Applied Soft Computing Journal 100 (2021) 106939

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

A feature disentangling approach for person re-identification via self-supervised data augmentation

Feng Chen [a], Nian Wang [a], Jun Tang [a,*], Fan Zhu [b]

[a] School of Electronics and Information Engineering, Anhui University, Anhui, Hefei, 230601, China
[b] Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 24 November 2019
Received in revised form 19 August 2020
Accepted 24 November 2020
Available online 4 December 2020

Keywords:
Person re-identification
Convolutional neural network
Self-supervised learning
Data augmentation
Feature disentangling

ABSTRACT

To address the problem of insufficient training data in person ReID, this paper proposes a data augmentation method based on image channels shuffling, by which a large volume of diversified training samples sharing similar edges can be produced. In the meantime, a soft label assignment strategy is designed to characterize the correlations between the original image and the generated counterparts. Furthermore, we design an encoder–decoder based learning structure for the person ReID task, where the encoder module tackles feature disentangling according to the introduced correlations, and the decoder module handles reconstruction using the combinations of decoupled features. Extensive experiments on four benchmark datasets demonstrate the effectiveness and robustness of the proposed method by attaining significant improvement over some state-of-the-art approaches. Source code is released at: https://github.com/flychen321/feature_disentangle_reid.

1. Introduction

Person re-identification (ReID) serves as an important component in surveillance systems and many other related intelligent applications. Given a probe image, the person ReID task aims to retrieve images with the same identity (ID) from a large collection of images captured from multiple disjoint cameras. Despite years of efforts, person ReID remains a very challenging task due to the significant variations in human poses, background cluttering and occlusions across different cameras.

Over the past few years, convolutional neural networks (CNNs) have been the dominant technique for most person ReID [1,2] studies with the prevalence of deep learning in the computer vision community. In practical scenarios, the fundamental challenges of person ReID include a large number of identity classes and extremely limited samples in each class. Many CNN based metric learning or representation learning approaches, such as triplet loss [3] and SVDNet [4], were proposed to address this issue. These methods can boost the generalization capability of the networks to a certain degree, nevertheless they are unable to fundamentally eliminate the negative impacts which are introduced by the inefficiency of training samples. Data augmentation has been considered as a commonly-used strategy for model enhancement, and the common approaches can be generally grouped into two categories. The first category mainly employs traditional image processing techniques, including random cropping, flipping, erasing, etc. These operations are beneficial in terms of increasing sample diversity within each identity class, however they are unable to encode any class-sensitive information during the transformation. The second category is devoted to the application of generative adversarial networks (GANs) [5]. With well-designed objective functions, GANs can generate diverse samples while being aware of specific discrepancies between the generated samples [6] (e.g., camera variations). However, the cost of training GANs is expensive, and it may take several hours or even longer. Besides, the learned model often produces undesirable samples (see Fig. 1), which leads to limited improvements on performance.

Recently, feature disentangling has been widely utilized for learning discriminative and representative features of highly variable inputs for the person ReID task. Pose-guided [8] and attribute-guided [9] ReID learning are classical approaches to decomposing features that leads to pose/domain invariant representations. However, when poses or attributes are introduced into ReID learning, auxiliary supervision is indeed utilized in an implicit way. Separate encoding [10] is another promising way to achieve feature decomposition. Combined with separate encoding, high-quality artificial images are produced to cover unseen variations in the training set by means of generative models. Nevertheless, the weakness underlying this approach is that the separation of encoding is defined in a brute-force manner, as the disentangled appearance and structure spaces are only defined conceptually and there are no explicit data support for them.

[*] Corresponding author.
E-mail address: tangjunahu@163.com (J. Tang).

https://doi.org/10.1016/j.asoc.2020.106939
1568-4946/© 2020 Elsevier B.V. All rights reserved.

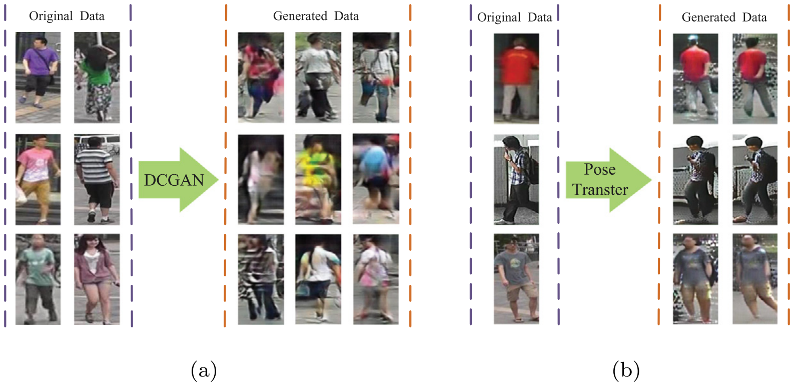


Fig. 1. Illustration of GAN based data augmentation. (a) Augmentation with DCGAN [7]. The quality of the generated samples is poor, and some of them are nearly corrupted. (b) Augmentation with pose-transfer GAN [8]. Pose-transfer GAN converts the original image into a specific pose, but this process may discard some important details (e.g. the generated images have no details on hands and feet).

Self-supervised learning aims to exploit the structural information of data itself to learn visual representations, which is an appealing and promising solution to learning from large-scale data without sufficient annotation. Motivated by the recent progress of self-supervised learning, we employ self-supervised learning to generate auxiliary data as an enhancement for the ReID model.

In light of the above discussions, we propose a novel method for person ReID with two critical improvements. Firstly, we design an efficient self-supervised data augmentation method via shuffling image channels (see Fig. 3), where a large number of images with significant appearance variations can be easily produced. Meanwhile, the generated images share the similar edges with the original one. The image generation procedure can be viewed as filling a common edge image with different contents. We then present a soft label assignment strategy in line with our data augmentation, which can provide a reasonable description for the relationship between a sample and its augmented ones. Secondly, based on the interpretable decoupled representation of augmented samples, we develop a separate encoder–decoder based network with multiple paths (see Fig. 2). In the training stage, a pedestrian image and its augmented one are simultaneously fed into the network. The encoder module decomposes a pedestrian image into two latent spaces, i.e., the edge-related space and the content-related space, which conforms to the interpretation of our augmentation strategy. The decoder module is responsible for the reconstruction of pedestrian images, which uses the different combinations of semantic features delivered by the encoder module as the inputs. Due to the consistency of edge-related features, the decoder module is designed to reconstruct the counterpart input images.

In this paper, we focus on the self-supervised data augmentation to facilitate the procedure of feature disentangling. Although conceptually related, our approach is different from the previous feature disentangling approach [10] in two aspects. Firstly, the composed images are produce using GAN-based generative models in [10], while our approach conducts image generation with a simple yet effective image channel shuffling operation. Secondly, our feature decoupling is driven by pixel-level information and it is actually based on single real-world images, while the decoupling of the appearance and structure spaces in [10] is performed at a conceptual level and images with different IDs are necessary for feature disentangling. Such two distinct characteristics make our approach more scalable.

The main contributions of this paper are as follows:

1. We propose a self-supervised data augmentation method based on image channels shuffling, which provides a highly efficient solution for producing diverse training samples. Meanwhile, the proposed approach provides the foundation for effective feature disentangling.

2. We introduce an adaptive soft label assignment strategy that supports the proposed data augmentation method, which is also useful for feature disentangling.

3. We develop an end-to-end trainable encoder–decoder model with multiple paths for the person ReID task. By embedding the encoder–decoder model into the feature disentangling and the image reconstruction workflows, the model eventually leads to discriminative feature embeddings.

2. Related work

In this section, we briefly review some related literatures from two aspects.

Deep learning based person ReID. Recently, many GAN based ReID methods have been proposed to cope with insufficiency of training data and domain discrepancy. Zheng et al. [7] proposed the pioneer work in applying GANs to data augmentation, and they also introduced a variant of label smoothing regularization to assign a uniform label distribution for the unlabeled generated samples. For the purpose of learning view-invariant representations, Zhong et al. [6] utilized cycle-GAN to achieve camera style adaption to reduce the discrepancy between different cameras. Liu et al. [8] used GANs to generate samples with rich pose styles to alleviate the influence of pose variations. In a somewhat opposite manner, Qian et al. [11] employed GANs to synthesize pedestrian images with some specific pose styles. Feature disentangling is another active topic on enhancing ReID model learning. Wang et al. [9] proposed an unsupervised cross-dataset ReID method, which transfers the labeled information of an existing dataset to an unlabeled target one via joint learning attribute-semantic and identity-discriminative feature representations. Mao et al. [12] proposed to learn correspondence representations from both semantic-components and color-texture distributions via deep pyramid matching, and the obtained representations are fused to achieve the person ReID task. Song et al. [13] developed a mask-guided contrastive attention model based on binary segmentation masks, thereby achieving the feature separation of body and background regions. Zhou et al. [14] proposed a two-branch based network to aggregate pose estimation and attribute recognition to improve the re-ID performance. Zheng et al. [10] presented a joint learning manner for the person ReID task, where a generative module and a discriminative module are interacted to boost the discrimination of decoupled features. In a broad sense, we can regard that the part-based models conduct feature decoupling along the spatial dimension.

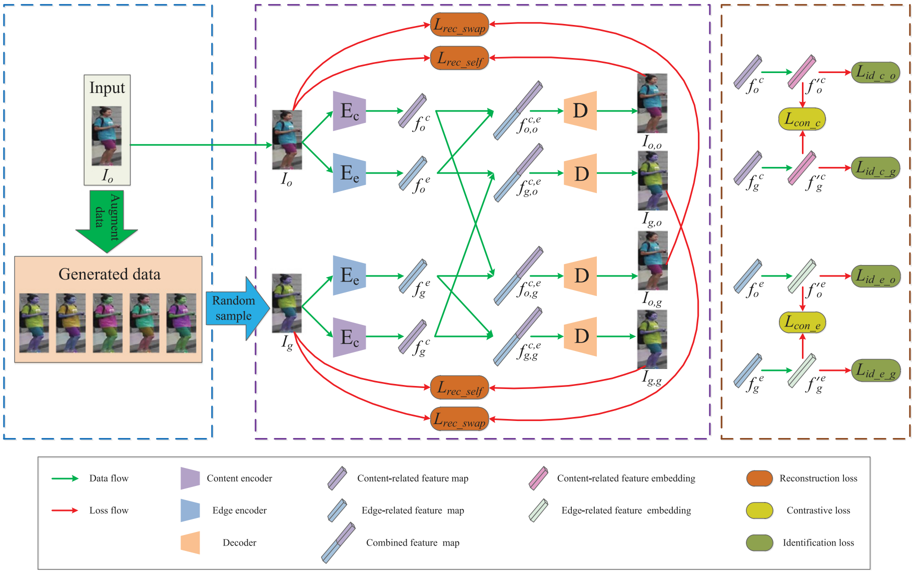


Fig. 2. Overview of our approach. The blue dashed box indicates the process of data augmentation. The purple and brown dashed boxes demonstrate the procedure of features disentangling. For clearer illustration, the process of calculating the contrastive loss and the identification loss is drawn in the brown dashed box, where the input feature maps are obtained from the encoding results illustrated in the purple dashed box. Best viewed in color.

Sun et al. [15] split a person image into a couple of parts and then learned their feature embeddings, respectively, and the overall similarity is computed by aggregating the similarities between corresponding parts. Wei et al. [16] split a person into three parts using the joints obtained by the pose estimator, and then explicitly leveraged the local and global cues of human body to generate discriminative and robust representations for pedestrians. Fu et al. [17] divided a person into different horizontal parts with multiple scales and consequently enhanced the discriminative ability of partial features in a hierarchical manner.

Self-supervised learning. With the growing research interest in self-supervised learning, its applications have been broaden to a variety of tasks, including action recognition [18], crowd counting [19,20], image recognition [21], person ReID [22], etc. Misra et al. [18] augmented video samples by shuffling the temporal order of video frames. As the augmented samples contain the implicit cues of incorrect orders, they are beneficial to enhance the perception of important temporal information. Wei et al. [21] presented an unsupervised approach to learning visual representations by solving jigsaw puzzles in an iterative manner, which is formulated as an optimization problem considering both absolute and relative positions. Chen et al. [23] introduced an unsupervised generative model integrating adversarial learning and self-supervised learning together, in which image rotation is used as a proxy to enhance representation learning. Addressing the problem of partial person ReID, Sun et al. [22] applied self-supervision to learning the visibility of regions, which is used to estimate the shared regions between two images and thus leads to the suppression of spatial misalignment.

3. Method

In this section, we present the details of our method, which is depicted in Fig. 2.

3.1. Data augmentation and soft label assignment

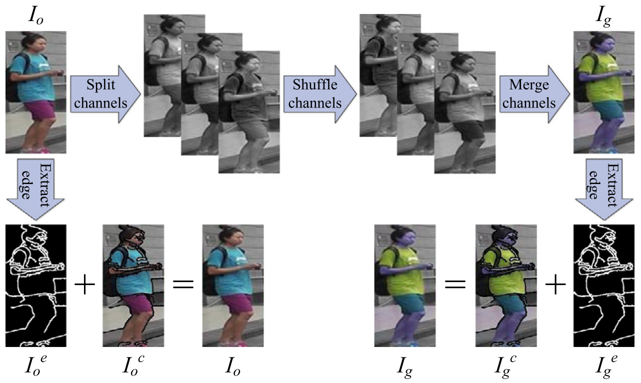


Fig. 3. Illustration of data augmentation. I_o and I_g denote the original image and the generated one, respectively. I_o^e and I_o^c represent the edge and content information of I_o, respectively. I_g^e and I_g^c have similar definitions. From the upper part, we can see that an artificial image I_g with different appearance can be generated by channel shuffling. From the lower part, we can observe that an image can be represented as an composition of the edge and content components at the pixel level, and the edge information of the original image and its corresponding generated image is similar (i.e. I_o^e and I_g^e are similar).

Intuitively, structural information such as related positions of body parts and geometric structures is mainly contained in the edges of a pedestrian image, and appearance information such as clothing color and texture is not that relevant to edges. The proposed data augmentation is designed based on this consideration. A color image can be viewed as a combination of three ordered channels. If the channel order is shuffled, we can obtain a generated image with significant appearance variations, as illustrated in Fig. 3. Meanwhile, the generated image and the original one share similar edges. Notably, this image augmentation method is of high time efficiency compared with GAN-based methods, as we only need to swap image channels.

Given a pedestrian image, we have up to 6 combinations (including the original one) by this approach and regard them as a group. From another perspective, we can imagine that a group of augmented images is generated by filling different contents into an edge-like image. This assumption enlightens us to design a model to disentangle similar component (edge) and different component (content) for a pair of input images composed of an original one and a generated one, consequently enabling

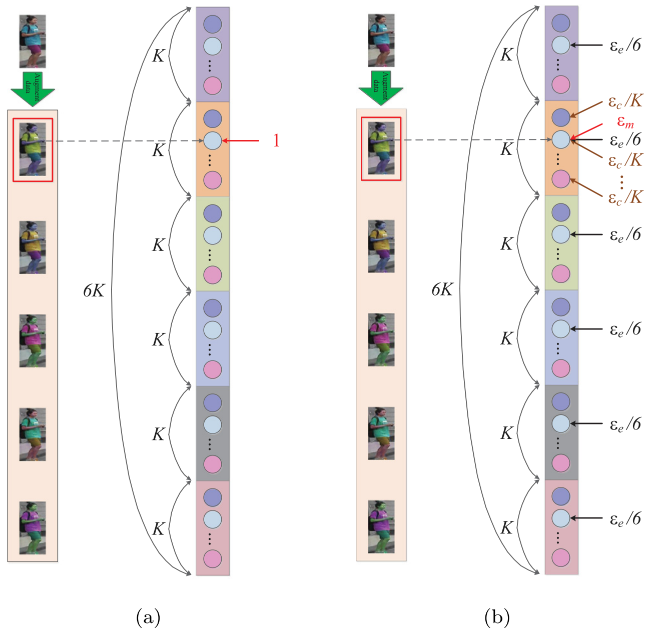


Fig. 4. Illustration of label assignment. (a) One-hot label. This strategy assigns the generated data distinct IDs with respect to channel order, which ignores the edge consistency between the original image and the corresponding generated ones, and the relationship between images with the same channel order. (b) Soft label. This strategy explores the relationship between the original and augmented data, which smooths the one-hot label by decomposing it into the main ingredient (ϵ_m), the content-related ingredient (ϵ_c) and the edge-related ingredient (ϵ_e).

the model to embed images into two separate latent spaces (edge-related and content-related) for better discrimination.

Next we introduce our soft label assignment strategy considering both the channel order and the relationship among a group of augmented images from a global viewpoint, which concerns the following computation of the identification loss. We commence by introducing the one-hot label for comparison. Let K denote the number of identities of the original dataset. As a group of augmented images have significant difference in appearance, we can define a label vector with 6K bits. Concretely, let $k \in \{1, 2, \dots, 6K\}$ be the identity number of labeled data and y be the position of $'1'$ in the one-hot label encoding of a given image, then we have a 6K-dim vector q and the ground truth distribution $q(\cdot)$ is defined by:

$$q(k) = \begin{cases} 0, & k \neq y \\ 1, & k = y \end{cases} \tag{1}$$

The label assignment is illustrated in Fig. 4(a). However, one-hot label strategy is not an applicable solution to assigning labels for our augmented images, as it overlooks the internal connections among them. We therefore propose a soft label method to tackle this problem. As shown in Fig. 4(b), we consider that the label distribution $q(\cdot)$ is determined by three ingredients, and it is formulated by:

$$q(k) = q_m(k) + q_c(k) + q_e(k), \tag{2}$$

where $q_m(k)$, $q_c(k)$ and $q_e(k)$ denote the distributions of the main ingredient, the content-related information, and the edge-related information, respectively. Specifically, the main ingredient is used to represent the identity information of the training sample, which is similar to $'1'$ in one-hot. The content-related component is used to represent the relationship between images with the same channel order, and the edge-related component is used to measure the edge similarity between an image and its corresponding generated ones. To satisfy the definition of probability distribution, the soft label distribution is subject to the constraint denoted by:

$$\epsilon_m + \epsilon_c + \epsilon_e = 1, \tag{3}$$

where ϵ_m, ϵ_c and ϵ_e denote the weights on the main ingredient, the content-related information, and the edge-related information, respectively.

$q_m(\cdot)$ is defined by:

$$q_m(k) = \begin{cases} 0, & k \neq y \\ \epsilon_m, & k = y \end{cases} \tag{4}$$

The definition of $q_m(\cdot)$ is similar to one-hot label in Eq. (1). ϵ_m denotes the weight of the main component and there is only one non-zero item in the 6K-dim soft label encoding.

$q_c(\cdot)$ emphasizes on the consistency of channel order. ϵ_c denotes the weight of the content-related component, which is used to represent the correlation of content-related information between images with the same channel order. For each image, there are K identities have the same channel order as it in the total 6K identities, thus ϵ_c is evenly divided by K bits in the soft label encoding. And it is defined by:

$$q_c(k) = \begin{cases} \epsilon_c/K, & k \text{ and } y \text{ have the same} \\ & \text{channel order} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$q_e(\cdot)$ concentrates on the similarity among a group of augmented images. ϵ_e denotes the weight of the edge-related component, which is used to represent the correlation of edge-related information between the original image and its corresponding augmented ones. For each image, there are 6 identities have similar edges with it in the total 6K identities, so ϵ_e is evenly divided by 6 bits in the soft label encoding. And it is denoted by:

$$q_e(k) = \begin{cases} \epsilon_e/6, & k \text{ and } y \text{ come from} \\ & \text{the same group} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Through the above definition, we present a flexible approach to label assignment depending on what information we are interested in. For example, if we intend to discover content-related information, we assign ϵ_c a larger value. Similarly, we set ϵ_e a larger value to pay more attention to extract edge-related information.

So far, we present a data augmentation approach using only self-contained information, and the discriminative cues of image samples are revealed from ID labels.

3.2. Network architecture

The proposed network architecture is in harmony with our data augmentation and soft label assignment strategy. As illustrated in Fig. 2, the network input consists of a pair of images, which include an original image and an image randomly selected from the augmented ones. For any input image, our network utilizes two encoders with the same structure but different parameters to extract semantic feature maps corresponding to content-related and edge-related information, respectively. Besides, a shared decoder is responsible for image reconstruction according to the feed of different combinations of encoded semantic features.

In the following, we elaborate the objective functions to optimize our network. Let I_o denote the input original image, and I_g represent the input augmented one.

Reconstruction loss. Our model utilizes a content encoder E_c and an edge encoder E_e to extract the content-related feature map f_o^c and the edge-related feature map f_o^e from the input image I_o, respectively. With similar notations, the model extracts f_g^c and f_g^e from the augmented image I_g. To guarantee that the obtained feature maps capture representative information, we use the decoder D to reconstruct each counterpart input image based on the concatenated feature maps yielded from itself. Concretely, D utilizes the concatenated feature maps denoted by $f_{o,o}^{c,e}$ and $f_{g,g}^{c,e}$ to reconstruct images $I_{o,o}$ and $I_{g,g}$ respectively, and we therefore need to enforce the similarity constraint between I_o and $I_{o,o}$, as well as I_g and $I_{g,g}$. This constraint is denoted by the pixel-wise L_2 loss:

$$L_{rec\_self} = \sum_{x \in \{o,g\}} \|I_x - D(E_c(I_x), E_e(I_x))\|_2 \,, \tag{7}$$

where $D(E_c(I_x), E_e(I_x))$ denotes the generated image $I_{x,x}$ corresponding to the combined feature map $f_{x,x}^{c,e}$.

If these two kinds of information are effectively decoupled, for a pair of input images, the content-related feature maps, i.e., f_o^c and f_g^c, should be distinct while the edge-related feature maps, i.e., f_o^e and f_g^e, should be identical. To impose this constraint, we swap the separated representations of the pairwise images, by concatenating f_o^c and f_g^e, f_g^c and f_o^e, respectively, to obtain the swapped representations $f_{o,g}^{c,e}$ and $f_{g,o}^{c,e}$ across two images. As f_o^e and f_g^e are theoretically identical, the reconstructed images $I_{g,o}$ and $I_{o,g}$ are desired to be the same as I_g and I_o, respectively. Formally, we have:

$$L_{rec\_swap} = \sum_{\substack{x,y \in \{o,g\} \\ x \neq y}} \left\| I_x - D(E_c(I_x), E_e(I_y)) \right\|_2 \,, \tag{8}$$

where $D(E_c(I_x), E_e(I_y))$ has the similar definition as described in Eq. (7). The overall reconstruction loss is defined by:

$$L_{rec} = L_{rec\_self} + L_{rec\_swap} \tag{9}$$

Contrastive loss. Meanwhile, we utilize the contrastive loss to speed up the convergence of network training, which is implemented in a feature-separated way. As we favor the distinction of the content-related features, L_{con_c} is defined by:

$$L_{con\_c} = max(margin_c - \left\| f_o^{'c} - f_g^{'c} \right\|_2, 0)^2, \tag{10}$$

where $f_o^{'c}$ is a feature embedding mapped from f_o^c through a fully connection layer and the other related notations have similar definitions, and $margin_c$ is a hyper-parameter indicating the minimum gap between $f_o^{'c}$ and $f_g^{'c}$. We encourage the consistency between the edge-related features, so L_{con_e} is given by:

$$L_{con\_e} = \left\| f_o^{'e} - f_g^{'e} \right\|_2^2, \tag{11}$$

And the overall contrastive loss is defined as:

$$L_{con} = \alpha_c L_{con\_c} + \alpha_e L_{con\_e}, \tag{12}$$

where α_c and α_e are the trade-off parameters.

Identification loss. To further enhance the identification power of the learned features, the identification loss is used to enable $f_o^{'c}$, $f_g^{'c}$, $f_o^{'e}$ and $f_g^{'e}$ to identify the correct categories of the input pairwise images, which is shown in the brown dashed box in Fig. 2. And we have:

$$L_{id} = \sum_{x \in \{o,g\}} L_{id\_c\_x} + \sum_{x \in \{o,g\}} L_{id\_e\_x}. \tag{13}$$

For each $L_{id\_c\_x}$ and $L_{id\_e\_x}$ in Eq. (13), we utilize the cross-entropy loss function to define them and the ground truth probability distribution is obtained by means of our proposed soft label strategy.

Altogether, the overall loss function is a weighted sum of the following losses:

$$L_{total} = L_{con} + \beta L_{rec} + \gamma L_{id}, \tag{14}$$

where β and γ control the relative importance of the corresponding objectives.

Table 1
Description of datasets in our experiments.

Datasets	Market	Duke	CUHK03		MSMT17
			Detected	Labeled	
Train images	12,936	16,522	7368	7365	32,621
Train IDs	751	702	767	767	1041
Test images	19,732	19,989	6728	6732	93,820
Test IDs	750	702	700	700	3060
Views	6	8	2	2	15

4. Experiments

To evaluate the performance of our proposed method, we conduct comprehensive experiments and ablation studies on four benchmark datasets, including Market-1501 (Market) [24], DukeMTMC-reID (Duke) [25], CUHK03 [26] and MSMT17 [27].

4.1. Datasets and evaluation metrics

The details of four benchmark datasets are described in Table 1. For the CUHK03 dataset, we follow the recently proposed protocol [28], which divides it into a training set with 767 identities and a test set with the remaining 700 identities.

We utilize two commonly-used metrics to evaluate algorithm performance, i.e., Rank-1 identification rate and mean Average Precision (mAP). All experiments are conducted under the single query setting.

4.2. Implementation details

In our approach, the encoders E_c and E_e share the same structure built upon DenseNet-121 [29] pre-trained on ImageNet [30]. We employ DenseNet-121 for its popularity and competitive performance. Compared with the alternative backbones such as VGG [31] and ResNet-50 [32], DenseNet-121 has larger receptive fields so as to better extract the global information of pedestrians. And we replace its global average pooling layer with an adaptive max pooling layer and remove its fully-connected layer to output feature maps with the size of 1024 × 8 × 4 (channel × height × width). Two fully-connected layers are sequentially added at the top of DenseNet-121. One is used to map the obtained convolutional feature maps to 512-dim feature embeddings (i.e., map f_o^c to $f_o^{'c}$, f_o^e to $f_o^{'e}$, etc.). And the other is utilized to output the classification probability distribution, where the number of neurons is set to 6 times the number of identities of the original training set ($6K$) according to our soft label assignment method. The decoder D consists of five transposed convolutional layers, which converts combined feature maps to reconstructed images with the same size of input ones. To make it clearer, we visualize the input and output sizes of each layer along a branch in Fig. 5. The detailed structure of D is shown in Table 2, where the parameter size is denoted by [$kernel\_height \times kernel\_width, channels\_num, stride, padding$], and the input and output sizes are denoted by [$channels\_num \times height \times width$].

We keep the aspect ratio of all images and resize them to the uniform size of 256 × 128 and utilize random horizontal flipping

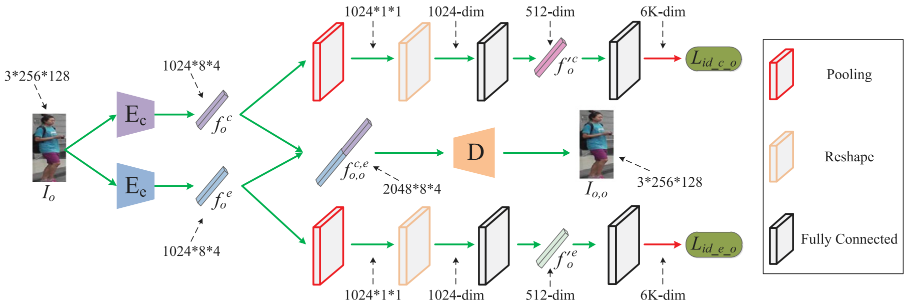


Fig. 5. Illustration of the input and output sizes of each layer along a branch of our network. The input and output sizes of other branches are the same as what we depict here.

Table 2
The detailed structure of the decoder.

Layer name	Parameters size	Input size	Output size
ConvTranspose0	[3 × 3, 512, 2, 1]	2048 × 8 × 4	512 × 16 × 8
ConvTranspose1	[3 × 3, 128, 2, 1]	512 × 16 × 8	128 × 32 × 16
ConvTranspose2	[3 × 3, 64, 2, 1]	128 × 32 × 16	64 × 64 × 32
ConvTranspose3	[3 × 3, 32, 2, 1]	64 × 64 × 32	32 × 128 × 64
ConvTranspose4	[3 × 3, 3, 2, 1]	32 × 128 × 64	3 × 256 × 128

Table 3
Hyper-parameters of loss functions.

Datasets	α_c	α_e	β	γ	$margin_c$
Market	0.2	0.6	0.2	0.3	2.5
Duke, CUHK03 and MSMT17	0.2	0.5	0.3	0.3	2.0

and random erasing for data augmentation. The batch size is set to 24 and the dropout probability is set to 0.5. The SGD optimizer is employed to train the network, and the learning rate starts with 0.01 for the basic DenseNet-121 layers and 0.1 for the two added fully-connected layers and all layers of the decoder. The learning rates are divided by 10 every 35 epochs and we train 90 epochs in all.

To make a fair comparison, we use the strategy of cross-dataset tuning to determine hyper-parameter values. Specifically, we determine the hyper-parameter values on Market according to the tuned results on Duke. Similarly, we set the hyper-parameter values on Duke, CHUK03, and MSMT17 based on the tuned results on Market. The obtained hyper-parameter values of the loss functions are summarized in Table 3, and the hyper-parameter values with respect to soft label assignment are reported in Table 4.

In the test stage, the 512-dim content-related and edge-related features are concatenated into a 1024-dim vector as the final pedestrian representation, and the Euclidean distance is used to compute the similarity between the probe image and the gallery.

4.3. Comparisons and discussions

4.3.1. Comparison with state-of-the-art

Here we report the performance comparison with other state-of-the-art methods, which is summarized in Table 5. Specifically, the implementation of all comparison methods is based on deep learning. It should be stressed that we only use single query and do not employ any other post-processing means. Our approach yields the competitive results on all the datasets. In comparison with the typical feature disentangling approaches [13–16] [10, 17], our method outperforms almost all the compared methods by a meaningful margin, which indicates that the introduced feature disentangling is more effective in learning discriminative representations. Compared with GAN-based data augmentation approaches [6–8,11], our method not only attains significant performance improvement but also surpasses them on time efficiency, as we augment data in an online way and do not need to train offline generative models.

4.3.2. Ablation studies and discussions

To fully present the effectiveness of three key components, i.e., data augmentation (DA), soft label assignment (SLA), and feature disentangling (FD), we make a systematical analysis on them by experiments under the same-dataset setting (i.e., training and testing on the same dataset). In these evaluations, we use the same settings as those in the presented approach for fair comparison without specification.

Baseline Our model degrades to a classification model that only contains one DenseNet-121 branch if data augmentation and feature disentangling are disabled. And we use it as the baseline. The number of neurons of the top fully-connected layer is set to the number of original identities. Meanwhile, the identification loss and the one-hot label are utilized to train the network.

Baseline+DA We apply the proposed data augmentation and assign the generated data different new IDs with respect to channel order. Accordingly, we set the number of neurons of the top fully-connected layer to 6 times the number of identities of the original training set. The identification loss and the one-hot label are utilized in the training stage.

Baseline+DA+SLA Here we use the parameter settings of $\epsilon_m = 0.7$, $\epsilon_c = 0.2$ and $\epsilon_e = 0.1$. This configuration also provides a reference to the usefulness of feature disentangling.

The ablation experimental results are summarized in Table 6. We can observe clear performance improvement when each component is applied one by one. It also verifies that the combination of these components is complementary and beneficial to achieve better performance.

To validate the usefulness of each loss function (i.e., reconstruction loss, contrastive loss and identification loss) discussed in Section 3.2, we perform some ablation experiments on Market and Duke by disabling the corresponding loss item in Eq. (14), respectively. The results are reported in Table 7. We can see that the performance drops dramatically when the identification loss is disabled. Among the three losses, only computing the identification loss needs label information. Therefore, our model degrades to an unsupervised one when we remove the identification loss item. Under this circumstance, the whole training process does not use any supervision information, consequently leading to significant performance decline. If the other two losses are disabled separately, we can also observe obvious performance degradation, which demonstrates that each loss function is useful for our model training.

Table 4
Hyper-parameters of soft label assignment.

Datasets	Content-related branch			Edge-related branch		
	ϵ_m	ϵ_c	ϵ_e	ϵ_m	ϵ_c	ϵ_e
Market	0.7	0.3	0.0	0.95	0.0	0.05
Duke, CUHK03 and MSMT17	0.7	0.3	0.0	0.9	0.0	0.1

Table 5
Comparison of our method with the state-of-the-art. Best results are indicated in bold.

Methods					CUHK03					
	Market		Duke		Detected		Labeled		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Verif + Identif [1] (TOMM'2018)	79.51	59.87	68.9	49.3	–	–	–	–	–	–
MGCAM [13] (CVPR'2018)	83.79	74.33	–	–	46.71	46.87	50.14	50.21	–	–
DCGAN [7] (ICCV'2017)	83.97	66.07	67.68	47.13	–	–	–	–	–	–
Triplet loss [3] (arXiv'2017)	84.9	69.1	78.6	59.2	–	–	–	–	–	–
Pose-transfer [8] (CVPR'2018)	87.65	68.92	78.52	56.91	41.6	38.7	45.1	42.0	–	–
Pose-Normalized [11] (CVPR'2018)	89.43	72.58	73.58	53.20	–	–	–	–	–	–
Camstyle [6] (CVPR'2018)	89.49	71.55	78.32	57.61	–	–	–	–	–	–
GLAD [16] (MM'2017)	89.9	73.9	–	–	–	–	–	–	61.4	34.0
LRDNN [14] (IJCAI'2019)	90.4	82.8	85.3	73.2	–	–	–	–	–	–
PCB [15] (ECCV'2018)	92.3	77.4	81.7	66.1	63.7	57.5	–	–	68.2	40.4
HPM [17] (AAAI'2019)	94.2	82.7	86.6	74.3	63.9	57.5	–	–	–	–
DG-Net [10] (CVPR'2019)	**94.8**	86.0	86.6	74.8	65.6	61.1	–	–	77.2	52.3
Our approach	94.72	**86.74**	**89.05**	**78.02**	**70.45**	**65.80**	**72.21**	**67.82**	**79.74**	**56.27**

Table 6
Results of the ablation experiments under same-datasets setting. Best results are indicated in bold.

Method					CUHK03					
	Market		Duke		Detected		Labeled		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	91.15	76.79	82.54	66.97	53.50	48.72	56.79	50.84	68.26	41.67
Baseline + DA	93.08	82.00	86.18	73.95	61.21	55.92	67.07	60.67	74.14	49.02
Baseline + DA + SLA	94.03	83.75	86.94	74.01	64.93	59.44	67.57	62.51	75.59	50.78
Our approach	**94.72**	**86.74**	**89.05**	**78.02**	**70.45**	**65.80**	**72.21**	**67.82**	**79.74**	**56.27**

Table 7
Results of the ablation experiments for loss functions. Best results are indicated in bold.

Method	Market		Duke	
	Rank-1	mAP	Rank-1	mAP
Our approach w/o Reconstruction loss	93.72	85.45	87.91	77.33
Our approach w/o Contrastive loss	94.08	85.37	87.64	77.00
Our approach w/o Identification loss	11.85	3.15	10.95	3.29
Our approach	**94.72**	**86.74**	**89.05**	**78.02**

To take our study one step further, we investigate the effect of single semantic features and the combined feature, respectively. The results are reported in Table 8. We can observe remarkable results when any single feature is employed, demonstrating the usefulness of feature disentangling. In particular, the combination of semantic features achieves substantial improvement over single features, which indicates that the overall training manner is able to boost ReID learning. Specifically, the results on CUHK03 are different from those on the other datasets, edge-related features achieve better performance than content-related features as the color variation is less than that in the other datsets, which in turn verifies the effectiveness of our feature disentangling. Meanwhile, the combined feature produces the best results, which is consistent with those presented in our paper.

Next we visualize some retrieval results in Fig. 6. We can see that content-related features are favorable to the pedestrian images with high correlation in contents. For instance, the cloth color in the 6th, 8th, and 11th columns is very close to that of the probe image. In contrast, edge-related features are sensitive to structure similarity and lead to more appearance errors. And the cloth color in the 6th, 9th, 12th, 13th columns is completely different from that of the probe image. Moreover, by combining two semantic features, some ambiguities caused by single features can be reduced.

4.3.3. More discussions on cross-dataset ReID

It is well-known there exists domain gap [33–35] between different person ReID datasets, and it usually leads to significant performance drop when we directly apply a ReID model trained on a dataset to another one. To further demonstrate the generalization ability of our approach, we perform comparative experiments under the cross-dataset setting, i.e., training and test on different datasets.

Table 8
Performance comparison in the case of different feature embeddings.

Feature					CUHK03					
	Market		Duke		Detected		Labeled		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
f_c	92.34	82.64	86.58	74.03	63.53	57.18	66.29	60.49	75.66	49.42
f_e	93.59	84.41	86.40	73.61	67.75	61.47	70.00	63.63	75.38	49.46
$[f_c, f_e]$	**94.72**	**86.74**	**89.05**	**78.02**	**70.45**	**65.80**	**72.21**	**67.82**	**79.74**	**56.27**

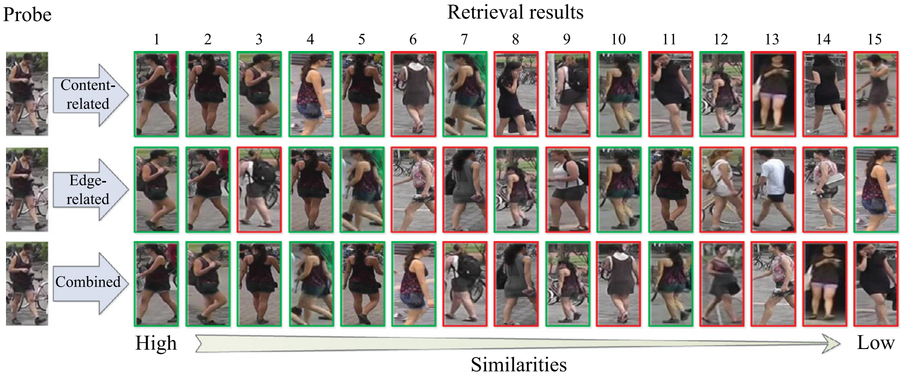


Fig. 6. Illustration of retrieval results on Market when different semantic features are used. Each row depicts a search result delivered by a specific semantic feature. Green boxes indicate correct matches and red boxes indicate error matches.

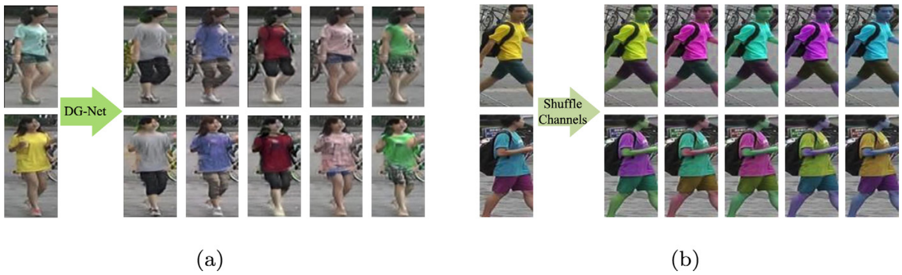


Fig. 7. Examples of generated images on Market-1501. (a) The samples generated by DG-Net can increase the diversity of both colors and clothes styles. (b) The samples produced by our approach mainly focus on enriching the diversity of colors.

Table 9
Results of the comparative experiments under the cross-dataset setting. Best results are indicated in bold.

Method	Market → Duke		Duke → Market	
	Rank-1	mAP	Rank-1	mAP
Verif + Identif [1] (TOMM'2018)	20.06	9.67	36.52	14.98
Triplet loss [3] (arXiv'2017)	28.37	15.61	34.03	14.59
DCGAN [7] (ICCV'2017)	23.15	10.47	43.80	19.21
Camstyle [6] (CVPR'2018)	26.0	12.7	51.9	21.7
PCB [15] (ECCV'2018)	37.1	21.2	52.9	25.4
DG-Net [10] (CVPR'2019)	**42.62**	**24.25**	**56.12**	26.83
Our approach	39.36	23.13	54.93	**27.58**

We choose some representative ReID methods from Table 5 for comparison. In order to guarantee the fairness of comparison, all comparative experiments are conducted under direct transferring (i.e., without any fine-tuning). The results of the comparative experiments are detailed in Table 9, where '→' denotes the transfer from the source to the target. We have two observations from the results. Firstly, all methods suffer from considerable performance degradation due to the domain gap. Secondly, our method achieves relatively better results under the cross-dataset setting. In comparison with [1,3,6,7] [15], we can see that our data augmentation and feature disentangling strategy have better generalization ability. The performance of our method is slightly worse than that of [10] on the whole. We think the reason is that the samples generated by [10] not only increase the diversity of colors, but also increase the diversity of clothes styles (see Fig. 7).

To take our study one step further, we perform ablation experiments under the cross-dataset setting. The experimental results are reported in Table 10 (The Baseline, DA, and SLA are the same as those in Section 4.3.2). We can see that clear performance gain is obtained when each component is utilized one by one. It also verifies that each component of our method is helpful to improve model performance under the cross-dataset setting.

Table 10
Results of the ablation experiments under the cross-dataset setting. Best results are indicated in bold.

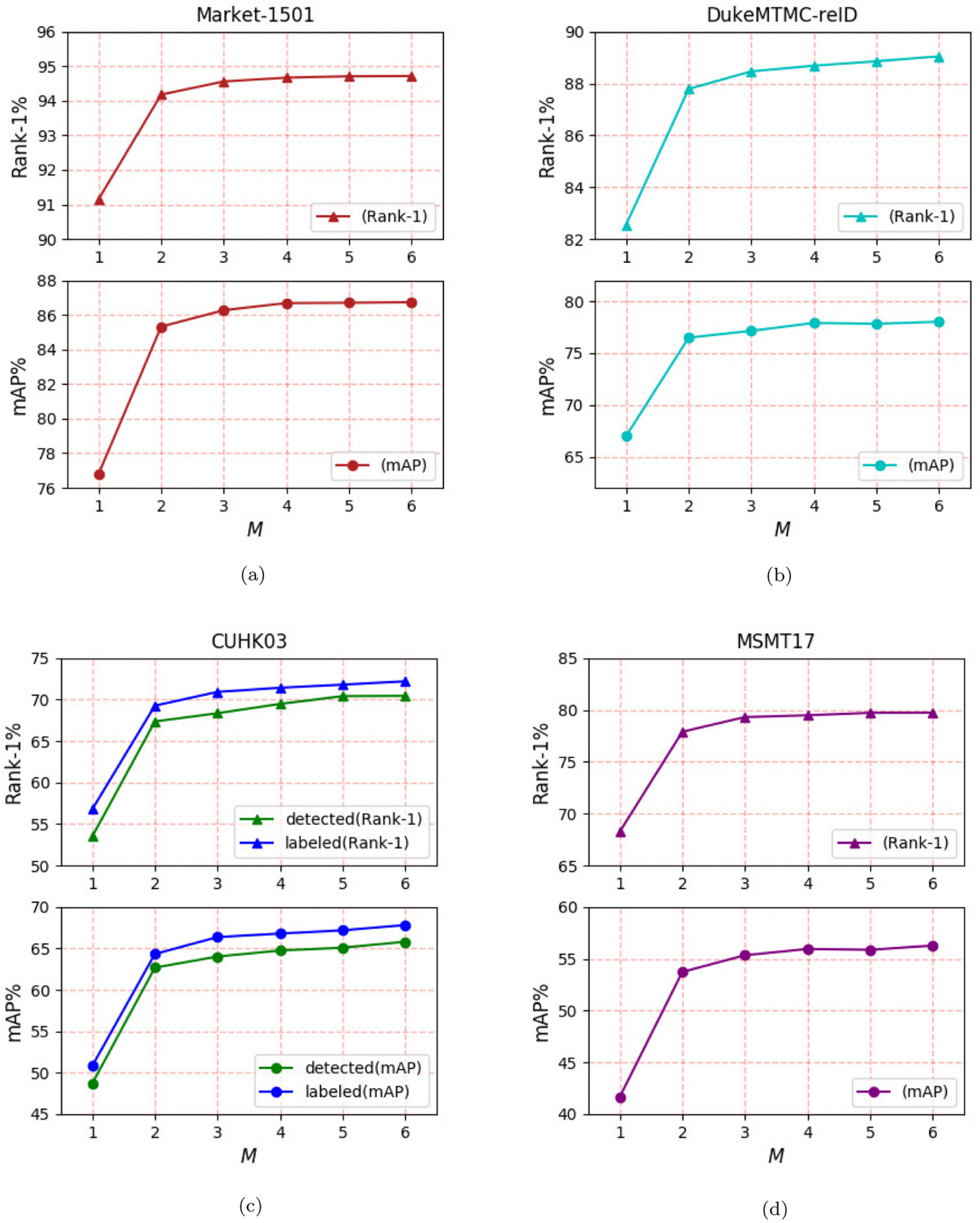
Method	Market → Duke		Duke → Market	
	Rank-1	mAP	Rank-1	mAP
Baseline	14.68	6.51	32.01	11.83
Baseline + DA	35.27	19.28	48.55	22.74
Baseline + DA + SLA	37.51	22.26	51.69	24.72
Our approach	**39.36**	**23.13**	**54.93**	**27.58**

4.4. Parameters analysis

In this part, we perform a quantitative analysis on some important parameters in our approach.

The number of channel order combinations. The number of channel order combinations is related to the number of augmented images. We evaluate our approach on four datasets with the variation of this parameter. The results are plotted in Fig. 8, which show that the best performance can be reached when we use 6 combinations.

Hyper-parameters of the loss functions. We study how changes in hyper-parameters of loss functions affect the performance on Market. According to the validation results demonstrated in Fig. 9, We can see that our approach can obtain



(a)

(b)

(c)

(d)

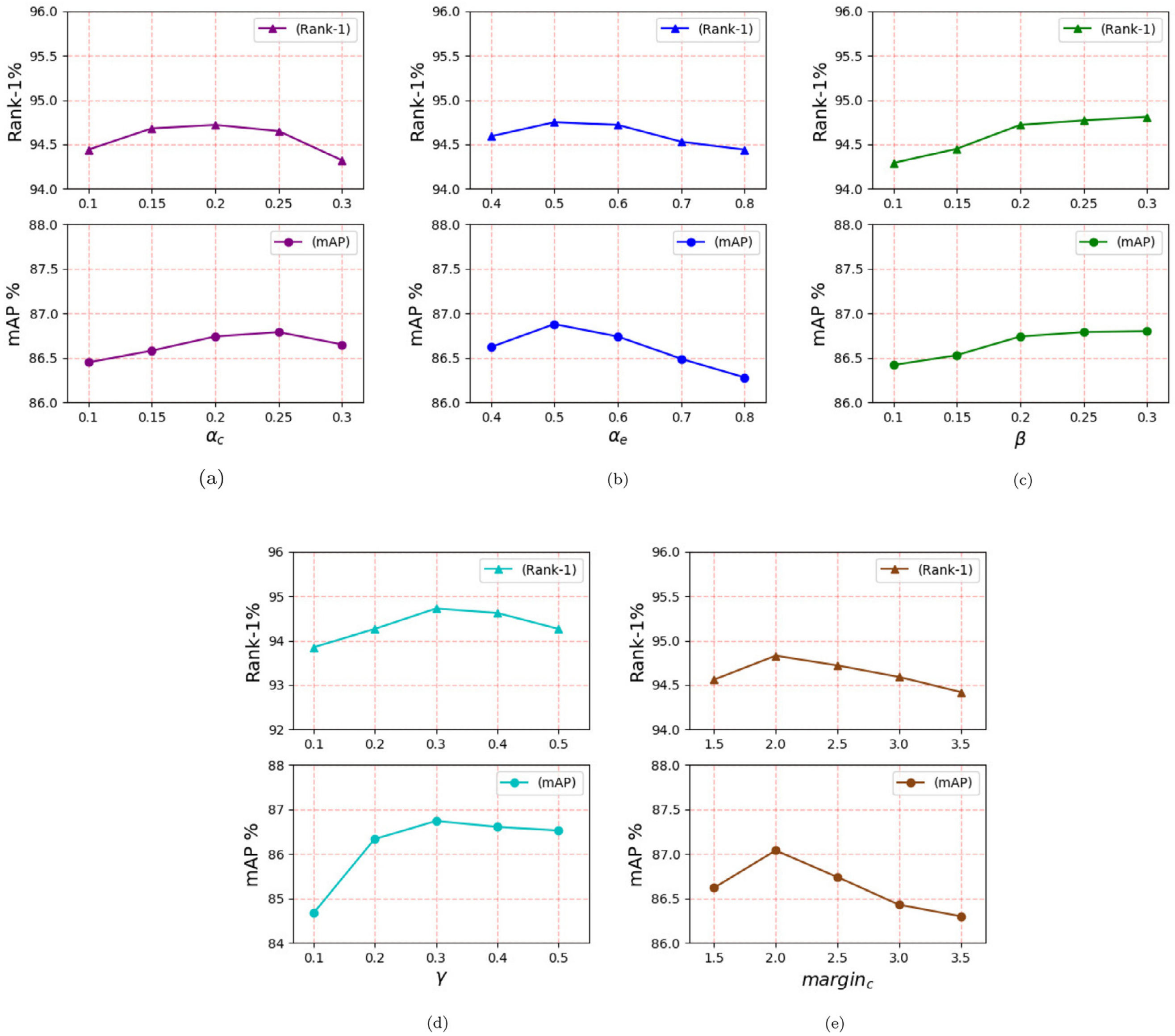
Fig. 8. The influence of the number of image channel combinations (denoted by M) on performance. (a) Market; (b) Duke; (c) CUHK03; (d) MSMT17.

satisfactory results in a wide range of each hyper-parameter, which demonstrates the robustness of our method.

Hyper-parameters in soft label assignment. Here we study the effect of changes in ϵ_m on performance on Market. As we set $\epsilon_e = 0.0$ and $\epsilon_c = 0.0$ for the content-related branches and the edge-related ones to decouple features respectively, we only need to search the value of ϵ_m according to the constraint defined in Eq. (3). We execute evaluation of one sort of branches by fixing the alternatives. As shown in Fig. 10, our approach can yield acceptable results in a wide fluctuation of each hyper-parameter, which prove the robustness of our method.

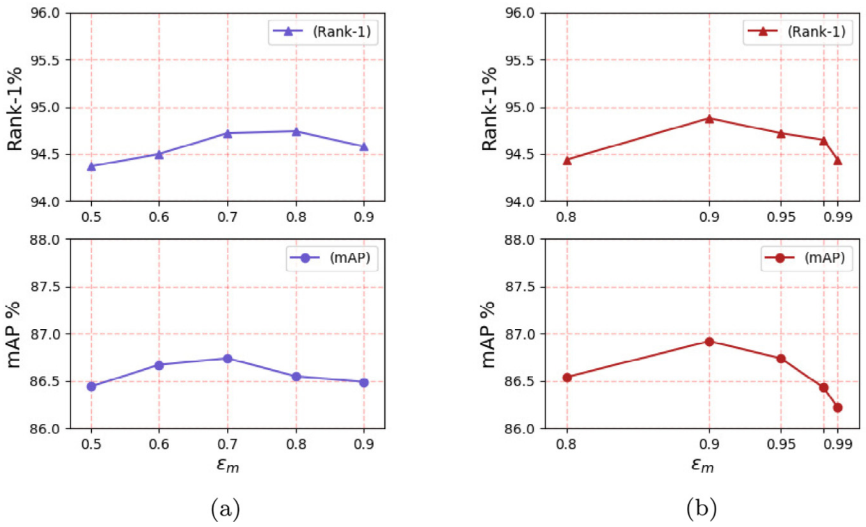
5. Conclusions

In this paper, we have presented a unified method by seamless integrating data augmentation and ReID learning. Our efficient self-supervised data augmentation method used image channels



(a) (b) (c) (d) (e)

Fig. 9. The influence of loss functions related super parameters on performance. (a) α_c; (b) α_e; (c) β; (d) γ; (e) $margin_c$.



(a) (b)

Fig. 10. The influence of soft label related super parameters on performance. (a) ϵ_m for the content-related branches; (b) ϵ_m for the edge-related branches.

shuffling to generate diverse image samples. Meanwhile, we proposed a soft label assignment strategy to characterize the correlations between the original image and the generated ones. Furthermore, we designed a network with the encoder–decoder structure to disentangle features according to the introduced correlations, which finally enhance the discriminative information of the learned features. Systematic comparison experiments and ablation studies were conducted to demonstrate the advantage of our approach.

CRediT authorship contribution statement

Feng Chen: Conceptualization, Methodology, Formal analysis, Data curation, Software, Visualization, Writing - original draft. **Nian Wang:** Supervision, Funding acquisition, Resources, Writing - review & editing. **Jun Tang:** Investigation, Methodology, Writing - review & editing, Supervision, Funding acquisition, Project administration. **Fan Zhu:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of China under grants 61672032 and 61772032, and by the National Key R&D Program of China under grant 2018YFC0807302.

References

[1] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 14 (1) (2018) 13.

[2] M. Zheng, S. Karanam, Z. Wu, R.J. Radke, Re-identification with consistent attentive siamese networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5735–5744.

[3] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv: Computer Vision and Pattern Recognition.

[4] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3800–3808.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[6] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5157–5166.

[7] Z. Zheng, L. Zheng, 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, pp. 3754–3762.

[8] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, Pose transferrable person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4099–4108.

[9] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2275–2284.

[10] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147.

[11] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 650–667.

[12] C. Mao, Y. Li, Y. Zhang, Z. Zhang, X. Li, Multi-channel pyramid person matching network for person re-identification, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7243–7250.

[13] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1179–1188.

[14] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, M. Gou, Lrdnn: Local-refining based deep neural network for person re-identification with attribute discerning, in: IJCAI, 2019, pp. 1041–1047, http://dx.doi.org/10.24963/ijcai.2019/146.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496.

[16] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, Glad: Global-local-alignment descriptor for pedestrian retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017, pp. 420–428.

[17] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8295–8302.

[18] I. Misra, C.L. Zitnick, M. Hebert, Shuffle and learn: unsupervised learning using temporal order verification, in: European Conference on Computer Vision, Springer, 2016, pp. 527–544.

[19] X. Liu, J. Van De Weijer, A.D. Bagdanov, Exploiting unlabeled data in cnns by self-supervised learning to rank, IEEE Trans. Pattern Anal. Mach. Intell.

[20] X. Liu, J. van de Weijer, A.D. Bagdanov, Leveraging unlabeled data for crowd counting by learning to rank, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7661–7669.

[21] C. Wei, L. Xie, X. Ren, Y. Xia, C. Su, J. Liu, Q. Tian, A.L. Yuille, Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1910–1919.

[22] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 393–402.

[23] T. Chen, X. Zhai, M. Ritter, M. Lucic, N. Houlsby, Self-supervised gans via auxiliary rotation loss, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12154–12163.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[25] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 17–35.

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

[27] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88.

[28] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[29] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[33] A. Torralba, A.A. Efros, Unbiased look at dataset bias, in: CVPR 2011, IEEE, 2011, pp. 1521–1528.

[34] J. Liu, Z.-J. Zha, D. Chen, R. Hong, M. Wang, Adaptive transfer network for cross-domain person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7202–7211.

[35] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, J.-H. Lai, Unsupervised person re-identification by soft multilabel learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2148–2157.