

Unsupervised Pre-training for Person Re-identification

Dengpan Fu¹ Dongdong Chen² Jianmin Bao^{2*} Hao Yang²
 Lu Yuan² Lei Zhang² Houqiang Li¹ Dong Chen²

¹University of Science and Technology of China ²Microsoft Research

fdpan@mail.ustc.edu.cn cddlyf@gmail.com lihq@ustc.edu.cn

{jianbao, haya, luyuan, leizhang, doch}@microsoft.com

Abstract

In this paper, we present a large scale unlabeled person re-identification (Re-ID) dataset “LUPerson” and make the first attempt of performing unsupervised pre-training for improving the generalization ability of the learned person Re-ID feature representation. This is to address the problem that all existing person Re-ID datasets are all of limited scale due to the costly effort required for data annotation. Previous research tries to leverage models pre-trained on ImageNet to mitigate the shortage of person Re-ID data but suffers from the large domain gap between ImageNet and person Re-ID data. LUPerson is an unlabeled dataset of 4M images of over 200K identities, which is 30× larger than the largest existing Re-ID dataset. It also covers a much diverse range of capturing environments (e.g., camera settings, scenes, etc.). Based on this dataset, we systematically study the key factors for learning Re-ID features from two perspectives: data augmentation and contrastive loss. Unsupervised pre-training performed on this large-scale dataset effectively leads to a generic Re-ID feature that can benefit all existing person Re-ID methods. Using our pre-trained model in some basic frameworks, our methods achieve state-of-the-art results without bells and whistles on four widely used Re-ID datasets: CUHK03, Market1501, DukeMTMC, and MSMT17. Our results also show that the performance improvement is more significant on small-scale target datasets or under few-shot setting.

1. Introduction

Model pre-training plays an indispensable role in person Re-identification (Re-ID). Compared to other vision tasks, as data collection and annotation for Re-ID is extremely difficult and expensive, existing public datasets all have a limited scale in terms of image number (largest MSMT17 [39], 126K images), person identities (largest Airport [25], 9,651

(a) Market1501

(b) DukeMTMC

Figure 1: Person Re-ID performance comparison of applying different pre-trained models on two methods: IDE [45] and MGN [38]. We report the results on different dataset scales for Market1501 and DukeMTC with small-scale setting. IN sup. and LUP unsup. are the supervised model trained on ImageNet and the unsupervised model trained on LUPerson, respectively.

identities) and captured environments (< 20 scenes, fixed cameras and resolution). To mitigate the shortage of person Re-ID data, previous research has tried to leverage models pre-trained on ImageNet and transfer the pre-trained feature to Re-ID tasks [22, 3, 34, 38, 13]. However, it is arguable if using ImageNet for pre-training is optimal, due to the large domain gap between ImageNet and person Re-ID data.

Inspired by the recent success of self-supervised learning [6, 8, 19], we make the first attempt towards large scale unsupervised pre-training for person Re-ID feature representation learning in this paper. Considering the limited scale of existing Re-ID datasets, we build a new Large-scale Unlabeled Person Re-ID dataset “LUPerson”. It consists of 4M person images of over 200K identities extracted from 46K YouTube videos, which is 30× larger than the largest existing Re-ID dataset MSMT [39]. Moreover, the collected videos cover a wide range of capturing environments (e.g., using fixed or moving cameras, under dynamic scenes, or having different resolutions), yielding a great data diversity which is essential for learning generic representation. We

*Corresponding author.

hope this study and the developed LUPerson dataset will serve as a solid baseline and motivate more feature representation learning researches for person re-identification.

Based on the LUPerson dataset, we systematically study the problem of unsupervised Re-ID feature learning. We find that directly applying the commonly used contrastive learning method such as MoCo v2 [9] does not work well for the Re-ID task. After a careful investigation, we discover some unique factors when applying unsupervised pre-training to the Re-ID task: 1) As a common data augmentation [19, 7, 1], the color distortion (*e.g.*, color jitter) is harmful to Re-ID feature learning. This is because the color information is a crucial clue for Re-ID. 2) To prevent contrastive learning from degrading to a trivial solution, a strong task-specific augmentation operation RandomErasing [49] is proved as beneficial as in supervised Re-ID training. 3) How to use a proper temperature parameter in the contrastive loss plays an important role in finding a balance between maintaining the discriminativity and mining the hard negatives.

We demonstrate the effectiveness of our pre-trained model on various person Re-ID datasets. Upon the strong MGN [38] baseline, our pre-trained model can improve the *mAP* by 3.5% on Market1501 [44], 2.7% on DukeMTMC [47], and 2.0% on MSMT17[39]; while achieving 2.9% *mAP* gain on CUHK03 [44] based on another strong BDB [10] baseline. These results are superior to all the state-of-the-art methods. Our results show that the performance improvement is even more significant with small-scale training samples for different datasets and baselines, as shown in Fig.1. Besides, our pre-trained model is also general to unsupervised Re-ID methods. Based on the strongest baseline SpCL [15], our pre-trained model consistently achieves remarkable improvements on various datasets. To the best of our knowledge, this is the first showing that large scale unsupervised pre-training can significantly benefit the person Re-ID task.

Our key contributions can be summarized as follows:

- We build a large-scale unlabeled dataset LUPerson, which consists of 4M images for over 200K identities, for unsupervised person Re-ID feature learning. This dataset is much larger than any existing public datasets and enables the first unsupervised pre-training for person Re-ID tasks.
- We make generic unsupervised pre-training possible for Re-ID tasks, by carefully investigating the crucial factors, such as data augmentation strategies and the temperature usage in the contrastive learning framework.
- The unsupervised representation learning is general to not only supervised Re-ID methods, but also un-

supervised Re-ID methods, and helps significantly improve their performance on different datasets.

2. Related Work

Supervised Person Re-ID. Most existing Re-ID approaches are based on supervised learning on labeled datasets. There are three typical categories of approaches: 1) learning a global feature from the whole image, with supervision imposed through a classification loss, *e.g.*, IDE model [45] and [33]; 2) using a hard triplet loss on the global feature to ensure a smaller distance for features of the same person, such as [22]; 3) learning a part-based feature instead. For example, Sun *et al.* [36] proposed to partition an image feature into multiple horizontal strips each learned with a separate classification loss, and Suh *et al.* [34] presented a part-aligned bi-linear representations. The MGN [38], known as one of the state-of-the-art (SOTA) methods, combined both a classification loss on the global feature and a triplet loss on the local features. Our pre-trained model can be used in the above three representative methods, and show better performance and generalization ability.

Unsupervised Person Re-ID. Recently, some works attempted to directly train an unsupervised Person Re-ID model without utilizing any labels on existing Re-ID datasets. BUC [27] proposes a bottom-up hierarchical clustering method to jointly optimize a network and the relationship among samples. Mutual Mean-Teaching (MMT) [14] adopts two networks and learns mutually to refine the hard and soft pseudo labels in the target domain to mitigate the effects of noisy pseudo labels. MMCL [37] formulates unsupervised person Re-ID as a multi-label classification task to progressively seek true labels. SpCL [15] is known as the SOTA approach in this literature. Its key idea is to fine-tune the network using pseudo labels generated from reliable clustering results trained with contrastive losses. In contrast to these methods, our work focuses on the unsupervised pre-training phase to learn a feature representation that can be generalized to either supervised or unsupervised Re-ID approaches.

Unsupervised Representation Learning. Benefit from contrastive learning [40, 6, 19], unsupervised pre-training can learn feature representations with comparable quality to that learned from supervised approaches. Specifically, Wu *et al.* [40] proposed to store representations within a memory bank. MoCo and MoCo v2 [19, 9] introduced a dynamic queue to maintain slowly updated representations to generate negative samples. On the contrary, SimCLR and SimCLR v2 [6, 7] proved that a projection head and rich data augmentations can also lead to advantageous visual representations even without these memory structures. Recently, BYOL [17] shows a good performance even without using any negative pairs. In this paper, we choose the MoCo v2 [8] framework as our unsupervised pre-training

method. Unfortunately, directly applying this framework on person Re-ID tasks does not work well, which requires an in-depth study to identify the key factors which make Re-ID tasks different from generic visual representation learning.

3. LUPerson: Large-scale Re-ID Dataset

Data is the life-blood of training deep neural network models and ensuring their success. For the person Re-ID task, sufficient and high-quality data are also indispensable for increasing the model’s generalization capability. A good Re-ID dataset requires not only a large amount of identities, each appearing in multiple cameras, but also a great diversity in terms of pose variation and capturing environments (*e.g.*, camera angles, resolutions, scenes, etc.). Unfortunately, developing and annotating such a large-scale Re-ID dataset is extremely difficult and expensive.

All existing Re-ID datasets are of limited scale and diversity. Table 1 lists the statistics of existing popular Re-ID datasets, including VIPeR [16], GRID [28], CUHK03 [26], Market-1501 [44], Airport [25], DukeMTMC [47], and MSMT17 [39]. As we can see, the largest dataset only consists of less than 0.2M images and the largest number of identities is less than 10K. Moreover, these datasets cover very limited scenes (< 20) and camera settings. As a result, it is hard to use such datasets to learn a high-quality feature representation which is as good as that learned in the generic image classification task which can utilize 1.2M images (*i.e.*, ImageNet-1k) or even 14M images (*i.e.*, ImageNet-22k).

In this work, we ask the question: *can we develop a person Re-ID dataset which is as large as ImageNet?* To address this question, we build **LUPerson**: a Large-scale Unlabeled Person Re-ID dataset, consisting of 4M images for more than 200K identities collected from 46K scenes. To our best knowledge, this is the largest scale person Re-ID dataset.

The development of LUPerson is inspired by the recent success of unsupervised pre-training on generic image classification tasks. Although LUPerson is an unlabeled dataset, the large number of identities and diverse capturing environments included in the dataset offer a great potential for learning a high-quality person Re-ID feature representation that can be utilized to boost the performance of all kinds of Re-ID CNN models. We will make the dataset publicly available to motivate more research works to advance the state of the art of person re-identification.

3.1. Data Collection and Processing

To build the dataset, we crawled over 70K streetview videos from YouTube by using queries like “cityname + streetview (or scene)”. To cover a large diversity, we chose

Figure 2: Some example images from our LUPerson dataset, which shows a strong diversity in terms of environment, scene, camera view, lighting, human pose, race, and age.

the names of top 100 big cities in the world¹ and collected about 730 (680 – 760) raw videos on average for each city. To ensure a high quality, we further filtered out some invalid videos by checking the following cases: 1) duplicated videos with the same name (YouTube key); 2) videos containing less than 100 frames; 3) static videos (we evenly sample 5 frames and treat this video as a static video if these frames are the same); 4) virtual reality (VR) videos. In total, 50,534 videos are remained after filtering.

We follow the process of building existing person Re-ID datasets, which extract each person instance detected in every image. We use YOLO-v5² trained on MS-COCO to detect persons in every sampled frame. Considering that some instances only show partially visible body, we apply HRNet [35] to detect the body key-points. Such key-points can be categorized into three types: head, upper body, lower body. In our specification, one person image is regarded as valid if it satisfies the following requirements: 1) head and upper body are visible; 2) lower body is partially visible if either hip or knee exists; 3) The height/width ratio should

¹Innovation City Index 2019: Top 100 Cities <https://www.innovation-cities.com/worlds-top-100-cities-for-innovation-2019/18841/>

²YOLO-v5 <https://github.com/ultralytics/yolov5>

Datasets	#images	#scene	#persons	environment	camera view	resolution	detector	crop size
VIPeR[16]	1,264	2	632	-	fixed	fixed	hand	128 × 48
GRID[28]	1,275	8	1,025	subway	fixed	fixed	hand	vary
CUHK03[26]	14,096	2	1,467	campus	fixed	fixed	DPM[12]+hand	vary
Market[44]	32,668	6	1,501	campus	fixed	fixed	DPM[12]+hand	128 × 64
Airport[25]	39,902	6	9,651	airport	fixed	fixed	ACF[11]	128 × 64
DukeMTMC[47]	36,411	8	1,852	campus	fixed	fixed	Hand	vary
MSMT17[39]	126,441	15	4,101	campus	fixed	fixed	FasterRCNN[32]	vary
LUPerson	4,180,243	46,260	> 200k	vary	dynamic	dynamic	YOLOv5	vary

Table 1: The statistics comparison between existing popular Re-ID datasets and our large scale LUPerson dataset. It shows that LUPerson is the current largest Re-ID dataset and has much better diversity.

be larger than 1.5 and less than 5; 4) The detection confidence must be larger than 0.72; 5) The bounding box width must be larger than 48 pixels. Besides, we extract person every 100 frames. After applying these rules, we finally get 4,180,243 person images. To make an estimation of the person number, we adopt the following strategy: we search all the frames in a video to find the frame with the max number of persons and take the max number as the number of person of this video, then we sum up the persons of all the videos and get the number 219,848. Some example images are given in Figure 2, it shows that our LUPerson has very diverse backgrounds and pose variations.

3.2. Comparison with Existing Datasets

Compared with existing datasets [16, 28, 26, 44, 25, 47, 39], LUPerson is superior in the following aspects.

1) *The numbers of images and identities.* To the best of our knowledge, LUPerson is the largest person Re-ID dataset and contains over 4M images of 200K identities, which is $30\times$ larger than MSMT17 [39].

2) *Diverse environment.* LUPerson is collected from much diverse environments (containing 46,260 scenes) with both static and dynamic camera views of street, campus, supermarket, sport fields, etc., while existing datasets were normally collected from a few fixed environments (*e.g.*, campus, or no more than 15 scenes [39]) and limited camera views. Diverse environments covered by LUPerson are essential for learning a generic Re-ID feature that can be generalized to real applications.

3) *Complex devices.* The raw videos of LUPerson are captured by diverse devices, such as variant types of video recorders, vlogs, smart phones, and surveillance cameras, in contrast to existing datasets which often share the same hardware property.

4) *Lighting variance.* The crawled videos span over a large range of time in a day, including morning, noon, and night, thus causing different lighting changes.

5) *Ethnic and pose diversity.* Since the raw videos are collected from 100 cities across the world, LUPerson has ethnic diversity. Due to the variance of camera views, it

also has large pose variations.

4. Unsupervised Pre-training for Person Re-ID

Based on the LUPerson dataset, we attempt to pre-train an unsupervised model for improving the generalization ability of the learned person Re-ID feature representation. Without loss of generality, we leverage the widely used contrastive learning method MoCoV2 [9] as a simple baseline. However, we find such a general setting does not work well on the person Re-ID task. It requires an in-depth study and re-design.

In this section, we will firstly introduce the contrastive learning method in section 4.1, and then present the key changes specifically for the person Re-ID task in section 4.2. Finally, we will show how the learned representation benefits existing person Re-ID methods in section 4.3.

4.1. Contrastive Learning

Contrastive learning can be regarded as a dictionary look-up problem. Given an encoded query q and a set of encoded key samples $\{k_1, \dots, k_K\}$ in the dictionary, the contrastive learning is essentially to encourage q to be similar to the positive samples $\{k_i^+\}$ and dissimilar to the negative samples $\{k_i^-\}$. Since no label information is available in unsupervised learning, we do not know which are positive or negative. Hence, existing methods (*e.g.*, MoCo [19])

Figure 3: Illustration of the Momentum Contrast mechanism for contrastive learning in MoCo [19].

adopt a self-learning strategy. Specifically, as shown in Figure 3, given an image sample x , two different augmentations $T_1(x), T_2(x)$ are used to generate two different samples. Then one sample is regarded as query q and the other is regarded as the positive sample k^+ . All the samples not augmented from the same image of q are then regarded as the negative samples k^- . Formally, the contrastive loss is defined as:

$$L_c = -\log \frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_{i=0}^{K-1} \exp(q \cdot k_i^- / \tau)}, \quad (1)$$

where τ is a temperature hyper-parameter, K is the number of negative samples.

To obtain (q, k) , separated encoders f_q and f_k are used respectively, and both of them consist of a base feature extractor network (e.g., ResNet50 backbone) and an extra projection head (e.g., two-layer MLP). Intuitively, the above contrastive loss in Equation 1 can be viewed as a $(K + 1)$ -way classification loss. As demonstrated in existing methods [6, 40], a large K is beneficial because a rich set of negative samples are covered. One key idea of MoCo is using a queue to maintain a large dictionary and progressively updating the samples in the dictionary with the latest ones. However, updating the key encoder with a large queue by back-propagation is intractable. To solve this issue, another key idea is using momentum updating scheme for f_k .

$$k := m \cdot k + (1 - m) \cdot q, \quad (2)$$

where k, q are the parameters of f_k, f_q , and m is a momentum coefficient.

4.2. Key Factors to Study

Compared with general image recognition tasks, person Re-ID is a more challenging task since it needs to seek more fine-grained details for robust person association. Empirically, we observe that directly applying MoCoV2 [9] cannot learn a good representation. To improve the performance, we systematically study some key factors: *data augmentation* and *temperature tuning strategy* in the contrastive loss.

Data Augmentation. Data augmentation is crucial to self-supervised contrastive learning. Though the effects of different augmentation operators have been extensively studied for general representation learning [42, 6, 19], we find the recommended augmentations are not optimal to the person Re-ID task.

Firstly, we re-investigate the effect of each augmentation operator used in MoCoV2 with respect to the transfer performance on the CUHK03 dataset. Here, we mainly focus on *color distortion augmentations* (including random grayscale, Gaussian blurring and color jitter), since common augmentations like cropping, resizing and flipping are still indispensable to the Re-ID task. The results are shown

in Table 2. Compared with the “default” strategy in MoCoV2, we exclude one augmentation at each time. For random grayscale and Gaussian blurring, removing each has limited impact to the performance. Thus, we will not study them in our augmentation. For color jitter, the performance is boosted by 0.6% in terms of *mAP* and 0.6% in terms of *cmc1* after it is disabled. This is because person Re-ID heavily relies on color information (e.g. color of clothes, color of bags) for association, while color jitter will harm its performance.

Then, we revisit the task-specific augmentation *RandomErasing* [49], which randomly selects a rectangle region in an image and erases its pixels with random values during training. It has been widely used in person Re-ID methods. As we can see in Table 2, adding RandomErasing can achieve about 0.8% gain in terms of *mAP*. Besides, relatively large strength of RandomErasing can help achieve a better result, as shown in Table 3. Note that the optimal RandomErasing strength on existing person Re-ID method is 0.4, which is smaller than the optimal value (0.6) in our unsupervised pre-training. This indicates that unsupervised pre-training benefits more from stronger data augmentation than vanilla person Re-ID training.

As a summary, we make two important changes in data augmentation for Re-ID contrastive learning: adding RandomErasing with high strength and removing color jitter.

Temperature Tuning Strategy. In the contrastive loss, the temperature hyper-parameter τ effectively weights different examples as shown in Equation 1. A too large value of τ will reduce the discriminativity between positive samples k^+ and negative samples k^- since it makes $\exp(q \cdot k^+ / \tau)$ close to $\exp(q \cdot k^- / \tau)$. By contrast, a too small value of τ would be harmful for the model to learn from hard negative samples, since it makes softmax output more spike towards the positive pair and cannot help the model learn from hard negatives. Hence, an appropriate temperature is critical to help the model learn a discriminative representation.

For this purpose, we search for an optimal τ with respect to the transfer performance on existing person Re-ID datasets (e.g., CUHK03 [26]). The result is shown in Table 4. Using the default $\tau = 0.2$ leads to *mAP* = 71.1; switching to a smaller τ , the accuracy increases to *mAP* = 74.7 ($\tau = 0.07$). The phenomena consistently appears in other target Re-ID datasets. More interestingly, in general image recognition [9], the accuracy is reduced from 66.2 to 62.9 when τ varies from 0.2 to 0.07. One possible reason is that person Re-ID data is inherently more fine-grained compared with general image recognition data, like ImageNet. In other words, Re-ID data has smaller inter-class variations, which make positive samples close to negative samples. To avoid reducing the discriminativity, a small temperature is better. It may suggest that a smaller temperature value is more appropriate for contrast learning on

Setting	Default	+RE	-GS	-GB	-CJ	-CJ+RE
mAP	73.4	74.2	73.2	73.3	74.0	74.7
cmc1	74.0	74.8	73.9	74.1	74.6	75.4

Table 2: Transfer performance on the CUHK03 dataset with different data augmentations. “+,” “-” mean with and without, and “RE, GS, GB, CJ” mean RandomErasing, GrayScale, GaussianBlurring, and ColorJitter respectively.

Max area	0.0	0.2	0.4	0.6	0.8
mAP	73.2	74.1	74.4	74.7	73.3
cmc1	73.8	74.3	75.3	75.4	73.7

Table 3: Transfer performance on the CUHK03 dataset with different RandomErasing strength, i.e., maximum erasing area.

	0.03	0.05	0.07	0.1	0.2	0.3
mAP	72.5	73.5	74.7	74.1	71.1	67.3
cmc1	73.4	74.0	75.4	73.9	71.5	67.4

Table 4: Transfer performance on the CUHK03 dataset with different temperature in the contrastive loss.

fine-grained recognition problems.

4.3. Transferring Features

The unsupervised pre-training performed on this large-scale dataset effectively learns a generic person Re-ID feature representation that can benefit existing supervised and unsupervised person Re-ID methods. As mentioned in [19], features produced by unsupervised pre-training can have different distributions compared with ImageNet supervised pre-training. Besides, the training hyper-parameters (*e.g.*, learning rate) of existing Re-ID methods are also tuned for supervised pre-training models, and thus may not be optimal for unsupervised pre-training models.

To address this issue, we add an extra batch normalization (BN) layer for re-calibration after anywhere the pre-trained feature is used. For example, in MGN [38], as three heads are appended after the backbone features, we add three extra BN layers before each head. With this strategy, we can use the same hyper-parameters as the ImageNet supervised counterpart.

5. Experiments

5.1. Implementation

Training details. We train MoCoV2 with Pytorch, and the images are resized to 256×128 . For image normalization, we use [0.3525, 0.3106, 0.3140], [0.2660, 0.2522, 0.2505] as mean and std, which are calculated from our LUPerson dataset. The pre-training models are trained with $8 \times V$ 100

GPUs for 200 epochs. The initial learning rate is 0.3 with batch size 2560. If not specified, the backbone network we used is ResNet50.

Dataset. To demonstrate the superiority of our pre-training models, we conduct extensive experiments on four target person Re-ID datasets, including CUHK03, Market, DukeMTMC and MSMT17. For CUHK03, following the new protocols proposed in [48], we split this dataset into two parts: 7,365 images with 767 identities for training and 6,732 images with 700 identities for testing, and we only use its labeled sub-set. For the other three datasets, we use their official settings.

Evaluation Protocol. In all the experiments, we follow the standard evaluation metrics: mean Average Precision (mAP) and the Cumulated Matching Characteristics top-1 (cmc1) metric.

5.2. Improving Supervised Re-ID Methods

In this section, we show the performance improvement by replacing the supervised pre-trained model on ImageNet with our unsupervised pre-trained model on LUPerson in three representative supervised Re-ID baselines: Trip [22], IDE [45] and MGN [38]. The Trip and IDE are our re-implementation based on open source and have comparable or better performance compared with the original papers’ claim. For MGN, we use its popular implementation in fast-reid [20]³, which can obtain considerable higher performance than what is reported in MGN [38].

Table 5 shows the detailed improvements over 4 popular person Re-ID datasets. It can be seen that, equipped with our new pre-trained model, all the three methods can have more than 4.2%, 3.5%, 2.7%, 2% improvement in terms of *mAP* on CUHK03, Market1501, DukeMTMC and MSMT17 respectively. As reference, we also compare to the baseline “unsupervised pre-trained model on ImageNet”. In most cases, the unsupervised ImageNet model is slightly better than the supervised one but much less than our model. On the one hand, it demonstrates the potential of unsupervised pre-training. On the other hand, pre-training on ImageNet is not compatible with pre-training on LUPerson, which further emphasizes the importance of building person-related dataset for person Re-ID.

5.3. Comparison on Small-scale and Few-shot

We further study how our pre-trained model benefits the cases where the target dataset has a smaller scale or just a few labels. This is especially important for real applications in which collecting a large labelled Re-ID dataset is so difficult. In this paper, we simulate two typical small dataset settings on the three target datasets: DukeMTMC, Market1501 and MSMT17. CUHK03 is not involved because it is too small.

³fast-reid: <https://github.com/JDAI-CV/fast-reid>

pre-train	Trip [22]	IDE [45]	MGN [38]	pre-train	Trip [22]	IDE [45]	MGN [38]
IN sup.	45.2/63.8	50.6/55.9	70.5/71.2	IN sup.	76.2/89.7	74.1/90.2	87.5/95.1
IN unsup.	55.5/61.2	52.5/57.7	67.1/67.0	IN unsup.	75.1/88.5	74.5/89.3	88.2/95.3
LUP unsup.	62.6/67.6	57.6/62.3	74.7/75.4	LUP unsup.	79.8/71.5	77.9/91.0	91.0/96.4

(a) CUHK03				(b) Market1501			
pre-train	Trip [22]	IDE [45]	MGN [38]	pre-train	Trip [22]	IDE [45]	MGN [38]
IN sup.	65.2/80.7	62.8/80.8	79.4/89.0	IN sup.	34.3/54.8	36.2/66.2	63.7/85.1
IN unsup.	65.4/81.1	63.4/81.6	79.5/89.1	IN unsup.	34.4/55.4	37.6/67.3	62.7/84.3
LUP unsup.	69.8/83.1	65.9/82.2	82.1/91.0	LUP unsup.	36.6/57.1	39.8/68.9	65.7/85.5

(c) DukeMTMC				(d) MSMT17			
--------------	--	--	--	------------	--	--	--

Table 5: Improvement by using different pre-trained models on three representative supervised Re-ID baselines. “IN sup.”, “IN unsup.” refer to supervised and unsupervised pre-trained model on ImageNet, “LUP unsup.” refers to unsupervised pre-trained model on LUPerson. The first number is *mAP* and the second is *cmc1*.

pre-train	small-scale					few-shot				
	10%	30%	50%	70%	90%	10%	30%	50%	70%	90%
IN sup.	53.1/76.9	75.2/90.8	81.5/93.5	84.8/94.5	86.9/95.2	21.1/41.8	68.1/87.6	80.2/92.8	84.2/94.0	86.7/94.6
IN unsup.	58.4/81.7	76.6/91.9	82.0/94.1	85.4/94.5	87.4/95.5	18.6/36.1	69.3/87.8	78.3/90.9	84.4/94.1	87.1/95.2
LUP unsup.	64.6/85.5	81.9/93.7	85.8/94.9	88.8/95.9	90.5/96.4	26.4/47.5	78.3/92.1	84.2/93.9	88.4/95.5	90.4/96.3

(a) Market1501										
pre-train	small-scale					few-shot				
	10%	30%	50%	70%	90%	10%	30%	50%	70%	90%
IN sup.	45.1/65.3	64.7/80.2	71.8/84.6	75.5/86.8	78.0/88.3	31.5/47.1	65.4/79.8	73.9/85.7	77.2/87.8	79.1/88.8
IN unsup.	48.1/66.9	65.8/80.2	72.5/84.4	76.3/86.9	78.5/88.7	32.4/48.0	65.3/80.2	73.7/85.1	77.7/87.8	79.4/89.0
LUP unsup.	53.5/72.0	69.4/81.9	75.6/86.7	78.9/88.2	81.1/90.0	35.8/50.2	72.3/83.8	77.7/87.4	80.8/89.2	82.0/90.6

(b) DukeMTMC										
pre-train	small-scale					few-shot				
	10%	30%	50%	70%	90%	10%	30%	50%	70%	90%
IN sup.	23.2/50.2	41.9/70.8	50.3/76.9	56.9/81.2	61.9/84.2	14.7/34.1	44.5/71.1	56.2/79.5	60.9/82.8	63.4/84.5
IN unsup.	22.6/48.8	40.4/68.7	49.0/75.0	55.7/79.9	60.9/83.0	13.2/29.2	41.4/67.1	53.3/77.6	59.1/81.5	62.4/83.8
LUP unsup.	25.5/51.1	44.6/71.4	53.0/77.7	59.5/81.8	63.7/85.0	17.0/36.0	49.0/73.6	57.4/80.5	62.9/83.5	65.0/85.1

(c) MSMT17										
------------	--	--	--	--	--	--	--	--	--	--

Table 6: Performance for small-scale and few-shot setting with MGN method for Market1501, DukeMTMC and MSMT17 respectively.

- **Small-scale.** We randomly select a certain percentage of IDs and all the images belonging to the sampled IDs would be included.
- **Few-shot.** We keep all the identities and randomly sample a certain percentage of images for each ID. During sampling, we try to ensure that each ID has a similar number of images.

We use MGN as the baseline method and show the results in Table 6 by varying the percentage from 10% 100%. As we can see, our pre-training model significantly boosts the performance of MGN in most cases when the training set is small, no matter the identities is less or each identity has few images.

Specifically, for the “small-scale” setting on Mar-

ket1501, which contains only 1,170 images for 75 persons (percentage= 10%, more details in supplementary materials), MGN with our pre-training model achieves 64.6 *mAP* on the testing set, which is 11.5 *mAP* higher than the ImageNet supervised counterpart.

For the “few-shot” setting, our pre-training model also boosts the performance of MGN with a remarkable margin, but the performance gain becomes a bit saturated with the decrease of image amounts. One possible reason is that quite limited images for each identity may weaken the ability of hard triplet loss.

5.4 Pre-training Data Scale

In generic image recognition tasks, more powerful models often rely on more high-quality data. Here, we study

scale	CUHK03	Market1501	DukeMTMC	MSMT17
12.5%	69.2/69.0	89.5/95.9	80.2/89.1	61.5/82.3
25%	72.1/71.5	90.1/96.2	80.9/90.1	63.1/83.3
50%	74.1/74.5	90.8/96.4	81.6/90.4	65.5/84.7
100%	74.7/75.4	91.0/96.4	82.1/91.0	65.7/85.5

Table 7: Comparison for different pre-training data scale, and the baseline method is MGN.

the impact of pre-training data scale. Specifically, we involve various percentages (12.5%, 25%, 50%, 100%) of LUPerson into unsupervised pre-training and then evaluate the finetuning performance on the target datasets. As shown in Table 7, the learned representation is much stronger with the increase of the pre-training data scale. But the performance tends to saturate, when the data scale is large enough. A larger network capacity would be necessary to leverage more data for further improving the performance.

5.5. Improving Unsupervised Re-ID Methods

pre-train	USL		UDA			
	M	D	D	M	M	D
IN sup.	72.4/87.8	64.9/80.3	76.4/90.1	67.9/82.3		
IN unsup.	72.9/88.6	62.6/78.8	77.1/90.6	66.3/81.6		
LUP unsup.	76.2/90.2	67.1/81.6	79.2/91.7	69.1/83.2		

Table 8: Improvement for unsupervised Re-ID method SpCL. M and D refer to Market1501 and DukeMTMC. Note that, we use the official released code and the performance obtained is slightly lower than the original paper.

Our pre-training model not only benefits the supervised person Re-ID methods, but also is indispensable in existing unsupervised person Re-ID methods. To verify it, we use state-of-the-art unsupervised baseline methods SpCL [15] with its two settings: unsupervised learning (USL) and unsupervised domain adaptation (UDA). We follow the common setting as SpCL [15], MMT [14] and MMCL [37], etc., and evaluate on Market1501 and DukeMTMC. The results are shown in 8. As we can see, our pre-training model can improve 3.8 *mAP* and 2.2 *mAP* on Market1501 and DukeMTMC respectively, which is a new state-of-the-art for USL in person Re-ID. For the UDA setting, our model can also boost M \rightarrow D and D \rightarrow M by 2.8% and 1.2% on *mAP*. It further demonstrates the superiority and generality of our pre-training model.

5.6. Comparison with State-of-the-Arts

We compare our results with existing state-of-the-art Re-ID methods in Table 9. Note that we do not apply any post-processing method like Re-Rank [48] in our approach. As we can see, we achieve state-of-the-art performance on Market1501, DukeMTMC and MSMT17 with considerable advantages, by simply applying our pre-training

Method	CUHK03	Market1501	DukeMTMC	MSMT17
PCB [36] (2018)	57.5/63.7	81.6/93.8	69.2/83.3	-
MGN [38] (2018)	67.4/68.0	86.9/95.7	78.4/88.7	-
MGN*	70.5/71.2	87.5/95.1	79.4/89.0	<u>63.7/85.1</u>
BOT [29] (2019)	-	85.9/94.5	76.4/86.4	-
DGNet [46] (2019)	-	86.0/94.8	74.8/86.6	52.3/77.2
IANet [23] (2019)	-	83.1/94.4	73.4/87.1	46.8/75.5
DSA [43] (2019)	75.2/78.9	87.6/95.7	74.3/86.2	-
Auto [31] (2019)	73.0/77.9	85.1/94.5	-	52.5/78.2
ABDNet [5] (2019)	-	88.3/95.6	78.6/89.0	60.8/82.3
OSNet [50] (2019)	67.8/72.3	84.9/94.8	73.5/88.6	52.9/78.7
SCAL [4] (2019)	72.3/74.8	89.3/95.8	79.6/89.0	-
P2Net [18] (2019)	73.6/78.3	85.6/95.2	73.1/86.5	-
MHN [2] (2019)	72.4/77.2	85.0/95.1	77.2/89.1	-
BDB [10] (2019)	76.7/79.4	86.7/95.3	76.0/89.0	-
SONA [41] (2019)	79.2/81.8	88.8/95.6	78.3/89.4	-
GCP [30] (2020)	75.6/77.9	88.9/95.2	78.6/87.9	-
SAN [24] (2020)	76.4/80.1	88.0/96.1	75.5/87.9	55.7/79.2
ISP [51] (2020)	74.1/76.5	88.6/95.3	80.0/89.6	-
GASM [21] (2020)	-	84.7/95.3	74.4/88.3	52.5/79.5
Ours(R50)+BDB	79.6/81.9	88.1/95.3	77.4/88.7	52.5/79.1
Ours(R50)+MGN	74.7/75.4	91.0/96.4	82.1/91.0	65.7/85.5
MGN(R101)	73.5/74.6	89.0/95.8	80.9/89.8	66.0/85.7
Ours(R101)+MGN	76.9/77.6	92.0/97.0	84.1/91.9	68.8/86.6

Table 9: Comparison with state of the arts. MGN* refers to the re-implementation of MGN in fast-reid. The best is marked as bold and the second is underlined.

ResNet50 model on MGN. On CUHK03, there is a noticeable gap between the baseline MGN and the state-of-the-art SONA [41]. Thus, we also apply our pre-trained model upon BDB [10], which achieves 2.9 *mAP* gains over BDB and helps beat SONA [41].

We also test our method on a large backbone ResNet101. Compared to results of ResNet50, we can see that a stronger backbone can learn better representation features, yielding better performance.

6. Conclusion

This paper releases a new large-scale person Re-ID dataset “LUPerson”, which may address the issue of limited scale and diversities in all existing Re-ID datasets. Based on such an unlabelled dataset, we make the first attempt to utilize unsupervised pre-training to learn a general person Re-ID feature representation. Experiments demonstrated the effectiveness and generalization ability of our pre-training model on both supervised and unsupervised Re-ID approaches. Furthermore, it helps achieve bigger gains in small-scale or few-shot Re-ID datasets. Certainly, there are some interesting topics motivated in this direction, including leveraging the temporal information of videos into the pre-training and developing an end-to-end unsupervised Re-ID model which can beat supervised ones.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (NSFC, 61836011).

References

- [1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 371–381, 2019. 8
- [3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018. 1
- [4] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9637–9646, 2019. 8
- [5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abdnnet: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8351–8361, 2019. 8
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 5
- [7] 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. 2
- [8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2
- [9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 4, 5
- [10] Zuo Zhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3691–3701, 2019. 2, 8
- [11] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1532–1545, 2014. 4
- [12] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009. 4
- [13] Dengpan Fu, Bo Xin, Jingdong Wang, Dongdong Chen, Jianmin Bao, Gang Hua, and Houqiang Li. Improving person re-identification with iterative impression aggregation. *IEEE Transactions on Image Processing*, 29:9559–9571, 2020. 1
- [14] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2019. 2, 8
- [15] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Advances in Neural Information Processing Systems*, 2020. 2, 8
- [16] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. 3, 4
- [17] 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. 2
- [18] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 3642–3651, 2019. 8
- [19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 4, 5, 6
- [20] Lingxiao He, Xingyu Liao, Wu Liu, Xinchun 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 *European Conference on Computer Vision*, pages 357–373. Springer, 2020. 8
- [22] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2, 6, 7
- [23] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019. 8
- [24] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, pages 11173–11180, 2020. 8
- [25] Srikrishna Karanam, Mengran Gou, Ziyang Wu, Angels Rates-Borras, Octavia Camps, and Richard J Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2(3):5, 2016. 1, 3, 4
- [26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 3, 4, 5

- [27] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8738–8745, 2019. 2
- [28] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *2013 IEEE International Conference on Image Processing*, pages 3567–3571. IEEE, 2013. 3, 4
- [29] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 2019. 8
- [30] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11839–11847, 2020. 8
- [31] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3750–3759, 2019. 8
- [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4
- [33] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 486–504, 2018. 2
- [34] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 1, 2
- [35] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3
- [36] 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*, 2018. 2, 8
- [37] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10981–10990, 2020. 2, 8
- [38] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 274–282. ACM, 2018. 1, 2, 6, 7, 8
- [39] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 1, 2, 3, 4
- [40] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2, 5
- [41] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3760–3769, 2019. 8
- [42] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 5
- [43] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019. 8
- [44] Liang Zheng, Liye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 2, 3, 4
- [45] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 1, 2, 6, 7
- [46] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 8
- [47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 3, 4
- [48] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 6, 8
- [49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2, 5
- [50] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019. 8
- [51] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. *ECCV*, 2020. 8