by forward-propagating the inferred factors into the 11 sub-networks of our multi-view generator network. The second row consists of the same subject's original GEI under 0° view, and 11 generated GEIs under all the corresponding views as the first row. The third row consists of another subject's original GEI under 0° view, and 11 generated GEIs under all of the corresponding views as the first two rows. We can observe that, under each view, the generated GEIs from the same subject are very similar, while the generated GEIs from different subjects can be easily distinguished.

on each training image. We denote these random patterns as $R0.5$, $R0.9$ and $B20$ respectively.

Figure 3 shows completion results. It is clear that even under severe occlusion, the proposed method can still get sharp results. We compute the average per-pixel difference between the recovered images and the corresponding ground truth images to measure the recovery quality. For $R0.5$, we have recover error 0.256, for $R0.9$ and $B20$, we have recover errors 0.404 and 0.328 respectively.

Note that this task is challenging for existing methods to learn useful representations [14], [20], [21]. Recently, [22] proposes to directly learn the generator model from incomplete data and shows promising initial results. However, their method is only applied to one front view, whereas the problem of how to efficiently learn multi-views remains unexplored. Our proposed model directly shares the latent factors across different views. Therefore, each sub-domain is forced to communicate with other domains to agree upon the common knowledge. Besides, each sub-domain only contains the images from one view, thus a better generator net can be learned.

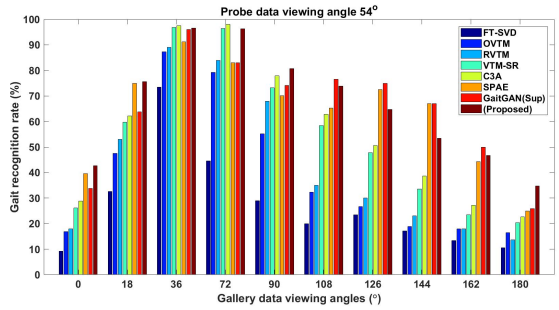
C. Experiments on the CASIA gait database B

The CASIA gait database B [27] is one of the largest multi-view gait databases. It contains 124 subjects captured from 11 viewing angles, namely 0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, and 180°. Under each view, there are ten gait sequences for each person, including six sequences of normal walking, two sequences of walking with a bag, and two sequences of walking with a coat. We experiment on the gait sequences of normal walking. Since Gait Energy Image (GEI) [28] is a popular gait feature, which is efficient for computation and is robust to noise, we employ it for our method's input and target images.

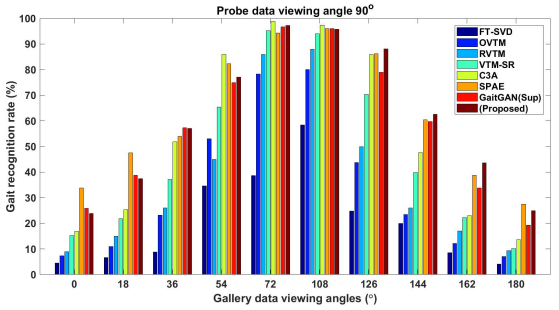
According to the experimental protocol defined in [16], [29], the database is divided into two groups: the first group of the 62 subjects is used for constructing the training set and the remaining 62 subjects are used for performance evaluation. Suppose the gallery view and the probe view are denoted by θ_G and θ_P, respectively. In the training phase, we use the gait sequences from the training group under all the 11 views to alternatively infer the common identity factors and learn the deconvolutional sub-network under each view. In the recognition phase, the sequences from the evaluating group under the gallery view θ_G are utilized to construct the gallery set and the sequences from the evaluating group under the probe view θ_P are utilized to form the probe set. Our method can utilize two kinds of information learned from the proposed model to perform multi-view gait recognition: (1) The shared factors (the latent factors $\mathbf{Z}$ in Figure 1). For both the gallery and probe GEIs, we infer their shared factors and use these factors for recognition. (2) The generation of GEIs. For both the gallery and probe GEIs, we first infer their shared factors, and then generate new GEIs by feeding the inferred latent factors into the sub-networks under some common views. Figure 4 shows the generated GEIs through our multi-view generative model. We employ these generated GEIs for further classification. Our method performs a match-score fusion of the above two kinds of information for the final multi-view gait recognition (see Section II.D for detail).

To evaluate the performance of our algorithm, we compared the proposed MvGSR with the following state-of-art methods: FT-SVD [30], OVTM [10], RVTM [11],VTM-SR [12] , C3A [6], SPAE [29], and GaitGAN [16]. As in experiments of those methods, the probe view θ_P is selected as 54°, 90° or 126°. For each selected probe view, we test on the gallery view θ_G from the rest 10 viewing angles except the corresponding probe view.

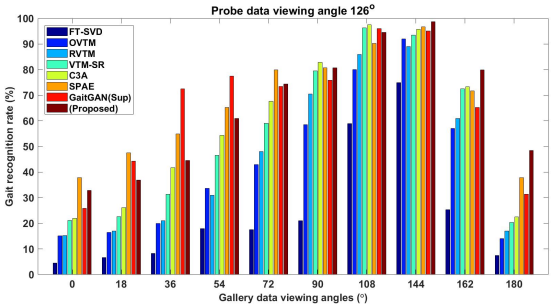
The experimental results are plotted in Figure 5, which reveals several interesting observations. (1) Most of the algorithms can obtain a high recognition rate when the viewing angle difference between the gallery and probe is small. The reason behind this high recognition rate is that gaits between two closer views share more common information, therefore only simple transformation of gaits is needed to have a good performance. (2) When the angle difference between the gallery and the probe is large, the deep learning gait recognition methods, including SPAE, GaitGAN and the proposed



(a)



(b)



(c)

Fig. 5. Comparisons of gait recognition rates for multi-view gait recognition using different methods. The viewing angles of the probe data in subfigures (a)-(c) are $54°$, $90°$ and $126°$, respectively. The viewing angles of the gallery data are the rest 10 viewing angles except the corresponding probe viewing angle.

MvGSR, perform better than the traditional linear methods, such as FT-SVD, OVTM, RVTM, VTM-SR, and C3A. This is because the deep learning methods can approximate highly non-linear transformations, which are important for this task with large viewpoint variation. (3) Generally speaking, the proposed MvGSR method performs better then the other two deep learning methods in most of the views. The proposed method cannot only improve the recognition rate when the viewpoint variation is not large, but it can also handle large viewpoint variation quite well.

In Table I, we further summarize the experimental results of three unsupervised methods, including C3A, SPAE, and the proposed MvGSR, which obtain good performances as shown in Figure 5, where we also include three supervised methods, ViDP [31], CNN [32], and GaitGAN. It is worth noting that

TABLE I

COMPARISONS OF THE AVERAGE RECOGNITION ACCURACY UNDER THE PROBE VIEWS $\theta_P = 54°, \theta_P = 90°, \theta_P = 126°$, WHERE THE GALLERY VIEW ARE THE REST 10 VIEWING ANGLES EXCEPT THE CORRESPONDING PROBE VIEWING ANGLE. THE VALUES IN THE RIGHT MOST COLUMN ARE THE AVERAGES RATES AT THE THREE PROBE VIEWS $\theta_P = 54°, \theta_P = 90°, \theta_P = 126°$.

Methods\Probe angles	$\theta_P = 54°$	$\theta_P = 90°$	$\theta_P = 126°$	Average
C3A	56.64%	54.65%	58.38%	56.56%
ViDP (Supervised)	64.2%	60.4%	65.0%	63.2%
CNN (Supervised)	77.8%	64.9%	76.1%	72.9%
SPAE	63.31%	62.1%	66.29%	63.9%
GaitGAN (Supervised)	64.52%	58.15%	65.73%	62.8%
MvGSR	66.53%	60.78%	65.19%	64.17%

it is generally unfair to directly compare the unsupervised method with the supervised one, since the latter one needs to utilize the label or other auxiliary information for training. As we can observe from Table I, our method, as an unsupervised generative model, obtains comparable results with these state-of-art methods, and even performs better than some supervised methods, such as ViDP and GaitGAN.

IV. CONCLUSION

This paper proposes to learn multi-view generator networks through shared latent factors. We argue that the learned shared representation can effectively encode the common knowledge across different views, and the domain specific generator networks can accurately obtain the domain related information. Therefore, the proposed method can naturally enforce the consensus and complementarity principles which a good shared representation should follow. We conduct qualitative experiments on face rotation and completion, demonstrating that our method can be effectively utilized for generation tasks. We also conduct quantitative comparisons with existing methods for gait recognition, showing that our method, though trained in unsupervised manner, is competitive or even becomes state-of-art in many cases.

Future directions. The proposed model can be further investigated in two ways: First, we can extend our framework for multi-modality learning in which data are coming from different source domains, e.g., text, audio, iris etc. Second, the input data to our model can be further relaxed so that the data in different views do not need to have one to one correspondence. We leave these as our future research.

ACKNOWLEDGMENT

This work was supported by DARPA SIMPLEX N66001-15-C-4035, ONR MURI N00014-16-1-2007, DARPA ARO W911NF-16-1-0579, DARPA N66001-17-2-4029, Natural Science Foundation of China No. 61703119, Natural Science Fund of Heilongjiang Province of China No. QC2017070 and Fundamental Research Funds for the Central Universities of China No. HEUCFM180405.

REFERENCES

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013. 1

[2] Y. Li, M. Yang, and Z. Zhang, "Multi-view representation learning: A survey from shallow methods to deep methods," *arXiv preprint arXiv:1610.01206*, 2016. 1

[3] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004. 1

[4] T. Sun and S. Chen, "Locality preserving cca with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, no. 5, pp. 531–543, 2007. 1

[5] X. Shen, Q. Sun, and Y. Yuan, "A unified multiset canonical correlation analysis framework based on graph embedding for multiple feature extraction," *Neurocomputing*, vol. 148, no. 19, pp. 397–408, 2014. 1

[6] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognition*, vol. 50, pp. 107–117, 2016. 1, 5

[7] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal processing letters*, vol. 17, no. 1, pp. 20–23, 2010. 1

[8] K. Wang, X. Xing, T. Yan, and Z. Lv, "Couple metric learning based on separable criteria with its application in cross-view gait recognition," in *Chinese Conference on Biometric Recognition*. Springer, 2014, pp. 347–356. 1

[9] X. Xing and K. Wang, "Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition," *Signal Processing*, vol. 125, pp. 329–335, 2016. 1

[10] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *IEEE 12th International Conference on Computer Vision Workshops*. IEEE, 2009, pp. 1058–1064. 1, 5

[11] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 2073–2076. 1, 5

[12] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 6, pp. 966–980, 2012. 1, 5

[13] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1883–1890. 1

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680. 1, 2, 5

[15] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," *arXiv preprint arXiv:1704.04086*, 2017. 1

[16] S. Yu, H. Chen, E. B. G. Reyes, and P. Norman, "Gaitgan: invariant gait feature extraction using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 30–37. 1, 5

[17] X. Zhou and B. Bhanu, "Feature fusion of face and gait for human recognition at a distance in video," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4. IEEE, 2006, pp. 529–532. 1

[18] X. Xing, K. Wang, and Z. Lv, "Fusion of gait and facial features using coupled projections for people identification at a distance," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2349–2353, 2015. 2

[19] L. Sigal, R. Memisevic, and D. J. Fleet, "Shared kernel information embedding for discriminative inference," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2852–2859. 2

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 2014. 2, 5

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 2, 4, 5

[22] T. Han, Y. Lu, S.-C. Zhu, and Y. N. Wu, "Alternating back-propagation for generator network." in *AAAI*, 2017, pp. 1976–1984. 2, 5

[23] X. Zhou and B. Bhanu, "Integrating face and gait for human recognition at a distance in video," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 5, pp. 1119–1137, 2007. 3

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015. 4

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 4

[26] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Comput.*, vol. 28, no. 5, pp. 807–813, May 2010. 4

[27] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition*, vol. 4. IEEE, 2006, pp. 441–444. 5

[28] Z. Lv, X. Xing, K. Wang, and D. Guan, "Class energy image analysis for video sensor-based gait recognition: A review," *Sensors*, vol. 15, no. 1, pp. 932–964, 2015. 5

[29] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, 2017. 5

[30] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Computer Vision-ECCV 2006*. Springer, 2006, pp. 151–163. 5

[31] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2034–2045, 2013. 6

[32] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 209–226, 2017. 6