Person Re-identification through Clustering and Partial Label Smoothing Regularization

Jean-Paul Ainam, Ke Qin, Guangchun Luo
qinke@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 new label smoothing regularization scheme for person re-identification. We first use an unsupervised method for discriminative learning representation. We apply a clustering algorithm on the learned feature to partition the training set into k groups of equal variance and derive a shared space for similar images. Secondly, a GAN model is fed with each cluster to produce samples with relatively similar features to the original space. Our method consists of assigning an adaptive smooth label distribution to each generated sample according to their original cluster. To train our model, we define a new objective function which takes into account the generated samples and fine-tuned a CNN baseline using the objective function. Our model learns to exploit the samples generated by the GAN model 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.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; **Matching**; *Ranking*; Cluster analysis;

KEYWORDS

Person ReID; Pedestrian Retrieval; GAN; Smooth Label; Unsupervised Learning

ACM Reference Format:
Jean-Paul Ainam, Ke Qin, Guangchun Luo. 2019. Person Re-identification through Clustering and Partial Label Smoothing Regularization. In *Proceedings of ACM International Conference on Big Data and Smart Computing (ICBDSC '19).* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Person re-ID is a task of building up correspondence between persons from various cameras and deciding if a given person has been seen by another camera. The issue has been broadly examined in the past and has achieved phenomenal results with machine learning based approach [2, 8, 38, 40]. Current deep learning approaches require a huge volume of labeled data for training. Such large dataset are not common in person re-identification; making it difficult to apply deep learning methods. One way this can be mitigated 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 [21] 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.

Our framework consists of three main steps. In the first step, we train a CNN model to learn feature representation and extract high dimensional vectors representing the feature maps of the training images. The extracted feature map is fed into a *k-means* clustering algorithm to separate similar images from dissimilar images. In the second step, each cluster set is used to train an image generator and output sample images with relatively similar feature representation. We then assign a label to generated samples using our regularization method. In the final step, we define a new loss function and introduces a noise linear layer into existing architecture, to adapt the network outputs against the noise GAN label distribution. Our model generalize well, and experimental results show that it outperforms previous state-of-art methods. The summary of our contributions are:

- the introduction of a new label smoothing regularization scheme for person re-id task.
- the adoption of clustering design over the feature map representation and a Partial Label Smoothing Regularization (PLSR) over generated images.
- and finally a semi-supervised learning representation design with PLSR that improves the person re-identification accuracy.

1.1 Generative Adversarial Network

Introduced by Goodfellow *et al.* [11], a GAN model consists of two different components: a generator (G) that generates an image and a Discriminator (D) that discriminates real images from generated images. They compete following the minimax two-player game. Radford et al. 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 for

ICBDSC '19, January 2019, Bali, Indonesia

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

realistic image generation [4, 21, 28, 43, 44], text-to-image generation [22]; video generation [27]; image-to-image generation [12], image inpainting [20], super-resolution [15] and many more were published. In this work, we use DCGAN [21] model to generate unlabeled images from the training set.

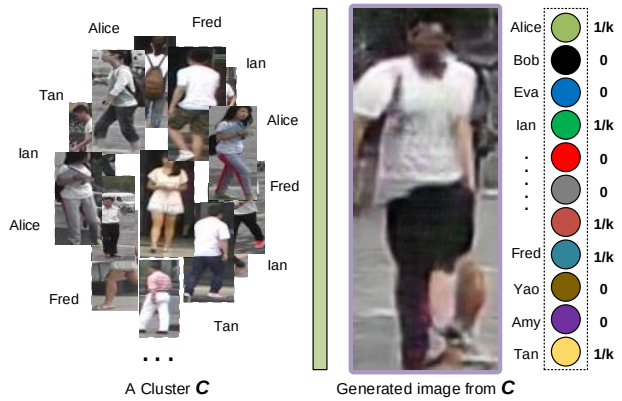

Figure 1: Our method: Clustering and Partial label smoothing distribution over generated samples

1.2 Person Re-Identification

Early works on person re-id, KISSME [14] [37], XQDA [17], MLAPG [18], LFDA [31], Similarity Learning [5], SILTP [17] and LBP [35] are based on metric distance and aim at learning inter-personal or intra-personal distances. However, recent works are CNN based with a goal to learn the best feature representation and distance metric. Furthermore, recent works like [3, 32, 33, 39, 42] are CNN and GAN based. Similar to our work, Zhedong *et al.* [40] show that a regularized method (LSRO) over GAN-generated data can improve person re-id and propose to assign uniform label distribution to unlabeled data. In addition, Zhong *et al.* [42] propose a camera style (CamStyle) adaptation method to regularize CNN training through the adoption of LSR and use CycleGAN [43] for image style generation. We show in section 2.3 how our model differs from [40] and [42].

2 MODELING

2.1 Unsupervised loss

We partition the training into k group objects with relatively similar features. To do this, we define an objective function $\mathcal{L}(\theta)$ such that.

$$\mathcal{L}(\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} || \mathcal{F}_i - \mu_k ||^2 \tag{1}$$

where $\mathcal{F} = \{x_{(1)}^{(1)}, x_{(2)}^{(2)} \ldots, x_{(n)}^{(m)}\} \in \mathbb{R}^{N\times M}$ represents the high dimensional feature vectors extracted from the last convolution layer given an input image $\boldsymbol{I}$ with $W \times H \times C$ shape dimension (C is the channel and $W \times H$ is the spatial size); μ_k is a cluster centroid and $||\ .\ ||$ the Euclidean distance between a feature data $x_{(i)}^{(i)}$; and N the number of cases.

Eq. 1 assures that the distance between each training sample and its assigned cluster center is small for each features $\mathcal{F}$. Using this objective function resulted in better clustering quality.
For DCGAN training, we define a loss function similar to [11] and minimize Eq. 2 with respect to the parameters of $G(z)$ and maximize Eq. 2 with respect to the parameters of $D(x)$.

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

2.2 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 $\boldsymbol{I}_i$. We define the cost function for real images as the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{K} \log p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = -\log p(y|\boldsymbol{x};\theta) \tag{3}$$

In general, neural network represents a function $f(x;\theta)$ which provides the parameters $\boldsymbol{w}$ for a distribution over y. So minimizing $\mathcal{L}(\theta)$ is equivalent to maximizing the probability of the ground-truth label $p(\tilde{y}_i = y_i|\boldsymbol{I}_i)$. where $\boldsymbol{\theta}$ represents the set of parameters of the network.

2.2.1 Label-smoothing regularization (LSR). Szegedy *et al.* [25] introduce a mechanism to regularize a layer by estimating the marginalized effect of label-dropout during training.

$$H(q\prime, q) = -\sum_{k=1}^{K} \log p(k) q\prime(k) \quad \text{and} \quad q\prime(k) = (1-\epsilon)\delta_{k,y} + \frac{\epsilon}{K} \tag{4}$$

where $\delta_{k,y}$ is Dirac delta.

Based on on [25] work, we introduce our new loss function for semi-supervised learning as a combination of cross entropy loss Eq. 3 and a modified version of LSR defined as followed.

Given $\boldsymbol{I}, z_{i,k} = 1$ if $\boldsymbol{I}_i \in C$ and $z_{i,k} = 0$ if $\boldsymbol{I}_i \notin C$. Here, $z_{i,k}$ are the unnormalized probabilities of the ith image generated from cluster 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 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

$$z_i = \frac{1}{k} z_{i,k} \qquad \text{for} \quad k \in \{1, 2, \ldots, K\} \tag{5}$$

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 C. Our loss for generated images is defined as:

$$\mathcal{L}_{PLS}(\theta) = -\sum_{i=1}^{K} \log p(\tilde{z}_i = z_i|\boldsymbol{I}_i) = -\log(p(z|\boldsymbol{x};\theta) \tag{6}$$

Combining Eq. 3 and Eq. 6, the proposed objective function $\mathcal{L}_{PLSR}$ is characterized by:

$$\mathcal{L}_{PLSR}(\theta) = -(1-\mathcal{Z}) \log\Big(p(y|\boldsymbol{x};\theta)\Big) - \frac{\mathcal{Z}}{K} \log\Big(p(z|\boldsymbol{x};\theta)\Big) \tag{7}$$

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

2.3 Discussion

Recently, Zheng *et al.* [40] propose Label Smoothing Regularization for Outliers (LSRO) while Zhong *et al.* [42] propose CamStyle. Although similar to our moldel, they differ from our model in two aspects.

First, the two models [40, 42] assign uniform label distribution to the generated images, i.e., equal probability distribution $\mathcal{L}_{LSR}(\epsilon = 1)$ for LSRO and $\mathcal{L}_{LSR}(\epsilon = 0.1)$ for CamStyle. Whereas our method assigns nonuniform label distribution for the generated images, i.e. $\mathcal{L}_{LSR}(\epsilon = \frac{1}{k_i})$ where k_i is the class set size of cluster i. In other words, an equal distribution to all generated images leads to an over-smooth when the number of classes is excessively large, but assigning adaptive label distribution based on their similarities is the best way to with such unfairness introduce by LSRO and CamStyle. Consequently, this strategy enables our model to be highly efficient in dealing with large amount of data. Our method *PLSR* learns the most discriminative features and can easily avoid the over-smooth similarity.

Second, thanks to *k-means* clustering algorithm, our model is able to keep the similarities and spread feature space through the generation of cluster images to improve the person re-identification accuracy. Compared to LSRO and CamStyle, we introduce an extra noise layer to match the noisy distribution introduced by the generated images. The parameters of this linear layer can be estimated as part of the training process and involve simple modification of current deep network architectures.

However, LSRO, CamStyle and our method share common practices such as (1) leveraging the training set by the generation of sample images using GAN models; (2) adopting Label Smooth Regularization (LSR) to alleviate the impact of noise introduced by the generated images; (3) finally, performing semi-supervised learning for person re-id using labeled and unlabeled data in a CNN-based approach.

3 EXPERIMENTS

Clustering: We use a CNN [1] model to learn good intermediate representation of the training set, extract high dimension feature representation from the last convolution layer and apply *k-means* algorithm to cluster the training set into k groups $(2, \ldots, 5)$ of similar images. *K-means* clustering algorithm is applied directly on feature maps. We found this way to be faster and better than clustering on raw data images.

To judge the goodness of our clustering algorithm, we performed a cluster quality metric [2] on a dataset and found the score higher for cluster $size = 3$. As a result, we use $k = 3$ for all the remaining experiments.

Generative Adversarial Network: Our DCGAN model follows the implementation details of [21]. The Generator G consists of four deconvolution operations with 5×5 filter size and a stride of 2. 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 $128 \times 128 \times 3$. Similarly, the Discriminator D consists of four convolution operations with 5×5 filter size and a stride of 2. We add a linear layer followed by a *sigmoid* to discriminate real images against fake images. The input shape of D includes sample images from G and real images from the training set. Each convolution and deconvolution layer is followed by a batch normalization and *ReLU* in both the generator and discriminator.

[1] In this work, we use ResNet model, but any other CNN could be considered instead

[2] Silhouette Coefficient [23]

Table 1: Dataset split details. The total number of images (*QueryImgs, GalleryImgs, TrainImgs*), together with the total number of identities (*TrainID, TestID*) are listed. 12, 000 **generated samples were added to each training**

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 2: Comparison results on Market-1501.

Methods	Single Query		Multi Query	
	R1	mAP	R1	mAP
DNS [34]	61.02	35.68	71.56	46.03
Gate Reid [26]	65.88	39.55	76.04	48.45
SOMAnet [3]	73.87	47.89	81.29	56.98
Verif.Identif [39]	79.51	59.87	85.47	70.33
DeepTransfer [8]*	83.7	65.5	89.6	73.80
LSRO [40]	83.97	66.07	88.42	76.10
(Ours) PLSR	**89.16**	**75.15**	**92.25**	**81.92**

Table 3: Comparison results with state-of-arts on CUHK03. We use single query setting and the detected subset

Methods	R1	R5	R10	mAP
Gated ReID [26]	68.1	88.1	94.6	58.8
SOMAnet [3]	72.40	92.10	95.80	-
SVDNet [24]	81.8	95.2	97.2	84.8
Verif.Identif. [39]	83.40	97.10	98.7	86.40
LSRO [40]	84.62	97.60	98.90	87.40
(Ours) PLSR	**91.03**	**98.22**	**99.26**	**94.21**

Table 4: Comparison results on DukeMTMCReID.

Methods	R1	R5	R10	mAP
BoW+KISSME [37]	25.13	-	-	12.17
XQDA (LOMO) [17]	30.75	-	-	17.04
LSRO [40]	67.68	-	-	47.13
OIM [30]	68.1	-	-	47.4
TriNet [10]*	72.44	-	-	53.50
SVDNet [24]	**76.7**	86.4	89.9	56.8
(Ours)PLSR	**76.53**	**88.15**	**91.02**	**60.79**

Figure 2: Rank accuracy for Market-1501, DukeMTMCReID, VIPeR and CUHK03 datasets. Comparison with the state of art result. Our model outperforms number of state-of-art results.

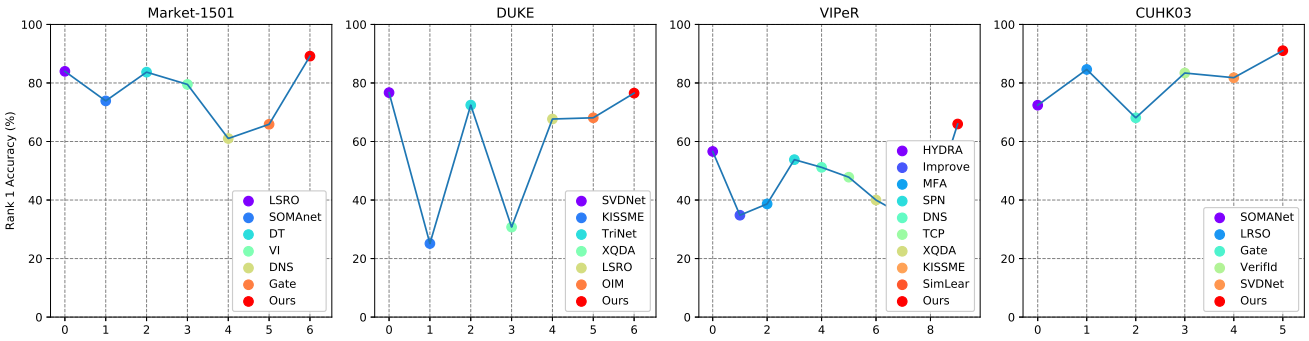

Table 5: Comparison results VIPeR dataset.

Methods	R1	R5	R10	R20
ImproveDeep [1]	34.81	63.61	75.63	84.49
KISSME [14]	34.81	60.44	77.22	86.71
Simil.Learning [5]	36.80	70.40	83.70	91.70
MFA (LOMO)[31]	38.67	69.18	80.47	89.02
XQDA (LOMO) [17]	40.00	68.13	80.51	91.08
TCP [6]	47.8	74.7	84.8	91.1
DNS [34]	51.17	**82.09**	**90.51**	**95.92**
SpindleNet [36]	53.80	74.1	83.2	92.1
HydraPlus-Net [19]	56.6	78.8	87.0	92.4
(Ours) PLSR	**65.98**	**81.49**	**88.45**	**95.25**

3.1 Implementation details

We use Resnet50 [9] as baseline, pre-trained on ImageNet and modify the last fully connected layer with the number of classes i.e. 751; 1, 367 and 702 units for Market-1501, CUHK03 and DukeMTMCReID respectively. We train the network for 130 epochs using stochastic gradient descent with a base learning rate lr of 0.01. We gradually decrease lr by a factor of $\gamma = 0.1$ after 40 epochs. We use a momentum of $\mu = 0.9$, weight decay of $\lambda = 5 \times 10^{-4}$ and a mini-batch size of 32. DCGAN model is trained for 30 epochs using Adam [13] with learning rate $lr = 0.0002$ and $\beta_1 = 0.5$. All the input images are resized to 256×256 before being randomly cropped into 224×224 with random horizontal flip. We scale the pixels in the range of 1 and -1 and apply zero-center by mean pixel and random erasing [41].

3.2 Evaluations

We use Cumulated Matching Characteristics (CMC) and mean average precision (mAP) as defined in [37] to evaluate the performance of our model. We use the L2 Euclidean distance to compute a similarity score for ranking and followed the evaluation protocol as defined by previous works [29, 39, 40]. We only report rank 1, 5 and 10 accuracy. Competitive results are also shown in black.

3.2.1 Comparison with the state of art. We evaluate the proposed model on four datasets and report the results on Table 2 3 4 5. '-' means that no reported results is available and '*' means the paper is available on ArXiv but not published.

Market-1501 [37] dataset contains 12, 936 images and 1501 identities. We used 751 identities for training and 750 identities for testing.. On this dataset, we achieved an **89.16%** rank 1 accuracy and **75.15%** mAP accuracy exceeding LSRO [40] by **5.19%** and **9.08%** respectively. Table 2 shows that our method outperforms previous works globally.

CUHK03 [16] datasets provides two image sets, one set is automatically detected by the deformable-part-model detector DPM [7], and the other set contains manually cropped bounding boxes. In this work, we use the detected samples, so misalignment, occlusions and body part missing are quite common, making the dataset more realistic. On **CUHK03**, we achieved a **91.03%** rank 1 accuracy and **94.21%** mAP accuracy exceeding LSRO [40] by **6.41%** and **6.81** on rank 1 and mAP respectively. As shown in Table 3, out method outperforms existing models.

On **DukeMTMCReID**, as shown in Table 4, we achieved an **76.53%** rank 1 accuracy and **60.79%** mAP accuracy. We exceed LSRO [40] results by **8.85%** and **13.66** on rank 1 and mAP accuracy respectively. However, SVDNet [24] exceeds our model by only **0.17%**.

On **VIPeR** dataset, our method achieves a **65.98%** rank 1 accuracy. 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 (PLSR) boots **1.23%~6.41%** rank 1 accuracy and **1.43%~6.81%** mAP on all datasets.

4 CONCLUSION

In this paper, we proposed Partial Label Smoothing Regularization (PSLR), a semi-supervised framework to address the over-smoothness problem found in current regularization methods. PSLR consists of three steps. Firstly, we trained a CNN for discriminative learning patterns from labeled data. For each training image, we extract a high *dim*-feature map from the last convolution layer and directly apply *k-means* clustering algorithm. Secondly, we used a GAN model to generate sample images for each given cluster. Each generated sample is therefore assigned a label using our regularization method. And finally, we defined a new objective function

and fine-tuned a baseline model. Extensive experiments on four large-scale datasets show the superiority of our method over a vast state-of-art methods.

5 ACKNOWLEDGMENT

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 also appreciate Yongsheng Peng and Yuyang Zhou for their valuable contributions.

REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks. 2015. An improved deep learning architecture for person re-identification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3908–3916. https://doi.org/10.1109/CVPR.2015.7299016

[2] Jon Almazán, Bojana Gajic, Naila Murray, and Diane Larlus. 2018. Re-ID done right: towards good practices for person re-identification. *CoRR* abs/1801.05339 (2018).

[3] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. 2018. Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. *Computer Vision and Image Understanding* 167 (2018), 50 – 62. https://doi.org/10.1016/j.cviu.2017.12.002

[4] D. Berthelot, T. Schumm, and L. Metz. 2017. BEGAN: Boundary Equilibrium Generative Adversarial Networks. *ArXiv e-prints* (March 2017). arXiv:cs.LG/1703.10717

[5] D. Chen, Z. Yuan, B. Chen, and N. Zheng. 2016. Similarity Learning with Spatial Constraints for Person Re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1268–1277.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. 2016. 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)*. 1335–1344.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (Sept 2010), 1627–1645.

[8] M. Geng, Y. Wang, T. Xiang, and Y. Tian. 2016. Deep Transfer Learning for Person Re-identification. *ArXiv e-prints* (Nov. 2016). arXiv:cs.CV/1611.05244

[9] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[10] A. Hermans, L. Beyer, and B. Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *ArXiv e-prints* (March 2017). arXiv:cs.CV/1703.07737

[11] Goodfellow J. Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Sherjil Ozair David, Courville Aaron, and Bengio Yoshua. 2014. Generative Adversarial Network. In *NIPS*. The Neural Information Processing Systems.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. 2016. Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv e-prints* (Nov. 2016). arXiv:cs.CV/1611.07004

[13] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[14] M. KÃűstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. 2012. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2288–2295. https://doi.org/10.1109/CVPR.2012.6247939

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *ArXiv e-prints* (Sept. 2016). arXiv:cs.CV/1609.04802

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 152–159. https://doi.org/10.1109/CVPR.2014.27

[17] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2197–2206. https://doi.org/10.1109/CVPR.2015.7298832

[18] S. Liao and S. Z. Li. 2015. Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3685–3693. https://doi.org/10.1109/ICCV.2015.420

[19] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. 2017. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. In *Proceedings of the IEEE international conference on computer vision*. 350–359.

[20] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. 2016. Context Encoders: Feature Learning by Inpainting. In *Computer Vision and Pattern Recognition (CVPR)*.

[21] A. Radford, L. Metz, and S. Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv e-prints* (Nov. 2015). arXiv:cs.LG/1511.06434

[22] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. 2016. Generative Adversarial Text to Image Synthesis. *ArXiv e-prints* (May 2016). arXiv:1605.05396

[23] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[24] Y. Sun, L. Zheng, W. Deng, and S. Wang. 2017. SVDNet for Pedestrian Retrieval. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 3820–3828. https://doi.org/10.1109/ICCV.2017.410

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. https://doi.org/10.1109/CVPR.2017.357

[26] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. 2016. Gated Siamese Convolutional Neural Network Architecture for Human Re-identification. In *ECCV*.

[27] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating Videos with Scene Dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 613–621. http://dl.acm.org/citation.cfm?id=3157096.3157165

[28] W. Wang, Q. Huang, S. You, C. Yang, and U. Neumann. 2017. Shape Inpainting using 3D Generative Adversarial Network and Recurrent Convolutional Networks. *ArXiv e-prints* (Nov. 2017). arXiv:cs.CV/1711.06375

[29] Lin Wu, Chunhua Shen, and Anton Hengel. 2016. Deep Linear Discriminant Analysis on Fisher Networks: A Hybrid Architecture for Person Re-identification. 65 (06 2016).

[30] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint Detection and Identification Feature Learning for Person Search. In *CVPR*.

[31] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. 2014. Person Re-Identification Using Kernel-Based Metric Learning Methods. In *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, 1–16.

[32] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. 2017. Cross-view Asymmetric Metric Learning for Unsupervised Person Re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*.

[33] C. Zhang, L. Wu, and Y. Wang. 2018. Crossing Generative Adversarial Networks for Cross-View Person Re-identification. *ArXiv e-prints* (Jan. 2018). arXiv:cs.CV/1801.01760

[34] L. Zhang, T. Xiang, and S. Gong. 2016. Learning a Discriminative Null Space for Person Re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1239–1248. https://doi.org/10.1109/CVPR.2016.139

[35] Y. Zhang and S. Li. 2011. Gabor-LBP Based Region Covariance Descriptor for Person Re-identification. In *2011 Sixth International Conference on Image and Graphics*. 368–371. https://doi.org/10.1109/ICIG.2011.40

[36] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. *Conference on Computer Vision and Pattern Recognition* (07 2017), 907–915.

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. 2015. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1116–1124. https://doi.org/10.1109/ICCV.2015.133

[38] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. 2017. Person Re-identification in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3346–3355. https://doi.org/10.1109/CVPR.2017.357

[39] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A Discriminatively Learned CNN Embedding for Person Re-identification. *ACM Transactions on Multimedia Computing Communications and Applications* (2017). https://doi.org/10.1145/3159171

[40] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*.

[41] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. 2017. Random Erasing Data Augmentation. *ArXiv e-prints* (Aug. 2017). arXiv:cs.CV/1708.04896

[42] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera Style Adaptation for Person Re-identification. In *CVPR*.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *ArXiv e-prints* (March 2017). arXiv:cs.CV/1703.10593

[44] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems 30*. 465–476.