Sparse Label Smoothing for Semi-supervised Person Re-Identification

Jean-Paul Ainam* jpainam@uacosendai-edu.net
Ke Qin† qinke@uestc.edu.cn
Guisong Liu‡ lgs@uestc.edu.cn
Guangchun Luo gcluo@uestc.edu.cn
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan, P.R. China, 611731

Abstract

In this paper, we propose a semi-supervised framework to address the over-smoothness problem found in current regularization methods. We carefully propose to derive a regularization method by constructing clusters of similar images. We propose Sparse Label Smoothing Regularization (SLSR) which consist of three steps. First, we train a CNN to learn discriminative patterns from labeled data. For each image, we extract the feature map from the last convolution layer and directly apply k-means clustering algorithm on the feature. Secondly, we train a GAN model for feature representation learning and generate sample images for each cluster. Each generated sample is assigned a label using our regularization method. Thirdly, we define a new objective function and fine-tuned two baseline models ResNet and DenseNet. Extensive experiments on four large-scale datasets Market-1501, CUHK03, DukeMTMC-ReID, and VIPeR show that our regularization method significantly improves the Re-ID accuracy compared to existing semi-supervised methods. On Market-1501 dataset, for instance, rank-1 accuracy is improved from 87.29% *to* 89.16% *for ResNet, and from* 90.05% *to* 92.43% *for DenseNet. The code is available at* `https://github.com/jpainam/SLS_ReID`

1. Introduction

Person re-identification is the process of establishing a correspondence between images of a person from multiple cameras, i.e. given a person; person re-id determines whether the person has been observed by another camera. The problem has been widely studied in the past and has achieved extraordinary results with deep learning based approaches [3, 11, 56, 58]. Modern deep learning methods require a large volume of labeled data for training to generalize well. Existing labeled data in person re-identification are limited in scale by the number of identities and by their size (30 images on average per identity). This lack of large datasets is a big challenge in applying deep learning technique to person re-identification. One way this can be lessened is by using unsupervised methods to train on data without labels. These methods learn features from the data which can then be used for supervised learning with small datasets.

In this work, we propose a semi-supervised framework that uses DCGAN [32] to generate data from clusters. These generated images are assigned a smooth label distribution based on their original cluster. We use the generated data in conjunction with the labeled data and define two losses, an unsupervised loss, and supervised loss. The model is trained to minimize the two losses.

As shown in Fig. 1; our framework has three main steps. The unsupervised step is fed with unlabeled data and output $\mathcal{F} \in \mathbb{R}^{N \times 2048}$ dimensional vectors representing the feature maps of N training images. The extracted feature maps $\mathcal{F}$ is then introduced into a *k-means* clustering algorithm to obtain k cluster sets. We use each cluster set to train an image generator [32] to generate sample images. As each generated image belongs to one of the clusters, we can assign a label to generated images through our regularization method. Finally, the semi-supervised step uses existing network architectures and introduces an extra linear layer, i.e. a noise layer which adapts the network outputs to match the noisy GAN label distribution. Our model can generalize well, and experiment results show that our method outperformed previous methods.

*Jean-Paul Ainam is also a lecturer at Cosendai University, Cameroon. He is currently a Ph.D student at University of Electronic Science and Technology of China.

†Corresponding author.

‡This author is an equal corresponding author.

In this paper, we make the following contributions:

- We propose a GAN-based model tailored for person re-identification task with a sparse label smoothing regularization (SLSR).
- We use an unsupervised learning approach to do clustering on the data and trained a GAN network to generated images for each cluster.
- We use partial smoothing label regularization over the generated images.
- We show that unsupervised representation learning with SLSR improves the person re-identification accuracy.

The rest of this paper is organized as follows. Section 2 surveys the related works in person re-identification. Section 3 presents the proposed regularization method. Section 4 presents the network architectures and the implementation details. Section 5 shows the experimental results and section 6 concludes the paper.

2. Related works

In this section, we describe the works relevant to our pipeline. These include a clustering algorithm, person re-identification task, a GAN model for unsupervised learning and a CNN model for semi-supervised learning.

2.1. Generative Adversarial Network

Generative Adversarial Network (GAN) was first introduced by Goodfellow et al. [18] and is 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. Radford et al. [32] 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 [6, 44, 62, 63]. GANs were applied to various interesting tasks such as realistic image generation [32], text-to-image generation [33]; video generation [41]; image-to-image generation [20], image inpainting [31], super-resolution [24] and many more. In this work, we use DCGAN [32] model to generate unlabeled images from the training set. We decide to choose DCGAN model after carefully contrasting various image generators. DCGAN architecture is very simple but yet generates more realistic images as shown in Fig 2.

2.2. Supervised and semi-supervised learning

Supervised learning is a well-studied problem in computer vision for image classification. Given training data $\mathcal{X} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$ where $y^{(i)}$ is the corresponding label on data $x^{(i)}$, supervised learning representation learns the mapping function $Y = F(X|\Theta)$ or the posterior distribution $P(Y|X)$. In contrast, unsupervised learning learns the intrinsic structure from unlabeled data. Semi-supervised learning is regarded as an unsupervised learning with some constraints on labels, or a supervised learning with additional information on the data distribution. Researchers also treat unsupervised learning as a sub task to supervised learning [58]. Lee *et al*. [25] train a supervised network with labeled and unlabeled data by assigning *pseudo-label* to unlabeled data. Yu *et al*. [50] and Wang *et al*. [46] propose unsupervised asymmetric metric learning to unsupervised person re-id. Papandreou *et al*. [30] propose Expectation-Maximization (EM) combining weak and strong labels under supervised and semi-supervised settings for image segmentation and Zheng *et al*. [58] train a semi-supervised network and assign uniform label distribution to generated samples. We depart from [32, 38, 58, 62] and propose to train a network in a semi-supervised fashion using a combination of two losses.

2.3. Person Re-Identification

Person re-id is viewed as an image retrieval problem and started as a multi-camera tracking research [45]. Some early works on person re-id focus on learning a metric and emphasize inter-personal distances or intra-personal distances (KISSME [22] [55], XQDA [27], MLAPG [28], LFDA [49] and Similarity learning [8]). Other works such as SILTP [27], LBP [53] use Color Histograms, Color Names or a combination of them to address the challenge variations in illumination and pose view-point. Recent works in person re-id are CNN based, and the goal is to jointly learn the best feature representation and distance metric. Zheng *et al*. [57] propose a siamese network with verification loss and identification loss and predicted the identities of a pair of input images. Many unsupervised methods with GAN-generated data have been developed [50, 51, 61] to address the problem of lack of large labeled datasets in person re-id. Barros *et al*. [5] introduce, for the first time in the re-identification field, the strategy of using synthetic data as a proxy for the real data and claim to recognize people independently of their clothing. Zhedong *et al*. [58] show that a regularized method (LSRO) over GAN-generated data can improve person re-id. Zhong *et al*. [61] propose a camera style (CamStyle) adaptation method to regularize CNN training through the adoption of LSR and use CycleGAN [62] for image generation. We show in section 3.3 how our model differs from [58] and

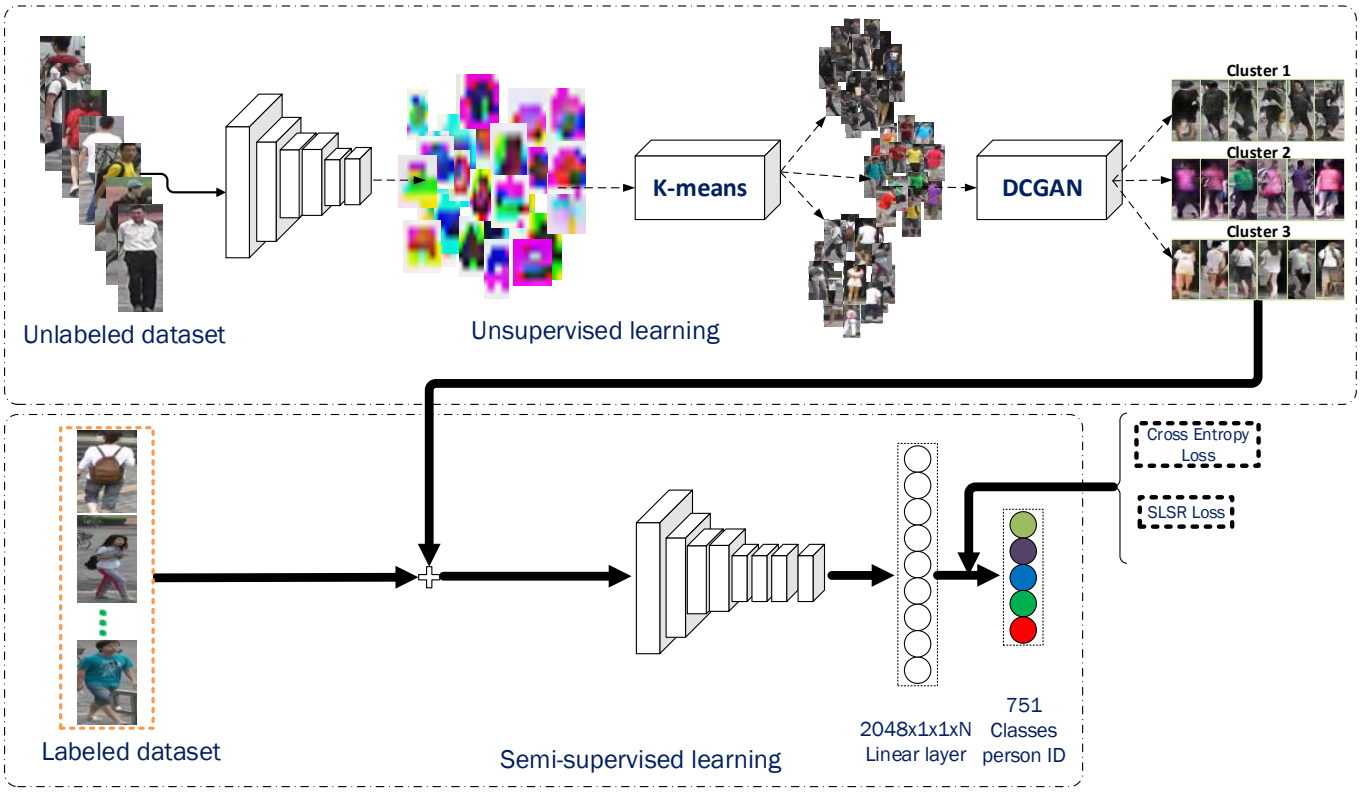

Figure 1. Our model consists of 3 steps: (1) Clustering on training data using unlabeled source dataset (Section 4.2). (2) For each cluster; train a DCGAN to generate images. Assign a partial label distribution to the generated images (Section 3). (3) Combine the partial labeled images with the training image .

[61].

3. Modeling

3.1. Unsupervised loss

We intend to partition the training sample into k groups of equal variance and find a share space among similar objects. Our goal is to produce k different clusters with relatively similar features. To do this, we define an objective function like that of *k-means* clustering [2, 14].

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} || \, z_i - \mu_k \, ||^2 \qquad (1)$$

where μ_k is a cluster center and $|| \, . \, ||$ the Euclidean distance between an embedded data z_i and the cluster center μ_k. If $\mathcal{F} = \{x_{(1)}^{(1)}, \ldots, x_{(n)}^{(m)}\} \in \mathbb{R}^{N \times M}$ denotes the set of feature vectors extracted from the last convolution layer and $\boldsymbol{I}$ the input image of shape $W \times H \times C$ (where C is the channel and $W \times H$ is the spatial size); a feature extraction function $\phi(\boldsymbol{x}_i; \theta)$ for N images is an $\phi : \mathbb{R}^{N \times W \times H \times C} \rightarrow \mathbb{R}^{N \times M}$ defined by $\phi(\boldsymbol{x}_i; \theta) = \boldsymbol{w}_k \cdot \boldsymbol{I} + b_k^I$. The weights or the parameters $(\boldsymbol{w}_k, b_k)$ are automatically learned from data. Minimizing Eq. 1 with respect to the network parameters θ results on:

$$\min_{\theta} \mathcal{L} = \sum_{i=1}^{N} \sum_{k=1}^{K} || \, \phi(\boldsymbol{x}_i; \theta) - \mu_k \, ||^2 \qquad (2)$$

where N is the number of cases, μ_k a centroid for cluster k. Learning the centroids such that, given a threshold γ, distances between similar vector $x_{(i)}^{(j)} \in \mathcal{F}$ are smaller than γ, while those between dissimilar vectors are greater than γ. Eq. 2 assures that the distance between each training sample and its assigned cluster center is small for each features $\mathcal{F}$. Using this objective function results in better clustering quality as shown in Fig. 3.

To generate realistic images, we define a loss function similar to [12] and minimize Eq. 3 with respect to the parameters of $G(z)$ while maximizing Eq. 3 with respect to

Figure 2. Samples of Generated Images and Original Images. The first two rows show the generated images by DCGAN. The last row shows the original images from Market-1501 training set.

the parameters of $D(x)$.

$$\mathcal{L}_{GAN} = \log D(x) + \log\Big(1 - D(G(z))\Big) \tag{3}$$

3.2. Semi-supervised loss

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 for label y_i. The network computes the probabilities of each label y_i:

$$p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^T \cdot \boldsymbol{x}_i)}{\sum_{k=1}^{N} \exp(\boldsymbol{w}_k^T \cdot \boldsymbol{x}_i)} \tag{4}$$

where $\boldsymbol{x}_i$ refers to the input vector from the previous layers, Given $k \in \{1, 2, \ldots, K\}$ the class labels for N training samples, we define the cost function for real images as the negative log-likelihood:

$$\mathcal{L}_{Entropy} = -\sum_{i=1}^{K} \log p(\tilde{y}_i = y_i|\boldsymbol{I}_i) \tag{5}$$

In general, neural network represents a function $f(x;\theta)$ which provides the parameters $\boldsymbol{w}$ for a distribution over y. So minimizing $\mathcal{L}_{Entropy}$ 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, Eq. 5 can be written as

$$\mathcal{L}_{Entropy}(\theta) = -\log p(y|\boldsymbol{x};\theta) \tag{6}$$

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

Regularization via Label Smoothing (LSR) Szegedy *et al*. [38] propose 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:

$$\delta_{k,y} = \begin{cases} 1 & k = y \\ 0 & k \neq y \end{cases} \tag{7}$$

For training image with ground-truth label y, Szegedy *et al*. [38] replace the label distribution $q(k|x) = \delta_{k,y}$ with

$$q\prime(k,y) = \begin{cases} (1-\epsilon)\delta_{k,y} & k = y \\ \frac{\epsilon}{K} & k \neq y \end{cases} \tag{8}$$

Departing from LSR [38], we introduce our loss function for semi-supervised learning as a combination of cross entropy (Eq. 6) and a modified version of LSR. Given I

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

Here, $z_{i,k}$ are the unnormalized probabilities of the ith image generated from cluster $\mathcal{C}$ with K classes. z_i represents a one-hot 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 $\boldsymbol{I}_i$ and normalize z_i so that $\sum_{k=1}^{K} z_{i,k} = 1$. To explicitly take into account our label regularization for $\boldsymbol{I}_i$, we change the network to produce

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

and we optimize $\sum_{i,k} \mathcal{L}(\tilde{z}_i, \frac{1}{k} z_{i,k})$ where k is the number of class label in cluster $\mathcal{C}$. Our loss for generated images can then be written as:

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

or simply written as

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

Combining Eq. 6 and Eq. 12, the proposed loss function $\mathcal{L}_{SLSR}$ is defined as:

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

Where K is the number of classes. For training images, we set $\lambda = 0$ and for the generated images, $\lambda = 1$

3.3. Discussion

Recently, Zheng *et al*. [58] propose Label Smoothing Regularization for Outliers (LSRO) and Zhong *et al*. [61]

propose CamStyle as a data augmentation approach. LSRO expands the training set with unlabeled samples generated by DCGAN [32] and assigns uniform LSR [38] to the generated samples i.e. $\mathcal{L}_{LSR}(\epsilon = 1)$ while CamStyle uses CycleGAN [62] 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 similars to our work, we argue that our method is different on two aspects:

1) LSRO [58] and CamStyle [61] fuse equal distribution to all generated images; this can leads to an over-smooth especially when the number of classes is excessively large. Our method however fuses generated images with adaptive label distribution over each cluster i.e $\mathcal{L}_{LSR}(\epsilon = \frac{1}{k_i})$ where k_i is the class set size of cluster i. In LSRO and CamStyle, dissimilar and similar images may be assigned relatively equal similarity value, while our method deals with such unfairness by considering generated images in the locality of each sample and propose a strategy to determine the appropriate candidates by using *k-means* clustering algorithm. The proposed SLSR 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-smooth similarity.

2) In our model, similarities are maintained and propagated through the network 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. Fig. 3 illustrates the effectiveness of our method and extensive experiments demonstrate the superiority of our method compared to LSRO [58] and CamStyle [61]. Our model 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 [18] models; (2) the adoption of Label Smooth Regularization (LSR) proposed by Szegedy *et al*. [38] to alleviate the impact of noise introduced by the generated images; (3) performing semi-supervised learning for person re-id using labeled and unlabeled data in a CNN-based approach.

4. Network Overview

4.1. Generative Adversarial Network

We follow the implementation details of [32]. 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 filter size and a stride of 2. We add a linear layer followed by $sigmoid$ 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 deconvolution layer is followed by a batch normalization [19] and *ReLU* in both the generator and discriminator.

Algorithm 1 Algorithm for SLSR Training

Input: $\mathcal{K}$: Number of clusters, $\mathcal{X}$: Training sample
Initialisation : initialize cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly
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 Eq. 6
for each sample *m* **do**
  Extract $x^{(n)}_{(m)}$ 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** $c^{(i)} := arg\min_j \mid\mid x^{(i)} - \mu_j \mid\mid$
  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 Eq. 2
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 Eq. 3
  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 Eq. 12

4.2. Clustering

It is well known that multiview data object admits a common clustering structure across view [23] and that person re-id is a cross-camera retrieval task across view. We aim at exploring such clustering structure propriety to generate images that model the correlation among similar views through the use of *k-means* and GAN. We apply *k-means* algorithm to cluster the training images into *k* clusters $(2, \ldots, 5)$. *K-means* clustering is a simple yet very effective unsupervised learning algorithm for data cluster-

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

Table 1. For each cluster size, we calculate the silhouette coefficient [35] 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 2 prove that we achieve higher accuracy for $k = 3$, in Market-1501 dataset

Figure 3. Generated samples from three cluster sets using DCGAN.We show that Identities with similar appearances are in the same cluster. We find that the color is the major learned feature for clustering

ing. It clusters data based on the Euclidean distance between data points. We train for 40 epochs a CNN network using a learning rate of 0.001 with a momentum of 0.9. We use ResNet50 [15] model to learn 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 goodness of our clustering algorithm, we consider the ground truth not known and perform an evaluation using the model itself. Table 1 shows the cluster quality metric Silhouette Coefficient [35] applied on Market-1501 dataset [55].

4.3. Convolutional Neural Network

We fine-tuned two baseline models Resnet50 [15] and DenseNet [17] pre-trained on ImageNet [36], we introduce an extra linear layer into the network which adapts the network outputs to match the noisy GAN label distribution i.e. $2048 \times 1 \times 1$ and $1024 \times 1 \times 1$ linear layer in Resnet50 and DenseNet baselines respectively. 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 $ReLU$.

5. Experiments

5.1. Person Re-ID datasets

We intensively evaluate our proposed model on four widely used datasets including Market-1501, CUHK03, DukeMTMC-ReID and VIPeR.

Market-1501 [55] is a large and most realistic dataset collected in front of a campus supermarket. It contains overlapping among the six cameras and images were automatically detected by the deformable part model (DPM) [10]. 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 [55] and use all the training set for the unsupervised step and one image per identity as validation image in the semi-supervised step.

CUHK03 [26] contains $13,164$ images and $1,467$ identities. The dataset provides two image sets, one set is automatically detected by the deformable-part-model detector DPM [10], 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 to make our model more realistic. The dataset is captured by six cameras, and each identity has an average of 4.8 images in each view.

DukeMTMC-ReID [58] is a dataset derived from the DukeMTMC [34] 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 of Zhedong *et al*. [58]. 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 unsupervised learning and randomly pick one image per identity for the validation set. The remaining images are used for the supervised learning step.

VIPeR [13] contains 632 pedestrian image pairs captured outdoor from two viewpoints. Each pair contains two images of the same individual cropped and scaled to 128x48 pixels. The datasets are divided into two equal subsets. To be fair in the comparison, we follow the testing strategy as defined in [13], [54]

5.2. Implementation details

We use Resnet50 [15] and DenseNet [17] as baselines, and modify the last fully connected layer with the number of classes i.e. 751; $1,367$ and 702 units for Market-

Cluster size	K = 2				K = 3				K = 4			
Generated	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
12,000	93.41	96.61	**97.62**	90.15	**93.82**	**96.73**	97.42	**90.20**	93.02	96.17	97.42	89.89
18,000	92.87	96.02	97.12	89.59	92.96	95.96	97.06	89.54	93.32	96.23	97.18	89.26
24,000	93.20	95.62	97.03	89.41	92.70	96.05	96.94	89.01	92.42	96.99	96.94	88.99
30,000	93.02	95.99	97.09	89.51	93.08	96.02	96.91	89.00	92.66	96.02	97.09	88.65
36,000	92.31	95.87	96.94	88.51	92.78	96.26	97.12	88.13	92.27	95.81	96.97	88.55

Table 2. Impact of the number of cluster in 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 in 3 different values of k by use Re-rank [59] with k-reciprocal encoding evaluation in single query setting. The best results with $k = 3$ is used for later experiments for all the datasets

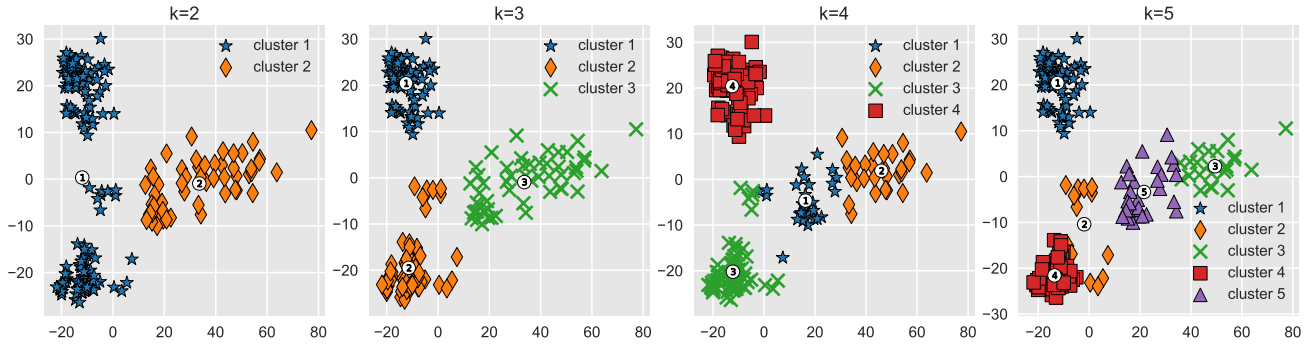

Figure 4. Visualization of extracted feature map $\mathcal{F}$ from ResNet50 on Market1501 dataset. Results of *k-means* clustering algorithm on $\mathcal{F}$ for $k = 2, \ldots, 5$. We arrive at a fair clustering view with $k = 2$ and $k = 3$. Best viewed in color.

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

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

1501, CUHK03 and DukeMTMCReID respectively. To train the network, we use stochastic gradient descent [7] and start with a base learning rate of $\eta^{(0)} = 0.01$ and gradually decrease it as the training progresses using $\eta^{(i)} = \eta^{(0)}(1+\gamma \cdot i)^{-p}$, where $\gamma = 0.1$, $p = 0.025$ and i is the current mini-batch iteration. We use a momentum of $\mu = 0.9$ and weight decay of $\lambda = 5 \times 10^{-4}$ and the mini-batch size of 32. We train the network for 130 epochs. To generate image samples, we train DCGAN for 30 epoch using Adam [21] with learning rate $lr = 0.0002$ and $\beta_1 = 0.5$.

Data preprocessing: For DenseNet baseline, all the input images are resized to 288×144 before being randomly cropped into 256×128 with random horizontal flip. We scale the pixels between 0 and 1. For Resnet50 baseline, input images are resized to 256×256 before being randomly cropped into 224×224 with random horizontal flip; we also scale the pixels in the range of 1 and -1. Zero-center by mean pixel and random erasing [60] are finally applied to both baselines to make the network more robust to variations and occlusions.

5.3. Evaluations

We use Cumulated Matching Characteristics (CMC) and mean average precision (mAP) as defined in [55] to evaluate the performance of our model. We use the L2 Euclidean distance to compute a similarity score for ranking or retrieval task as in previous works [47, 57, 58].

Re-ranking: Recent works [4, 43, 59] choose to perform an additional re-ranking to improve ReID accuracy.

In this work, we report re-ranking results based on Zhong *et al*. [59] method with k-reciprocal encoding, which combines the original L2 distance and Jaccard distance.

SLS+DenseNet and SLS+ResNet represent our method with DenseNet and ResNet baseline models respectively. SLS+Rerank is our DenseNet model with re-ranking [59].

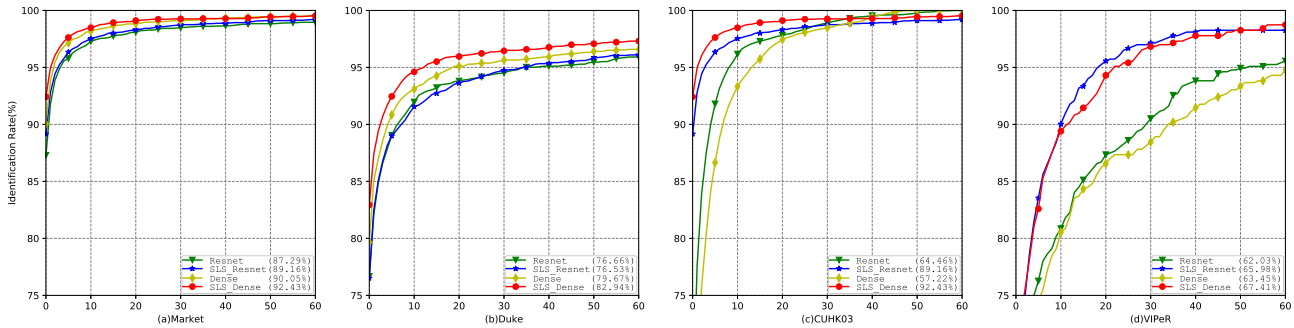

Figure 5. CMC curve for (a)Market-1501, (b) DukeMTMCReID, (c) CUHK03 and (d)VIPeR datasets. Comparison of the baselines with our own variations of deep architectures. For instance, in VIPeR dataset, our models SLS˙Resnet and SLS˙Denset repectively improve the baselines Resnet and DensetNet by a large margin.

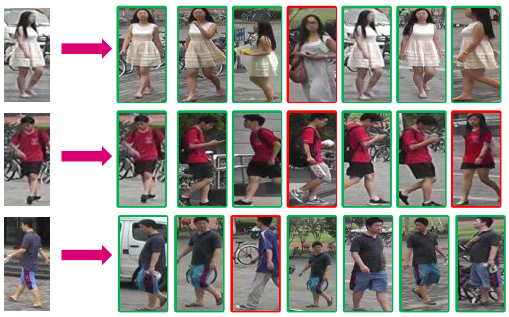

Figure 6. Sample image retrieval on Market-1501 dataset using our framework. The images in the first column are the query images. The images in the right columns are the retrieved images. The retrieval images are sorted according to the similarity scores from left to right. We use re-ranking [59] with k-reciprocal enconding.

We only report rank 1, 5 and 10 accuracy.

Methods	R1	R5	R10	mAP
BoW+KISSME [55]	25.13	-	-	12.17
XQDA (LOMO) [27]	30.75	-	-	17.04
LSRO [58]	67.68	-	-	47.13
OIM [48]	68.1	-	-	47.4
TriNet [16]*	72.44	-	-	53.50
SVDNet [37]	**76.7**	86.4	89.9	56.8
ResNet Basel.	76.66	87.83	91.47	58.35
DenseNet Basel.	79.67	89.85	92.86	63.19
(Ours) SLS+ResNet	76.53	**88.15**	**91.02**	**60.79**
(Ours) SLS+DenseNet	**82.94**	**91.69**	**94.43**	**67.78**
(Ours) SLS+Rerank	**86.66**	**92.91**	**94.97**	**83.35**

Table 6. Comparison results of the state-of-arts methods on DukeMTMCReID. We add 12,000 generated images to the train set. We show that our methods is superior to previous works. * paper on ArXiv but not published

5.3.1 Comparison with the state of art

In this section, we compare our results with state-of-art methods in Tables 4 5 6 7.

On **Market-1501** dataset our method achieves an **89.16%** rank 1 accuracy and **75.15%** mAP accuracy exceeding LSRO [58] by **5.19%** and **9.08%** respectively. Our method with both SLSR and re-ranking [59] with k-reciprocal encoding further improves rank 1 and mAP accuracy to **93.82%** and **90.20%** respectively. Table 4 shows that our method outperforms previous works globally.

On **CUHK03** dataset, we achieve a **91.03%** rank 1 accuracy and **94.21%** mAP accuracy which are close by **0.77%** to the best result reported by HydraPlus-Net [29]. Our method exceeds LSRO [58] by **6.41%** and **6.81** on rank 1 and mAP respectively. Table 5 shows that our method outperforms previous.

On **DukeMTMCReID** dataset, not many reported results exist on this dataset as shown in Table 6. Yet, our method achieves an **82.94%** rank 1 accuracy and **67.78%** mAP accuracy exceeding existing works. Compared to LSRO [58], our ResNet rank 1 accuracy exceeds their result by **8.85%**. SVDNet [37] exceeds our ResNet model by **0.17%**; but our model with DenseNet still exceeds their result by **6.24%**.

On **VIPeR** dataset, our method achieve a **67.41%** and **65.98%** rank 1 accuracy with DenseNet and ResNet respectively. We improve the baseline by **3.95%** for rank 1 accuracy and achieve competitive results for rank 5, 10 and 20.

Compared to previous works in general, our method (SLSR) boots **1.23%∼6.41%** rank 1 accuracy and **1.43%∼6.81%** mAP on all datasets.

6. Conclusion

In this paper, we propose sparse label smoothing regularization for person re-identification. We use unsupervised learning to do clustering on unlabeled data. For each cluster

	Single Query			
Methods	R1	R5	R10	mAP
BoW+KISSME [55]	44.42	-	-	20.76
FisherNet [47]	48.15	-	-	29.94
Simil.Learning [8]	51.90	-	-	26.35
DiscNullS [52]	61.02	-	-	35.68
Gate Reid [40]	65.88	-	-	39.55
MR B-CNN [39]	66.36	85.01	90.17	41.17
Cross-GAN [51]*	72.15	-	94.3	48.24
SOMAnet [5]	73.87	88.03	92.22	47.89
HydraPlus-Net [29]	76.9	91.3	94.5	-
Verif.Identif [57]	79.51	-	-	59.87
SVDNet [37]	82.3	92.3	95.2	62.1
DeepTransfer [11]*	83.7	-	-	65.5
LSRO [58]	83.97	-	-	66.07
TGP-ReID [3]*	92.2	97.9	-	81.2
ResNet Basel.	87.29	95.57	96.94	69.70
DenseNet Basel.	90.05	96.82	98.10	74.16
(Ours) SLS+ResNet	**89.16**	**95.78**	**97.33**	**75.15**
(Ours) SLS+DenseNet	**92.43**	**97.27**	**98.39**	**79.08**
(Ours) SLS+Re-rank	**93.82**	**96.73**	**97.42**	**90.20**
	Multi Query			
DiscNullS [52]	71.56	-	-	46.03
Gate Reid [40]	76.04	-	-	48.45
SOMAnet [5]	81.29	92.61	95.31	56.98
Verif.Identif [57]	85.47	-	-	70.33
LSRO [58]	88.42	-	-	76.10
DeepTransfer [11]*	89.6	-	-	73.80
TGP-ReID [3]*	94.7	98.6	-	87.3
ResNet Basel.	91.27	96.85	98.19	76.94
DenseNet Basel.	92.90	97.89	98.69	81.22
(Ours) SLS+ResNet	**92.25**	**97.51**	**98.34**	**81.92**
(Ours) SLS+DenseNet	**94.06**	**98.16**	**98.84**	**85.20**

Table 4. Comparison results of the state-of-art methods on Market-1501. We add 12,000 generated images to the train set. '-' means that no reported results is available and '*' means the paper is available on ArXiv but not published

set, we train a GAN to generate images similar to the cluster set. We derive our strategy based on the intuition that each image represent a point in some high-dimensional feature space, and that similar images are close point and share the same feature space, sufficient to be assigned similar label according to their cluster. We use *k-means* clustering algorithm to partition similar images from dissimilar images and assign SLS to generated images. We finally train a CNN baseline using our SLSR loss function. Our model learns to exploit the samples generated by DCGAN to boost the performance of the person re-id by improving generalization. Extensive evaluations were conducted on four large-scale datasets to validate the advantage of the proposed model on

Methods	R1	R5	R10	mAP
KISSME [22]	11.7	33.3	48.0	-
DeepReID [26]	19.89	50.00	64.00	-
ImprovedDeep [1]	44.96	76.01	83.47	-
XQDA (LOMO) [27]	46.25	78.90	88.55	-
SI-CI [42]	52.20	84.30	94.8	-
DiscNullS [52]	54.7	80.1	88.30	-
FisherNet [47]	63.23	89.95	92.73	44.11
MR B-CNN [39]	63.67	89.15	94.66	-
Gated ReID [40]	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 [37]	81.8	95.2	97.2	84.8
Cross-GAN [51]*	83.23	-	96.73	-
Verif.Identif. [57]	83.40	97.10	98.7	86.40
DeepTransfer [11]*	84.10	-	-	-
LSRO [58]	84.62	97.60	98.90	87.40
TriNet [16]	87.58	98.17	-	-
HydraPlus-Net [29]	**91.8**	**98.4**	99.1	-
ResNet Basel.	75.11	94.97	97.87	83.91
DenseNet Basel.	67.55	90.78	95.26	77.85
(Ours) SLS+ResNet	91.03	98.22	**99.26**	**94.21**
(Ours) SLS+DenseNet	83.61	96.95	98.88	**89.47**
(Ours) SLS+Rerank	**92.95**	**97.79**	**99.27**	**95.10**

Table 5. Comparison result with state-of-arts on CUHK03. We add 18,000 generated images to the training set and used single query setting on the detected subset. '-' means that no reported results is available. * paper on ArXiv but not published

Methods	R1	R5	R10	R20
ImproveDeep [1]	34.81	63.61	75.63	84.49
KISSME [22]	34.81	60.44	77.22	86.71
Simil.Learning [8]	36.80	70.40	83.70	91.70
MFA (LOMO) [49]	38.67	69.18	80.47	89.02
XQDA (LOMO) [27]	40.00	68.13	80.51	91.08
TCP [9]	47.8	74.7	84.8	91.1
Cross-GAN [51]*	49.28	-	**91.66**	93.47
DiscNullS [52]	51.17	**82.09**	90.51	95.92
SSM [4]	53.73	-	91.49	**96.08**
SpindleNet [54]	53.80	74.1	83.2	92.1
HydraPlus-Net [29]	56.6	78.8	87.0	92.4
ResNet Basel.	62.03	75.00	80.22	86.71
DenseNet Basel.	63.45	72.78	79.11	86.23
(Ours) SLS+ResNet	**65.98**	81.49	88.45	95.25
(Ours) SLS+DenseNet	**67.41**	81.01	88.61	93.51

Table 7. Comparison results with state-of-arts on VIPeR dataset. 8,000 generated images are added to the train set. We provide results of the ResNet and DenseNet fine-tuned baselines. * paper on ArXiv but not published

existing models. Tables 4 5 6 7 show the superiority of the model over a wide variety of state-of-art methods.

7. Acknowledgements

This work is supported by the Ministry of Science and Technology of Sichuan province (Grant No. 2017JY0073) and Fundamental Research Funds for the Central Universities in China (Grant No. ZYGX2016J083). We 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 *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3908–3916, June 2015.

[2] E. Aljalbout, V. Golkov, Y. Siddiqui, and D. Cremers. Clustering with Deep Learning: Taxonomy and New Methods. *ArXiv e-prints*, Jan. 2018.

[3] J. Almazán, B. Gajic, N. Murray, and D. Larlus. Re-id done right: towards good practices for person re-identification. *CoRR*, abs/1801.05339, 2018.

[4] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3356–3365, July 2017.

[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. *Computer Vision and Image Understanding*, 167:50 – 62, 2018.

[6] D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary Equilibrium Generative Adversarial Networks. *ArXiv e-prints*, Mar. 2017.

[7] L. Bottou. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[8] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1268–1277, June 2016.

[9] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, June 2016.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010.

[11] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep Transfer Learning for Person Re-identification. *ArXiv e-prints*, Nov. 2016.

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[13] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 09/2007 2007.

[14] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[16] A. Hermans, L. Beyer, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *ArXiv e-prints*, Mar. 2017.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] G. J. Ian, P.-A. Jean, M. Mehdi, X. Bing, S. O. David, C. Aaron, and B. Yoshua. Generative adversarial network. In *NIPS*. The Neural Information Processing Systems, 2014.

[19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv e-prints*, Nov. 2016.

[21] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[22] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, June 2012.

[23] A. Kumar, P. Rai, and H. Daumé, III. Co-regularized multi-view spectral clustering. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 1413–1421, USA, 2011. Curran Associates Inc.

[24] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *ArXiv e-prints*, Sept. 2016.

[25] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 07 2013.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, June 2014.

[27] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, June 2015.

[28] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3685–3693, Dec 2015.

[29] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[30] 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 *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1742–1750, 2015.

[31] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv e-prints*, Nov. 2015.

[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. *ArXiv e-prints*, May 2016.

[34] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, abs/1609.01775, 2016.

[35] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[37] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3820–3828, Oct 2017.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, July 2016.

[39] E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion Bilinear Convolutional Neural Networks for Person Re-Identification. *ArXiv e-prints*, Dec. 2015.

[40] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[41] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 613–621, USA, 2016. Curran Associates Inc.

[42] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1288–1296, June 2016.

[43] J. Wang, S. Zhou, J. Wang, and Q. Hou. Deep ranking model by large adaptive margin learning for person re-identification. *Pattern Recognition*, 74:241 – 252, 2018.

[44] W. Wang, Q. Huang, S. You, C. Yang, and U. Neumann. Shape Inpainting using 3D Generative Adversarial Network and Recurrent Convolutional Networks. *ArXiv e-prints*, Nov. 2017.

[45] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34:3–19, 2013.

[46] 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 Transactions on Neural Networks and Learning Systems*, 28(1):57–70, Jan 2017.

[47] L. Wu, C. Shen, and A. Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. 65, 06 2016.

[48] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[49] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Computer Vision – ECCV 2014*, pages 1–16, Cham, 2014. Springer International Publishing.

[50] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[51] C. Zhang, L. Wu, and Y. Wang. Crossing Generative Adversarial Networks for Cross-View Person Re-identification. *ArXiv e-prints*, Jan. 2018.

[52] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1239–1248, June 2016.

[53] Y. Zhang and S. Li. Gabor-lbp based region covariance descriptor for person re-identification. In *2011 Sixth International Conference on Image and Graphics*, pages 368–371, Aug 2011.

[54] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. *Conference on Computer Vision and Pattern Recognition*, pages 907–915, 07 2017.

[55] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, Dec 2015.

[56] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3346–3355, July 2017.

[57] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *ACM Transactions on Multimedia Computing Communications and Applications*, 2017.

[58] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[59] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[60] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random Erasing Data Augmentation. *ArXiv e-prints*, Aug. 2017.

[61] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[62] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *ArXiv e-prints*, Mar. 2017.

[63] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017.