MULTI BRANCH SIAMESE NETWORK FOR PERSON RE-IDENTIFICATION

Asad Munir *Niki Martinel* *Christian Micheloni*

University of Udine, Italy

ABSTRACT

To capture robust person features, learning discriminative, style and view invariant descriptors is a key challenge in person Re-Identification (re-id). Most deep Re-ID models learn single scale feature representation which are unable to grasp compact and style invariant representations. In this paper, we present a multi branch Siamese Deep Neural Network with multiple classifiers to overcome the above issues. The multi-branch learning of the network creates a stronger descriptor with fine-grained information from global features of a person. Camera to camera image translation is performed with generative adversarial network to generate diverse data and add style invariance in learned features. Experimental results on benchmark datasets demonstrate that the proposed method performs better than other state of the arts methods.

***Index Terms*—** Person Re-Identification, Cycle-GAN

1. INTRODUCTION

Person re-identification (re-id) is a challenging task that requires to recover person of interest's (probe) images from an image gallery set across multiple disjoint cameras. Due to its importance in video surveillance applications, the problem of re-id is gaining more and more attention. Variations like changing in view point, background clutters, different camera domains and occlusions make re-id a very difficult task.

To resolve these issues, existing techniques focus on either robust feature representations [1, 2] or learn optimal matching metrics [3, 4]. Currently, deep learning based methods [5, 6] with the combination of the above mentioned solutions provide superior results outperforming traditional handcrafted low level feature representations for re-id. With the rapid growth of deep learning convolutional neural networks designed for image classification retrieving robust and impressive feature representations of person image in person re-id is more reliable. On contrary to classification, the learned descriptor discriminates between unseen similar images as the training and testing classes (identities) are different in re-id.

Different researches [7, 8] aim to design a better metric learning loss functions apart from feature learning including triplet loss, triplet hard loss, quadruplet loss etc for improving generalization of the model. These metric learning losses have higher performance than classification losses because of dissimilar identities in testing stage. Classification based approaches need to calculate distance matrix of features for unseen person images during inference time that creates mismatching due to different categories (i.e person identities) in training and testing. To avoid this mismatching and learn more robust global features for person re-id, we propose a siamese network (metric learning) based on classification loss. For the better use of classification loss and to overcome the mismatch of features during testing, we add multiple classifiers to learn more disrciminative features from person images.

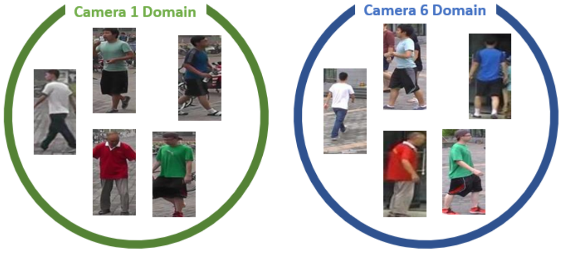

Fig. 1. Domain (style) difference between camera 1 and camera 6 of Market1501 dataset.

The problem of person re-id undergoes many variations in images such as pose variations (different views of a single person) and domain variations (different camera domains i.e camera environment and illuminations) as shown in Fig.1. To learn these type of variations many generative adversarial network (GAN) [9] based approaches [10, 11, 12, 13, 14] has been proposed. In these methods, new data is generated with the help of GAN and is added in original training data to make it more robust to these variations. Generative models with pose variations and style (camera domains) variations have significantly improved the performance of person re-id.

In the proposed approach, the learned features are more discriminative and robust to overcome the pose variations. Different camera domains have different environments and illuminations (i.e indoor and outdoor cameras) and produce images of their own style. To learn these variations, we generate augmented data for every camera style. Since CycleGAN translates images between two domains, we trained CycleGAN [15] models for each pair of cameras. Generated samples are added to the original data with a soft labeling

[16] which are improving the performance of the proposed approach. We propose a multi classifier siamese network integrated with CycleGAN [15] to learn discriminative features for person re-identification. Our contributions are as follows

- A Multi branch (classifier) siamese network with classification loss to learn the most discriminative features for person images.
- CycleGAN is integrated with proposed siamese network to capture the style variations and to enhance the network performance.

By introducing the above contributions, the proposed approach produces better results than the existing methods on benchmark datasets for person re-id as mentioned in experimental results section.

2. MULTI-BRANCH SIAMESE NETWORK

2.1. Problem Definition and Notations

Let a set of n training images $\{I_i\}_{i=1}^{n}$ be acquired by a camera network with corresponding identities labels $\{y_i\}_{i=1}^{n}$. Also let $\{(I_a, y_a), (I_p, y_p)\}$ and $\{(I_a, y_a), (I_n, y_n)\}$ denote the positive and negative training pairs such that $y_a = y_p$ and $y_a \neq y_n$. The task is to retrieve similar images from the gallery set of a camera, given a probe image in different camera. The training process of re-id is same as the general image classification problem when using cross entropy classification loss. Re-id learns better parameters for the last classification (fc) layer and build a discriminative global feature representation to predict the person identity labels. Based on these global features, a distance matrix is computed to differentiate between the persons during testing stage since training and testing sets have no overlap in identities.

2.2. Proposed Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) are deeper and efficient in learning feature representations and accurate to train if they consist of shorter connections between layers. Leveraging on such outcomes, image encoders are defined upon ResNet-50 architecture in the proposed network. We modify the last downsampling block to make spatial size of convolutional feature maps larger before global average pooling layer. We set the stride to 1 by following the work of R-FCN [17]. At the end of image encoder, we add a 1×1 convolutional layer to reduce the feature size from 2048 to 1024. This added convolutional layer learns the most discriminative global features from the entire person image. Two images features from encoders and convolution layers are feeded into element-wise subtraction, element-wise square, batch normalization and fully connected (fc) layer to calculate the similarity score. To predict the person identity from features at convolutional layer, multiple classifiers (fully connected layers) are added along with batch normalization and Rectified Linear Unit (ReLU) layers. Since in back-propagation the gradients from the classifiers gather into previous convolutional layers, thus the classifiers are responsible to focus the learned model on global features for computing distance matrix [18]. The overall architecture of the network is shown in Fig 3. Both the image encoders and added convolutional layers are sharing weights since they are performing the same task.

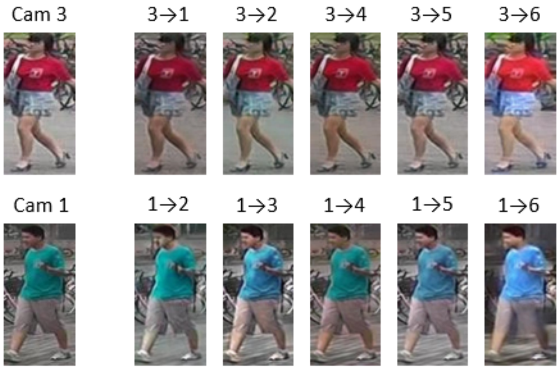

Fig. 2. Camera style transferred images by Cycle-GAN from one camera to all other cameras in Market-1501 dataset.

2.3. Camera to Camera Style Translation

We employ CycleGAN [15] to generate new samples in each camera style. The goal of the CycleGAN is to learn two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$ such that the distribution X is indistinguishable from distribution of images from $F(Y)$ using adversarial loss and two discriminators D_X and D_Y. The overall loss function of CycleGAN is:

$$\begin{aligned} L(G, F, D_X, D_Y) = & \; L_{GAN}(D_Y, G, X, Y) \\ & + L_{GAN}(D_X, F, Y, X) \\ & + \lambda L_{cyc}(G, F) \end{aligned} \tag{1}$$

where $L_{GAN}(D_Y, G, X, Y)$ and $L_{GAN}(D_X, F, Y, X)$ are the adversarial loss functions for both mapping functions. $L_{cyc}(G, F)$ is the cycle consistency loss to reconstruct the image after cycle mapping and is given by:

$$L_{cyc}(G, F) = ||F(G(X)) - X||_1 + ||G(F(Y)) - X||_1 \tag{2}$$

λ is a weight parameter in eq 1 while X and Y are images from two different camera domains. To preserve the color consistency between input and output images, identity mapping loss is added which is expressed as:

$$L_{idt}(G, F) = ||F(X) - X||_1 + ||G(Y) - Y||_1 \tag{3}$$

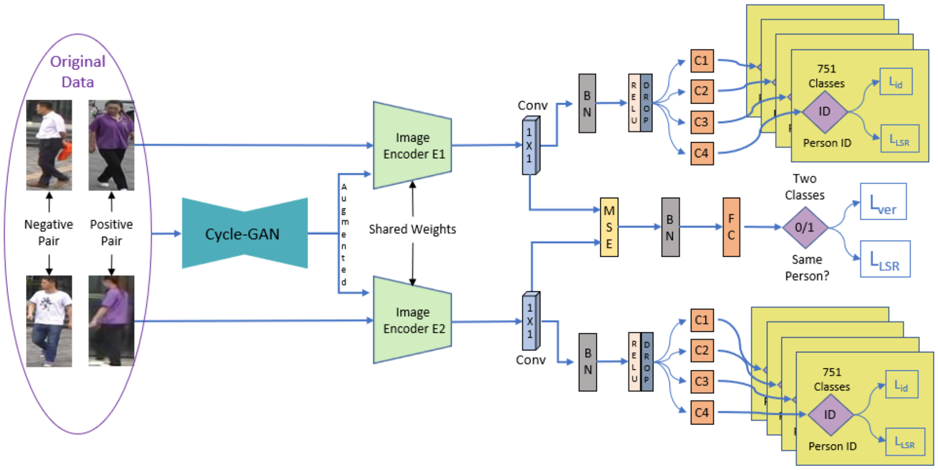

Fig. 3. Overview of our framework. $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the final loss. We show 751 classes from Market-1501 [19] dataset.

We train CycleGAN [15] models for every pair of cameras in the datasets and follow the settings and networks architectures used in CamStyle [13]. Given a re-id dataset consisting of images collected from M cameras, we generate $M - 1$ new images for every image in training set and refer them as style augmented images as shown in Fig 2. Since the contents of original images are preserved in augmented images so we assign the same identity labels to newly generated samples. Along with the original training images, we use style augmented images in training to make the network robust to style variations.

2.4. Training

we use cross entropy classification losses for the training of original images. Two types of losses named as verification loss L_{ver} and identity loss L_{id} are used for similarity and identity learning respectively in the network. L_{ver} is the binary cross entropy loss and is given as:

$$L_{ver} = -C \log d(x_1, x_2) - (1 - C)(1 - \log d(x_1, x_2)) \quad (4)$$

where x_1, x_2 represent the two input person images and $d(x_1, x_2)$ is the output score of the network. C is the ground-truth label i.e if x_1, x_2 belongs to same person then $C = 1$ and $C = 0$ otherwise. To predict the identity of the person image, we use cross entropy loss L_{id} which is written as:

$$L_{id} = -\sum_{c=1}^{C} \log(p(c))q(c) \quad (5)$$

where $p(c)$ is the output probability of the input belonging to class c and C is the total number of classes (person identities) in the dataset. $q(c)$ is the ground truth distribution and it is expressed as:

$$q(c) = \begin{cases} 1 & c = y \\ 0 & c \neq y \end{cases} \quad (6)$$

The generated augmented samples contain noise so they cannot be treated as real samples. To address this issue, we apply the label smoothing regularization (LSR) [16] for the augmented samples to assign soft labels to them. The redefinition of eq 6 is

$$q_{LSR}(c) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{C} & c = y \\ \frac{\epsilon}{C} & c \neq y \end{cases} \quad (7)$$

where $\epsilon \in [0, 1]$ and we use $\epsilon = 0.1$ in our work. With eq 7 cross entropy loss becomes L_{LSR} loss written as

$$L_{LSR} = -(1 - \epsilon) \log p(y) - \frac{\epsilon}{C} \sum_{c=1}^{C} \log p(c) \quad (8)$$

The overall loss function is the addition of all the losses from every classifier to train the whole network on original and augmented samples.

$$L = L_{ver} + L_{id} + L_{LSR} \quad (9)$$

L_{id} and L_{LSR} losses are calculated at the output of each classifier which is represented in Fig 3. All these losses are added to make the final loss. In testing stage, only image encoder is used along with the added convolutional layer to calculate image features and Euclidean distances between prob and gallery images.

3. EXPERIMENTS

Datasets We conduct our experiments on two benchmark datasets Market-1501 [19] and DukeMTMC-reID [20]. The

Table 1. Statistics of two re-id benchmark datasets

Benchmark	Item	Total	Train	Test
Market-1501	ID	1501	751	750
	Image	32668	12936	19281
DukeMTMC-ReID	ID	1404	702	702
	Image	36411	16522	17661

statistics of these two datasets are shown in table 1. We adopt standard data split setting and single query test.

Implementation Details We implemented the proposed model using Pytorch. Resnet50 [21] is used as image encoder E1 and E2 pretrained on ImageNet with the settings mentioned in section 2.2. The network is optimized by Stochastic Gradient Descent (SGD) with momentum 0.9. The initial learning rates are 0.001 and 0.1 for image encoders and all the other layers respectively, and they are divided by 10 after 80 epochs as we train 100 epochs in total. The batch size is set to 64 with positive-negative ratio and generated-original ratio are $1 : 3$. All the images are resized to 256×128 with random cropping and random horizontal flipping data augmentations. The dropout probability is 0.5. For the training of Cycle-GAN [15] we followed the setting used in Cam-style [13].

Table 2. Comparisons to the state-of-the-art re-id methods on Market-1501 and DukeMTMC-ReID. The top 1 and 2 results are in red and blue.

Methods	Reference	Market		DukeMTMC	
		mAP	R1	mAP	R1
SVDNet [22]	ICCV17	62.1	82.3	56.8	76.7
DPFL [23]	ICCV17	73.1	88.9	60.6	73.2
BraidNet [24]	CVPR18	69.5	83.7	59.5	76.4
PSE [25]	CVPR18	69.0	87.7	62.0	79.8
MLFN [26]	CVPR18	74.3	90.0	62.8	81.0
Range-s [27]	ICIP19	81.0	90.7	70.1	81.9
IDCL [18]	CVPRW19	73.3	89.2	59.1	79.4
LSRO [16]	ICCV17	66.1	84.0	47.1	67.7
PT [11]	CVPR18	68.9	87.7	56.9	78.5
PN-GAN [10]	ECCV18	72.6	89.4	53.2	73.6
CAM-Style [13]	CVPR18	71.5	89.4	57.6	78.2
OURS	-	75.9	91.1	62.9	82.2

3.1. Comparison with the State-of-Arts Methods

Table 2 shows the performance comparison with the previous methods. The dashed line in table 2 is splitting two types of methods based on training data. The methods below the dashed line add extra augmented data generated by Generative Adversarial Networks (GANs) [9] for training compared to the above methods which are trained on original data only. Range-s [27] has better Mean Average Precision (mAP) because it is based on re-ranking algorithm while we present our results without any type of re-ranking. The posted results of IDCL [18] are based only on multi branch strategy as we are using that type of multi classifier technique (table 3, 3^{rd} row). We perform better from all other methods in terms of Rank 1 Accuracy (R1). As Compared to GAN methods (below the dashed line), Our performance with the augmented data is much higher than them in case of both measurements. The limitation of the proposed method is the number of cameras because large number of cameras have very high computational cost when calculating style transfer between each pair of cameras.

3.2. Component Analysis

We divide the proposed network into four components to make an ablation study and verify the effectiveness of each component. The Muti-C in table 3 represents four classifiers and removing it means we are using a single classifier. The existence of L_{ver} makes the network a siamese structure, without which the network is only a single line structure. We carry out experiments on Market dataset with different combinations. The detailed results are described in table 3.

Table 3. Component Analysis of the proposed Multi Classifier Siamese Network on Market-1501 dataset in terms of mAP(%) and top-1 accuracy(%).

Networks	Components				Market	
	L_{ver}	L_{id}	Multi-C	Cycle-GAN	mAP	R1
baseline single	×	✓	×	×	65.8	85.6
baseline siamese	✓	✓	×	×	71.5	88.4
proposed(no ver)	×	✓	✓	×	73.3	89.2
proposed(no cgan)	✓	✓	✓	×	75.0	90.2
proposed	✓	✓	✓	✓	**75.9**	**91.1**

4. CONCLUSION

In this work, We propose a Multi Branch (classifier) Siamese Network along with Cycle-GAN for person re-identification. With multiple classifiers and losses, proposed network learns robust global features at the added convolutional layers. Multiple identity losses are merged with verification loss to build a stronger and disciminative descriptor . To overcome the camera style variations, we generate augmented data with the help of cycle-GAN. During training, augmented data is utilized by providing soft labeling loss function along with original data. Experimental results demonstrate the benefits of the proposed method in enhancing the performance of person re-id on two benchmark datasets.

Acknowledgement

This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866)

5. REFERENCES

[1] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[2] Ziyan Wu, Yang Li, and Richard J Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[3] Niki Martinel, Christian Micheloni, and Gian Luca Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Transactions on Image Processing*, 2015.

[4] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[5] Yanbei Chen, Xiatian Zhu, and Shaogang Gong, "Person re-identification by deep learning multi-scale representations," in *ICCVW*, 2017.

[6] Yiluan Guo and Ngai-Man Cheung, "Efficient and deep person re-identification using multi-level similarity," in *CVPR*, 2018.

[7] Niki Martinel, Gian Luca Foresti, and Christian Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *CVPRW*, 2019.

[8] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[10] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue, "Pose-normalized image generation for person re-identification," in *ECCV*, 2018.

[11] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu, "Pose transferrable person re-identification," in *CVPR*, 2018.

[12] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al., "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *NIPS*, 2018.

[13] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018.

[14] Asad Munir, Gian Luca Foresti, and Christian Micheloni, "Generating domain and pose variations between pair of cameras for person re-identification," in *ICDSC*, 2019.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[16] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *CVPR*, 2017.

[17] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016.

[18] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li, "In defense of the classification loss for person re-identification," in *CVPRW*, 2019.

[19] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[20] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*, 2016.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[22] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017.

[23] Yanbei Chen, Xiatian Zhu, and Shaogang Gong, "Person re-identification by deep learning multi-scale representations," in *ICCVW*, 2017.

[24] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang, "Person re-identification with cascaded pairwise convolutions," in *CVPR*, 2018.

[25] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *CVPR*, 2018.

[26] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang, "Multi-level factorisation net for person re-identification," in *CVPR*, 2018.

[27] Guile Wu, Xiatian Zhu, and Shaogang Gong, "Person re-identification by ranking ensemble representations," in *ICIP*, 2019.