Unleashing Potential of Unsupervised Pre-Training with Intra-Identity Regularization for Person Re-Identification

Zizheng Yang Xin Jin Kecheng Zheng Feng Zhao*
University of Science and Technology of China
{yzz6000,jinxustc,zkcys001}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

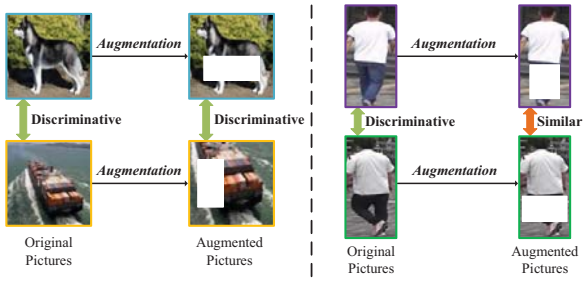
Abstract

Existing person re-identification (ReID) methods typically load the pre-trained ImageNet weights for initialization directly. However, as a fine-grained classification task, ReID is more challenging and there exists a large domain gap between ImageNet classification. Inspired by the great success of self-supervised representation learning with contrastive objectives, in this paper, we design an Unsupervised Pre-training framework for re-identification (UP-ReID) based on the contrastive learning (CL) pipeline. During the pre-training, we attempt to address two critical issues for learning fine-grained ReID features: (1) the augmentations in the CL pipeline usually distort the discriminative clues in person images, and (2) the fine-grained local features of person images are not fully-explored. Therefore, we introduce an ***intra-identity*** *(I^2-)regularization in the UP-ReID, which is instantiated as two constraints coming from the global image and local patch aspects, respectively. A global consistency constraint is enforced between augmented and original person images to increase robustness to augmentation, while an intrinsic contrastive constraint among local patches of each image is employed to fully explore the local discriminative clues. Extensive experiments on multiple popular Re-ID datasets, PersonX, Market1501, CUHK03, and MSMT17, demonstrate that our UP-ReID pre-trained model can significantly benefit the downstream ReID fine-tuning and achieve state-of-the-art performance.*

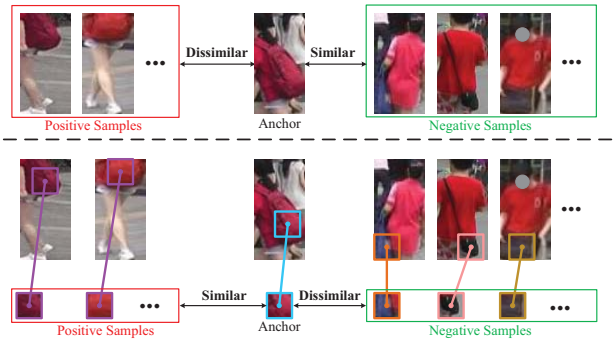
1. Introduction

As a fine-grained classification problem, person re-identification (ReID) aims at identifying a specific person across non-overlapping camera views. Existing ReID methods have achieved a remarkable success in both supervised [25,27,35,37,42,46] and unsupervised [10,13,14,26,30,45] domains. Most of these approaches directly leverage the weights pre-trained on ImageNet for model initialization, which may not be optimal for ReID tasks, resulting in poor fine-tuning performance and slow convergence [14, 42]. The main reasons stem from two aspects: inapplicable pre-training method (ImageNet is more like a coarse-grained classification) and large domain gap between ImageNet and ReID datasets. Thus, how to efficiently pre-train a good ReID-specific initialization network is still under-explored.

Unsupervised pre-training has achieved a fast development with the great success of contrastive learning [1, 4, 6,



(a) Left: the two augmented images are still discriminative for general classification tasks. Right: the discriminative attributes of the two person images are ruined by augmentation for person ReID.



(b) Top: a case that uses the global features for a failed ReID, where the positive samples share dissimilar appearance but the negative samples have a similar appearance instead. Bottom: a case that uses the fine-grained discriminative attributes, such as backpacks and bags, for a successful ReID where the person images are distinguishable and independent to clothing.

Figure 1. Two critical issues in the existing contrastive learning-based pre-training methods, which should be well solved in the ReID-specific pre-training framework.

*Corresponding Author.

7, 18], which is taken as a pretext work, serving for different downstream supervised or unsupervised ReID fine-tuning algorithms. Beyond the general pre-training task, this paper aims to propose a ReID-specific pre-training framework (*e.g.*, pre-training a ResNet50 [19] for learning discriminative ReID representations) on a large-scale unlabeled dataset. The pioneering work in [12] makes the first attempt on ReID pre-training and introduces a new large-scale unlabeled ReID dataset LUPerson. However, it directly transfers the general pre-training process based on contrastive learning that designed for ImageNet classification to ReID task, which ignores the fact that ReID is a fine-grained classification problem. This solution faces the following two critical issues:

The first one comes from the used augmentations in the existing contrastive learning pipeline, which could possibly damage the discriminative attributes of person images. As shown in Figure 1a, different from the coarse-grained classification problem on ImageNet, the discriminative attributes of person images are prone to be destroyed by the augmentation operations. For example, in the ImageNet classification, although the augmentations applied to the pictures (*e.g.*, dogs and ships) may cause the lack of regional information, the remaining parts are still discriminative enough to support the model for distinguishing them. However, when applying the same augmentations to person images in ReID, it will provoke a disaster, the most discriminative attributes (*i.e.*, trousers color) of person images are destroyed, making them indistinguishable.

The second one is that the fine-grained information of person images is not fully explored in previous pre-training methods. They typically only care about the learning of image-level global feature representations. Nevertheless, as a fine-grained classification task, ReID needs detailed local features in addition to global ones for the accurate identity matching [40, 42, 45]. As illustrated in Figure 1b, the local fine-grained clues (*e.g.*, backpacks, cross-body bags) are more helpful than global features w.r.t distinguishing different persons.

To address the above issues, we introduce an *intra-identity (I^2-)* regularization in the proposed ReID-specific pre-training framework UP-ReID. It consists of a *global consistency* constraint between augmented and original person images, and an *intrinsic contrastive* constraint among local patches of each image. Specifically, we first enforce a global consistency to make the pre-training model be more invariant to augmentations. We feed the augmented images as well as the original images into the model and then narrow the similarity distance between them in distributions. Second, we propose an *intrinsic contrastive* constraint for the local information exploration. Instead of directly feeding the holistic augmented images, we partition them into multiple patches and then send these patches along with the holistic images to the network. After that, we compute an intrinsic contrastive loss among patches to encourage the model to learn both fine-grained and semantic-aware representations. Moreover, based on the prior knowledge that human body is horizontally symmetric, we establish a hard mining strategy for the calculation of this loss, which makes the training stable and thus improves the generalization ability of the pre-trained model.

We summarize our main contributions as follows:

- To the best of our knowledge, the proposed UP-ReID is the first attempt toward a ReID-specific pre-training framework by explicitly pinpointing the difference between the general pre-training and ReID pre-training.
- Considering the particularity of ReID tasks, an intra-identity (I^2-)regularization is introduced in our UP-ReID, which is instantiated from the global image level and local patch level.
- In the I^2-regularization, a global consistency is first enforced to increase the robustness of pre-training to data augmentations. An intrinsic contrastive constraint with prior-based hard mining strategy among local patches of person images is further introduced to fully explore the local discriminative clues.

Extensive experiments on multiple widely-used ReID benchmarks demonstrate the effectiveness of the proposed UP-ReID, which outperforms other state-of-the-art pre-training methods by prominent margins and could benefit a series of downstream ReID-related tasks.

2. Related Work

2.1. Person ReID

Fully-supervised ReID approaches. Fully-supervised ReID methods are based on supervised learning with labeled datasets and have achieved a great success [27, 28, 35, 46]. These works can be divided into two mainstream branches. One focuses on designing effective optimization metrics (i.e., metric learning) for person ReID, such as hard triplet loss [23] and circle loss [39]. On the other hand, learning fine-grained features is also a popular branch. PCB [40] and MGN [42] both leverage local features of pedestrian images by manually splitting each holistic image into multiple sub-parts to achieve accurate person ReID.

Unsupervised ReID approaches. There are two typical categories of unsupervised person ReID: Unsupervised Domain Adaptation (UDA)-based methods and Domain Generalization (DG)-based methods. 1) UDA could handle the domain gap issue when the target domain data are accessible, which aims to learn a generic model from both labeled source data and unlabeled target data. The UDA-based methods can be further categorized into three main

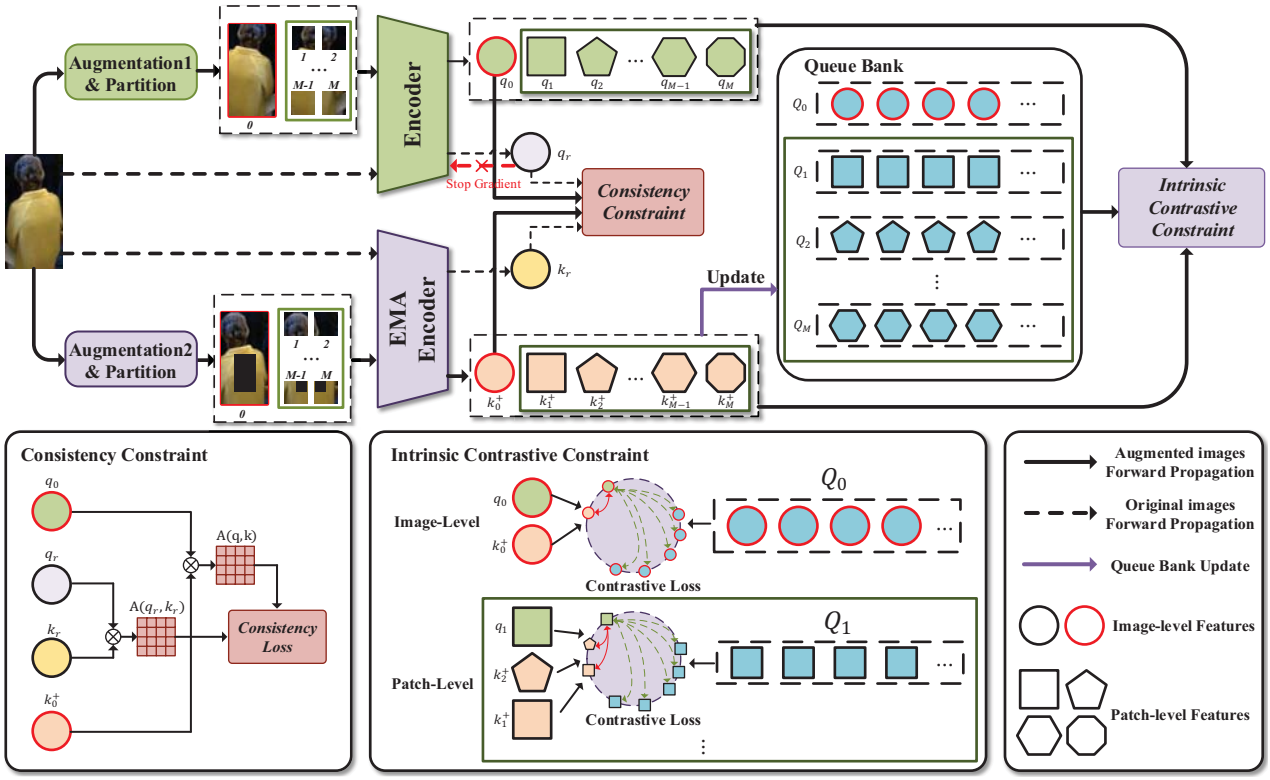


Figure 2. Architecture of the proposed UP-ReID. Given an input image, we can get two different groups of augmented instances after two different augmentations and partition. Then, we feed them into the online encoder and EMA encoder respectively, together with the original images. A consistency loss is computed to narrow the gap between the similarity distribution of the augmented images and that of the original images. We also compute an intrinsic contrastive loss based on a delicately designed hard mining strategy. The EMA encoder features are used to update the queue bank. The online encoder is optimized by the gradient of the total loss, while the EMA encoder is updated by momentum-based moving average of the online encoder.

classes: style transfer based works [11, 44, 50], attribute recognition based works [34, 43] and pseudo labeling based works [13, 14, 36]. 2) DG is designed for a more challenging case where the target domain data are unavailable. Jin *et al.* [26] designs a Style Normalization and Restitution (SNR) module to enhance the identity-relevant features and filter out the identity-irrelevant features for improving the model's generalization ability. In addition, meta-learning [48] is also employed as a popular way to achieve person ReID specific domain generalization. Contrastive learning-based methods [3, 24] also achieves a great success in purely unsupervised ReID on small-scale datasets. However, all these methods generally load the pre-trained ImageNet weights for initialization, and ignore the gap between the ImageNet classification and the fine-grained ReID task.

2.2. Self-Supervised Representation Learning

Based on the recently popular contrastive learning, the unsupervised pre-training has achieved a great success, and many representative works have achieved comparable or even slightly better performance than supervised works. MoCo [18] and MoCo v2 [6] design a dynamic queue and introduce a momentum update mechanism to optimize a key encoder progressively. SimCLR [4] and SimCLR v2 [5] also achieve great performance with a large batch size, rich data augmentations and a simple but effective projection head. BYOL [16] and SimSiam [7] further achieve great performance even without negative pairs. SwAV [1] replaces comparison between pairwise samples with comparison between cluster assignments of multiple views.

The work of [12] proposes a new large-scale unlabeled dataset "LUPerson" which is large enough to support pre-training and makes the first attempt to pre-train specific models for person ReID initialization. However, since the work merely migrates the approach of pre-training models on ImageNet to ReID directly, it suffers from the instability issue (see Figure 1a) caused by augmentation and lacked of the exploration of fine-grained discriminative information of pedestrian images (see Figure 1b). In this work, we study how to design a pre-training framework that avoids data augmentation interference while fully using fine-grained local information for discriminative representation learning.

3. Unsupervised Pre-training for ReID

Person ReID training typically contains two procedures of pre-training and fine-tuning: (a) the model (*e.g.*, ResNet50) is first *pre-trained* unsupervisedly on a large-scale dataset (*e.g.*, LUPerson [12]) with a pretext task, (b) and then the pre-trained model is utilized to initialize the backbone and *fine-tuned* with small-scale labeled or unlabeled person ReID datasets (*e.g.*, Market1501 [49]). In this paper, we focus on the first phase, *i.e.*, how to pre-train a ReID-friendly model in an unsupervised manner.

We first overview the whole pipeline of our UP-ReID in Section 3.1, and then introduce the proposed I^2-regularization for pre-training, which comprises a global consistency constraint (Section 3.2) and an intrinsic contrastive constraint (Section 3.3). Last but not least, a prior-based hard mining strategy employed for local feature enhancement is discussed in Section 3.4.

3.1. Overview

As illustrated in Figure 2, UP-ReID has two encoders: an online encoder f_q and a momentum-based moving averaging (EMA) update encoder f_k. Both f_q and f_k are composed of a feature encoder and a projection head. The feature encoder is the model to be pre-trained (*e.g.*, ResNet50), and the projection head is a multi-layer perceptron. The online encoder f_q will be updated by back-propagation, while the EMA encoder f_k will be slowly progressed through momentum-based moving average of the f_q, which is $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$. θ_k, θ_q represent the parameters of f_k, f_q, and m means the momentum coefficient.

Given an input image x, we can get two different views of x after two different augmentations: a query view $x_{q,0}$ and a key view $x_{k,0}$. Unlike previous contrastive learning methods that only take the augmented images $x_{q,0}$ and $x_{k,0}$ as the input, we also feed the original image x into the network as shown in Figure 2. Then, we enforce a consistency loss $\mathcal{L}_{consist.}$ to narrow down the distance between the similarity distribution of the augmented images and that of the original images in a mini-batch, which is described in detail in Section 3.2.

Moreover, before feeding $x_{q,0}$ and $x_{k,0}$ into the network, we partition each of them into M non-overlapping patches. Note that, all $2M$ patches are partitioned from the same person image x actually. Then, we feed these patches along with the entire augmented images into the online encoder and EMA encoder. An intrinsic contrastive loss $\mathcal{L}_{inc}$ is computed over them to learn both fine-grained local representations and the semantic image-level representations, which is discussed in detail in Section 3.3. For a better fine-grained information exploration, a hard mining strategy is further introduced to the calculation of the intrinsic contrastive loss, which is presented in Section 3.4. Ultimately, the total optimization objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{consist.} + \mathcal{L}_{inc}. \tag{1}$$

Additionally, a dynamic queue bank is constructed to store the feature representations of previous mini-batches and provide sufficient negative samples for the current mini-batch training. In practice, we prepare a queue for the image-level features, *i.e.*, Q_0, and a queue for each patch-level local features, *i.e.*, Q_i, $i \in \{1, ..., M\}$. All of these queues together constitute the queue bank and they will be dynamically updated using the features extracted by the EMA encoder.

3.2. Consistency over Augmented-Original Images

Data augmentation plays a crucial role in contrastive learning. However, discriminative attributes of pedestrian images are very likely to be ruined by various augmentation operations (see Figure 1a). Due to the visual distortions caused by augmentation, a sample may be less similar to its positive instances but more similar to its negative samples instead, which inevitably imposes a negative effect on the pre-training process.

To alleviate this problem, we turn to the original images for help. Although the identity-related features are possibly destroyed in the augmented images, those discriminative clues still remain in the original person images, *i.e.*, raw images before augmentation. Thus, we propose to use the similarity between the original images as ground truth to supervise the images that go through the data augmentation, *i.e.*, maintain the consistency before and after data augmentations.

For a mini-batch of input person images x_r, two groups of images x_q and x_k are generated after two different augmentations, which are then fed into the network to produce the online encoder features q and EMA encoder features k respectively: $q = f_q(x_q)$ and $k = f_k(x_k)$. The similarity distribution is computed as:

$$A(q, k) = q \cdot k^T, \tag{2}$$

where q and k have been normalized by the normalization layer followed by the projection head, and $A(\cdot)$ denotes the inter-instance similarity calculation function between two batches of images after two different kinds of augmentation.

Similarly, we perform the same operation on the original input images x_r, which is expressed as $q_r = f_q(x_r)$ and $k_r = f_k(x_r)$. Then, we calculate the inter-instance similarity distribution of the original images:

$$A(q_r, k_r) = q_r \cdot k_r^T. \tag{3}$$

After that, we employ a maximum mean discrepancy (MMD) [15] metric to measure the difference between two distributions and construct a consistency loss based on it:

$$\mathcal{L}_{consist.} = MMD\left(A(q, k),\ A(q_r, k_r)\right). \tag{4}$$

Note that the calculated similarity distribution over the original images $A(q_r, k_r)$ just serves as the ground truth to supervise that of the augmented images $A(q, k)$ and does not participate in the update. So, there is no gradient back-propagation for the features of the original images. The consistency loss $\mathcal{L}_{consist.}$ helps the model to deduce and restore the discriminative local regions that are distorted by data augmentations, and further encourages the model to learn discriminative feature representations between different instances.

3.3. Intrinsic Contrastive Constraint

To explore the intrinsic properties of a person image, we also introduce an intrinsic contrastive constraint in our UP-ReID framework. Before feeding the augmented images of $x_{q,0}$ and $x_{k,0}$ into the network (here we use subscript '0' to denote the holistic person image), we partition each of them into M non-overlapping patches uniformly,

$$\{x_{q,1}, ..., x_{q,M}\} = P(x_{q,0}), \tag{5}$$

$$\{x_{k,1}, ..., x_{k,M}\} = P(x_{k,0}), \tag{6}$$

where P represents the partition operation, $x_{q,i}$ denotes the i-th patch partitioned from $x_{q,0}$, and $x_{k,i}$ denotes the i-th patch partitioned from $x_{k,0}$. Then, we group them together to obtain two sets: $\mathcal{X}_q = \{x_{q,i}\}_{i=0}^{M}$ and $\mathcal{X}_k = \{x_{k,i}\}_{i=0}^{M}$.

Taking the image set $\mathcal{X}_q$ as an example for illustration, it comprises an image-level holistic instance $x_{q,0}$ and M patch-level local instances $x_{q,i}$ ($i \in \{1, ..., M\}$). All of them come from the same input image x and belong to the same instance, *i.e.*, the input x. In short, $x_{q,0}$ contains the image-level global information while $x_{q,i}$ ($i \in \{1, ..., M\}$) highlights the local information.

As shown in Figure 2, we feed $\mathcal{X}_q$ and $\mathcal{X}_k$ into the online encoder f_q and EMA encoder f_k, respectively, *i.e.*, $q_i = f_q(x_{q,i})$ and $k_i^+ = f_k(x_{k,i})$, $i \in 0, 1, ..., M$. To learn semantic-aware representations from the holistic images, we enforce a InfoNCE [32] loss over the global features, which is formulated as:

$$\mathcal{L}_g = -\log \frac{\exp(q_0 \cdot k_0^+ / \tau_1)}{\exp(q_0 \cdot k_0^+ / \tau_1) + \sum_{j=0}^{N-1} \exp(q_0 \cdot k_{0,j}^- / \tau_1)}, \tag{7}$$

where τ_1 is the temperature hyper-parameter, $k_{0,j}^-$ is the negative sample in the image-level feature queue Q_0, and N is the total number of negative samples in Q_0.

For the local fine-grained representation learning, we calculate a patch-wise contrastive loss over the patch-level instances. For the feature q_i, we denote its positive sample as k_p^+ and negative queue as Q_n. Formally, the patch-wise contrastive loss for the i-th patch p_i is defined as:

$$\mathcal{L}_{p_i} = -\log \frac{\exp(q_i \cdot k_p^+ / \tau_2)}{\exp(q_i \cdot k_p^+ / \tau_2) + \sum_{j=0}^{N-1} \exp(q_i \cdot k_{n,j}^- / \tau_2)}, \tag{8}$$

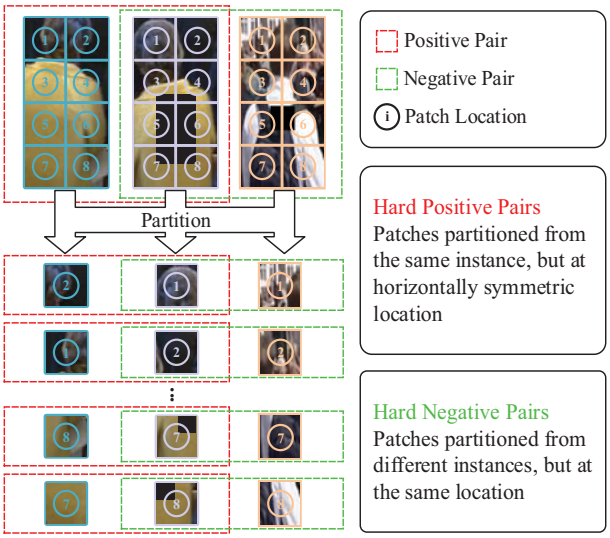


Figure 3. Illustration of our hard mining strategy. We choose two horizontally symmetric patches partitioned from the same instance as a positive pair, and two patches partitioned from different instances but at the same patch location as a negative pair.

where $k_{n,j}^-$ is the negative sample in Q_n, and τ_2 is the temperature hyper-parameter. The details about the selection of k_p^+ and Q_n will be described in Section 3.4.

In order to fully explore the discriminative information contained in each body part of a pedestrian, we compute the aforementioned contrastive loss for each patch-level feature and take the weighted average sum of them as the final constraint. That is, the intrinsic contrastive loss is a weighted sum of $\mathcal{L}_g$ and multiple $\mathcal{L}_{p_i}$:

$$\mathcal{L}_{inc} = \lambda_g * \mathcal{L}_g + \lambda_p * \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_{p_i}, \tag{9}$$

where λ_g and λ_p are the weighting parameters.

3.4. Hard Mining for Local Feature Exploration

For the patch-level feature q_i in Eq. 8, k_i^+ and Q_i should be the positive sample and negative queue, respectively, *i.e.*, $k_p^+ = k_i^+, Q_n = Q_i$, corresponding to same patch region as q_i. For better representation learning, based on the prior knowledge that human body is horizontally symmetric, we further develop an effective hard mining method to select the positive sample and negative queue for each patch-level feature, which is shown in Figure 3.

Hard Negative Queue Selection. The same body part of different persons could be discriminative, such as hair color and shoes color. Hence, for q_i in Eq. 8, we choose the patches partitioned from different instances but at the same location as the negative samples (*i.e.*, $Q_n = Q_i$).

Hard Positive Sample Selection. Considering the prior knowledge that persons are horizontally symmetric, we

choose two horizontally symmetric patches partitioned from the same instance as a positive pair. Specifically, in Eq. 8, we select the feature $k^+_{i\_hs}$ (*i.e.*, the horizontally symmetrical patch feature corresponding to position i) as the positive sample of q_i.

Intuitively, the human body structure and clothing are mostly horizontally symmetric, which indicates that two symmetric patches of the same person image contain very similar visual representative patterns (*e.g.*, color, texture). This is important for person ReID. Thus, choosing them as a positive pair to pre-train the model is reasonable. On the other hand, due to the different capture environments caused by camera angles or human postures, a pedestrian image may not be completely symmetric, which means two symmetric patches have similar primary visual information but there are still differences in details. Hence, choosing them as a positive pair can improve the model's ability to identify similar visual representation patterns under different situations, which further helps the model recognize the same pedestrian in various environments.

Given that there are still some extreme cases that are totally inconsistent with the prior knowledge of horizontal symmetry of the pedestrian pictures (*e.g.*, pedestrian pictures taken from the side), we also select the same position patch of the other view person image (*i.e.*, k^+_i) as one of the positive samples of q_i. So, the patch-wise contrastive loss in Eq. 8 is modified to:

$$\mathcal{L}_{p_i} = -\log \frac{\sum_{k^+_p \in \mathcal{P}(i)} \exp(q_i \cdot k^+_p / \tau_2)}{\sum_{k^+_p \in \mathcal{P}(i)} \exp(q_i \cdot k^+_p / \tau_2) + \sum_{j=0}^{N-1} \exp(q_i \cdot k^-_{i,j} / \tau_2)}, \tag{10}$$

where $\mathcal{P}(i) = \{k^+_{i\_hs}, k^+_i\}$, $k^-_{i,j} \in Q_i$.

4. Experiments

4.1. Implementation

Training details. For fair comparison, we use ResNet50 as the pre-trained backbone model and SGD as the optimizer. The input images are resized to 256×128. The mini-batch size is set to 800, and the initial learning rate is 0.1. In our experiments, M is set to 8, N is set to 65536, m is set to 0.9, τ_1 and τ_2 are both set to 0.1, λ_g and λ_p are set to 0.8 and 0.2. The pre-training models are trained with 8×2080Ti GPUs for 3 weeks under PyTorch framework.

Augmentation and Patch Partition. Data augmentation plays a crucial role in self-supervised contrastive learning. We utilize the same augmentation operations as [12]. As for partition, we adopt image-level partition strategy. Specifically, we first partition a holistic image into multiple horizontal stripes, and then divide each stripe vertically into two patches uniformly. It is necessary to emphasize that we apply the global-level augmentation (*i.e.*, augmentation followed by partition) rather than patch-level augmentation (*i.e.*, partition followed by augmentation). Because the global-level augmentation is closer to the realistic data variation and will not break the inherent consistency among patches partitioned from the same person image.

Datasets. We pre-train our model on "LUPerson" [12] dataset. To demonstrate the superiority of our pre-trained model, we conduct extensive downstream experiments on four public ReID datasets, including CUHK03 [28], Market1501 [49], PersonX [38], and MSMT17 [44]. Note that we do not use DukeMTMC [51] to avoid ethical issues.

Evaluation Protocols. Following the standard evaluation metrics, we use the cumulative matching characteristics at Rank1 and mean average precision (mAP) to evaluate the performance.

4.2. Improvement on Supervised ReID

In this section, we show the superiority of our UP-ReID by comparing with the model unsupervisedly pre-trained on LUPerson by MoCo v2 [12] and the commonly used supervised pre-trained model on ImageNet in three representative supervised ReID approaches: Batch DropBlock Network (BDB) [9], Strong Baseline (BOT) [31] and Multiple Granularity Network (MGN) [42]. The BDB is re-implemented based on the open source code. As for BOT and MGN, we implement them in fast-reid [20].

Table 1 shows the improvements in the three selected supervised ReID methods on four popular person ReID datasets. It can be seen that, compared to initializing with MoCo v2, the MGN with UP-ReID has achieved **12.2%**, **0.7%**, **1.9%**, **0.4%** improvements in terms of Rank1 on CUHK03, Market1501, PersonX, MSMT17, respectively; BOT also has achieved **2.8%**, **0.2%**, **0.7%**, **2.7%** improvements in terms of Rank1 on these four datasets.

Figure 4 shows the comparison of the convergence speed of applying different pre-trained models in method BDB at the early stage of fine-tuning. UP-ReID outperforms both MoCo v2 and INSUP with faster convergence on all three datasets. The performance enhancement is more noticeable on PersonX (see Figure 4c). On the Market1501 where the advantage is not obvious, UP-ReID still holds the lead of MoCo v2 by 1.7% mAP improvement (see Figure 4b).

4.3. Improvement on Unsupervised ReID

Our pre-trained model can also benefit unsupervised ReID methods. To demonstrate this, we test our pre-trained model on SpCL [14]. We evaluate the performance on Market1501 and PersonX.

In Table 2, M means purely unsupervised training on Market1501, and P → M means unsupervised domain adaptation whose source dataset is PersonX and target dataset is Market1501. As we can see, UP-ReID outperforms MoCo v2 by **2.9%**, **6.3%** in terms of mAP and **2.2%**, **2.5%** in terms of Rank1 on M and P → M, respectively. It fur-

Table 1. Comparison of three representative supervised ReID methods using different pre-trained models in terms of mAP/Rank1 (%). "INSUP" refers to the supervised pre-trained model on ImageNet, "MoCo v2" and "UP-ReID" refer to the MoCo v2 and our UP-ReID pre-trained models on LUPerson, respectively. More comparison results can be found in **Appendix**.

Model	BDB [9]	BOT [31]	MGN [42]
INSUP	76.7/79.4	62.0/63.9	70.5/71.2
MoCo v2	78.9/81.5	66.7/66.3	74.7/75.4
UP-ReID	**79.6/82.6**	**68.7/69.1**	**85.3/87.6**

(a) CUHK03

Model	BDB [9]	BOT [31]	MGN [42]
INSUP	86.7/95.3	85.7/94.3	87.5/95.1
MoCo v2	88.1/95.3	87.6/94.9	91.0/96.4
UP-ReID	**88.5/95.3**	**88.1/95.1**	**91.1/97.1**

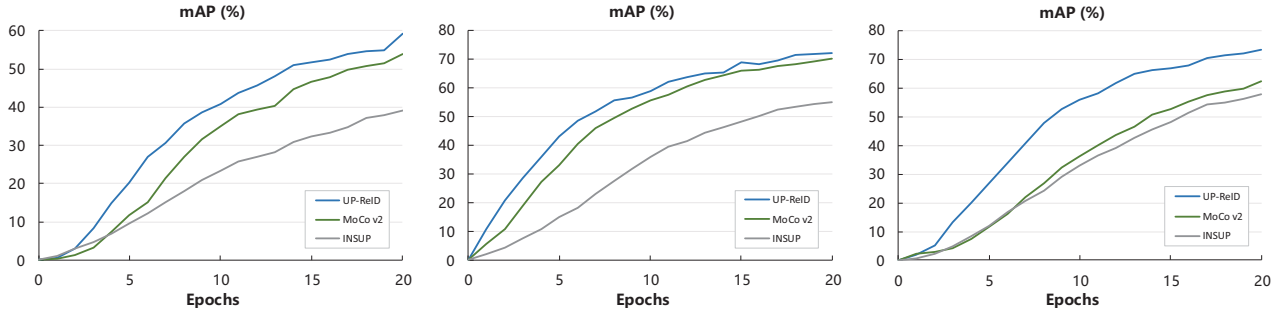
(b) Market1501

Model	BDB [9]	BOT [31]	MGN [42]
INSUP	84.4/95.1	86.7/94.8	85.3/94.3
MoCo v2	84.8/95.2	86.5/94.6	85.8/94.2
UP-ReID	**86.1/95.3**	**88.0/95.3**	**89.7/96.1**

(c) PersonX

Model	BDB [9]	BOT [31]	MGN [42]
INSUP	49.2/77.4	53.4/76.8	61.5/84.0
MoCo v2	51.2/78.1	53.2/75.4	62.9/83.9
UP-ReID	**52.4/78.7**	**56.2/78.1**	**63.3/84.3**

(d) MSMT17



(a) mAP learning curve on CUHK03 (b) mAP learning curve on Market1501 (c) mAP learning curve on PersonX

Figure 4. mAP learning curves with different pre-trained models in BDB on three datasets (CUHK03, Market1501, and PersonX) with the same training schedule. More comparison results can be found in **Appendix**.

ther verifies that UP-ReID can achieve better superiority and generalization capability for person ReID. Note that we implement SpCL by official OpenUnReid [14].

Table 2. Performance (%) comparisons of using different pre-trained models on unsupervised ReID method SpCL.

Model	M		P → M	
	mAP	Rank1	mAP	Rank1
INSUP	73.1	88.1	73.8	88.0
MoCo v2	72.2	87.8	72.4	88.4
UP-ReID	**75.1**	**90.0**	**78.7**	**90.9**

4.4. Comparison with State-of-the-Art Methods

In this section, we compare our results with state-of-the-art methods on CUHK03 and Market1501 datasets. Notice that we do not use any additional modules like IBN-Net or post-processing methods like Re-Rank [52]. We just simply apply UP-ReID pre-trained vanilla ResNet50 on MGN. As shown in Table 3, MGN equipped with UP-ReID ResNet50 outperforms all compared methods on both datasets.

4.5. Ablation Study

In this section, we perform comprehensive ablation studies to demonstrate the effectiveness of our designs in the proposed UP-ReID. Here we fine-tune different pre-trained models with supervised ReID method MGN [42] on CUHK03 to validate the effectiveness of each component.

Table 3. Performance (%) comparisons with state-of-the-art approaches on CUHK03 and Market1501. The best results are marked as bold and the second ones are masked by underline. We show more comparison results in **Appendix**.

Methods	CUHK03		Market1501	
	mAP	Rank1	mAP	Rank1
PCB [40] (ECCV'18)	57.5	63.7	81.6	93.8
OSNet [53] (ICCV'19)	67.8	72.3	84.9	94.8
P2Net [17] (ICCV'19)	73.6	78.3	85.6	95.2
SCAL [2] (ICCV'19)	72.3	74.8	89.3	95.8
DSA [46] (CVPR'19)	75.2	78.9	87.6	95.7
GCP [33] (AAAI'20)	75.6	77.9	88.9	95.2
SAN [27] (AAAI'20)	76.4	80.1	88.0	<u>96.1</u>
ISP [54] (ECCV'20)	74.1	76.5	88.6	95.3
GASM [21] (ECCV'20)	-	-	84.7	95.3
RGA-SC [47] (CVPR'20)	<u>77.4</u>	<u>81.1</u>	88.4	<u>96.1</u>
HOReID [41] (CVPR'20)	-	-	84.9	94.2
AMD [8] (ICCV'21)	-	-	87.1	94.8
TransReID [22] (ICCV'21)	-	-	<u>89.5</u>	95.2
PAT [29] (CVPR'21)	-	-	88.0	95.4
MGN+UP-ReID (Ours)	**85.3**	**87.6**	**91.1**	**97.1**

Effectiveness of the Consistency Constraint and the Intrinsic Contrastive Constraint. Our UP-ReID consists of two key constraints: the consistency constraint (CC) and the intrinsic contrastive constraint (ICC). We evaluate the benefits of them in Table 4, where "Baseline" stands for "MoCo v2". Specifically, (b) Baseline with CC and (c) Baseline with ICC outperform the (a) Baseline by **4.4%/4.8%** and **6.7%/8.2%** in terms of mAP/Rank1 on CUHK03, respectively. With both two constraints, (d) UP-ReID achieves **85.3%(+10.6%)** mAP and **87.6%(+12.2%)** Rank1 on CUHK03, which demonstrates that CC and ICC are complementary and both vital to UP-ReID, jointly resulting in a superior performance.

We also evaluate the effectiveness of each components of our UP-ReID in terms of the convergence speed on CUHK03. Figure 5 plots the mAP learning curves of four different pre-trained models with MGN. As we can see, the (b) Baseline with CC and (c) Baseline with ICC achieve faster convergence than (a) Baseline. More importantly, (d) the UP-ReID with both constraints (*i.e.*, ICC and CC) achieves faster convergence than both (b) and (c) which only have one constraint.

The experimental results demonstrate that both the consistency constraint and the intrinsic contrastive constraint contribute to a better visual representation. The former is designed to counter the augmentation perturbations, and the latter is designed for detailed information exploration.

Table 4. The ablation results of several variants of UP-ReID pre-trained models that are fine-tuned on CUHK03. The values in the brackets are the improvement compared to the Baseline.

Model	CC	ICC	mAP	Rank1
(a) Baseline	×	×	74.7	75.4
(b) Baseline w CC	√	×	79.1(+4.4)	80.2(+4.8)
(c) Baseline w ICC	×	√	81.4(+6.7)	83.6(+8.2)
(d) UP-ReID	√	√	**85.3**(+10.6)	**87.6**(+12.2)

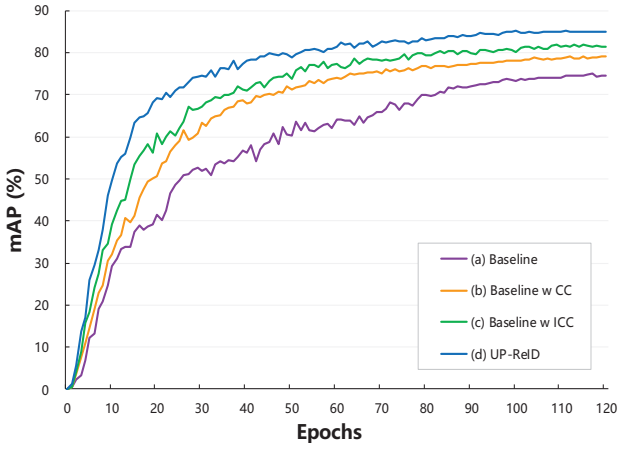


Figure 5. mAP learning curves of CUHK03 in MGN with four different pre-trained UP-ReID models.

Effectiveness of the Hard Mining Strategy. For better representation learning, we introduce a hard mining (HM) strategy to the intrinsic contrastive constraint. As shown in Table 5, the UP-ReID without hard mining strategy (*i.e.*, replace Eq. 10 with Eq. 8) has a 4.6%/4.5% drop in mAP/Rank1. Obviously, our hard mining strategy improves the discrimination capability of the pre-trained model.

Different from the previous works [23], we select positive and negative pairs based on the prior knowledge that persons are horizontally symmetric instead of an online way. We further investigate the influence of different hard mining strategies and show more results in **Appendix**.

Table 5. Effectiveness of the hard mining strategy for ICC in our UP-ReID on CUHK03.

Model	mAP	Rank1	Rank5
UP-ReID w/o HM	80.7	83.1	93.1
UP-ReID w HM	**85.3**	**87.6**	**95.4**

Influence of the Number of the Patch-Level Instances. Note that each patch-level instance is partitioned from the corresponding image-level instance. Different patch number (M) means different patch size. We investigate the influence of the patch-level instance number in the intrinsic contrastive constraint. As described in Table 6, M=8 outperforms M=4 by **4.0%/4.5%** in mAP/Rank1 on CUHK03, which also surpasses M=12 by **4.6%/5.4%** in mAP/Rank1. When M=8, each patch-level instance has a proper size, which is neither too large to ignore discriminative attributes, nor too small to introduce unnecessary noise.

Table 6. Results of different number of patches in ICC.

Model	mAP	Rank1	Rank5
UP-ReID w $M = 4$	81.3	83.1	92.6
UP-ReID w $M = 12$	80.7	82.2	92.4
UP-ReID w $M = 8$	**85.3**	**87.6**	**95.4**

5. Conclusion

In this paper, to address the two critical issues in applying contrastive learning to ReID pre-training tasks, we propose a ReID-specific pre-training framework UP-ReID with an intra-identity regularization, which consists of a global consistency constraint and an intrinsic contrastive constraint. Moreover, we introduce a hard mining strategy to explore the local information for a better representation learning. Extensive experiments demonstrate that UP-ReID can improve the downstream works performance with higher precision and much faster convergence.

Acknowledgments. This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12, and the JKW Research Funds under Grant 20-163-14-LZ-001-004-01. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 3

[2] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, pages 9637–9646, 2019. 7

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. *arXiv preprint arXiv:2103.16364*, 2021. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 3

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 1, 3

[8] Xiaodong Chen, Xinchen Liu, Wu Liu, Xiao-Ping Zhang, Yongdong Zhang, and Tao Mei. Explainable person re-identification with attribute-guided metric distillation. In *ICCV*, pages 11813–11822, 2021. 7

[9] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *ICCV*, pages 3691–3701, 2019. 6, 7

[10] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. *arXiv preprint arXiv:2103.11568*, 2021. 1

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018. 3

[12] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, pages 14750–14759, 2021. 2, 3, 4, 6

[13] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 1, 3

[14] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *arXiv preprint arXiv:2006.02713*, 2020. 1, 3, 6, 7

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, 2012. 4

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[17] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3642–3651, 2019. 7

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[20] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. FastReID: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 6

[21] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *ECCV*, pages 357–373. Springer, 2020. 7

[22] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 7

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 8

[24] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *ICCV*, pages 8526–8536, 2021. 3

[25] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *AAAI*, volume 34, pages 11165–11172, 2020. 1

[26] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 1, 3

[27] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, volume 34, pages 11173–11180, 2020. 1, 2, 7

[28] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2, 6

[29] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 7

[30] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, volume 33, pages 8738–8745, 2019. 1

[31] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 0–0, 2019. 6, 7

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[33] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *AAAI*, volume 34, pages 11839–11847, 2020. 7

[34] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *ICCV*, pages 8080–8089, 2019. 3

[35] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, pages 486–504, 2018. 1, 2

[36] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *PR*, 102:107173, 2020. 3

[37] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 402–419, 2018. 1

[38] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, pages 608–617, 2019. 6

[39] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 2

[40] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 2, 7

[41] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020. 7

[42] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 1, 2, 6, 7

[43] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, pages 2275–2284, 2018. 3

[44] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 3, 6

[45] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, pages 3633–3642, 2019. 1, 2

[46] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019. 1, 2, 7

[47] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 7

[48] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, pages 6277–6286, 2021. 3

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 4, 6

[50] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 3

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 6

[52] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 7

[53] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 7

[54] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, pages 346–363. Springer, 2020. 7