IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 9, 2019, accepted February 14, 2019, date of publication February 25, 2019, date of current version March 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901599

Sparse Label Smoothing Regularization for Person Re-Identification

JEAN-PAUL AINAM[1,2], KE QIN[1], GUISONG LIU[1], AND GUANGCHUN LUO[1]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Adventist Cosendai University, Nanga-Eboko BP. 04, Cameroon

Corresponding authors: Ke Qin (qinke@uestc.edu.cn) and Guisong Liu (lgs@uestc.edu.cn)

This work was supported in part by the Ministry of Science and Technology of Sichuan Province under Grant 2017JY0073, and in part by the Fundamental Research Funds for the Central Universities in China under Grant ZYGX2016J083.

ABSTRACT Person re-identification (re-id) is a cross-camera retrieval task which establishes a correspondence between images of a person from multiple cameras. Deep learning methods have been successfully applied to this problem and have achieved impressive results. However, these methods require a large amount of labeled training data. Currently, the labeled datasets in person re-id are limited in their scale and manual acquisition of such large-scale datasets from surveillance cameras is a tedious and labor-intensive task. In this paper, we propose a framework that performs intelligent data augmentation and assigns the partial smoothing label to generated data. Our approach first exploits the clustering property of existing person re-id datasets to create groups of similar objects that model cross-view variations. Each group is then used to generate realistic images through adversarial training. Our aim is to emphasize the feature similarity between generated samples and the original samples. Finally, we assign a non-uniform label distribution to the generated samples and define a regularized loss function for training. The proposed approach tackles two problems 1) how to efficiently use the generated data and 2) how to address the over-smoothness problem found in current regularization methods. The extensive experiments on four large-scale datasets show that our regularization method significantly improves the re-id accuracy compared to existing methods.

INDEX TERMS Computational and artificial intelligence, artificial neural network, feature extraction, image retrieval.

I. INTRODUCTION

Person re-identification is the problem of identifying persons across images using different cameras or across time using a single camera. Automatic person re-id has become essential in surveillance systems due to the rapid expansion of large-scale and distributed multi-camera systems. However, many issues such as view point variations, dramatic variations in visual appearance, unstable light conditions, human pose variations, clothing similarity, background clutter and occlusions still prevent the task of achieving high accuracy. Despite the increasing attention given by researchers to solve the person re-id problem, it has remained a challenging task in practical environments.

Current approaches to solving person re-id are based on Convolutional Neural Network (CNN) and generally follow a verification or identification framework. A verification framework [27], [45], [59] usually takes a pair of images as input and outputs a similarity score while an identification framework [28], [36], [50], [60] learns a robust and discriminative feature representation from a single input image and predicts the person identity.

In general, CNN-based approaches to person re-id task received remarkable improvements and presented potentials for practical usage in modern surveillance system. However, CNN based methods require a large volume of labeled data for training to generalize. Furthermore, existing labeled datasets in person re-identification are limited in their scale by the number of the training images and by the number of images available for each identity. For example, Market-1501 dataset [58] contains 12, 936 training images and 751 identities, with 17 images on average per identities (i.e. 12, 936/751). Moreover, the need of large datasets becomes obvious as the task of labeling is manual, particularly tedious and labor-intensive. In addition, it involves manual selection of identities and association of images from different cameras with various view points, illumination, occlusions and body pose changes. This lack of large datasets

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaochun Cheng.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

is a big challenge in applying deep learning techniques to person re-id. Therefore, it is very important to find intelligent way to increase the training set.

Recently, Generative Adversarial Networks (GAN) [17] models have been particularly popular due to their ability to generate realistic-looking images via adversarial training. Thus, they can be used to solve the problem of lack of large datasets by generating synthesized unlabeled images which can be used in conjunction with the training set. However, transferring unlabeled images from the generated set to the training set is a challenging task and remains unresolved. Early studies to solve this problem adopted simplistic approaches. For instance, ''All in one'' [40] method assigns a single new label i.e., $K + 1$, to every generated sample. And, ''Pseudo Label'' [23] assigns the maximum class probability predictions of a pre-trained CNN model to the generated sample. Similarly, [16], [60], [63] proposed to use Label Smooth Regularization (LSR) to assign labels to fake samples. LSR was proposed in the 1980s and recently revisited in [43] as a mechanism to reduce over-fitting by estimating a marginalized effect over non-ground truth labels y during training by assigning small value to y instead of 0. Specifically, [60] extends LSR to outliers (LSRO) by assigning uniform label distribution (i.e. $\frac{1}{K}$) to generated images. This choice was made to avoid classifying generated samples into one of the existing categories. However, we argue that generated images have considerable visual differences and assigning same labels to all would lead to ambiguous predictions. This claim is also supported by [16]. Along this line, [16] proposed to assign labels based on the normalized class predictions over all pre-defined classes. We find that [16]'s method is similar to ''Pseudo label'' [23] and besides, empirical experiments conducted by [60] showed that LSRO is superior to ''All in one'' and ''Pseudo-label''.

One major drawback of all existing LSR approaches such as LSRO is that, they can easily lead to over-smoothness especially when the number of classes is excessively large. For instance, in a practical environment with thousand of identities, uniform label smoothing approach will assign value close to 0 and will fail to model the underlying relationships between the labeled and unlabeled data samples. In this work, we attempt to overcome this shortcoming by dynamically associating unlabeled samples with a subset of the class label distribution during the training process. Inspired by clustering that leverages the underlying patterns within data, we propose a novel label assigning approach called Sparse Label Smoothing Regularization (SLSR) which delivers significant performance boost in person re-identification, specifically for large-scale dataset.

In this paper, we make the following contributions:

1) We propose a GAN-based model tailored for person re-identification task with Sparse Label Smoothing Regularization (SLSR).
2) We use k-means to do clustering on the training set, generate GAN-based samples for each cluster and use partial smoothing label regularization over the generated images.
3) Using extensive experiments, we show that feature representation learning with SLSR improves the person re-identification accuracy.

The rest of this paper is organized as follows. Section II surveys the related works in person re-identification. Section III presents the proposed regularization method. Section IV presents the framework architecture; section V shows the implementation details and the experimental results and section VI concludes the paper.

II. RELATED WORKS

In this section, we describe the works relevant to our pipeline. These works include person re-identification and Generative Adversarial Network.

A. PERSON RE-IDENTIFICATION

Related works in person re-id can be roughly divided into two groups: distance metric learning and deep machine learning based approaches. The first group, also known as discriminative distance metric focuses on learning local and global feature similarities by leveraging inter-personal and intra-personal distances [6], [21], [29], [52], [56], [58]. The second group is CNN-based with a goal to jointly learn the best feature representation and a distance metric. Some feature based learning approaches [8], [24], [42] first decompose the images into three parts. Each part is then passed into a number of sub-networks for feature extraction. The three parts are finally fused at the fully connected layers and jointly contribute to the training process using a triplet loss function. Other methods [27], [45], [59] used a Siamese convolutional neural network architecture for simultaneously learning a discriminative feature and a similarity metric. Given a pair of input images, they predict if it belongs to the same subject or not through a similarity score. To improve the similarity score, [32], [61] proposed to optimize the evaluation metrics commonly used in person re-id.

Recently, [54], [60], [63] proposed to address the problem of lack of large datasets in person re-id by training a GAN [17] model to generate samples and a CNN model for identification task. It was particularly observed that, generated images with smooth labels can improve person re-id accuracy when they are combined with the training samples.

Following the success of attention mechanisms in Natural Language Processing, [28], [30], [36], [50] explored its application to the person re-id problem by proposing various forms of attentions. In details, [30] proposed an end-to-end Comparative Attention Network (CAN) to progressively compare the appearance of a pair of images and determine whether the pair belongs to the same person. During training, a triplet of raw images is fed into CAN for discriminative feature learning and local comparative visual attention generation. Reference [28] proposed a CNN architecture for jointly learning soft and hard attention. The two attention

mechanisms with feature representation learning are simultaneously optimized. In addition, [36] proposed gradient-based attention mechanism to solve the problem of pose and illumination found in person re-id problem in a triplet architecture and [50] recommended Co-attention based comparator to learn a co-dependent feature of an image pair by attending to distinct regions relative to each pair. Reference [59] proposed a Siamese network with verification loss and identification loss and predicted the identities of a pair of input images.

Many semi-supervised and unsupervised methods based on GAN have been developed [5], [53], [54], [63] to address the problem of lack of large labeled dataset in person re-id. Reference [5] introduced, for the first time in the re-id field, the strategy of using synthetic data as a proxy for the real data and claim to recognize people independently of their clothing. Reference [60] showed that a regularized method (LSRO) over GAN-generated data can improve the person re-id accuracy by assigning uniform label distribution to generated samples. Reference [63] proposed a camera style (CamStyle) adaptation method to regularize CNN training through the adoption of LSR and used CycleGAN [64] for image generation. Similarly, [23] trained a supervised network with labeled and unlabeled data by assigning *pseudo-label* to unlabeled data and [48], [53] proposed unsupervised asymmetric metric learning to unsupervised person re-id. In addition, [33] proposed Expectation-Maximization (EM) combining weak and strong labels under supervised and semi-supervised settings for image segmentation. Reference [25] proposed a semi-supervised region metric learning method to improve the person re-id task performance under imbalanced unlabeled data using label propagation with cross person score distribution alignment and discriminative region-to-region metric. Recently, [26] proposed a domain adaptation method to address the problem of lack of exhaustive identity label. Their proposed model jointly learns per-camera tracklet association and cross-camera tracklet correlation by maximising the discovery of tracklet across camera views and by exploiting the underlying re-id discriminative information in an end-to-end optimization.

Building from [35], [43], [60], [64], we propose a label assignment strategy that assigns partial label distribution to generated samples. We intend to use the training data in conjunction with GAN generated images to train the network using a regularized loss function.

We show in section III-C how our model differs from [60] and [63].

B. GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Network (GAN) is first introduced by [17] and described as a framework for estimating generative models via an adversarial process. GAN consists of two different components: a generator (G) that generates an image and a Discriminator (D) that discriminates real images from generated images. The two networks compete following the minimax two-player game. This kind of learning is called Adversarial Learning. Reference [35] proposed Deep Convolutional GAN (DCGAN) and certain techniques to improve the stability of GANs. The trained DCGAN showed competitive performance over unsupervised algorithms for image classification tasks. Multiple variants of GANs were published in the literature and were applied to various interesting tasks such as realistic image generation [35], text-to-image generation [37]; video generation [46]; image-to-image generation [19], image inpainting [34], super-resolution [22] and many more. In this work, we use DCGAN [35] model to generate unlabeled images for each cluster set. We chose DCGAN model after carefully contrasting various image generators. DCGAN architecture is very simple but yet generates more realistic images as illustrated in Figure 3.

III. OUR APPROACH

In this section, we present our proposed framework.

A. CLUSTERING THE TRAINING SET

We intend to partition the training samples into K groups of equal variance and find a shared feature space among similar objects. Our goal is to produce K different clusters with relatively similar features. To do this, we defined an objective function like that of *k-means* clustering [2], [13].

$$\mathcal{L}_{clustering} = \sum_{i=1}^{N} \sum_{k=1}^{K} || z_i - \mu_k ||^2 \tag{1}$$

where N is the number of cases, μ_k the cluster center and $||\ .\ ||$ the Euclidean distance between an embedded data z_i and the cluster center μ_k. In our experiments, we replaced z_i by the output feature map produced by a pre-trained model. Equation 1 learns the centroid such that, given a threshold γ, distances between similar feature vector are smaller than γ, while those between dissimilar feature vector are greater than γ. This ensures that distance between generated samples and a subset of the training images is small. We argue that using a generative model on similar objects effectively contributes in maintaining the complex relationships between unlabeled and labeled data, minimizes the affinity distance between the two sample sets and approximates the actual training data. In addition, experimental results have shown that using the intermediary feature representation of a pre-trained CNN model instead of the raw image results in better clustering quality.

To generate realistic images from each cluster, we defined a loss function similar to [11] and minimized Equation 2 with respect to the parameters of $G(z)$ and maximized Equation 2 with respect to the parameters of $D(x)$.

$$\mathcal{L}_{GAN} = \log D(x) + \log\Big(1 - D(G(z))\Big) \tag{2}$$

B. SPARSE LABEL DISTRIBUTION SCHEME

Let $p(\tilde{y}_i = y_i|\boldsymbol{I}_i)$ be a vector class probabilities produced by the neural network for an input image $\boldsymbol{I}_i$ and $\boldsymbol{w}_i$ the combination of weight and bias terms to be learned. The network

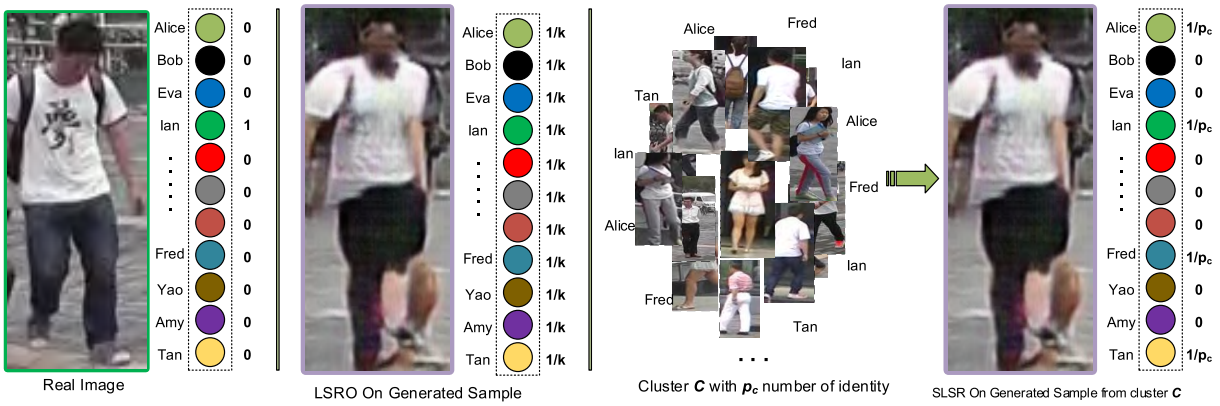


FIGURE 1. Real image (left) uses one-hot vector to encode the label information. LSRO (middle) uses a uniform label distribution $\frac{1}{k}$ on generated samples, while SLSR (right) uses partial label distribution drawn from the label distribution of the cluster of origin for label information.

computes the probabilities of each input image using:

$$p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^T \cdot \boldsymbol{x}_i)}{\sum_{k=1}^{N} \exp(\boldsymbol{w}_k^T \cdot \boldsymbol{x}_i)} \tag{3}$$

where $\boldsymbol{x}_i$ is the input vector from previous layers. Given N training samples, we define the cost function for real images as the negative log-likelihood:

$$\mathcal{L}_{xent} = -\sum_{i=1}^{N} \log p(\tilde{y}_i = y_i|\boldsymbol{I}_i) \tag{4}$$

In general, neural network represents a function $f(x;\theta)$ which provides the parameters $\boldsymbol{w}$ for a distribution over y. So minimizing $\mathcal{L}_{xent}$ is equivalent to maximizing the probability of the ground-truth label $p(\tilde{y}_i = y_i|\boldsymbol{I}_i)$. For a given person with identity y, Equation 4 can be written as

$$\mathcal{L}_{xent}(\theta) = -\log p(y|\boldsymbol{x};\theta) \tag{5}$$

where $\boldsymbol{\theta}$ represents the set of parameters of the whole network to be learned.

Regularization via Sparse Label Smoothing (SLSR) [43] proposed a mechanism to regularize a classifier by estimating a marginalized effect over non-ground truth labels $q(k|x)$ during training by assigning small value to y instead of 0. $q(k|x) = \delta_{k,y}$ where $\delta_{k,y}$ is Dirac delta:

$$\delta_{k,y} = \begin{cases} 1 & k = y \\ 0 & k \neq y \end{cases} \tag{6}$$

For training image with ground-truth label y, [43] replaced the label distribution $q(k|x) = \delta_{k,y}$ with

$$q\prime(k,y) = \begin{cases} (1-\epsilon)\delta_{k,y} & k = y \\ \dfrac{\epsilon}{k} & k \neq y \end{cases} \tag{7}$$

where $\epsilon \in [0,1]$ is the smoothing parameter. When $\epsilon = 0$, Equation 7 can be reduced to Equation 6. Then, the cross-entropy loss in Equation 5 is re-defined as

$$\mathcal{L}_{LSR} = -(1-\epsilon)\log p(y|x;\theta) - \frac{\epsilon}{K}\sum_{i=1}^{K} \log p(y_i|x;\theta) \tag{8}$$

Departing from [43], we introduce our loss function for the feature representation learning as a combination of cross entropy and a modified version of LSR. Given an identity I

$$z_{i,c} = \begin{cases} 1 & \boldsymbol{I}_i \in \mathcal{C} \\ 0 & \boldsymbol{I}_i \notin \mathcal{C} \end{cases} \tag{9}$$

Here, $z_{i,c}$ are the unnormalized probabilities of an image generated using cluster $\mathcal{C}$ with p_c number of classes. z_i represents a one-hot encoding vector where every entry k is equal to 1 if the class label k belongs to $\mathcal{C}$ and 0 if not. We consider the ground-truth distribution over the generated image and normalize z_i so that $\sum_{i=1}^{N} z_{i,c} = 1$. To explicitly take into account our label regularization, we changed the network to produce

$$z_i = \frac{1}{p_c} z_{i,c} \qquad \text{for} \quad c \in \{1, 2, \ldots, K\} \tag{10}$$

Figure 1 illustrates our proposed label distribution scheme. We finally optimize $\sum_i \mathcal{L}(\tilde{z}_i, \frac{1}{p_c} z_{i,c})$. Our loss for generated images is written as:

$$\mathcal{L}_{SLS} = -\sum_{i=1}^{p_c} \log p(\tilde{z}_i = z_i|\boldsymbol{I}_i) \tag{11}$$

or simply written as

$$\mathcal{L}_{SLS}(\theta) = -\log(p(z|\boldsymbol{x};\theta) \tag{12}$$

Combining Equation 5 and Equation 12, the proposed regularized loss function $\mathcal{L}_{SLSR}$ is defined as:

$$\mathcal{L}_{SLSR}(\theta) = -(1-\lambda)\log\Big(p(y|\boldsymbol{x};\theta)\Big) - \frac{\lambda}{K}\log\Big(p(z|\boldsymbol{x};\theta)\Big) \tag{13}$$

For training images, we set $\lambda = 0$ and for the generated images, $\lambda = 1$

C. DISCUSSION

Recently, [60] proposed Label Smoothing Regularization for Outliers (LSRO) and [63] proposed CamStyle as a data augmentation technique. LSRO expands the training set with unlabeled samples generated by DCGAN [35] and assigns uniform LSR [43] to a generated sample i.e. $\mathcal{L}_{LSR}(\epsilon = 1)$ while CamStyle uses CycleGAN [64] to generate new training samples according to camera styles and assigns $\mathcal{L}_{LSR}(\epsilon = 0.1)$ to style-transferred images. Although LSRO and CamStyle are similar to our work, we argue that our method is different on two aspects:

1) LSRO [60] and CamStyle [63] assign equal smoothing label distribution to all generated images; this can lead to over-smoothness especially when the number of classes is excessively large. However, our method assigns an adaptive smoothing label distribution to a generated sample based on the label distribution of its cluster c i.e $\mathcal{L}_{LSR}(\epsilon = \frac{1}{p_c})$ where p_c is the number of class identity in cluster c. In SLSR, $\epsilon = \frac{1}{p_c}$ is not unique and depends on p_c. This is opposed to $\epsilon = 1$ and $\epsilon = 0.1$ used in LSRO and CamStyle, respectively. Moreover, in LSRO and CamStyle, dissimilar and similar images may be assigned relatively equal similarity value, while our method deals with such unfairness by considering a generated image in the locality of real samples and proposes a strategy to determine the appropriate candidates by using *k-means* clustering algorithm. A non-uniform label distribution is assigned to generated images according to their cluster of origin. This enables our model to be highly efficient in dealing with large amount of data while being robust to noise as well. Our method SLSR learns the most discriminative features and can easily avoid the over-smoothness problem.

2) In our model, similarities are maintained and propagated through the framework by the concatenation of similar images into one homogeneous feature space. Leveraging feature space for each cluster can substantially improve the performance of person re-identification compared with using single-label distribution over all classes. Figure 1 illustrates the label distribution of SLSR and LSRO and clearly describes the uniform distribution of LSRO versus the non-uniform distribution of SLSR. Comparative studies in Tables 6 7 8 9 ascertain the effectiveness of our method and extensive experiments demonstrate its superiority compared to LSRO [60] and CamStyle [63]. In addition, our framework introduces an extra noise layer to match the noisy GAN label distribution. The parameters of this linear layer can be estimated as part of the training process and involve simple modification of current deep network architectures.

LSRO, CamStyle and our method SLSR share some common practices such as (1) enhancing the training set by the generation of fake images using GAN [17] models; (2) the adoption of Label Smooth Regularization (LSR) proposed by [43] to alleviate the impact of noise introduced by the generated images; (3) performing an end-to-end training for person re-id using labeled and unlabeled data in a CNN-based approach.

Algorithm 1 Algorithm for SLSR Training

Input: $\mathcal{K}$: Number of clusters, $\mathcal{X}$: Training samples
Initialisation: Randomly initialize the cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$
Draw m samples $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}\}$ from the training data $\mathcal{X}$ and train a CNN for **I** iteration using Equation 5
for each sample m **do**
  Extract $x_{(m)}^{(n)}$ feature map from the last conv layer
end for
Let $\mathcal{F} \in \mathbb{R}^{N \times M}$ be the feature maps for all samples
repeat
  for **every** $x^{(i)} \in \mathcal{F}$ **set** $c^{(i)} := arg\min_j || x^{(i)} - \mu_j ||$
  for **each** j **set** $\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)}=j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=j\}}$
until convergence
for each image $x_i \in \mathcal{X}$, assign x_i to μ_k using Equation 1
for each clusters k_i **do**
  Train a GAN with m example $\{\eta^{(1)}, \ldots \eta^{(m)}\}$ drawn from the cluster k_i and m samples $\{z^{(1)}, \ldots, z^{(m)}\}$ drawn from noise prior $P_g(Z)$ using Equation 2
  Generate sample images and assign *sparse label smoothing distribution to the generated image*
end for
Add the generated images to the training set and train a CNN using Equation 12

We also compared SLSR properties with LSRO, ''Pseudo Label'' and ''All-in-one'' methods. The overall comparison of our approach SLSR with the closely related methods is summarized in Table 1. Existing strategies to label GAN-based images in person re-id include ''Pseudo label'' [23], LSRO [60] and ''All in one'' [40]. SLSR and LSRO adopt smooth vector while ''All in one'' and ''Pseudo label'' adopt one hot vector. The difference is that, LSRO label contribution on pre-defined classes is the same, with a fixed and manually assigned value of $\frac{1}{k}$ while SLSR dynamically assigns label and considers their similarities. This ensures different label contribution on the pre-defined classes and accurately models practical environment settings.

IV. FRAMEWORK OVERVIEW

Our framework consists of three steps as illustrated in Figure 2 and includes (1) a clustering step using *k-means* clustering algorithm, (2) a generative adversarial training step for image generation and finally, (3) an identity classification training task using the original training set in conjunction with the generated set.

A. CLUSTERING

It is well known that multi-view data object admits a common clustering structure across view and that person re-id is a cross-camera retrieval task across view. We aim at exploring such clustering propriety to generate images that model

TABLE 1. Properties comparison between LSRO, All-in-One, Pseudo label and our method (SLSR).

Methods	Label distribution	Label contribution	Label source	Label assignment
All-in-One [40]	One Hot Encoding	Same	Manual	Static
Pseudo Label [23]	One Hot Encoding	Different	Probability	Dynamic
LSRO [60]	Smooth Encoding	Same	Manual	Static
SLSR	**Smooth Vector**	**Different**	**Similarity**	**Dynamic**

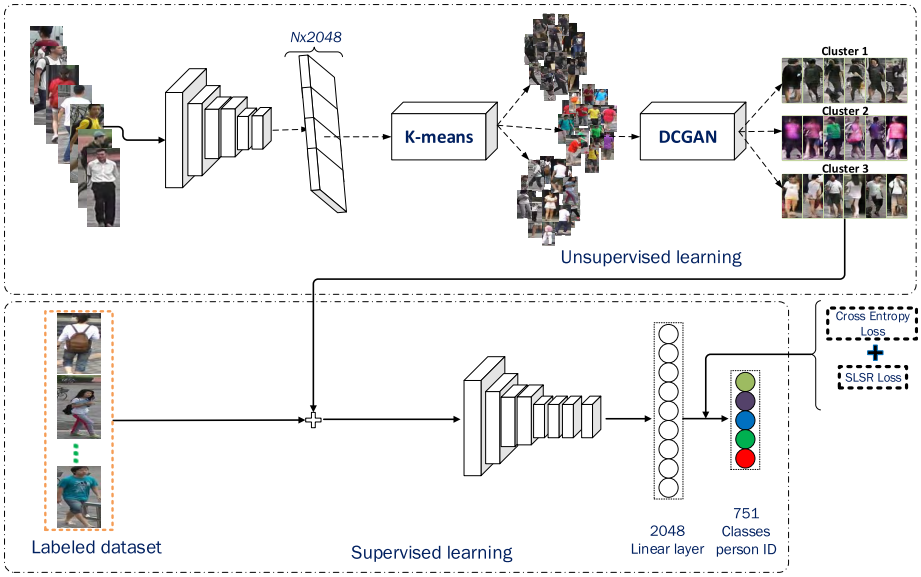


FIGURE 2. Our model consists of 3 steps: (1) Clustering on training data using unlabeled source dataset (Section IV-A). (2) For each cluster; train a DCGAN to generate images. Assign a partial label distribution to the generated images (Section III). (3) Combine the partial labeled images with the training image.

cross-view variations through the use of *k-means* clustering algorithm and GAN. We apply *k-means* algorithm to cluster the training images into K clusters $(2, \ldots, 5)$ as illustrated in Figure 4. *K-means* clustering is a simple yet very effective unsupervised learning algorithm for data clustering. It clusters data based on the Euclidean distance between data points. We trained a CNN network for 40 epochs using a learning rate of 0.001 with a momentum of 0.9. We use ResNet50 [14] model to learn a good intermediate representation and later extract high dimension features representation from the last convolutional layer. *K-means* clustering algorithm is applied to the set of feature map. We found this way to be faster and better than clustering on raw data images.

To judge the effectiveness of our clustering algorithm, we considered the ground truth not known and performed an evaluation using the model itself. Table 2 shows the cluster quality metric Silhouette Coefficient [39] applied on Market-1501 dataset [58]. We found Silhouette Coefficient higher for $K = 3$ and $K = 4$ showing that good cluster is achieved with these values of K. In the next sections, we use $K = 3$ for all the remaining experiments.

B. GENERATIVE ADVERSARIAL NETWORK

In this second step of our framework, we used Deep Convolution Generative Adversarial Network (DCGAN) [35] to generate data from clusters. We followed the implementation details of [35]. The Generator G consists of a Deconvolutional Network (DNN) made of $8 \times 8 \times 512$ linear function, a series of four deconvolution operations with a filter size of 5×5 and a stride of 2, and one tanh function. The input shape of G is a 100-dim uniform distribution Z scaled in the range of $[-1, 1]$ and the output shape a sample image of size $128 \times 128 \times 3$. The Discriminator D consists of Convolutional Neural Network (CNN) formed by four convolution functions with 5×5 filters and a stride of 2. We added a linear layer followed by a *sigmoid* function to discriminate real images against fake images. The input shape includes sample images from G and real images from the training set. Each convolution and

TABLE 2. For each cluster size, we calculate the silhouette coefficient [39] using mean intra-cluster distance (a) and mean nearest-cluster distance (b) ($\frac{b-a}{max(a,b)}$). The silhouette coefficient is generally higher when clusters are dense and well separated (best value is 1 and the worse value is -1). We show that this score is higher for cluster *size* $= 3$. Results from Table 4 prove that we achieve higher accuracy for $k = 3$, on Market-1501 dataset.

Number of clusters	Average silhouette score
2	51.75%
3	**70.03%**
4	68.49%
5	61.76%

FIGURE 3. Sample images generated from three clusters using DCGAN. The first column shows the original images from the cluster set and the remaining columns show samples generated from the corresponding cluster. We show that identities with similar features also generate fake samples with similar features and that color is a major learned feature.

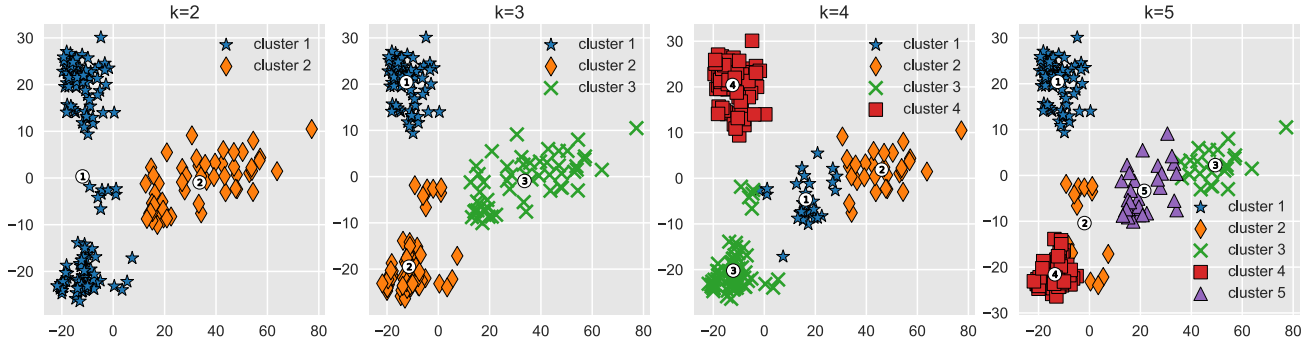


FIGURE 4. Visualization of extracted feature map $\mathcal{F}$ from ResNet on Market1501 dataset. Results of *k-means* clustering algorithm on $\mathcal{F}$ for $k = 2, \ldots, 5$. We arrive at a fair clustering view with $k = 3$ and $k = 4$. Best viewed in color.

deconvolution layer is followed by a batch normalization [18] and *ReLU* in both the generator and discriminator.

C. CONVOLUTIONAL NEURAL NETWORK

In the last step of the framework, we fine-tuned the ResNet [14] baseline model pre-trained on ImageNet, we introduced an extra linear layer into the network which adapts the network outputs to match the noisy GAN label distribution. The network was able to adjust the weights based on the error when we add a linear layer on top of the softmax layer rather than a non-linear such as *tanh* or *ReLU*. We used the generated data in conjunction with the labeled data and defined a loss function with a regularization term. The model is trained to minimize the loss function.

V. EXPERIMENTS

In this section, we performed experiments on four widely adopted person re-id datasets. The evaluation code is available at https://github.com/jpainam/SLS_ReID and is mainly conducted on Market-1501 dataset.

A. PERSON RE-ID DATASETS

Table 3 gives detailed information of the testing/training split strategy adopted during the experiments on Market-1501, CUHK03, DukeMTMC-ReID and VIPeR datasets.

Market-1501 [58]: is a large and most realistic dataset collected in front of a campus supermarket. It contains overlapping views among the six cameras and images were automatically detected by the Deformable Part Model (DPM) [9]. The dataset contains 12, 936 images with 751 identities in the training set and 19, 732 images with 750 identities in the test set. We follow the standard data separation strategy as described in [58] and use all the training set for the clustering step and one image per identity as validation image in the last step.

TABLE 3. Dataset split details. The total number of images (*QueryImgs, GalleryImgs, TrainImgs*), together with the total number of identities (*TrainID, TestID*) are listed.

Dataset	Market	CUHK03	VIPeR	Duke
#IDs	1501	1,467	632	1404
#Images	36,036	14,097	1,264	36,411
Cameras	6	2	2	8
TrainID	751	1367	316	702
TrainImgs	12,936	13,113	625	16,522
TestID	750	100	316	702
QueryImgs	3,368	984	632	2,228
GalleryImgs	19,732	984	316	17,661

CUHK03 [27]: contains 13, 164 images and 1, 467 identities. The dataset provides two image sets, one set is automatically detected by the Deformable Part Model [9], and the other set contains manually cropped bounding boxes. Misalignment, occlusions and body part missing are quite common in the detected set. In this work, we use the detected set as it is more realistic. The dataset is captured by six cameras, and each identity has an average of 4.8 images in each view.

DukeMTMC-ReID [60]: is a dataset derived from the DukeMTMC [38] dataset for multi-target tracking. The original dataset consists of a video data set recorded by 8 synchronized cameras over 2, 000 unique identities. In this paper, we use the subset as defined by [60]. It contains 16, 522 training images with 702 identities and 17, 661 test images with 702 identities. We follow the partition settings of the Market-1501 dataset and use all the training images for the first step and randomly pick one image per identity as validation set. The remaining training images are used for the supervised learning step.

VIPeR [12]: contains 632 pedestrian image pairs captured outdoor from two viewpoints. Each pair contains two images of the same individual cropped and scaled to 128×48 pixels. The datasets are divided into two equal subsets. To be fair in the comparison, we follow the testing strategy as defined in [12] and [57].

B. IMPLEMENTATION DETAILS

We modified ResNet50 [14] last fully connected layer with the number of classes i.e. 751; 1, 367 and 702 units for Market-1501, CUHK03 and DukeMTMCReID respectively. To train the network, we used stochastic gradient descent and start with a base learning rate of $\eta^{(0)} = 0.01$ and gradually decrease it as the training progresses using the inverse policy $\eta^{(i)} = \eta^{(0)}(1 + \gamma \cdot i)^{-p}$, where $\gamma = 0.1$, $p = 0.025$ and i is the current mini-batch iteration. We used a momentum of $\mu = 0.9$ and weight decay of $\lambda = 5 \times 10^{-4}$ and the mini-batch size of 32. We trained the network for 130 epochs. To generate image samples, we trained DCGAN for 30 epoch using Adam [20] with learning rate $lr = 0.0002$ and $\beta_1 = 0.5$.

Data preprocessing: All the input images are resized to 256×256 before being randomly cropped into 224×224 with random horizontal flip. We scaled the pixels between -1 and 1. Finally, pixels are zero-centered by subtracting their mean in each dimension and random erasing [62] is applied to make the network more robust to variations and occlusions.

C. BASELINE MODELS COMPARISON

We also compared SLSR and LSRO using our baseline. At first glance, our baseline already outperforms LSRO as it is reported in Table 5. Our baseline model fine-tuned ResNet model with an extra linear layer for the noisy data distribution and introduced a 512-bottleneck layer before the softmax layer while the baseline model used by LSRO makes no change to the existing ResNet architecture. For a fair comparison, we evaluated LSRO model on our baseline and showed the results of the experiments in Table 5. For instance, on Market1501 dataset, our baseline model improves LSRO by a factor of 4.66% on rank-1 accuracy and by a factor of 8.88% on mAP accuracy. This shows that the architectural design of our baseline also benefits LSRO. Such baseline can be adopted to improve the overall person re-id accuracy. Using the same baseline, we still observed a slight performance improvement. On Market-1501 dataset for example, under single query setting, SLSR slightly outperforms LSRO by a factor of 0.2% on mAP accuracy and 0.53% on rank-1 accuracy while under multi-query setting, SLSR outperforms LSRO by a factor of 2.05% on mAP accuracy and 0.83% on rank-1 accuracy. This improvement is explained by the relatively small size of the label distribution in Market1501 dataset. We recall that Market1501 dataset [58] contains 751 identities for 12, 936 training images. In this case, LSRO will assign a reasonable smooth value of 0.001 (1/751) while our method with 3 clusters will assign a relative value of 0.004. The two values are relatively closed. So, during training, the two models can converge identically. Nonetheless, in order to verify the effectiveness of the proposed method on a large class dataset and verify its robustness against the over-smoothness problem, we conducted an empirical study on CUHK03 dataset [27]]. As a quick reminder, CUHK03 dataset [27] contains 1367 identities for 13, 113 images, making it one of the largest dataset in person re-id in term of label distribution. The comparison of the results in Table 5 clearly shows that our model stands out from LSRO when the class label distribution is large. In details, we achieved a rank-1 accuracy improvement of 2.27% and a mAP accuracy improvement of 4.19%. We conclude that our model can better handle practical environment scenario with thousands of labels.

D. THE IMPACT OF USING DIFFERENT NUMBER OF CLUSTER

The impact of using different numbers of clusters and different number of synthesized images during training is also evaluated and reported in Table 4. We performed an ablation study and a performance comparison using 6000, 8000, 12000, 18000 and 24000 unlabeled images and expected the model to increasingly learn discriminative pattern from these data.

TABLE 4. Impact of the number of cluster on Market-1501 dataset. As the number of cluster gets larger, the accuracy drops. In general, we find that a large k decreases the training error but increases the validation/testing error. We show results of applying SLSR for 3 different values of k with no re-ranking [61] and single query setting. The best results are obtained with $K = 3$ and $K = 4$. $K = 3$ is used for experiments on all the datasets.

Cluster size	K = 2				K = 3				K = 4				K = 5			
Generated	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
6,000	88.30	95.90	97.50	74.08	90.59	96.58	97.86	77.56	88.92	95.93	97.62	74.31	87.32	94.00	96.31	65.88
8,000	88.98	95.75	97.56	74.35	91.18	96.94	98.13	78.43	90.08	96.73	98.01	76.25	88.03	94.65	96.46	67.34
12,000	**89.99**	**96.41**	**98.04**	**75.47**	**92.43**	**97.27**	**98.39**	**79.08**	**91.36**	**97.06**	**98.22**	**79.14**	**88.48**	**95.96**	**97.56**	**73.62**
18,000	89.49	96.17	97.62	75.63	91.95	96.70	98.24	78.94	91.06	96.85	98.07	78.30	88.56	95.75	97.26	74.00
24,000	89.49	96.08	97.53	75.10	91.15	96.43	97.71	78.21	91.05	96.79	98.19	77.40	87.85	95.25	96.91	72.78

TABLE 5. Comparison results with LSRO using our baseline. We applied LSRO loss on our baseline on Market-1501 dataset without re-ranking. We show that the architectural design of our baseline also benefits LSRO. SQ stands for Single Query and MQ for Multi-Query.

Methods	Market1501 SQ		Market1501 MQ		CUHK03	
	R1	mAP	R1	mAP	R1	mAP
Our Baseline	87.29	69.70	91.27	76.94	75.11	83.91
LSRO + Original Baseline [60]	83.97	66.07	88.42	76.10	84.62	87.40
LSRO + Our Baseline	88.63 ↑4.66	74.95↑8.88	91.42↑ 3.00	79.87↑3.77	88.76↑4.14	90.02↑2.62
SLSR	**89.16↑5.19**	**75.15↑9.08**	**92.25↑3.83**	**81.92↑5.82**	**91.03**	**94.21**

However, the results show that as the number of generated samples increases, the person re-id performance improves by a factor of 1.25% but reaches saturation with 12, 000 generated samples. We note that the number of training images in Market-1501 dataset is 12, 936. As a result, we make two remarks. First, the addition of different numbers of fake samples steadily improves the baseline. We find that the peak performance is achieved by roughly doubling the number of training samples with fake samples. Compared with LSRO where the peak performance is achieved when $2 \times GAN$ i.e. 24, 000 images are added, our approach only requires 12, 000 to reach peak performance. Also, increasing the number of GAN images beyond 12, 000 does not improve the accuracy. The network reaches early convergence thanks to SLSR. In addition, the number of cluster affects the rank-1 accuracy. In fact, if $K = 1$, the approach resembles LSRO; with $K > 2$ and $K < 5$, we observe accuracy improvement over the baseline but a drop in accuracy with $K > 5$. As the number of cluster increases, the learning procedure tends to converge towards assigning a single ground truth label to the fake samples similar to 'Pseudo label' scheme, which is not desirable. Therefore, we conclude that a trade-off is recommended to avoid poor regularization of partial label distribution.

E. EVALUATIONS

We adopted the widely used Cumulative Matching Curve (CMC) metric for quantitative evaluations. We used the standard protocol to ensure fair comparison between the proposed method and the state-of-the-art methods. The test protocols are as follow.

For VIPeR dataset, we randomly divide the dataset into training and testing sets, each set containing half of the available individuals. In the test set, we randomly select one image of a person from camera 1 as a query image and one image of the same person from camera 2 as a gallery image. For CUHK03 dataset, we followed the standard protocol used by [7] and for Market-1501 dataset, we used the standard evaluation protocol as defined by [58]. And, for DukeMTMC-ReID we used the standard evaluation protocol defined in [60]. Both single-query and multi-query matching results are reported on Market-1501 dataset while only single query evaluation is adopted for CUHK03, VIPeR and DukeMTMC-ReID datasets. Rank-1, rank-5, rank-20 accuracy and Mean Average Precision (mAP) are computed to evaluate the performance of all the methods. For each image in the query set, we first compute the L2 distance between the query image and all the gallery images using the output feature produced by our trained network, and we return the top-n nearest images in the gallery set. If the returned list contains an image of the same person at a given position k, then this query is considered as success at rank-k.

Re-Ranking: Recent works [4], [61] choose to perform an additional re-ranking to improve ReID accuracy. In this work, we report re-ranking results using re-ranking with k-reciprocal encoding [61], which combines the original L2 distance and Jaccard distance. Re-ranking with k-reciprocal encoding approach assumes that there are multi positive samples in the gallery. So, re-ranking approach will fail to improve the performance in small datasets such as ViPER and CUHK03 datasets. In this work, we did not report these results. In Tables 6 7 8 9, SLSR represents our method and SLSR + RR represents our model with re-ranking [61].

F. COMPARISON WITH THE STATE OF ART

In this section, we compare our results with state-of-art methods and report the results in Tables 6 7 8 9.

On **Market-1501** dataset our method achieved an **89.16%** rank-1 accuracy and **75.15%** mAP accuracy exceeding

TABLE 6. Comparison result with state-of-arts on CUHK03. '-' means that no reported results is available. * paper on ArXiv but not published.

Methods	R1	R5	R10	mAP
KISSME [21]	11.7	33.3	48.0	-
DeepReID [27]	19.89	50.00	64.00	-
TAUDL [26]	44.7			31.2
ImprovedDeep [1]	44.96	76.01	83.47	-
XQDA (LOMO) [29]	46.25	78.90	88.55	-
SI-CI [47]	52.20	84.30	94.8	-
DNS [55]	54.7	80.1	88.30	-
FisherNet [49]	63.23	89.95	92.73	44.11
MR B-CNN [44]	63.67	89.15	94.66	-
Gated ReID [45]	68.1	88.1	94.6	58.8
SOMAnet [5]	72.40	92.10	95.80	-
SSM [4]	72.7	92.4	96.1	-
SVDNet [41]	81.8	95.2	97.2	84.8
Cross-GAN [54]*	83.23	-	96.73	-
Verif.Identif. [59]	83.40	97.10	98.7	86.40
DeepTransfer [10]*	84.10	-	-	-
LSRO [60]	84.62	97.60	98.90	87.40
TriNet [15]	87.58	98.17	-	-
HydraPlus-Net [31]	**91.8**	**98.4**	99.1	-
(Ours) SLSR	**91.03**	**98.22**	**99.26**	**94.21**

TABLE 7. Comparison results of the state-of-arts methods on DukeMTMCReID. We show that our methods is superior to previous works. * paper on ArXiv but not published.

Methods	R1	R5	R10	mAP
BoW+KISSME [58]	25.13	-	-	12.17
XQDA (LOMO) [29]	30.75	-	-	17.04
TAUDL [26]	61.7			43.5
LSRO [60]	67.68	-	-	47.13
OIM [51]	68.1	-	-	47.4
TriNet [15]*	72.44	-	-	53.50
SVDNet[41]	**76.7**	86.4	89.9	56.8
(Ours) SLSR	**76.53**	**88.15**	**91.02**	**60.79**
(Ours) SLSR+RR	**82.67**	**89.72**	**93.00**	**79.23**

LSRO [60] by a factor of **5.19%** on rank-1 accuracy and by a factor of **9.08%** on mAP accuracy. Our method with both SLSR and re-ranking [61] with k-reciprocal encoding further improves rank-1 and mAP accuracy from 89.16% to **91.54%** and from 75.15% to **88.09%** respectively. Table 8 shows that our method outperforms many existing works.

On **CUHK03** dataset (Table 6), we achieved a **91.03%** rank-1 accuracy and **94.21%** mAP accuracy which are close by a factor of **0.77%** to the result reported by HydraPlus-Net [31]. Our method exceeds LSRO [60] by a factor of **6.41%** on rank-1 accuracy and by a factor of **6.81%** on mAP.

Not many reported results exist on **DukeMTMCReID** dataset, as shown in Table 7. Yet, our method achieved a **76.53%** rank-1 accuracy and **60.79%** mAP accuracy exceeding existing works. Compared to LSRO [60], our rank-1 accuracy exceeds their result by a factor of **8.85%**. SVDNet [41] exceeds our model by a small factor of **0.17%**.

We also achieved competitive result on a small dataset such as **VIPeR** dataset, Specifically, our method achieved a **65.98%** rank 1 accuracy.

TABLE 8. Comparison results of the state-of-art methods on Market-1501. '-' means that no reported results is available and '*' means the paper is available on ArXiv but not published.

	Single Query			
Methods	R1	R5	R10	mAP
BoW+KISSME [58]	44.42	-	-	20.76
FisherNet [49]	48.15	-	-	29.94
Simil.Learning [6]	51.90	-	-	26.35
DNS [55]	61.02	-	-	35.68
TAUDL [26]	63.7			41.2
Gate Reid [45]	65.88	-	-	39.55
MR B-CNN [44]	66.36	85.01	90.17	41.17
Cross-GAN [54]*	72.15	-	94.3	48.24
SOMAnet [5]	73.87	88.03	92.22	47.89
HydraPlus-Net [31]	76.9	91.3	94.5	-
Verif.Identif [59]	79.51	-	-	59.87
SVDNet [41]	82.3	92.3	95.2	62.1
DeepTransfer [10]*	83.7	-	-	65.5
LSRO [60]	83.97	-	-	66.07
TGP-ReID [3]*	92.2	97.9	-	81.2
(Ours) SLSR	**89.16**	**95.78**	**97.33**	**75.15**
(Ours) SLSR+RR	**91.54**	**95.37**	**96.62**	**88.09**
	Multi Query			
Methods	R1	R5	R10	mAP
DNS [55]	71.56	-	-	46.03
Gate Reid [45]	76.04	-	-	48.45
SOMAnet [5]	81.29	92.61	95.31	56.98
Verif.Identif [59]	85.47	-	-	70.33
LSRO [60]	88.42	-	-	76.10
DeepTransfer [10]*	89.6	-	-	73.80
TGP-ReID [3]*	94.7	98.6	-	87.3
(Ours) SLSR	**92.25**	**97.51**	**98.34**	**81.92**
(Ours) SLSR+RR	**94.18**	**98.06**	**98.78**	**90.10**

TABLE 9. Comparison results with state-of-arts on VIPeR dataset.

Methods	R1	R5	R10	R20
ImproveDeep [1]	34.81	63.61	75.63	84.49
KISSME [21]	34.81	60.44	77.22	86.71
Simil.Learning [6]	36.80	70.40	83.70	91.70
MFA (LOMO)[52]	38.67	69.18	80.47	89.02
XQDA (LOMO) [29]	40.00	68.13	80.51	91.08
Cross-GAN [54]*	49.28	-	**91.66**	93.47
DNS [55]	51.17	**82.09**	90.51	95.92
SSM [4]	53.73	-	91.49	**96.08**
SpindleNet [57]	53.80	74.1	83.2	92.1
HydraPlus-Net [31]	56.6	78.8	87.0	92.4
(Ours) SLSR	**65.98**	**81.49**	**88.45**	**95.25**

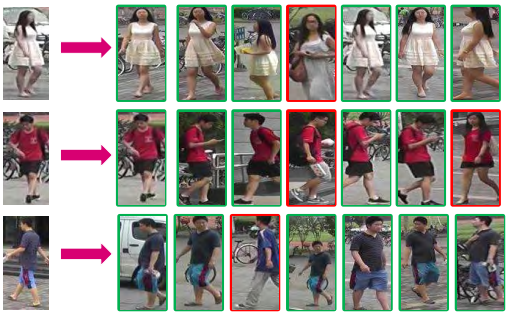


FIGURE 5. Sample images retrieved from Market-1501 dataset using our framework. The images in the first column are the query images. The images in the right columns are the retrieved images. The retrieved images are sorted according to the similarity scores from left to right. We use re-ranking [61] with *k*-reciprocal encoding.

VI. CONCLUSION

In this paper, we proposed Sparse Label Smoothing Regularization (SLSR) for solving the person re-identification problem. We proposed to use generated samples in conjunction with training samples to improve the re-id accuracy and proposed a labeling approach for generated samples. We emphasized on the fact that a fair labeling approach on synthesized images should consider the underlying relationship between the training and the generated samples. We proposed SLSR as a pipeline to train a CNN model with labeled and synthesized images. We clustered the training images using an intermediary feature representation of a pre-trained CNN model and generate images for each cluster. The generated images are assigned smooth label according to the label distribution of the cluster used for DCGAN stream. Through ablation, we show that SLRS can address the problem of over-smoothness found in current regularization methods. Extensive evaluations were conducted on four large-scale datasets to validate the advantage of the proposed model on existing models. Tables 6 7 8 9 show the superiority of the model over a wide variety of state-of-art methods.

ACKNOWLEDGEMENTS

The authors appreciate Yongsheng Peng, Eldad Antwi-Bekoe for their useful contributions and Yuyang Zhou for the management of the GPUs during experiments.

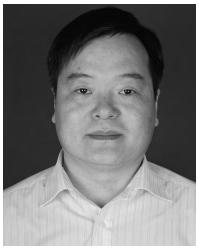
REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Boston, MA, USA, Jun. 2015, pp. 3908–3916.
[2] E. Aljalbout, V. Golkov, Y. Siddiqui, and D. Cremers. (2018). "Clustering with deep learning: Taxonomy and new methods." [Online]. Available: https://arxiv.org/abs/1801.07648
[3] J. Almazán, B. Gajic, N. Murray, and D. Larlus. (2018). "Re-id done right: towards good practices for person re-identification." [Online]. Available: https://arxiv.org/abs/1801.05339
[4] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3356–3365.
[5] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep CNNs in re-identification," *Comput. Vis. Image Understand.*, vol. 167, p. 50– 62, Feb. 2018.
[6] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1268–1277.
[7] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.
[8] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.
[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[10] M. Geng, Y. Wang, T. Xiang, and Y. Tian. (2016)"Deep transfer learning for person re-identification." [Online]. Available: https://arxiv.org/abs/1611.05244
[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[12] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. 10th IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, Sep. 2007, pp. 1–7.
[13] J. A. Hartigan, *Clustering Algorithms*. Hoboken, NJ, USA: Wiley, 1975.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[15] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737
[16] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang. (2018). "Multi-pseudo regularized label for generated samples in person re-identification." [Online]. Available: https://arxiv.org/abs/1801.06742
[17] G. J. Ian *et al.*, "Generative adversarial network," in *Proc. NIPS*, 2014, pp. 241–258.
[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37. Jul. 2015, pp. 448–456.
[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2016). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004
[20] D. Kingma and J. Ba. (2018). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[21] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2288–2295.
[22] C. Ledig *et al.* (2016). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802
[23] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML Workshop Challenges Represent. Learn. (WREPL)*, 2013.
[24] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. CVPR*, Jul. 2017, pp. 7398–7407.
[25] J. Li, A. J. Ma, and P. C. Yuen, "Semi-supervised region metric learning for person re-identification," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 855–874, 2018.
[26] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 772–788.
[27] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 152–159.
[28] W. Li, X. Zhu, and S. Gong. (2018). "Harmonious attention Network for person re-identification." [Online]. Available: https://arxiv.org/abs/1802.08122
[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 2197–2206.
[30] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE TIP*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[31] X. Liu *et al.*, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Aug. 2017, pp. 350–359.
[32] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Boston, MA, USA, Jun. 2015, pp. 1846–1855.

[33] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 2015, pp. 1742–1750.
[34] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2016, pp. 2536–2544.
[35] A. Radford, L. Metz, and S. Chintala. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434
[36] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, and H. Qi, "Person re-identification using visual attention," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 201, pp. 4242–4246.
[37] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. (2016). "Generative adversarial text to image synthesis." [Online]. Available: https://arxiv.org/abs/1605.05396
[38] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. (2016). "Performance measures and a data set for multi-target, multi-camera tracking." [Online]. Available: https://arxiv.org/abs/1609.01775
[39] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, p. 53– 65, Nov. 1987.
[40] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. NIPS*, 2016, pp. 2234–2242.
[41] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3820–3828.
[42] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. (2017). "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)." [Online]. Available: https://arxiv.org/abs/1711.09349
[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[44] E. Ustinova, Y. Ganin, and V. Lempitsky. (2015). "Multiregion bilinear convolutional neural Networks for person re-identification." [Online]. Available: https://arxiv.org/abs/1512.05300
[45] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural Network architecture for human re-identification," in *Proc. ECCV*, Sep. 2016, pp. 791–808.
[46] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 613–621.
[47] F.-Q. Wang, W.-M. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1288–1296.
[48] Y. Wang, W. Zhang, L. Wu, X. Lin, and X. Zhao, "Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 57–70, Jan. 2017.
[49] L. Wu, C. Shen, and A. Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognit.*, vol. 65, pp. 238–250, May 2016.
[50] L. Wu, Y. Wang, J. Gao, and D. Tao. (2018). "Deep Co-attention based comparators for relative representation learning in person re-identification." [Online]. Available: https://arxiv.org/abs/1804.11027
[51] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. CVPR*, 2017, pp. 3415–3424.
[52] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Computer Vis. – ECCV*. Cham, Switzerland: Springer, 2014, pp. 1–16.
[53] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Aug. 2017, pp. 994–1002.
[54] C. Zhang, L. Wu, and Y. Wang. (2018). "Crossing generative adversarial Networks for cross-view person re-identification." [Online]. Available: https://arxiv.org/abs/1801.01760
[55] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.
[56] Y. Zhang and S. Li, "Gabor-LBP based region covariance descriptor for person re-identification," in *Proc. 6th Int. Conf. Image Graph. (ICIG)*, Aug. 2011, pp. 368–371.
[57] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 907–915.
[58] L. Zheng, L. Sheng, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.
[59] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, p. 13, Jan. 2017.
[60] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in Vitro*," in *Proc. IEEE Int. Conf. Comput. Vis.*, Aug. 2017, pp. 3754–3762.
[61] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with *k*-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1318–1327.
[62] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. (2017). "Random erasing data augmentation." [Online]. Available: https://arxiv.org/abs/1708.04896
[63] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. CVPR*, 2018, pp. 5157–5166.
[64] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks." [Online]. Available: https://arxiv.org/abs/1703.10593

JEAN-PAUL AINAM received the B.Sc. degree in software engineering from Cosendai University, in 2012, and the M.Sc. degree in computer science from Babcock University, Nigeria, in 2014. He is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China under the supervision of Prof. K. Qin. He is also a Lecturer with Cosendai University. Over the last three years, he has taught number of courses in Cosendai University. His research interests include neural networks and deep machine learning, especially computer vision.

KE QIN was born in Sichuan, China, in 1980. He received the master's and Ph.D. degrees in computer science from the University of Electronic Science and Technology, China (UESTC), in 2010 and 2006, respectively. He was also a Visiting Scholar with Carleton University, Ottawa, Canada, in 2008, and also with the University of California at Santa Barbara, Santa Barbara, USA, in 2017. He is currently an Associate Professor with UESTC. He authored more than 40 refereed journal and conference publications. His research interests include neural networks and machine learning.



GUISONG LIU received the B.S. degree in mechanics from Xi'an Jiaotong University, Xi'an, China, in 1995, the M.S. degree in automatics, and the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2000 and 2007, respectively. He was a Visiting Scholar with Humbolt University at Berlin, in 2015. He is currently a Professor with the School of Computer Science and Engineering, UESTC. He has published more than 50 journal and conference papers. His research interests include pattern recognition, neural networks, and machine learning.

GUANGCHUN LUO received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, in 2004, where he has been a Full Professor with the Faculty of Computer Science, since 2009. He published more than 80 papers. His current research interests include artificial intelligence and big data.