# Coupled Marginalized Auto-Encoders for Cross-Domain Multi-View Learning

**Shuyang Wang**[1], **Zhengming Ding**[1], **and Yun Fu**[1,2]
[1]Department of Electrical & Computer Engineering,
[2]College of Computer & Information Science,
Northeastern University, Boston, MA, USA
{shuyangwang, allanding, yunfu}@ece.neu.edu

## Abstract

In cross-domain learning, there is a more challenging problem that the domain divergence involves more than one dominant factors, e.g., different view-points, various resolutions and changing illuminations. Fortunately, an intermediate domain could often be found to build a bridge across them to facilitate the learning problem. In this paper, we propose a Coupled Marginalized Denoising Auto-encoders framework to address the cross-domain problem. Specifically, we design two marginalized denoising auto-encoders, one for the target and the other for source as well as the intermediate one. To better couple the two denoising auto-encoders learning, we incorporate a feature mapping, which tends to transfer knowledge between the intermediate domain and the target one. Furthermore, the maximum margin criterion, e.g., intra-class compactness and inter-class penalty, on the output layer is imposed to seek more discriminative features across different domains. Extensive experiments on two tasks have demonstrated the superiority of our method over the state-of-the-art methods.

## 1 Introduction

Many real-world samples can be approached through different views/modalities, especially in image classification. For example, face images can be captured with different poses, lighting conditions, or even with makeup [Wang and Fu, 2016]; or face images can be obtained from different sensors which provide Visible and Near-Infrared features [Ding *et al.*, 2015]. Naturally, the comparison of different types of heterogeneous data or knowledge across domains extensively exists in many computer vision problems. For example, facial sketch based recognition [Zhang *et al.*, 2011] is one of the most well-studied cross-domain learning problems. Also, cross-view action recognition [Liu *et al.*, 2011] utilized training data captured by one camera and applied to recognize test data from another camera. Since the spanned feature spaces are quite different, it is very difficult to directly compare images across domains, and it becomes a major challenge to represent and relate data across different domains.

In cross-domain learning, we usually have two domains with different distributions, which are dominant with one factor, e.g., different view-points, various resolutions and large age gap. Cross-domain learning aims to seek a common latent space, where domain shift is well reduced. However, when the distribution divergences involve more than one factors, it becomes a more challenging problem to mitigate the large divergence across two domains. Fortunately, we can find an intermediate domain to bridge the gap smoothly. Take kinship verification for example. Parents and child not only have their own specific difference in appearance, but also suffer an age gap. Therefore, it is hard to handle kinship verification in such case. However, we always can find an intermediate domain, that is, the young parents, whose age is close to his/her child. We can observe that young parents (YP) and old parents (OP) would be more similar in appearance but in different ages, whilst the young child (YC) and his/her young parents (YP) share a similar age distribution. Although such intermediate domain YP builds a bridge between OP and YC, it meanwhile brings in one more domain to make the learning problem more complicated.

Recently, there are several kinds of techniques to deal with cross-domain learning problem, including feature adaptation learning, classifier adaptation learning and dictionary learning. Among them, feature adaptation learning [Ding and Fu, 2014; Zhao and Fu, 2015] intends to seek a common feature space, where the domain divergence would be mitigated. Classifier adaptation learning [Wu and Jia, 2012] aims to train a classifier on one domain then adapt to the other domain. Dictionary learning [Huang and Wang, 2013] is designed to build one dictionary or two as the bases to generate more discriminative features for two domains. Most recently, deep learning [Dong *et al.*, 2014; Schroff *et al.*, 2015] has attracted much attention in many applications, which aims to build deep structures to capture more discriminative information.

In this paper, we propose a Coupled Marginalized Denoising Auto-encoders framework, whose core idea is to build two types of marginalized denoising auto-encoders for effective feature extraction (Figure 1). Specifically, the intermediate dataset is treated as one of two views in one domain, therefore, one domain has two views while the other domain only has one view. This problem can be defined as *Cross-domain Multi-view Learning*. To sum up, the major contributions of this paper are two-fold as follows:

- Coupled marginalized denoising auto-encoders have been proposed to extract features for each domain. To better couple two auto-encoders, a feature mapping scheme is adopted to alleviate one divergence factor between the intermediate one and another domain. Specifically, a feature mapping matrix is proposed to project the hidden layers of them into a common space.
- The maximum margin criterion, i.e., intra-class compactness and inter-class penalty on the output layer, is imposed on these two auto-encoders to endow discriminative ability. With the learned mapping matrix, we can transform the hidden of two domains into one space and generate the output with the same decoding parameters.

## 2 Related Works

This section mainly discusses the related works from two perspectives: method-based one, which is auto-encoder; and application-based, which are kinship verification and person re-identification.

**Auto-encoder** (AE) [Ranzato *et al.*, 2008] is known as a basic building block with single hidden layer to constitute a deep structure. The identical input and target framework makes the neurons in the hidden layer an identity-preserved representation of input data. Furthermore, denoising auto-encoder (DAE) is trained to have denoising ability by involving the reconstruction of clean input from partially corrupted one with artificially added noise [Vincent *et al.*, 2010]. However, there is a crucial limitation of DAE, which is high computational cost due to non-linear optimization. To this end, [Chen *et al.*, 2012] proposed marginalized DAE (mDAE), which replaces the encoder and decoder with one linear transformation matrix. mDAE provides a closed-form solution for the parameters thus eliminates the use of other optimization algorithms, e.g., stochastic gradient descent, back-propagation. The DAE is sped up by two orders of magnitude subsequently. In this paper, we also adopt the idea of mDAE to fast the feature learning, however, we still preserve the encoder and decoder in the neural networks. Besides, we propose coupled marginalized DAEs to handle cross-domain problems.

**Person Re-identification** has been well-studied recently, due to its important application in video surveillance. There are mainly two groups of methods: one is metric learning, which focuses on learning effective metrics to measure the similarity between two images [Zheng *et al.*, 2013; Koestinger *et al.*, 2012]. The other research efforts focus on learning expressive advanced features, e.g., salience features [Zhao *et al.*, 2013a; 2013b] and mid-level features [Zhao *et al.*, 2014]. Most recently, [Jing *et al.*, 2015] designed coupled dictionary learning to address the challenge on matching two different views with different resolutions. Differently, our proposed coupled auto-encoders adopt a lite-version deep structure to extract more effective features across multiple domains.

**Kinship Verification** was first been tackled by [Fang *et al.*, 2010], whose goal is to determine whether there is a kin relation between a pair of given face images. It is still a challenging problem in computer vision, as kinship suffers different types of variations, e.g., large age gap between parent and children. Xia et al. introduced young parent dataset as an intermediate domain to facilitate kinship learning [Xia *et al.*, 2011]. We also evaluate our algorithm with the intermediate domain, however, it is the first attempt to address kinship verification with coupled marginalized auto-encoders.

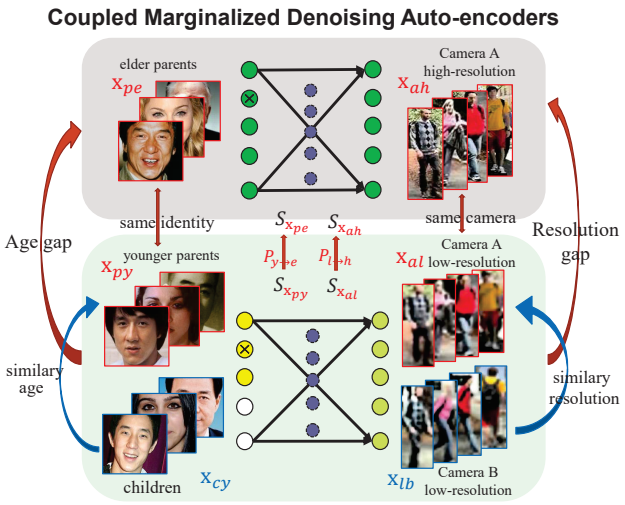


Figure 1: Framework of our proposed algorithm on two applications, e.g., kinship verification and Person re-identification. Specifically, two coupled mDAEs are learned with a projection matrix $P_{y\to e}/P_{l\to h}$ between two hidden layers of gallery $S_{x_{pe}}/S_{x_{ah}}$ and intermediate sets $S_{x_{py}}/S_{x_{al}}$.

## 3 Coupled Marginalized Auto-encoders

### 3.1 Motivation & Overview

In this section, we reiterate our problem scenario, to make it general, where gallery set and probe set lie in different views and different domains (Figure 1). In this situation, we seek an intermediate set, which can be treated as the bridge to mitigate the distribution difference between gallery and probe sets. Take kinship verification for example. The final goal is to verify whether a given image pair (old parent, young child) has kin relation or not. The young parents photos and children photos are considered as similar ages while the old parents photos have much larger ages. Therefore, the gap between old and young parents images is mainly age while the gap between young parents and children is mainly identity difference within biologic heredity. Consequently, this learning process significantly reduces the large gap between distributions to facilitate the kinship verification problem. Also in the person re-identification problem, where gallery set is HR images from camera A and probe is LR images from camera B. We can easily obtain an intermediate set with similar resolution to probe images by down-sampling the images from camera A and hence will have same viewpoint with camera A.

To this end, we propose Coupled Marginalized Denoising Auto-encoders to simultaneously diminish the gap between intermediate to gallery and probe, so that our model

can significantly reduce domain shift between gallery and probe. Figure 1 gives an illustration of our framework, and the application on kinship verification problem and person re-identification problem. In kinship verification, two auto-encoders are built for elder face (old parents photos) and younger face (young parents and children photo), respectively (AE-e, AE-y), which serve as domain adaptation to learn a latent feature space for younger face. A mapping projection $P_{y\to e}$ is learned to couple two hidden layers of gallery $S_{x_{pe}}$ and intermediate sets $S_{x_{py}}$. Note that only hidden layer for old parents and young parents are associated with this coupled term, since these two sets only differ in age. Thus a projection matrix which can map young face to elder one is learned and could be used on the hidden layer of children to map the children sample from AE-y to AE-e. Then a discriminative constraint is developed on the output layer of AE-e to preserve more supervised information.

### 3.2 Denoising Auto-encoder Revisit

Given the $D$-dimension input visual descriptor $x \in \mathbb{R}^D$. The auto-encoder involves two transformations: "input→hidden units $h \in \mathbb{R}^d$", and "hidden units→reconstructed output $\hat{x} \in \mathbb{R}^D$" as encoder and decoder:

$$h = \sigma(Wx + b_1); \quad \hat{x} = \sigma(W^{\mathrm{T}}h + b_2) \tag{1}$$

where $W$ is a $d \times D$ weight matrix, and $b_1 \in \mathbb{R}^d$, $b_2 \in \mathbb{R}^D$ are offset vectors. $\sigma$ is a non-linear activation function.

Recently, marginalized denoising auto-encoder (mDAE) [Chen *et al.*, 2012] was proposed to learn a linear transformation matrix $W$ to replace the encoding and decoding steps, and achieved comparable performance with the original auto-encoder. To make the proposed model more flexible, in comparison, we still preserve encode and decode steps but in a linearized way as:

$$\frac{1}{2n}\sum_{i=1}^{n}\|x_i - MM^{\mathrm{T}}\tilde{x}_i\|_2^2, \tag{2}$$

where $\tilde{x}_i$ is the corrupted version of $x_i$. We can treat $M^{\mathrm{T}}\tilde{x}_i$ is the encoding step, while $MM^{\mathrm{T}}\tilde{x}_i$ as the decoding step. The solution to above objective depends on the randomly corrupted features of each input. To lower the variance, mDAE minimized the overall squared loss of $m$ corrupted versions:

$$\frac{1}{2mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\|x_i - MM^{\mathrm{T}}\tilde{x}_{i,j}\|_2^2, \tag{3}$$

where $\tilde{x}_{i,j}$ is the $j$-th corrupted version of $x_i$. Define $X = [x_1, \cdots, x_n]$, its $m$-times repeated version $\bar{X}$ and its corrupted version $\tilde{X}$. Eq. (3) then can be reformulated as

$$\frac{1}{2mn}\|\bar{X} - MM^{\mathrm{T}}\tilde{X}\|_{\mathrm{F}}^2, \tag{4}$$

which has the closed-form solution for ordinary least squares.

### 3.3 Coupled Marginalized Auto-encoders

When dealing with cross-domain multi-view data problem, we aim to build multiple auto-encoders for different domains, respectively. In this way, each auto-encoder could better uncover more information inside each domain. Assume we have $\{X_h, X_l, Y_l\}$ three datasets, where $X_h$ and $X_l$ are from the same view but two domains, while $X_l$ and $Y_l$ are from the same domain but different views. That is, $X_l$ is a bridge to connect $X_h$ and $Y_l$. We build two marginalized auto-encoders to extract features from $X_h$ and $\{X_l, Y_l\}$, respectively. For simplicity, we set $Z_l = [X_l, Y_l]$. The coupled marginalized auto-encoders learning could be formalinized as:

$$\mathcal{L}_a = \|\bar{X}_h - M_h M_h^{\mathrm{T}}\tilde{X}_h\|_{\mathrm{F}}^2 + \|\bar{Z}_l - M_l M_l^{\mathrm{T}}\tilde{Z}_l\|_{\mathrm{F}}^2, \tag{5}$$

where $M_h$ and $M_l$ are the two transformation matrices for two encoders. However, the two auto-encoders are learned individually, therefore, it is essential to couple two auto-encoders to effective knowledge transfer.

As we mentioned before, $X_h$ and $X_l$ are from the same view but different domains. For example, $X_h$ is the high-resolution data while $X_l$ is the low-resolution data in the same view; or $X_h$ and $X_l$ are the same person but in different ages, e.g., old parents and young parents. Therefore, there should be a high correlation across them and we propose a cross-domain mapping to mitigate the domain shift. To this end, we have the following objective function as:

$$\mathcal{L}_m = \|M_l^{\mathrm{T}}X_l - PM_h^{\mathrm{T}}X_h\|_{\mathrm{F}}^2, \tag{6}$$

where $P$ is the feature mapping matrix, which transforms the hidden layer of one domain to that of the other domain.

Furthermore, the supervised information of positive pairs and negative pairs are very essential to build two discriminative coupled auto-encoders. Assume we have the positive pair in two domains $\{X_h^p, Y_l^p\}$ and negative pairs $\{X_h^n, Y_l^n\}$. We aim to couple the output of positive pairs similar while keeping the output of negative pairs far away. To this end, we propose these discriminative terms:

$$\begin{aligned}\mathcal{L}_d = \quad & \lambda_1\|M_l M_l^{\mathrm{T}}Y_l^p - M_l P M_h^{\mathrm{T}}X_h^p\|_{\mathrm{F}}^2 \\ & -\lambda_2\|M_l M_l^{\mathrm{T}}Y_l^n - M_l P M_h^{\mathrm{T}}X_h^n\|_{\mathrm{F}}^2,\end{aligned} \tag{7}$$

where $\lambda_1$ and $\lambda_2$ are the trade-off parameters. $X_h^p$ and $X_h^n$ are first encoded with $M_h$, then mapped to the other domain and further decoded with $M_l$. To sum up, we propose our coupled marginalized auto-encoders learning:

$$\min_{M_l, M_h, P} \mathcal{L}_a + \alpha\mathcal{L}_m + \mathcal{L}_d, \tag{8}$$

where $\alpha$ is the balanced parameter.

### 3.4 Optimization

To solve the proposed objective function (8), we apply an iterative optimization scheme to update three variables $M_l$, $M_h$ and $P$ one by one. The detailed updating steps are:

**Update** $M_h$:

$$\begin{aligned}\min_{M_h} \|\bar{X}_h - M_h M_h^{\mathrm{T}}\tilde{X}_h\|_{\mathrm{F}}^2 + \alpha\|M_l^{\mathrm{T}}X_l - PM_h^{\mathrm{T}}X_h\|_{\mathrm{F}}^2 \\ +\lambda_1\|M_l M_l^{\mathrm{T}}Y_l^p - M_l P M_h^{\mathrm{T}}X_h^p\|_{\mathrm{F}}^2 \\ -\lambda_2\|M_l M_l^{\mathrm{T}}Y_l^n - M_l P M_h^{\mathrm{T}}X_h^n\|_{\mathrm{F}}^2,\end{aligned} \tag{9}$$

which has a closed-form solution as:

$$\begin{aligned}A_h M_h + B_h M_h P^{\mathrm{T}}P - C_h = 0, \\ \Rightarrow B_h^{-1}A_h M_h + M_h P^{\mathrm{T}}P - B_h^{-1}C_h = 0,\end{aligned} \tag{10}$$

which can be solved with Liapunov function. $A_h = \tilde{X}_h\tilde{X}_h^{\mathrm{T}} - \bar{X}_h\tilde{X}_h^{\mathrm{T}} - \tilde{X}_h\bar{X}_h^{T}$, $B_h = \alpha X_h X_h^{\mathrm{T}} + \lambda_1 X_h^p X_h^{p^T} - \lambda_2 X_h^n X_h^{n^T}$ and $C_h = \alpha X_h X_l^{\mathrm{T}} + \lambda_1 X_h^p Y_l^{p^{\mathrm{T}}} - \lambda_2 X_h^n Y_l^{n^{\mathrm{T}}}$. Ideally, the repeated number $m$ would be $\infty$, so that the denoising transformation $M_h$ could be effectively learned from infinitely copies of noisy data. Fortunately, the matrices $P_h = \tilde{X}_h\tilde{X}_h^{\mathrm{T}}$ and $Q_h = \bar{X}_h\tilde{X}_h^{\mathrm{T}}$ converge to their expected values as $m \to \infty$. Therefore, $A_h$ can be calculated as:

$$A_h = \mathbb{E}(P_h) - \mathbb{E}(Q_h) - \mathbb{E}(Q_h)^{\mathrm{T}}, \tag{11}$$

where the expectations $\mathbb{E}(P_h)$ and $\mathbb{E}(Q_h)$ can be easily computed through mDAE [Chen *et al.*, 2012].

**Update** $M_l$:

$$\begin{aligned}\min_{M_h} \|\bar{Z}_l - M_l M_l^{\mathrm{T}}\tilde{Z}_l\|_{\mathrm{F}}^2 + \alpha\|M_l^{\mathrm{T}}X_l - PM_h^{\mathrm{T}}X_h\|_{\mathrm{F}}^2 \\ +\lambda_1\|M_l M_l^{\mathrm{T}}Y_l^p - M_l P M_h^{\mathrm{T}}X_h^p\|_{\mathrm{F}}^2 \\ -\lambda_2\|M_l M_l^{\mathrm{T}}Y_l^n - M_l P M_h^{\mathrm{T}}X_h^n\|_{\mathrm{F}}^2,\end{aligned} \tag{12}$$

which also has a closed-form solution as:

$$\begin{aligned}A_l M_l + B_l M_l P^{\mathrm{T}}P - C_l = 0, \\ \Rightarrow B_l^{-1}A_l M_l + M_l P^{\mathrm{T}}P - B_l^{-1}C_l = 0,\end{aligned} \tag{13}$$

which can be solved with Liapunov function. $A_l = \tilde{Z}_l\tilde{Z}_l^{\mathrm{T}} - \bar{Z}_l\tilde{Z}_l^{\mathrm{T}} - \tilde{Z}_l\bar{Z}_l^{\mathrm{T}}$, $B_l = \alpha X_l X_l^{\mathrm{T}} + \lambda_1 Y_l^p Y_l^{p^{\mathrm{T}}} - \lambda_2 Y_l^n Y_l^{n^T}$ and $C_l = \alpha X_l X_h M_h P^{\mathrm{T}} + \lambda_1 Y_l^p X_h^{p^{\mathrm{T}}} - \lambda_2 Y_l^n X_h^{n^{\mathrm{T}}}$. And $A_l$ can also be calculated in the same way to $A_h$.

**Update** $P$:

$$\begin{aligned}\min_{P} \alpha\|M_l^{\mathrm{T}}X_l - PM_h^{\mathrm{T}}X_h\|_{\mathrm{F}}^2 \\ +\lambda_1\|M_l M_l^{\mathrm{T}}Y_l^p - M_l P M_h^{\mathrm{T}}X_h^p\|_{\mathrm{F}}^2 \\ -\lambda_2\|M_l M_l^{\mathrm{T}}Y_l^n - M_l P M_h^{\mathrm{T}}X_h^n\|_{\mathrm{F}}^2,\end{aligned} \tag{14}$$

whose solution is also closed-form and can be represented as:

$$P = A_p B_p^{-1}, \tag{15}$$

where $A_p = M_l^{\mathrm{T}}(\alpha X_l X_h^{\mathrm{T}} + \lambda_1 Y_l^p X_h^{p^{\mathrm{T}}} - \lambda_2 Y_l^n X_h^{n^{\mathrm{T}}})M_h$ and $B_p = M_h^{\mathrm{T}}(\alpha X_h X_h^{\mathrm{T}} + \lambda_1 Y_l^p X_h^{p^{\mathrm{T}}} - \lambda_2 Y_l^n X_h^{n^{\mathrm{T}}})M_h$.

When iterative updating is finished, the new features for probe and gallery are calculated as $Y_l^{new} = M_l M_l^T Y_l$, and $X_h^{new} = M_l P M_h^T X_h$, respectively, for later tasks.

## 4 Experiments

We evaluate our approach on two applications, e.g., person re-identification and kinship verification.

### 4.1 Experimental Setting

To evaluate the effectiveness of proposed method in SR person re-identification, we mainly compare our approach with two types of related methods, e.g., metric learning methods and feature learning methods. The metric learning methods include large margin nearest neighbor (LMNN) [Weinberger *et al.*, 2005], information theoretic metric learning (ITML) [Davis *et al.*, 2007], KISS metric learning (KISSME) [Koestinger *et al.*, 2012] and probabilistic relative distance comparison (PRDC) [Zheng *et al.*, 2013]. The compared feature learning methods include symmetry-driven

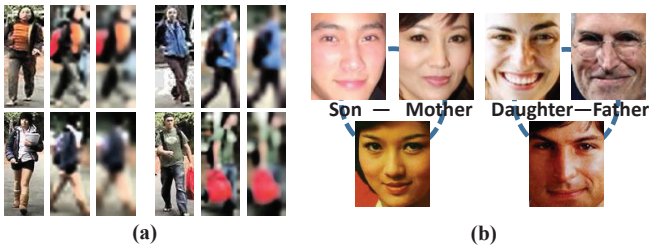


Figure 2: (a) Images in VIPeR, high resolution gallery images from camera A (left), followed by two low resolution probe images from camera B with down sampling rate $1/4$ and $1/8$. (b) Sample images in UB KinFace. Each group consists images for children (top-left), old parents (top-right) and young parents (lower) as the bridge.

accumulation of local features (SDALF) [Farenzena *et al.*, 2010], unsupervised salience learning (eSDC) [Zhao *et al.*, 2013b], salience matching (SalMatch) [Zhao *et al.*, 2013a], and mid-level filters [Zhao *et al.*, 2014]. The state-of-the-art dictionary learning method SLD$^2$L [Jing *et al.*, 2015] for person re-identification is also included. All compared methods are performed with the online available code provided by the authors, except for SLD$^2$L, whose results are copied from the original paper.

In person re-identification experiments, we adopt a fusion strategy to jointly learn the proposed model on account of both patch-based and image-based features. We directly use the patch feature provide by Zhao et al. [Zhao *et al.*, 2013b]. However, due to the well-known misalignment problem, the matching cannot be done directly between the corresponding patches in the probe image and gallery images. Therefore, for each probe patch, the neighbors of the corresponding patches in gallery images should also be searched and calculate each pair's distance. The overall similarity between a pair of probe image and gallery image can be estimated with this adjacency searching scheme. One problem still unsolved with this phenomenon is that when we train the patch based model, the pairwise samples in two auto-encoders may not actually corresponding to each other, which will be considered as noise or outliers for our model. To this end, two processing steps are introduced in our model. First, we adopt a weighted scheme to solve the misalignment, where one patch is reconstructed with all patches with different weights, therefore, we could find the best matched patch to boost the performance. What's more, besides of only comparing the patch-based features, the image-based matching is also conducted with common used ELF descriptor [Gray and Tao, 2008]. Then the final score for $i$-th probe is obtained by adding patch-base and image-base scores together.

For kinship verification, two transfer learning methods, i.e., Transfer Subspace Learning (TSL) [Si *et al.*, 2010] and KVTL [Xia *et al.*, 2011] are compared. Note that KVTL was introduced with UB KinFace dataset to particularly dealing with this task. Besides, the evaluation is also conducted on two state-of-the-art coupled dictionary learning methods, i.e., SCDL [Wang *et al.*, 2012] and CDFL [Huang and Wang, 2013] for comparison. Moreover, a most recent proposed

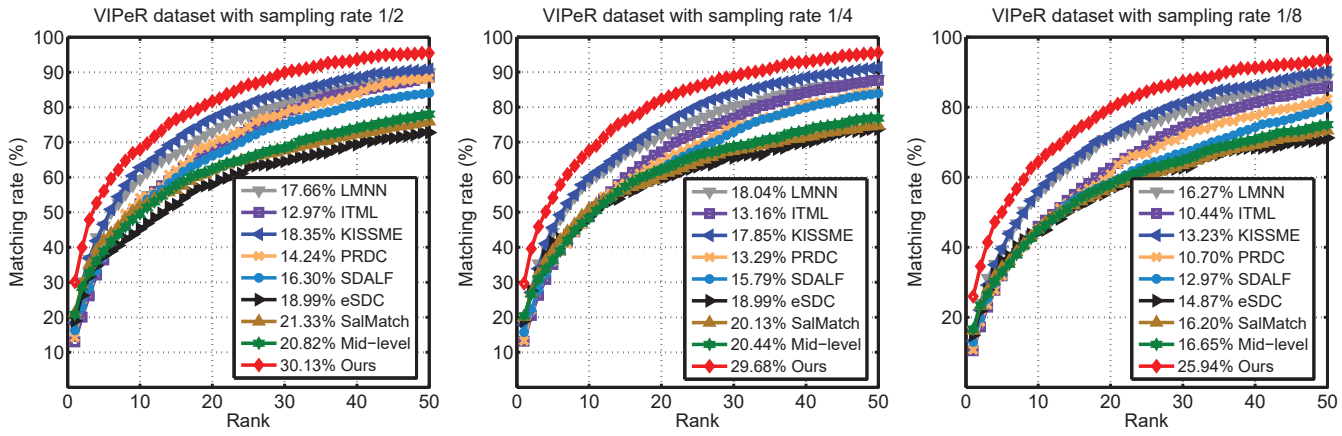


Figure 3: Experiments results on VIPeR dataset with down sampling rate $1/2$ (left), $1/4$ (middle) and $1/8$ (right). Rank-1 matching rate is marked before each approach.

Table 1: Top r ranked matching rates (%) on the VIPeR dataset with sampling rate of 1/8

| Methods | $r = 1$ | $r = 5$ | $r = 10$ | $r = 20$ |
|---|---|---|---|---|
| PRDC | 10.69 | 31.84 | 45.19 | 60.82 |
| LMNN | 16.27 | 39.37 | 55.06 | 71.58 |
| ITML | 10.44 | 31.84 | 45.95 | 62.53 |
| KISSME | 13.23 | 39.56 | 56.01 | 71.90 |
| SDALF | 12.97 | 33.29 | 44.49 | 58.39 |
| eSDC | 14.87 | 36.08 | 44.30 | 56.96 |
| SalMatching | 16.20 | 34.24 | 45.06 | 56.96 |
| Mid-level | 16.65 | 32.91 | 44.87 | 57.91 |
| $SLD^2L$ | 16.86 | 41.22 | 58.06 | 79.00 |
| Ours | **25.95** | **50.00** | **64.37** | **79.75** |

Neighborhood Repulsed Metric Learning (NRML) for kinship verification [Lu *et al.*, 2014] is also compared. For SCDL, CDFL and NRML, we conduct experiments with the online available code, while for TSL and KVTL, the matching results under same setting are copied from original paper.

There are 3 parameters in our model including $\alpha$, $\lambda_1$ and $\lambda_2$, which are tuned through 5-fold cross validation. Specifically, we set them as $\alpha = 1$, $\lambda_1 = 1.4$, $\lambda_2 = 0.4$ for VIPeR, and $\alpha = 10$, $\lambda_1 = 10$, $\lambda_2 = 0.1$ for UB KinFace dataset.

## 4.2 Person Re-identification

**VIPeR Dataset** [Gray *et al.*, 2007] was collected in outdoor academic environment by two cameras from different views. It contains 632 pedestrians with each having a pair of images. All images are normalized to $128\times48$.

In the protocol of person re-identification, we follow the down-sampling operations in [Jing *et al.*, 2015] to generate 632 low-resolution images from camera B. For each pedestrian pair, there is one HR image from camera A and one generated LR image from camera B. Figure 2(a) shows four pairs of images in different resolutions in VIPeR dataset. Then, the evaluation setting follows [Gray and Tao, 2008], where half of the dataset, i.e., 316 image pairs, are randomly split for training, and the remaining half for testing. In the testing, HR images from camera A are used as gallery image set and those LR images from camera B are constructed as probe set. For each probe image, every gallery images are matched to obtain the rank. Rank-$r$ matching rate means the expectation of the correctly matches at rank $r$, and the CMC curve is the cumulated matching rate at all ranks. We conduct 10 trials of evaluation to achieve stable results.

As mentioned above, patch-based and image-based features are both utilized in our framework for person re-identification. Specifically, for image-based, we use Gray and Tao's ELF descriptor [Gray *et al.*, 2007].[1] Other compared metric learning based methods also conduct on this representation, since it is widely used by existing person re-identification techniques. For patch-based feature, we follow the extraction process in [Zhao *et al.*, 2013b][2], thus each patch was represented by a vector with 672 dimension.

Table 1 reports the matching rates in Rank-1, 5, 10 and 20 with sampling rate of $1/8$. We can observe that the matching results are severely dropped, compared with those reported in the original paper due to the low resolution challenge. The performances of our model always surpass these compared methods, and the Rank-1 rate is significantly improved, which verifies the effectiveness of our proposed approach for person re-identification. More detailed comparison results are plotted in Figure 3, with the dataset at different sampling rates (1/2, 1/4, and 1/8). It is observed from the CMC curves that our approach consistently achieves higher matching results at all down-sampling rates. Figure 4(a) shows the matching rates of our approach and its two components.

It is worth to note that the computation time of our approach is proportional to the feature dimension, and the number of patches. Our experiments run on a computer with an Intel I7 quad-core 3.4GHZ CPU and 8GB memory. The computation time of learning coupled auto-encoders on VIPeR dataset is about six minutes, thanks to the close-form solu-

[1] http://www.eecs.qmul.ac.uk/~rlayne/downloads_qmul_elf_descriptor.html
[2] http://www.ee.cuhk.edu.hk/~rzhao/

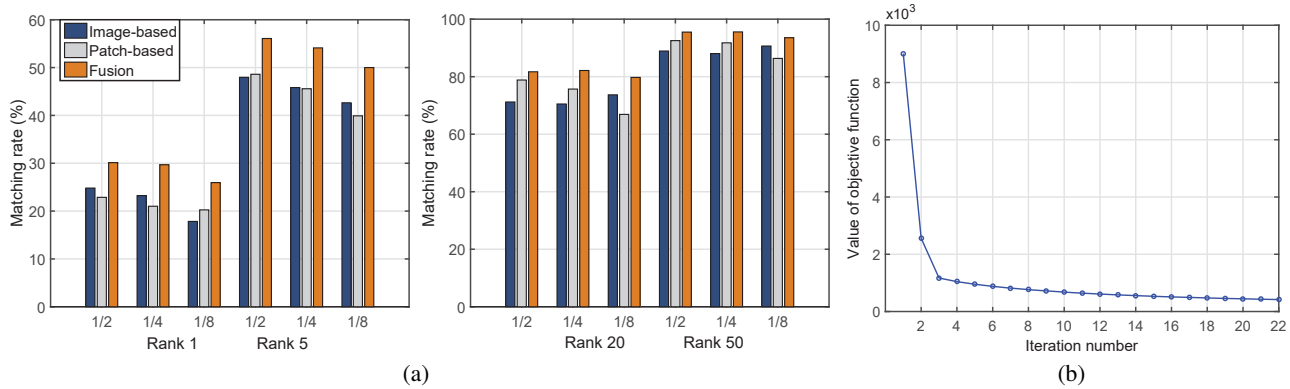


Figure 4: (a) Matching rates of image-based model, patch-based model and the fusion model on VIPeR dataset at sampling rate equal to 1/2, 1/4, 1/8, respectively. (b) Convergence curve of our proposed method on VIPeR dataset.

tion of mDAEs. Further, we also evaluate the convergence of our proposed algorithm on VIPeR dataset (Figure 4(b)), which shows a rapid convergence.

### 4.3 Kinship Verification

Currently, UB KinFace [Shao *et al.*, 2011][3] is the only dataset collected with children, young parents and old parents. The dataset consists of 600 images which can be separated into 200 groups (two persons each group). Each group is composed of child and parent, while each parent has their young and old images. All images in the database are real-world images of public figures downloaded from the Internet.

In the following experiments, we follow the feature extracting setting with [Xia *et al.*, 2011]. The cropped faces (Figure 2(b)) are first obtained with facial landmark detection, and aligned to canonical faces using an affine transform. We then extract the Gabor features (5 scales and 8 directions) from the face image after illumination normalization.

We conduct two evaluation protocols on this dataset: one is kinship verification and the other is child-old parent matching. First, the 200 groups are randomly split into five folds with 40 pairs each fold, then the two protocols are both performed with five-fold cross validation. For the verification protocol, 40 positive pairs and 40 negative pairs are generated using the testing 40 pairs at each fold. The true child-parent pairs are positive examples, while the children with randomly selected non-corresponding parents form negative pairs. Those 80 pairs are given to be classified into true or false pairs. The classification process is simply using Euclidean distance and ROC curve to produce the verification accuracy (area under curve). For the child-old parent matching, similar as person re-identification problem, the Rank-$r$ recognition rates are reported on the 40 child probe and parent gallery pairs at each fold. The results of five-fold cross-validation on both protocols are provided in Table 2. Both the kinship verification rate and Rank-$r$ matching rates show our method's advantage. Take the poor quality of this dataset's "wild" images into consideration, the improvements are significant enough to demonstrate our proposed method's effectiveness.

[3] http://www1.ece.neu.edu/~yunfu/research/Kinface/Kinface.htm

Table 2: Verification accuracy (left column) and Top r ranked matching rates (%) (right two columns) on UB KinFace dataset

| Methods | ACC | Rank 10 | Rank 20 |
|---|---|---|---|
| NRML | 55.50±4.01 | 30.00 | 57.50 |
| CDFL | 61.25±3.26 | 35.50 | 57.50 |
| SCDL | 59.00± 5.55 | 37.50 | 62.50 |
| TSL | 56.11±2.72 | N/A | N/A |
| KVTL | 56.67±6.93 | N/A | N/A |
| Ours | **63.25±2.44** | **45.00** | **75.00** |

## 5 Conclusion

In this paper, we proposed Coupled Marginalized Denoising Auto-encoders for cross-domain learning, where we built two marginalized denoising auto-encoders, each for one domain to extract discriminative features. To better align two auto-encoders, we designed a feature mapping matrix to transform the hidden layer features of one domain close to that of the other. In this way, the mapping matrix could better couple two domains to mitigate the domain shift. Furthermore, two supervised terms, intra-class and inter-class regularizers on the output of one of the auto-encoders were developed to generate discriminative output for two domains. Specifically, the mapped hidden-layer features of one domain was decoded with the weight matrix of the other auto-encoder, then two discriminative terms were imposed on the output of two domains. Experimental results on two real-world applications, i.e., kinship verification and person re-identification, demonstrated the superiority of our method, by comparing with the state-of-the-art algorithms.

## 6 Acknowledgements

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

## References


[Chen *et al.*, 2012] Minmin Chen, Zhixiang Xu, Fei Sha, and Kilian Q Weinberger. Marginalized denoising autoencoders for domain adaptation. In *ICML*, pages 767–774, 2012.

[Davis *et al.*, 2007] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007.

[Ding and Fu, 2014] Zhengming Ding and Yun Fu. Low-rank common subspace for multi-view learning. In *ICDM*, pages 110–119. IEEE, 2014.

[Ding *et al.*, 2015] Zhengming Ding, Ming Shao, and Yun Fu. Missing modality transfer learning via latent low-rank constraint. *TIP*, 24(11):4322–4334, 2015.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[Fang *et al.*, 2010] Ruogu Fang, Kevin D Tang, Noah Snavely, and Tsuhan Chen. Towards computational models of kinship verification. In *ICIP*, pages 1577–1580. IEEE, 2010.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367. IEEE, 2010.

[Gray and Tao, 2008] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008.

[Gray *et al.*, 2007] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, volume 3. Citeseer, 2007.

[Huang and Wang, 2013] De-An Huang and Yu-Chiang Frank Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*, pages 2496–2503. IEEE, 2013.

[Jing *et al.*, 2015] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, pages 695–704, 2015.

[Koestinger *et al.*, 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012.

[Liu *et al.*, 2011] Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, pages 3209–3216. IEEE, 2011.

[Lu *et al.*, 2014] Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *TPAMI*, 36(2):331–345, 2014.

[Ranzato *et al.*, 2008] Marc'aurelio Ranzato, Y lan Boureau, and Yann L. Cun. Sparse feature learning for deep belief networks. In *NIPS*, pages 1185–1192. 2008.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[Shao *et al.*, 2011] Ming Shao, Siyu Xia, and Yun Fu. Genealogical face recognition based on ub kinface database. In *CVPRW*, pages 60–65. IEEE, 2011.

[Si *et al.*, 2010] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *TKDE*, 22(7):929–942, 2010.

[Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.

[Wang and Fu, 2016] Shuyang Wang and Yun Fu. Face behind makeup. In *AAAI*, 2016.

[Wang *et al.*, 2012] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*, pages 2216–2223. IEEE, 2012.

[Weinberger *et al.*, 2005] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005.

[Wu and Jia, 2012] Xinxiao Wu and Yunde Jia. View-invariant action recognition using latent kernelized structural svm. In *ECCV*, pages 411–424. Springer, 2012.

[Xia *et al.*, 2011] Siyu Xia, Ming Shao, and Yun Fu. Kinship verification through transfer learning. In *IJCAI*, volume 22, page 2539, 2011.

[Zhang *et al.*, 2011] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520. IEEE, 2011.

[Zhao and Fu, 2015] Handong Zhao and Yun Fu. Dual-regularized multi-view outlier detection. In *IJCAI*, pages 4077–4083, 2015.

[Zhao *et al.*, 2013a] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *ICCV*, pages 2528–2535. IEEE, 2013.

[Zhao *et al.*, 2013b] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593. IEEE, 2013.

[Zhao *et al.*, 2014] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151. IEEE, 2014.

[Zheng *et al.*, 2013] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 35(3):653–668, 2013.