Large-Scale Pre-training for Person Re-identification with Noisy Labels

Dengpan Fu[1] Dongdong Chen[3] Hao Yang[2] Jianmin Bao[2]*
Lu Yuan[3] Lei Zhang[4] Houqiang Li[1] Fang Wen [2] Dong Chen[2]

[1]University of Science and Technology of China [2]Microsoft Research, [3]Microsoft Cloud AI, [4]IDEA

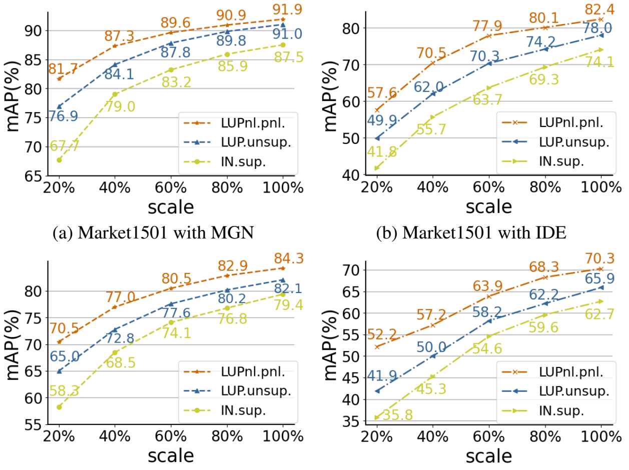
fdpan@mail.ustc.edu.cn cddlyf@gmail.com lihq@ustc.edu.cn
{jianbao,haya,luyuan,fangwen,doch}@microsoft.com, leizhang@idea.edu.cn

Abstract

This paper aims to address the problem of pre-training for person re-identification (Re-ID) with noisy labels. To setup the pre-training task, we apply a simple online multi-object tracking system on raw videos of an existing unlabeled Re-ID dataset "LUPerson" and build the ***N****oisy* ***L****abeled variant called "LUPerson-****NL****". Since theses ID labels automatically derived from tracklets inevitably contain noises, we develop a large-scale* ***P****re-training framework utilizing* ***N****oisy* ***L****abels (****PNL****), which consists of three learning modules: supervised Re-ID learning, prototype-based contrastive learning, and label-guided contrastive learning. In principle, joint learning of these three modules not only clusters similar examples to one prototype, but also rectifies noisy labels based on the prototype assignment. We demonstrate that learning directly from raw videos is a promising alternative for pre-training, which utilizes spatial and temporal correlations as weak supervision. This simple pre-training task provides a scalable way to learn SOTA Re-ID representations from scratch on "LUPerson-NL" without bells and whistles. For example, by applying on the same supervised Re-ID method MGN, our pre-trained model improves the mAP over the unsupervised pre-training counterpart by* ***5.7%, 2.2%, 2.3%*** *on CUHK03, DukeMTMC, and MSMT17 respectively. Under the small-scale or few-shot setting, the performance gain is even more significant, suggesting a better transferability of the learned representation. Code is available at* https://github.com/DengpanFu/LUPerson-NL.

1. Introduction

A large high-quality labeled dataset for person re-identification (Re-ID) is labor intensive and costly to create. Existing fully labeled datasets [25, 52, 58, 61] for person Re-ID are all of limited scale and diversity compared to other vision tasks. Therefore, model pre-training becomes a crucial approach to achieve good Re-ID performance. However, due to the lack of large-scale Re-ID dataset, most previous methods simply use the models pre-trained on the crowd-labeled ImageNet dataset, resulting in a limited improvement because of the big domain gap between generic images in ImageNet and person-focused images desired by the Re-ID task. To mitigate this problem, the recent work [12] has demonstrated that unsupervised pre-training on a web-scale unlabeled Re-ID image dataset "LUPerson" (sub-sampled from massive streeview videos) surpasses that of pre-training on ImageNet.



(a) Market1501 with MGN
(b) Market1501 with IDE
(c) DukeMTMC with MGN
(d) DukeMTMC with IDE

Figure 1. Comparing person Re-ID performances of three pre-trained models on two methods (IDE [59] and MGN [51]). Results are reported on Market1501 and DukeMTC, with different scales under the small-scale setting. *IN.sup.* refers to the model supervised pre-trained on ImageNet, *LUP.unsup.* is the model unsupervised pre-trained on LUPserson, and *LUPnl.pnl.* is the model pre-trained on our LUPerson-NL dataset using our proposed PNL.

In this paper, our hypothesis is that *scalable ReID pre-training methods that learn directly from raw videos can generate better representations*. To verify it, we propose the *noisy labels guided person Re-ID pre-training*, which leverages the spatial and temporal correlations in videos as weak

[*] Corresponding author.

supervision. This supervision is nearly cost-free, and can be achieved by the tracklets of a person over time derived from any multi-object tracking algorithm, such as [56]. In particular, we track each person in consecutive video frames, and automatically assign the tracked persons in the same tracklet to the same Re-ID label and vice versa. Enabled by the large amounts of raw videos in LUPerson [12], publicly available data of this form on the internet, we create a new variant named "LUPerson-**NL**" with derived pseudo Re-ID labels from tracklets for pre-training with noisy labels. This variant totally consists of $10M$ person images from $21K$ scenes with noisy labels of about $430K$ identities.

We demonstrate that contrastive pre-training of Re-ID is an effective method of learning from this weak supervision at large scale. This new **P**re-training framework utilizing **N**oisy **L**abels (**PNL**) composes three learning modules: (1) a simple *supervised learning module* directly learns from Re-ID labels through classification; (2) a *prototype-based contrastive learning module* helps cluster instances to the prototype which is dynamically updated by moving averaging the centroids of instance features, and progressively rectify the noisy labels based on the prototype assignment. and (3) a *label-guided contrastive learning module* utilizes the rectified labels subsequently as the guidance. In contrast to the vanilla momentum contrastive learning [7,12,19] that treats only features from the same instance as positive samples, our label-guided contrastive learning uses the rectified labels to distinguish positive and negative samples accordingly, leading to a better performance. In principle, joint learning of these three modules make the consistency between the prototype assignment from instances and the high confident (rectified) labels, as possible as it can.

The experiments show that our PNL model achieves remarkable improvements on various person Re-ID benchmarks. Figure 1 indicates that the performance gain from our pre-trained models is consistent on different scales of training data. For example, upon the strong MGN [51] baseline, our pre-trained model improves the mAP by $4.4\%, 4.9\%$ on Market1501 and DukeMTMC over the ImageNet supervised one, and $0.9\%, 2.2\%$ over the unsupervised pre-training baseline [12]. Moreover, the gains are even larger under the small-scale and few-shot settings, where the labeled Re-ID data are extremely limited. To the best of our knowledge, we are the first to show that large-scale noisy label guided pre-training can significantly benefit person Re-ID task.

Our key contributions can be summarized as follows:

- We propose noisy label guided pre-training for person Re-ID, which incorporates supervised learning, prototype-based contrastive learning, label-guided contrastive learning and noisy label rectification to a unified framework.
- We construct a large-scale noisy labeled person Re-ID dataset "LUPerson-NL" as a new variant of "LUPerson". It is by far the largest noisy labeled person Re-ID dataset without any human labeling effort.
- Our models pre-trained on LUPerson-NL push the state-of-the-art results on various public benchmarks to a new limit without bells and whistles.

2. Related Work

Supervised Person Re-ID. Most studies of person Re-ID employ supervised learning. Some [6, 21, 55] introduce a hard triplet loss on the global feature, ensuring a closer feature distance for the same identity, while some [45, 59, 60] impose classification loss to learn a global feature from the whole image. There are also some other works that learn part-based local features with separate classification losses. For example, Suh *et al*. [46] presented part-aligned bi-linear representations and Sun *et al*. [48] represented features as horizontal strips. Recent approaches investigate learning invariant features concerning views [34], resolutions [31], poses [32], domains [22,23], or exploiting group-wise losses [36] or temporal information [18, 27] to improve performance. The more advantageous results on public benchmarks are achieved by MGN [51], which learns both global and local features with multiple losses. In [40], Qian et al further demonstrated the potential of generating cross-view images for person re-indentification conditioned on normalized poses. In this paper, we focus on model pre-training, and our pre-trained models can be applied to these representative methods and boost their performance.

Unsupervised Person Re-ID. To alleviate the lack of precise annotations, some works resort to unsupervised training on unlabeled datasets. For example, MMCL [49] formulates unsupervised person Re-ID as a multi-label classification to progressively seek true labels. BUC [33] jointly optimizes the network and the sample relationship with a bottom-up hierarchical clustering. MMT [14] collaboratively trains two networks to refine both hard and soft pseudo labels. SpCL [15] designs a hybrid memory to unify the representations for clustering and instance-wise contrastive learning. Both MMT [14] and SpCL [15] rely on explicit clustering of features from the whole training set, making them quite inefficient on large datasets like MSMT17. Since the appearance ambiguity is difficult to address without supervision, these unsupervised methods have limited performance. One alternative to address this issue is introducing model *pre-training on large scale data*. Inspired by the success of self-supervised representation learning [4,5,7,17,19,28,53], Fu *et al*. [12] proposed a large scale unlabeled Re-ID dataset, LUPerson, and illustrated the effectiveness of its unsupervised pre-trained models. In this work, we further try to make use of *noisy labels* from video tracklets to improve the pre-training quality through large-scale weakly-supervised pre-training.

Weakly Supervised Person Re-ID. Several approaches also employ weak supervision in person Re-ID training.

Instead of requiring bounding boxes within each frame, Meng *et al*. [38] rely on precise video-level labels, which reduces annotation cost but still need manual efforts to label videos. On the contrary, we resort to noisy labels that can be automatically generated from tracklets on a much larger scale. Some [8, 29, 50] also leverage tracklets to supervise the training of Re-ID tasks. But unlike these approaches, we are proposing a ***large-scale pre-training*** strategy for person Re-ID, by both building a new very large-scale dataset and devising a new pre-training framework: the new dataset, LUPerson-NL, is even larger than LUPerson [12] and has large amount of noisy Re-ID labels; The new framework, PNL, combines supervised learning, label-guided contrastive learning and prototype based contrastive learning to exploit the knowledge under large-scale noise labels. Most importantly, our pre-trained models have demonstrated remarkable performance and generalization ability, helping achieve state-of-the-art results superior to all existing methods on public person Re-ID benchmarks.

3. LUPerson-NL: LUPerson With Noisy Labels

Supervised models based on deep networks are always data-hungry, but the labeled data they rely on are expensive to acquire. It is a tremendous issue for person Re-ID task, since the human labelers need to check across multiple views to ensure the correctness of Re-ID labels. The data shortage is partially alleviated by a recently published dataset, LUPerson [12], a dataset of unlabeled person images with a significantly larger scale than previous person Re-ID datasets. Unsupervised pre-trained models [12] on LUPerson have demonstrated remarkable effectiveness without utilizing additional manual annotations, which arouses our curiosity: *can we further improve the performance of pre-training directly by utilizing temporal correlation as weak supervision?* To verify this, we build a new variant of LUPerson on top of the raw videos from LUPerson and assign label to each person image with automatically generated tracklet. We name it **LUPerson-NL** with **NL** standing for **N**oisy **L**abels. It consists of $10M$ images with about $430K$ identities collected from $21K$ scenes. To our best knowledge, this is the largest person Re-ID dataset constructed without human labelling by far. Our **LUPerson-NL** will be released for scientific research only, while any usage for other purpose is forbidden.

3.1. Constructing LUPerson-NL

We utilize the off-the-shelf tracking algorithm [56] [1] to detect persons and extract person tracklets from the same raw videos of [12]. We assign each tracklet with a unique class label. The detection is not perfect: *e.g*. the bounding boxes may only cover partial bodies without heads or upper parts. Human pose estimation [47] is thus appended that helps filter out imperfect boxes by predicting landmarks.

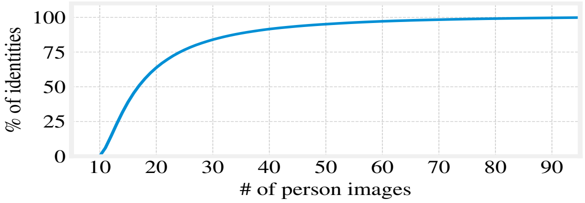
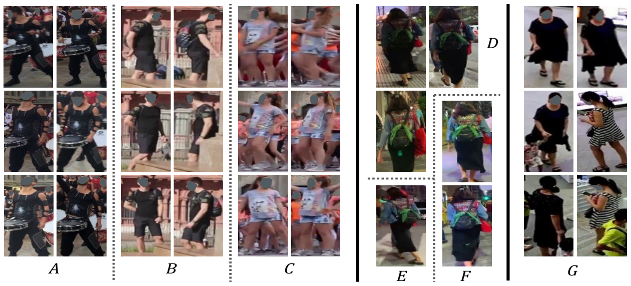


Figure 2. Identity distribution of LUPerson-NL. A curve point (X, Y) indicates $Y\%$ of identities each has less than X images.



(a) Correctly labeled identities (b) Noise-I (c) Noise-II

Figure 3. Besides the correctly labeled identities as shown by (a), there are two types of labeling errors in LUPerson-NL. `Noise-I`: same person labeled as different identities, *e.g*. D, E and F shown in (b). `Noise-II`: different persons labeled as the same identity, *e.g*. G shown in (c).

We track every person in the video frame by frame. In order to guarantee both the sufficiency and diversity, we adopt the following strategy: i) We first remove the person identities that appear in too few frames, *i.e*. no more than 200; ii) Within the tracklet of each identity, we then perform sampling with a rate of one image per 20 frames to reduce the number of duplicated images. Thus we can make sure that there would be at least 10 images associating to each identity. Through this filtering procedure, we have collected $10,683,716$ images of $433,997$ identities in total. They belong to $21,697$ videos which are less than the videos that [12] uses, due to our extra filtering strategy for more reliable identity labels. Thus, LUPerson-NL is very different from LUPerson, as it adopts very different sampling and post-processing strategies, not to mention the noisy labels driven from the spatial-temporal information.

3.2. Properties of LUPerson-NL

LUPerson-NL is advantageous in following aspects: **Large amount of images and identities.** We detail the statistics of existing popular person Re-ID datasets in Table 1. As we can see, the proposed LUPerson-NL, with over $10M$ images and $433K$ noisy labeled identities, is the second largest among the listed. Indeed, SYSU30K has more images, but it extracts images from only *1K TV program videos frame by frame*, making it less competitive in variability and less compatible in practice, the pre-training performance comparison can be found at supplementary materials. Besides, LUPerson-NL was constructed without human labeling effort, making it more suitable to scale-up. **Balanced distribution of identities.** We illustrate the cu-

[1] FairMOT: `https://github.com/ifzhang/FairMOT`

Datasets	#images	#scene	#persons	labeled	environment	camera view	detector	crop size
VIPeR [16]	1,264	2	632	yes	-	fixed	hand	128 × 48
GRID [35]	1,275	8	1,025	yes	subway	fixed	hand	vary
CUHK03 [30]	$14,096$	2	$1,467$	yes	campus	fixed	DPM [11]+hand	vary
Market [58]	$32,668$	6	$1,501$	yes	campus	fixed	DPM [11]+hand	128 × 64
Airport [25]	$39,902$	6	$9,651$	yes	airport	fixed	ACF [10]	128 × 64
DukeMTMC [61]	$36,411$	8	$1,852$	yes	campus	fixed	Hand	vary
MSMT17 [52]	$126,441$	15	$4,101$	yes	campus	fixed	FasterRCNN [42]	vary
SYSU30K [50]	29,606,918	1,000	30,508	weakly	TV program	dynamic	YOLOv2	vary
LUPerson [12]	$4,180,243$	$46,260$	$> 200k$	no	vary	dynamic	YOLOv5	vary
LUPerson-NL	$10,683,716$	$21,697$	$\simeq 433,997$	noisy	vary	dynamic	FairMOT [56]	vary

Table 1. Comparing statistics among existing popular Re-ID datasets. LUPerson-NL is by far the largest Re-ID dataset with better diversity without human labeling effort. SYSU30K is partly annotated by human annotator.

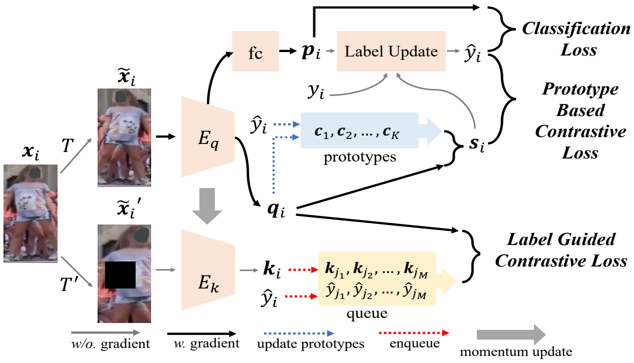


Figure 4. The overview of our PNL framework. It comprises a supervised classification module, a prototype based contrastive learning module, and a label-guided contrastive learning module.

mulative percentage of identities with respect to the number of their corresponding person images as a curve in Figure 2. A point (X, Y) on the curve represents that there are in total $Y\%$ identities in LUPerson-NL, that each of them has less than X images. It can be observed that: i) about 75% of all the identities in LUPerson-NL have a person image number within $[10, 25]$; ii) the percentage of identities that have more than 50 person images each, occupy only a very small portion of about 6.4% $(27,767/433,997)$ in LUPerson-NL. These observations all show that our LUPerson-NL is well balanced in terms of identity distribution, making it a suitable dataset for person Re-ID tasks.

In spite of our dedicatedly designed tracking and filtering strategies as proposed in Sec 3.1, the identity labels we obtained can never be very accurate due to the technical upper bounds of current tracking methods. Figure 3 visualizes the two noise types in LUPerson-NL that are caused by different labeling errors, which are `Noise-I`, where the same person is split into different tracklets and is mistaken as different persons; and `Noise-II`, that different persons are recognized as the same person.

4. PNL: Pre-training with Noisy Labels for Person Re-ID

Based on the new LUPerson-NL dataset with large scale noisy labels, we devise a novel **P**retraining framework with **N**oisy **L**abels for person Re-ID, namely **PNL**.

Denote all the data samples from LUPerson-NL as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, with n being the size of the dataset, $\boldsymbol{x}_i$ a person image and $y_i \in \{1, \ldots, K\}$ its associated identity label. Here K represents the number of all identities that are recorded in LUPerson-NL.

Inspired by recent methods [4, 5, 7, 17, 19, 28], our PNL framework adopts Siamese networks that have been fully investigated for contrastive representation learning. As shown by Figure 4, given an input person image $\boldsymbol{x}_i$, we first perform two randomly selected augmentations (T, T'), producing two augmented images $(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_i')$. We feed one of them, $\tilde{\boldsymbol{x}}_i$, into an encoder E_q to get a *query feature* $\boldsymbol{q}_i$; while the other one, $\tilde{\boldsymbol{x}}_i'$, is fed into another encoder E_k to get a *key feature* $\boldsymbol{k}_i$. Following [19], we design E_k to be a momentum version of E_q, *i.e.* the two encoders E_k and E_q share the same network structure, but with different weights. The weights in E_k are exponential moving averages of the weights in E_q. During training, weights of E_k are refreshed through a momentum update from E_q. And the detailed algorithm can be found at supplementary materials.

4.1. Supervised Classification

Since the raw labels $\{y_i\}_{i=1}^n$ in LUPerson-NL contain lots of noises as illustrated in previous section, they have to be rectified during training. Let $\hat{y}_i$ be the rectified label of image $\boldsymbol{x}_i$. As long as $\hat{y}_i$ is given, it would be intuitive that we train classification based on the corrected label $\hat{y}_i$. In particular, we would append a classifier to transform the feature from E_q into probabilities $\boldsymbol{p}_i \in \mathbb{R}^K$ with K being the number of classes. Then we impose a classification loss

$$\mathcal{L}_{ce}^i = -\log(\boldsymbol{p}_i[\hat{y}_i]). \tag{1}$$

However, the acquisition of $\hat{y}_i$ is not straight-forward. We resort to *prototypes*, the moving averaged centroids of features from training instances, to accomplish this task.

4.2. Label Rectification with Prototypes

As depicted by Figure 4, we maintain prototypes as a dictionary of feature vectors $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$, where K is the number of identities, $\boldsymbol{c}_k \in \mathbb{R}^d$ is a prototype representing a class-wise feature centroid. In each training step, we would first evaluate the similarity score s_i^k between the query feature $\boldsymbol{q}_i$ and each of the current prototypes $\boldsymbol{c}_k$ by

$$s_i^k = \frac{\exp(\boldsymbol{q}_i \cdot \boldsymbol{c}_k/\tau)}{\sum_{k=1}^{K}\exp(\boldsymbol{q}_i \cdot \boldsymbol{c}_k/\tau)}. \tag{2}$$

Let $\boldsymbol{p}_i$ be the classification probability given by the classifier with weights updated in the previous step. The rectified label $\hat{y}_i$ for this step is then generated by combining both the prototype scores $\boldsymbol{s}_i = \{s_i^k\}_{k=1}^{K}$ and the classification probability $\boldsymbol{p}_i$ as

$$\begin{aligned} \boldsymbol{l}_i &= \frac{1}{2}(\boldsymbol{p}_i + \boldsymbol{s}_i), \\ \hat{y}_i &= \begin{cases} \arg\max_j \boldsymbol{l}_i^j & \text{if } \max_j \boldsymbol{l}_i^j > T, \\ y_i & \text{otherwise.} \end{cases} \end{aligned} \tag{3}$$

Here we compute a soft pseudo label $\boldsymbol{l}_i$ and convert it to a hard one $\hat{y}_i$ based on a threshold T. If the highest score in $\boldsymbol{l}_i$ is larger than T, the corresponding class would be selected as $\hat{y}_i$, otherwise the original raw label y_i would be kept.

4.3. Prototype Based Contrastive Learning

The newly rectified label $\hat{y}_i$ can then be used to supervise the cross-entropy loss $\mathcal{L}_{ce}^i$ for classification as formulated by Equation 1. Besides, it also helps train prototypes $\boldsymbol{c}_k$ in return. In specific, we propose a *prototype based contrastive loss* $\mathcal{L}_{pro}^i$ to constrain that the feature of each sample should be closer to the prototype it belongs to. We formulate the loss as

$$\mathcal{L}_{pro}^i = -\log\frac{\exp(\boldsymbol{q}_i \cdot \boldsymbol{c}_{\hat{y}_i}/\tau)}{\sum_{j=1}^{K}\exp(\boldsymbol{q}_i \cdot \boldsymbol{c}_j/\tau)}, \tag{4}$$

with $\boldsymbol{q}_i$ being the query feature from E_q, τ being a hyperparameter representing temperature.

All the prototypes are maintained as a dictionary, with step-wise updates following a momentum mechanism as

$$\boldsymbol{c}_{\hat{y}_i} = m\boldsymbol{c}_{\hat{y}_i} + (1-m)\boldsymbol{q}_i. \tag{5}$$

4.4. Label-Guided Contrastive Learning

Instance-wise contrastive learning proved to be very effective in self-supervised learning [4, 5, 7, 17, 19]. It learns instance-level feature discrimination by encouraging similarity among features from the same instance, while promoting dissimilarity between features from different instances. The instance-wise contrastive loss is given by

$$\mathcal{L}_{ic}^i = -\log\frac{\exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_i^+/\tau)}{\exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_i^+/\tau) + \sum_{j=1}^{M}\exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_j^-/\tau)}, \tag{6}$$

with $\boldsymbol{q}_i$ being the query feature of current instance i. $\boldsymbol{k}_i^+ (= \boldsymbol{k}_i)$ is the positive key feature generated from the momentum encoder E_k. It is marked *positive* since it shares the same instance with $\boldsymbol{q}_i$. $\boldsymbol{k}_*^- \in \mathbb{R}^d$, on the contrary, are the rest features stored in a queue that represent *negative* samples. The queue has a size of M. At the end of each training step, the queue would be updated by en-queuing the new key feature and de-queuing the oldest one.

Such instance-level contrastive learning is far from perfect, as it neglects the relationships among different instances. For example, even though two instances depict the same person, it would still strengthen the gap between their features. Instead, we propose a *label guided contrastive learning* module, making use of the rectified labels $\hat{y}_i$ to ensure more reasonable grouping of contrastive pairs.

We redesign the queue to additionally record labels $\hat{y}_i$. Represented by $\mathcal{Q} = [(\boldsymbol{k}_{j_t}, \hat{y}_{j_t})]_{t=1}^{M}$, our new queue accepts not only a key feature $\boldsymbol{k}_i$ but also its rectified label $\hat{y}_i$ during update. These newly recorded labels help better distinguish positive and negative pairs. Let $\mathcal{P}(i)$ be the new set of positive features and $\mathcal{N}(i)$ the new set of negative features: features in $\mathcal{P}(i)$ share the same rectified label with the current instance i while features in $\mathcal{N}(i)$ do not. Our label guided contrastive loss can be given by

$$\mathcal{L}_{lgc}^i = \frac{-1}{|\mathcal{P}(i)|}\log\frac{\sum\limits_{\boldsymbol{k}^+\in\mathcal{P}(i)}\exp\left(\frac{\boldsymbol{q}_i\cdot\boldsymbol{k}^+}{\tau}\right)}{\sum\limits_{\boldsymbol{k}^+\in\mathcal{P}(i)}\exp\left(\frac{\boldsymbol{q}_i\cdot\boldsymbol{k}^+}{\tau}\right) + \sum\limits_{\boldsymbol{k}^-\in\mathcal{N}(i)}\exp\left(\frac{\boldsymbol{q}_i\cdot\boldsymbol{k}^-}{\tau}\right)}, \tag{7}$$

with

$$\begin{aligned} \mathcal{P}(i) &= \{\boldsymbol{k}_{j_t} | \hat{y}_{j_t} = \hat{y}_i, \forall(\boldsymbol{k}_{j_t}, \hat{y}_{j_t}) \in \mathcal{Q}\} \cup \{\boldsymbol{k}_i\}, \\ \mathcal{N}(i) &= \{\boldsymbol{k}_{j_t} | \hat{y}_{j_t} \neq \hat{y}_i, \forall(\boldsymbol{k}_{j_t}, \hat{y}_{j_t}) \in \mathcal{Q}\}, \end{aligned} \tag{8}$$

where $\boldsymbol{k}_i$ and $\hat{y}_i$ are the key feature and the rectified label of the current instance i.

Finally we combine all the components above to pretrain models on LUPerson-NL with the following loss

$$\mathcal{L}^i = \mathcal{L}_{ce}^i + \lambda_{pro}\mathcal{L}_{pro}^i + \lambda_{lgc}\mathcal{L}_{lgc}^i. \tag{9}$$

We set $\lambda_{pro} = \lambda_{lgc} = 1$ during training.

5. Experiments

5.1. Implementation

Hyper-parameter settings. We set the hyper-parameters $\tau = 0.1$ and $T = 0.8$. The momentum m for updating both the momentum encoder E_k and the prototypes is set to 0.999. More hyper-parameters exploration and training details can be found at supplementary materials.

Dataset and protocol. We conduct extensive experiments on four popular person Re-ID datasets: CUHK03, Market, DukeMTMC and MSMT17. We adopt their official settings, except CUHK03 where its labeled counterpart with new protocols proposed in [62] is used. We follow the standard evaluation metrics: the mean Average Precision (mAP) and the Cumulated Matching Characteristics top-1 (cmc1).

pre-train	Trip [21]	IDE [59]	MGN [51]
IN sup.	45.2/63.8	50.6/55.9	70.5/71.2
IN unsup.	55.5/61.2	52.5/57.7	67.1/67.0
LUP unsup.	62.6/67.6	57.6/62.3	74.7/75.4
LUPnl pnl.	**69.1/73.1**	**68.3/73.5**	**80.4/80.9**

(a) CUHK03

pre-train	Trip [21]	IDE [59]	MGN [51]
IN sup.	76.2/89.7	74.1/90.2	87.5/95.1
IN unsup.	75.1/88.5	74.5/89.3	88.2/95.3
LUP unsup.	79.8/71.5	77.9/91.0	91.0/96.4
LUPnl pnl.	**81.2/91.4**	**82.4/92.8**	**91.9/96.6**

(b) Market1501

pre-train	Trip [21]	IDE [59]	MGN [51]
IN sup.	65.2/80.7	62.8/80.8	79.4/89.0
IN unsup.	65.4/81.1	63.4/81.6	79.5/89.1
LUP unsup.	69.8/83.1	65.9/82.2	82.1/91.0
LUPnl pnl.	**71.0/84.7**	**70.3/85.0**	**84.3/92.0**

(c) DukeMTMC

pre-train	Trip [21]	IDE [59]	MGN [51]
IN sup.	34.3/54.8	36.2/66.2	63.7/85.1
IN unsup.	34.4/55.4	37.6/67.3	62.7/84.3
LUP unsup.	36.6/57.1	39.8/68.9	65.7/85.5
LUPnl pnl.	**41.4/61.6**	**44.0/72.0**	**68.0/86.0**

(d) MSMT17

Table 2. Comparing three supervised Re-ID baselines using different pre-trained models. “IN sup.”/“IN unsup.” indicates model that is supervisely/unsupervisely pre-trained on ImageNet; “LUP unsup.” is the model unsupervisely pre-trained on LUPerson; “LUPnl pnl.“ refers to the model that pre-trained on LUPerson-NL using our PNL framework. All results are shown in *mAP*/*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
LUPnl pnl.	**72.4/88.8**	**85.2/94.2**	**88.3/95.5**	**90.1/96.2**	**91.3/96.4**	**42.0/61.6**	**83.7/94.0**	**88.1/95.2**	**90.5/96.3**	**91.6/96.4**

(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
LUPnl pnl.	**60.6/75.8**	**74.5/86.3**	**78.8/88.3**	**81.6/89.5**	**83.3/91.2**	**52.2/64.1**	**77.7/87.9**	**81.1/89.6**	**83.2/91.1**	**84.1/91.3**

(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
LUPnl pnl.	**28.2/51.1**	**47.7**/71.2	**55.5**/77.2	**61.6/81.8**	**66.1**/84.8	**24.5/42.7**	**53.2/74.4**	**62.2/81.0**	**65.8/83.8**	**67.4/85.3**

(c) MSMT17

Table 3. Comparing pre-trained models on three labeled Re-ID datasets, under the *small-scale* setting and the *few-shot* setting, with different usable data percentages. “LUPnl pnl.” is our model pre-trained on LUPerson-NL using PNL. Results are shown in *mAP*/*cmc1*.

5.2. Improving Supervised Re-ID

To evaluate our pre-trained model based on LUPerson-NL with respect to supervised person Re-ID tasks. we conduct experiments using three representative supervised Re-ID baselines with different pre-training models. These baseline methods include two simpler approaches driven only by the triplet loss (Trip [21]) or the classification loss (IDE [59]), as well as a stronger and more complex method MGN [51] that use both triplet and classification losses.

We report results in Table 2, where the abbreviations {“IN”, “LUP”, “LUPnl”} represent ImageNet [43], LUPerson [12] and our LUPerson-NL respectively; while the {“sup.”, “unsup.”, “pnl.”} stand for the {“supervised”, “unsupervised”, and “pretrain with noisy label”} pre-training methods. *e.g.* the **“LUPnl pnl.”** in the bottom rows of Table 2 all refer to our model, which is pre-trained on our LUPerson-NL dataset using our PNL framework.

From Table 2 we can see, for all of the three baseline methods, our pre-trained model improves their performances greatly on the four popular person Re-ID datasets. Specifically, the improvements are at least 5.7%, 0.9%, 1.2% and 2.3% in terms of *mAP* on CUHK03, Market1501, DukeMTMC and MSMT17 respectively.

Note that even though the performance of the baseline MGN on Market1501 has been extremely high, our model still brings considerable improvement over it. The other way around, our pre-trained models obtain more significant improvements on relatively weak methods (Trip and

IDE), unveiling that model initialization plays a critical part in person Re-ID training.

Our noisy label guided pre-training models are also significantly advantageous over the previous "LUPerson unsup" models, which emphasizes the superiority of our PNL framework and our LUPerson-NL dataset.

5.3. Improving Unsupervised Re-ID Methods

Our pre-trained model can also benefit unsupervised person Re-ID methods. Based on the state-of-the-art unsupervised method SpCL [15], we explore different pre-training models utilizing two settings proposed by SpCL: the pure unsupervised learning (USL) and the unsupervised domain adaptation (UDA). Results in Table 4 illustrate that our pre-trained model outperforms the others in all UDA tasks, as well as the USL task on DukeMTMC dataset. In the USL task on Market1501, we achieve the second best scores slightly lower than the LUPerson model [12].

5.4. Comparison on Small-scale and Few-shot

Following the same protocols proposed by [12], we conduct experiments under two small data settings: the *small-scale* setting and the *few-shot* setting. The small-scale setting restricts the percentage of usable identities, while the few-shot setting restricts the percentage of usable person images each identity has. Under both settings, we vary the usable data percentages of three popular datasets from $10\% \sim 100\%$. We compare different pre-trained models under these settings with MGN as the baseline method. The results shown in Table 3 verify the consistent improvements brought by our model on all the datasets under both settings.

Besides, the results in Table 3 show that the gains of our pre-trained models are even larger under a more limited amount of labeled data. For example, under the "small-scale" setting, our model outperforms "LUPerson unsup" by 7.8%, 7.1% and 2.7% on Market1501, DukeMTMC and MSMT17 respectively with 10% identities. The improvements rise to 15.6%, 16.4% and 6.5% under the "few-shot" setting with 10% person images.

Most importantly, our pre-trained "LUPnl pnl" model helps achieve advantageous results with a *mAP* of 72.4 and a *cmc1* of 88.8, using only 10% labeled data from the Market1501 training set. The task is really challenging, considering that the training set composes only $1,170$ images belonging to 75 identities; while evaluations are performed on a much larger testing set with $19,281$ images belonging to 750 identities. We consider these results extremely appealing as they demonstrate the strong potential of our pre-trained models in real-world applications.

5.5. Comparison with other pre-training methods

We compare our proposed PNL with some other popular pre-training methods in Table 5. LUP [12] is a varient of MoCoV2 for person Re-ID based on unsupervised contrastive learning, while SupCont [26] considers both supervised and constrastive learning. Our PNL outperforms all these rep-resentative pre-training methods, indicating the superiority of our proposed method.

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
LUPnl pnl.	75.6/89.3	**68.1/82.0**	**80.7/92.2**	**72.2/84.9**

Table 4. Performances of different pre-trained models on the unsupervised Re-ID method SpCL [15] under two unsupervised task settings: the pure unsupervised learning (USL) and the unsupervised domain adaptation (UDA). Here M and D refer to the Market1501 dataset and the DukeMTMC dataset respectively.

method	SupCont [26]	LUP [12]	PNL(ours)
MSMT17	66.5/84.7	65.3/84.0	68.0/86.0

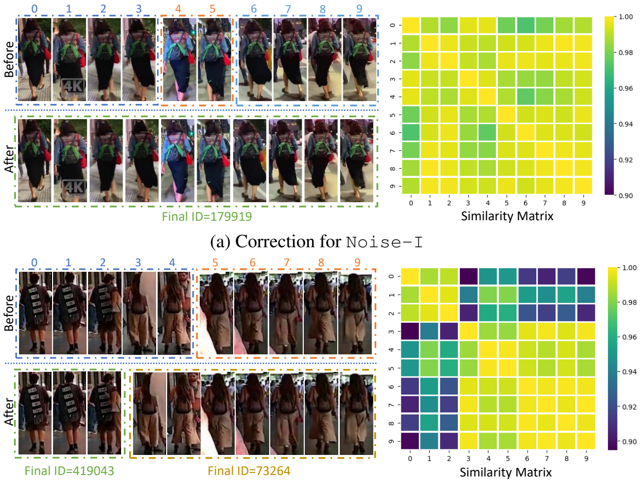
Table 5. Performance comparison for different pre-training methods on LUPerson-NL dataset.

#	ce	ic	pro	lgc	20%	40%	100%
1	✓				32.0/56.1	45.0/69.5	62.7/83.0
2		✓			34.5/59.5	47.9/72.6	65.3/84.0
3	✓	✓			37.6/62.6	49.6/73.5	66.5/84.7
4	✓		✓		35.7/59.1	48.5/72.4	65.8/84.1
5	✓			✓	38.5/63.0	50.9/74.5	67.1/85.2
6	✓	✓	✓		39.0/63.4	51.7/74.4	67.4/85.4
7	✓		✓	✓	**39.6/63.7**	**51.9/75.0**	**68.0/86.0**

Table 6. Ablating components of PNL on MSMT with data percentages 20%, 40% and 100% under the small scale setting. ce: supervised classification; ic: instance-wise contrastive learning; pro: prototypes for both prototype-based contrastive learning and label rectification; lgc: label-guided contrastive learning.

5.6. Ablation Study

We also investigate the effectiveness of each designed component in PNL through ablation experiments. Results shown by Table 6 illustrate the efficacy of our proposed components. We have the following observations: **i)** Training with an instance-wise contrastive loss $\mathcal{L}_{ic}^{i}$ (row 2) without using any labels leads to even better performance than training with a classification loss $\mathcal{L}_{ce}^{i}$ (row 1) that utilizes the labels from LUPerson-NL, implying that the noisy labels in LUPerson-NL would misguide representation learning if directly adopted as supervision. **ii)** Jointly training with both losses $\mathcal{L}_{ce}^{i}$ and $\mathcal{L}_{ic}^{i}$ (row 3) improves over using only one loss (row 1, row 2), suggesting that learning instance-wise discriminative representations complements label supervision. **iii)** The prototypes which contribute to both prototype-based contrastive learning and the label correction, are very important under various settings, as verified by comparing row 1 with row 4; row 3 with row 6; and row 5 with row 7. **iv)** Our label-guided contrastive learning



(a) Correction for Noise-I

(b) Correction for Noise-II

Figure 5. Visualizing the label correction functionality of our PNL framework with respect to the two noise types from LUPerson-NL. Person images in the same rectangle indicate that they are recognized as the same identity. The right-hand similarity matrices are calculated based on the image features all learned using our PNL framework with label correction.

component consistently outperforms the vanilla instance-wise contrastive learning under various settings, as verified by comparing row 3 with row 5, and row 6 with row 7. **v)** Combining all our components (supervised classification, prototypes and label-guided contrastive learning) together leads to the best performance as shown by row 7.

5.7. Label Correction

Our PNL can indeed correct noisy labels. We demonstrate two typical examples in Figure 5 visualizing the label corrections with respect to the two kinds of noises. As we can see, in Figure 5a the same person are marked as three different persons in LUPerson-NL due to Noise-I in labels. After our PNL pre-training, these three tracklets are merged together since their trained features are very close, as verified by the right-hand similarity matrix. In Figure 5b different persons are labeled as the same identity in LUPerson-NL due to Noise-II in labels. After PNL training, these mis-labeled person identities are all correctly re-grouped into two identities, which can also be reflected by the right-hand similarity matrix.

we also ablate the label correction module in Table 7 with different settings, and observe it can improve the performance. It also validates the importance of combining label rectification with label-guided contrastive learning together, where more accurate positive/negative pairs can be leveraged.

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

We compare our results with current state-of-the-art methods on four public benchmarks. We don't apply any post-processing techniques such as IIA [13] and RR [62]. To ensure fairness, we adopt ResNet50 as our backbone and does not compare with methods that rely on stronger backbones (results with stronger backbones *e.g*. ResNet101 can be found at supplementary materials). Results in Table 8 verify the remarkable advantage brought by our pre-trained models. Without bells and whistles, we achieve state-of-the-art performance on all four benchmarks, outperforming the second with clear margins.

setting	*ce+pro*		*ce+pro+lgc*	
	w/o. *lc*	w. *lc*	w/o. *lc*	w. *lc*
MSMT17	64.8/83.4	65.8/84.1	66.7/85.0	68.0/86.0

Table 7. Ablating the label correction. lc: "label correction".

Method	CUHK03	Market1501	DukeMTMC	MSMT17
MGN† [51] (2018)	70.5/71.2	87.5/95.1	79.4/89.0	63.7/85.1
BOT [37] (2019)	-	85.9/94.5	76.4/86.4	-
DSA [57] (2019)	75.2/78.9	87.6/95.7	74.3/86.2	-
Auto [41] (2019)	73.0/77.9	85.1/94.5	-	52.5/78.2
ABDNet [3] (2019)	-	88.3/95.6	78.6/89.0	60.8/82.3
SCAL [2] (2019)	72.3/74.8	89.3/95.8	79.6/89.0	-
MHN [1] (2019)	72.4/77.2	85.0/95.1	77.2/89.1	-
BDB [9] (2019)	76.7/79.4	86.7/95.3	76.0/89.0	-
SONA [54] (2019)	79.2/81.8	88.8/95.6	78.3/89.4	-
GCP [39] (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 [63] (2020)	74.1/76.5	88.6/95.3	80.0/89.6	-
GASM [20] (2020)	-	84.7/95.3	74.4/88.3	52.5/79.5
ESNET [44] (2020)	-	88.6/95.7	78.7/88.5	57.3/80.5
LUP [12](2020)	79.6/81.9*	91.0/96.4	82.1/91.0	65.7/85.5
Ours+MGN	80.4/80.9	**91.9/96.6**	**84.3/92.0**	**68.0/86.0**
Ours+BDB	**82.3/84.7**	88.4/95.4	79.0/89.2	53.4/79.0

Table 8. Comparison with the state of the art. Numbers of MGN† come from a re-implementation based on FastReID, which are even better than the original. Numbers of PNL marked by * are obtained on BDB, the rest without the * mark are obtained on MGN. We show best scores in bold and the second scores underlined.

6. Conclusion

In this paper, we demonstrate that large-scale Re-ID representation can be directly learned from massive raw videos by leveraging the spatial and temporal information. We not only build a large-scale noisy labeled person Re-ID dataset **LUPerson-NL** based on tracklets of raw videos from LUPerson without using manual annotations, but also design a novel weakly supervised pretraining framework **PNL** comprising different learning modules including supervised learning, prototypes-based learning, label-guided contrastive learning and label rectification. Equipped with our pre-trained models, we push existing benchmark results to a new limit, which outperforms unsupervised pre-trained models and ImageNet supervised pre-trained models by a large margin.

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

References

[1] 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

[2] 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

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8351–8361, 2019. 8

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 4, 5

[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. 2, 4, 5

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 4, 5

[8] Zhirui Chen, Jianheng Li, and Wei-Shi Zheng. Weakly supervised tracklet person re-identification by deep feature-wise mutual learning. *arXiv preprint arXiv:1910.14333*, 2019. 3

[9] Zuozhuo 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. 8

[10] 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

[11] 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

[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. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 1, 2, 3, 4, 6, 7, 8

[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. 8

[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

[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, 7

[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. 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, 4, 5

[18] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9647–9656, 2019. 2

[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. 2, 4, 5

[20] 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

[21] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 6

[22] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9527–9536, 2019. 2

[23] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020. 2

[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, Ziyan 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, 4

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020. 7

[27] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3958–3967, 2019. 2

[28] Junnan Li, Caiming Xiong, and Steven Hoi. Mopro: Webly supervised learning with momentum prototypes. In *International Conference on Learning Representations*, 2020. 2, 4

[29] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1770–1782, 2019. 3

[30] 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. 4

[31] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8090–8099, 2019. 2

[32] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7919–7929, 2019. 2

[33] 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

[34] Fangyi Liu and Lei Zhang. View confusion feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6639–6648, 2019. 2

[35] 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. 4

[36] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4976–4985, 2019. 2

[37] 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

[38] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2019. 3

[39] 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

[40] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018. 2

[41] 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

[42] 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

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[44] Dong Shen, Shuai Zhao, Jinming Hu, Hao Feng, Deng Cai, and Xiaofei He. Es-net: Erasing salient parts to learn more in re-identification. *IEEE Transactions on Image Processing*, 2020. 8

[45] 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

[46] 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. 2

[47] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[48] 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

[49] 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

[50] Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, and Liang Lin. Weakly supervised person re-id: Differentiable graphical learning and a new benchmark. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2142–2156, 2020. 3, 4

[51] 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, 8

[52] 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, 4

[53] 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

[54] 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

[55] Ye Yuan, Wuyang Chen, Yang Yang, and Zhangyang Wang. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 354–355, 2020. 2

[56] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv e-prints*, pages arXiv–2004, 2020. 2, 3, 4

[57] 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

[58] Liang Zheng, Liyue 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. 1, 4

[59] 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

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):1–20, 2017. 2

[61] 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. 1, 4

[62] 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. 5, 8

[63] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. *ECCV*, 2020. 8