Neurocomputing 421 (2021) 244–259

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# A diversified shared latent variable model for efficient image characteristics extraction and modelling

Hao Xiong [a,*], Yuan Yan Tang [b], Fionn Murtagh [c], Leszek Rutkowski [d,e], Shlomo Berkovsky [a]

[a] Australian Institute of Health Innovation, Macquarie University, North Ryde NSW 2113, Australia
[b] Faculty of Science and Technology, University of Macau, Macau 999078, China
[c] Centre for Mathematics and Data Science, University of Huddersfield, United Kingdom
[d] Institute of Computational Intelligence, Czesto-chowa University of Technology, ul.Armii Krajowej 36, 42-200 Czestochowa, Poland
[e] Information Technology Institute, University of Social Sciences, Poland

ARTICLE INFO

*Article history:*
Received 23 November 2019
Revised 4 September 2020
Accepted 20 September 2020
Available online 30 September 2020
Communicated by Steven Hoi

*Keywords:*
Latent Variable Model
Multi-view Learning
Diversity-encouraging Prior
Variational Inference

ABSTRACT

An object can be consisting of various attributes, such as illuminance, appearance, shape, orientation, etc. Separately extract these attributes has enormous value in visual effects modeling, attribute-specific retrieval and recognition. Essentially, these attributes can be fairly abstract and thus need labels to extract. However, sometimes the labels of these attributes may not be available with training data. A solution to this problem is projecting the observed data into a lower dimension latent subspace, such that each observed data can be represented by a latent variable. After that, the dimensions of a latent variable can be segmented into different parts by weighting the kernel automatic relevance determination (ARD) parameters. Consequently, the latent variable is segmented into different parts each of which corresponds to the main attribute. In real life scenery, the attributes of an object may vary significantly from case to case. For instance, a single face can probably be under different illuminance conditions. Taking into account the diversity of these attribute variations, we propose the Diversified Shared Latent Variable Model (DSLVM) to extract and manipulate object attributes in an unsupervised way. More specifically, we initially set up two views that share the same latent variables. Then, two Diversity Encouraging (DE) priors are applied to the inducing points of each model view. Here, the inducing points are a small representative dataset that explains the observed data in its entirety. Meanwhile, the exploited diversity encouraging priors are able to cover more diverse characteristics of the attributes. The defined objective function is computed by variational inference. Extensive experiments on different datasets demonstrate that our method can accurately deal with various object.



## 1. Introduction

Extracting and modeling objects' characteristics separately is vital in movie special effects, video games, augmented reality. However, the labels of these characteristics for training the corresponding model could be scarce or even not available. Therefore, we propose the diversified shared latent variable model in an unsupervised way to robustly extract the attributes without the availability of labels.

Statistical methods, such as PCA based methods [43,44,18,20,25,15,26], are often employed to project the observed data into a lower dimensional space. Therefore, the latent space captures the features of characteristics from the observed data. The object can then be modeled by sampling on the projected continuous latent space. In this scenario, the mapping between the observed data and its subspace is linear, but the correlation between the observed data and latent space is not always linear. [27,45] proposed Gaussian process-based models to enable non-linear processes using a kernel function. However, these methods are still unable segmenting the latent space and thus differentiate the characteristics of objects.

Meanwhile, some works [9,39,12] have attempted to learn a shared latent variable to represent the observed data with multiple views. However, most shared latent variable model based approaches are not able to perform a discrete segmentation on share latent variable. In contrast, Damianou et al. [9] proposed Manifold Relevance Determination (MRD) to capture the structure underlying the high-dimension face data by segmenting latent space into shared and private subspaces, useful for separately modeling appearance and illuminance from the face image. After sampling from the shared subspace, a new face image is generated

* Corresponding author.
*E-mail address:* hao.xiong@mq.edu.au (H. Xiong).

https://doi.org/10.1016/j.neucom.2020.09.035
0925-2312/

under various illumination conditions while retaining an unchanged facial appearance. Similarly, sampling the private subspace only changes the facial appearance.

However, [9] does not model shaded faces very well. Theoretically, when sampling the shared dimensions around the subspace of a dark shadow, the model should also generate a new face with dark shadows. However, MRD tends to generate a new face without any shadows. Furthermore, by varying the private subspace, MRD is also inclined to generate a suboptimal face that is blurred or with an altered appearance. The root cause of these phenomena is that the latent variables generated by MRD focus on face images with salient features. That is to say, illuminated faces are more salient and featured than shaded faces. Thus, illuminated faces attract more latent variables than shaded faces. As a consequence, the trained latent subspace from MRD cannot account for all possible illumination conditions and appearance variations.

To overcome the above issues of MRD method, we propose DSLVM to robustly extract object characteristics by concurrently exploiting multi-view learning and DE priors. Specifically, we construct a diversified shared latent variable model associated with two views under a Gaussian process framework. These two views share the same latent variable but separately estimate two ARD parameter vectors. Here, the effect of the ARD parameters is to determine the shared and private latent spaces, where the shared dimensions model the variations of a common attribute such as illuminance, orientation. Likewise, private dimensions model private attribute like object appearance. The main contributions of our work are threefold:

- For each view, the newly designed DE prior is introduced to the inducing points, which is a small representative set of a large amount of observed data. As a consequence, the inducing points become more diverse and hence capture more distinct features of each attribute from observed data. For instance, the illuminance conditions in a given face dataset are probably not evenly distributed. As a result, the learned model is inclined to biased without DE prior.
- The model is learned in an unsupervised way (no labels required). The newly defined objective function of our model is not solvable, so we apply the variational inference to approximate the objective function with a lower bound; The solution is then obtained by maximizing the lower bound.
- We experiment with three different databases including faces under various lighting conditions and 3D chairs with various orientations. The experimental results show that our method outperforms the baselines. It further demonstrate the robustness and effectiveness of our approach for image characteristics extraction and modelling in an unsupervised way.

We first review existing works on latent variable models. Then, we provide a brief introduction to MRD before demonstrating our proposed DSLVM. Finally, we perform modeling experiments on three representative databases.

## 2. Related work

Latent variable models are broadly classified into single latent space based models, hierarchical latent variable models and shared latent variable models. In this section, we review some previous latent variable based approaches for both single view and multi-view learning.

Gaussian process latent variable model (GP-LVM) [27] can be considered as a classical latent variable model which is often used in dimension reduction in high dimensional data. The latent variables are regarded as low dimensional features of data. The whole model is optimized rather than integrating out the latent variables. Meanwhile, GPLVM can be seen as a nonlinear extension of the linear probabilistic PCA (PPCA)[32]. After that, a Bayesian Gaussian process latent variable model [11] is proposed to variationally integrate out the latent variables in the GPLVM. By exploiting Jensen's inequality and variational inference, the solution to the log marginal likelihood of data is tractable. In order to speed up computation, inducing points [8,35,7,4] are further exploited.

The hierarchical latent variable model refers to build up multiple latent spaces which are hierarchically stacked. Neil et al. [28] proposed a tree-like hierarchical latent variable model to express conditional independencies in the data as well as the manifold structure. Then, [10] imitated the concept of deep neural network and built multiple latent spaces layer by layer to model complex data. Meanwhile, recurrent neural networks (RNNs) [14,19,6,41] can well process sequential data. As a consequence, the works [36,33] proposed RNN-like hierarchical latent variable models for dialogue generation and audio processing.

Shared latent variable model, in general, refers to the latent variables that are shared by two or multiple views. In early multi-view learning, canonical correlation analysis (CCA) [16,42,46] draws considerable attention. In CCA, two projection matrices are learned so that the two views can be projected into a common subspace. Afterward, several variants of CCA were proposed as the extensions. In [2], a sparsity prior was added into the traditional CCA. Besides, the $l_1$ loss and a Student-$t$ [31] were added to limit the influence of outliers and noise. Rather than focusing on two views learning, the works [34,5] further extend the aforementioned methods to multi-view learning. As a matter of fact that these methods are based on linear projection function, which can not be applicable to nonlinear input features. More recently, multi-view learning were applied to recognition tasks such as human action recognition [22], facial expression recognition [23,1,30] and object recognition [17]. For instance, Muhammad et al. [22] proposed to compute multi-view features from horizontal and vertical gradients. Then, combined with features extracted from pre-trained CNN network, a best feature set can be selected and be fed into Naive Bayes classifier for human action recognition.

Some methods aim to perform supervised multi-view analysis. For instance, [21] proposed Multi-view Discriminative Analysis (MvDA) which maximizes the between-class and minimizes the within-class variations in a common subspace. Here, MvDA essentially extends Linear Discriminant Analysis (LDA) [3] to multi-view case. Besides, there are some Generalized Multiview Analysis (GMA) [37] based methods proposed for multiview learning. A case in point is the Generalized Multiview LDA (GMLDA). Likewise, GMLDA tried to unite different views of the same class and separate the content of different classes in a common subspace. Another example of GMA is the GM Locality Preserving Projections (GMLPP). Based on the work of LPP [32], it is able to find a discriminative data manifold using labels. However, these supervised learning based methods are not practical since the training labels are not always available.

Guoli et al. [38] proposed multi-modal similarity Gaussian process latent variable model (m-SimGP) to learn the nonlinear mapping functions between heterogeneous modalities and the shared latent space. Based on m-SimGP, multi-modal regularized similarity GPLVM (m-RSimGP) was further encouraging similar/dissimilar shared latent variables to be similar/dissimilar. Later on, the same authors [39] proposed multimodel distance-preserved similarity GPLVM (m-DSimGP) to preserve the intra-modal global similarity structure, and multimodel regularized similarity GPLVM (m-RSimGP). Recently, Guoli *et al.* proposed the harmonized multimodal learning with Gaussian process latent variable models [40], which developed a new learning scheme "Harmonization" so that the latent variables can be learned jointly from all modalities by exploiting strong complementarity among different modalities. By referring to the framework of autoencoder, [29] introduced other nonlinear projections from observation to shared latent space. Meanwhile, a discriminative prior was integrated to encourage latent variables from same/ different classes to be close/far. Besides, shared latent variable model can also be used in face recognition tasks. In [13], Stefano et al. built a model that learned a discriminative manifold shared by multiple views of facial expression. Then, facial expression recognition was performed in the expression manifold.

In general, the existing shared latent variable models focus on the tasks such as image classification, audio-visual speech recognition, image captioning and cross modal retrieval. Therefore, they are unable to synthesize images properly. Though methods with a single or hierarchical latent space can synthesize images, they are unable to segment the latent space and thus cannot extract characteristics of images. To the best of our knowledge, MRD [9] is the only shared latent variable model that is able to extract and separate image characteristics by segmenting the latent space. Therefore, we exploit MRD as a baseline in our experiment.

## 3. Manifold relevance determination

In MRD, $Y \in R^{n\times d}$ (with columns $\{y_{:j}\}_{j=1}^{d}$) denotes one view of the observed data, where n is the number of data points and d is the dimensionality of each data point in Y. These data are associated with latent variables $X \in R^{n\times q}$ for the sake of dimensionality reduction. Then, the likelihood function is defined as:

$$p(Y|X) = \prod_{j=1}^{d} p(y_{:j}|X), \tag{1}$$

where $y_{:j}$ represents the $j$th column of $Y$ and

$$p(y_{:j}|X) = N(y_{:j}|0, K_{ff} + \sigma^{-1} I_n). \tag{2}$$

$K_{ff}$ is an $n \times n$ kernel matrix, and the kernel function is an exponentiated quadratic (RBF):

$$k(x_{i,:}, x_{k,:}) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{j=1}^{q}\alpha_j (x_{i,j} - x_{k,j})^2\right), \tag{3}$$

where each $x_{i,:}$ is the $i$th row of X, and $K_{ff} = k(x_{i,:}, x_{k,:})$. The marginal likelihood function is defined as:

$$p(Y) = \int p(Y|X)p(X)\mathrm{d}X. \tag{4}$$

Since Y is observed data which is supposed to be noisy, another latent variable F is introduced that is considered as the unnoisy version of Y, and

$$p(Y|F) = \prod_{j=1}^{d} N(y_{:j}|f_{:j}, \sigma^{-1} I_n). \tag{5}$$

Here, $f_{:j}$ is the $j$th column of variable F which is the same size as Y. For the sake of dimensionality reduction, the conditional distribution in Eq. (1) becomes

$$p(F|X) = od_{j=1}^{d} N(f_{:j}|0, K_{ff}). \tag{6}$$

The latent variables are i.i.d. and the prior distribution p(X) is obtained by selecting a fully factorized latent space prior:

$$p(X) = \prod_{j=1}^{q} N(x_{:j}|0, I_q), \tag{7}$$

where $x_{:j}$ refers to $j$th column of X.

In MRD, another view $Z \in R^{n\times d_Z}$ of the observed data will be incorporated into the same model. However, it still assumes a single latent variable $X$ shared by the two views $Y$ and $Z$. Likewise, the nonlinear mapping from latent variable $X$ to observations $Y$ and $Z$ is the same.

## 4. Diversified shared latent variable model

In this section, we explain our diversified shared latent variable model by extracting the characteristics of illuminance and appearance from faces. Initially, we build up two views. The first view $Y \in R^{n\times p}$ contains $n$ face images, each of which has $p$ pixels. These face images consist of several distinct individuals captured under various possible lighting conditions. Likewise, an analogous view $Z \in R^{n\times p}$ is set up and defined. This provides two variations in the views $Y$ and $Z$, which are the illuminance conditions and the subject's appearance, respectively.

### 4.1. The model

Our model contains only a single latent variable $X \in R^{n\times q}$. With the mappings $\left\{f_d^Y\right\}_{d=1}^{p} : X \mapsto Y$ and $\left\{f_d^Z\right\}_{d=1}^{p} : X \mapsto Z$, the shared latent variable $X$ provides a low dimensional representation of either $Y$ or $Z$. Following [11], we assume that the data are corrupted by additive Gaussian noise $\epsilon^{\{Y,Z\}} \sim N\left(0, \sigma_\epsilon^{\{Y,Z\}} I\right)$,

$$\begin{aligned} y_{nd} &= f_d^Y(x_n) + \epsilon_{nd}^Y, \\ z_{nd} &= f_d^Z(x_n) + \epsilon_{nd}^Z, \end{aligned} \tag{8}$$

where $\{y, z\}_{nd}$ denotes the dimension $d$ of point $n$ in $Y$ and $Z$.

Moreover, our latent functions $f_d^Y$ and $f_d^Z$ are selected to be independent draws of a zero-mean GP with an ARD covariance function of the form:

$$\begin{aligned} k^Y(x_{i,:}, x_{k,:}) &= \left(\sigma_f^Y\right)^2 \exp\left(-\tfrac{1}{2}\sum_{j=1}^{q}\alpha_j^Y (x_{i,j} - x_{k,j})^2\right), \\ k^Z(x_{i,:}, x_{k,:}) &= \left(\sigma_f^Z\right)^2 \exp\left(-\tfrac{1}{2}\sum_{j=1}^{q}\alpha_j^Z (x_{i,j} - x_{k,j})^2\right). \end{aligned} \tag{9}$$

In the GPLVM framework, each generative mapping is modeled as a product of an independent GP that is parameterized by a covariance matrix $K_{ff}^{\{Y,Z\}}$ evaluated over the latent variable $X$, so that

$$\begin{aligned} p\left(F^Y|X, \theta^Y\right) &= \prod_{j=1}^{p} N\left(f_{:j}^Y|0, K_{ff}^Y\right), \\ p\left(F^Z|X, \theta^Z\right) &= \prod_{j=1}^{p} N\left(f_{:j}^Z|0, K_{ff}^Z\right), \end{aligned} \tag{10}$$

where $F^Y = \left\{f_{:j}^Y\right\}_{j=1}^{p}, F^Z = \left\{f_{:j}^Z\right\}_{j=1}^{p}$, and $\theta^{Y,Z}$ collectively denote the parameters of covariance matrices $K_{ff}^{\{Y,Z\}}$ and the noise variances $\epsilon^{Y,Z}$. Besides, each element in covariance matrices $K_{ff}^{\{Y,Z\}}$ is computed by covariance function (Eq. 9).

This leads to a likelihood function under the model:

$$p(Y, Z|X, \theta) = \prod_{\kappa=\{Y,Z\}} \int p(\kappa|F^\kappa)p(F^\kappa|X, \theta^\kappa)dF^\kappa. \tag{11}$$

Then, the joint marginal likelihood requires integration over the latent variable X in Eq. (11) and is:

$$p(Y,Z|\theta) = \int p(Y,Z|X,\theta)p(X)dX. \tag{12}$$

Since X appears non-linearly in the inverse of the covariance matrices $K_{ff}^{Y}$ and $K_{ff}^{Z}$, the integration over the latent variable X is intractable. To overcome this problem, we derive an approximate Bayesian training and inference procedure by variationally marginalizing out $X$.

To facilitate the variational inference (explained below), likelihoods $p\left(F^{Y}|X,\theta^{Y}\right)$ and $p\left(F^{Z}|X,\theta^{Z}\right)$ are augmented with $m$ inducing points $U^{Y}$ and $U^{Z}$, then $p(Y,Z|X,\theta)$ is augmented as:

$$p\left(Y,Z|U^{\{Y,Z\}},X,\theta\right) = \prod_{\kappa=\{Y,Z\}} \int p(\kappa|F^{\kappa})p(F^{\kappa}|U^{\kappa},X,\theta^{\kappa})dF^{\kappa}. \tag{13}$$

Here, the inducing points $U^{Y},U^{Z} \in R^{m\times p}$ are evaluated at the pseudo-inputs $\overline{X}^{Y},\overline{X}^{Z} \in R^{m\times q}$, respectively. The marginal GP priors over the inducing points $U^{Y}$ and $U^{Z}$ are:

$$p\left(U^{\{Y,Z\}}|\overline{X}^{\{Y,Z\}}\right) = \prod_{j=1}^{q} p\left(u_{:,j}^{\{Y,Z\}}|\overline{X}^{\{Y,Z\}}\right) \tag{14}$$

$$= \prod_{j=1}^{q} N\left(u_{:,j}^{\{Y,Z\}}|0,K_{uu}^{\{Y,Z\}}\right), \tag{15}$$

where $K_{uu}^{\{Y,Z\}} = k\left(\overline{X}_{j,:}^{\{Y,Z\}},\overline{X}_{k,:}^{\{Y,Z\}}\right)$.

Since the pseudo-inputs $\overline{X}^{\{Y,Z\}}$ are variational parameters, they can be dropped from the expressions. Then, the joint marginal distribution $p(Y,Z)$ is:

$$p(Y,Z) = \prod_{\kappa=\{Y,Z\}} \int p(\kappa|F^{\kappa})p(F^{\kappa}|U^{\kappa},X)p(U^{\kappa})dU^{\kappa}dF^{\kappa}p(X)dX. \tag{16}$$

#### 4.1.1. Diversity encouraging prior

However, it transpires that the latent variables $X$ are highly likely to be similar after training. Hence, we introduce a DE prior with respect to the latent variable repulsion property due to the fact that real-world attributes may be much more diverse, e.g various illuminance conditions, different human ethnicities, than expected.

In our model, a kernel-based diversity prior $K_{\phi}$ is defined as follows:

$$K(\phi_i,\phi_j;\rho) = \int_{\chi} f(x|\phi_i)^{\rho} f(x|\phi_j)^{\rho} dx, \tag{17}$$

Here, a probability product kernel is constructed to define the repulsion. Therefore, every kernel element is expressed as the inner product of probability distributions. Besides, the latent variables follow the variational distribution:

$$q(X) = \prod_{i=1}^{n} N(x_{i,:}|u_{i,:},S_i), \tag{18}$$

in which $S_i$ are diagonal matrices. Note that the variational distributions $q(X)$ will be explained in the next section. Then, a diversity prior $K(\phi_i,\phi_j;\rho)$ can be applied to the variational distributions $q(X)$ to allow for repulsion. For simplicity, let $\rho = 1$:

$$K(q(x_{i,:}),q(x_{j,:});1) = (2\pi)^{-\frac{D}{2}}\left(\prod_{d=1}^{D}\frac{1}{S_{id}+S_{jd}}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(u_{id}-u_{jd})}{S_{id}+S_{jd}}\right), \tag{19}$$

where $D$ refers to the dimensionality of one latent variable and $S_{id}$ is the $d$th element on the diagonal of the covariance matrix $S_i$. The latent variables selected with respect to such prior are supposed to cover multiple distinct aspects of attributes instead of focusing on the most salient ones.

Therefore, the new objective function is:

$$G(\theta) = \log\left(p(Y,Z|\theta)|K_{\phi}|^{\lambda}\right), \tag{20}$$

where $\theta$ refers to the hyperparameters in our proposed model. Furthermore, $\lambda > 0$ is used to balance the weights between likelihood measurements and the DE prior.

### 4.2. Variational inference

The full training process maximizes the objective function $G(\theta)$. Since the maximization of $G(\theta)$ is intractable, a variational lower bound $F_{v}(q)$ is applied and maximized to approximate the true marginal likelihood. This can be achieved with the aid of a variational distribution that factorizes as $q(\Theta)q(X)$, where $q(X) \sim N(\mu,S)$.

For simplicity, we drop the hyperparameters $\theta$ from the expressions. By applying Jensen's inequality, the variational bound $F_{v}(q) \leqslant \log\left(p(Y,Z)|K_{\phi}|^{\lambda}\right)$ is derived:

$$F_{v}(q) = \int q(\Theta)q(X)\log\left(\frac{\prod_{\kappa=\{Y,Z\}} p(\kappa|F^{\kappa})p(F^{\kappa}|U^{\kappa},X)p(U^{\kappa})}{q(\Theta)}\right) dFdUdX + \int q(X)\log\frac{p(X)}{q(X)}dX + \lambda\log|K_{\phi}|. \tag{21}$$

Here, $q(\Theta)$ is further factorized as $q\left(\Theta^{Y}\right)q\left(\Theta^{Z}\right)$, which means we have a variational distribution $q(\Theta^{\kappa})_{\{\kappa=Y,Z\}}$ for each view $Y$ and $Z$. After inserting $q(\Theta)$ back into $F_{v}(q)$, we have:

$$F_{v}(q) = L_Y + L_Z - KL[q(X)||p(X)] + \lambda\log|K_{\phi}|. \tag{22}$$

Now, $F_{v}(q)$ contains the terms: $L_Y, L_Z$, a KL term, and the DE prior. Here, the KL term is first introduced and the variational distribution $q(X)$ in it follows:

$$q(X) = \prod_{i=1}^{n} N\left(x_{i,:}|\mu_{i,:},S_i\right), \tag{23}$$

where each covariance matrix $S_i$ is diagonal.

Since both $q(X)$ and $p(X)$ are Gaussian distributions, the KL term can be easily calculated:

$$KL[q(X)||p(X)] = \frac{1}{2}\sum_{i=1}^{n} tr\left(\mu_{i,:}\mu_{i,:}^{T} + S_i - \log S_i\right) - \frac{nq}{2}. \tag{24}$$

We then compute the terms $L_Y$ and $L_Z$, which have similar expressions, so only the exact expression of $L_Y$ is given here:

$$L_Y = \int q\left(\Theta^{Y}\right)q(X)\log\left(\frac{p\left(Y|F^{Y}\right)p\left(F^{Y}|U^{Y},X\right)p\left(U^{Y}\right)}{q\left(\Theta^{Y}\right)}\right) dFdUdX, \tag{25}$$

We next present the details to compute $L_Y$, and the approach to compute $L_Z$ is exactly the same.

Meanwhile, the variational distribution $q\left(\Theta^{Y}\right)$ is of the form:

$$q\left(\Theta^{Y}\right) = q\left(U^{Y}\right)p\left(F^{Y}|U^{Y},X\right), \tag{26}$$

where $q\left(U^Y\right)$ is a free-form distribution. After putting $q\left(\Theta^Y\right)$ back into $L_Y$, we have:

$$L_Y = \int p\left(F^Y|U^Y,X\right) q\left(U^Y\right) q(X) \log\left(\frac{p(Y|F^Y)p(U^Y)}{q(U^Y)}\right) dFdUdX. \tag{27}$$

Here, Eq. 27 can be decomposed as:

$$L_Y = \int p\left(F^Y|U^Y,X\right) q\left(U^Y\right) q(X) \log p\left(Y|F^Y\right) dFdUdX + \int q\left(U^Y\right) \log\frac{p(U^Y)}{q(U^Y)} dU. \tag{28}$$

Remember that the variables $y^Y_{:j}, f^Y_{:j}$ and $u^Y_{:j}$ are independent. Therefore, $L_Y$ can be further factorised as:

$$L_Y = \int \prod_{j=1}^{q} p\left(f^Y_{:j}|u^Y_{:j},X\right) q\left(u^Y_{:j}\right) q(X) \sum_{j=1}^{q} \log p\left(y^Y_{:j}|f^Y_{:j}\right) dFdUdX + \int \prod_{j=1}^{q} q\left(u^Y_{:j}\right) \sum_{j=1}^{q} \log\frac{p\left(u^Y_{:j}\right)}{q\left(u^Y_{:j}\right)} dU. \tag{29}$$

Now, some items can be integrated out and $L_Y$ is further simplified as:

$$L_Y = \int p\left(f^Y_{:j}|u^Y_{:j},X\right) q\left(u^Y_{:j}\right) q(X) \sum_{j=1}^{q} \log p\left(y^Y_{:j}|f^Y_{:j}\right) df^Y_{:j} du^Y_{:j} dX + \int q\left(u^Y_{:j}\right) \sum_{j=1}^{q} \log\frac{p\left(u^Y_{:j}\right)}{q\left(u^Y_{:j}\right)} du^Y_{:j}. \tag{30}$$

For further simplification and ease of reading, let $\langle\cdot\rangle_p$ be shorthand for the expectation with respect to the distribution $p$. Then, $L_Y$ is re-expressed as:

$$L_Y = \sum_{j=1}^{q}\left(\int q\left(u^Y_{:j}\right) q(X) \left\langle \log p\left(y^Y_{:j}|f^Y_{:j}\right)\right\rangle_{p\left(f^Y_{:j}|u^Y_{:j},X\right)} \mathrm{dXd}u^Y_{:j} + \left\langle \log\frac{p\left(u^Y_{:j}\right)}{q\left(u^Y_{:j}\right)}\right\rangle_{q\left(u^Y_{:j}\right)}\right). \tag{31}$$

Clearly, the essential part of $L_Y$ is behind the sum operator. Instead of the whole $L_Y$, we focus on this part only. Assume $L_Y = \sum_{j=1}^{q} L^Y_j$, then we have:

$$L^Y_j = \int q\left(u^Y_{:j}\right) q(X) \left\langle \log p\left(y^Y_{:j}|f^Y_{:j}\right)\right\rangle_{p\left(f^Y_{:j}|u^Y_{:j},X\right)} \mathrm{dXd}u^Y_{:j} + \left\langle \log\frac{p\left(u^Y_{:j}\right)}{q\left(u^Y_{:j}\right)}\right\rangle_{q\left(u^Y_{:j}\right)}. \tag{32}$$

Now, $L^Y_j$ contains an integration operation and a KL term. Evidently, $\left\langle \log p(y_{:,i}|f_{:j})\right\rangle_{p(f_{:j}|u_{:j},X)}$ is the difficult part in computing $L^Y_j$, so this essential term is computed first as:

$$\left\langle \log p(y_{:,i}|f_{:j})\right\rangle_{p(f_{:j}|u_{:j},X)} = \left\langle \log N\left(y_{:j}|a^Y_j, (\sigma^Y_\epsilon)^2 I_n\right)\right\rangle_{q(X)} - \frac{1}{2\left(\sigma^Y_\epsilon\right)^2} tr\left(K^Y + \left(K^Y_{uu}\right)^{-1} K^Y_{uf} K^Y_{fu}\right), \tag{33}$$

where $a^Y_j = K^Y_{fu}\left(K^Y_{uu}\right)^{-1} K^Y_{uf}$.

Now, inserting Eq. 33 back into Eq. 32, the expression of $L^Y_j$ (Eq. 32) becomes:

$$L^Y_j = \underbrace{\int q\left(u^Y_{:j}\right) \log\frac{e^{\left\langle \log N\left(y_{:j}|a^Y_j,(\sigma^Y_\epsilon)^2 I_n\right)\right\rangle_{q(X)}} p\left(u^Y_{:j}\right)}{q\left(u^Y_{:j}\right)} \mathrm{d}u^Y_{:j}}_{\text{Term 1[Errormml:mo](KL-like quantity[Errormml:mo])}} - \frac{1}{2\left(\sigma^Y_\epsilon\right)^2} tr\left(\left\langle K^Y_{ff}\right\rangle_{q(X)} + \left(K^Y_{uu}\right)^{-1}\left\langle K^Y_{uf} K^Y_{fu}\right\rangle_{q(X)}\right). \tag{34}$$

Here, Eq. 34 contains a Kullback–Leibler (KL) like quantity that involves the variational distribution $q\left(u^Y_{:j}\right)$. Its term 1 is a Kullback–Leibler (KL) like quantity which is constantly smaller or equal to 0. When variational distribution $q\left(u^Y_{:j}\right)$ equals to its numerator in term 1, term 1 becomes 0 thus maximizing Eq. (34). So, optimal $q\left(u^Y_{:j}\right)$ is:

$$q\left(u^Y_{:j}\right) = e^{\left\langle \log N\left(y_{:j}|a^Y_j,(\sigma^Y_\epsilon)^2 I_p\right)\right\rangle_{q(X)}} p\left(u^Y_{:j}\right). \tag{35}$$

With the optimal variational distribution $q\left(u^Y_{:j}\right)$ in hand, $L^Y_j$ can be upper bounded by $\widehat{L}^Y_j$ by applying *reversing Jensen's inequality* [24]:

$$\widehat{L}^Y_j = \log\int e^{<\log N\left(y_{:j}|a^Y_j,(\sigma^Y_\epsilon)^2 I_p\right)>_{q(X)}} p\left(u^Y_{:j}\right) du^Y_{:j} - \frac{1}{2\left(\sigma^Y_\epsilon\right)^2} tr\left(\left\langle K^Y_{ff}\right\rangle_{q(X)} + \left(K^Y_{uu}\right)^{-1}\left\langle K^Y_{uf} K^Y_{fu}\right\rangle_{q(X)}\right). \tag{36}$$

Now, $q\left(u^Y_{:j}\right)$ is optimally eliminated and the whole objective function is tractable. A brief summary of the proposed algorithm is shown in Algorithm 1.

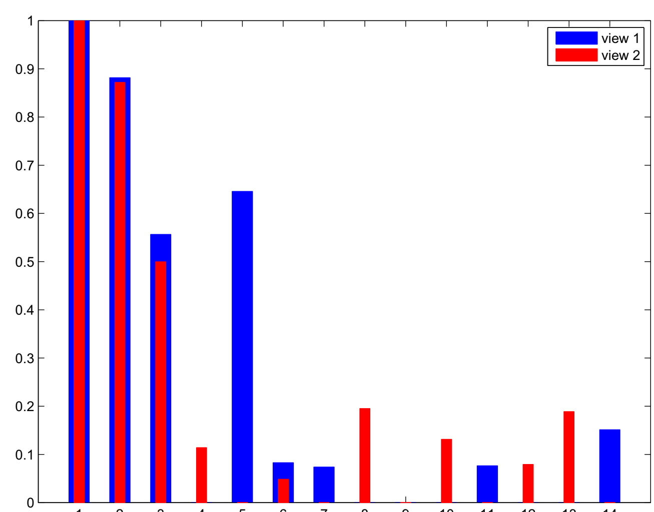



**Fig. 1.** The ARD parameter values. The dimensionality of the latent variable is 14, and two sets of ARD parameters are estimated. The blue and red bars refer to values of ARD parameters for view 1 and view 2, respectively. Here, the shared dimensions are 1, 2, 3, 6, while the private dimensions are the remaining.

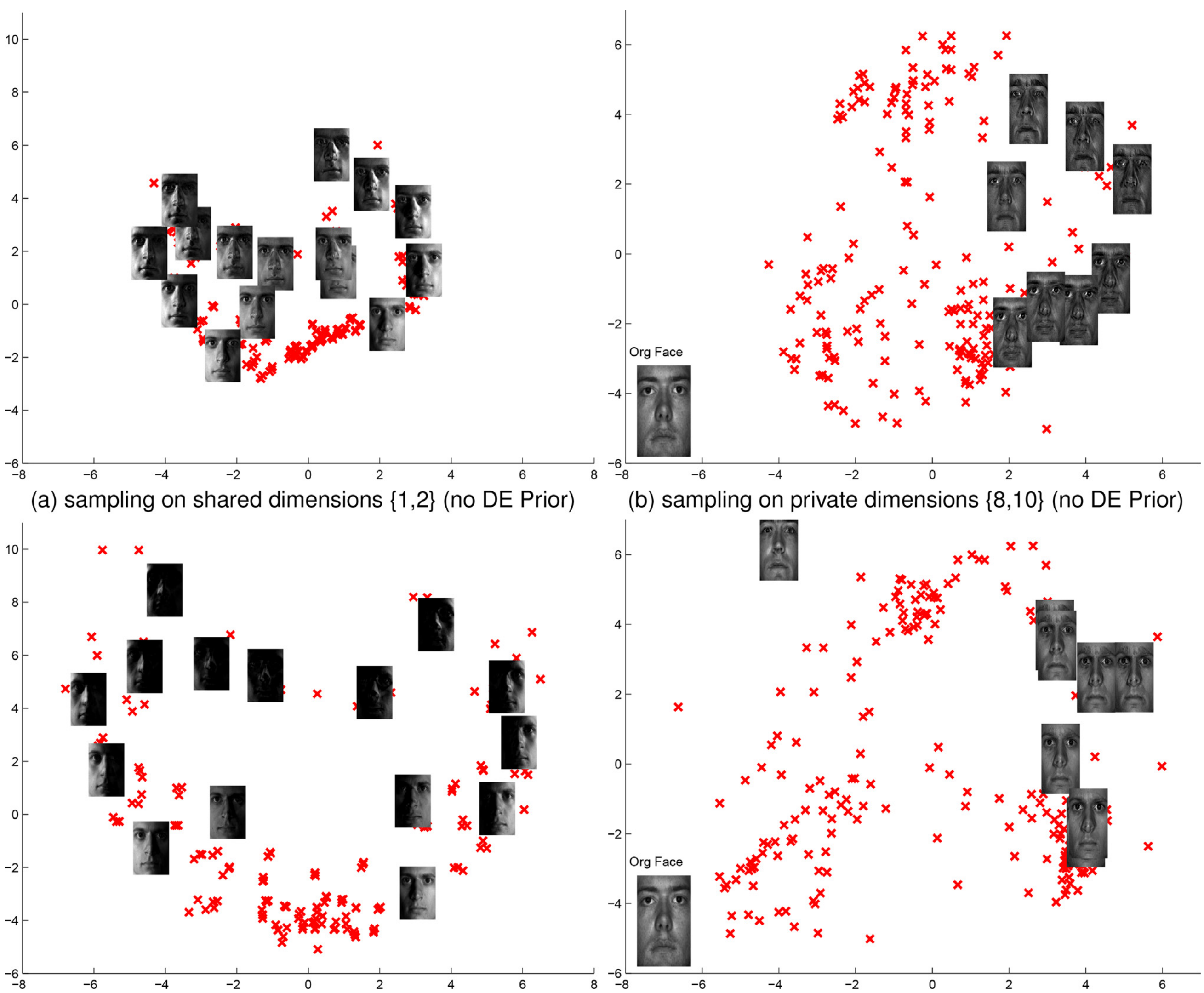
**Algorithm 1.** Diversified Shared Latent Variable Model

**Input:** observed data $Y$ (view 1) and $Z$ (view 2)
**Output:** optimised model parameters $\theta$
1: **Definition:**
define kernel-based diversity prior $K_\phi$ (Eq. 17).
2: **Define Objective Function:**
$G(\theta) = \log\left(p(Y,Z|\theta)|K_\phi|^\lambda\right)$ Eq. (20)
3: **Variational Inference:**
introduce variational distributions $q(\Theta)$ and $q(X)$
4: **Variational Bound:**
obtain variational bound $F_v(q)$ (Eq. 21) of $G(\theta)$ using variational distributions $q(\Theta)$ and $q(X)$
5: compute varitional bound $F_v(q)$ using Eq. 22–36 to optimize model parameters $\theta$ for image characteristics extraction and modelling.

### 4.3. Characteristics extraction and modelling

For ease of explanation, we use face images for illustration. Initially, we build two corresponding views with the same illumination conditions but different subject appearances. Either view has two variations, which are the illumination condition and facial appearance. Based on our model, shared latent variables with two sets of ARD weights $w^Y = \{w_d^Y\}_{d=1}^q$ and $w^Z = \{w_d^Z\}_{d=1}^q$ are learned (shown in Fig. 1). Note that the dimension of the shared latent variable is 14 (q = 14).

The ARD weights play a pivotal role in determining the responsibility of each dimension from the latent variable. We can segment the latent variable $X$ into $\left(X^Y, X^s, X^Z\right)$, in which $X^s \in R^{n\times q_s}$ is the shared subspace that models the face illuminance conditions. For $w_d^Y, w_d^Z > \delta$, the dimensions of shared subspace $X^s$ can be selected when $w_d^Y$ and $w_d^Z$ are highly similar. Here, $\delta$ is any number greater than zero. In contrast, $X^Y$ and $X^Z$ are the two private spaces




(a) sampling on shared dimensions {1,2} (no DE Prior)
(b) sampling on private dimensions {8,10} (no DE Prior)
(c) sampling on shared dimensions {1,2} (with DE Prior)
(d) sampling on private dimensions {8,10} (with DE Prior)

**Fig. 2.** The effects of diversity-encouraging priors (DE priors). (a), (c) By sampling and projecting the shared dimensions $\{1,2\}$, the shared dimensions generated with DE priors are illustrated to capture more lighting conditions including darker illuminations. (b), (d) Sampling private dimensions should alter face appearance but keep the original lighting condition unchanged. However, the illumination of sampled faces in (b) has been changed compared to those in (d). Note that the org face in the bottom left corner of either (b) or (d) is the face for which the private dimensions are altered.

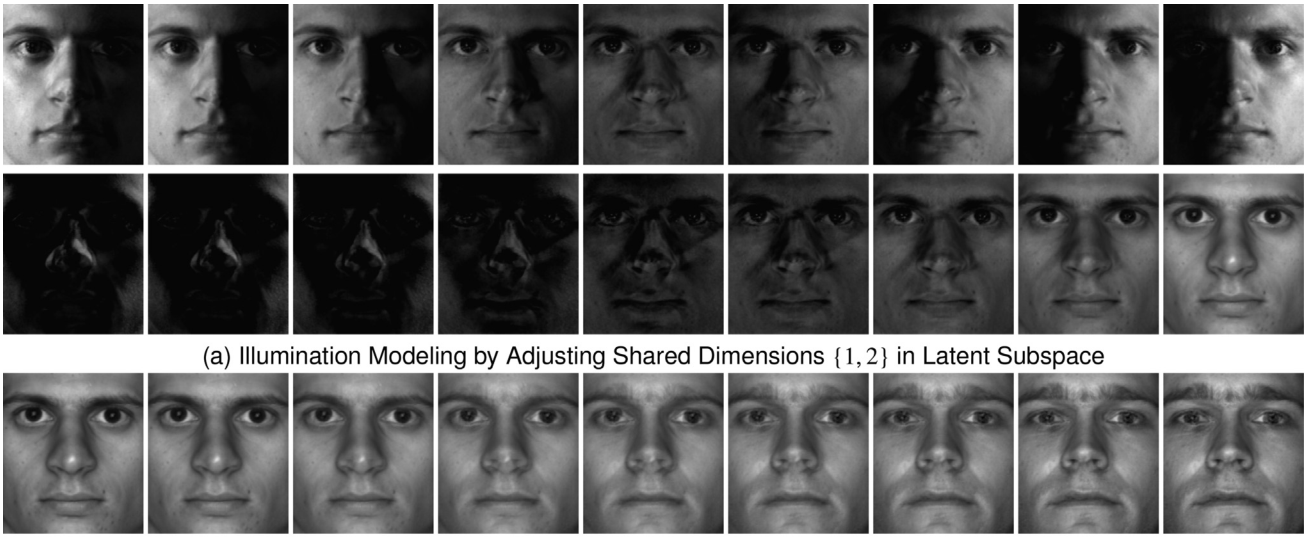



**Fig. 3.** The effects of adjusting shared and private dimensions. Note that modeling the face illumination with the shared dimensions does not alter the facial appearance and the face morphing is similar.

to model the face appearances in each view. Thus, it can be seen from Fig. 1 that the dimensions $\{1,2,3\}$ are supposed to model the face illumination variations. Likewise, the dimensions $\{8,10\}$ or $\{5,14\}$ comprise the private dimensions that control the face appearance for face morphing.

Here, the latent subspace is continuous. By sampling the shared dimensions in the latent subspace and mapping them back to the observed data space, we can create a new face with a certain illumination condition. An example of illumination variation modeling is illustrated in Fig. 3(a). Here, the mapping from latent space to observed data space is namely the inference stage described in [9]. Note that the lighting effects vary along the x-axis (top row) and y-axis (bottom row). By changing the shared dimensions, the face illumination is altered. Likewise, the private dimensions are sampled to show the face morphing effects (Fig. 3(b)).

To demonstrate the necessity for the DE prior, we compare the face illumination modeling effects before and after applying the DE prior. Instead of sampling the latent subspace arbitrarily, we only sample the trained latent variables and then map them back to the observed data space. This is because each trained latent variable corresponds to an observed face, which serves as the ground truth for ease of comparison. As shown in Fig. 2,3, the shared dimensions $\{1,2\}$ and private dimensions $\{8,10\}$ of the latent dimensions are much more diverse after applying the DE prior. Without the DE prior, the illuminance modeling in Fig. 2(a) fails to model shaded faces. The reason for is that these shaded faces do not have much saliency or many features to capture. Therefore, the introduced DE prior enables all possible illumination conditions to be covered (shown in Fig. 2(c)). Also, the facial appearance of the org face in Fig. 2(b) is modeled by altering private dimensions. However, it turns out that such modeling not only changes the face appearance but also alters its illumination. It is clear that the illumination after appearance modeling in Fig. 2(b) is darker than the org face. In contrast, the face appearance modeling with the diversity constraint shown in Fig. 2(d) maintains the same illumination conditions while changing the appearance. Further examples of face illuminations are shown in Fig. 4.

**Fig. 4.** Illustration of differences before and after exploiting the DE prior. Top row: ground truth. Middle row: our approach. Bottom row: MRD. It is clear that the face images generated by our method (middle row) are closer to the ground truth.

## 5. Experimental results

### 5.1. Yale and Multi-PIE face datasets

We performed illumination variation modeling and face morphing experiments using the Yale[1] and Multi-PIE[2] face databases. The Yale dataset includes 30 subjects, each of which is associated with 64 illumination conditions, i.e., 1920 test faces in total. The Multi-PIE face database contains a number of faces from which we randomly selected 48 different faces for testing, each of which was illuminated from 20 different viewpoints (960 test faces in total). Note that the only variations in these datasets are the lighting directions and face appearances. Since our model enables high-dimensional data processing, no feature extraction is required and the raw face data can be accepted directly. The number of pixels in each raw dataset is 32256 and 133563 for Yale and Multi-PIE, respectively.

1 http://vision.ucsd.edu/content/yale-face-database
2 http://www.multipie.org/

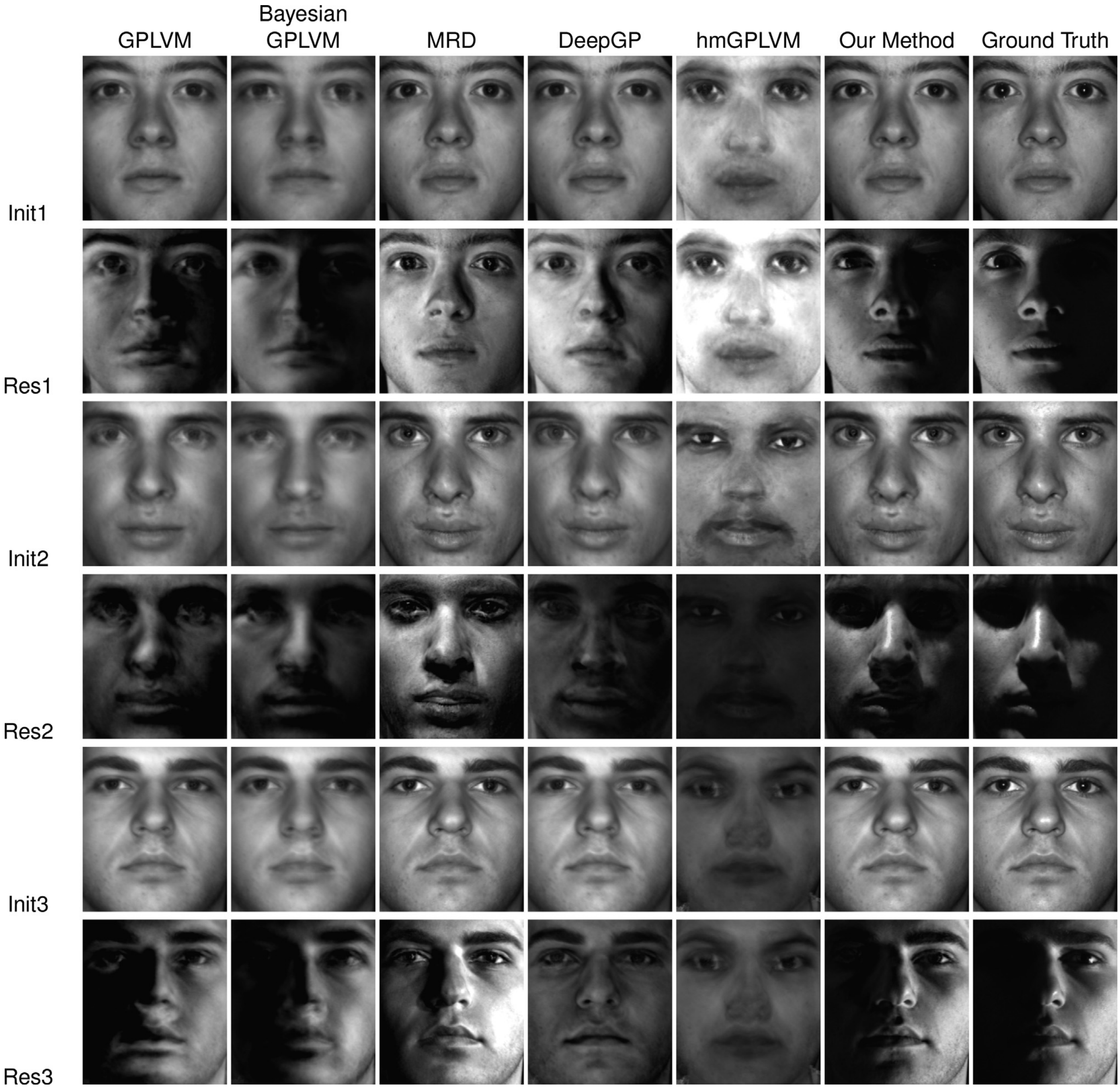



**Fig. 5.** Face illumination modeling effects. Six modeling examples from the Yale dataset are shown, and every example consists of two rows for demonstration purposes. For each example, its initial status without illumination effects is displayed in the top row (denoted as Init). The illumination modeling result is shown in the bottom row (denoted as Res). The illumination modeling effects of our method in the bottom rows are very close to the ground truth.

At the training stage, taking the Yale face database as an example, we simultaneously build two views $Y$ and $Z$, where both $Y$ and $Z$ include 3 subjects. Now, each view is composed of 192 images that correspond to 3 distinct faces under 64 various illumination conditions. Here, $Y, Z \in R^{n \times d}$, where $n = 192$ and $d = 32256$. Therefore, the 30 subjects in the Yale face dataset are divided into 5 groups. Then, the images in the $Y$ view are reordered so that every image $y_n$ from the view $Y$ randomly correspond to each $z_n$ in $Z$ under the same illumination conditions. As a result, the images are only matched between two views in terms of the illumination condition rather than subject appearance. The model is then forced to learn the correspondence between face illumination variations. Likewise, the 48 subjects from the Multi-PIE database are partitioned into 4 groups, in which every group has 12 subjects. As a consequence, the Y and Z views in each group respectively contain 120 face images composed of 6 subjects associated with 20 lighting conditions.

After optimization, the latent variables $X \in R^{192 \times 14}$ are automatically segmented into shared and private parts based on the derived ARD weights $\{w^Y, w^Z\}$. Here, the shared dimensions of the latent variable model face illumination variations, while the private part controls face appearance and, accordingly, achieves morphing. In our experiments on the Yale database, the shared part of the latent variable generally consists of the dimensions $\{1, 2, 3\}$, whilst the private part tends to be the dimensions $\{8, 10\}$ or $\{5,14\}$. However, this is not always the case, and these may vary depending on the specific datasets. As explained in Section 3, the ARD weights directly determine the shared and private

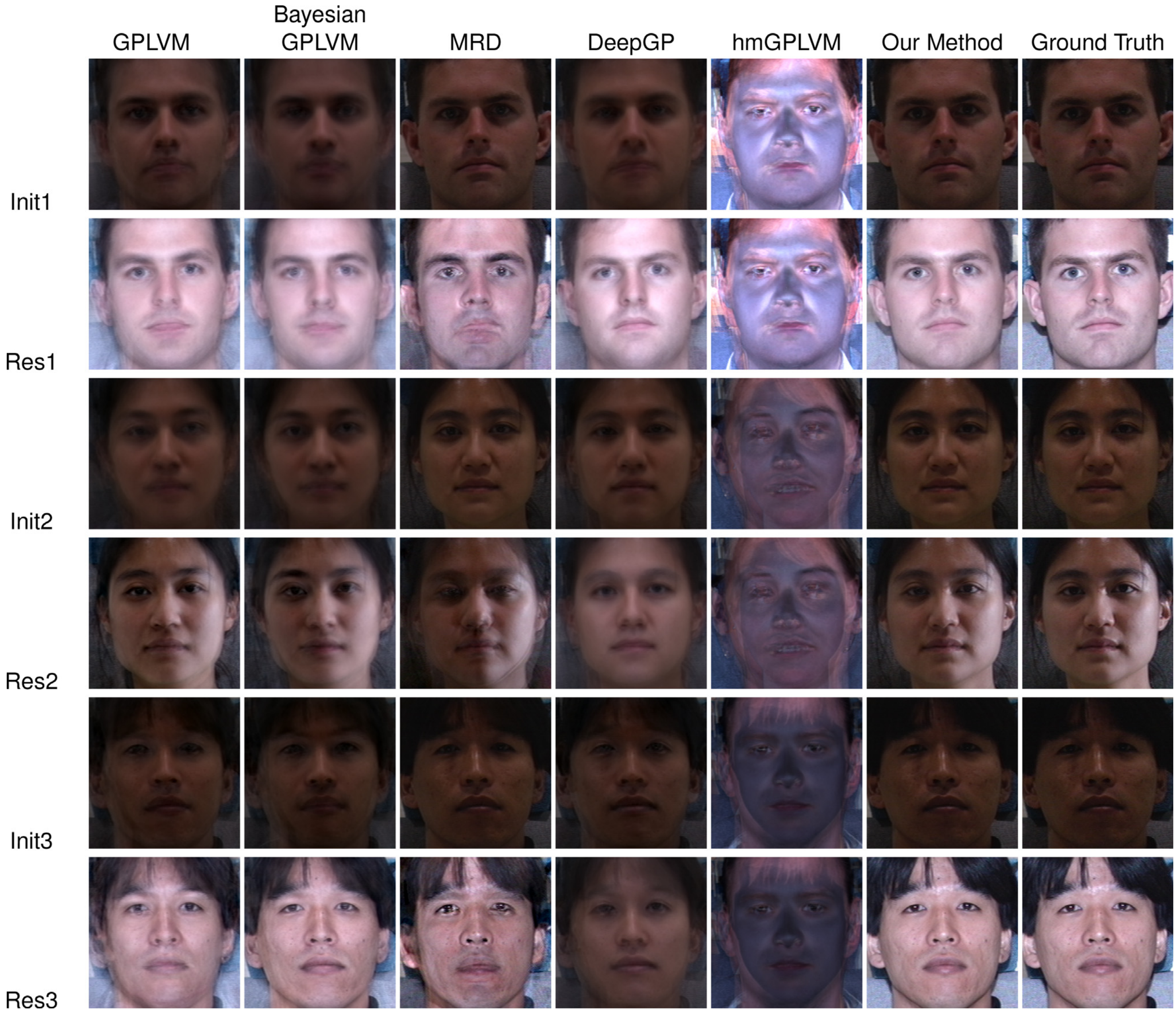



**Fig. 6.** Face illumination modeling effects. Six modeling examples from the Multi-PIE dataset are shown, and every example consists of two rows for demonstration purposes. For each example, its initial status without illumination effects is displayed in the top row (denoted as Init). The illumination modeling result is shown in the bottom row (denoted as Res). The illumination modeling effects of our method in the bottom rows are very close to the ground truth.

dimension selections. The scalar $\delta$ introduced in Section 3 is empirically set at 0.005. This ensures that only the private dimensions with higher ARD weights are selected, while the other insignificant private dimensions are neglected. Similarly, the above procedure is also applicable to the Multi-PIE face database.

First, illumination modeling of the Yale and Multi-PIE databases was performed using five different methods: Manifold Relevance Determination (MRD) [9], Gaussian Process Latent Variable Model (GPLVM) [27], Bayesian Gaussian Process Latent Variable Model (BGPLVM) [45], Deep Gaussian Process (DeepGP) [10] and Harmonized Multimodal Learning With Gaussian Process Latent Variable Models (hmGPLVM) [40]. MRD is mentioned above, and GPLVM aims to project the high dimensional data into low-dimensional latent variables. After obtaining the latent variables, the face image illumination can be modeled by sampling certain dimensions in the latent subspace and mapping them back to the observed face data space. A fully Bayesian model was then proposed based on GPLVM, namely BGPLVM. Since BGPLVM contains non linear terms that render the integration over latent variables intractable, the variational inference was adopted here to compute the objective function. DeepGP built up multiple latent spaces hierarchically. hmGPLVM developed a novel learning scheme called Harmonization to learn the latent variables shared by two views. Note that GPLVM, BGPLVM and DeepGP are single view-based modeling methods and are unable to obtain the dimensions that depict the illumination alone. For GPLVM, BGPLVM and DeepGP, we alter the same dimensions on latent variables as the shared subspace obtained by our proposed model. Meanwhile, hmGPLVM was exploited for cross-modal retrieval task but we used it to synthesize images by projecting the latent variables back to observed data space.

The effects of illumination variation modeling by altering shared dimensions are illustrated in Fig. 5 and Fig. 6. hmGPLVM clearly perform worst because it was originally developed for cross-modal retrieval task. Meanwhile, GPLVM, BGPLVM and DeepGP exploit single-view training. Hence, the information about the latent variable dimensions is unlikely to be automatically segmented and will be blended. When the face image variations


**Fig. 7.** Face morphing effects. Examples of face morphing from Yale dataset are illustrated. To achieve the morphing effect, only the private dimensions of shared latent variables are changed from target to reference faces. Our method is clearly capable of performing face morphing.

include subject appearance and illumination conditions, these methods fail to capture the illumination characteristic alone. Furthermore, it can be seen that MRD cannot model the faces under dark illumination conditions. In comparison with lit faces, the shaded faces do not have many features or saliency. Accordingly, the latent subspace generated by MRD cannot cover the information of these shaded faces. Instead, by employing the DE prior and performing training from multiple views, our method robustly models more face illumination conditions.

We next demonstrate face morphing using our proposed model (Fig. 7 and Fig. 8). It can be seen that changing the values of the selected private latent subspace only alters the face appearance. Given two faces, the indices of dimensions from their private space are supposed to be the same. Therefore, by gradually altering the values on private dimensions from the target face to the reference face, face morphing is achieved. Our proposed model clearly performs well on both color and black-and-white face images. Furthermore, our method is capable of performing face morphing between distinct ethnic backgrounds. In contrast, the other methods have the following limitations: 1) the face appearance is blurred, and 2) the final face appearance created by face morphing does not look like the ground truth. The root cause of these outcomes is that the existing methods are unable to thoroughly and completely model the face appearance, especially for the dataset containing a large variety of ethnicities. Our approach overcomes these limitations by exploiting the DE prior. see Table 1.

We exploit structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) for quantitative evaluations with the other three methods. Here, PSNR and SSIM respectively measure the quality of modeled faces, and the similarity between modeled faces and their ground truth. In the Yale database, there are 30 subjects, and each $Y$ and $Z$ view contains 15 subjects. Since the view $Y$ is intended to help view $Z$ segment the latent space, we only perform evaluations on the 15 subjects in the view $Z$. In illumination modeling, each subject in the view $Z$ has 64 illumination conditions, from which we select the fully lit face as the initial starting face for modeling. Then, there should be 63 SSIM and PSNR values for a single subject. We average all 63 SSIM and PSNR values to use a single average SSIM and PSNR value for a subject, giving rise to 15 SSIM and PSNR values for the illumination modeling results for the Yale database. Likewise, for the face morphing evaluation, we take two consecutive subjects in the Yale face database each time from view $Z$. For two consecutive subjects, we perform face morphing on every two faces with the same illumination conditions. Hence, we have 64 face morphing results for a pair of subjects. Similarly, the SSIM and PSNR evaluations are performed as for the illumination modeling evaluation. The same process is then repeated for the evaluation of the Multi-PIE database. As shown in Fig. 9, Fig. 10 and Table 2, our proposed method significantly outperforms the others on both the Yale and Multi-PIE databases.

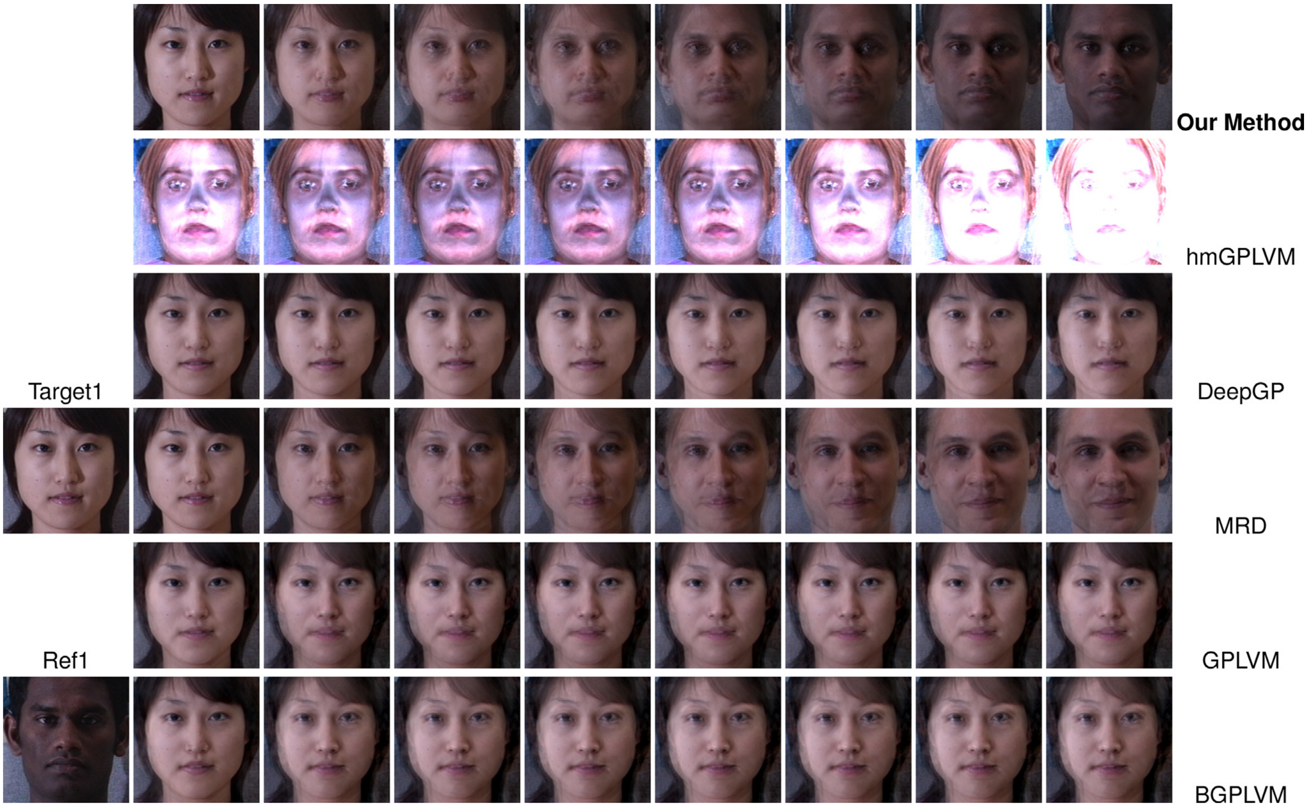



**Fig. 8.** Face morphing effects. Examples of face morphing from Multi-PIE database are illustrated. To achieve the morphing effect, only the private dimensions of shared latent variables are changed from target to reference faces. Our method is clearly capable of performing face morphing.

**Table 1**
A summary of symbols in Section 4.

| | | |
|---|---|---|
| $Y \in R^{n\times p}$ | Observed data of first view | $y_{:,j}$ is the jth column of Y (n rows and p columns) |
| $Z \in R^{n\times p}$ | Observed data of an analogous view | $z_{:,j}$ is the jth column of Z |
| $\kappa = \{Y, Z\}$ | Two views | Denote observed data $Y, Z$ as a whole |
| $X \in R^{n\times q}$ | Shared latent variables of observed data $Y, Z$ | $x_{i,:}$ is the ith row of X (n rows and q columns) |
| $F^{\{Y,Z\}} \in R^{n\times p}$ | Unnoisy version of $Y, Z$ | $f_{:,j}$ is the jth column of F |
| $U^{\{Y,Z\}} \in R^{m\times p}$ | Inducing points | Reduce model training time |
| | A small representative subset of $Y, Z$ | $u_{:,j}$ is the jth column of U |
| $\overline{X}^{\{Y,Z\}} \in R^{m\times q}$ | Pseudo-inputs | Latent variables of inducing points $U$ |
| | | $\overline{X}^{\{Y,Z\}}_{i,:}$ is the ith row of pseudo-inputs $\overline{X}$ |
| $K^{\{Y,Z\}}_{ff}$ | Covariance matrices | its element is computed by covariance function (Eq. 9) |
| $K^{\{Y,Z\}}_{uu}$ | Covariance matrix | Computed using pseudo-inputs $\overline{X}^{\{Y,Z\}}$ |

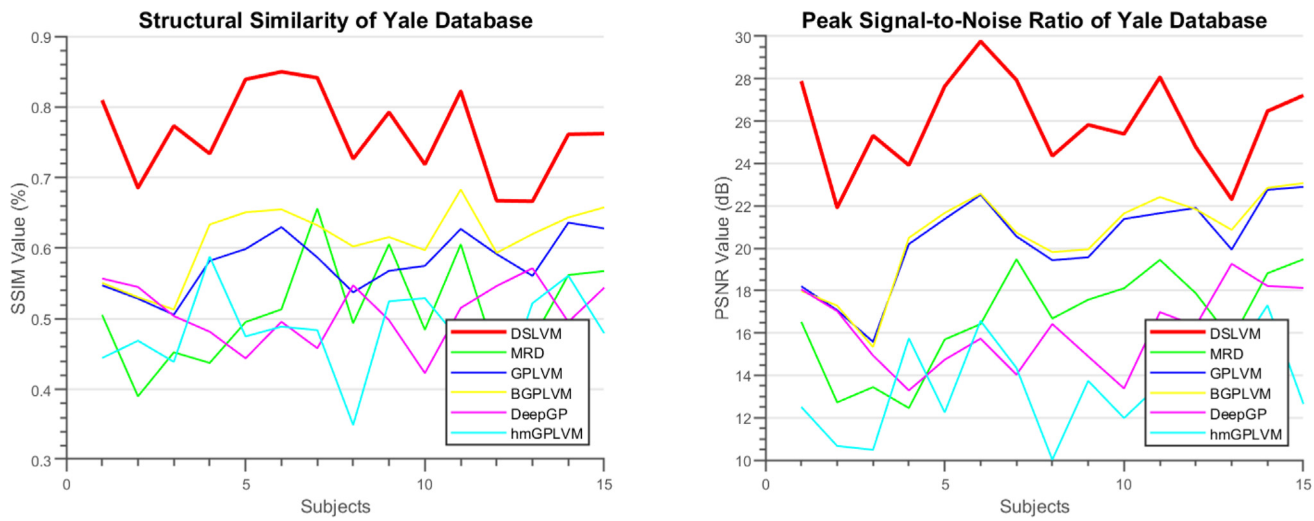
### 5.2. 3D chairs dataset

The 3D chair dataset [3] contains 1393 3D CAD models of different chairs. Meanwhile, each chair has 62 different viewpoints. For experimental validation, we randomly select 20 different chairs from the dataset. Since each chair has 62 viewpoints, we have 1240 images in total for evaluation. Moreover, the 20 sets of chairs were divided into 5 groups each of which contains 4 chairs. Then, each view of our method DSLVM has 2 chairs. By doing this, our method is able to separately extract chair viewpoint and appearance in an unsupervised manner. Here, the dimension of latent variables is set as 7.

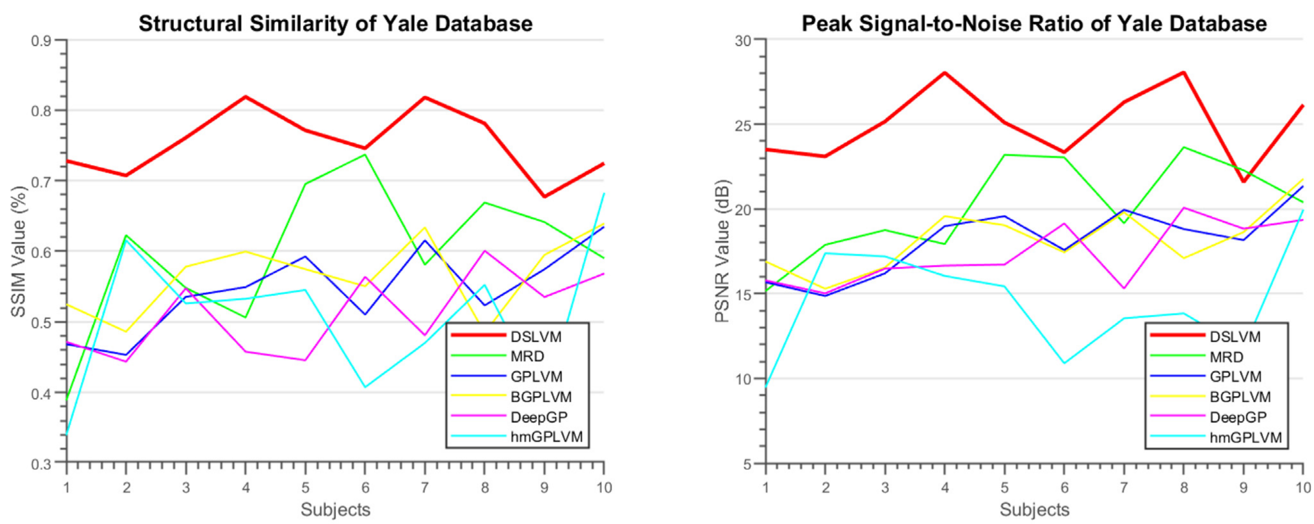
In Fig. 11, an initial chair image is given. Then, a number of chair images with different viewpoints are used to provide viewpoint information for new 3D chair synthesis. More specifically, given a query chair with a certain viewpoint, the proposed model is supposed to extract its viewpoint information and thus generate a new chair of initial chair image with the same viewpoint as query chair, but does not change its appearance. It is clear that our proposed method DSLVM outperformed the others due to the exploitation of diversity encouraging prior.

In addition to qualitative comparison, we also performed a quantitative evaluation with SSIM and PSNR. Specifically, given a chair image, we only altered its shared part of latent variables to change its viewpoint. Since we have 62 viewpoints for each chair image, there were 61 images (excluding initial one) synthesized

[3] https://www.di.ens.fr/willow/research/seeing3Dchairs/




(a) SSIM and PSNR of Illumination Modeling on Yale Database




(b) SSIM and PSNR of Face Morphing on Yale Database

**Fig. 9.** Illustration of the quantitative evaluations on the Yale database. Top row: evaluations of face illumination modeling. Bottom row: evaluations of face morphing. The evaluated results are displayed with PSNR and SSIM values. The red lines indicate the results of our proposed method, which indicate that our approach substantially enhances the accuracy of illumination modeling and face morphing.

after changing the shared parts. Then, we used these synthesized images and their ground truth for SSIM and PSNR evaluations. Evidently, our method outperformed others in terms of SSIM and PSNR (as shown in Table 3).

## 6. Discussion

In the experiments, we test our method and baselines on three datasets for face illumination modelling, face morphing and 3D chair viewpoint modelling. Overall, our method performs best qualitatively and quantitatively. Unsurprisingly, the single latent space based models GPLVM, BGPLVM and DeepGP cannot segment the latent variables at all so they are unable to extract the attributes from image appropriately. Besides, the shared latent variable models are usually exploited for image classification, audio-visual speech recognition, image captioning and cross modal retrieval. Here, the novel shared latent variable model hmGPLVM we used is for cross model retrieval. Therefore, it can be seen that hmGPLVM cannot synthesize images properly and accordingly fail to extract attributes from images.

Our method, considering the diversity prior, can cover more diversified attribute variations and discriminatively separate the attributes. In contrast, MRD without the diversity prior generally focuses on certain aspects of the attributes and thus ignores other aspects. For example, the latent variables generated by MRD tend to pay more attention to the facial appearance attribute rather than to the facial illumination. As a result, some face images synthesized by MRD in Figs. 4–6 have high contrast and look clearer (these are more relevant to face appearance). However, compared with our method, MRD falls short in separating the attributes and thus the images synthesized by it are generally distinct from the ground truth (as shown in Figs. 4, 5, 6, 7, 8 and 11).

Here, GPLVM, BGPLVM, MRD and our method primarily optimized the latent variables, parameters of the kernel function and variational distributions during the computation. Consequently, the computational complexities of these four methods were similar. Although our method defined a new diversity prior, the parameters of this prior for optimization are limited. As a result, the inclusion of the diversity prior did not significantly increase the computation time. During the optimization, DeepGP and

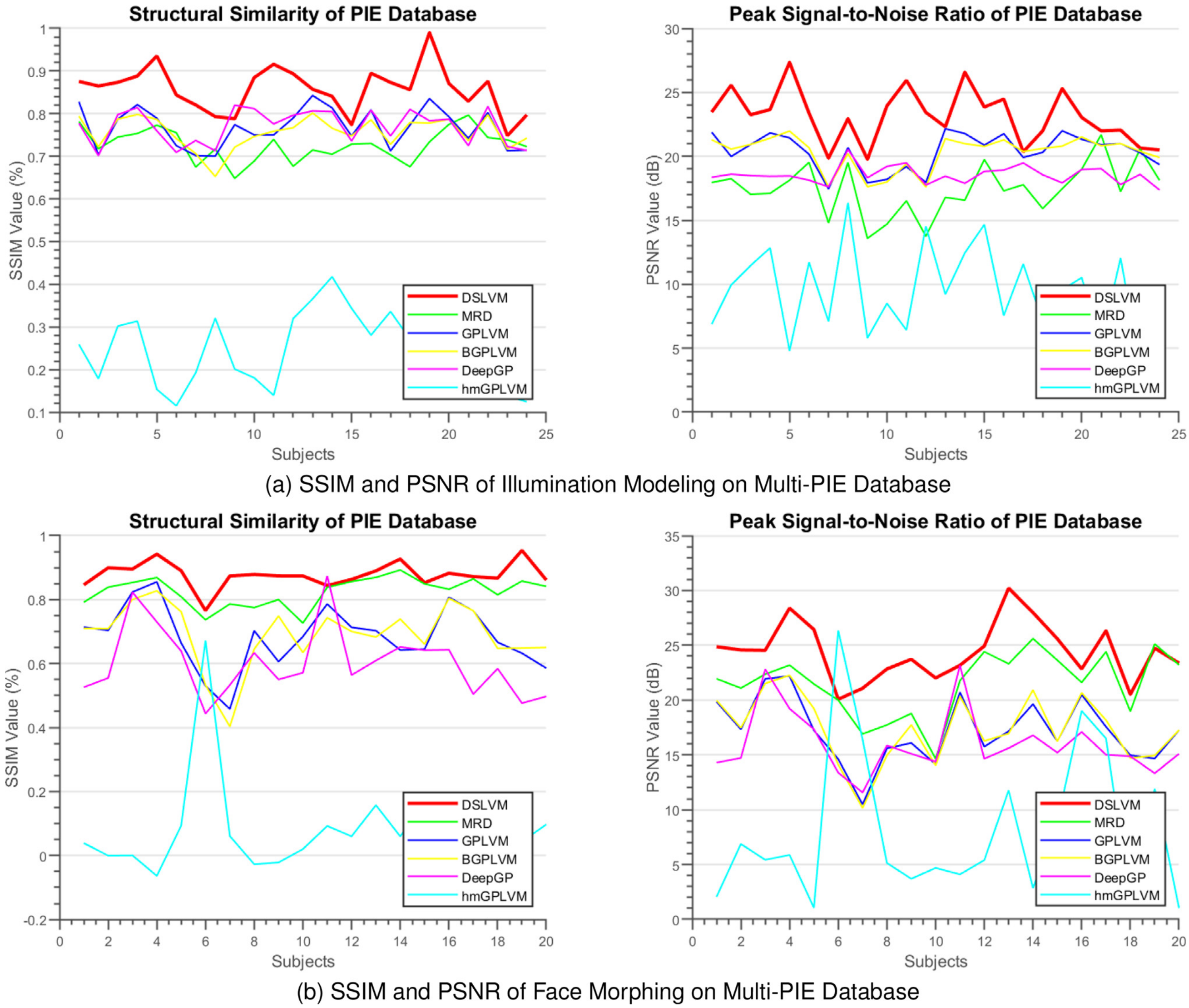



**Fig. 10.** Illustration of the quantitative evaluations on the Multi-PIE database. Top row: evaluations of face illumination modeling. Bottom row: evaluations of face morphing. The evaluated results are displayed with PSNR and SSIM values. The red lines indicate the results of our proposed method, which indicate that our approach substantially enhances the accuracy of illumination modeling and face morphing.

**Table 2**
Comparison of face illumination modeling and face morphing performances on the Yale and Multi-PIE databases.

| | Face Illumination Modeling | | | | Face Morphing | | | |
|---|---|---|---|---|---|---|---|---|
| | Yale Face | | Multi-PIE | | Yale Face | | Multi-PIE | |
| | Avg SSIM | Avg PSNR (higher is better) | Avg SSIM | Avg PSNR | Avg SSIM | Avg PSNR (higher is better) | Avg SSIM | Avg PSNR |
| DSLVM | **0.76** | **25.92** | **0.86** | **23.15** | **0.75** | **25.03** | **0.88** | **24.40** |
| hmGPLVM | 0.48 | 13.16 | 0.25 | 9.29 | 0.51 | 14.55 | 0.08 | 8.02 |
| DeepGP | 0.51 | 16.10 | 0.77 | 18.55 | 0.51 | 17.34 | 0.60 | 15.97 |
| MRD | 0.51 | 16.69 | 0.73 | 17.46 | 0.59 | 20.14 | 0.83 | 21.50 |
| GPLVM | 0.58 | 20.33 | 0.77 | 20.38 | 0.55 | 18.11 | 0.68 | 17.19 |
| BGPLVM | 0.61 | 20.57 | 0.76 | 20.29 | 0.56 | 18.21 | 0.69 | 17.39 |

hmGPLVM have similar types of parameters as GPLVM, BGPLVM and MRD. However, DeepGP exploited multi-layered latent spaces to build up the model and hmGPLVM defined a new Harmonization learning scheme, which substantially increased the number of parameters for computation. As a consequence, DeepGP and hmGPLVM required a longer computation time.

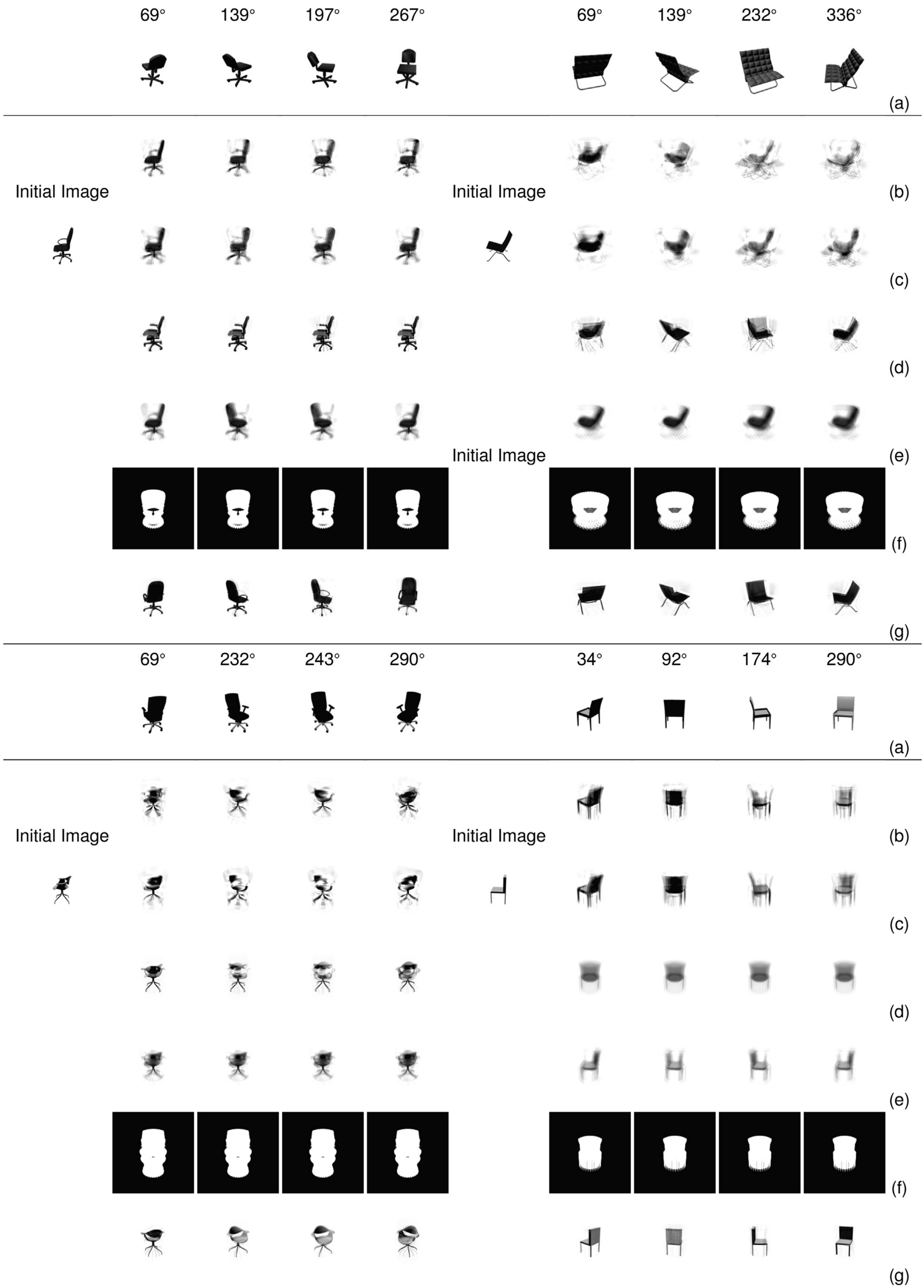



**Fig. 11.** Synthesis of 3D chairs with reference to the query images of different viewpoints (a). The **Initial Images** were re-rendered respectively by methods (b) GPLVM, (c) BGPLVM, (d) MRD and (e) DeepGP and (f) hmGPLVM and (g) DSLVM with reference to the viewpoints of the query images.

**Table 3**
Evaluation on the chair images from 3D Chair dataset.

| 3D Chair Dataset | | |
|---|---|---|
| | Avg SSIM (higher is better) | Avg PSNR |
| GPLVM | 0.862 | 17.131 |
| BGPLVM | 0.866 | 17.379 |
| MRD | 0.893 | 20.161 |
| DeepGP | 0.894 | 18.518 |
| hmGPLVM | 0.014 | 0.437 |
| DSLVM | **0.947** | **28.634** |

## 7. Conclusion

This paper investigates unsupervised characteristics extraction and modelling by jointly exploiting multi-view learning and the diversity property under the shared latent variable model framework. Unlike traditional single view-based modeling, the introduced multiple views can handle the variabilities that exist in image attributes. Specifically, multi-view learning helps automatically segment the latent space into the shared and private latent subspaces, where the shared subspace refers to the dimensions of the latent variable and aims to model variations of common attributes. The private subspace is the remaining dimensions of the latent variable that control the other attributes. Additionally, a diversity encouraging prior is introduced to capture distinguishing characteristics from the observed images. This renders our approach superior for accurately modeling variations of object attributes under more complex and varied circumstances.

Our model define a new objective function with diversity encouraging priors. However, since the objective function is initially not tractable, the variational inference is employed to approximate it by deriving a lower bound. Then, the solution is obtained by maximizing the derived lower bound. To test the effectiveness and robustness of our proposed model, we perform experiments on three different datasets: 3D Chair dataset with various chairs from different viewpoints, Multi-PIE (color) and Yale (black-and-white) face databases that capture individuals from different countries under a large range of lighting positions. The experiments show that our model is applicable to extract characteristics of a wide range of objects including chairs, faces, etc. Despite the effectiveness and robustness of our approach for image characteristics extraction, it is limited to separate and model two attributes (such as face appearance, illumination) out of the image. In the future, we aim to upgrade our model so that it can extract multiple characteristics out of an image at the meantime.

## CRediT authorship contribution statement

**Hao Xiong:** Methodology, Formal analysis, Writing - original draft. **Yuan Yan Tang:** Supervision. **Fionn Murtagh:** Supervision. **Leszek Rutkowski:** Validation. **Shlomo Berkovsky:** Writing - review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] S. AliKhan, M. Ishtiaq, M. Nazir, M. Shaheen, Face recognition under varying expressions and illumination using particle swarm optimization, J. Comput. Sci. 28 (2018) 94–100.
[2] C. Archambeau, F.R. Bach, Sparse probabilistic projections, Proc. Adv. Neural Inf. Process. Syst (2009) 73–80.
[3] C.M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag 4, USA, 2006.
[4] J.Q. Candela, C.E. Rasmussen, A unifying view of sparse approximate Gaussian process regression, J. Mach. Learn. Res. 6 (2005) 1939–1959.
[5] K. Chaudhuri, S.M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proc. 26th Annu. Int. Conf. Mach. Learn., 2009, pp. 129–136.
[6] T.B. Chen, V.W. Soo, A comparative study of recurrent neural network architectures on learning temporal sequences, in: IEEE International Conference on Neural Networks, 1996.
[7] L. Csato, Gaussian processes – iterative sparse approximations. PhD thesis, Aston University, 2002.
[8] L. Csato, M. Opper, Sparse on-line Gaussian processes, Neural Comput. 3 (2002) 641–668.
[9] A. Damianou, C.H. Ek, M. Titsias, N. Lawrence, Manifold relevance determination, International Conference on Machine Learning (2012) 145–152.
[10] A. Damianou, N. Lawrence, Deep gaussian process, International Conference on Artificial Intelligence and Statistics (2013) 207–215.
[11] A.C. Damianou, M.K. Titsias, N.D. Lawrence, Variational inference for uncertainty on the inputs of Gaussian process models, J. Mach. Learn. Res. 17 (2016) 1–62.
[12] S. Eleftheriadis, O. Rudovic, M. Pantic, Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition, IEEE Trans. Image Process. 24 (2015) 189–204.
[13] S. Eleftheriadis, O. Rudovic, M. Pantic, Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition, IEEE Tran. Image Process. 24 (2015) 189–204.
[14] J.L. Elman, Finding structure in time, Cogn. Sci. 14 (1990) 179–211.
[15] Z. Fan, Y. Xu, W. Zuo, J. Yang, J. Tang, Z. Lai, D. Zhang, Modified principal component analysis: an integration of multiple similarity subspace models, IEEE Trans. Neural Netw. Learn. Syst. 25 (2014) 1538–1552.
[16] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.
[17] N. Hussain, M.A. Khan, M. Sharif, S.A. Khan, A.A. Albesher, T. Saba, A. Armaghan, A deep neural network and classical features based scheme for objects recognition: an application for machine inspection, Multimedia Tools Appl. (2020), https://doi.org/10.1007/s11042-020-08852-3.
[18] I.T. Jolliffe, Principal Component Analysis, Springer, Berlin, 2002.
[19] M. Jordan, Attractor dynamics and parallelism in a connectionist sequential machine, Annual Conference of the Cognitive Science Society (1986) 531–546.
[20] F. Ju, Y. Sun, J. Gao, Y. Hu, B. Yin, Image outlier detection and feature extraction via l1-norm-based 2d probabilistic pca, IEEE Trans. Image Process. 24 (2015) 4834–4846.
[21] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, Proc. Eur. Conf. Comput. Vis. (2012) 808–821.
[22] M.A. Khan, K. Javed, S.A. Khan, T. Saba, U. Habib, J.A. Khan, A.A. Abbasi, Human action recognition using fusion of multiview and deep features: an application to video surveillance, Multimedia Tools Appl. (2020), https://doi.org/10.1007/s11042-020-08806-9.
[23] S.A. Khan, A. Hussain, M. Usman, Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features, Multimedia Tools Appl. 77 (2018) 1133–1165.
[24] N.J. King, N.D. Lawrence, Fast variational inference for Gaussian process models through kl-correction, in: European Conference on Machine Learning, 2006.
[25] Z. Lai, W. Wong, Y. Xu, J. Yang, D. Zhang, Approximate orthogonal sparse embedding for dimensionality reduction, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 723–735.
[26] Z. Lai, Y. Xu, Q. Chen, J. Yang, D. Zhang, Multilinear sparse principal component analysis, IEEE Trans. Neural Netw. Learn. Syst. 25 (2014) 1942–1950.
[27] N.D. Lawrence, Gaussian process latent variable models for visualisation of high dimensional data, Neural Inf. Process. Syst. (2004) 329–336.
[28] N.D. Lawrence, A.J. Moore, Hierarchical gaussian process latent variable models, in: International Conference on Machine Learning, 2007, pp. 481–488.
[29] J. Li, B. Zhang, D. Zhang, Shared autoencoder Gaussian process latent variable model for visual classification, IEEE Trans. Neural Networks Learn. Syst. 29 (2018) 4272–4286.
[30] A. Munir, A. Hussain, S.A. Khan, M. Nadeem, S. Arshid, Illumination invariant facial expression recognition using selected merged binary patterns for real world images, Optik 158 (2018) 1016–1025.
[31] M.A. Nicolaou, Y. Panagakis, S. Zafeiriou, M. Pantic, Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2014, pp. 1522–1526.
[32] X. Niyogi, Locality preserving projections, in: Proc. Neural Inf. Process. Syst, 2004, p. 153.
[33] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, D. Eck, A hierarchical latent vector model for learning long-term structure in music, in: International Conference on Artificial Intelligence and Statistics, 2018.
[34] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Proc. SIKDD Conf. Data Mining Data Warehouse, 2010, pp. 1–4.
[35] M. Seeger, C.K.I. Williams, N.D. Lawrence, Fast forward selection to speed up sparse Gaussian process regression, in, in: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003.

[36] I.V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, Y. Bengio, A hierarchical latent variable encoder-decoder model for generating dialogues, Thirty-First AAAI Conference on Artificial Intelligence (2017) 3295–3301.
[37] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2012, pp. 2160–2167.
[38] G. Song, S. Wang, Q. Huang, Q. Tian, Similarity Gaussian process latent variable model for multi-model data analysis, International Conference on Computer Vision (2015) 4050–4058.
[39] G. Song, S. Wang, Q. Huang, Q. Tian, Multimodal similarity Gaussian process latent variable model, IEEE Trans. Image Process. 26 (2017) 4168–4181.
[40] G. Song, S. Wang, Q. Huang, Q. Tian, Harmonized multimodal learning with gaussian process latent variable models, IEEE Trans. Pattern Anal. Mach. Intell. (2019), https://doi.org/10.1109/TPAMI.2019.2942028.
[41] B. Ster, Selective recurrent neural network. neural processing letters, IEEE Trans. Image Process. 38 (2013) 1–15.
[42] J. Sun, S. Keates, Canonical correlation analysis on data with censoring and error information, IEEE Trans. Neural Netw. Learn. Syst. 24 (2013) 1909–1919.
[43] Y. Sun, X. Tao, Y. Li, J. Lu, Robust 2d principal component analysis: a structured sparsity regularized approach, IEEE Trans. Image Process. 24 (2015) 2515–2526.
[44] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. Roy. Stat. Soc. 61 (1999) 611–622.
[45] M.K. Titsias, N.D. Lawrence, Bayesian Gaussian process latent variable model, International Conference on Artificial Intelligence and Statistics (2010) 844–851.
[46] Y.H. Yuan, Q.S. Sun, Multiset canonical correlations using globality preserving projections with applications to feature extraction and recognition, IEEE Trans. Neural Netw. Learn. Syst. 25 (2014) 1131–1146.

**Hao Xiong** received the PhD degree in computer science from the University of Sydney in 2018. He is currently a research fellow in machine learning within the Australian Institute of Health Innovations within the Faculty of Medicine and Health Sciences, Macquarie University, NSW, Australia. His research interests include bioinformatics, medical image analysis, latent variable models, image processing, and computer vision.

**Yuan Yan Tang** (Life Fellow, IEEE) is currently a Chair Professor with the Faculty of Science and Technology, University of Macau, and a Professor/Adjunct Professor/Honorary Professor with several institutes, including Chongqing University, Concordia University, and Hong Kong Baptist University. He has published over 400 academic articles. He has authored/coauthored over 25 monographs/books/book chapters. His current research interests include wavelets, pattern recognition, image processing, and cybersecurity. He is a Fellow of IAPR. He is the Founder and the Chair of Pattern Recognition Committee in the IEEE SMC. He is the Founder and the Editor-in-Chief of the International Journal on Wavelets, Multiresolution, and Information Processing and an Associate Editor of several international journals.

**Fionn Murtagh** received the B.A. and B.A.I. degrees in mathematics and engineering science and the M.Sc. degree in computer science from Trinity College Dublin, Dublin, Ireland, the Ph.D. degree in mathematical statistics from the Université Pierre and Marie Curie, Paris VI University, Paris, France, and the Habilitation from the University of Strasbourg, Strasbourg, France. He is currently a Professor of computer science with the Royal Holloway, University of London, Egham, U.K. He is also currently the Director of the Science Foundation Ireland's research funding programs in information and communications technologies, renewable energies, materials science, and other areas. Dr. Murtagh is a Member of the Royal Irish Academy, a Fellow of the International Association for Pattern Recognition, and a Fellow of the British Computer Society.

**Leszek Rutkowski** (F'05) received the M.Sc. and Ph.D. degrees from the Wroclaw University of Technology, Wroclaw, Poland, in 1977 and 1980, respectively, and the Honoris Causa degree from the AGH University of Science and Technology, Kraków, Poland, in 2014. He has been with the Czestochowa University of Technology, Czestochowa, Poland, since 1980, where he is currently a Professor and the Director of the Institute of Computational Intelligence. From 1987 to 1990, he held a visiting position with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK, USA. He has authored over 200 publications and 6 books. His current research interests include stream data mining, computational intelligence, and pattern classification. Dr. Rutkowski was elected as a member of the Polish Academy of Sciences, in 2004. He has received the IEEE Fellow Membership Grade for contributions to neurocomputing and flexible fuzzy systems in 2004. He was a recipient of the IEEE Transactions on Neural Networks Outstanding Paper Award in 2005. He is the Founding Chair of the Polish Chapter, IEEE Computational Intelligence Society, which received the 2008 Outstanding Chapter Award.

**Shlomo Berkovsky** is the leader of the Precision Health research stream at Macquarie University. The stream focusses on the use of machine learning methods to develop patient models and personalised predictions of diagnosis and care. Shlomo also studies how sensors and physiological responses can predict medical conditions, and how clinicians and patients interact with health technologies. His areas of expertise include user modelling, online personalisation, and persuasive technologies.