Deep Localized Metric Learning

Yueqi Duan, Jiwen Lu, *Senior Member, IEEE*, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— Metric learning has been widely used in many visual analysis applications, which learns new distance metrics to measure the similarities of samples effectively. Conventional metric learning methods learn a single linear Mahalanobis metric, yet such linear projections are not powerful enough to capture the nonlinear relationships. Recently, deep metric learning approaches, such as discriminative deep metric learning and deep transfer metric learning, have been introduced to fully exploit the nonlinearity of samples by learning hierarchical nonlinear transformations. However, these methods only learn holistic metrics over the input space and are limited for the heterogeneous data sets, where data varies locally. In this paper, we propose a deep localized metric learning approach for visual recognition by learning multiple fine-grained deep localized metrics. We first learn K local subspaces and one holistic subspace with the K-auto-encoders-based clustering. Then, given an input pair, we compute its localized distance on each learned subspace and obtain the final distance representation. Finally, we train the entire neural networks to ensure the distances of positive pairs smaller than negative pairs by a large margin. Experimental results on three visual recognition applications, including face recognition, person re-identification, and scene recognition, show that our DLML outperforms most existing metric learning approaches.**

***Index Terms*— Deep metric learning, local metric learning, K-auto-encoders, visual recognition.**

I. Introduction

VISUAL recognition has attracted much attention in computer vision, which is widely applicable in numerous applications, such as face recognition [1]–[3], person re-identification [4]–[6], image classification [7]–[9] and many others. As a representative pattern recognition task, there are two main classes of approaches to improve the recognition performance: 1) extracting more discriminative descriptors, and 2) designing more effective distance metrics. The first aims to separate different objects in the feature level, and the second is to obtain a new fine-grained distance metric for better similarity measurement of descriptors. In this paper, we mainly focus on the second category.

Generally, the similarity measurements are task-specific because different datasets usually subject to varying distributions. Unlike hand-crafted distance metrics which perform the same function on all the tasks ignoring the differences in data distribution, metric learning approaches which aim to learn distance metrics from the dataset are more data-adaptive and obtain better performance. A variety of metric learning methods have been proposed in recent years [3], [10]–[17], where most approaches learn a single linear Mahalanobis metric through the labeled training dataset or the positive and negative input pairs, such as large margin nearest neighbor (LMNN) [15], information-theoretic metric learning (ITML) [14] and logistic discriminant metric learning (LDML) [10]. However, a single Mahalanobis metric only learns linear transformation, which is not powerful enough to capture the nonlinear correlations of the samples. Kernel tricks are usually employed to address the limitation by first mapping the dataset into the high-dimensional space and then learning the metrics on the transformed space [5], [18]–[20]. However, these approaches usually suffer from scalability problem as they fail to obtain the explicit nonlinear projections.

More recently, several deep metric learning approaches have been proposed by learning hierarchical nonlinear projections, which present outstanding performance such as discriminative deep metric learning (DDML) [3], deep transfer metric learning (DTML) [17], deep coupled metric learning (DCML) [21] and multi-manifold deep metric learning (MMDML) [22]. However, these methods only learn holistic metrics for measuring the similarities over the input space, which may not be able to handle the data varying locally. Inspired by the fact that localized metric learning exploits the local specificities and deep learning presents outstanding nonlinear capability, we propose a deep localized metric learning (DLML) method by learning multiple deep localized metrics, so that the learned metrics are more fine-grained for local subspaces. We consider an example in face recognition for a clearer illustration of the advantages. Some faces are more easily classified for another mistakenly. For example, we tend to misrecognize an Asian female for another Asian female, rather than an European male. Learning holistic metrics may lead to a high misclassification rate under these situations as they suffer from small inter-class distances, while localized metric learning emphasizes the confusing "Asian females" and "European males" local subspaces, and is therefore reasonable to improve the classification performance. Fig. 1 illustrates the pipeline of the proposed approach. We first train one holistic subspace to minimize the reconstruction error of all input samples, and train K local spaces with the K-Auto-Encoders (KAEs) based

Manuscript received December 28, 2016; revised April 4, 2017; accepted May 28, 2017. Date of publication June 1, 2017; date of current version October 24, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001; in part the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090; in part by the National 1000 Young Talents Plan Program; in part by the National Basic Research Program of China under Grant 2014CB349304; in part by the Ministry of Education of China under Grant 20120002110033; and in part by the Tsinghua University Initiative Scientific Research Program. This paper was recommended by Associate Editor W. Zuo. *(Corresponding author: Jiwen Lu.)*

The authors are with the Department of Automation, Tsinghua University, Beijing 100084, China, and also with the State Key Laboratory of Intelligent Technologies and Systems, and Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China (e-mail: duanyq14@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

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

Digital Object Identifier 10.1109/TCSVT.2017.2711015

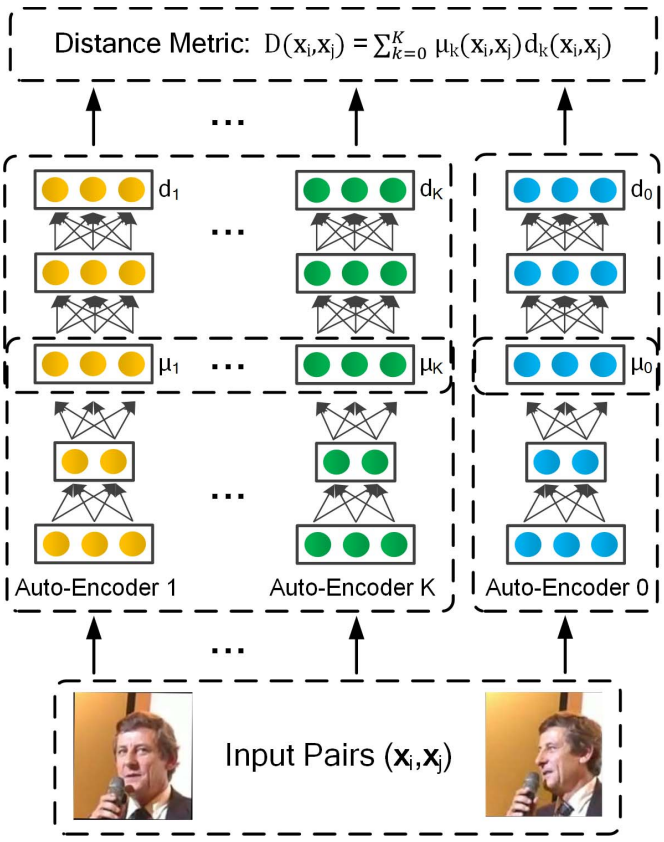


Fig. 1. The flowchart of the proposed DLML approach for visual recognition. We first train the Auto-Encoders to obtain K local subspaces and one holistic subspace with all the training samples in an unsupervised manner. Then, we connect a fully connection network to each Auto-Encoder to learn a hierarchical nonlinear metric for each subspace. Given a pair of input samples $\mathbf{x}_i$ and $\mathbf{x}_j$, we calculate its localized distances $d_k(\mathbf{x}_i, \mathbf{x}_j)$ using the outputs of the entire networks, and fuse the $K+1$ distances into the final distance metric $D(\mathbf{x}_i, \mathbf{x}_j)$. The weight functions of the localized metrics are determined by the reconstruction errors of the Auto-Encoders. Lastly, the parameters of the entire networks are trained to ensure the distances of positive pairs smaller than negative pairs by a large margin.

clustering. Then, we connect a fully connection network to each Auto-Encoder to obtain multiple deep localized metrics in the learned subspaces. The outputs of the deep neural networks are utilized to calculate localized pairwise distances, and the reconstruction errors of the Auto-Encoders determine the weight functions. Lastly, we train the entire deep neural networks to ensure the final distances of positive pairs smaller than those of negative pairs by a large margin. Extensive experimental results on three different visual recognition tasks including face recognition, person re-identification and scene recognition show that the proposed DLML approach outperforms most existing metric learning methods.

We organized the rest of the paper as follows: Section II briefly review two related topics. Section III details the proposed DLML approach. Section IV introduces the experimental results on three different visual recognition applications, and Section V concludes the paper.

II. Related Work

In this section, we briefly review two related topics: metric learning and visual recognition.

A. Metric Learning

The purpose of metric learning is to learn new distance metrics in order to reduce the distances between positive pairs and increase the distances between negative pairs. Most existing metric learning methods learn a single linear Mahalanobis metric for similarity measurement [10], [14], [15], [23], [24]. For example, Davis *et al.* [14] proposed an information-theoretic metric learning (ITML) method by leveraging Mahalanobis distances and multivariate Gaussian distribution in an information-theoretic setting. Guillaumin *et al.* [10] presented a logistic discriminant metric learning (LDML) approach to learn distance metric with logistic discriminant from labeled face pairs. Koestinger *et al.* [23] proposed a KISS metric embedding (KISSME) method to learn a distance metric with equivalence constraints, following the KISS (keep it simple and straightforward) principle. However, a single linear transformation cannot capture the nonlinear relationships between sample pairs, which are quite common in the real-world applications. To address this, kernel tricks are usually applied for nonlinear transformations [5], [18], [19], yet they cannot obtain the explicit functions and face scalability problem. More recently, several deep metric learning methods have been proposed to address the limitation by learning hierarchical nonlinear transformations [3], [17]. For example, Hu *et al.* [3] proposed a discriminative deep metric learning (DDML) method by learning a distance metric with a deep neural network. They also presented a deep transfer metric learning (DTML) [17] approach by transfering information from the source domain to the target domain for cross-domain recognition. However, these deep metric learning approaches only learn holistic distance metrics over the input space, which are too restrictive for heterogeneous datasets especially when samples vary locally.

B. Visual Recognition

There are mainly two categories of approaches for visual recognition: feature representation and metric learning. Feature representation aims to extract discriminative features to separate different objects in the feature level, which should be robust to illuminations, rotations, viewpoints and occlusions [1], [2], [7]–[9], [25]–[32]. For example, Ahonen *et al.* [25] proposed a local binary pattern (LBP) based face representation. Ma *et al.* [27] proposed a domain transfer ranked support vector machines (DTRSVM) method by relaxing the constraint to the mean of positive pairs and constructing a discriminative model. Zhao *et al.* [28] studied properties for good filters and patch clusters, and learned mid-level filters from the collected local patches. Lu *et al.* [30] presented a compact binary face descriptor (CBFD) method by learning evenly-distributed and energy-saving binary descriptors in an unsupervised manner. More recently, a number of convolutional neural networks (CNNs) based approaches have been proposed, which obtain the state-of-the-art. Representative CNN features include AlexNet [7], DeepID [1], VGG [2], [8], FaceNet [32], GoogLeNet [33] and ResNet [9]. There have also been numerous metric learning methods proposed in the last decade to address the visual recognition

problem [3]–[5], [10], [11], [14], [15], [23], [29], [34]–[38]. For example, Cai *et al.* [34] presented a deep nonlinear metric learning with independent subspace analysis (DNLML-ISA) method by using an ISA network. Cui *et al.* [11] proposed a pairwise-constrained multiple metric learning (PMML) approach to fuse face region descriptors. Weinberger *et al.* [15] presented a large margin nearest neighbor (LMNN) approach to learn a distance metric k-nearest neighbor classification. Nguyen *et al.* [35] proposed a cosine similarity metric learning (CSML) approach to learn a transformation for cosine similarity measurement. Pedagadi *et al.* [37] learned distance metric for pedestrian re-identification by utilizing unsupervised PCA and supervised local Fisher (LF) discriminant analysis for dimension reduction. Paisitkriangkrai *et al.* [38] proposed a structured learning based approach by exploiting multiple visual features through metric ensembles.

III. Proposed Approach

In this section, we first propose the K-Auto-Encoders based clustering, and then present the localized pairwise distance and the deep localized metric learning. Lastly, we introduce the implementation details of the proposed method.

A. K-Auto-Encoders Based Clustering

There have been numerous metric learning approaches proposed recently, yet there are three key limitations of the existing metric learning approaches:

1) Most existing metric learning approaches simply learn a single linear transformation using the Mahalanobis metric. However, subjects usually lie in a nonlinear manifold in real-world visual applications, where a linear metric may not be able to fully capture the relationship between the input pairs.
2) Kernel tricks are usually employed to learn a nonlinear discriminative mapping in an implicit high-dimensional feature space, yet kernel-based approaches usually suffer from scalability problems.
3) Deep metric learning methods learn hierarchical nonlinear projections which present outstanding capability of capturing the nonlinearity of samples. However, existing deep metric learning methods such as DDML [3] and DTML [17] only learn holistic metrics, which is restrictive for heterogeneous datasets varying locally.

Inspired by the fact that localized metric learning approaches learn a set of local metrics by dividing metric learning from a clustering process and present outstanding performance, we propose a K-Auto-Encoders (KAEs) based deep localized metric learning method with deep neural networks to address the limitations above. Localized metric learning enables unrestricted multiple mappings, and deep learning provides powerful hierarchical nonlinear projections. More specifically, we employ K networks for local metric learning which are labeled by 1 to K in Fig. 1, and one network for holistic metric learning labeled by 0. For the local networks, we utilize KAEs to perform a K-clustering process, which is the key step for local metric learning. For the holistic network, we simply learn the Auto-Encoder to minimize the reconstruction loss.

We train the Auto-Encoders with all the images in an unsupervised manner. We especially explain the details of training local KAEs. K-means is one of the most effective unsupervised clustering methods, which optimizes in an iterative two-step procedure: 1) quantizing each sample point into a clustering center, and 2) updating each cluster with the related samples. Referencing K-means, we train our KAEs iteratively with the similar two steps. First, we cluster each sample $\mathbf{x}_n$ into the specific Auto-Encoder which minimizes its reconstruction loss $\epsilon_{nk} = ||\Delta \mathbf{x}_{nk}||_2$. Then, we update the parameters of each Auto-Encoder using the corresponding samples to minimize their reconstruction loss. Fig. 2 explains the detailed procedures of training KAEs.

The Auto-Encoders learn $K + 1$ subspaces through the training samples, where K of them are local subspaces trained by parts of the set and the other is the holistic subspace. The parameters of the Auto-Encoders will be further trained with the entire networks for better localized metrics. There are three main advantages of the proposed KAEs based clustering. First, it effectively learns K local subspaces which comprehensively captures the latent manifold of the training samples from multiple angles. Second, the relationship between samples and each local subspace can be directly described by the reconstruction error of the Auto-Encoder. Third, compared with other unsupervised clustering approaches such as K-means whose parameters are fixed in the metric learning procedure, the proposed KAEs based method constructs entire deep networks in localized metric learning, where the parameters of the Auto-Encoders are further trained by the objective function of metric learning to obtain better subspaces.

B. Localized Pairwise Distance

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ be the N samples from the training set, where $\mathbf{x}_n \in \mathbb{R}^d$ $(1 \leq n \leq N)$ is the nth training sample. We first train the local KAEs using the iterative two-step procedure as well as the holistic Auto-Encoder with all the input samples in an unsupervised manner. Then, we connect each Auto-Encoder with a fully connection network to construct $K + 1$ deep neural networks in total, where K networks are for the local metrics and one network learns the holistic metric. All the networks share the same structure but different in parameters. We denote each entire network (including the Auto-Encoder and the fully connection layers) has $M+1$ layers and $\mathbf{h}_k^{(m)}$ represents the mth layer of the kth network, where $k = 0, 1, \cdots, K$ and $m = 0, 1, \cdots, M$. In the kth network, $\mathbf{h}_k^{(0)}$ represents the input samples. Suppose $\mathbf{W}_k^{(m)}$ and $\mathbf{b}_k^{(m)}$ are the projection matrix and the bias vector between the $(m-1)$th layer and the mth layer of the kth network, respectively, so that the mth layer can be computed as:

$$\mathbf{h}_k^{(m)} = \phi(\mathbf{W}_k^{(m)} \mathbf{h}_k^{(m-1)} + \mathbf{b}_k^{(m)}), \tag{1}$$

where ϕ represents the nonlinear activation function, such as tanh, sigmoid and ReLU.

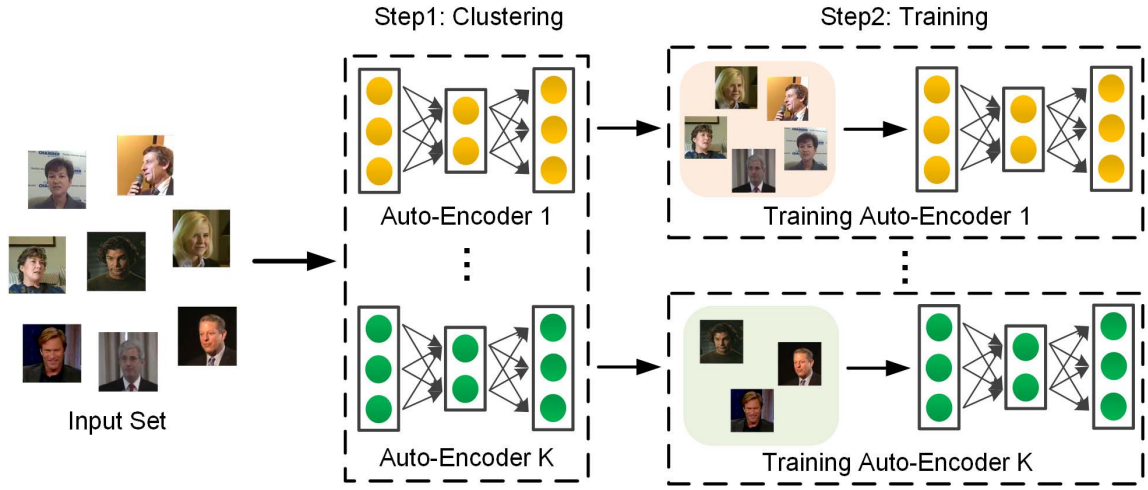


Fig. 2. The detailed explanation of training KAEs. The training procedure is divided into two steps: 1) clustering each input sample into one of the KAEs with the minimum reconstruction loss, and 2) training the parameters of the associate Auto-Encoder with the corresponding samples. We iteratively execute the two steps until convergence.

For each pair of input samples $\mathbf{x}_i$ and $\mathbf{x}_j$, we can obtain K local representations and one holistic representation for each of them. We denote $f_k(\mathbf{x}_i)$ as the representation of $\mathbf{x}_i$ in the kth network and $f_k(\mathbf{x}_j)$ as the representation of $\mathbf{x}_j$, where f_k is the output of the kth network $\mathbf{h}_k^{(M)}$. Therefore, there are $K+1$ distances under different metrics in total:

$$d_k(\mathbf{x}_i, \mathbf{x}_j) = ||f_k(\mathbf{x}_i) - f_k(\mathbf{x}_j)||_2, \tag{2}$$

where $d_0(\mathbf{x}_i, \mathbf{x}_j)$ is calculated with the global metric and the others are with the local metrics.

In order to integrate the $K+1$ distances into the final distance, we define a weight function μ_k for each metric according to the reconstruction error of the Auto-Encoder as follows:

$$\mu_k(\mathbf{x}_i, \mathbf{x}_j) = \frac{s(\epsilon_{ik}, \epsilon_{jk})}{\sum_{l=0}^{K} s(\epsilon_{il}, \epsilon_{jl})}, \tag{3}$$

where

$$s(\epsilon_{ik}, \epsilon_{jk}) = \exp(-\frac{\epsilon_{ik} + \epsilon_{jk}}{2\sigma^2}). \tag{4}$$

The reconstruction error can be considered as the energy of the samples on each subspace, where better projections lead to smaller energies and should deliver greater significance. As each input sample has different energy distribution on local and holistic subspaces, the weight functions are designed to enlarge the influence of the metrics with small reconstruction error in a smooth manner, and to lower the significance of the metrics with large reconstruction error. We regularize the weights $\mu_k(\mathbf{x}_i, \mathbf{x}_j)$ to scale its L1-norm to 1.

With the weight functions, the final distance can be computed as follows:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=0}^{K} \mu_k(\mathbf{x}_i, \mathbf{x}_j) d_k(\mathbf{x}_i, \mathbf{x}_j). \tag{5}$$

The set of K determines the comprehensiveness of exploiting local subspaces. Larger K leads to a better description of the input manifold, yet it may suffer from severer overfitting and higher computational costs. The proposed localized metric learning method degenerate into conventional deep metric learning if we set $K = 0$.

C. Deep Localized Metric Learning

Having obtained the pairwise distance through the neural networks, we hope to define an objective function for training the parameters, where positive pairs have smaller distances than negative pairs by a relatively large margin. We assume that the distance margin is 2δ where δ is a positive number. Given a preset parameter l, the distances between positive pairs should be less than $l - \delta$ while larger than $l + \delta$ for negative pairs. Fig. 3 shows an intuitive illustration of the proposed localized metric learning method. We formulate the following constraint to realize the above motivations:

$$\delta - y_{ij}(l - D(\mathbf{x}_i, \mathbf{x}_j)) < 0, \tag{6}$$

where y_{ij} represents the label information, equaling to 1 for positive pairs and -1 for negative pairs. The formulation (6) can be rewritten as $D(\mathbf{x}_i, \mathbf{x}_j) < l - \delta$ for positive pairs and $D(\mathbf{x}_i, \mathbf{x}_j) > l + \delta$ for negative pairs, which clearly demonstrates the physical meaning of the proposed method.

Inspired by the above motivations, we formulate the following objective function to learn deep localized metrics:

$$\begin{aligned} \min_{\mathbf{W}_k^{(m)}, \mathbf{b}_k^{(m)}} J &= J_1 + \lambda J_2 \\ &= \sum_{i,j} f(\delta - y_{ij}(l - D(\mathbf{x}_i, \mathbf{x}_j))) \\ &\quad + \lambda \sum_{k=0}^{K} \sum_{m=1}^{M} (||\mathbf{W}_k^{(m)}||_F^2 + ||\mathbf{b}_k^{(m)}||_2^2), \end{aligned} \tag{7}$$

where

$$f(x) = \frac{1}{\eta} \log(1 + \exp(\eta x)) \tag{8}$$

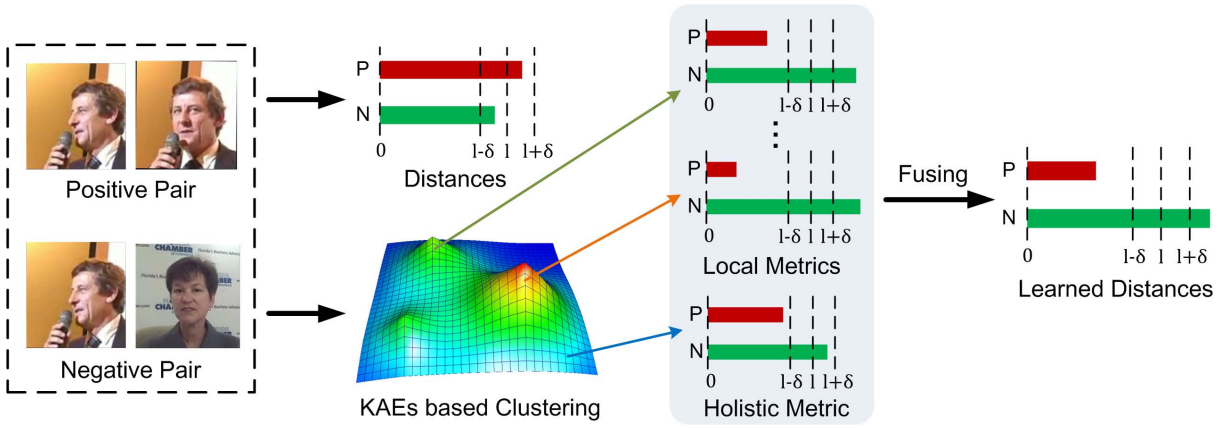


Fig. 3. An intuitive illustration of the proposed localized metric learning method. Given a positive input pair and a negative input pair, we can obtain their distances in the original space. Because of the large intra-class variations caused by poses, illuminations and expressions, the distance of the positive pair may even larger than the negative pair. We first train K local Auto-Encoders and one holistic Auto-Encoder for subspace clustering, and then correspondingly learn K local metrics and one holistic metric to ensure the learned final distance of the positive pair less than $l-\delta$, while larger than $l+\delta$ for the distance of the negative pair.

is the generalized logistic approximation of the function $\max(x, 0)$.

The physical meaning of J_1 is to ensure a margin of 2δ between the distances of positive pairs and negative pairs. J_2 aims to regularize the parameters of the $K+1$ neural networks. λ is to balance the two terms.

We apply the batch gradient scheme to update the parameters of the networks with the objective function, and we can calculate the derivatives of $\mathbf{W}_k^{(m)}$ and $\mathbf{b}_k^{(m)}$ as follows:

$$\frac{\partial J}{\partial \mathbf{W}_k^{(m)}} = \sum_{i,j}(\mathbf{\Phi}_{k,ij}^{(m)}(\mathbf{h}_{k,i}^{(m-1)})^T + \mathbf{\Phi}_{k,ji}^{(m)}(\mathbf{h}_{k,j}^{(m-1)})^T) + \lambda \mathbf{W}_k^{(m)} \tag{9}$$

$$\frac{\partial J}{\partial \mathbf{b}_k^{(m)}} = \sum_{i,j}(\mathbf{\Phi}_{k,ij}^{(m)} + \mathbf{\Phi}_{k,ji}^{(m)}) + \lambda \mathbf{b}_k^{(m)} \tag{10}$$

where $\mathbf{h}_{k,i}^{(m)}$ and $\mathbf{h}_{k,j}^{(m)}$ represent the mth layer of the kth neural network under the inputs of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. As the learning rates of the Auto-Encoders are set relatively small, we ignore the derivatives of the weights μ_k.

For the last layer $\mathbf{h}_k^{(M)}$ of the kth network, we calculate the updating functions $\mathbf{\Phi}_{k,ij}^{(m)}$ and $\mathbf{\Phi}_{k,ji}^{(m)}$ as follows:

$$\begin{aligned}\mathbf{\Phi}_{k,ij}^{(M)} &= \mu_k(\mathbf{x}_i, \mathbf{x}_j) f'(c)(\mathbf{h}_{k,i}^{(M)} - \mathbf{h}_{k,j}^{(M)}) \circ \phi'(\mathbf{z}_{k,i}^{(M)}) \\ \mathbf{\Phi}_{k,ji}^{(M)} &= \mu_k(\mathbf{x}_i, \mathbf{x}_j) f'(c)(\mathbf{h}_{k,j}^{(M)} - \mathbf{h}_{k,i}^{(M)}) \circ \phi'(\mathbf{z}_{k,j}^{(M)}),\end{aligned}$$

and for other layers, the updating functions are formulated as follows:

$$\begin{aligned}\mathbf{\Phi}_{k,ij}^{(m)} &= (\mathbf{W}_k^{(m+1)})^T \mathbf{\Phi}_{k,ij}^{(m+1)} \circ \phi'(\mathbf{z}_{k,i}^{(m)}) \\ \mathbf{\Phi}_{k,ji}^{(m)} &= (\mathbf{W}_k^{(m+1)})^T \mathbf{\Phi}_{k,ji}^{(m+1)} \circ \phi'(\mathbf{z}_{k,j}^{(m)}),\end{aligned} \tag{11}$$

where $\circ$ represents the element-wise multiplication and c, $\mathbf{z}_{k,i}^{(M)}$ and $\mathbf{z}_{k,j}^{(m)}$ are defined as follows:

$$\begin{aligned}c &\triangleq \delta - y_{ij}(l - D(\mathbf{x}_i, \mathbf{x}_j)) \\ \mathbf{z}_{k,i}^{(m)} &\triangleq \mathbf{W}_k^{(m)} \mathbf{h}_{k,i}^{(m-1)} + \mathbf{b}_k^{(m)} \\ \mathbf{z}_{k,j}^{(m)} &\triangleq \mathbf{W}_k^{(m)} \mathbf{h}_{k,j}^{(m-1)} + \mathbf{b}_k^{(m)}.\end{aligned} \tag{12}$$

With the derivatives above, we use the gradient descent algorithm with back-propagation to update the parameters of the neural networks:

$$\mathbf{W}_k^{(m)} = \mathbf{W}_k^{(m)} - \nu \frac{\partial J}{\partial \mathbf{W}_k^{(m)}} \tag{13}$$

$$\mathbf{b}_k^{(m)} = \mathbf{b}_k^{(m)} - \nu \frac{\partial J}{\partial \mathbf{b}_k^{(m)}}, \tag{14}$$

where ν represents the learning rate of the networks.

Algorithm 1 details the approach of our DLML.

D. Implementation Details

In this subsection, we present the implementation details of training Auto-Encoders and the fully connection networks in the proposed method.

1) Auto-Encoders: In the proposed method, we train one holistic Auto-Encoder and K local Auto-Encoders for localized metric learning. We apply the sigmoid as the activation function. For the holistic Auto-Encoder, we directly learn the parameters with all the training samples to minimize their reconstruction losses. For the local KAEs, as it is easy to cluster all the samples into one of the KAEs if we directly employ random initialization, we design a more fine-grained initialization approach. More specifically, we first learn the parameters by training with all input samples on each of KAEs under different random initialization. This procedure obtains K similar but different holistic subspaces, which are then applied to the iterative two-step learning approach. The learned parameters of KAEs are utilized as the initialization of the metric learning procedure.

2) Fully Connection Networks: In our experiments, we simply apply a normalized random initialization for the fully connection part of the networks, where $\mathbf{b}_k^{(m)}$ is set as 0 and $\mathbf{W}_k^{(m)}$ is set under the uniform distribution:

$$X \sim U[-\frac{\sqrt{6}}{\sqrt{p_k^{(m)} + p_k^{(m-1)}}}, \frac{\sqrt{6}}{\sqrt{p_k^{(m)} + p_k^{(m-1)}}}], \tag{15}$$

Algorithm 1 DLML

Input: Training set $\mathbf{X}$, number of local subspaces K, number of fully connection networks layers M, parameters λ and υ, distance threshold l, margin δ, convergence error ε and iteration number T.

Output: Parameters of the neural networks $\mathbf{W}_k^{(m)}$ and $\mathbf{b}_k^{(m)}$.

Initialize the parameters $\mathbf{W}_k^{(m)}$ and $\mathbf{b}_k^{(m)}$.
Train the holistic Auto-Encoder to minimize the reconstruction error.
Iteratively learn KAEs with the two-step procedure.
for $iter = 1, 2, \cdots, T$ **do**
  Randomly select a sample pair $\mathbf{x}_i$ and $\mathbf{x}_j$ with the label y_{ij}.
  for $m = 1, 2, \cdots, M$ **do**
    Obtain $\mathbf{h}_{k,i}^{(m)}$ and $\mathbf{h}_{k,j}^{(m)}$ with forward-propagation.
  end for
  for $m = M, M-1, \cdots, 1$ **do**
    Update $\mathbf{W}_k^{(m)}$ using (13) with back-propagation.
    Update $\mathbf{b}_k^{(m)}$ using (14) with back-propagation.
  end for
  Calculate J_t using (7).
  if $t > 1$ and $|J_t - J_{t-1}| < \varepsilon$ **then**
    break.
  end if
end for
end return $\mathbf{W}_k^{(m)}$ and $\mathbf{b}_k^{(m)}$.

where $p_k^{(m)}$ represents the dimension of the mth layer in the kth network.

IV. Experiments

We evaluated the proposed DLML method on three different visual analysis tasks including face recognition, person re-identification and scene recognition. The followings describe the experimental details and results.

A. Face Recognition

In this subsection, we conducted experiments on two widely used face recognition datasets including labeled faces in the wild (LFW) [39] and YouTube Face (YTF) [40].

The LFW dataset [39] contains 5749 subjects with total 13233 face images, which were captured from the web in wild conditions, varying from ages, poses, resolutions, illuminations and many others. Fig. 4 shows some example faces from the LFW dataset. The supervised learning procedure on the LFW dataset can be divided into image restricted setting and image unrestricted setting. We applied the former setting by only utilizing the pairwise labels. We conducted our experiments on the "View 2" dataset, following the standard evaluation protocol [39], including 6000 image pairs with half of them matched and the others mismatched. They were divided into 10 folds with 300 matched pairs and 300 mismatched pairs in each fold. We aligned each facial image with a conventional 2D affine transformation and then cropped into 80×150 to remove the background.

Fig. 4. Face samples of the LFW dataset. Each column represents the same person.

Fig. 5. Face samples of the YTF dataset. Each column represents the same person.

The YTF dataset [40] consists of 3415 videos of 1595 different subjects collected from the YouTube website, and each video clip has 181.3 frames on average. The videos are varying from poses, illuminations and expressions. Fig. 5 shows some example faces from the YTF dataset. We also followed the standard image restricted face verification protocol, which contained 10 folds of 5000 video pairs in total. In each fold, half of the video pairs are positive and the other half are negative.

1) Feature Representation: Before applying the proposed DLML approach, we first extracted features for facial images of the LFW and YTF datasets following [3].

For the LFW dataset, we extracted three types of features including local binary patterns (LBP) [25], dense SIFT (DSIFT) [41] and sparse SIFT (SSIFT) [10]. We summarize the details of these features as follows:

1) LBP: We divided each facial image into several non-overlapping 10×10 regions. As the image size is 80×150, there are 120 ($=8 \times 15$) regions in total for each face. We extracted a uniform 59-dimensional LBP feature for each region, which were concatenated into a 7080-dimensional vector.
2) DSIFT: We extracted a 128-dimensional SIFT descriptor densely on each non-overlapping 16×16 region, and obtained 45 ($=5 \times 9$) descriptors for each face. They were concatenated into a 5760-dimensional vector.
3) SSIFT: We computed the SSIFT descriptors on nine facial landmarks under three scales for each facial image, and they were concatenated into a 3456-dimensional vector.

Moreover, we followed [3] by using the square root of each feature to evaluate the combined features. We applied

the whitened PCA (WPCA) approach to reduce the feature dimension to 500 for redundancy reduction [42]. The nearest neighbor classifier with cosine similarity is applied for classification.

For the YTF dataset, we directly applied the provided baseline descriptors from the dataset [40], which included LBP [25], Four-Patch LBP (FPLBP) [43] and Center-Symmetric LBP (CSLBP) [44]. We simply averaged the features in the same video clip to construct the final representation as faces were pre-aligned and cropped through facial landmarks. Moreover, we also utilized WPCA to reduce the dimension of each feature to 500.

In our experiments, we fixed the structure of the Auto-Encoders as [500→400→300→400→500], following by a two-layer fully connection to construct the entire networks, whose dimension is set as [500→400→300]. We fixed the K to 4 for LFW and YTF. The distance threshold l, the margin δ and parameter λ are fixed as 3, 1 and 10^{-2} respectively for all the experiments by experience.

2) Comparison With Existing Methods: We compared our DLML with existing methods under the image-restricted setting of the LFW and YTF datasets.[1] The compared methods consist of metric learning methods and descriptor based methods. Typical metric learning methods include cosine similarity metric learning (CSML) [35], pairwise-constrained multiple metric learning (PMML) [11], discriminative deep metric learning (DDML) [3], deep nonlinear metric learning with independent subspace analysis (DNLML-ISA) [34], large margin multi-metric learning (LM3L) [45] and deep transfer metric learning with Auto-Encoder regularization (DTML-AE) [17]. Descriptor based methods mainly include convolutional deep belief network (CDBN) [46], Fisher vector [47], pose adaptive filter (PAF) [48], compact binary face descriptors (CBFD) [30] and adaptive probabilistic elastic matching (APEM) [49].

Table I tabulates the recognition accuracy and the standard error, and Fig. 6 shows the ROC curves of the proposed DLML compared with the state-of-the-art approaches under the image restricted setting of the LFW dataset. Similarly, Table II and Fig. 7 show the accuracy and the ROC curves on the YTF dataset. We see that our DLML outperforms the existing state-of-the-art metric learning approaches such as PMML, DNLML-ISA, DDML, LM3L and DTML-AE. Unlike the shallow metric learning methods that simply learn a Mahalanobis metric or apply kernel tricks for nonlinear transformation, our DLML learns deep nonlinear projections which present stronger discriminative power. Compared with deep metric learning approaches including DNLML-ISA, DDML and DTML-AE, our DLML learns multiple deep localized metrics to capture the local specificities, where the metrics for local regions are described more precisely. Moreover, the combined descriptor shows better discriminativeness and obtains better verification accuracy on both LFW and YTF.

[1]It is noted that deep learning methods have achieved outstanding results on LFW and YTF with the image-unrestricted setting, where there are mainly two key differences: 1) outside data is not allowed for training under the image-restricted setting, and 2) only weak pairwise label information is exploited instead of the strong supervision.

TABLE I

COMPARISON OF MEAN VERIFICATION ACCURACY AND THE STANDARD ERROR (%) UNDER THE IMAGE-RESTRICTED SETTING OF THE LFW DATASET

Method	Accuracy
Sub-SML [50]	89.90 ± 0.38
VMRS [51]	91.10 ± 0.59
Fisher vector [47]	87.47 ± 1.49
CSML+SVM [35]	80.00 ± 0.37
CBFD [30]	87.23 ± 1.68
STFRD+PMML [11]	89.35 ± 0.50
PAF [48]	87.77 ± 0.51
APEM (LBP) [49]	81.97 ± 1.90
APEM (SIFT) [49]	81.88 ± 0.94
APEM (Fusion) [49]	84.08 ± 1.20
DNLML-ISA (SSIFT) [34]	86.17 ± 0.40
DNLML-ISA [34]	88.50 ± 0.40
DDML (SSIFT) [3]	87.83 ± 0.93
DDML (Combine) [3]	90.68 ± 1.41
CDBN [46]	86.88 ± 0.62
CDBN+Hand-crafted [46]	87.77 ± 0.62
LM3L [45]	89.57 ± 1.53
DTML-AE [17]	88.23 ± 0.45
DLML (SSIFT)	89.95 ± 1.05
DLML (Combine)	$\mathbf{92.22 \pm 1.51}$

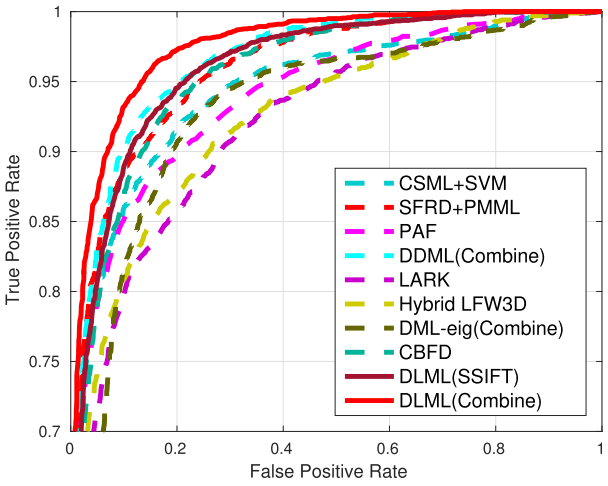


Fig. 6. Comparison of ROC curves under the image-restricted setting of the LFW dataset.

3) Comparison of Different Clustering Methods: Besides the proposed KAEs, other clustering algorithms can also be used in the proposed framework. We compared the proposed KAEs to the commonly used K-Means approach on the LFW dataset by first clustering features using K-Means and then learning the deep localized metrics. Table III shows that KAEs achieves better result than K-Means. The main reason is that KAEs performs quantization by learning K subspace projections rather than K centroids, which presents stronger descriptive power and robustness. How to develop more elaborate quantization methods to combine with our proposed framework to further improve the performance is an interesting future work.

4) Comparison of Different Weights: Generally, $\mu_k(\mathbf{x}_i, \mathbf{x}_j)$ aims to define the weight of the kth local subspace for the input pair $\mathbf{x}_i$ and $\mathbf{x}_j$, where the subspaces with smaller

TABLE II

COMPARISON OF MEAN VERIFICATION ACCURACY AND THE STANDARD ERROR (%) UNDER THE IMAGE-RESTRICTED SETTING OF THE YTF DATASET

Method	Accuracy
LBP [25]	75.90 ± 1.40
MBGS (LBP) [40]	76.40 ± 1.80
VSOF+OSS (Adaboost) [52]	79.70 ± 1.80
PHL+SILD (LBP) [53]	80.20 ± 1.30
STFRD+PMML [11]	79.48 ± 2.52
APEM (LBP) [49]	77.44 ± 1.46
APEM (SIFT) [49]	78.54 ± 1.42
APEM (Fusion) [49]	79.06 ± 1.51
DDML (LBP) [3]	81.26 ± 1.63
DDML (Combine) [3]	82.34 ± 1.47
LM3L [45]	81.30 ± 1.20
DLML (LBP)	83.35 ± 1.18
DLML (Combine)	$\mathbf{84.94 \pm 1.06}$

TABLE III

MEAN VERIFICATION ACCURACY (%) COMPARISON OF DIFFERENT CLUSTERING METHODS UNDER THE IMAGE-RESTRICTED SETTING OF THE LFW DATASET

Method	Accuracy
DLML (K-Means)	88.18
DLML (KAEs)	89.95

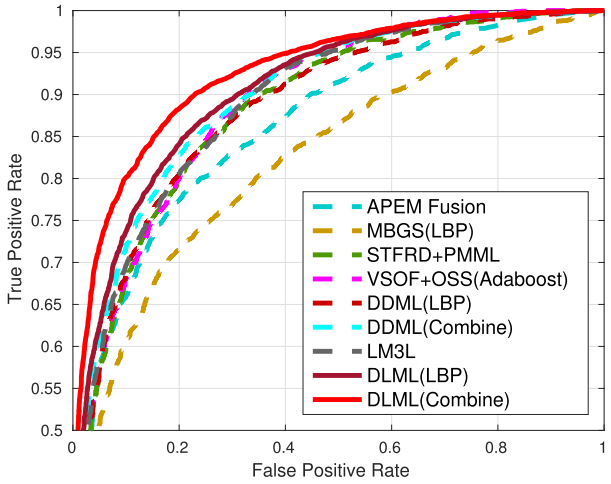


Fig. 7. Comparison of ROC curves under the image-restricted setting of the YTF dataset.

reconstruction loss would have higher weights. We apply the formulation in 4 for smoothness, where other formulations are also applicable. We tested another straightforward formulation by setting $s(\epsilon_{ik}, \epsilon_{jk}) = \frac{1}{\epsilon_{ik}+\epsilon_{jk}}$, and the experimental results on LFW and YTF are shown in Table IV. We observe that there is only a slight drop in accuracy.

5) Computational Time: Our hardware configuration comprises of a 2.8 G-Hz CPU and a 32G RAM, and we have not applied GPU for acceleration. We tested the training time on the LFW dataset with $K = 4$, and it took 861.9s to train the entire networks, where the average training time for each network was 172.4s. We also evaluated the computational time for testing, and it took 0.2 second on the LFW dataset. For the widely used deep metric learning method of DDML [3], the training and test time are 33.8 and 0.1 seconds on LFW, respectively.

TABLE IV

MEAN VERIFICATION ACCURACY (%) COMPARISON OF DIFFERENT μ ON LFW AND YTF

Method	LFW	YTF
DLML (New)	89.07	82.58
DLML (Original)	89.95	83.35

Fig. 8. Examples of the VIPeR dataset. Each column represents the same person captured from two different viewpoints.

B. Person Re-Identification

Person re-identification is a conventional visual analysis task, where images of person captured from multiple non-overlapping cameras need to be recognized. The image pairs usually suffer from large variation of poses, illuminations, viewpoints and occlusions, which leads large intra-class differences.

We conducted the experiments of DLML on the VIPeR dataset [54]. It contains 632 subjects under two different viewpoints, which are cropped to 128×48 to remove the background information. The dataset has a large viewpoint variation and relatively low resolution, which makes it a challenging dataset in person re-identification. Fig. 8 shows some examples of the VIPeR dataset. As person re-identification suffers from more variations compared with the face recognition tasks, we fixed the number of subspaces as 5 by experience.

We followed two widely used evaluation protocols [54] to test the proposed DLML. In the first protocol, 632 pairs of images are randomly divided into 316 pairs for training and the remaining 316 pairs for testing. In the second protocol, only 100 images pairs are used as the training set and the other 532 pairs are used as the test set. We tested the matching rates under different ranks to evaluate the proposed method.

1) Feature Representation: We utilized the local maximal occurrence (LOMO) feature to describe each image. LOMO is a commonly used pedestrian descriptor which extracts the HSV and scale invariant local ternary pattern (SILBP) features for a high-level description. LOMO exploits the

TABLE V

COMPARISON OF MATCHING RATES (%) UNDER DIFFERENT RANKS WITH METRIC LEARNING APPROACHES ON THE VIPeR DATASET (P = 316)

Method	Rank-1	Rank-5	Rank-10	Rank-20
PCCA [55]	19.27	48.89	64.91	80.28
rPCCA [5]	22.00	54.80	71.00	85.30
KISSME [23]	19.60	35.00	62.20	77.00
LMNN [15]	34.97	65.01	78.13	88.76
ITML [14]	24.64	35.93	48.76	60.08
kLFDA [5]	32.30	65.80	79.70	90.90
LOMO+XQDA [29]	40.00	68.13	80.51	91.08
DLML (LOMO)	**43.16**	**71.66**	**86.73**	**94.85**

TABLE VI

COMPARISON OF MATCHING RATES (%) UNDER DIFFERENT RANKS WITH METRIC LEARNING APPROACHES ON THE VIPeR DATASET (P = 100)

Method	Rank-1	Rank-5	Rank-10	Rank-20
PCCA [55]	9.27	24.89	37.43	52.89
rPCCA [5]	20.79	44.83	56.29	70.88
KISSME [23]	19.96	43.50	55.36	69.98
LMNN [15]	22.45	45.03	57.38	70.08
ITML [14]	19.34	36.89	46.25	59.07
kLFDA [5]	21.52	43.84	56.11	69.55
LOMO+XQDA [29]	23.18	44.98	58.01	71.23
DLML (LOMO)	**25.89**	**49.31**	**61.48**	**74.84**

TABLE VII

COMPARISON OF MATCHING RATES (%) UNDER DIFFERENT RANKS WITH PERSON RE-IDENTIFICATION APPROACHES ON THE VIPeR DATASET (P = 316)

Method	Rank-1	Rank-5	Rank-10	Rank-20
PRDC [57]	15.66	38.42	53.86	70.09
McMCML [6]	28.83	59.34	75.82	88.51
MidFilter [28]	29.11	52.34	65.95	78.80
LADF [58]	29.88	61.04	75.98	88.10
SalMatch [59]	30.16	52.31	75.31	86.71
kBiCov [60]	31.11	58.33	70.71	82.44
PolyMap [61]	36.80	70.40	83.70	91.70
SIR-CIR [56]	35.76	66.81	84.03	92.04
DLML (LOMO)	**43.16**	**71.66**	**86.73**	**94.85**

TABLE VIII

COMPARISON OF MATCHING RATES (%) UNDER DIFFERENT RANKS WITH PERSON RE-IDENTIFICATION APPROACHES ON THE VIPeR DATASET (P = 100)

Method	Rank-1	Rank-5	Rank-10	Rank-20
PRDC [57]	9.12	24.19	34.40	48.55
RPML [36]	10.90	26.70	37.70	51.60
MtMCML [6]	12.33	31.64	45.13	61.11
LADF [58]	12.90	30.30	42.70	58.00
PolyMap [61]	17.40	41.60	55.30	70.80
DLML (LOMO)	**25.89**	**49.31**	**61.48**	**74.84**

horizontal occurrence information and presents robustness to view point changes by applying the Retinex transformation and scale invariant local descriptors, which fits for the person re-identification tasks. Moreover, we reduced the dimension of the feature to 500 using WPCA, and reflected each training image horizontally for data augmentation. We applied the same network structure as the face recognition task.

2) Comparison With Metric Learning Methods: We compared the proposed DLML with existing metric learning methods, which included PCCA [55], regularized PCCA (rPCCA) [5], KISSME [23], large margin nearest neighbor classifier (LMNN) [15], Information theoretic metric learning (ITML) [14], kernel local Fisher discriminant analysis (kLFDA) [5] and cross-view quadratic discriminant analysis (XQDA) [29].

Table V and Table VI shows the experimental results of our DLML and the state-of-the-art metric learning methods on the VIPeR dataset. XQDA [29] achieves an outstanding performance on both settings of the VIPeR dataset, yet it only learns a single quadratic metric which may not be powerful enough to describe the complicated nonlinear relationship. The proposed DLML method learns multiple deep localized metrics which presents stronger discriminative power and obtains better performance on both P = 316 and P = 100 settings of the VIPeR dataset.

3) Comparison With Person Re-Identification Methods: We also compared the proposed DLML approach with commonly used person re-identification methods on the VIPeR dataset. Table VII and Table VIII show the experimental results under P = 316 and P = 100 settings respectively. We observe that our LDML achieves outstanding matching rates on both settings of the VIPeR dataset, which illustrates the effectiveness of the proposed approach. SIR-CIR [56] exploits the connection between single-image representation (SIR) and cross-image representation (CIR) by using a joint learning framework on CNN. Instead of studying both SIR and CIR, the proposed DLML mainly focuses on the CIR task to learn distance metrics for image pairs, exploiting multiple deep localized metrics, and obtains comparable results on the VIPeR dataset.

C. Scene Recognition

In this subsection, we evaluated the proposed DLML approach on the MIT Indoor-67 dataset [62], which is a popular indoor scene recognition dataset. The MIT Indoor-67 dataset consists of 16520 images of 67 indoor scenes, which has small inter-class variations and is challenging for classification. Fig 9 shows some examples from the MIT Indoor-67 dataset. We followed the standard evaluation protocol [62] by approximately using 80 images per class for training and 20 images for testing. We randomly selected equal number of positive pairs and negative pairs from the training set for each class to train the metric networks.

1) Feature Representation: We utilized the commonly used PlaceNet [63] feature as the representation of each indoor scene image. PlaceNet is a convolutional neural network which trains on the Places dataset with over 7 million pictures. Through the pre-trained network, we can obtain a 4096-dimensional feature for each image. Also, we utilized WPCA to reduce the feature dimension into 800 to remove the redundancy. We fixed $K = 8$ in this experiment, and set the structure of Auto-Encoders and fully connection networks

Fig. 9. Examples of "computer room", "livingroom" and "video store" from the MIT Indoor-67 dataset. Images from each row represent the same category, and the dataset suffers from small inter-class variations.

TABLE IX

COMPASION OF RECOGNITION RATE (%) WITH CONVENTIONAL SCENE RECOGNITION METHODS ON THE MIT INDOOR-67 DATASET

Method	Accuracy
ROI [62]	26.05
MMDL [64]	50.15
DSFL [65]	52, 24
BOP [66]	46.10
IFV [66]	60.77
IFV+BOP [66]	63.10
Mode-Seeking [68]	64.03
Mode-Seeking+IFV [68]	66.87
ISPR [67]	50.10
ISPR+IFV [67]	68.50
DLML (PlaceNet)	**83.47**

TABLE X

COMPARISON OF RECOGNITION RATE (%) WITH THE CNN METHODS ON THE MIT INDOOR-67 DATASET

Method	Accuracy
PlaceNet [63]	68.24
MOP-CNN [69]	68.90
HybridNet [63]	70.80
URDL+CNNaug [70]	71.90
DSFL+CNN [65]	76.23
MPP-FCR2 (7 scales) [71]	75.67
MPP+DSFL [71]	80.78
CFV (VGG-19) [72]	81.00
CS (VGG-19) [73]	82.24
DLML (PlaceNet)	**83.47**

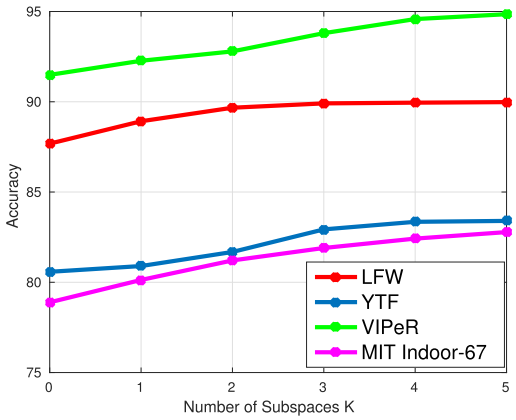


Fig. 10. Accuracy of DLML under different number of subspaces (from 0 to 5) on LFW, YTF, VIPeR (P = 316, Rank-20) and MIT Indoor-67.

as [800→700→600→700→800] and [800→700→600], respectively.

2) Comparison With Existing Methods: We first compared our DLML with conventional scene recognition methods. The compared approaches mainly include regions of interest (ROI) [62], max-margin multiple-instance dictionary learning (MMDL) [64], discriminative and shareable feature learning (DSFL) [65], improved Fisher vectors (IFV) [66] and important spatial pooling regions (ISPR) [67]. Table IX shows the experimental results of the proposed DLML and conventional scene recognition methods on the MIT Indoor-67 dataset. Our DLML outperforms these methods by at least 15% on accuracy. The conventional methods fail to extract effective features as deep learning is not applied to exploit the semantic information. The proposed DLML utilize the PlaceNet feature and learn deep localized metrics, which leads to strong discriminative power.

Then, we compared the state-of-the-art CNN approaches on the MIT Indoor-67 dataset, and Table X illustrates the experimental results. We observe that the proposed DLML method largely improves the performance of the PlaceNet feature by over 15%, and achieves the state-of-the-art performance on the MIT Indoor-67 dataset. As the inter-class variations are relatively small in this dataset, there are a number of local regions where subjects from different class are similar in representation. The proposed DLML approach locates the ambiguous local regions and enlarges the distances of negative pairs, which effectively relieve the problem of small inter-class distances and improve the distinctiveness of the learned feature.

D. Number of Subspaces

In our DLML, we learned $K + 1$ deep metrics in total, where K of them were local metrics and the other was a holistic metric. The localized metric learning will degenerate to single metric learning for $K = 0$, and become a combination of two holistic metrics for $K = 1$. In order to illustrate the effectiveness of localized metric learning, we tested the number of subspaces on all the databases including LFW, YTF, VIPeR and MIT Indoor-67, and Fig. 10 illustrates the performance of the proposed DLML under different number of subspaces. We observe that the accuracy increases at first with larger number of subspaces for all the datasets, and then becomes still when K is relatively large. The reason is that more localized information is exploited with larger K, yet it will be saturated when learning a large number of subspaces. In the training procedure of KAEs, we only train each autoencoder with the features belonging to it. When K is too large, all the features would be quantized to a few autoencoders, where the rest of autoencoders have no samples for training. In this situation, the subspaces is saturated and obtains the similar performance with the small number of K.

Moreover, the VIPeR and MIT Indoor-67 datasets suffer from more variations, so that the optimal K is larger.

E. Analysis

The above experiments suggest the following four key observations:

1) The proposed DLML learns discriminative metrics on multiple local subspaces, which obtains more fine-grained distance metrics over the input space. In many visual analysis tasks, the input set may suffer from small inter-class or large intra-class distances in local regions, and the localized metric learning emphasizes the confusing subspaces to improve the distinctiveness of the feature.
2) The proposed DLML applies deep neural networks to learn complicated distance metrics from the input set. Compared with conventional metric learning approaches which learns a linear Mahalanobis metric, our DLML presents stronger discriminative power by learning hierarchical nonlinear transformations.
3) The performance is improved with the increase of the number of the learned subspaces at first, and it will be saturated if the number is too large. Generally, more complicated datasets require a larger K for finely description.
4) Combining different feature descriptors lead to better recognition performance, as the combination provides more complete information of the input images.

V. Conclusion

In this paper, we have presented a deep localized metric learning (DLML) approach for visual recognition. Our DLML learns K local subspaces and one holistic subspace through the K-Auto-Encoders (KAEs) based clustering, and obtains deep localized metrics with deep neural networks. The proposed DLML achieves better or very competitive performance on three different visual analysis tasks including face recognition, person re-identification and scene recognition, which shows its effectiveness and wide applicability.

References

[1] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, 2014, pp. 1891–1898.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, vol. 1. 2015, pp. 1–12.

[3] J. Hu, J. Lu, and Y. P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. CVPR*, Jun. 2014, pp. 1875–1882.

[4] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.

[5] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. ECCV*, 2014, pp. 1–16.

[6] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[10] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. ICCV*, 2009, pp. 498–505.

[11] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. CVPR*, 2013, pp. 3554–3561.

[12] J. Bohné, Y. Ying, S. Gentric, and M. Pontil, "Large margin local metric learning," in *Proc. ECCV*, 2014, pp. 679–694.

[13] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *J. Mach. Learn. Res.*, vol. 6, pp. 937–965, Jun. 2005.

[14] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. ICML*, 2007, pp. 209–216.

[15] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[16] J. Lu, G. Wang, and P. Moulin, "Localized multifeature metric learning for image-set-based face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 529–540, Mar. 2016.

[17] J. Hu, J. Lu, Y. P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 5576–5588, Dec. 2016.

[18] I. W. Tsang, J. T. Kwok, C. Bay, and H. Kong, "Distance metric learning with kernels," in *Proc. ICANN*, 2003, pp. 126–129.

[19] D. Y. Yeung and H. Chang, "A kernel approach for semisupervised metric learning," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 141–149, Jan. 2007.

[20] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 1950–1962, Sep. 2015.

[21] V. E. Liong, J. Lu, Y. P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[22] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *Proc. CVPR*, 2015, pp. 1137–1145.

[23] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, 2012, pp. 2288–2295.

[24] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. ICCV*, 2013, pp. 2408–2415.

[25] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[26] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 898–903, 2012.

[27] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proc. ICCV*, 2013, pp. 3567–3574.

[28] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. CVPR*, 2014, pp. 144–151.

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015, pp. 2197–2206.

[30] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, 2014, pp. 1701–1708.

[32] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[33] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[34] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou, "Deep nonlinear metric learning with independent subspace analysis for face verification," in *Proc. ACM MM*, 2012, pp. 749–752.

[35] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. ACCV*, 2010, pp. 709–720.

[36] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. ECCV*, 2012, pp. 780–793.

[37] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. CVPR*, 2013, pp. 3318–3325.

[38] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. CVPR*, 2015, pp. 1846–1855.

[39] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[40] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, 2011, pp. 529–534.

[41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[42] S. Ul Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. BMVC*, 2012, pp. 1–12.

[43] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," *ECCVW*, Oct. 2008, pp. 1–14.

[44] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *Proc. ACCV*, 2009, pp. 88–97.

[45] J. Hu, J. Lu, J. Yuan, and Y. P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Proc. ACCV*, 2014, pp. 252–267.

[46] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. CVPR*, 2012, pp. 2518–2525.

[47] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. BMVC*, 2013, pp. 1–12.

[48] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. CVPR*, 2013, pp. 3539–3545.

[49] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. CVPR*, 2013, pp. 3499–3506.

[50] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *Proc. ICCV*, 2013, pp. 3208–3215.

[51] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *Proc. ICCV*, 2013, pp. 1960–1967.

[52] H. Méndez-Vázquez, Y. Martínez-Díaz, and Z. Chai, "Volume structured ordinal features with background similarity measure for video face recognition," in *Proc. ICB*, 2013, pp. 1–6.

[53] M. Kan, D. Xu, S. Shan, W. Li, and X. Chen, "Learning prototype hyperplanes for face verification in the wild," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3310–3316, Aug. 2013.

[54] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. PETS*, vol. 3. Oct. 2007, p. 5.

[55] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. CVPR*, 2012, pp. 2666–2672.

[56] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. CVPR*, 2016, pp. 1288–1296.

[57] W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[58] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. CVPR*, 2013, pp. 3610–3617.

[59] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. ICCV*, 2013, pp. 2528–2535.

[60] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image Vis. Comput.*, vol. 32, no. 6, pp. 379–390, 2014.

[61] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. CVPR*, 2015, pp. 1565–1573.

[62] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. CVPR*, 2009, pp. 413–420.

[63] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 487–495.

[64] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," in *Proc. ICML*, 2013, pp. 846–854.

[65] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," in *Proc. ECCV*, 2014, pp. 552–568.

[66] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. CVPR*, 2013, pp. 923–930.

[67] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Proc. CVPR*, 2014, pp. 3726–3733.

[68] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. NIPS*, 2013, pp. 494–502.

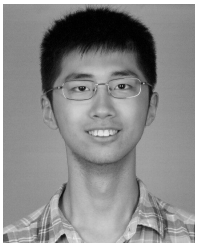
[69] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. ECCV*, 2014, pp. 392–407.

[70] B. Liu, J. Liu, J. Wang, and H. Lu, "Learning a representative and discriminative part model with deep convolutional features for scene recognition," in *Proc. ACCV*, 2014, pp. 643–658.

[71] D. Yoo, S. Park, J. Y. Lee, and I. S. Kweon. (2014). "Fisher kernel for deep neural activations." [Online]. Available: https://arxiv.org/abs/1412.1628

[72] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. CVPR*, 2015, pp. 3828–3836.

[73] G. S. Xie, X. Y. Zhang, S. Yan, and C. L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017.



Yueqi Duan received the B.E. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2010, where he is currently working toward the Ph.D. degree with the Department of Automation. His research interests include visual recognition, feature learning, and binary descriptor.

Jiwen Lu received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from Xian University of Technology, Xian, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. From 2011 to 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His research interests include computer vision, pattern recognition, and machine learning. He has authored or co-authored over 150 scientific papers in these areas, 42 of which were IEEE transactions papers. He was a recipient of the National 1000 Young Talents Plan Program in 2015. He is the Workshop Chair/Special Session Chair/Area Chair for over ten international conferences. He serves as an Associate Editor of *Pattern Recognition Letters*, *Neurocomputing*, and IEEE ACCESS; a Managing Guest Editor of the *Pattern Recognition and Image and Vision Computing*; a Guest Editor of *Computer Vision and Image Understanding*, and an elected member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society.

Jianjiang Feng received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007, respectively. From 2008 to 2009, he was a Post-Doctoral Researcher with the PRIP Laboratory, Michigan State University. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing. His research interests include fingerprint recognition and computer vision. He is an Associate Editor of *Image and Vision Computing*.

Jie Zhou received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. He has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as the IEEE PAMI, TIP, and CVPR. His research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of *International Journal of Robotics and Automation* and two other journals.