Hierarchical Invariant Feature Learning with Marginalization for Person Re-Identification

Rahul Rama Varior, *Student Member, IEEE,* Gang Wang, *Member, IEEE*

arXiv:1511.09150v1 [cs.CV] 30 Nov 2015

***Abstract*—This paper addresses the problem of matching pedestrians across multiple camera views, known as person re-identification. Variations in lighting conditions, environment and pose changes across camera views make re-identification a challenging problem. Previous methods address these challenges by designing specific features or by learning a distance function. We propose a hierarchical feature learning framework that learns invariant representations from labeled image pairs. A mapping is learned such that the extracted features are invariant for images belonging to same individual across views. To learn robust representations and to achieve better generalization to unseen data, the system has to be trained with a large amount of data. Critically, most of the person re-identification datasets are small. Manually augmenting the dataset by partial corruption of input data introduces additional computational burden as it requires several training epochs to converge. We propose a hierarchical network which incorporates a marginalization technique that can reap the benefits of training on large datasets without explicit augmentation. We compare our approach with several baseline algorithms as well as popular linear and non-linear metric learning algorithms and demonstrate improved performance on challenging publicly available datasets, VIPeR, CUHK01, CAVIAR4REID and iLIDS. Our approach also achieves the state-of-the-art results on these datasets.**

***Index Terms*—Person re-identification, Marginalization, Invariant features, Hierarchical feature learning, Metric Learning.**

I. Introduction

Matching pedestrians across multiple non-overlapping camera views is a research problem that has gained a lot of interest in recent years. It has become an integral part of surveillance, human tracking and human retrieval. Figure 1 shows some example images of pedestrians captured from such non-overlapping camera views. The objective of this problem is to identify the matching image(s) from a set of gallery images for a given probe image, thereby saving labor intensive work of searching through the entire set of images for identifying the correct match. Main approaches that address this problem concentrate on developing a feature representation [1], [2], [3] and [4] for the images or learning a distance metric [5], [6], [7] [8] and [9] so that images belonging to the same person are closer to each other in a feature space.

Despite the efforts of several researchers over the years, person re-identification still remains a challenging problem. In this paper, we address two major challenges in person re-identification. First, environmental conditions such as varying illumination, backgrounds and changes in pose across camera views, resolution and image quality cause significant change in appearance for images of same individual across camera views. Previous methods [2], [14] address these challenges by designing features specific for each of these aspects. We propose a hierarchical network that can learn invariant representations from labeled image pairs across different views. Local features are first extracted from the images. Inspired by the success of kernel based methods in many computer vision problems [15] and [16], we use a non-linear kernel function to map the input features to a kernel space. Further, a linear mapping is learned to extract patterns in this kernel space so that for corresponding parts of matching images, the extracted patterns are *close* to each other. To learn the linear mapping, we take advantage of labeled image pairs and enforce the invariance constraint for the mapped kernel features across different views.

Fig. 1: Some images taken from standard person re-identification datasets such as VIPeR [10], CUHK01 [11], iLIDS [12] and CAVIAR4REID [13]. The objective of person re-identification is to match a given probe image to a set of gallery images. **Best viewed in color**

The second problem addressed in this paper is the lack of labeled training data. For a good generalization to unseen data, complex systems must be trained with a large amount of data. Most of the existing person re-identification datasets are small. Augmenting the data explicitly by partial corruption based on a specific corruption distribution can be adopted. But augmenting the data makes the system computationally expensive as it requires several epochs over the entire training

R. Rama Varior and G. Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798. (e-mail: {rahul004,wanggang}@ntu.edu.sg.)

set to converge. Motivated by the technique used in [17], we propose a novel hierarchical network which incorporates the marginalization technique that can reap the benefits of training on augmented dataset without explicitly adding more data. The technique is to marginalize out the noise of the corrupted inputs. The proposed framework is related to the kernel Local Fisher Discriminant Analysis (kLFDA) framework in [18] which can be possibly extended to a two layer structure. But a potential difficulty is applying the marginalization technique over the LFDA framework. Therefore, we adapt the SVM Metric Learning (SVMML) in [6] by incorporating the marginalization technique. This leads to an improved metric learning algorithm which suits better for smaller person re-identification datasets.

To the best of our knowledge, this is the first work that proposes a hierarchical invariant feature learning framework coupled with the SVM metric learning framework simultaneously incorporating marginalization technique to address the problem of lack of training data for person re-identification. In short, main contributions of this paper are;

- We propose a new learning framework that captures invariant information from local patches of image pairs for person re-identification. A linear transformation is learned in the kernel space by enforcing invariance constraint for local patch features extracted from labeled image pairs.
- We propose a novel feature learning framework that incorporates marginalization that can reap the benefits of training on *infinite* data without explicitly augmenting them. At the second layer of the proposed hierarchical network, we adapt the SVMML by incorporating marginalization. Marginalization helps in achieving better generalization over unseen data.
- We show that the proposed method outperforms several baselines and popular kernel based algorithms for VIPeR [10], CUHK01[11], CAVIAR4REID [13] and iLIDS [12] re-identification datasets. Our approach also achieves state-of-the-art results on these datasets.

The rest of the paper is organized as follows. Section II gives a brief overview of the related works on metric learning and feature learning for person re-identification as well as some of the recent deep learning frameworks. Detailed description of our approach with formulations are given in Section III. In Section IV, we show our experimental results. We analyze different baselines in Section V and Section VI concludes this paper.

II. Related Work

Existing works for person re-identification focus on several aspects of the problem such as developing a feature representation [19], [2], [3], [4], [20], metric learning [6], [5], [8], [7] for distance computation and learning mid-level representations [21], [1], [22]. [23], [24] focus on finding the salient patches and rank the matching images using rank SVM. Majority of the works [2], [14], [24] focus on developing feature representation based on texture, color, shape, regions and interest points. Below we give some of the related works in Metric Learning and Feature Learning for person re-identification

A. Metric Learning

Some of the prominent metric learning algorithms proposed for person re-identification are [11], [25], [6], [7], [8] etc. Regularized LFDA was proposed for person re-identification in [7]. The objective is to maximize the inter-class separability and to minimize the within class variance. To address the non-linearities in feature space, it was proposed to use kernel based dimensionality reduction techniques in [26]. SVM Metric learning was proposed in [6], and the idea is to learn a decision boundary that is adaptive to the data samples. In [18], several kernel-based metric learning methods for person re-identification were evaluated and kernel based LFDA was found to be performing the best for several re-identification datasets. Deep metric learning architectures has also been proposed for person re-identification in [5] and [1]. In [5], data augmentation is employed for improved performance. Deep architectures have to be trained with a large amount of data for better generalization to unseen data. However, none of the above metric learning algorithms address the problem of lack of labeled training data in re-identification datasets. Hence, we propose a marginalized metric learning framework based on SVMML objective function instead of techniques such as artificial augmentation of the dataset.

B. Feature Learning

In [1] a deep architecture was proposed to learn filter pairs that can handle photometric and geometric transformations. It consists of a single convolutional layer with max-pooling and a patch matching layer that multiplies the feature responses of the convolutional layer at different horizontal offsets. [5] also contains a feature learning framework based on Convolutional Neural Networks (CNN). The similarity of the features from the CNN is measured using a Cosine similarity function. A recent work [27] also proposes a CNN based deep architecture for person re-identification. In addition to convolution and max-pooling, a Cross-Input Neighborhood Difference is employed to compute relationships between images from two views. In another related work [22], mid-level filters were learned from patch clusters to achieve cross view invariance. But variations in illumination, viewing angle and other environmental variables affect the cross view invariance of the features as well as demand non-linear mappings for the features. Therefore, we propose an invariant feature learning framework in kernel space which can be seen as learning a flexible non-linear transformation in the original feature space [28]. This can be very effective for overcoming the non-linearities due to the aforementioned challenges. In addition to that, deep architectures required enormous amount of data to learn robust representations. Our approach also has the advantage of training on large amount of data without explicit augmentation as we employ marginalization.

C. Deep Learning

Some of the prominent articles in deep learning [29], [30], [31] suggest that multiple layers with non-linear transformations can be efficient in directly modeling complex functions

mapping the input to output. Several successful algorithms have been proposed [32], [33], [34], [29] to train such large networks such as deep belief networks and stacked auto-encoders. In-order to achieve a better generalization, such systems need to be trained with a large amount of data. Artificial augmentation of the data by partial corruption was proposed in [35]. But augmented data creates additional computational burden. To circumvent the complexity due to augmented data, marginalized Denoising Auto-Encoder (**mDAE**) was proposed in [17], [36]. This is a variant of the traditional denoising auto-encoder [30]. The key technique is to marginalize out the noise by adding a regularization term and thereby reaping the benefits of training on *infinite* data without explicitly corrupting the original data. Inspired by this work, we propose a novel marginalized invariant feature learning framework as well as marginalized metric learning framework based on the **SVMML** formulation.

III. Approach

The proposed framework has a two-layer structure. In the first layer, the input features are first mapped into a kernel space. A linear mapping is learned to extract stable structures and patterns from the exemplar responses. Due to environmental conditions such as varying illumination, backgrounds and changes in pose across camera views, successful re-identification requires invariant features from different views. Therefore, the extracted patterns must be *close* to each other for corresponding parts of matching image pairs. This is achieved by enforcing an invariance constraint to the learned features. We do not extract any discriminative information by using negative pairs at this stage as the discriminative capability of local stripes may be insufficient. Additionally, since most of the person re-identification datasets are small, labeled data is scarce. Therefore, the overall objective function is also adapted by the marginalization technique which further boosts the generalization capability. Further, the image representation is obtained by concatenating the mapped local features of that image. In the second layer, the features of the whole image is fed into a metric learning framework. We adapt the SVMML framework by incorporating marginalization so that with the limited amount of training data available, better performance can be achieved. The visualization of the process pipeline is shown in figure 2. Below, we explain our approach in detail.

A. Features

Each image is divided equally into 6 non-overlapping horizontal stripes as in [37], [8] and [18]. For each of these horizontal stripes, following features were extracted.

1) LBP: Texture patterns were captured by Local Binary Pattern (LBP) [38] histograms computed with 8-neighbors at a radius 1 and 16-neighbors at a radius 2. The dimensions of the representations are 59 and 243 for 8-neighbors and 16-neighbors respectively.

2) Color Histograms: Color information was captured by computing 16 bin histograms for each of the RGB, HS and YUV color channels respectively. The total dimension of the color representation will be 128.

An $l1$ normalization is performed for each of these channels individually and concatenated to form the final feature representation. This leads to a 430 dimensional representation for each of the horizontal stripes. Computing the histogram over horizontal stripes can give the advantage of translational invariance which is significant in person re-identification problems due to the pose changes. For a fair comparison, these features were used for all the performance comparison against the baseline methods as well as the metric learning methods. Once the local features (horizontal stripes) are obtained, these features are projected into a kernel space by using them as anchor points (exemplars) and a linear transformation is learned by enforcing an invariance constraint.

B. Learning invariant features

The objective of learning a linear transformation in the kernel space is to capture stable structures and patterns which are invariant across views. Let z_{ϕ_i} and z_{ψ_j} be the mapped responses for the corresponding stripes of two matching images in the probe and gallery respectively. They can be mathematically expressed as

$$z_{\phi_i} = f_\theta(\phi_i) = W^{(1)}\phi_i + b^{(1)} \tag{1}$$

$$z_{\psi_j} = f_\pi(\psi_j) = W^{(3)}\psi_j + b^{(3)} \tag{2}$$

where $\phi_i \in \mathbb{R}^D$ denotes the exemplar response vector of one of the stripes of a probe image and $\psi_j \in \mathbb{R}^D$ denotes the exemplar response vector of the corresponding matching stripe of the matching image pair in the gallery. $W^{(1)}$, $b^{(1)}$, $W^{(3)}$ and $b^{(3)}$ are the learned linear transformation parameters. Since the probe and gallery images are from different sources, we learn separate transformations for each of them as in [1], [39]. The objective of the mapping is to ensure z_{ϕ_i} and z_{ψ_j} to be *close* to each other. Thus the overall loss can be formulated as

$$\begin{aligned} l(\phi_i, f_\theta(\phi_i), \psi_j, f_\pi(\psi_j)) &= \left\| \phi_i - (W^{(2)} z_{\phi_i} + b^{(2)}) \right\|^2 \\ &+ \left\| \psi_j - (W^{(4)} z_{\psi_j} + b^{(4)}) \right\|^2 \\ &+ \left\| z_{\phi_i} - z_{\psi_j} \right\|_2^2 \end{aligned} \tag{3}$$

Here, $W^{(1)}, b^{(1)}, W^{(2)}$ and $b^{(2)}$ are the weight and bias parameters for the transformation and reconstruction steps for probe exemplar response vector and $W^{(3)}, b^{(3)}, W^{(4)}$ and $b^{(4)}$ are the weight and bias parameters for the transformation and reconstruction steps for gallery exemplar response vector. The first and second terms on the right hand side (RHS) of (3) are the reconstruction terms for the inputs. These terms are essential to avoid learning trivial solutions for the weight and bias parameters. $f_\theta(\phi_i)$ and $f_\pi(\psi_j)$ denotes the mappings shown in equation (1) and (2) respectively. The objective function was split for the probe and gallery and an alternative optimization scheme was adopted since the invariance term contains parameters from two sources. The final objective

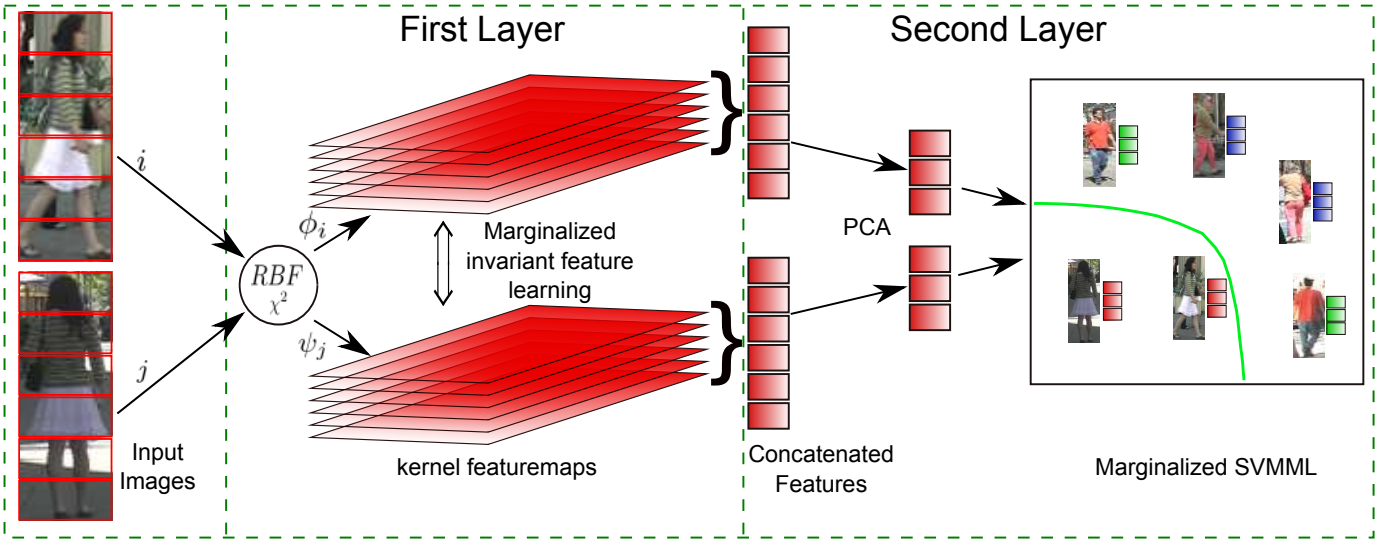

Fig. 2: The process pipeline of the proposed framework. Images are divided into horizontal stripes and features are extracted. By using the D exemplars the RBF-χ^2 kernel mapping is performed. Further these features are fed into a marginalized invariant feature learning framework to extract invariant features. After the feature extraction, the representation of individual stripes are concatenated to form the whole image representation and a PCA is used to reduce the dimensionality of the image representation. Further, marginalized SVMML formulation is used to learn the metric and to rank the matches. **Best viewed in color**

functions to be solved for the probe and gallery training data can be written as

$$\begin{aligned} l(\phi_i, f_\theta(\phi_i)) &= \left\| \phi_i - (W^{(2)} z_{\phi_i} + b^{(2)}) \right\|^2 \\ &+ \left\| z_{\phi_i} - z_{\psi_j} \right\|_2^2 \end{aligned} \tag{4}$$

$$\begin{aligned} l(\psi_i, f_\pi(\psi_i)) &= \left\| \psi_i - (W^{(4)} z_{\psi_i} + b^{(4)}) \right\|^2 \\ &+ \left\| z_{\phi_i} - z_{\psi_j} \right\|_2^2 \end{aligned} \tag{5}$$

C. Marginalization

To learn a robust representation and to have good generalization capability, the system needs to be trained with a large amount of data. Since most of the person re-identification datasets are small, the learned transformations will most likely overfit and may not have good generalization capability to unseen data. The concept of denoising auto-encoder (**DAE**) [35] was introduced to learn robust representation from limited training data. In a DAE, the training data is artificially corrupted based on a specific corruption distribution to create more training data. The corrupted data is sampled from the conditional distribution $p(\tilde{\phi}|\phi)$ and the commonly used corruption distribution is the additive Gaussian distribution. Let $\tilde{\phi}_i$ be the corrupted copy of ϕ_i. The process of explicitly corrupting the data ϕ_i creates multiple copies $\tilde{\phi}_i^1, ..., \tilde{\phi}_i^m$ of the data and the total amount of data becomes m fold. For a large m, the computational complexity increases to a large extent as it requires several epochs over the entire training set to converge.

Marginalized Denoising Auto-Encoder (**mDAE**) [17] was proposed to address the additional computational complexity due to augmented data. The technique is to marginalize out the corruption during auto-encoder training. In the proposed framework, the key difference for our objective function is that, in addition to the auto-encoder loss, an invariance constraint is enforced between the learned representations to achieve cross-view invariance. Therefore, we adapt the mDAE by incorporating the invariance term in our framework. Besides, the proposed framework consists of two parallel networks paired by the invariance term. Therefore it requires an alternate optimization scheme. Below, we present the derivation for one of the parallel networks and for the other network, it can be worked out similarly.

As $m \to \infty$, the average loss over the entire corpus along with the corrupted data becomes the expectation of the loss. Therefore optimizing the objective for *infinite* training data is equivalent to optimizing the below objective function.

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{p(\tilde{\phi}_i|\phi_i)} \left[l(\phi_i, f_\theta(\tilde{\phi}_i)) \right] \tag{6}$$

where $\tilde{\phi}_i$ is the corrupted version of the i^{th} exemplar. $l(.)$ denotes the proposed loss function. The loss function is approximated by its Taylor expansion at the mean of the corruption, $\mu_\phi = \mathrm{E}_{p(\tilde{\phi}|\phi)}[\tilde{\phi}]$. The proposed loss function is a sum of the Auto-Encoder loss and Invariance loss. The Taylor expansion of each of these losses can be considered independently and the following derivation can be applied to both of them.

For simplicity of notation, here $\phi \in \mathbb{R}^D$ denotes one of the data points in a training set of $\Phi = \{\phi_1, ..., \phi_n\}$ and $\tilde{\phi}$

denotes its corrupted version. The Taylor expansion yields the following learning objective.

$$\begin{aligned} l(\phi, f_\theta(\tilde{\phi})) &\approx l(\phi, f_\theta(\mu_\phi)) \\ &+ (\tilde{\phi} - \mu_\phi)^T \bigtriangledown_{\tilde{\phi}} l \\ &+ \frac{1}{2}(\tilde{\phi} - \mu_\phi)^T \bigtriangledown^2_{\tilde{\phi}} l(\tilde{\phi} - \mu_\phi) \end{aligned} \tag{7}$$

Taking the expectation on both sides for (7),

$$\begin{aligned} \mathrm{E}_{p(\tilde{\phi}|\phi)}\left[l(\phi, f_\theta(\tilde{\phi}))\right] &\approx l(\phi, f_\theta(\mu_\phi)) \\ &+\mathrm{E}_{p(\tilde{\phi}|\phi)}\left[(\tilde{\phi} - \mu_\phi)^T\right] \bigtriangledown_{\tilde{\phi}} l \\ +\frac{1}{2} tr\left(\mathrm{E}\left[(\tilde{\phi} - \mu_\phi)(\tilde{\phi} - \mu_\phi)^T\right] \bigtriangledown^2_{\tilde{\phi}} l\right) \end{aligned} \tag{8}$$

The second term on the RHS of (8) vanishes since the expectation of the left hand side (LHS) yields $\mathrm{E}_{p(\tilde{\phi}|\phi)}[\tilde{\phi}] = \mu_\phi$. Therefore, Taylor expansion at the mean μ_ϕ is critical. Further simplification results in the following objective.

$$\mathrm{E}\left[l(\phi, f_\theta(\tilde{\phi}))\right] \approx l(\phi, f_\theta(\mu_\phi)) + \frac{1}{2} tr\left(\Sigma_\phi \bigtriangledown^2_{\tilde{\phi}} l\right) \tag{9}$$

where $\Sigma_\phi = \mathrm{E}\left[(\tilde{\phi} - \mu_\phi)(\tilde{\phi} - \mu_\phi)^T\right]$ is the variance of the corruption distribution and $\bigtriangledown^2_{\tilde{\phi}} l$ is the Hessian of $l(.)$ with respect to $\tilde{\phi}$. Σ_ϕ is a diagonal matrix as the assumption is that the corruption is applied to each dimension independently. Therefore, computing the Hessian for higher dimensional data simplifies to computing only its diagonal elements. The d^{th} dimension of the Hessian matrix's diagonal can be obtained by the straightforward application of chain rule as shown below.

$$\begin{aligned} \frac{\partial^2 l}{\partial(\tilde{\phi}(d))^2} &= \left(\frac{\partial z}{\partial \tilde{\phi}(d)}\right)^T \frac{\partial^2 l}{\partial z^2} \left(\frac{\partial z}{\partial \tilde{\phi}(d)}\right) \\ &+ \left(\frac{\partial l}{\partial z}\right)^T \frac{\partial^2 z}{\partial(\tilde{\phi}(d))^2} \end{aligned} \tag{10}$$

where $\tilde{\phi}(d)$ represents the d^{th} dimension of $\tilde{\phi}$ and z is the latent representation. The corruption is applied to each dimension of ϕ independently, hence the second derivative with respect to each dimension of ϕ.

Following [40], the last term in (10) can be dropped and the Hessian can be approximated as

$$\frac{\partial^2 l}{\partial(\tilde{\phi}(d))^2} \approx \sum_{h=1}^{D_h} \frac{\partial^2 l}{\partial z_h^2} \left(\frac{\partial z_h}{\partial \tilde{\phi}(d)}\right)^2 \tag{11}$$

Substituting (11) in (9) yields the objective function below.

$$l(\phi, f_\theta(\mu_\phi)) + \frac{1}{2} \sum_{d=1}^{D} \sigma^2_{\phi(d)} \sum_{h=1}^{D_h} \frac{\partial^2 l}{\partial z_h^2} \left(\frac{\partial z_h}{\partial \tilde{\phi}(d)}\right)^2 \tag{12}$$

where D and D_h are the input and hidden layer dimensions respectively. z_h denotes the h^{th} dimension of the latent representation of the input. The value of $\sigma^2_{\phi(d)}$ for additive Gaussian is σ_d^2 and the mean μ_ϕ is ϕ.

Further, we need to compute the second term on the RHS of (12) for the proposed loss function. Since the proposed loss function consists of the Auto-Encoder and Invariance terms, it involves computing $\frac{\partial^2 l}{\partial(\tilde{\phi}(d))^2}$ for both the Auto-encoder term as well as the Invariance term. Details are given below.

The final objective function is obtained by adapting the objective function in (4) and (5) according to the marginalization technique. Let R^p_θ be the marginalization term obtained by applying (11) over (4) for the probe data and R^g_π be the corresponding marginalization term for the gallery data. The exemplar responses ϕ and ψ respectively for the probe and gallery input features are fed into two separate networks paired by the invariance term. The objective function for the probe ($l^{(p)}$) can be obtained by substituting R^p_θ in (12).

$$\begin{aligned} l^{(p)} &= l(\phi, f_\theta(\phi)) + \frac{1}{2} \sum_{d=1}^{D} \sigma_d^2 \mathrm{R}^p_\theta \\ &= l(\phi, f_\theta(\phi)) + \sum_{d=1}^{D} \sigma_d^2 \left(\sum_h (w^{(1)}_{hd})^2\right) \\ &+ \sum_{d=1}^{D} \sigma_d^2 \sum_h \left(\left(\sum_d (w^{(2)}_{hd})^2\right) (w^{(1)}_{hd})^2\right) \end{aligned} \tag{13}$$

Similarly, the objective function for gallery ($l^{(g)}$) is as shown below.

$$\begin{aligned} l^{(g)} &= l(\psi, f_\pi(\psi)) + \frac{1}{2} \sum_{d=1}^{D} \sigma_d^2 \mathrm{R}^g_\pi \\ &= l(\psi, f_\pi(\psi)) + \sum_{d=1}^{D} \sigma_d^2 \left(\sum_h (w^{(3)}_{hd})^2\right) \\ &+ \sum_{d=1}^{D} \sigma_d^2 \sum_h \left(\left(\sum_d (w^{(4)}_{hd})^2\right) (w^{(3)}_{hd})^2\right) \end{aligned} \tag{14}$$

where $w^{(1)}_{hd}$ corresponds to an element in the matrix $W^{(1)}$ and similarly for the other matrices.

The last term in (13) and (14) can be obtained from the marginalized denoising auto-encoder cost function $l(\phi, f_\theta(\tilde{\phi}))$ as shown in [17]. The marginalization penalty for the invariance term can be derived by taking the second derivative of the invariance loss function with respect to the input ϕ for each of its dimensions d. Let l_{inv} be the invariance term. The derivation with respect to a data point is shown below.

$$\begin{aligned} l_{inv} &= \left\| z_{\phi_i} - z_{\psi_j} \right\|_2^2 \\ &= \sum_h \left(\sum_d (w^{(1)}_{hd} \phi_i(d)) + b^{(1)}_h - \sum_d (w^{(3)}_{hd} \psi_j(d)) + b^{(3)}_h\right)^2 \end{aligned} \tag{15}$$

$$\frac{\partial l_{inv}}{\partial \tilde{\phi}_i(d)} = 2\sum_h \left(\sum_d (w_{hd}^{(1)} \phi_i(d)) + b_h^{(1)} - \sum_d (w_{hd}^{(3)} \psi_j(d)) + b_h^{(3)} \right) w_{hd}^{(1)} \tag{16}$$

$$\frac{\partial^2 l_{inv}}{\partial (\tilde{\phi}_i(d))^2} = 2\left(\sum_h (w_{hd}^{(1)})^2 \right) \tag{17}$$

The derivation of $\left(\frac{\partial^2 l_{inv}}{\partial (\tilde{\phi}_i(d))^2} \right)$ can be done as shown above which will result in the second term on the RHS of (14).

$$\frac{\partial^2 l_{inv}}{\partial (\tilde{\psi}(d))^2} = 2\left(\sum_h (w_{hd}^{(3)})^2 \right) \tag{18}$$

Further, the weight decay term, $\lambda \left\| W^{(i)} \right\|_2^2$ for $\{i = 1, \cdots, 4\}$ is also added to the respective objective functions, (13) and (14) with a penalty λ. Hence the final objective function can be given as

For Probe

$$\begin{aligned} l^{(p)} &= \left\| \phi_i - (W^{(2)} z_{\phi_i} + b^{(2)}) \right\|^2 + \left\| z_{\phi_i} - z_{\psi_j} \right\|_2^2 \\ &+ \sum_{d=1}^{D} \sigma_d^2 \left(\sum_h (w_{hd}^{(1)})^2 \right) \\ &+ \sum_{d=1}^{D} \sigma_d^2 \left(\sum_d (w_d^{(2)})^2 \right) (w_{hd}^{(1)})^2 \\ &+ \lambda \left\| W^{(1)} \right\|_2^2 + \lambda \left\| W^{(2)} \right\|_2^2 \end{aligned} \tag{19}$$

For Gallery

$$\begin{aligned} l^{(g)} &= \left\| \psi_i - (W^{(4)} z_{\psi_i} + b^{(4)}) \right\|^2 + \left\| z_{\phi_i} - z_{\psi_j} \right\|_2^2 \\ &+ \sum_{d=1}^{D} \sigma_d^2 \left(\sum_h (w_{hd}^{(3)})^2 \right) \\ &+ \sum_{d=1}^{D} \sigma_d^2 \left(\sum_d (w_d^{(4)})^2 \right) (w_{hd}^{(3)})^2 \\ &+ \lambda \left\| W^{(3)} \right\|_2^2 + \lambda \left\| W^{(4)} \right\|_2^2 \end{aligned} \tag{20}$$

D. Metric Learning

The second layer of the proposed system is a marginalized Metric Learning framework based on the SVM Metric Learning (SVMML). The objective of SVMML is to compute a decision boundary which is locally adaptive to the data samples. In our framework, we adapt the traditional SVMML by incorporating marginalization so that the benefits of training on large amount of data can be achieved. The objective function in equations (19) and (20) are solved alternatively and the parameters for the probe and gallery are learned. By using equations (1) and (2), the probe and gallery exemplars are mapped into an invariant feature space. Once z_{ϕ_i} and z_{ψ_i} are obtained, the global image representation is obtained by concatenating the mapped exemplar responses of the 6 horizontal stripes in the image. The global image representation is projected into a lower dimensional space by PCA. Let D_{ml} be the dimension of the projected space. These act as inputs to the proposed framework.

Let $k_i \in \mathbb{R}^{D_{ml}}$ be the global image representation of a probe image and $k_{i'} \in \mathbb{R}^{D_{ml}}$ be an image representation from the gallery set. The second order decision function as given in [6] can be written as

$$\begin{aligned} f(k_i, k_{i'}) &= \frac{1}{2} k_i^T A k_i + \frac{1}{2} k_{i'}^T A k_{i'} + k_i^T B k_{i'} \\ &+ c^T (k_i + k_{i'}) + b \end{aligned} \tag{21}$$

where A and B are real, symmetric, positive semi definite (PSD) and negative semi definite (NSD) matrices respectively. c is a d - dimensional vector and b is the bias term. In practice, the authors of [6] apply a $log - exp$ transformation to the loss function in equation (21) and also omit the term $c^T(k_i + k_{i'})$ in their implementation.

Therefore, the final objective function of SVMML is

$$l_{ml} = g(f(k_i, k_{i'})) = log(e^{f(k_i, k_{i'})} + 1) \tag{22}$$

where $g(.)$ denotes the $log - exp$ transformation.

Eventhough k_i and $k_{i'}$ are mapped responses from two sources, the invariance term projects them into the same feature space. Hence the marginalization can be directly applied to the loss function in equation (22) without having two separate objective functions for probe and gallery data.

Similar to the derivation of the marginalization term for the invariance loss in equation (17), the second derivative of l_{ml} with respect to each dimension of k_i has to be computed. Below, we show the derivation w.r.t one training example. This can be generalized to the entire training dataset and the final loss function for the metric learning framework can be obtained.

$$\frac{\partial l_{ml}}{\partial \tilde{k_i}(d)} = \frac{1}{2} \frac{\partial g}{\partial f} \odot (A k_i + A^T k_i + k_{i'}^T B^T) \tag{23}$$

Here, $\odot$ denotes the element-wise multiplication. The second derivative of l_{ml} w.r.t each dimension of k_i is given by,

$$\frac{\partial^2 l_{ml}}{\partial (\tilde{k_i}(d))^2} = \frac{1}{2} \frac{\partial g}{\partial f} \odot csum(A + A^T) + \frac{1}{2} \frac{\partial^2 g}{\partial f^2} \odot (A k_i + A^T k_i + k_{i'}^T B^T) \odot (A k_i + A^T k_i + k_{i'}^T B^T) \tag{24}$$

Here, $csum(.)$ denotes the column-wise sum of a matrix. The partial derivatives

$$\frac{\partial g}{\partial f} = \frac{1}{1 + e^{-f(k_i, k_{i'})}}$$

$$\frac{\partial^2 g}{\partial f^2} = \left(\frac{1}{1+e^{-f(k_i,k_{i'})}}\right) \times \left(\frac{1}{1+e^{f(k_i,k_{i'})}}\right)$$

The final loss function for the metric learning framework can be obtained by substituting equation (24) in equation (22)

$$l_{ml} = log(e^{f(k_i,k_{i'})}+1) + \frac{1}{2}\sum_{d=1}^{D_{ml}} \sigma_{k_i}^2 \frac{\partial^2 l_{ml}}{\partial(\tilde{k_i}(d))^2} \qquad (25)$$

D_{ml} denotes the dimension of the image representation. We do not use a low-rank projection for the metric learning framework. However, decomposing A and B to $A = MM^T$ and $B = -NN^T$ can be helpful in learning a PSD and NSD low-rank matrices respectively. Further Frobenius norm regularization for A and B were added to the objective function in equation (25).

E. Optimization

The objective functions in (13) and (14) are minimized alternatively for the parameters $W^{(i)}$ and $b^{(i)}$ in (4) and (5) respectively. The first network, as explained in (13) is optimized for κ iterations while keeping the parameters of (14) fixed and vice-versa. L-BFGS gradient based minimization is adopted for optimizing the cost function and the total number of iterations is kept as 300. The main parameters of the experiment were empirically determined by cross-validation as done in [18] on the VIPeR dataset and kept same for other datasets. They are, the dimension of the linearly projected space $D_h = 800$, $\lambda = 1 \times 10^{-7}$, $\sigma_d = 0.1$.

Objective function in equation (25) can be optimized by gradient projection algorithms. If a low-rank projection is required, optimization has to be done for M and N. Similar to the first layer, number of iterations is kept as 300. Main parameters are determined by cross-validation on the VIPeR dataset and kept same for others. They are, dimension of the global image representation $D_{ml} = 400$, $\sigma_{k_i} = 0.01$ and the frobenius regularization penalty , $\lambda_{ml} = \{1 \times 10^{-8}, 1 \times 10^{-7}\}$ for A and B respectively.

IV. Experiments

Our approach was evaluated on four challenging publicly available datasets which are characterized by ample variation in their environments, pose and illumination. For a fair comparison, we use the same features and experimental settings for all the baselines. The amount of supervision for each of the datasets is also kept the same and the results are reported as Cumulative Matching Characteristics (CMC), which shows the probability of identifying the correct match at different ranks.

The main baselines which can be considered as the variants of our approach are listed below.

1) **2 layer kLFDA** - To show that the proposed architecture with marginalization technique has better generalization capability over unseen data, we develop a baseline by extending the kLFDA [18] into two layers.
2) **no_Marg** - To prove that marginalization technique helps in achieving better performance, we compare the proposed approach to its variant without marginalization at both layers. We use SVMML at the second layer for a fair comparison. This makes the overall a system an autoencoder with invariance term coupled with the traditional SVMML metric learning framework
3) **no_Inv** - To show that the proposed objective function with invariance is crucial for person re-identification performance, we conduct experiments without enforcing the invariance constraint in our objective function. Removing the invariance term makes the first layer objective function a simple marginalized Denoising Auto-encoder. At the second layer, the proposed marginalized SVMML is used to learn the metric.
4) **marg_SVMML** - To prove that the proposed marginalized metric learning framework achieves better performance, we couple the proposed marginalized framework in the first layer with the traditional SVMML.
5) **marg_kLFDA** - This framework is similar to the above. Instead of using the traditional SVMML framework at the second layer, we use kLFDA at the second layer.

A summary of the above baselines is given in table I for easy reference. In addition to the above baselines, our approach was also compared with popular linear and kernel based metric learning algorithms. Finally, we compare the proposed algorithm with the state-of-the-art algorithms in all datasets.

A. Evaluation Methodology

All the experiments were conducted in the single-shot setting as done in [18], [41], i.e. for each query image from the probe set is compared to one image from the gallery set. For all the datasets, the images are divided into 6 non-overlapping horizontal stripes and the features are extracted as explained in section III-A. Since the datasets are small, all the horizontal stripes from the training images are used as anchor points (exemplars) for the RBF-χ^2 kernel mapping. Further, the a transformation is learned by solving the objective functions in equations (13) and (14) with the invariance constraint between the exemplar pairs in the kernel space. The linear transformation projects the mapped kernel responses to an 800 dimensional space. To obtain the global image representation, the mapped responses of the 6 horizontal stripes are concatenated. PCA is employed to reduce the 4800 dimensional responses to a 400 dimensional representation. Finally, the global image representation is fed into the marginalized metric learning framework to learn the metric. For all the comparison with the baselines as well as popular metric learning algorithms, the features as mentioned in section III-A were used.

B. Datasets

1) VIPeR Dataset: VIPeR dataset [10] is one of the most challenging datasets for person re-identification. It contains 1264 images of 632 pedestrians captured from two different camera views. Image resolution is 128×48 and significant variations in illumination and pose can be observed in this dataset. In our experiments and for the comparisons in the tables, 316 image pairs were used as training images. For each of the training image pairs, feature pairs are generated which

TABLE I: Summary of the baseline approaches.
L1 - Layer - I , L2 - Layer II , Kernel - × means No kernel is used

Baseline	Kernel - L1	Invariance - L1	Marginalization - L1	Kernel - L2	Marginalization - L2	Metric Learning
2 layer kLFDA	RBF-χ^2	×	×	RBF-χ^2	×	kLFDA
no_Marg	RBF-χ^2	✔	×	×	×	SVMML
no_Inv	RBF-χ^2	×	✔	×	✔	SVMML
marg_SVMML	RBF-χ^2	✔	✔	×	×	SVMML
marg_kLFDA	RBF-χ^2	✔	✔	RBF-χ^2	×	kLFDA
Ours	RBF-χ^2	✔	✔	×	✔	SVMML

results in a total of 1896 such pairs. All the local features were used as exemplars and features were transformed into the kernel space which led to a 3792 dimensional representation for each of the stripes.

We compare our approach with popular kernel based and other non-linear metric learning algorithms proposed for person re-identification and it can be seen from Table IIIa that our method outperforms all the other metric learning approaches for person re-identification. Table IIa shows the performance comparison of our approach against the baselines. It can be seen that the proposed invariant feature learning framework with marginalization can achieve better results than its variants.

The performance comparison with the state-of-the-art methods is shown in Table V. The proposed algorithm beats all the other methods individually at all ranks. Combining different methods with complementary features and ensemble of metric learning algorithms have also been studied before. As shown in table V, the metric ensembles proposed in [41] gives a matching rate of 45.9% at Rank 1. Combination of our approach with [6] outperforms all the existing methods for VIPeR dataset achieving state-of-the-art results. Results show that our features are complementary to the features used in [6]. It should be noted that our approach alone has comparable performance with the results obtained by [41] and the combination of [22] and [6] at higher ranks.

2) CAVIAR4REID Dataset: CAVIAR4REID [13] is another challenging dataset for evaluating person re-identification algorithms. This dataset is extracted from the well known CAVIAR dataset for evaluating pedestrian tracking and detection algorithms. It contains a total of 1220 images of 72 different individuals captured from arbitrary viewpoints under varying illumination. Among the 72 pedestrians, 50 of them appear in both camera views and the remaining 22 appear only in one camera view. The image resolution varies from 17 × 39 to 72 × 144. In our experiments, we resize each image into 128 × 48 and extract the features as mentioned in Section III-A. For all the comparisons, the dataset was split into halves for training and testing which results in images of 36 pedestrians for training and single-shot experiment setting was adopted [18].

Performance comparison of our approach against the baselines is shown in table IIb. Proposed algorithm outperforms all the baselines. The performance comparison against popular linear and non-linear metric learning algorithms are shown in table IIIb. We also compare our approach to the state-of-the-art approaches for single-shot setting and show the results in Table VI. It can be seen that MFA achieves slightly better result than ours at rank 1. However, at higher ranks we outperform all the other methods. The proposed algorithm combined with the kLFDA [18] outperforms all the other methods at all ranks and sets the state-of-the-art performance in the single shot setting for this dataset.

3) iLIDS Dataset: i-LIDS is another multi-shot re-identification dataset captured at a busy airport arrival hall. There are a total of 476 images of 119 people. The number of images per person varies from 2 to 8. The images undergo large illumination change, considerable change in view angle and are largely occluded which makes this dataset more challenging. For all the comparisons, the dataset was split into halves for training and testing which results in images of 59 pedestrians for training.

Table IIc shows the performance comparison of our method against the baseline approaches. It can be seen that the proposed learning framework achieves the best results. Table IIIc shows the performance comparison of the proposed algorithm with other metric learning algorithms. Table VII shows the comparison of our approach with the state-of-the-art approaches in single-shot setting for this dataset. The proposed algorithm individually is comparable to [41] at higher ranks. Combination of our algorithm with kLFDA in [18] achieves state-of-the-art results at higher ranks for this dataset, but at rank 1, we observed that [41] outperforms ours. We believe that this is due to an efficient mechanism to learn weights for individual features or learned metric for boosting rank 1 performance, proposed in [41].

4) CUHK01 Dataset: CUHK01 is a re-identification dataset with 3884 images of 971 individuals captured at two different views in a campus environment. For each individual, there are two images in each view. All the images are manually cropped and normalized to 160×60 pixels. Variation in pose, illumination makes this re-identification dataset considerably challenging. The dataset is split into halves for training and testing which leads to 485 individuals for training and rest for testing.

Table IId shows the performance comparison of our method against the baseline approaches. It can be seen that the proposed learning framework achieves the best results. Table IIId shows the performance comparison of our method with other metric learning algorithms. Table VIII shows the comparison of our approach with the state-of-the-art approaches in single-shot setting. It can be seen that some recent works [41]

TABLE II: Performance Comparison of our approach with the baseline algorithms on the VIPeR, CAVIAR4REID, iLIDS and CUHK01 datasets. Proposed framework outperform all the baselines and other variants of this method except for VIPeR dataset where marg_kLFDA performs slightly better than our method at rank 20.

(a) ViPER

Method	Rank 1	Rank 5	Rank 10	Rank 20
2 layer kLFDA	32.7	65.2	79.1	90.2
no_Inv	35.3	70.6	82.5	91.7
no_Marg	33.2	67.7	79.8	90.9
marg_SVMML	34.8	69.4	82.6	91.2
marg_kLFDA	36.4	70.4	83.6	**93.4**
Ours	**39.3**	**73.0**	**84.6**	92.5

(b) CAVIAR4REID

Method	Rank 1	Rank 5	Rank 10	Rank 20
2 layer kLFDA	36.7	65.2	77.8	92.1
no_Inv	36.8	70.8	85.7	95.7
no_Marg	34.1	70.0	83.9	93.9
marg_SVMML	38.3	71.3	84.6	95.2
marg_kLFDA	**40.2**	73.1	85.9	**96.3**
Ours	39.9	**73.0**	**86.9**	95.7

(c) iLIDS

Method	Rank 1	Rank 5	Rank 10	Rank 20
2 layer kLFDA	37.3	65.1	76.5	89.3
no_Inv	37.0	65.9	77.8	88.3
no_Marg	35.0	62.5	75.6	87.5
marg_SVMML	36.1	66.3	78.9	89.5
marg_kLFDA	39.1	68.4	**81.7**	91.2
Ours	**39.5**	**70.4**	81.0	**91.4**

(d) CUHK01

Method	Rank 1	Rank 5	Rank 10	Rank 20
2 layer kLFDA	26.8	49.6	60.4	71.3
no_Inv	26.5	52.9	64.0	74.5
no_Marg	25.6	48.9	59.0	70.1
marg_SVMML	27.2	49.4	62.1	71.8
marg_kLFDA	28.2	50.5	60.8	71.5
Ours	**29.2**	**54.7**	**66.3**	**77.6**

TABLE III: Performance Comparison of our approach with popular linear and kernel based metric learning algorithms on the VIPeR, CAVIAR4REID, iLIDS and CUHK01 datasets. Proposed framework outperform all the other algorithms except for CAVIAR4REID dataset where rPCCA performs slightly better than our method at rank 20.

(a) ViPER

Method	Rank 1	Rank 5	Rank 10	Rank 20
PCCA [37]	19.6	51.5	68.2	82.9
rPCCA	22.0	54.8	71.0	85.3
LFDA w/o kernel	19.7	46.7	62.1	77.0
KISSME [25]	23.8	52.9	67.1	80.5
SVMML	27.0	60.9	75.4	87.3
kLFDA	32.3	65.8	79.7	90.9
MFA	32.2	66.0	79.7	90.6
Ours	**39.3**	**73.0**	**84.6**	**92.5**

(b) CAVIAR4REID

Method	Rank 1	Rank 5	Rank 10	Rank 20
PCCA [37]	33.4	67.2	83.1	95.7
rPCCA	34.0	67.5	83.4	**95.8**
LFDA w/o kernel	31.7	56.1	70.4	86.9
KISSME [25]	31.4	61.9	77.8	92.5
SVMML	25.8	61.4	78.6	93.6
kLFDA	35.9	63.6	77.9	91.2
MFA	38.4	69.0	83.6	95.1
Ours	**39.9**	**73.0**	**86.9**	95.7

(c) iLIDS

Method	Rank 1	Rank 5	Rank 10	Rank 20
PCCA [37]	24.1	53.3	69.2	84.8
rPCCA	28.0	56.5	71.8	85.9
LFDA w/o kernel	32.2	56.0	68.7	81.6
KISSME [25]	28.0	54.2	67.9	81.6
SVMML	20.8	49.1	65.4	81.7
kLFDA	36.9	65.3	78.3	89.4
MFA	32.1	58.8	72.2	85.9
Ours	**39.5**	**70.4**	**81.0**	**90.7**

(d) CUHK01

Method	Rank 1	Rank 5	Rank 10	Rank 20
PCCA [37]	17.9	41.2	54.8	69.3
rPCCA	21.8	47.9	60.8	73.8
LFDA w/o kernel	15.7	34.4	44.6	56.6
KISSME [25]	10.3	27.2	37.5	49.7
SVMML	13.5	32.5	43.7	57.3
kLFDA	26.1	49.4	58.4	71.8
MFA	27.2	47.7	58.4	70.2
Ours	**29.2**	**54.7**	**66.3**	**77.6**

outperforms a combination of ours and [18] for iLIDS dataset at Rank 1 and CUHK01 dataset at all ranks.

V. Performance Analysis

For a detailed evaluation of our approach, we conducted experiments with some baseline methods.

A. Ours vs 2 layer kLFDA

First, we extended the kLFDA framework in [18] to two layers. Compared to the performance achieved by the proposed framework, it can be seen that marginalization and invariance can perform better than the 2-layer kLFDA framework. This indicates that, advantages of training on large amount of data can be achieved by the proposed framework. It should

TABLE IV: Performance at different ranks on the VIPeR dataset with different amount of training data. p denotes the number of training samples and r denotes the rank. It can be seen that the proposed method outperforms the best performing metric learning algorithm for person re-identification.

Algorithm	p=100				p=200				p=432				p=532			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
SVMML [6]	9.7	30.1	43.6	60.0	16.3	43.9	58.7	75.0	33.5	68.9	84.0	92.8	46.4	82.6	91.1	97.4
kLFDA [18]	15.9	38.8	52.7	67.9	25.4	53.3	68.3	82.3	45.2	81.2	91.4	97.2	56.4	90.8	96.6	99.3
MFA [18]	15.5	38.3	52.0	67.7	25.0	52.7	68.3	82.3	45.0	80.6	91.2	97.2	56.3	89.9	96.1	98.9
Ours	**17.7**	**43.2**	**58.1**	**70.7**	**27.5**	**59.3**	**72.9**	**86.3**	**49.0**	**83.5**	**92.0**	**98.5**	**62.0**	**91.6**	**96.8**	**99.4**

also be noted that extending kLFDA to two layers does not give significant advantages over kLFDA directly applied to features. Individual local stripes from image pairs may not have enough discriminative capability compared to the whole image representation. Therefore, at the first stage, we extract only invariant features. Comparing the baselines marg_kLFDA and 2 layer kLFDA, it can be inferred that extracting invariant information is more suitable at the local patch level.

B. Ours without Invariance criterion (no_Inv)

The next baseline is no_Inv where no invariance criterion is enforced. Intuitively, cross-view invariance is a key cue for person re-identification. This was validated from our experiments and results are reported in table II for all the datasets. While comparing to the final results obtained by the proposed framework, it can be seen that the invariance criterion improves the performance by around $2-4\%$ for all the datasets.

C. Ours without Marginalization (no_Marg, marg_SVMML and marg_kLFDA)

The next baseline we chose was the variant of the proposed approach without marginalization to show how much gain can be achieved by this technique. Since most of the person re-identification datasets are small, incorporating marginalization should give a better generalization capability. To analyze the advantages of marginalization, experiments were conducted in three stages. First, we conduct experiments on all datasets without marginalization on both layers. It can be seen from table II that, our approach substantially outperforms the baseline no_Marg.

Further, we develop a variant without marginalization at the second layer, i.e. at the metric learning stage. Instead, the traditional kLFDA and SVMML were used for learning the metric. From table II, it can be observed that, performance of marg_SVMML is close to the performance achieved by 2 layer kLFDA for iLIDS and CUHK01 datasets at all ranks but better for VIPeR and CAVIAR4REID. Even though, the SVMML performance is inferior to kLFDA (table III) when applied directly to the features, it can be well inferred from the results in table II that the proposed invariant feature learning framework with marginalization substantially helps the SVMML framework to perform better. But the performance of marg_kLFDA is better than marg_SVMML which is a combination of the traditional SVMML with the first layer of the proposed framework. It is also noteworthy that marg_kLFDA can perform better than the 2 layer kLFDA.

Finally, the proposed marginalized SVMML is used as the metric learning framework. From table II, it can be seen that the proposed metric learning framework outperforms marg_SVMML in all scenarios indicating that marginalized SVMML can be advantageous. In most of the cases, marginalized SVMML outperforms marg_kLFDA. For VIPeR and CUHK01 dataset, there is a notable improvement for the proposed framework over marg_kLFDA, but for others, the improvement is not substantial. At some ranks marg_kLFDA outperforms ours by a very small margin of $0.1-0.4\%$. But in VIPeR and CUHK01, significant improvements can be seen at higher ranks. These results show that advantages of data augmentation can be achieved by the proposed framework based on marginalization and it achieves a better generalization over unseen data. Figure 3 shows a visualization of some of the results obtained by the proposed approach and marg_SVMML. It can be seen that, many of the detections by marg_SVMML in the top $5-10$ ranks were successfully detected at Rank 1 by our approach.

D. Performance variation with number of training data on VIPeR

More experiments were conducted on the VIPeR dataset using different number of training samples to validate the advantages of the proposed marginalized feature learning and metric learning framework. Table IV shows the results of the experiment and it can be seen that the proposed framework outperforms all the best performing metric learning methods for person re-identification. When $p=100$, as shown in the table IV, it can be seen that the proposed algorithm outperforms the traditional SVMML by a very large margin. Compared to MFA and kLFDA, eventhough the rank 1 performance is improved by only 2%, at higher ranks, it can be seen that the performance improvement is significant $(3-6\%)$. Similarly, when $p=200$, the performance improvement at higher ranks is improved by around $(3-7\%)$ at higher ranks.

E. Comparison with state-of-the-art methods

Comparison with state-of-the-art methods is given in table V - VIII for different datasets. It can be seen that, individually, the proposed method outperforms all the recent approaches for person re-identification such as [27], [20], [4] and [18]. For clear distinction between individual methods and ensemble

TABLE V: Performance Comparison of state-of-the-art algorithms for the VIPeR dataset. Proposed approach when combined with [6] outperforms all the existing state-of-the-art methods for VIPeR dataset. The performance of the proposed algorithm individually is also comparable to the previous state-od-the-art methods at higher ranks. Results for [41] and [27] were taken from the CMC graphs in the respective literature. NA - Not Available in the literature

Method	Rank 1	Rank 5	Rank 10	Rank 20
Kernel Descriptors [16]	18.1	44.0	59.8	77.5
LFDA [7]	24.1	51.2	67.1	82.0
SSCDL [39]	25.6	53.7	68.1	83.6
PatMatch [24]	26.9	47.5	62.3	75.6
Mid-level [22]	29.1	52.3	65.9	79.9
SVMML [6]	29.4	63.3	76.3	88.1
VWCM [42]	30.7	63.0	76.0	88.6
SalMatch [23]	30.2	52.3	65.5	79.1
Deep ML [5]	28.2	59.3	73.5	86.4
DL 2015 [27]	34.8	63.7	75.8	NA
CMWCE [20]	37.6	68.1	81.3	90.2
Salient Color names [4]	37.8	68.5	81.2	90.4
Ours	**39.3**	**73.0**	**84.6**	**92.5**
Mid-level [22] + SVMML [6]	43.4	73.0	84.9	93.7
Metric Ensembles[41]	45.9	77.5	88.9	95.8
marg_kLFDA + SVMML[6]	46.5	74.1	86.4	95.1
Ours + SVMML[6]	**47.9**	**79.7**	**90.2**	**95.9**

methods, we have separated them by a horizontal split in tables V, VII and VIII. However, combination of several methods are becoming popular for person re-identification in the recent literatures [22], [41]. It can be seen that, for VIPeR dataset, a combination of [22] and [6] achieved 43.4% at Rank 1 where as another ensemble approach, [41] achieved 45.9% at Rank 1. A combination of ours with [6] achieved the new state-of-the-art for this dataset - 47.9% at Rank 1. But for other datasets, such as iLIDS and CUHK01, it can be seen that [41] outperforms the combination proposed by us at Rank 1 and all ranks respectively. In [41], a learning algorithm is proposed to combine the matching scores obtained from different set of features for a metric learning so as to improve the lower rank performances. However, in our framework, matching scores obtained for same features using different metric learning framework are added together after rescaling them from $0-1$ for each query image. We would like to point out that the algorithm proposed in [41] is in fact complementary to ours in the sense that for fusing the scores obtained from different features (or different metric learning algorithms), similar learning approaches can be used. But since this is beyond the scope of the proposed algorithm, the obtained results with out any learning mechanism for fusion is reported.

TABLE VI: Performance Comparison of state-of-the-art algorithms for the CAVIAR4REID dataset in the single-shot setting. The proposed approach outperforms all the methods except MFA at rank 1. A combination of our approach with kLFDA [18] achieves the state-of-the-art results for CAVIAR4REID dataset.

Method	Rank 1	Rank 5	Rank 10	Rank 20
LFDA [7]	32.0	56.3	70.7	87.4
kLFDA [18]	35.9	63.6	77.9	91.2
SVMML [6]	31.2	62.8	78.5	94.2
MFA [18]	40.2	70.2	83.9	95.1
Ours	39.9	73.2	88.4	95.7
Ours + kLFDA [18]	**45.1**	**76.8**	**88.9**	**97.4**

TABLE VII: Performance Comparison of state-of-the-art algorithms for the iLIDS dataset in the single-shot setting. Proposed approach performs comparably to [41]. The combination of ours with kLFDA [18] achieves state-of-the-art results at higher ranks but performs inferior to Metric Ensembles at rank 1. We believe that this is due to the learned score combining mechanism proposed in [41].

Method	Rank 1	Rank 5	Rank 10	Rank 20
PRDC [8]	37.8	63.7	75.1	88.4
kLFDA [18]	38.0	65.1	77.4	89.2
Ours	**39.5**	**70.4**	**81.0**	**91.4**
Metric Ensembles [41]	**50.3**	71.9	80.6	91.3
Ours + kLFDA [18]	47.7	**73.1**	**84.9**	**93.9**

TABLE VIII: Performance Comparison of state-of-the-art algorithms for the CUHK01 dataset in the single-shot setting. The proposed approach outperforms all the methods at all ranks. Results for [41] and [27] were taken from the CMC graphs in the respective literature.

Method	Rank 1	Rank 5	Rank 10	Rank 20
SDALF [3]	9.9	22.6	30.3	41.0
eSDC [24]	19.7	32.7	40.3	50.6
LMNN [9]	13.5	31.3	42.3	54.1
ITML [43]	16.0	35.2	45.6	59.8
SalMatch [23]	28.5	45.9	55.7	68.0
Midlevel [22]	**34.3**	**55.1**	65.0	74.9
Ours	29.2	54.7	**66.3**	**77.6**
Metric Ensembles [41]	**53.4**	**76.7**	**84.4**	**90.1**
Ours + kLFDA [18]	39.5	62.1	74.3	82.9

VI. Conclusion

We proposed a novel invariant feature learning framework with marginalization for person re-identification. To handle the non-linearities introduced by variations in pose, illumination and environment, Local features extracted from the images and are first transformed into a kernel space. A linear transformation is learned in this kernel space to capture invariant information by using labeled image pairs. Since the amount of labeled pairs is less, we propose a novel objective function with marginalization to reap the benefits of training on *infinite*

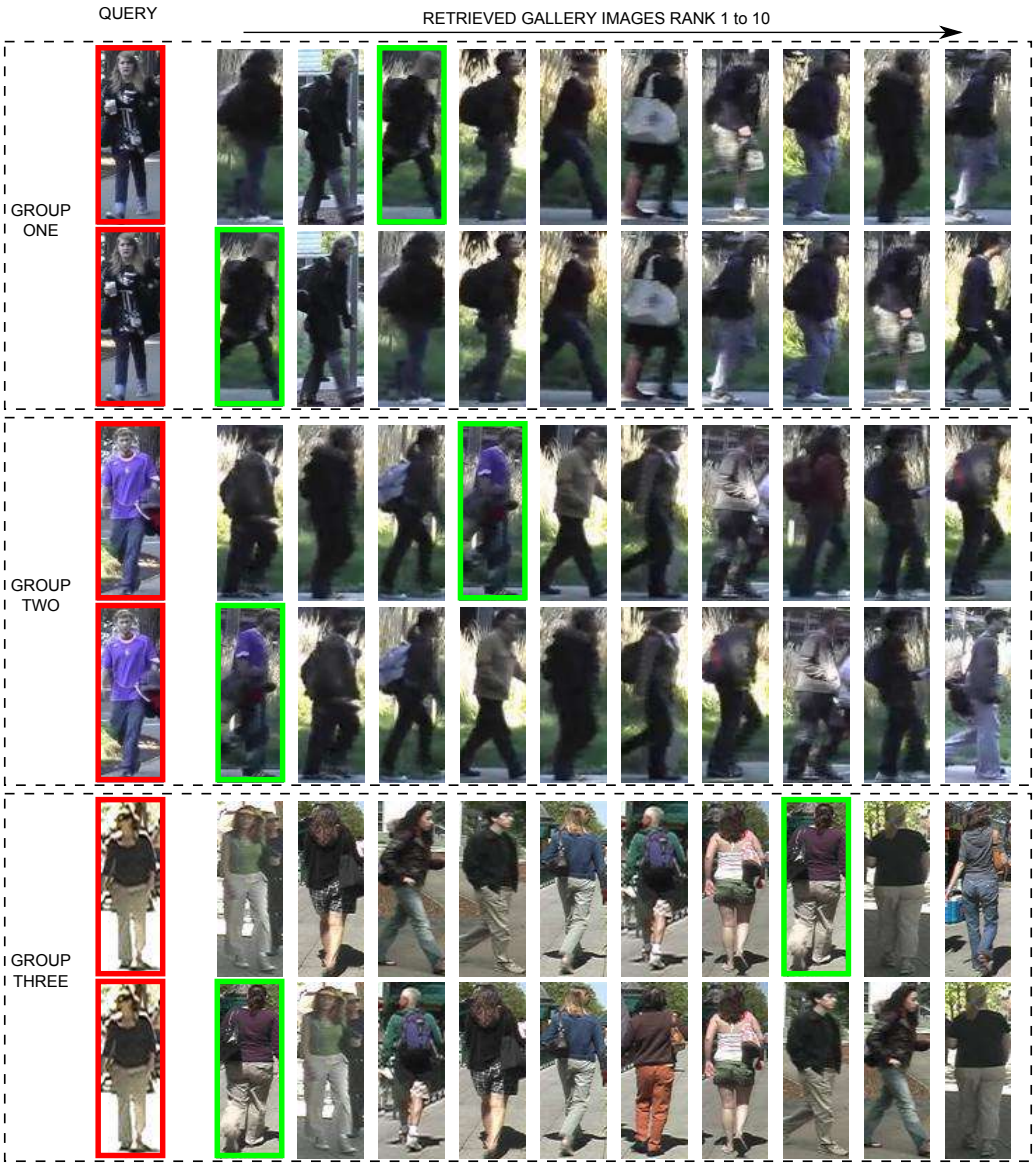

Fig. 3: Visualization of some results on SVMML and SVMML with marginalization. In each of the groups, as shown in the figure, the first row shows the top 10 retrieved matches for a query using marg_SVMML. The second row shows the top 10 matches retrieved by the marginalized SVMML. It can be seen that several images within the top 5 ranks or top 10 ranks of marg_SVMML were retrieved successfully at rank 1 by the proposed framework.

data. These mapped local features are concatenated to form the whole image representation and fed into a metric learning framework for classification. To achieve better generalization over the test set, we proposed a marginalized metric learning framework based on the popular SVM Metric Learning. The proposed approach was tested on four challenging publicly available datasets and our experiments show that learning invariant representations from the labeled images can improve the person re-identification performance. Additionally, the marginalization technique is shown to be advantageous when there is a lack of training data. Our comparison with state-of-the-art algorithms show that our method has a lot of future prospects.

REFERENCES

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *Proceedings of the British Machive Vision Conference (BMVC)*, 2012.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[4] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision (ECCV)*. 2014.

[5] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2014.

[6] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[7] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[8] W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[9] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, 2009.

[10] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2007.

[11] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2012.

[12] W. S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the British Machive Vision Conference (BMVC)*, 2009.

[13] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.

[14] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision (ECCV)*. 2008.

[15] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[16] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[17] M. Chen, K. Q. Weinberger, F. Sha, and Y. Bengio, "Marginalized denoising auto-encoders for nonlinear representations," in *Proceedings of the International conference on Machine learning (ICML)*, 2014.

[18] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision (ECCV)*. 2014.

[19] X. Wang and R. Zhao, "Person re-identification: System design and evaluation overview," in *Person Re-Identification*, 2014.

[20] Y. Yang, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Color models and weighted covariance estimation for person re-identification," *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2014.

[21] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *European Conference on Computer Vision (ECCV) - Workshops and Demonstrations*, 2012.

[22] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identfiation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[23] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[24] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[25] M. Kostinger, M. Hirzer, P. Wohlhart, P.M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[26] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proceedings of Asian Conference on Computer Vision (ACCV)*. 2010.

[27] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] J. Wang, K. Sun, F. Sha, S. Marchand-Maillet, and A. Kalousis, "Two-stage metric learning," in *Proceedings of the International conference on Machine learning (ICML)*, 2014.

[29] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, 2009.

[30] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, 2010.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012.

[32] G. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, 2006.

[33] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, 2006.

[34] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems (NIPS)*, 2007.

[35] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the International conference on Machine learning (ICML)*, 2008.

[36] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proceedings of the International conference on Machine learning (ICML)*. 2012.

[37] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[38] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002.

[39] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[40] Y. LeCun, L. Bottou, G. Orr, and K. Muller, "Efficient backprop," in *Neural Networks: Tricks of the trade*, 1998.

[41] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015.

[42] Z. Zhang, Y. Chen, and V. Saligrama, "A novel visual word co-occurrence model for person re-identification," in *European Conference on Computer Vision Workshop on Visual Surveillance and Re-Identification (ECCV Workshop)*, 2014.

[43] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.