Towards Precise Intra-camera Supervised Person Re-identification

Menglin Wang¹, Baisheng Lai², Haokun Chen¹, Jianqiang Huang², Xiaojin Gong¹, Xian-Sheng Hua²
¹Zhejiang University, Hangzhou, China
²Alibaba Group, Hangzhou, China
lynnwang6875@gmail.com, baisheng.lbs@alibaba-inc.com, chenhaokun@zju.edu.cn, jianqiang.hjq@alibaba-inc.com, gongxj@zju.edu.cn, huaxiansheng@gmail.com

Abstract
Intra-camera supervision (ICS) for person re-identification (Re-ID) assumes that identity labels are independently annotated within each camera view and no inter-camera identity association is labeled. It is a new setting proposed recently to reduce the burden of annotation while expect to maintain desirable Re-ID performance. However, the lack of inter-camera labels makes the ICS Re-ID problem much more challenging than the fully supervised counterpart. By investigating the characteristics of ICS, this paper proposes camera-specific non-parametric classifiers, together with a hybrid mining quintuplet loss, to perform intra-camera learning. Then, an inter-camera learning module consisting of a graph-based ID association step and a Re-ID model updating step is conducted. Extensive experiments on three large-scale Re-ID datasets show that our approach outperforms all existing ICS works by a great margin. Our approach performs even comparable to state-of-the-art fully supervised methods in two of the datasets.

1 Introduction
Person re-identification (Re-ID) is the task of matching images of the same person across disjoint cameras. Because of its significance in surveillance, this task has attracted broad research interest in recent years. Most previous works focus on fully supervised [Sun et al., 2018; Zhang et al., 2019; Chen et al., 2019; Luo et al., 2019] and unsupervised [Deng et al., 2018; Zhong et al., 2019a; Fan et al., 2018; Wu et al., 2019] settings. The performance of supervised person Re-ID has been greatly improved due to the development of deep learning techniques. However, these methods need a large amount of full annotations that are expensive and time-consuming to obtain, making them unscalable to real-world applications. Conversely, unsupervised methods require no annotations but their performance is still far from satisfactory.

This paper aims to learn a person Re-ID model under intra-camera supervision (ICS), which is a supervised setting proposed very recently [Zhu et al., 2019; Qi et al., 2019b]. It assumes that identity labels are independently annotated within each camera and no inter-camera identity association is labeled. Since the ID association across cameras is known as the most time-consuming part for manual annotation, ICS can greatly save annotation costs and make the Re-ID techniques more scalable. Nevertheless, the lack of inter-camera labels brings up more challenges when dealing with appearance variations in different cameras, leading to a performance inferior to the supervised counterparts.

In order to bridge the performance gap, we address the ICS person Re-ID problem by adopting the BNNeck [Luo et al., 2019] augmented ResNet-50 [He et al., 2016], which is simple but shown to be effective in the fully supervised Re-ID task, as our backbone. Upon the backbone we construct two branches, respectively, for intra- and inter-camera learning. The intra-camera learning aims to learn a discriminative feature representation via leveraging the per-camera annotations. In ICS, the per-camera independent labeling manner results in some identity classes containing only a few instances. Concerning this characteristic, we exploit non-parametric classifiers [Xiao et al., 2017; Wu et al., 2018] and design a camera-specific way to perform ID classification within each camera view. The non-parametric classifiers are implemented via an external memory bank, which additionally inspires us to propose a quintuplet loss. This loss takes advantage of information within each mini-batch and over global training batch to boost the intra-camera learning performance. The inter-camera learning aims to improve the Re-ID model by mining ID relationships across cameras. To this end, we design a graph-based strategy for ID association and pseudo labeling, which are further used to learn a better Re-ID model by directly training the BNNeck augmented network in a fully supervised manner.

Although both intra- and inter-camera learning perspectives are considered in existing ICS works [Qi et al., 2019a; Qi et al., 2019b; Zhu et al., 2019], our approach distinguishes itself from the others in the following aspects:

- We propose camera-specific non-parametric classifiers and a quintuplet loss for intra-camera learning. These designs are customized not only for the characteristics of ICS but also for the memory assisted network architecture. They enable our intra-camera learning module to achieve a Re-ID performance better than all existing ICS models that consider both intra- and inter-camera
learning parts.

- In inter-camera learning, our graph-based association step produces desirable pseudo labeling results, which enables us to directly apply the BNNeck augmented network architecture to train the Re-ID model in a fully supervised manner. Riding the wave of architectures successfully applied in the supervised Re-ID task boosts the performance further.

- Extensive experiments on three large-scale Re-ID datasets including Market-1501 [Zheng et al., 2015], DukeMTMCReID [Ristani et al., 2016; Zheng et al., 2017], and MSMT17 [Wei et al., 2018], show that the proposed approach outperforms previous ICS works by a great margin. Our performance is even comparable to fully supervised methods on the first two datasets.

2 Related Work

2.1 Person Re-identification

Fully supervised person Re-ID has made significant progress relying on the success of deep learning techniques. However, it remains to be an unsolved problem due to challenges arising from cluttered background, occlusion, as well as variations in illumination, pose, and viewpoint. Recent methods have exploited part-based features [Sun et al., 2018], human semantics [Zhang et al., 2019], attention mechanisms [Chen et al., 2019], or data generation [Zheng et al., 2019] to tackle the challenges. These methods often lead to complex network architectures. An exceptional work is Bag of Tricks (BoT) [Luo et al., 2019] that achieves the state-of-the-art Re-ID performance by applying some training tricks on a baseline network. Inspired by it, we construct our network upon the BoT model, i.e. the BNNeck augmented ResNet-50, to keep our backbone simple yet effective.

Unsupervised person Re-ID has attracted a lot of research interest in recent years. Existing methods can be roughly categorized into two groups. One is based on domain adaptation techniques [Deng et al., 2018; Zhong et al., 2019a; Ding et al., 2019] that transfer knowledge from labeled source domain to unlabeled target domain. The other group is purely unsupervised that requires no external labeled data. These methods usually perform a step to associate IDs across cameras via clustering [Fan et al., 2018; Lin et al., 2019] or graph [Wu et al., 2019] based strategies. Our intra-camera supervised work also adopts a graph-based ID association step. But, in contrast to use a graph-weighted loss [Wu et al., 2019], we formulate the association as a problem of finding connected components in a graph.

Semi-supervised person Re-ID aims to learn a Re-ID model from both labeled and unlabeled data [Yang et al., 2019]. Intra-camera supervision (ICS) is a special semi-supervised setting proposed very recently. All existing ICS works address the Re-ID problem by considering both intra- and inter-camera learning parts. For supervised intra-camera learning, [Qi et al., 2019a; Qi et al., 2019b] take the extensively used triplet loss [Hermans et al., 2017] while [Zhu et al., 2019] designs a multi-branch structure to learn classifiers. For inter-camera learning, [Qi et al., 2019a] develops a multi-camera adversarial learning approach to reduce the cross-camera data distribution discrepancy. [Qi et al., 2019b] utilizes a soft-labeling scheme, and [Zhu et al., 2019] adopts a multi-label learning strategy. In contrast, we propose different learning strategies in both parts and achieve much higher Re-ID performance.

2.2 Parametric and Non-parametric Classifiers

Parametric classifiers in this work refer to those implemented by fully connected (FC) layers in a deep neural network (DNN), usually trained with a cross-entropy softmax classification loss. Such classifiers have been extensively used in fully supervised person Re-ID [Sun et al., 2018; Zhai et al., 2019; Luo et al., 2019]. However, they have the following drawbacks: 1) The training process pays much attention on learning parameters for the FC layers that are abandoned during inference for person Re-ID [Zhai et al., 2019], making the learned feature representation less discriminative for test data. 2) The classifiers cannot be learned effectively when there are a large number of identities while each identity only has a small number of instances [Xiao et al., 2017].

Non-parametric classifiers in a DNN are implemented via an external memory bank and a non-parametric variant of the softmax function. It is first proposed in a fully supervised person search task [Xiao et al., 2017] and extensively adopted in unsupervised [Wu et al., 2018; Zhong et al., 2019a; Zhong et al., 2019b] and semi-supervised [Chen et al., 2018; Yang et al., 2019] learning. A common challenge in these tasks is that the number of classes is huge but each class contains only one or few examples. A DNN equipped with the non-parametric classifiers makes its parameters independent to the class number so that the training process entirely focus on the feature representation learning. Nevertheless, the non-parametric model may overfit more easily when training data is abundant enough.

3 The Proposed Method

The intra-camera supervision assumes that identity labels are independently annotated within each camera view and no inter-camera identity association is provided. Suppose there are $C$ cameras in a dataset. We denote the set of the $c$-th camera by $D_c = \{(x_i, y_i, c_i)\}_{i=1}^{M_c}$, in which image $x_i$ is annotated with an identity label $y_i \in \{1, \cdots, N_c\}$ and a camera label $c_i = c \in \{1, \cdots, C\}$. $M_c$ and $N_c$ are, respectively, the number of total images and IDs in this camera view. $N = \sum_{c=1}^{C} N_c$ is the total ID number directly accumulated over all cameras. It should be noted that the identities in different cameras are partially overlapped. That is, a same person may appear in two or more camera views, but it could be assigned with different IDs due to the per-camera independent labeling manner. Given this training set $D = \bigcup_{c=1}^{C} D_c$, we aim to learn a person Re-ID model that can well discriminate both intra- and inter-camera identities.

To this end, we develop our method from both intra- and inter-camera learning perspectives. The overall framework is shown in Figure 1. An image is first fed into a backbone
network for feature extraction. The extracted feature goes through an additional feature embedding layer and then classified by camera-specific non-parametric classifiers that are implemented via an external memory bank, together with an ID classification loss and a quintuplet loss, for intra-camera learning. The memory bank stores the centroid feature of each ID, which is of moderate discrimination ability after intra-camera learning. Then, the ID centroids are used for ID association and pseudo labeling across cameras. In inter-camera learning, the same backbone is adopted to extract features, along with a classifier parameterized by a FC layer to classify images into their pseudo identity classes.

3.1 Intra-camera Learning

When considering the Re-ID problem within an individual camera view, it can be treated as a fully supervised classification task. Therefore, it is reasonable to formulate the intra-camera learning as a multi-task classification problem and adopt a multi-branch architecture as done in [Zhu et al., 2019]. The network architecture is designed to share a feature extraction backbone and append with multiple classification branches, each of which corresponds to a specific camera view. This architecture is capable of learning feature representations that are discriminative within cameras and also somewhat discriminative across cameras. However, the parametric classifiers implemented via fully-connected layers in the branches could become ineffective when some IDs contain only a couple of samples, which is a common situation in the intra-camera supervised setting.

Camera-specific Non-parametric Classifiers

To alleviate the above-mentioned problem, we adopt non-parametric classifiers for intra-camera learning. As illustrated in Figure 1, our network consists of a feature extraction backbone, a FC embedding layer, together with an external memory bank. Based upon this network structure, we design camera-specific non-parametric classifiers to perform the classification tasks within each camera view.

Formally, when an image is input, the FC embedding layer outputs a $d$-dimensional feature vector. The memory bank $K \in \mathbb{R}^{d \times N}$ stores the up-to-date features of all accumulated IDs and each column corresponds to one ID. During back-propagation, the memory bank is updated by

$$K[j] \leftarrow \mu K[j] + (1 - \mu) f(x_i),$$

(1)

where $K[j]$ is the $j$-th column of the memory. $f(x_i)$ is a $L_2$ normalized feature extracted from image $x_i$ that belongs to the $j$-th ID. $\mu \in [0, 1]$ is an updating rate. After each update, $K[j]$ is scaled to having unit $L_2$ norm. The updated feature in each column can be interpreted as the centroid of an identity class in the feature space, which is a $d$-dimensional unit hypersphere.

Given the image $x_i$, together with its annotated intra-camera identity label $y_i$ and camera label $c_i$. The corresponding global ID index $j$ is obtained by $j = A + y_i$, where $A = \sum_{k=1}^{c_i-1} N_k$ is the total ID number accumulated from the first to the $c_i - 1$-th camera view. Then, the probability of classifying $x_i$ into the $j$-th ID is defined by a non-parametric softmax function

$$p(j|x_i) = \frac{\exp(K[j]^T f(x_i)/\tau)}{\sum_{k=A+c_i}^{A+N_i} \exp(K[k]^T f(x_i)/\tau)},$$

(2)

where $\tau$ is the temperature controlling the smoothness of probability distribution.

Note that the non-parametric classifier defined above is camera-specific, because the sum in the denominator is over the IDs within the same camera view only. In contrast to existing works [Xiao et al., 2017; Wu et al., 2018; Zhong et al., 2019a; Ding et al., 2019] that considers all entries in their memory bank, our formulation only takes those belonging to the same camera into account while ignores the IDs in all other cameras. Thus, each non-parametric classifier is responsible for the classification task within a specific camera.
The objective of camera-specific ID classification, termed as the intra-camera ID loss, is to minimize the negative log-likelihood over all training images. That is,

\[
\mathcal{L}_{\text{Intra-ID}} = -\sum_{c=1}^{C} \left( \frac{1}{|D_c|} \sum_{x_i \in D_c} \log p(j|x_i) \right),
\]

where the normalization coefficient \(\frac{1}{|D_c|}\) is placed to balance the various number of images under different cameras.

### A Hybrid Mining Quintuplet Loss

Inspired by recent fully supervised methods [Luo et al., 2019; Chen et al., 2019; Zhang et al., 2019] that combine ID loss and triplet loss together to learn precise Re-ID models, we also incorporate a metric learning loss as a supplement to the intra-camera ID loss to boost the performance. Instead of directly using the triplet loss that only samples locally within each mini-batch, we propose a quintuplet loss, which can take advantage of information not only in each mini-batch but also over the global training batch to enhance the intra-camera compactness and inter-ID separability.

Specifically, in each mini-batch, we randomly select \(P\) identities and \(K\) instances of each identity, as the common practice [Hermans et al., 2017]. For each anchor image \(x_a\), we design a hybrid mining scheme that selects two instances, and two identity centroids to form a quintuplet. The positive and negative instances are sampled to be the hardest ones within a mini-batch. In addition, we choose the positive ID centroid and the nearest negative ID centroid from the memory bank. Note that all instances and centroids are selected from the same camera as the anchor. Then, the intra-camera quintuplet loss is defined as follows:

\[
\mathcal{L}_{\text{Intra-Quint}} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[ m_1 + \max_{p=1,\ldots,K} \|g(x^*_a) - g(x^p)\|^2 
\right. \\
- \min_{n=1,\ldots,K;j=1,\ldots,P;j \neq i} \|g(x^*_a) - g(x^j)\|^2 \\
+ \left. \left[ m_2 + \|f(x^*_a) - K[A + y_a]\|^2 \\
- \min_{j=1,\ldots,N_1} \|f(x^*_a) - K[A + j]\|^2 \right] \right],
\]

where \(m_1\) and \(m_2\) are two margins, \(g(x_a), g(x_p),\) and \(g(x_n)\) are, respectively, the features of the anchor, positive and negative instances output from the global average pooling (GAP) layer in the backbone. \(f(x_a)\) is the anchor’s feature produced from the FC embedding layer as above-mentioned. Taking features from different layers is inspired by the BNNek structure in [Luo et al., 2019] and shown to be effective in our experiments. In addition, \(A + y_a\) is the global ID index in \(K\) given the intra-camera ID \(y_a\), \(\|\cdot\|_+ = \max(0, \cdot)\), and \(\|\cdot\|\) is the Euclidean distance.

### The Loss for Intra-camera Learning

In summary, the loss for intra-camera learning is the sum of the camera-specific ID classification loss and the quintuplet loss:

\[
\mathcal{L}_{\text{Intra}} = \mathcal{L}_{\text{Intra-ID}} + \mathcal{L}_{\text{Intra-Quint}}.
\]
reID [Ristani et al., 2016; Zheng et al., 2017], and MSMT17 [Wei et al., 2018]. To simulate the ICS setting, we generate intra-camera identity labels based on the provided full annotations. Table 1 lists the numbers of cameras, IDs, and images contained in each dataset, as well as the accumulated total identity number under intra-camera supervision ($\#ID_{ICS}$), the averaged image-per-person (IP) value, together with the averaged image-per-camera-per-person (ICP) value. For performance evaluation, we adopt the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP), as the common practice.

$$\text{mAP}(\%) \quad \text{Rank-1}(\%)$$

Table 1: Statistics of each dataset. $\#Cam$, $\#ID$, $\#Img$, and $\#ID_{ICS}$ are the number of cameras, IDs, images, and accumulated IDs under ICS, respectively. IP is the averaged image-per-person value and ICP is the averaged image-per-camera-per-person value.

| Dataset      | $\#Cam$ | $\#ID$ | $\#Img$ | $\#ID_{ICS}$ | IP  | ICP  |
|--------------|---------|--------|---------|--------------|-----|------|
| Market1501   | 6       | 751    | 12,936  | 3,262        | 17.23| 3.97 |
| DukeMTMC-ReID| 8       | 702    | 16,522  | 2,196        | 23.54| 7.52 |
| MSMT17       | 15      | 1,041  | 32,621  | 4,821        | 31.34| 6.77 |

Table 2 presents the comparison results in terms of mAP(%) and Rank-1(%). From the results we observe that the camera-specific non-parametric classifiers ($M_3$) outperform the camera-agnostic counterpart ($M_2$) by a great margin, showing that the camera-specific constraint plays an important role. In addition, the camera-specific non-parametric classifiers ($M_3$) perform consistently better than the parametric counterpart ($M_1$) on all datasets. Especially, it improves the performance by a significant margin (mAP +14.2% and Rank-1 + 9.3%) on Market1501 which has a smaller ICP value than the other two datasets, as reported in Table 1. It indicates that the non-parametric classifier is superior when identity classes contain fewer examples.

4.2 Ablation Study

**Effectiveness of The Camera-specific Non-parametric Classifiers in Intra-camera Learning**

We conduct a series of experiments to validate the effectiveness of each component proposed in our method. First, we are curious about how well the camera-specific non-parametric classifiers perform. Therefore, three model variants are investigated, including $M_1$: a multi-branch network structure [Zhu et al., 2019], in which each branch uses a classifier parameterized by a FC layer and optimized with a cross-entropy ID loss, for intra-camera learning; $M_2$: a non-parametric classifier but not camera-specific, that is, any image can be classified into all accumulated ID classes; $M_3$: the proposed camera-specific non-parametric classifiers with the intra-camera ID loss only.

| Models   | Market1501 | DukeMTMC-ReID | MSMT17 |
|----------|------------|---------------|--------|
|          | mAP Rank-1 | mAP Rank-1    | mAP Rank-1 |
| $M_1$    | 55.0 76.8  | 58.9 75.3     | 25.1 50.7 |
| $M_2$    | 30.7 45.5  | 29.1 33.3     | 6.0 10.4 |
| $M_3$    | **69.2 86.1** | **61.9 78.0** | **25.7 52.1** |
| $M_4$    | 71.9 86.6  | 64.0 79.1     | 28.1 54.3 |
| $M_5$    | 72.3 87.2  | 64.7 79.7     | 28.9 55.5 |
| $M_6$    | 83.6 93.1  | 72.0 83.6     | 31.3 57.7 |
| $M_7$    | 85.9 94.1  | 77.2 87.4     | 52.2 75.3 |

Table 2: Comparison of the different model variants. $M_1$-$M_5$ are all intra-camera learning models, in which $M_1$ is a multi-branch parametric classification architecture, $M_2$ is a camera-agnostic non-parametric classifier, and $M_3$ is our proposed camera-specific non-parametric classifiers. $M_4$ is $M_3$ with an additional triplet loss and $M_5$ is $M_3$ with the proposed quintuplet loss. $M_6$ is the full model containing both intra- and inter-camera learning. $M_7$ is a fully supervised version.

**Effectiveness of The Inter-camera Learning Part**

After the comparisons of all intra-camera learning components, we validate the effectiveness of the inter-camera learning part. To this end, we add this part on $M_5$ to get the full model $M_6$ and present the results in Table 2. Meanwhile, Table 2 also provides the results obtained by training the inter-
### 4.3 Comparison with the State-of-the-Arts

In this section, we compare our approach (named as Precise-ICS) with all existing ICS person Re-ID methods, including MTML [Zhu et al., 2019], PCSL [Qi et al., 2019b] and ACAN [Qi et al., 2019a]. The comparison results are presented in Table 3 in terms of mAP($\%$), Rank-1($\%$), Rank-5($\%$), and Rank-10($\%$). From the results we observe that the proposed approach outperforms the other ICS methods by a great margin. More specifically, the mAP is 14.2%, 18.5%, and 10.6% higher and the Rank-1 accuracy is 6.1%, 11.9%, and 9.4% higher than the best performances obtained by other methods on Market1501, DukeMTMC-ReID, and MSMT17 respectively.

| Methods | Market1501 | DukeMTMC-ReID | MSMT17 |
|---------|------------|--------------|--------|
|         | Prec/Rec   | Prec/Rec     | Prec/Rec |
| Fully supervised |               |             |        |
| OSNet [Zhou et al., 2019] | 84.9/94.8 | 73.5/88.6 | 52.9/78.7 |
| DCGNet [Zheng et al., 2019] | 86.0/94.8 | 74.8/86.6 | 52.3/77.2 |
| BoT [Luo et al., 2019] | 85.9/94.5 | 76.4/86.4 | - - |
| PCB [Sun et al., 2018] | 81.6/93.8 | 69.2/83.3 | 40.4/68.2 |

| Unsupervised |               |             |        |
| ECN [Zhong et al., 2019a] | 43.0/75.1 | 40.4/63.3 | 10.2/30.2 |
| AE [Ding et al., 2019] | 58.0/81.6 | 46.7/67.9 | 11.7/32.3 |
| BUC [Lin et al., 2019] | 38.3/66.2 | 27.5/47.4 | - - |
| UGA [Wu et al., 2019] | 70.3/87.2 | 53.3/75.0 | 21.7/49.5 |

| Intra-camera supervised |               |             |        |
| MTML [Zhu et al., 2019] | 65.2/85.3 | 50.7/71.7 | 18.6/44.1 |
| PCSL [Qi et al., 2019b] | 69.4/87.0 | 53.5/71.7 | 20.7/48.3 |
| ACAN [Qi et al., 2019a] | 50.6/73.3 | 45.1/67.6 | 12.6/33.0 |
| Precise-ICS: $M_5$ (Ours) | 72.3/87.5 | 64.7/79.7 | 28.9/55.5 |
| Precise-ICS: $M_6$ (Ours) | 83.6/93.1 | 72.0/83.6 | 31.3/57.7 |

Table 3: Comparison with state-of-the-art methods. 'Precise-ICS' is the approach proposed in this work, $M_5$ refers to the model with our intra-camera learning only and $M_6$ is our full model considering both intra- and inter-camera learning. Note that no re-ranking is used during training or evaluation.

| Model | Market1501 | DukeMTMC-ReID | MSMT17 |
|-------|------------|--------------|--------|
| $M_6$ | 96.4/75.9 | 90.1/74.3 | 86.3/38.3 |

Table 4: Precision and recall of the ID pairs associated by our approach.

camera learning branch in our network with entire ground-truth labels, which in essence is the fully supervised counterpart model $M_7$. This model indicates the upper bound performance that can be achieved by our Re-ID network architecture.

The results show that our inter-camera learning part makes significant improvements, especially on Market1501 and DukeMTMC-ReID. The improvements are benefited from our intra-camera learning components that equip the Re-ID model with an outstanding discrimination capability, enabling us to gain reliable ID association results. As shown in Table 4, the precision and recall of the ID pairs associated by our approach are pretty high on Market1501 and DukeMTMC-ReID, but relatively low on MSMT17 that is a dataset much more complicated than the others. The correctly associated IDs make the inter-camera learning effective. As indicated in Table 2, the performance of our full model ($M_6$) is even close to the supervised model ($M_7$) on Market1501 and DukeMTMC-ReID.

### 5 Conclusion

In this paper, we have proposed a new approach to address the person Re-ID problem under intra-camera supervision. The proposed network consists of a simple yet effective feature extraction backbone, together with two branches for intra- and inter-camera learning respectively. According to the per-camera labeling nature of ICS, we propose camera-specific non-parametric classifiers and a hybrid mining quintuplet loss...
for intra-camera learning. The designed components exploit per-camera labels thoroughly so that our intra-camera learning part only can perform better than existing ICS methods. Benefited from the discrimination ability gained in this part, the inter-camera learning module boosts the Re-ID performance further by mining ID relationship across cameras. Our full model outperforms all ICS methods by a large margin, greatly reducing the gap to the fully supervised counterparts.

References

[Chen et al., 2018] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In ECCV, 2018.

[Chen et al., 2019] 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 ICCV, 2019.

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

[Ding et al., 2019] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. arXiv preprint arXiv:1907.04194, 2019.

[Ester et al., 1996] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996.

[Fan et al., 2018] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications, 14(4):83, 2018.

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

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

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[Krizhevsky et al., 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[Lin et al., 2019] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In AAAI, 2019.

[Luo et al., 2019] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In CVPRW, 2019.

[Qi et al., 2019a] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao. Adversarial camera alignment network for unsupervised cross-camera person re-identification. arXiv preprint arXiv:1908.00862, 2019.

[Qi et al., 2019b] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao. Progressive cross-camera soft-label learning for semi-supervised person re-identification. arXiv preprint arXiv:1908.05669, 2019.

[Ristani et al., 2016] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, 2016.

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

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

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

[Wu et al., 2019] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z. Li. Unsupervised graph association for person re-identification. In ICCV, 2019.

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

[Yang et al., 2019] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Deep semi-supervised person re-identification with external memory. In ICME, 2019.

[Zhai et al., 2019] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. In defense of the classification loss for person re-identification. In CVPRW, 2019.

[Zhang et al., 2019] Zhizheng Zhang, Cuiying Lan, Wenjun Zeng, and Zhiho Chen. Densely semantically aligned person re-identification. In CVPR, 2019.

[Zhang et al., 2015] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In ICCV, 2015.

[Zhang et al., 2017] Zhe dong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In ICCV, 2017.

[Zhang et al., 2019] Zhe dong Zheng, Xiaodong Yang, Zhi ding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In CVPR, 2019.

[Zhong et al., 2019a] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In CVPR, 2019.
in memory for person re-identification. *arXiv preprint arXiv:1908.00485*, 2019.

[Zhou et al., 2019] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019.

[Zhu et al., 2019] Xiangping Zhu, Xiatian Zhu, Minxian Li, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification: A new benchmark. In *ICCVW*, 2019.