SECRET: Self-Consistent Pseudo Label Refinement for Unsupervised Domain Adaptive Person Re-Identification

Tao He[*1,2], **Leqi Shen**[*1,2], **Zhenhua Guo**[3], **Yuchen Guo**[†2], **Guiguang Ding**[†1,2]
[1] School of Software, Tsinghua University, Beijing, China
[2] Beijing National Research Center for Information Science and Technology (BNRist)
[3] Alibaba Group
{kevin.92.he, lunarshen, cszguo, yuchen.w.guo}@gmail.com, dinggg@tsinghua.edu.cn

Abstract

Unsupervised domain adaptive person re-identification aims at learning on an unlabeled target domain with only labeled data in source domain. Currently, the state-of-the-arts usually solve this problem by pseudo-label-based clustering and fine-tuning in target domain. However, the reason behind the noises of pseudo labels is not sufficiently explored, especially for the popular multi-branch models. We argue that the consistency between different feature spaces is the key to the pseudo labels' quality. Then a **SE**lf-**C**onsistent pseudo label **R**efin**E**men**T** method, termed as SECRET, is proposed to improve consistency by mutually refining the pseudo labels generated from different feature spaces. The proposed SECRET gradually encourages the improvement of pseudo labels' quality during training process, which further leads to better cross-domain Re-ID performance. Extensive experiments on benchmark datasets show the superiority of our method. Specifically, our method outperforms the state-of-the-arts by $\mathbf{6.3}\%$ in terms of mAP on the challenging dataset MSMT17. In the purely unsupervised setting, our method also surpasses existing works by a large margin. Code is available at https://github.com/LunarShen/SECRET.

1 Introduction

Person re-identification (Re-ID) is to match persons across non-overlapping cameras. Due to the laborious human labeling efforts in supervised person Re-ID methods (Luo et al. 2019; Wang et al. 2018; Sun et al. 2018), unsupervised domain adaptive (UDA) person Re-ID has become an active research field in recent years. UDA Re-ID aims at learning on an unlabeled target domain with only labeled data in source domain. Currently, there are roughly two ways to tackle the problem: (1) generative-model-based methods (Wei et al. 2018; Deng et al. 2018), in which generative models like GAN are used to translate the source domain data to the target domain together with their corresponding labels, so that supervised methods can be performed with the generated data. (2) pseudo-label-based methods (Fu et al. 2019; Ge, Chen, and Li 2020), which firstly pre-trains a model on source domain data by supervised methods and then alternates between generating pseudo labels by clustering and fine-tuning with pseudo labels in the target domain. Benefiting from exploring relations between samples in the target domain, pseudo-label-based ones (Fu et al. 2019; Ge, Chen, and Li 2020) have achieved better performance and are attracting more attention.

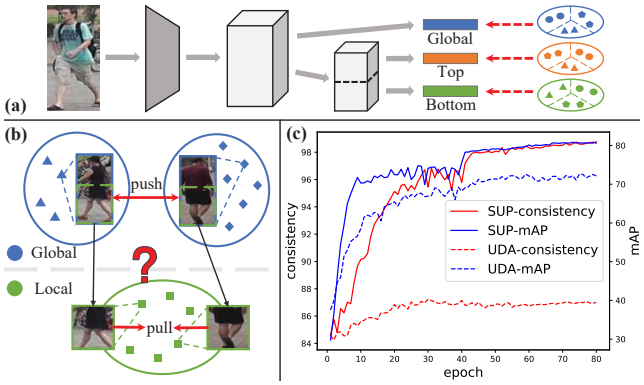


Figure 1: **(a)** A multi-branch CNN model. **(b)** Clustered results of global and local space. Shapes denote pseudo labels. **(c)** The consistency and mAP curves of a multi-branch model on Market-1501 during training epochs. *SUP*: supervised setting, where the three features are all supervised by the ground truth ID. *UDA*: unsupervised domain adaptive setting, where the three features are independently supervised by their respective pseudo labels.

Despite the fact that the pseudo-label-based methods have obtained promising performance, the key issue: the quality of the pseudo labels, is still unexplored. If the generated pseudo labels exactly match the ground truths, then the performance of UDA Re-ID methods will reach the supervised counterparts. Therefore, improvements on the quality of pseudo labels will potentially lead to a great performance gain. In this work, we move along this line to directly optimize pseudo labels' quality during the training process. Multi-branch is a popular pseudo-label-based method (Fu et al. 2019) which can explore global and local feature spaces simultaneously (as shown in Figure 1(a)). We argue that the consistency of different feature spaces is a key to improving performance. By consistency, we mean that different feature spaces should induce the same label

[*]equal contribution
[†]corresponding author

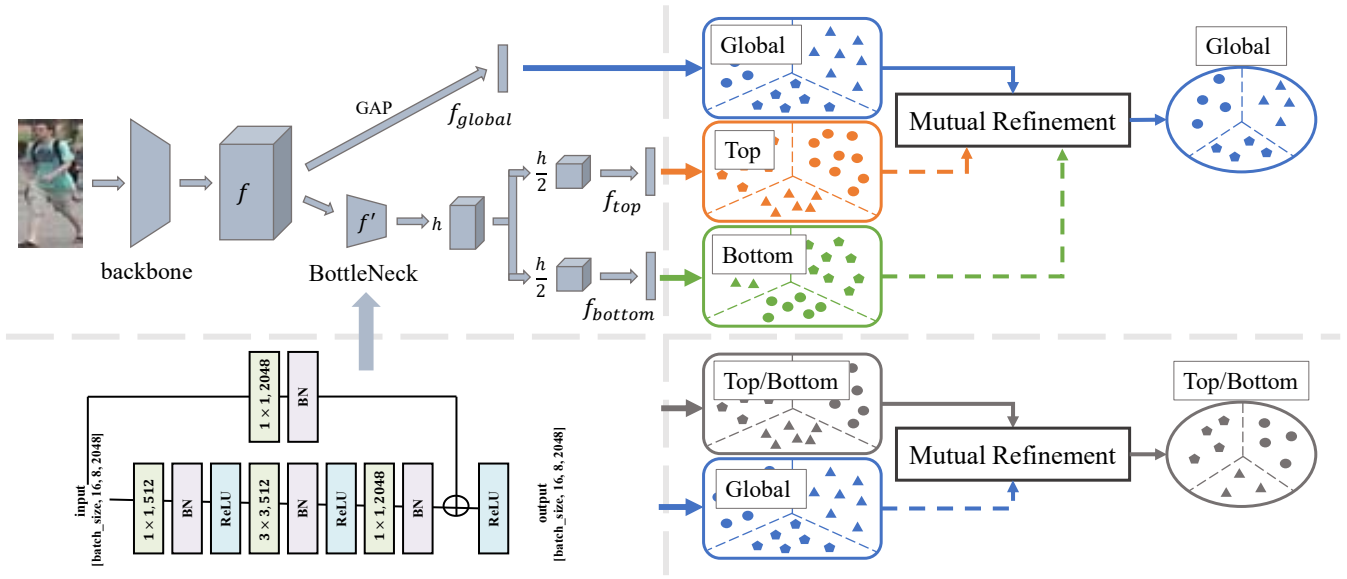


Figure 2: The overall framework of the proposed self-consistent pseudo label refinement (SECRET) method. It is composed of a backbone network, a global and local feature extraction module and a mutual refinement strategy of pseudo labels. The global feature space is refined by the two local feature spaces. Each of the two local feature spaces is only refined by the global feature space. The refined pseudo labels for each feature space are used as supervision signals to fine-tune the network.

space. In the supervised setting, the consistency is kept by the same supervision signals (such as person IDs) (Sun et al. 2018; Wang et al. 2018). However, in the unsupervised setting, there is no ground truth label to supervise each feature space. So simply ignoring the consistency and supervising each feature space by its own pseudo labels will lead to the limited performance.

To demonstrate the inconsistency problem, Figure 1(b) shows the clustering results of features from global and local branches. In the global space, the two persons (denoted by blue triangles and diamonds) can be easily distinguished. The training signals for global branch is "*push*". But in the (bottom) local space, they are very similar, thus clustered into the same group. The training signals for local branch is "*pull*". In this way, it is confusing to train the model with the inconsistent signals. Figure 1(c) also shows the consistency and mAP of a multi-branch model during the training epochs in the supervised and pseudo-label-based UDA settings. The consistency is measured by the agreement of label spaces induced by the global and two local feature spaces (details will be discussed in the experiment section). For the supervised setting, as the consistency becomes better, the performances measured by mAP is also improved. But for the UDA setting, the consistency is only slightly improved during the training epochs, so the performance is relatively lower than the supervised counterpart.

Motivated by the importance of consistency, we propose to improve the quality of pseudo labels by keeping consistency of different feature spaces in the UDA person Re-ID task. However, it is nontrivial to achieve this goal. Because there is no ground truth IDs in target domain for UDA task, it is infeasible to apply the same strategy as supervised methods. Moreover, global features usually represent a holistic view of a person, while local features pay more attention to specific parts or details. If there is no constraint on the consistency of these feature spaces, clustered results obtained from different feature spaces will easily disagree with each other, leading to the poor performance. Therefore, in order to keep consistency, we propose to mutually refine the pseudo labels generated by different feature spaces. Due to the fact that different feature spaces characterize the input instance from different aspects, in each feature space, we only remain the instances together with their pseudo labels that are in agreement with other feature spaces. In this way, the supervision signals for different feature spaces will gradually be consistent, leading to consistent feature spaces.

In summary, our contributions are as follows: (1) We are the first to reveal that the consistency of different feature spaces is a key to unsupervised domain adaptive person Re-ID. By keeping the consistency, the quality of pseudo labels will be improved. (2) We adopt a multi-branch network and design a self-consistent pseudo label refinement method to gradually improve the consistency of global and local feature spaces. (3) The overall method is evaluated on benchmark datasets, Market-1501, DukeMTMC-reID and MSMT17. Experimental results validate the consistency assumption and show significant improvements over the state-of-the-arts. In the more challenging unsupervised setting, our method also surpasses existing works by a large margin.

2 Related Works

Currently, unsupervised domain adaptive (UDA) person Re-ID can be roughly categorized into two classes: the GAN-based translation method and the pseudo-label-based fine-

tuning method. GAN-based methods (Wei et al. 2018; Deng et al. 2018; Chen, Zhu, and Gong 2019; Huang et al. 2019) first translate the labeled source domain data to the target domain, and then apply supervised methods in the target domain with translated labeled data. But the quality of the translated data cannot be well controlled, thus the performances are still very low. Besides, the computational requirement is also high.

The second approach first pre-trains a model on source domain data by supervised methods and then alternates between clustering and fine-tuning in the target domain. This approach shows promising results than the GAN-based approach in recent works (Fan et al. 2018; Fu et al. 2019; Zhang et al. 2019; Zhong et al. 2019; Ge, Chen, and Li 2020; Zhong et al. 2020b; Wang and Zhang 2020). PUL (Fan et al. 2018) first introduced the clustering and fine-tuning pipeline in Re-ID. To obtain more reliable pseudo labels, SSG (Fu et al. 2019) enhanced similarity measurement by human part features. MMT (Ge, Chen, and Li 2020) adopted mean teacher (Tarvainen and Valpola 2017) and mutual learning (Zhang et al. 2018). SpCL (Ge et al. 2020) used the intersection of stringent and lax parameters of clustering algorithm. It should be noted that these methods generally focus on the *individual* features, no matter on global or part level, and neglect the mutual information between them. Our method aims to improve consistency between *different* feature spaces by mutual learning from each other, which potentially lead to better performance.

3 Method

3.1 Overview

Figure 2 shows an overview of the proposed self-consistent pseudo label refinement (SECRET) method. In order to get different feature spaces, and gradually obtain self-consistent pseudo labels from multiple feature spaces, specifically: (1) we adopt a multi-branch network architecture to simultaneously obtain global and local features for an input person image; (2) three types of features are independently clustered by DBSCAN (Ester et al. 1996) algorithm; (3) clustered results are filtered by others, leading to more consistent results; (4) the three groups of pseudo labels are simultaneously used as the supervision signals for each branch to fine-tune the network.

3.2 Network Architecture

We adopt ResNet (He et al. 2016) as backbone. For a given image I, the feature map obtained from backbone is f, after a global average pooling (GAP), the global feature f_{global} will be a 2048 dimensional vector. As for local features, we first add a lightweight bottleneck on top of the feature map f to produce $f^{'}$, and then horizontally split $f^{'}$ into two parts. After global average pooling, the resulting features f_{top} and f_{bottom} are both 2048 dimensional vectors. The bottleneck for local features is similar to the building block in ResNet. The structure and detailed parameters are shown in the bottom left of Figure 2. Therefore, for an input image I, the outputs are global feature f_{global}, top local feature f_{top} and bottom local feature f_{bottom}.

Figure 3: An example illustration of self-consistent pseudo label refinement. *left*: local feature is refined by global feature. *right* : global feature is refined by local feature.

The network architecture is used in both source domain pre-training and target domain fine-tuning. At inference time, by default, only the global features are used (SECRET), so there is no additional cost compared with plain ResNet. The bottleneck in local branch only brings cost at training time. If additional cost can be afforded, the combination of global and local features can even improve the performance (SECRET-Joint).

3.3 Mutual Refinement of Pseudo Labels

The most important part of the proposed SECRET is the mutual refinement of pseudo labels. It is executed after independent clustering on global features and two local features in each training epoch. Then the refined pseudo labels for each branch will be used to fine-tune the whole network.

Algorithm 1: Mutual refinement of pseudo labels

Input:
$\mathcal{G}$: data set induced from global feature space
$\mathcal{L}_t$: data set induced from local top feature space
$\mathcal{L}_b$: data set induced from local bottom feature space
K: hyper-parameters to control the strictness
Output: optimized data set $\mathcal{G}^{'}$, $\mathcal{L}_t^{'}$ and $\mathcal{L}_b^{'}$

/* Filter global set $\mathcal{G}$ with local top set $\mathcal{L}_t$ */
$\mathcal{G}_t$ = call Algorithm 2 with arguments ($\mathcal{G}$, $\mathcal{L}_t$, K)
/* Filter global set $\mathcal{G}$ with local bottom set $\mathcal{L}_b$ */
$\mathcal{G}_b$ = call Algorithm 2 with arguments ($\mathcal{G}$, $\mathcal{L}_b$, K)
/* Intersect the filtered results $\mathcal{G}_1$ and $\mathcal{G}_2$ */
$\mathcal{G}^{'} = \{(x, y) \mid \forall\, (x, y) \in \mathcal{G}_t \; and \; (x, y) \in \mathcal{G}_b\}$
/* Filter local top set $\mathcal{L}_t$ with global set $\mathcal{G}$ */
$\mathcal{L}_t^{'}$ = call Algorithm 2 with arguments ($\mathcal{L}_t$, $\mathcal{G}$, K)
/* Filter local top set $\mathcal{L}_b$ with global set $\mathcal{G}$ */
$\mathcal{L}_b^{'}$ = call Algorithm 2 with arguments ($\mathcal{L}_b$, $\mathcal{G}$, K)

Figure 3 shows the motivation of refining pseudo labels by different feature spaces. If only the global features are considered, it is very likely to cluster the right two images into a group, as they are very similar from a holistic view. Using the erroneous clustered results as supervision signals to train the network will lead to poor performance. But if we also involve some local features, such as the specialized features for the top part of a person in the right of Figure 3, the difference in details will be emphasized, so that it will be easily distinguished between these two persons using local features. In this way, local features can be used to refine the clustered results of global features. Similarly, global features can also be used to refine the results of local features. As shown in the left of Figure 3, only clustering on local features of the bottom part cannot easily distinguish between persons wearing skirt and shorts, but with the help of global features, their differences are amplified.

Algorithm 2: Noisy instance elimination

Input:
$\mathcal{T}$: target set to be refined
$\mathcal{R}$: reference set used to refine target set
K: hyper-parameters to control the strictness
Output: optimized data set $\mathcal{T}^{'}$

$\mathcal{T}^{'} = \emptyset$
for *each pseudo label* l *in* $\mathcal{T}$ **do**
  /* Get all instances in $\mathcal{T}$ with pseudo label l */
  $\mathcal{T}_l = \{(x, y) \mid \forall\, (x, y) \in \mathcal{T}$ *and* $y = l\}$
  /* For each instance in $\mathcal{T}_l$, get the corresponding pseudo label in $\mathcal{R}$ */
  $\mathcal{P}^l = \{(x, y) \mid \forall\, (x, y) \in \mathcal{R}$ *and* $x \in \mathcal{T}_l\}$
  for *each pseudo label* m *in* $\mathcal{P}^l$ **do**
    $\mathcal{P}^l_m = \{(x, y) \mid \forall\, (x, y) \in \mathcal{P}^l$ *and* $y = m\}$
    /* Only remain the dominating instances, which is controlled by K */
    $\mathcal{T}^m_l = \{(x, y) \mid \forall\, (x, y) \in \mathcal{T}_l$ *and* $x \in \mathcal{P}^l_m\}$
    if $\frac{|\mathcal{P}^l_m|}{|\mathcal{T}_l|} > K$ **then**
      $\mathcal{T}^{'} = \mathcal{T}^{'} \cup \mathcal{T}^m_l$

The core idea of the mutual refinement procedure is to only remain the instances together with their pseudo labels that are in agreement with other feature spaces. Algorithm 1 is the pseudo-code for the mutual refinement procedure. For global feature space, we first refine its pseudo labels by local top and local bottom features individually, and then use the intersection of these two as the results. For each local feature space, only the global feature refines it. Algorithm 2 is the pseudo-code for eliminating the noisy instances by calculating the distribution of target feature space on the reference feature space. The hyper-parameter K controls the degree of agreement between two feature spaces. If K is large, only the instances with high agreement will be remained. Otherwise, instances with low agreement will also be kept. We also prove that the mutual refinement algorithm guarantees the improvement of consistency between different feature spaces. Details are in the supplementary materials. Compared with model training, the cost of mutual refinement is very low. In our experiment, it only takes about 1s in a whole training epoch, which takes about 200s.

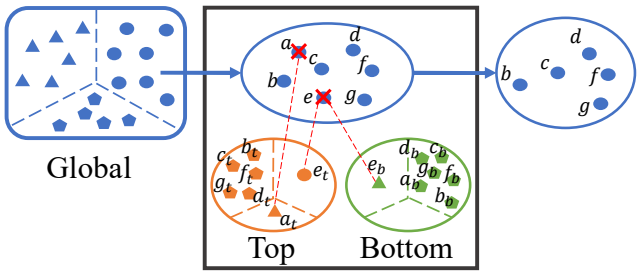


Figure 4: A toy example of the mutual refinement procedure for the global feature space.

Figure 4 shows a toy example of the refinement process of global space using two local spaces. In the clustered results of global features, for a given pseudo label (circle in global feature space) and its corresponding instances a to g, from the viewpoint of local features of these instances, the pseudo labels are a_t to g_t (for top feature) and a_b to g_b (for bottom feature). The minorities in the top feature space (a_t and e_t) and bottom feature space (e_b) will be eliminated from the original global feature space. The refined global feature space of the given pseudo label now contains only five instances (b, c, d, f, g).

3.4 Loss Function

Source Domain For N^s labeled instances in source domain, each is associated with a ground truth label. For each of the feature in $(f_g, f_{top}, f_{bottom})$, we simultaneously apply both cross-entropy loss and triplet loss.

Target Domain For N^t unlabeled instances in target domain, the feature set for all instances is as follows:

$$\mathcal{F} = \begin{cases} \mathcal{F}_g = \left\{f_g^1, ..., f_g^{N^t}\right\} \\ \mathcal{F}_{top} = \left\{f_{top}^1, ..., f_{top}^{N^t}\right\} \\ \mathcal{F}_{bottom} = \left\{f_{bottom}^1, ..., f_{bottom}^{N^t}\right\} \end{cases} \tag{1}$$

At T-th epoch, after first running DBSCAN (Ester et al. 1996) algorithm independently on $\mathcal{F}_g$, $\mathcal{F}_{top}$ and $\mathcal{F}_{bottom}$, and then conducting mutual refinement of pseudo labels as Section 3.3, instances with their improved pseudo labels in the target domain will be as follows:

$$X^t = \left\{x_i^t : (y_{global}^i, y_{top}^i, y_{bottom}^i; 1 \le i \le N_T^t)\right\} \tag{2}$$

Note that $N_T^t < N^t$, as noisy instances that found by clustering algorithm and the refinement procedure, are discarded from the fine-tuning data set. But we observe that in the later epochs, very few instances are eliminated due to the

Table 1: Experimental results of state-of-the-arts UDA methods and the proposed SECRET.

	Duke-to-Market		Market-to-Duke		Market-to-MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SPGAN (Deng et al. 2018)	22.8	51.5	22.3	41.1	–	–
HHL (Zhong et al. 2018)	31.4	62.2	27.2	46.9	–	–
ECN (Zhong et al. 2019)	43.0	75.1	40.4	63.3	–	–
PDA-Net (Li et al. 2019)	47.6	75.2	45.1	63.2	–	–
CR-GAN (Chen, Zhu, and Gong 2019)	54.0	77.7	48.6	68.9	–	–
PCB-PAST (Zhang et al. 2019)	54.6	78.4	54.3	72.4	–	–
SSG (Deng et al. 2018)	58.3	80.0	53.4	73.0	13.2	31.6
MMCL (Wang and Zhang 2020)	60.4	84.4	51.4	72.4	15.1	40.8
SNR (Jin et al. 2020)	61.7	82.8	58.1	76.3	–	–
ECN++ (Zhong et al. 2020b)	63.8	84.1	54.4	74.0	15.2	40.4
AD-Cluster (Zhai et al. 2020)	68.3	86.7	54.1	72.6	–	–
HGA (Zhang et al. 2021)	70.3	89.5	67.1	80.4	25.5	55.1
MMT (Ge, Chen, and Li 2020)	71.2	87.7	65.1	78.0	22.9	49.2
SpCL (Ge et al. 2020)	76.7	90.3	68.8	**82.9**	25.4	51.6
UNRN (Zheng et al. 2021)	78.1	91.9	<u>69.1</u>	<u>82.0</u>	25.3	52.4
SECRET	79.8	92.3	67.1	80.3	24.3	49.9
SECRET-Joint	79.9	92.3	68.2	81.5	25.4	51.2
SECRET(MT)	<u>82.9</u>	<u>93.1</u>	68.8	81.7	<u>31.2</u>	<u>59.7</u>
SECRET-Joint(MT)	**83.0**	**93.3**	**69.2**	<u>82.0</u>	**31.7**	**60.0**

high consistency between global and local feature spaces (as shown in Figure 5). Similar to the loss functions in source domain, we also apply cross-entropy loss and triplet loss in target domain for each of f_g, f_{top} and f_{bottom}.

4 Experiments

4.1 Evaluation Setting and Metrics

The proposed SECRET is evaluated on the popular benchmark datasets: Market-1501 (Zheng et al. 2015), DukeMTMC-reID (Ristani et al. 2016) and MSMT17 (Wei et al. 2018). In the setting of unsupervised domain adaptive person Re-ID, we first pre-train the model in the source domain with annotated data, and then alternates between clustering and pseudo label fine-tuning in the target domain without annotation. Following the common setting, three adaptation tasks are set up: Market-to-Duke, Duke-to-Market and Market-to-MSMT. Mean average precision (mAP) and rank-1 accuracy are adopted to evaluate the performance of the proposed SECRET. No post-processing technique is adopted, such as re-ranking (Zhong et al. 2017) or average query expansion (Chum et al. 2007).

4.2 Implementation Details

We use ResNet-50 as our backbone. The input images are resized to 256×128. Random flip, padding, and random crop are used as data augmentation in both source domain pre-training and target domain fine-tuning. Random erase (Zhong et al. 2020a) is only used in target domain fine-tuning. We randomly sample 4 instances per ground truth (in pre-training) or pseudo label (in fine-tuning) in a mini-batch, resulting in batch size 64. In pre-training, the initial learning rate is set to 3.5×10^{-4}, and decays by 0.1 at 40 and 70 epoch, and 80 epochs in total. In fine-tuning, clustering-and-pseudo-label-fine-tuning runs 80 epochs in total. In each epoch, the model is trained for 400 iterations on all datasets after clustering. The learning rate is set to 3.5×10^{-4}. The hyper-parameters K in filtering pseudo labels of global and local features is set to be 40%. We also analyze the sensitivity of K in Section 4.4. By default, the evaluation results are reported on the global feature only.

4.3 Comparisons with State-of-the-arts

We compare our proposed SECRET with the recent advances in UDA person Re-ID. The results are shown in Table 1. In order to make a fair comparison, We also adopt mean teacher (Tarvainen and Valpola 2017) to stabilize the training process, which is denoted as SECRET(MT) and SECRET-Joint(MT). The state-of-the-arts usually implement the mean teacher by moving average of model weights (Ge, Chen, and Li 2020), or memory bank (Ge et al. 2020). The proposed SECRET and SECRET-Joint show competitive performance with recent baselines. When mean teacher is adopted, SECRET(MT) and SECRET-Joint(MT) outperform all baselines by a large margin on Duke-to-Market and Market-to-MSMT17, and slightly improved on Market-to-Duke. Specifically, SECRET-Joint(MT) achieves an improvement of 6.3% mAP over the best baseline on the challenging setting Market-to-MSMT17.

Table 2 shows the results of a more challenging unsupervised setting, where there is no labeled source domain and the network is initialized by ImageNet (Deng et al. 2009) pre-trained model. Both SECRET and SECRET-Joint (with their MT counterparts) show significant improvement over state-of-the-arts on Market and MSMT17. On Duke, they are slightly lower than SpCL, but also surpass other base-

lines by a large margin.

4.4 Ablation Studies

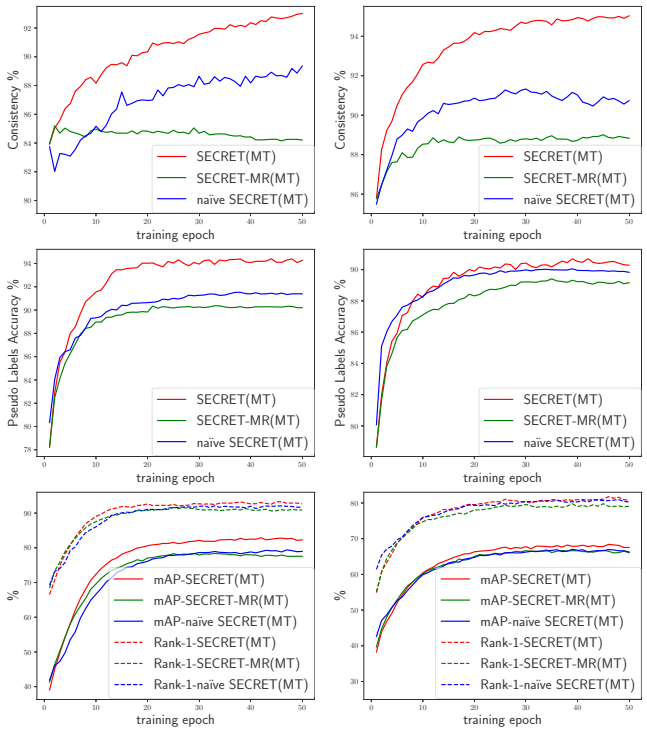
In this section, we evaluate each component of the proposed SECRET. Compared baselines are as follows:

- **Baseline**: ResNet-50; Clustering and fine-tuning on the global feature.
- **SECRET *w/o* Mutual Refinement — SECRET-MR**: ResNet-50 with the proposed two local branches; clustering on each feature individually; fine-tuning with its own pseudo labels of each branch.
- **naïve SECRET**: A naïve way to keep consistent; clustering and fine-tuning work on the concatenation of global and two local features; others are same as Baseline.
- **SECRET *w/o* Top — SECRET-T**: mutual refinement only works for global and bottom branch, others are same as SECRET.
- **SECRET *w/o* Bottom — SECRET-B**: mutual refinement only works for global and top branch, others are same as SECRET.
- **SECRET**: ResNet-50 with the proposed two local branches; clustering on each feature individually; mutual refinement works for all the global and two local branches. It is the full version of our proposed method.

Effectiveness of the mutual refinement procedure As shown in Table 3: (1) Compared with SECRET-MR, the performance of naïve SECRET is slightly lower. It indicates that the naïve approach of forcing the label space of global and local to be exactly the same may be harmful. (2) The performance of SECRET and its two variants, SECRET-T and SECERT-B, are better than SECRET-MR and baseline. It validates effectiveness of the proposed mutual refinement method. (3) As top and bottom feature characterize different aspects of the inputs, mutual refinement with global and only one local feature space (SECRET-T and SECRET-B) results in lower performance than the full SECRET.

Consistency Analysis of SECRET In order to further verify the assumption of consistency made in Section 1 and analyze the reason behind the good performance of SECRET, we also design two metrics: accuracy and consistency of pseudo labels. The accuracy of pseudo labels induced by a feature space is obtained by setting the label of a given cluster by its dominating ground truth label. Then instances in the cluster with that label are clean, and others are noisy. Then the accuracy of the cluster is $\frac{\#\ clean\ instances}{\#\ all\ instances}$. The overall accuracy is the mean accuracy of all clusters. The definition of consistency of different feature spaces is based on the accuracy. For two pseudo label sets $\mathcal{P}$ and $\mathcal{Q}$ induced from two different feature spaces, if we regard any one label set as the ground truth, and calculate the accuracy of the other set against it, the consistency between $\mathcal{P}$ and $\mathcal{Q}$ can be defined as the mean accuracy of $\mathcal{Q}$ against $\mathcal{P}$ and $\mathcal{P}$ against $\mathcal{Q}$. Then the overall consistency of all feature spaces is the mean of all pairs of feature spaces.

Figure 5 shows these two metrics together with mAP and Rank-1 during the training epochs. All methods here adopt the mean teacher strategy, so for simplicity we omit



(a) Duke-to-Market (b) Market-to-Duke

Figure 5: Consistency Analysis of SECRET: the consistency between different feature spaces, the accuracy of pseudo labels, and the mAP and Rank-1 during training epochs.

MT. SECRET-MR is with no consistency constraint. Naïve SECRET is a naïve way to keep consistent, where clustering and fine-tuning work on the concatenation of all feature spaces. With mutual refinement of pseudo labels, the consistency of SECRET is much higher than SECRET-MR and naïve SECRET. The high consistency leads to the high-quality pseudo labels, and eventually obtains high performance. We also observe that the MT version of naïve SECRET performs slightly better than SECRET-MR, which could be ascribed to the more robust representation by mean teacher strategy. Nevertheless, there is still no obvious evidence that the naïve way to keep consistency is useful.

Feature Selection at Inference Time As the proposed model can simultaneously generate one global feature and two local features, there are multiple choices of features at inference time: only global feature (default setting), only top local feature, only bottom local feature and a combination of these three features. The combination can be implemented by a weighted sum of distance individually calculated from three different features. For simplicity, we use the same weight η for both local features:

$$d_{i,j} = d_{i,j}^{global} + \eta \cdot d_{i,j}^{top} + \eta \cdot d_{i,j}^{bottom} \quad (3)$$

Experimental results of each feature are shown in Table 4. Local feature alone (top or bottom) leads to much poor performance. This makes sense because local features are specialized in local details, and are not as discriminative as

Table 2: Experimental results of state-of-the-arts unsupervised Re-ID methods and the proposed SECRET.

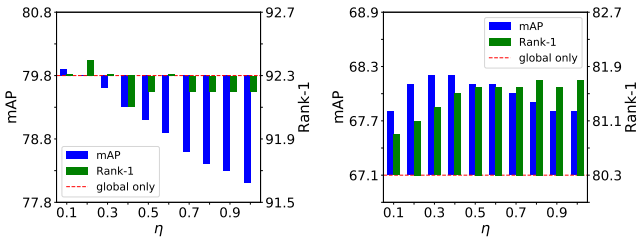
	Market		Duke		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SSL (Lin et al. 2020)	37.8	71.7	28.6	52.5	–	–
BUC (Lin et al. 2019)	38.3	66.2	27.5	47.4	–	–
MMCL (Wang and Zhang 2020)	45.5	80.3	40.2	65.2	11.2	35.4
HCT (Zeng et al. 2020)	56.4	80.0	50.7	69.6	–	–
SpCL (Ge et al. 2020)	72.6	87.7	**65.3**	**81.2**	19.1	42.3
SECRET	78.7	91.7	63.2	77.4	25.8	53.4
SECRET-Joint	79.3	92.2	64.1	78.6	26.4	53.7
SECRET(MT)	80.8	92.1	63.1	77.4	30.5	60.3
SECRET-Joint(MT)	**81.0**	**92.6**	63.9	77.9	**31.3**	**60.4**

Table 3: Evaluation results of the proposed mutual refinement methods on Market-1501 and DukeMTMC-reID.

	Duke-to-Market		Market-to-Duke	
	mAP	Rank-1	mAP	Rank-1
Baseline	67.3	85.1	56.5	73.2
SECRET-MR	72.4	88.2	61.0	76.0
naïve SECRET	71.5	88.5	58.7	74.1
SECRET-T	75.4	90.1	61.1	76.0
SECRET-B	74.8	89.9	61.6	76.3
SECRET	79.8	92.3	67.1	80.3
SECRET(MT)	**82.9**	**93.1**	**68.8**	**81.7**

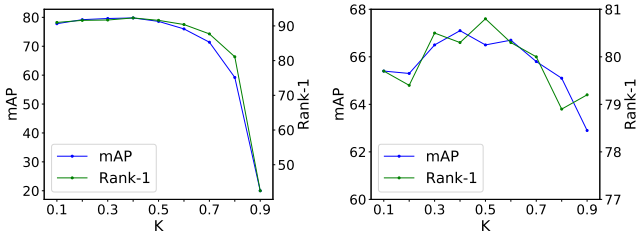
Table 4: Evaluation of different features at inference time on Market-1501 and DukeMTMC-reID.

	Duke-to-Market		Market-to-Duke	
	mAP	Rank-1	mAP	Rank-1
SECRET(-Global)	79.8	92.3	67.1	80.3
SECRET-Local-Top	65.6	86.6	58.5	76.5
SECRET-Local-Bottom	68.2	85.5	55.7	74.1
SECRET-Joint	**79.9**	**92.3**	**68.2**	**81.5**



(a) Duke-to-Market (b) Market-to-Duke

Figure 6: Sensitivity Analysis of Hyper-parameter η



(a) Duke-to-Market (b) Market-to-Duke

Figure 7: Sensitivity Analysis of Hyper-parameter K

global features. The global features alone show a much better performance.

Figure 6 shows the experimental results of different weight parameters. $\eta = 1.0$ means all features are equally important, while $\eta = 0.0$ means only using the global feature (denoted by the red line in the figure). Results of different η show small fluctuation (mAP from 78.1 to 79.9 for Duke-to-Market, from 67.1 to 68.2 for Market-to-Duke). For Duke-to-Market, the joint feature cannot bring significant improvements, and the performance of global feature alone has already been close to the best joint performance. For Market-to-Duke, compared with global only, 1.1% improvement in mAP can be obtained by the best joint feature.

Therefore, we can draw conclusions: (1) global feature is good enough and can also save computation cost. (2) if one can afford the additional cost, joint of global and local features can slightly improve the performance.

Sensitivity Analysis of Hyper-parameter K We also test the sensitivity of hyper-parameter K in the mutual refinement procedure. If K is extremely large, it is very easy to filter out most instances from a given pseudo label, which leads to insufficient training data. On the contrary, if K is small enough, very few instances will be eliminated from a given pseudo label, then the quality of pseudo labels cannot be effectively improved. Therefore, setting a moderate value for K will be a general choice. Figure 7 shows the experimental results of different hyper-parameters K. As K is set to be larger, the mAP results first improve and then drop. The best results are obtained at $K = 40\%$ for both Duke-to-Market and Market-to-Duke.

5 Conclusion and Future Work

In this work, we propose a self-consistent pseudo label refinement method for unsupervised domain adaptive person Re-ID. The key is to preserve consistency between global

and local features, so that the quality of pseudo labels will be improved, leading to a performance gain. Extensive experiments on benchmark datasets show that our method outperforms the state-of-the-arts by a large margin in most cases. In the future, more sophisticated solutions for improving the consistency can be exploited. For example, as the network architecture for local feature extraction is quite simple in this work, a more complex structure can be adopted in the local branch. In addition, other side information that is nearly free to obtain can also be used for improving consistency.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.U1936202, 61925107, No.61971260, U21B2013) and National Key R&D Program of China (No.2020AAA0105500). This work was also partially supported by Alibaba Innovative Research(AIR).

References

Chen, Y.; Zhu, X.; and Gong, S. 2019. Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chum, O.; Philbin, J.; Sivic, J.; Isard, M.; and Zisserman, A. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 994–1003.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, 226–231.

Fan, H.; Zheng, L.; Yan, C.; and Yang, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4): 1–18.

Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; and Huang, T. S. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 6112–6121.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *International Conference on Learning Representations*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; and Li, H. 2020. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID. In *Advances in Neural Information Processing Systems*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.; Wu, Q.; Xu, J.; and Zhong, Y. 2019. SBSGAN: Suppression of inter-domain background shift for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9527–9536.

Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; and Zhang, L. 2020. Style Normalization and Restitution for Generalizable Person Re-Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3140–3149.

Li, Y. J.; Lin, C. S.; Lin, Y. B.; and Wang, Y. C. F. 2019. Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8738–8745.

Lin, Y.; Xie, L.; Wu, Y.; Yan, C.; and Tian, Q. 2020. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3390–3399.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, 480–496.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, 1195–1204.

Wang, D.; and Zhang, S. 2020. Unsupervised Person Re-identification via Multi-label Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10981–10990.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 274–282.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 79–88.

Zeng, K.; Ning, M.; Wang, Y.; and Guo, Y. 2020. Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13657–13665.

Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; and Tian, Y. 2020. AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9018–9027.

Zhang, M.; Liu, K.; Li, Y.; Guo, S.; Duan, H.; Long, Y.; and Jin, Y. 2021. Unsupervised Domain Adaptation for Person Re-identification via Heterogeneous Graph Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3360–3368.

Zhang, X.; Cao, J.; Shen, C.; and You, M. 2019. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8222–8231.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4320–4328.

Zheng, K.; Lan, C.; Zeng, W.; Zhang, Z.; and Zha, Z.-J. 2021. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3538–3546.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1318–1327.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020a. Random Erasing Data Augmentation. In *AAAI*, 13001–13008.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018. Generalizing A Person Retrieval Model Hetero- and Homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 598–607.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2020b. Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.