Patch-based Discriminative Feature Learning for Unsupervised Person Re-identification

Qize Yang[1,3] , Hong-Xing Yu[1] , Ancong Wu[2] , and Wei-Shi Zheng[*1,4]

[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]School of Electronics and Information Technology, Sun Yat-sen University, China
[3]Accuvision Technology Co. Ltd.
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
yangqz@mail2.sysu.edu.cn, xKoven@gmail.com, wuancong@mail2.sysu.edu.cn, wszheng@ieee.org

Abstract

While discriminative local features have been shown effective in solving the person re-identification problem, they are limited to be trained on fully pairwise labelled data which is expensive to obtain. In this work, we overcome this problem by proposing a patch-based unsupervised learning framework in order to learn discriminative feature from patches instead of the whole images. The patch-based learning leverages similarity between patches to learn a discriminative model. Specifically, we develop a PatchNet to select patches from the feature map and learn discriminative features for these patches. To provide effective guidance for the PatchNet to learn discriminative patch feature on unlabeled datasets, we propose an unsupervised patch-based discriminative feature learning loss. In addition, we design an image-level feature learning loss to leverage all the patch features of the same image to serve as an image-level guidance for the PatchNet. Extensive experiments validate the superiority of our method for unsupervised person re-id. Our code is available at https://github.com/QizeYang/PAUL.

1. Introduction

Person re-identification (re-id) aims to match the underlying identity of a person from non-overlapping camera views. Because of its important applications in security and surveillance, person re-id has been drawing lots of attention from both academia and industry. In the past decades, most of the existing re-id works focus on distance metric learning [16, 3, 45, 54, 38, 39, 36] and feature learning [10, 20, 53, 2]. Particularly, deep learning [18, 28, 34, 43, 40, 37, 1, 5, 25, 55] has been adopted to re-id community and achieved significant progress. However, most existing re-id methods require tremendous labeled dataset which limits its scalability and usability in the real-world application scenario, because it is expensive and difficult to manually label a large scale dataset. Some recent works focus on using unsupervised learning to address the scalability problem by improving the hand-crafted features [10, 8, 24, 20], clustering [46, 47, 7]. State-of-the-art unsupervised re-id methods transfer the knowledge from a labelled source re-id dataset [40, 37, 1, 5, 25]. However, these methods are limited on the image level while the gaps of images between different datasets are significant. Thus they still yield weak performances.

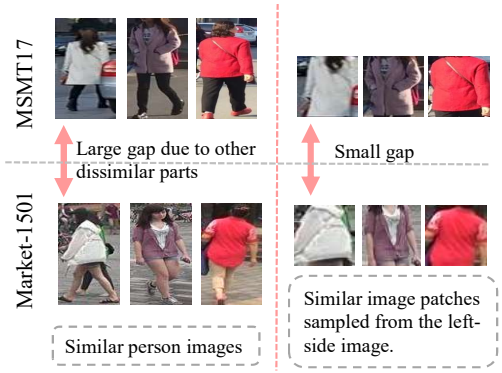


Figure 1. Some image samples of MSMT17 [40] and Market-1501 [53]. It is easier to find that if two images are similar, then their patches would probably also be similar. And the gap of the similar patches would be smaller than the similar images

While the label information of unlabelled data is absent, we find an interesting observation that as shown in Figure 1, if two images are similar, then their patches would probably also be similar. Based on this observation, a patch-based discriminative feature learning model would be more generalizable and can learn discriminative patch feature among different datasets. This inspires us to develop a patch-based unsupervised person re-id model to learn discriminative patch feature instead of image feature. Although some

*Corresponding author

part-based person re-id models [31, 50, 49, 27, 52, 29] have been proposed to study local discriminative feature and outperform those global feature learning methods [30, 12, 48], it is still a largely unsolved problem for person re-id to extract discriminative local features on unlabelled data.

In this work, we develop a **p**atch-b**a**sed **u**nsupervised **l**earning framework (PAUL) for person re-id, and this framework is designed specially for learning discriminative patch feature on unlabelled datasets, which can be divided in three parts as follows. A patch discriminative feature learning network (PatchNet) is designed to select patches from the feature map and learn discriminative feature for each patch. In PatchNet, we propose a **p**atch-bas**e**d **d**iscriminative fe**a**ture learning **l**oss (PEDAL) to guide the PatchNet for learning the patch feature on unlabeled datasets by pulling the features of similar patches together and pushing the dissimilar patches away. Simultaneously, we generate the surrogate positive samples by random image transformation for each image and mine hard negative samples in a mini-batch by cyclic ranking to compose a triplet, and then we develop an **i**mage-level **p**atch **f**eature learning **l**oss (IPFL) to leverage all the patch features of the same image to provide image-level guidance.

The main contributions of this work can be summarized as follows: (1) We demonstrate for the first time how to effectively extract discriminative patch-based local feature on unlabelled data for unsupervised person re-id. (2) To overcome the problem of lacking an effective guidance on unlabeled datasets, we propose PEDAL and IPFL to provide effective guidance for a deep model to train so it can learn discriminative features for unsupervised re-id.

We have evaluated the proposed method on two large-scale datasets including Market-1501 [53] and DukeMTMC-reID [26]. Our method significantly outperforms the existing methods for unsupervised person re-id.

2. Related Work

Supervised person re-id. Most existing person re-id methods employ supervised learning and base on learning distance metric or subspace [16, 3, 45, 54, 38, 39, 36], learning view-invariant discriminative feature [10, 20, 53, 2] or deep learning [18, 28, 34, 43, 19, 31]. However, supervised learning methods rely on substantial labeled training data and manually labeling are time-consuming and may not be reliable, which limits the scalability and practicability of supervised learning methods.

Particularly, some part based model [31, 50, 49, 27, 52, 29] for person re-id has been studied for tackling the misalignment of the person image or learning local features. These part based person re-id methods identify that the local feature learning methods is more generalizable and effective to the unseen identities. However, these methods are designed for supervised learning. Although the pre-trained part based model may be generalizable, these methods lack of an effective guidance on unlabeled datasets.

Unsupervised person re-id. Although Handcrafted appearance feature [10, 8, 24, 20] can be directly applied for unsupervised person re-id, the performance is typically weak because it is very challenging to design a view-invariant feature. To achieve the view-invariance, recent methods attempt to improve the feature [51, 35, 22, 15, 14], or mine underlying labels in the unlabelled data [7, 23, 46, 47]. In particular, Yu *et al*. [46, 47] propose an unsupervised asymmetric distance metric learning based on asymmetric K-means clustering to achieve the view-invariance. However, the pseudo labels of the images obtained by clustering may be noisy, because it may assign the same pseudo label to similar images with different identities, making it more difficult to distinguish the similar person.

Recently, unsupervised person re-id by cross-domain transfer learning [40, 37, 1, 5, 25, 55] is proposed to leverage other labeled datasets to improve the performance of model on the target dataset. Particularly, Wei *et al*. [40] propose to use GAN [9] to bridge the domain gap for person re-id. Wang *et al*. [37] propose to share the source domain knowledge through attributes learned from labeled source data and transfer such knowledge to unlabeled target data. The attribute labels that describe local sematic information is somewhat similar to the appearance information of image patches, while our proposed method does not require the additional attribute labels. Bak *et al*. [1] propose a three-step domain adaptation technique that takes advantage of synthetic data. However, the image adaptation process makes it complex to generalize a model to a new unlabeled dataset. Furthermore, it is difficult for these transfer learning to generalize because the gaps of the person images between different datasets are significant. In contrast, our method learns discriminative features on the patch level which is more straightforward and generalizable.

3. Method

3.1. Overview of the PAUL

In this work, we present a novel patch-based discriminative feature learning framework to utilize the common patches of different datasets and mine the discriminative features on an unlabeled dataset. This framework includes a PatchNet that aims to learn generalizable and discriminative patch feature, and two complementary losses which provide guidance for the PatchNet on the unlabeled dataset, as shown in Figure 2.

The PatchNet is mainly composed of a CNN backbone and the patch generation network (PGN) which can generate different patches from the feature map. Then the network is separated into several branches, appended with an average pooling layer and a convolutional layer for each

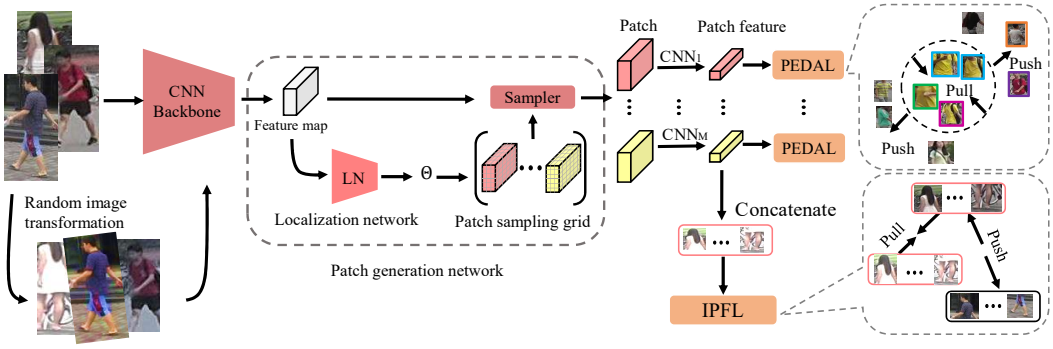


Figure 2. An illustration of the PAUL. The PatchNet is mainly composed of a CNN backbone and the patch generation network. First, we generate surrogate positive samples by using random image transformations. Next, we generate M patches for each feature map by using patch generation network (PGN) which can be split into three parts including the localization network (LN), the patch sampling grid, and the sampler. The PEDAL is designed to pull similar patches together and push the dissimilar patches. The IPFL is designed to pull the real sample and the surrogate positive samples together while pushing hard negative samples away.

branch. The PatchNet is pre-trained on the other labeled dataset initially so as to leverage the shared appearance knowledge of common image patches. Although the PGN is not our main contribution, it is the foundation of our method, so we introduce the PGN firstly in Sec. 3.2.

In order to provide effective guidance for PatchNet to learn more discriminative patch feature on the unlabeled dataset, we propose a **p**atch-bas**e**d **d**iscriminative fe**a**ture learning **l**oss (PEDAL) to pull similar patches together and push the dissimilar patches away.

Simultaneously, we develop a **i**mage-level **p**atch **f**eature learning **l**oss (IPFL) to leverage all the patch features of the same image to provide image-level guidance. Since there are no label information available to compose a triplet on the unlabeled dataset, we concatenate all the patch features of the same image to mine hard negative samples in a mini-batch by cyclic ranking, and generate the surrogate positive samples for each image.

3.2. Patch Generation Network

We sample patches from a relatively small size feature map instead of the image, because it is more efficient and can reduce the computation and the complexity of the CNN [31, 19]. To this end, we introduce a spatial transformation network [13] to form the PGN which can automatically sample the patches from the feature map. As shown in Figure 2, the PGN can be split into three parts including a localization network (LN), patch sampling grids, and the sampler. First, the LN takes the input feature map and predicts M spatial locations parameterized by a set of affine transformation parameters $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, \ldots, \boldsymbol{\theta}_M]$. The LN is composed of a convolutional layer and two fully-connected layers. We initialize the bias of the last fully-connected layer of LN such that the patches are sampled from different spatial regions and capture different cues for the person image at initial state. Then, each predicted transformation parameter $\boldsymbol{\theta}_m$ is used to compute a sampling grid, which is a set of points where the input feature map should be sampled to form the patches. The final step is sampling such that we can get M patches for each image. We refer the readers to [13] for more details.

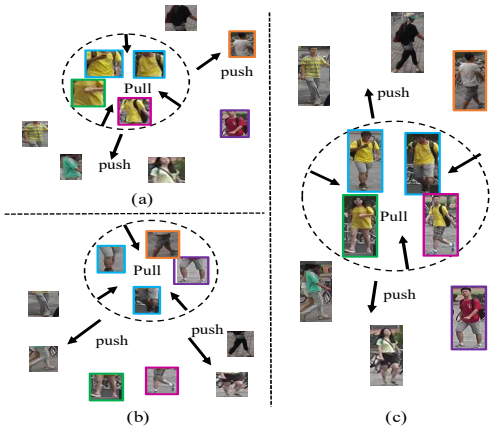


Figure 3. Illustration of discriminative feature learning on an unlabeled dataset. Different color borders means different identities. (a), (b) The person image can be divided into several patches. We learn discriminative patch feature by pulling the features of similar *patches* together and pushing dissimilar ones away. (c) In contrast, if directly pulling the similar *images* together in the feature space, visually similar person images with different identities would be pulled closer, which blurs the identity information of the person image. (Best viewed in color)

3.3. Patch-based Discriminative Feature Learning

In this section, our goal is to guide the PatchNet to learn discriminative patch feature on an unlabeled dataset. Note that the PGN generates M patches for each image feature

map and these different patches of the same image are located at different spatial regions. These different regions may contain different parts of body which have different semantic information [31, 50, 49, 27, 29], so it is better to encode these different patches of the same image by using different CNN branches and perform discriminative feature learning independently for different branches.

In supervised learning, we want the features of the same class to be closer in feature space while staying far away from other classes, so that the feature would be more discriminative [21, 41]. Our novel unsupervised patch feature learning approach is patch-level discriminative learning. We propose to pull the similar patches close while pushing those dissimilar patches away in feature space, which is illustrated in Figure 3.

Let $\mathbf{x}_i^m$ represents the feature of the m-th patch of the i-th image in a mini-batch, we need to compare each unlabeled patch to all the m-th patches of the other images so as to discover the visually similar patches, which is hardly tractable in the mini-batch optimization based deep learning. Therefore, we maintain a patch feature memory bank W^m for storing these patch features [44, 42]. Let $W^m = \{\mathbf{w}_j^m\}_{j=1}^N$, where N is the number of the training images, for each $\mathbf{w}_j^m$, we update it during training on the unlabeled dataset by

$$\mathbf{w}_{j,t}^m = \begin{cases} (1-l) \times \mathbf{w}_{j,t-1}^m + l \times \mathbf{x}_{j,t}^m, & t > 0, \\ \mathbf{x}_{j,t}^m, & t = 0, \end{cases} \quad (1)$$

where t is the training epoch and l is the updating rate of the $\mathbf{w}_{j,t}^m$, $\mathbf{x}_{j,t}^m$ is the up-to-date patch feature. Particularly, $t = 0$ means that we initialize all the memory bank before training on the unlabeled dataset, then keep updating batch-by-batch by using Eq. 1 during training, so that the $\mathbf{w}_j^m$ can be the online approximation of $\mathbf{x}_j^m$[44, 42].

Now, for each $\mathbf{x}_i^m$, we can obtain a set $\mathcal{K}_i^m$ of k nearest patches of $\mathbf{x}_i^m$ by computing the l_2 pairwise distance with $\{\mathbf{w}_j^m\}_{j=1}^N$, then the PEDAL can be formulated as follows:

$$\mathcal{L}_c^m = -\log \frac{\sum_{\mathbf{w}_j^m \in \mathcal{K}_i^m} e^{-\frac{s}{2}\left\|\mathbf{x}_i^m - \mathbf{w}_j^m\right\|_2^2}}{\sum_{j=1, j\neq i}^N e^{-\frac{s}{2}\left\|\mathbf{x}_i^m - \mathbf{w}_j^m\right\|_2^2}}, \quad (2)$$

where s is the scaling number. Minimizing $\mathcal{L}_c^m$ encourages the model to pull similar patches $\mathcal{K}_i^m$ close to $\mathbf{x}_i^m$ while pushing dissimilar patches $\{\mathbf{w}_j^m | \mathbf{w}_j^m \notin \mathcal{K}_i^m\}$ away from $\mathbf{x}_i^m$ in feature space. By this way, the model can learn how to map those visually similar patches closer so as to mine more visual consistent clues for these similar patches. Hence, the feature of these patches would be more discriminative.

Discussion. Compared to pulling the similar patches together, pulling the features of the similar person images together (as shown in Figure 3 (c)) would blur identity information of person image, making it ineffective to distinguish the similar images of different identities. In contrast, we divide the person image into several patches, so that different patches of the same image may contain different information of the person. Pulling the similar patches together could mine the latent discriminative clues for these similar patches, such as "yellow T-shirt" in Figure 3 (a). On the other hand, pulling the similar patches together may encounter the same problem as pulling the feature of the whole image, getting difficult to distinguish these similar patches. But the identity information is not simply encoded in the feature of patches, more importantly in the combination of patches. That means even if some patches of different identities are pulled together, we still can distinguish these identities by other patches. As shown in Figure 3 (a), although the model pulls the features of the patches of different identities with a yellow T-shirt together, we still can distinguish these pedestrians by the patches locating at the trousers (Figure 3 (b)). This mechanism is similar to use multi attributes to help identify the person.

Figure 4. Illustration of the cyclic ranking. We compute the ranking list $\mathcal{N}_i$ for the target image $\mathbf{x}_i$, Then we traverse the ranking list $\mathcal{N}_i$ in order, and we compute the ranking list for $\mathbf{x}_j \in \mathcal{N}_i$ until we find a hard negative sample. (Best viewed in color)

3.4. Image-level Patch Feature Learning

We develop an image-level loss additionally to further exploit image-level latent discriminative information which could be mined with the help of the discriminative patch features. An effective way is to minimize the intra-class gap while simultaneously maximize the inter-class gap in the feature space of the whole image. To this end, we introduce a cyclic ranking to mine the hard negative samples in a mini-batch, and we generate surrogate positive samples via a series of image transformations. Then, we develop a triplet-based loss function.

Mining hard negative samples in a mini-batch. Intuitively, if there are two image samples of the same iden-

tity in a mini-batch, then they are probably among the mutually nearest neighbors to each other. On the contrary, if two samples in a mini-batch are not the mutual neighbors to each other, this inconsistency indicates that they may have different identities. Based on the above discussion, we develop a cyclic ranking to mine hard negative samples in a mini-batch which is illustrated in Figure 4. Given a mini-batch of sample features $\{\mathbf{x}_i\}_{i=1}^{B}$, the ranking result of each sample $\mathbf{x}_i$ can be generated based on the pairwise similarity measure. We use the l_2 distance to measure the pairwise similarities, and thus we can get the ranking list $\mathcal{N}_i$ of $\mathbf{x}_i$. Then we traverse the ranking list $\mathcal{N}_i$ in order. For each negative sample candidate $\mathbf{x}_j \in \mathcal{N}_i$, we use the same method to compute the ranking list. Finally, if the $\mathbf{x}_i$ is **not** the top-r nearest neighbor of $\mathbf{x}_j$, then we argue that the $\mathbf{x}_j$ is likely to be a negative sample of $\mathbf{x}_i$. Furthermore, since hard negative pairs are more effective to learn a discriminative feature, we only consider the first (hardest) negative sample candidate $\mathbf{x}_j$ which matches the above condition. We denote this negative candidate as $\mathbf{n}_i$.

Discussion. The probability of the images to be the same identity in a mini-batch is very low when randomly sampling a few images from the dataset which includes a large number of images and identities [55], even when they are the mutual neighbors. But this is *not* impossible. In other words, it is difficult to estimate the likely binary label for the mutual neighbors without further scrutinization. Hence, we need a mechanism to determine the hard negative pair with higher confidence. The cyclic ranking provides such a mechanism to unsupervisedly mine the hard negative pair with a simple yet reasonable principle, which is important in the unsupervised discovery of the latent label information.

Surrogate positive samples. In [6], Dosovitskiy *et al*. propose using surrogate training data to learn discriminative feature for CNN under unsupervised setting and declare each set of transformed image patches to be a class. Similarly, in our experiments, we define a family of random transformation to generate the surrogate positive samples, including crop, scaling, rotation, brightness, contrast, and saturation of an image. Then we generate one surrogate positive sample for each real sample. Compared to Dosovitskiy *et al*. [6], the difference is that we perform such random transformation on images rather than image patches, and for each training epoch, we randomly generate one surrogate positive image for each anchor image to form a positive pair. Here, we give the definition of the IPFL as follows,

$$\mathcal{L}_v = \max\left\{\|\mathbf{x_i} - \mathbf{p}_i\|_2 - \|\mathbf{x_i} - \mathbf{n}_i\|_2 + m, 0\right\}, \quad (3)$$

where m is margin of the IPFL, $\mathbf{p}_i$ is a surrogate positive sample feature.

Figure 5. Some image examples of Market-1501, DukeMTMC-reID and MSMT17 dataset. Images of each column represent the same identity collected from different camera views.

3.5. Training the PatchNet on Unlabeled Dataset

As shown in Figure 2, we generate one surrogate positive sample for each image by using a series of random transformations, these surrogate positive samples are only used for computing IPFL. After that, the PGN generates M patches based on the feature map for each image. Finally, we compute the PEDAL for each patch and compute the IPFL for each triplet. The total loss function for each image in a mini-batch of our model can be formulated as

$$\mathcal{L} = \mathcal{L}_v + \lambda \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_c^m, \quad (4)$$

where λ controls the weight of the PEDAL.

4. Experiments and Analysis

4.1. Dataset and Evaluation Protocol

For validating the effectiveness of our proposed method, we carried out experiments on two large scale person re-id datasets, including Market-1501 [53], DukeMTMC-reID [26]. These two datasets are large-scale and have various variations including viewpoint change, occlusion, illumination, pose, background clustering. Specifically, the Market-1501 dataset contains 32668 images of 1501 identity, each of which was captured by at most six cameras. All the person images were detected automatically from video sequences. The DukeMTMC-reID dataset has 8 camera and 36411 labeled images belonging to 1404 identities. This dataset was constructed form the multi-camera tracking dataset DukeMTMC by random selection of manually tracklet bounding boxes. We followed the standard training/test split and evaluated the single-query test evaluation settings. For performance measurement, we used the cumulative matching characteristic (CMC) and the mean Average Precision (mAP).

4.2. Implementation Details

We chose the ResNet-50 [11] as our CNN backbone model which is pre-trained on the ImageNet dataset [4], and we removed the last fully-connected layer and the stride of last residual block is set to 1. The dimension of the output

Table 1. Performance (%) comparison on Market-1501 dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
LOMO [20]	27.2	41.6	49.1	8.0
Bow [53]	35.8	52.4	60.3	14.8
UMDL [25]	34.5	52.6	59.6	12.4
PUL [7]	45.5	60.7	66.7	20.5
CAMEL [46]	54.5	-	-	26.3
PTGAN [40]	38.6	-	66.1	-
SPGAN + LMP [5]	57.7	75.8	82.4	26.7
TJ-AIDL [37]	58.2	74.8	81.1	26.5
HHL [55]	62.2	78.8	84.0	31.4
DECAMEL [47]	60.2	76.0.	81.1	32.4
SyRI [1]	65.7	-	-	-
PAUL (Ours)	**68.5**	**82.4**	**87.4**	**40.1**

Table 2. Performance (%) comparison on DukeMTMC dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
LOMO [20]	12.3	21.3	26.6	4.8
Bow [53]	17.1	28.8	34.9	8.3
UMDL [25]	18.5	31.4	37.6	7.3
PUL [7]	30.0	43.4	48.5	16.4
PTGAN [40]	27.4	-	50.7	-
SPGAN + LMP [5]	46.4	62.3	68.0	26.2
TJ-AIDL [37]	44.3	59.6	65.0	23.0
HHL [55]	46.9	61.0	66.7	27.2
PAUL (Ours)	**72.0**	**82.7**	**86.0**	**53.2**

feature of each branch is 256. The patch generation network (PGN) was initialized such that the feature map would be divided into M equal-sized horizontal stripes. For the **p**atch-bas**e**d **d**iscriminative fe**a**ture learning **l**oss (PEDAL), we only considered the the top-10 ($k = 10$) nearest patches of other images and the updating rate l of the memory bank was set to 0.1. The scaling number s was set to 10 for DukeMTMC-reID and 30 for Market-1501, respectively, to ensure the convergence of the model as suggested in [33]. For the **i**mage-level **p**atch **f**eature learning **l**oss (IPFL), we generated one surrogate positive sample for each image, r is set to 3, and the margin is set as 2 empirically. The weight λ of $\mathcal{L}_c$ is set to 2. We used the MSMT17 [40] to pre-train the PatchNet. Then we fixed the PGN and we did not use the MSMT17 again during training on other dataset. During training on unlabeled dataset, the images were randomly sampled from the training set of dataset and resized to 384×128. Each mini-batch was composed of 40 real samples, and 40 surrogate positive samples additionally which were only used for computing the IPFL. We used the SGD [32] as our optimization algorithm, and the learning rate was set to 0.0001 initially and decayed by 0.1 every 50 epochs. We trained the model on the unlabeled datasets for 60 epochs. During testing, we concatenated the patch features of the same image together to compute the pairwise distance. The random transformations we used to generate the surrogate positive samples are as follows:

- Crop: randomly cropping each image into a size from 70% to 95% of the original size;
- Rotation: randomly rotating of the image by an angle up to 10 degrees;
- Contrast, saturation, and brightness: randomly changing it from 80% to 120% of the original image for each image;

4.3. Comparison with the State-of-the-art

We compared the proposed method with hand-crafted features (including LOMO [20], BoW [53] and UMDL [25]) and the state-of-the-art unsupervised learning methods for person re-id. The results of the comparisons on Market-1501 dataset are presented on Table 1 and the results on DukeMTMC-reID dataset are shown in Table 2. We can see that our proposed method outperforms the compared methods significantly on both datasets.

Specifically, the clustering-based methods (PUL [7], CAMEL [46] and DECAMEL [47]) may assign the same pseudo label to similar images of different identities, On the contrary, even though some patches of the image pull some patches of other identities together, there are still other patches of the image to provide discrimination.

Compared with transfer learning methods (including PTGAN [40], SPGAN+LMP [5],TJ-AIDL [37], HHL [55] and SyRI [1]), our proposed method outperforms the these methods with a significantly margin. The main reason can fall into two aspects. (1) The gap between the images of the source and target domains is larger compared to the gap between image patches, so it is harder to transfer an image based feature learning model to the target domain. (2) The PatchNet can learn discriminative features by optimizing the PEDAL and the IPFL on an unlabeled dataset. We note that the SyRI [1] uses the CUHK03 [18], DukeMTMC-reID and synthetic data (totally 3379 identity) to train their model by a three-step domain adaptation, while our proposed method can be directly trained on the unlabeled dataset without bells and whistles. Also note that TJ-AIDL [37] is somewhat related to the patch-based discriminative feature learning, because the similar patches of different images are likely sharing the same attribute information. However, TJ-ALDL requires extra attribute labels, limiting its scalability.

In summary, our method can learn the discriminative features in a patch level which is more generalizable among different datasets. In contrast, existing methods could not achieve this goal. In addition, with the guidance of the PEDAL and the IPFL, the PatchNet can be conveniently trained on an unlabeled dataset.

4.4. Ablation Study

We perform abaltion study to evaluate the effectiveness of each component in our method. The experimental results are reported in Table 3. The comparison of the "ResNet-50" and "PatchNet" validates that patch-based feature learning method is more generalizable and it can learn more discriminative feature on an unlabeled dataset because the patches are common among different datasets. We also evaluate the

Figure 6. A visualization of the nearest patches in the patch-based discriminative feature learning loss (Eq. (2)). We also show the whole images corresponding to the patches. Blue bounding box indicates the same identity. Red bounding box indicates the location of the patch.

Table 3. Ablation study (%). The results of "ResNet-50" and "PatchNet" mean we directly test the model (pre-trained on MSMT17) without training on unlabeled datasets.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
ResNet-50 [11] †	46.6	22.7	52.6	33.1
ResNet-50 [11] + $\mathcal{L}_c$ ‡	25.6	11.8	29.5	15.4
PatchNet (Baseline)	59.3	31.0	65.7	45.6
PatchNet+$\mathcal{L}_c$	66.2	38.0	70.6	52.1
PatchNet+$\mathcal{L}_v$	65.4	37.6	67.1	48.0
PatchNet+$\mathcal{L}_v$ + $\mathcal{L}_c$ (PAUL)	**68.5**	**40.1**	**72.0**	**53.2**

† Image-level feature learning.
‡ Here, $\mathcal{L}_c$ means pulling the features of the similar images together.

PEDAL and the IPFL separately and jointly to validate its effectiveness. We can observe that our proposed method perform better than baseline model with a significantly margin.

The effectiveness of PEDAL. We train the PatchNet only with the $\mathcal{L}_c$ to validate its effectiveness. From Table 3, we can see that the "PatchNet+$\mathcal{L}_c$" outperforms the "PatchNet" significantly on both dataset. The main reason is that the PEDAL can provide effective guidance for PatchNet to further refine the feature of patches on unlabeled datasets. We also can observe that if we pull the image features of similar person image together (i.e. "ResNet-50+$\mathcal{L}_c$"), the performance is worse than the "ResNet-50". This experiment shows that pulling the image feature together would blur the identity information, making it less discriminative to similar person.

The effectiveness of IPFL. The result of "PatchNet+$\mathcal{L}_v$" is clearly better than the baseline on both datasets. This is because the IPFL can provide effective learning guidance to the PatchNet by generating surrogate positive samples and mining hard negative samples in a mini-batch.

The effectiveness of the combination of PEDAL and IPFL. As shown in Table 3, the combination of PEDAL and IPFL achieves the best results compared to all other variants. This validates that the two losses are mutually complementary, since they function in different level aspects, i.e. the patch level and the image level. Specifically, the PEDAL can lead the PatchNet to learn discriminative patch feature on the unlabeled datasets, Therefore, concatenated feature (i.e. the image feature) would be more discriminative, facilitating the hard negative mining process of IPFL.

Table 4. Analysis on the proposed method with different patch generation schemes (%). For each patch generation schemes, we train the PatchNet on the MSMT17 and then perform unsupervised training with PEDAL and IPFL.

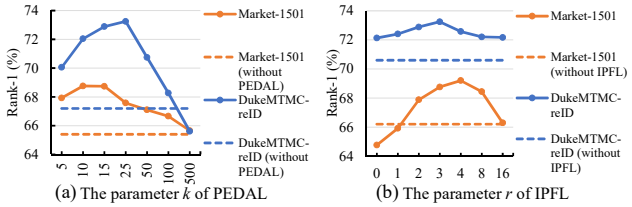
Generation schemes	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
Randomly	24.1	14.8	16.0	10.9
Equally	66.6	38.5	56.2	39.4
PGN	**68.5**	**40.1**	**72.0**	**53.2**
4 patches	67.3	39.5	70.4	50.6
6 patches	**68.5**	**40.1**	**72.0**	**53.2**
8 patches	66.7	37.2	70.6	51.8

4.5. Further Analysis

Visualization. To further understand the patch-based discriminative learning, we show a typical case of the nearest patches for different branches in Figure 6. In the upper row in Figure 6, we observe that our model pulls closer the red T-shirt patches which are very likely from the same identity. Although we may also pull closer the red T-shirt patches from other persons, they typically have other discriminative patches, e.g. different pants. The other discriminative patch features would be learned in other branches (see the lower row in Figure 6 for the pants patches). Therefore, although depending only on one patch is not identity-discriminative enough, combining different patches could further boost the discriminability by helping to distinguish these partly similar persons.

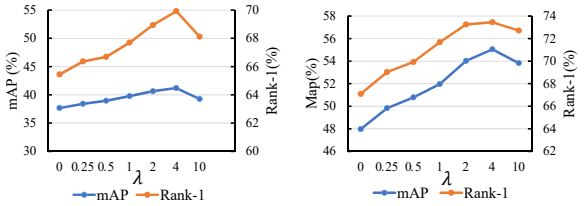
Analysis on the PGN. To validate the effectiveness of the PGN, we compare the PGN with two patch generation schemes, i.e. randomly selecting M patches from the feature map or dividing the feature map into M horizontal stripes equally. We also analyze the patch number M. The results are shown in Table 4. We can observe that the performance of the PGN is better than the randomly selection and the equal horizontal partition, because the PGN is learnable and thus it can adaptively adjust the locations of the patches to find more effective patches according to the dataset.

The effect of the parameter k of PEDAL. The parameter k of PEDAL decides how many patches of other images are regarded as the similar patches of the target patch, then the model would pull these patches closer and push other patches away. As shown in Figure 7 (a), if the k is too small, the IPFL may miss some similar patches so the performance



(a) The parameter k of PEDAL (b) The parameter r of IPFL

Figure 7. Experimental analysis on the parameter k of PEDAL (a) and the parameter r of IPFL (b). These experiments were carried out on Market-1501 and DukeMTMC-reID.



(a) Market-1501 (b) DukeMTMC-reID

Figure 8. Analysis on the weight of the PEDAL.

degrades. By contrast, if the k is too large, the PEDAL maybe pull some significantly dissimilar patches of other images closer, result in a worse performance. As shown in Figure 7 (a), the proposed PEDAL *consistently* improves the baseline on the two tested datasets when $k \in [5, 100]$, with optimal performances achieved consistently when $k \in [10, 25]$.

The effect of the parameter r of IPFL. The smaller r means the top-n nearest neighbor is easier to be regarded as the hard negative sample of $\mathbf{x}_i$. Particularly, $r = 0$ means the top-1 nearest of $\mathbf{x}_i$ is directly regarded as the hard negative sample. In this case, it is easily to regard another image of the same identity as the the hard negative sample. But if the r is too large, the negative sample is easy to distinguish so that the CNN can not benefit from it much. Specifically, for the parameter r, the proposed IPFL consistently provides improvements when $r \in [2, 8]$ on both datasets, as shown in Figure 7 (b).

The weight of the PEDAL. The analysis on the weight of the PEDAL is reported in Figure 8. Combined with Table 3, we can observe that the combination of the two losses can achieve a better result. The PEDAL provides the patch-level guidance to learn a more discriminative feature and the IPFL serves as the pairwise guidance for PatchNet. Additionally, we can observe that the PEDAL contributes more effective guidance for the PatchNet.

The universality of the proposed method. To further validate the universality of our proposed method, more experimental results on other person Re-ID datasets (including CUHK01 [17], CUHK03 [18], and VIPER [10]) are presented in Table 5. We can see that our method is universally effective on other datasets with the same parameter values.

Analysis on the pre-trained dataset. In order to evaluate the effect of the pre-trained dataset, we pre-train the PatchNet on other datasets and compare with other methods that use the same labeled dataset. The result is reported in Table 6, where we can observe that our method significantly outperforms the compared methods. By comparing Table 6 and Table 2, we can see that the MSMT17 significantly improves the performance on DukeMTMC-reID dataset. This may because the DukeMTMC-reID has more common patches with MSMT17 compared to the Market-1501.

Table 5. Performance (%) comparison with currunt state-of-the-art method on CUHK01, CUHK03, and VIPER. We fixed the same parameter values as the experiments on Market-1501 and DukeMTMC-reID)

Methods	CUHK03 [18]	CUHK01 [17]	VIPER [10]
CAMEL [46]	31.9	57.3	30.9
PatchNet (baseline)	45.4	69.9	40.8
PAUL (ours)	**52.3**	**73.3**	**45.2**

Table 6. The performance (%) of our proposed method when the PatchNet was pre-trained on the Market-1501 or the DukeMTMC-reID dataset.

Source	DukeMTMC-reID		Market-1501	
Target	Market-1501		DukeMTMC-reID	
Methods	Rank-1	mAP	Rank-1	mAP
PUL [7]	45.5	20.5	30.0	16.4
PTGAN [40]	38.6	-	27.4	-
TJ-AIDL [37]	58.2	26.5	44.3	23.0
HHL [55]	62.2	31.4	46.9	27.2
PAUL (Ours)	**66.7**	**36.8**	**56.1**	**35.7**

5. Conclusion

In this paper, we demonstrate the effectiveness of the local feature learning in unsupervised re-id by proposing a novel unsupervised patch-based discriminative learning which enables the local feature learning in an unlabeled re-id dataset. Specifically, we propose a patch-based unsupervised learning framework (PAUL), in which the PatchNet is designed to sample patches from a feature map of the person image and to learn discriminatively the patch feature on an unlabeled re-id dataset. To this end, we develop a patch discriminative feature learning loss to provide effective guidance to learn discriminative patch feature on the unlabeled re-id dataset. Simultaneously, we further propose a image-level patch feature learning loss to mine the latent pairwise relationship of the whole unlabeled images with the guidance of the patches. Extensive experiments validate the effectiveness of our proposed method as well as each learning component in our model.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC(61522115), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), and the Royal Society Newton Advanced Fellowship (NA150459).

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, 2018.

[2] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *CVIU*, 2013.

[3] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2018.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018.

[6] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*, 2014.

[7] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*, 2017.

[8] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[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] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

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

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[14] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised $l1$ graph learning. In *ECCV*, 2016.

[15] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015.

[16] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[17] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

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

[21] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[22] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014.

[23] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017.

[24] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[25] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016.

[26] 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.

[27] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[28] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, 2016.

[29] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. *arXiv preprint arXiv:1804.07094*, 2018.

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

[31] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. *arXiv preprint arXiv:1711.09349*, 2017.

[32] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

[33] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: l 2 hypersphere embedding for face verification. In *ACMMM*, 2017.

[34] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[35] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014.

[36] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016.

[37] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning

for unsupervised person re-identification. *arXiv preprint arXiv:1803.09786*, 2018.

[38] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[39] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *TPAMI*, 2016.

[40] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[41] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[42] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[43] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[44] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017.

[45] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[46] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[47] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *TPAMI (DOI 10.1109/TPAMI.2018.2886878)*, 2019.

[48] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. *arXiv preprint arXiv:1706.00384*, 6, 2017.

[49] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[50] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[51] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[52] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

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

[54] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013.

[55] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, 2018.