Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning

Yu Wu[1,2], Yutian Lin[2], Xuanyi Dong[2], Yan Yan[2], Wanli Ouyang[3], Yi Yang[1,2*]
[1]SUSTech-UTS Joint Centre of CIS, Southern China University of Science and Technology
[2]University of Technology Sydney [3]The University of Sydney
yu.wu-3@student.uts.edu.au; yi.yang@uts.edu.au

Abstract

We focus on the one-shot learning for video-based person re-Identification (re-ID). Unlabeled tracklets for the person re-ID tasks can be easily obtained by pre-processing, such as pedestrian detection and tracking. In this paper, we propose an approach to exploiting unlabeled tracklets by gradually but steadily improving the discriminative capability of the Convolutional Neural Network (CNN) feature representation via stepwise learning. We first initialize a CNN model using one labeled tracklet for each identity. Then we update the CNN model by the following two steps iteratively: 1. sample a few candidates with most reliable pseudo labels from unlabeled tracklets; 2. update the CNN model according to the selected data. Instead of the static sampling strategy applied in existing works, we propose a progressive sampling method to increase the number of the selected pseudo-labeled candidates step by step. We systematically investigate the way how we should select pseudo-labeled tracklets into the training set to make the best use of them. Notably, the rank-1 accuracy of our method outperforms the state-of-the-art method by 21.46 points (absolute, i.e.*, 62.67% vs. 41.21%) on the MARS dataset, and 16.53 points on the DukeMTMC-VideoReID dataset*[1].

1. Introduction

Person re-identification (re-ID) aims at spotting the person-of-interest from different cameras. In recent years, person re-ID on the large-scale video data, such as surveillance videos, has attracted significant attention [10, 20, 28, 32, 35]. Most proposed approaches rely on the fully annotated data, *i.e*., the identity labels of all the tracklets from multiple cross-view cameras. However, it is impractical to annotate very large-scale surveillance videos due to the dramatically increasing cost. Therefore, semi-supervised methods [21, 34] are of particular interest. This work mainly focuses on the one-shot setting, in which only one tracklet is labeled for each identity.

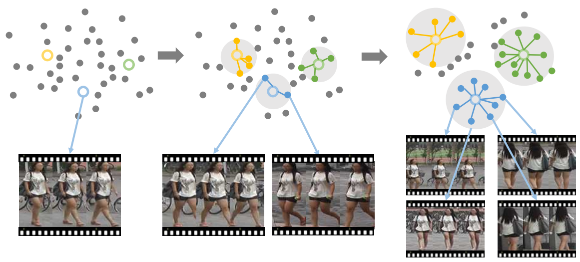

Figure 1. An illustration of the unlabeled data sampling procedure in the feature space. The hollow point and solid point denote the labeled tracklet and unlabeled tracklet, respectively. The pseudo label of each unlabeled tracklet is assigned by its nearest labeled neighbor (indicated by the colored line). Different colors represent different identities. Samples in the shade will be incorporated into training. We adopt the easy and reliable pseudo-labeled tracklets for updating at the beginning and difficult ones in subsequence.

The key challenge for the one-shot video-based person re-ID is the label estimation for the abundant unlabeled tracklets [7, 34]. A typical approach is to generate the pseudo labels for the unlabeled data at first. The initial labeled data and some selected pseudo-labeled data are considered as an enlarged training set. Lastly, this new training set is adopted to train the re-ID model.

Most existing methods employ a static strategy to determine the quantity of selected pseudo-labeled data for further training. For example, Fan *et al*. [7] and Ye *et al*. [34] compare the prediction confidences of pseudo-labeled samples with a *pre-defined* threshold. The samples with higher confidence over the fixed threshold are then selected for the subsequent training. During iterations, these algorithms se-

*Corresponding author.
[1]The code is publicly available at: https://yu-wu.net

lect a fixed and large number of pseudo-labeled data from beginning to end. However, it is inappropriate to keep the threshold fixed in the one-shot setting. In this case, the initial model may be not robust due to the very few training samples. Only a few of pseudo-label predictions are reliable and accurate at the initial stage. If one still selects the same number of data as that in the later stages, it will inevitably involve many unreliable predictions. Updating the model with excessive not-yet-reliable data would hinder the subsequent improvement of the model.

In this paper, to better exploit the unlabeled data in one-shot video-based person re-ID, we propose the stepwise learning method **EUG** (**E**xploit the **U**nknown **G**radually). Initially, a CNN model is trained on the one-shot labeled tracklet. EUG then iteratively updates the CNN by two steps, the label estimation step and the model update step. In the first step, EUG generates the pseudo labels for unlabeled tracklets, and selects some of pseudo-labeled tracklets for training according to the prediction reliability. The selected subset is continuously enlarged during iterations according to a sampling strategy. In the second step, EUG re-trains the CNN model on both the labeled data and the sampled pseudo-labeled subset. Particularly, as illustrated in Figure 1, EUG starts with a small-size subset of pseudo-labeled tracklets, which includes only the most reliable and easiest ones. In the subsequent stages, it gradually selects a growing number of pseudo-labeled tracklets to incorporate more difficult and diverse data. This is different from existing methods [21, 34].

To characterize the proposed progressive approach in one-shot person re-ID, we intensively investigate two significant aspects, *i.e.*, how the progressive sampling strategy benefits the label estimation and which sampling criterion is effective for the confidence estimation in person re-ID. For the first aspect, we find that if we enlarge the sampled subset of pseudo-labeled data in a more conservative way (at a slower speed), the model achieves a better performance. If we enlarge the subset in a more aggressive way (at a faster speed), the model achieves a worse performance. Note that the previous static sampling strategy can be viewed as an extremely aggressive manner. For the second aspect, we investigate the gap between the classification measures and retrieval evaluation metrics. We find that the sampling criteria highly affect the performance of the proposed method. Instead of the classification measures, a distance-based sampling criterion for the reliability estimation may yield promising performance in person re-ID.

Our contributions are summarized as follows:

- We propose a progressive method for one-shot video-based person re-ID to better exploit the unlabeled tracklets. This method adopts a dynamic sampling strategy to uncover the unlabeled data. We start with reliable samples and gradually include diverse ones, which significantly makes the model robust.
- We apply a distance-based sampling criterion for label estimation and candidates selection to remarkably improve the performance of label estimation.
- Our method achieves surprisingly superior performance on the one-shot setting, outperforming the state-of-the-art by 21.46 points (absolute) on MARS and 16.53 points (absolute) on DukeMTMC-VideoReID.

2. Related Works

Extensive works have been reported to address the video-based person re-ID problem. One simple solution is using image-based re-ID methods, and obtaining video representations by pooling the frame features [10, 16, 20].

Supervised Video-based Person Re-ID. Recently, a number of deep learning methods are developed [24, 29, 32, 35, 39, 40]. The typical architecture is to combine CNN and RNN to learn a video representation or the similarity score. In [40], temporal attention information and spatial recurrent information are used to explore contextual representation. Another commonly used architecture is the Siamese network architecture [18, 29, 32], which also achieve reasonably good performance.

Semi-Supervised Video-based Person Re-ID. Most works of semi-supervised person re-ID are based on image [1, 8, 19, 22]. The approaches of these works include dictionary learning, graph matching, metric learning, *etc*. To the best of our knowledge, there are three works aiming at solving the semi-supervised video-based re-ID task. Zhu *et al*. [41] proposed a semi-supervised cross-view projection-based dictionary learning (SCPDL) approach. A limitation is that this approach is only suitable for datasets that only captured by two cameras.

There are two recent works designed for one-shot video re-ID task [21, 34]. Although [21, 34] claim them as *unsupervised* methods, they are *one-shot* methods in experiments, as they require at least one labeled tracklet for each identity. They assume that the tracklets are obtained by tracking, and this process is automatic and unsupervised. Different tracklets from one camera with a long-time interval are assumed representing different identities. However, to conduct experiments in existing datasets, both methods require the annotation of at least a sample for each identity. To be more rigorous, we take this problem as a one-shot task. Ye *et al*. [34] propose a dynamic graph matching (DGM) method, which iteratively updates the image graph and the label estimation to learn a better feature space with intermediate estimated labels. Liu *et al*. [21] update the classifier with K-reciprocal Nearest Neighbors (KNN) in the gallery set, and refine the nearest neighbors by apply negative sample mining with KNN in the query set. While graph-based semi-supervised learning [33] could possibly be adopted for one-shot person Re-ID, it is time-consuming to solve a linear system for each query.

Progressive Paradigm. Curriculum Learning (CL) is proposed in [2], which progressively obtains knowledge from easy to hard samples in a pre-defined scheme. Kumar *et al*. [15] propose Self-Paced Learning (SPL) which takes curriculum learning as a regularization term to update the model automatically. The self-paced paradigm is theoretically analyzed in [13, 23]. Some works manage to apply the progressive paradigm in the computer vision area [5, 6, 27]. We are inspired by these progressive algorithms. Compared with the existing SPL and CL algorithms, we incorporated the retrieval measures (the distance in feature space) into the learning mechanism, which well fits the evaluation metric for person re-ID. Moreover, most previous SPL and CL works mainly focus on the supervised and semi-supervised task. Few are used in the one-shot learning setting.

3. The Progressive Model

3.1. Preliminaries

We first introduce the necessary notations. Let $\mathcal{L} = \{(x_1, y_1), ..., (x_{n_l}, y_{n_l})\}$ be the labeled dataset, and $\mathcal{U} = \{(x_{n_l+1}), ..., (x_{n_l+n_u})\}$ be the unlabeled dataset, where x_i and y_i denotes the i-th tracklet data and its identity label, respectively. We thus have $|\mathcal{L}| = n_l$ and $|\mathcal{U}| = n_u$ where $|\cdot|$ is the cardinality of a set. Following recent works [7, 17, 38], we take the training process as an identity classification task. For training on the labeled dataset, we have the following objective function:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \sum_{i=1}^{n_l} \ell(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{1}$$

where ϕ is an embedding function, parameterized by $\boldsymbol{\theta}$, to extract the feature from the data x_i. CNN models [3, 4, 9, 11, 30, 31] are usually used as the function ϕ. f is a function, parameterized by $\boldsymbol{w}$, to classifier the embedded feature $\phi(\boldsymbol{\theta}; x_i)$ into a k-dimension confidence estimation, in which k is the number of identities. ℓ denotes the suffered loss on the label prediction $f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i))$ and its ground truth identity label y_i.

To exploit abundant unlabeled tracklets with pseudo labels, we consider the following objective function in the one-shot re-ID problem:

$$\begin{aligned} \min_{\boldsymbol{\theta},\boldsymbol{w},s_i,\hat{y}_i} & \sum_{i=1}^{n_l} \ell(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i), y_i)) + \\ & \sum_{i=n_l+1}^{n_l+n_u} s_i \ell(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i), \hat{y}_i)), \end{aligned} \tag{2}$$

where $\hat{y}_i$ denotes the machine generated pseudo labels for the i-th unlabeled data. $s_i \in \{0, 1\}$ is the selection indicator for the unlabeled sample x_i, which determine whether the suffered loss of pseudo-labeled data $(x_i, \hat{y}_i)$ is adopted in optimizing. We use $\boldsymbol{s}$ to indicate the vertical concatenation of all s_i.

In the evaluation stage, for both of query data and gallery data, we only use $\phi(\boldsymbol{\theta}; \cdot)$ to embed each tracklet into the feature space. The query result is the ranking list of all gallery data according to the Euclidean Distance between the query data and each gallery data, *i.e.*, $||\phi(\boldsymbol{\theta}; x_q) - \phi(\boldsymbol{\theta}; x_g)||_2$, where x_q and x_g denote the query tracklet and the gallery tracklet, respectively.

3.2. Framework Overview

In this work, we propose a stepwise learning method to exploit the unlabeled data gradually and steadily. We adopt an alternative algorithm to solve the Eq. (2). Specifically, we first optimize $\boldsymbol{\theta}$ and $\boldsymbol{w}$, and then optimize $\hat{y}$ and $\boldsymbol{s}$, *i.e.*, the model updating and the label estimating.

Let $\mathcal{S}$ denote the set of selected pseudo-labeled candidates. We can obtain $\mathcal{S}$ by:

$$\mathcal{S} = \{(x_i, \hat{y}_i) | s_i = 1, n_l + 1 \leq i \leq n_l + n_u\}. \tag{3}$$

Our approach first trains an initial model on the labeled data $\mathcal{L}$, and then the initial model is applied to predict pseudo labels $\hat{y}$ on the unlabeled data. In subsequence, according to a label reliability evaluation criterion, we generate the selection indicators $\boldsymbol{s}$ in order to obtain the candidates set $\mathcal{S}$ via Eq. (3). In the model update step, the set $\mathcal{S}$ along with the initial labeled set $\mathcal{L}$ is regarded as the new training set $\mathcal{D}$, *i.e.*, $\mathcal{D} = \mathcal{L} \cup \mathcal{S}$. The set $\mathcal{D}$ will be utilized to re-train the model so as to make the model more robust. During training iterations, the candidates set $\mathcal{S}$ in each step is enlarged continuously. In this way, we can progressively learn a more stable model.

To be specific, for our progressive strategy EUG, we adopt an end-to-end CNN model with temporal average pooling (**ETAP-Net**) as the feature embedding function ϕ. The ETAP-Net is an adaption of ResNet-50 architecture for video inputs, where we add a fully-connected layer and a temporal average pooling layer before the classification layer. As shown in Figure 2, for each tracklet, all frames are processed to obtain frame-level feature embedding. The frame features within a tracklet are then element-wise averaged as the tracklet feature representation by the temporal average pooling layer. In the label estimation step, for each unlabeled video tracklet, the pseudo label is assigned by the identity label of its nearest labeled neighbor in the tracklet feature space. The distance between them is considered as the dissimilarity cost, which is used to measure the reliability of its pseudo label.

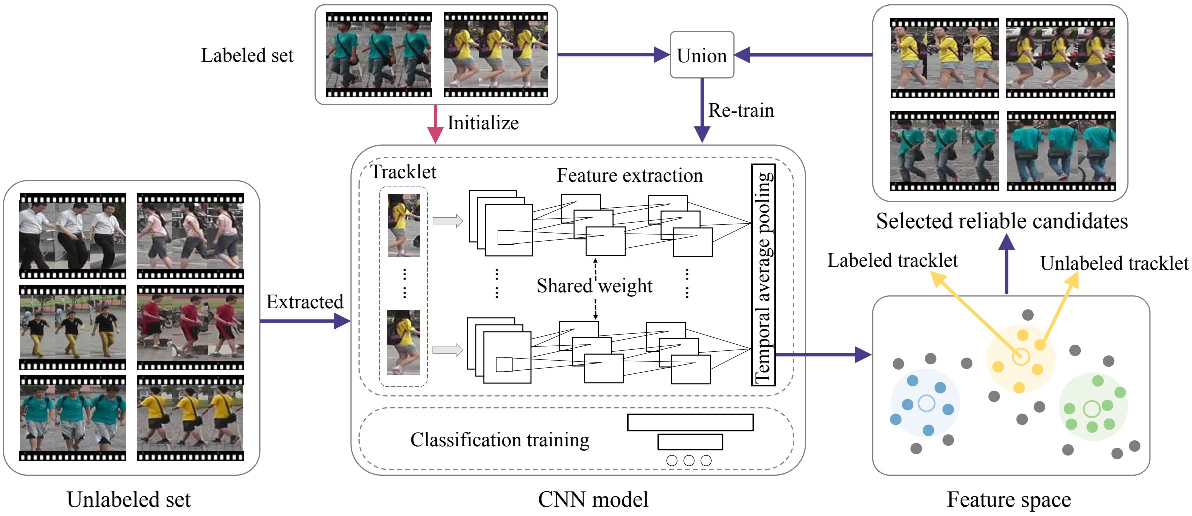

Figure 2. Overview of the framework. Different colors represent different identity samples. The CNN model is initially trained on the labeled one-shot data. For each iteration, we (1) select the unlabeled samples with reliable pseudo labels according to the distance in feature space and (2) update the CNN model by the labeled data and the selected candidates. We gradually enlarge the candidates set to incorporating more difficult and diverse tracklets. For a tracklet, each frame feature is first extracted by the CNN model and then temporally averaged as the tracklet feature. We take the training process as an identity classification task, and regard the evaluation as a retrieval problem on the features of the test tracklets.

3.3. Progressive and Effective Sampling Strategy

It is crucial to obtain the appropriately selected candidates $\mathcal{S}$ to exploit the unlabeled data. In this procedure, two significant aspects are mainly considered: First, how to ensure the reliability of selected pseudo-labeled samples? Second, what is an effective sampling criterion on the unlabeled data for one-shot person re-ID?

Discussion on Sampling Strategy. The reliability of pseudo labels originates from two main challenges in the one-shot learning setting. (1) the initial labeled data are too few to depict the detailed underlying distribution. (2) learning a CNN model on a not-yet-reliable training set may not improve the re-ID performance. The interplay of these two factors hinders the further performance improving. Therefore, it is irrational to incorporate excessive pseudo-labeled data into training at the initial iteration.

Discussion on Sampling Criterion. The previous works sample the unlabeled data from confident to uncertain ones according to the classification loss [5, 6, 27]. However, the loss from classification prediction does not well fit the retrieval evaluation. Moreover, it is far away to train a robust identity classifier in the one-shot setting, where each class has only one sample for training. The classifier may easily over-fit the one-shot labeled data and may not learn the intrinsic distinction in classification. Therefore, the classification prediction may be not reliable on an unseen sample.

Our Stepwise Solution. To address aforementioned two problems, we propose (1) a dynamic sampling scheme, which progressively increases the number of selected pseudo-labeled samples; (2) an effective sampling criterion, which takes the distance in the feature space as a measure of reliability.

The proposed dynamic sampling scheme steadily increases the size of selected candidates set $|\mathcal{S}|$ during iterations. It starts with a small proportion of pseudo-labeled data at the beginning stages, and then incorporates more diverse samples in the following stages. As the training iteration goes, the reliability of pseudo labels grows steadily, because the re-ID model becomes more robust and discriminative. Therefore, more pseudo-labeled candidates can be adopted into training.

For sampling criterion, instead of classification prediction, we adopt the Nearest Neighbors (NN) classifier for the label estimation. For the one-shot setting, the NN classifier in the feature space may be a better choice, since similar input data always have similar feature representations. The NN classifier assigned the label of each unlabeled data by its nearest labeled neighbor in feature space. We define the confidence of label estimation as the distance between the unlabeled data and its nearest labeled neighbor. For the candidates selection, we select some of top reliable pseudo-labeled data according to their label estimation confidence.

Algorithm 1 Exploit the Unknown Gradually

Input: Labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$, enlarging factor $p \in (0,1)$, initialized CNN model $\boldsymbol{\theta}_0$.
Output: The best CNN model $\boldsymbol{\theta}^*$.

Initialize the selected pseudo-labeled data $\mathcal{S}_0 \leftarrow \emptyset$, sampling size $m_1 \leftarrow p \cdot n_u$, iteration step $t \leftarrow 0$, best validation performance $V^* \leftarrow 0$
while $m_{t+1} \leq |\mathcal{U}|$ **do**
    $t \leftarrow t+1$
    Update training set: $\mathcal{D}_t \leftarrow \mathcal{L} \cup \mathcal{S}_{t-1}$
    Train the CNN model $(\boldsymbol{\theta}_t, \boldsymbol{w}_t)$ based on $\mathcal{D}_t$.
    Generate the selection indicators $\boldsymbol{s}_t$ via Eq. (5)
    Update $\mathcal{S}_t$ based on $\boldsymbol{s}_t$ via Eq. (3)
    Update the sampling number: $m_{t+1} \leftarrow m_t + p \cdot n_u$
end while
for i $\leftarrow$ 1 **to** T **do**
    Evaluate $\boldsymbol{\theta}_i$ on the validation set $\rightarrow$ performance V_i
    if $V_i > V^*$ **then**
        $V^*, \boldsymbol{\theta}^* \leftarrow V_i, \boldsymbol{\theta}_i$
    end if
end for

More formally, we define the **dissimilarity cost** for each unlabeled data $x_i \in \mathcal{U}$ as:

$$d(\boldsymbol{\theta}; x_i) = \min_{x_l \in \mathcal{L}} ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x_l)||_2, \tag{4}$$

The cost is the minimum l_2 distance between the unlabeled data x_i and an arbitrary labeled data $x_l \in \mathcal{L}$ in the feature space parameterized by $\boldsymbol{\theta}$. The dissimilarity cost is considered as the criterion for measuring the confidence of pseudo-labeled data. For the candidates selection, at the iteration step t, we sample the pseudo-labeled candidates into training by setting the selection indicator $\boldsymbol{s}_t$ as follows:

$$\boldsymbol{s}_t = \arg \min_{||s||_0 = m_t} \sum_{i=n_l+1}^{n_l+n_u} s_i d(\boldsymbol{\theta}; x_i), \tag{5}$$

where the m_t denotes the size of selected pseudo-labeled set. As the iteration step t increases, we enlarge the size of sampled pseudo-labeled data by set $m_t = m_{t-1} + p \cdot n_u$. $p \in (0,1)$ is the **enlarging factor** which indicates the speed of enlarging the candidates set during iterations. Eq. (5) selects the top m_t nearest unlabeled data for all the labeled data at the iteration step t. As described in Algorithm 1, we evaluate the model $\phi(\boldsymbol{\theta}_t; \cdot)$ on the validation set at each iteration step and output the best model. In the one-shot experiment, we take another video-based person re-ID training set as the validation set.

How to find a proper enlarging factor p? An *aggressive* choice is to set p to a very large value, which urges m_t to increase rapidly. As a result, the sampled pseudo-labeled candidates may not be reliable enough to train a robust CNN model. A *conservative* option is to set p to a very small value, which means m_t progressively enlarges with a small change in each step. This option tends to result in a very stable increase in the performance and a promising performance in the end. The disadvantage is that it may require an excessive number of stages to touch great performance.

4. Experiments

4.1. Datasets and Settings

The MARS dataset [36] is the largest video dataset for the person re-identification task captured in a university campus. The dataset contains 17,503 tracklets for 1,261 identities and 3,248 distractor tracklets, which are captured by six cameras. This dataset is split into 625 identities for training and 636 identities for testing. Every identity in the training set has 13 video tracklets on average and 816 frames on average. The bounding boxes are detected and tracked using the Deformable Part Model (DPM) and GMMCP tracker.

The DukeMTMC dataset [26] is a large-scale dataset aiming for multi-camera tracking. This dataset was captured in outdoor scenes with noisy background and suffers from illumination, pose, and viewpoint change and occlusions. To conduct our experiment, here we use a subset of DukeMTMC as the **DukeMTMC-VideoReID** [2] dataset specially for video-based re-ID. Since this dataset is manual annotated, each identity only has one tracklet under a camera. We crop pedestrian images from the videos for 12 frames every second to generate a tracklet. The dataset is split following the protocol in [37], *i.e.*, 702 identities for training, 702 identities for testing, and 408 identities as the distractors. Totally, we generate 369,656 frames of 2,196 tracklets for training, and 445,764 frames of 2,636 tracklets for testing and distractors.

Evaluation Metrics. We use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) to evaluate the performance of each method. For each query, its average precision (AP) is computed from its precision-recall curve. The mAP is calculated as the mean value of average precisions across all queries. We report the Rank-1, Rank-5, Rank-20 scores to represent the CMC curve. These CMC scores reflect the retrieval precision, while the mAP reflects the recall.

Experiment Setting. For one-shot experiments, we use the same protocol as [21]. In both datasets, we randomly choose one tracklet in camera 1 for each identity as initialization. If there is no tracklet recorded by camera 1 for one identity, we randomly select one tracklet in the next camera to make sure each identity has one video tracklet for initialization. Note that as discussed in Section 2, [21, 34] are the same one-shot setting in experiments.

[2] DukeMTMC-VideoReID is available at `https://yu-wu.net`

Methods	MARS				DukeMTMC-VideoReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
Baseline (one-shot)	36.16	50.20	61.86	15.45	39.60	56.84	66.95	33.27
DGM+IDE[34]	36.81	54.01	68.51	16.87	42.36	57.92	69.31	33.62
Stepwise[21]	41.21	55.55	66.76	19.65	56.26	70.37	79.20	46.76
EUG ($p = 0.30$)	42.77	56.51	67.17	21.12	63.82	78.64	87.04	54.57
EUG ($p = 0.20$)	48.68	63.38	72.57	26.55	68.95	81.05	89.46	59.50
EUG ($p = 0.15$)	52.32	64.29	73.08	29.56	69.08	81.19	88.88	59.21
EUG ($p = 0.10$)	57.62	69.64	78.08	34.68	70.79	83.61	89.60	61.76
EUG ($p = 0.05$)	**62.67**	**74.94**	**82.57**	**42.45**	**72.79**	**84.18**	**91.45**	**63.23**
Baseline (supervised)	80.75	92.07	96.11	67.39	83.62	94.59	97.58	78.34

Table 1. Comparison with the state-of-the-art methods on MARS and DukeMTMC-VideoReID. All the methods are conducted based on the same backbone model ETAP-Net. Baseline (one-shot) is the initial model trained on one-shot labeled data. p is the enlarging factor that indicates the enlarging speed of the sampled subset. At the bottom we provide the Baseline (supervised) result as a upper bound where 100% training data are labeled.

Implementation Details. We use PyTorch [25] for all experiments. As discussed in Section 3.2, we take ETAP-Net as our basic CNN model for training on video-based re-ID. In experiments, we take ImageNet [14] pre-trained ResNet-50 model with last classification layer removed as the initialization of ETAP-Net. For training as a classification task for each identity, an additional fully-connected layer with batch normalization [12] and a classification layer are appended at the end of the model. The parameters of the first three residual blocks of ResNet-50 are kept fixed in training to save GPU memory and boost iterations. In training, we randomly sample 16 frames as the input for each tracklet. In label estimation and evaluation steps, all the frames are processed by the CNN model to get the representations for each tracklet, which are further l_2 normalized and used to calculate the Euclidean distance. We adopt the stochastic gradient descent (SGD) with momentum 0.5 and weight decay 0.0005 to optimize the parameters for 70 epochs with batch size 16 in each iteration. The overall learning rate is initialized to 0.1 and changed to 0.01 in the last 15 epochs.

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

We compare our method to DGM [34] and Stepwise [21] on the one-shot task. Note that although [21, 34] claim them as *unsupervised* methods, they are actually *one-shot* methods in experiments, because they require at least one labeled tracklet for each identity. Since the performances of both works were reported based on hand-crafted features, to make a fair comparison, we reproduce their methods using the same backbone model ETAP-Net (ResNet-50) as ours. The re-ID performance on MARS and DukeMTMC-VideoRe-ID are summarized in Table 1. On the MARS dataset, we achieve surprising result with rank-1 accuracy 62.67%, mAP 42.45% with enlarging factor 0.05, which greatly outperform the state-of-the-art result by 21.46 points and 22.8 points (absolute), respectively. The great performance gap between [21, 34] and ours is due to the excessive not-yet-reliable pseudo-labeled data incorporated at the first iteration. The estimation errors are accumulated during iterations and thus limit the further enhancement.

Moreover, Baseline (one-shot) and Baseline (supervised) are our initial model and the upper bound model, respectively. Baseline (one-shot) takes only the one-shot labeled data as the training set and do not exploit the unlabeled data. Baseline (supervised) is conducted on the fully supervised setting that all tracklets in the dataset are labeled and adopted in training. Specifically, we achieve 26.51 points and 33.19 points rank-1 improvements over the Baseline (one-shot) on MARS and DukeMTMC-VideoReID, respectively.

4.3. Algorithm Analysis

Analysis on the sampling criteria. As mentioned in Section 3.3, some previous works such as SPL take the classification loss as the criterion. The label estimation and evaluation performances of sampling by classification loss and by dissimilarity cost are illustrated in Figure 3 and Table 2. From the figure, we observe the huge performance gaps for both label estimation and evaluation. The label estimations of both criteria achieve similar and high precision at the beginning stage. However, the label estimation accuracy gap between two criteria gradually enlarges. As a result, the performance of the classification loss criterion is only enhanced to a limited extent and drops quickly in the subsequence. Table 2 shows the evaluation performance differences of the two criteria with different enlarging factors. With the same enlarging factor, the criterion of sampling by dissimilarity cost always leads to the superior performance. When the enlarging factor is set to 0.05, the best rank-1 accuracy on evaluation for classification loss and dissimilarity cost is 48.33% and 62.67%, respectively.

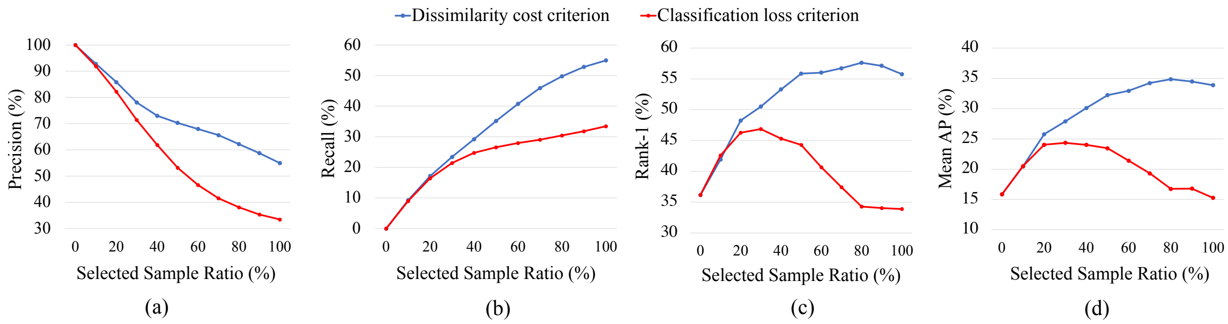

Figure 3. Comparison with two sampling criteria on MARS when the Enlarging Factor $p = 0.1$. (a) and (b): Precision and recall of the pseudo label prediction of selected pseudo-labeled candidates during iterations with different sampling criteria. (c) and (d): Rank-1 accuracy and mAP of person re-ID on the evaluation set during iterations with different sampling criteria. The x-axis stands for the percentage of selected data from entire unlabeled data for updating. Each solid point indicates an iteration step.

Enlarging Factor	Criteria	rank-1	rank-5	mAP
$p = 0.05$	Classification	48.33	62.67	25.35
	Dissimilarity	62.67	74.94	42.45
$p = 0.10$	Classification	46.86	60.25	24.23
	Dissimilarity	57.62	69.64	34.68
$p = 0.15$	Classification	46.53	60.12	24.03
	Dissimilarity	52.32	64.29	29.56
$p = 0.20$	Classification	45.91	59.95	23.56
	Dissimilarity	48.68	63.38	26.55
$p = 0.30$	Classification	41.86	56.01	20.24
	Dissimilarity	42.77	56.51	21.12

Table 2. Comparison of the two criteria on MARS. The "Classification" and "Dissimilarity" denotes the EUG methods with the classification loss criterion and the dissimilarity cost criterion, respectively. Note for that with the same enlarging factors, the dissimilarity cost criterion always lead to a superior performance.

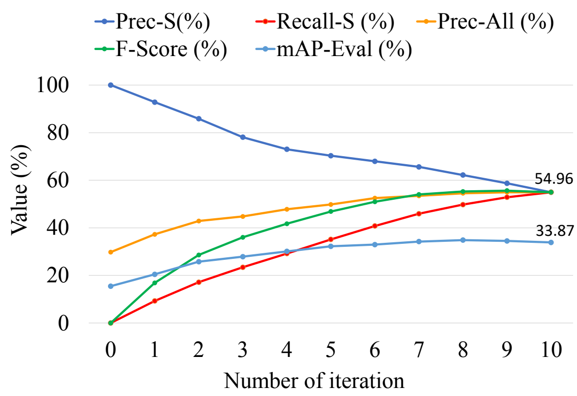

Figure 4. The label estimation performance with the enlarging factor = 0.1 over iterations on MARS. "Prec-S", "Recall-S" and "F-Score" denote the label estimation precision, recall and F-score for the *selected* pseudo-labeled candidates. "Prec-All" denotes the overall label estimation precision for *all* the unlabeled data. "mAP-Eval" represents the mAP performance of the evaluation on the test set. Note that on all the unlabeled data the overall label estimation accuracy is constantly increasing, which indicates the model learns much information throughout iterations.

Analysis over iterations. Figure 4 illustrates the label estimation performance and evaluation performance over iterations. At the initial iteration, the precision of pseudo label for the selected subset (blue line) is relatively high, since EUG only adopts a few of most reliable samples. In later stages, as EUG gradually incorporates more difficult and diverse samples, the precision drops along with the recall (red line) rising. In spite of the descending of precision, the F-score of label estimation (green line) continuous increases. Throughout iterations, the precision of pseudo label estimation for all the unlabeled data (orange line) constantly increases from 29.8% to 54.96%, which indicates the model grows robust steadily. At the last few iterations, the evaluation performance stops to increase, because the gain of adding new samples is offset by the loss of excessive pseudo label errors.

Analysis on the enlarging factor. For the iteration t, $t * p$ percent of unlabeled tracklets with reliable pseudo labels are sampled for updating the model. The effectiveness of enlarging factor p is shown in Figure 5. Two conclusions can be inferred: First, the model always achieves a better performance if we enlarge the selected set at a slower speed. The huge gaps among the five curves show that the great impact of the enlarging factor. Second, we observe that the gaps among the five curves are relatively small in the first several iterations and gradually enlarge in the later iterations. It shows the estimation errors are accumulated during iterations. This is because that the performance of the trained CNN model highly depends on the reliability of the training set. As a result, the evaluation performances appear obvious different in the last few iterations.

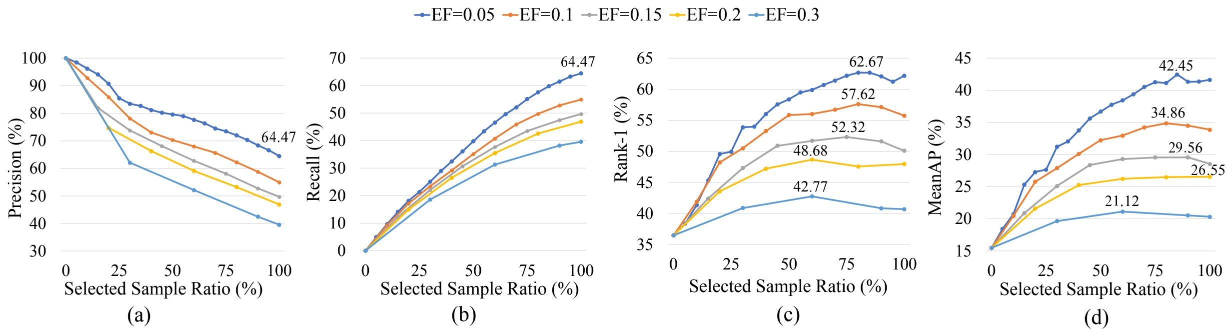

Figure 5. Comparison with different value of enlarging factor on MARS. (a) and (b) : Precision and recall of the pseudo label prediction of selected candidates with different enlarging factors. (c) and (d) : Rank-1 and mAP of person re-ID on the evaluation set with different enlarging factors. "EF" denotes the enlarging factor. The x-axis stands for the ratio of selected data from entire unlabeled data for updating. Each solid point indicates an iteration step. Note for that the lower enlarging factor is beneficial for improving performance.

Figure 6. The selected pseudo-labeled tracklets for an identity example on MARS with the enlarging factor $p = 0.1$. Error estimated samples are in red rectangles. All the tracklets incorporated in the former iterations are naturally selected by later ones. For this identity, one tracklet is missed, and four false samples are selected. Observe that the tracklet selected is easy and reliable at the beginning stage and difficult and diverse in the later stage.

4.4. Visualization

We visualize the selected samples for an identity during iterations in Figure 6. Since the initial tracklets is captured from the side view of the pedestrian, the two unlabeled tracklets captured from the same side are easily selected in iteration 0. In iteration 1 and 2, some tracklets in the behind or front view of the pedestrian are selected. The above tracklets are relatively easier for sampling. Further, in iteration 5 and 6, video tracklets suffering from obstructing and color variance are sampled. In iteration 7, samples with pedestrian of small size and dark background are selected. It's clear that the samples are selected from easy to hard, from similar to diverse. Note that there is no tracklet selected for this identity in iteration 3 and 4, which indicates the huge difficulty gap. There are also four mismatches in iteration 5, 6, and 7, in which the pedestrian is very similar to the ground truth identity, with the same pink shirt, gray pants, and long hair.

5. Conclusion

Label estimation for unlabeled tracklets is crucial for one-shot person re-ID. The challenge in the one-shot setting is that the pseudo labels are not reliable enough, which prevents the trained model from improving robust. To solve this problem, we propose a dynamic sampling strategy to start with easy and reliable unlabeled samples and gradually incorporating diverse tracklets for updating the model. We found that if we enlarge the selected set at a slower speed, the model achieves a better performance. In addition, we present a sampling criterion to remarkably improving the performance of label estimation. Our method surpasses the state-of-the-art method by 21.46 points (absolute) in rank-1 accuracy on MARS, and 16.53 points (absolute) on DukeMTMC-VideoReID. In sum, the proposed method is effective in exploiting the unlabeled data and reducing the annotation work load for one-shot video-based person re-ID.

Acknowledgment. Yi Yang is the recipient of a Google Faculty Research Award. Wanli Ouyang is supported by SenseTime Group Limited. We acknowledge the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centers Programme for funding this research.

References

[1] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *CVPR*, 2017. 2

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. 3

[3] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 3

[4] X. Dong, J. Huang, Y. Yang, and S. Yan. More is less: A more complicated network with less inference complexity. In *CVPR*, 2017. 3

[5] X. Dong, D. Meng, F. Ma, and Y. Yang. A dual-network progressive approach to weakly supervised object detection. In *ACM Multimedia*, 2017. 3, 4

[6] H. Fan, X. Chang, D. Cheng, Y. Yang, D. Xu, and A. G. Hauptmann. Complex event detection by identifying reliable shots from untrimmed videos. In *ICCV*, 2017. 3, 4

[7] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*, 2017. 1, 3

[8] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino. Semi-supervised multi-feature learning for person re-identification. In *AVSS*, 2013. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[10] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2

[11] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *CVPR*, 2017. 3

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6

[13] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014. 3

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 6

[15] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 3

[16] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 2

[17] Y. Lin, L. Zheng, and W. Y. a. Y. Y. Zheng, Zhedong and. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017. 3

[18] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017. 2

[19] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *ICCV*, 2014. 2

[20] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. *CVPR*, 2017. 1, 2

[21] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017. 1, 2, 5, 6

[22] A. J. Ma and P. Li. Semi-supervised ranking for re-identification with few labeled image pairs. In *ACCV*, 2014. 2

[23] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong. Self-paced co-training. In *ICML*, 2017. 3

[24] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016. 2

[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017. 6

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 5

[27] H. Shen, S.-I. Yu, Y. Yang, D. Meng, and A. Hauptmann. Unsupervised video adaptation for parsing human motion. In *ECCV*, 2014. 3, 4

[28] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, Dec 2016. 1

[29] L. Wu, C. Shen, and A. v. d. Hengel. Deep recurrent convolutional networks for video-based person re-identification: an end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016. 2

[30] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*. IEEE, 2016. 3

[31] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 3

[32] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. *ICCV*, 2017. 1, 2

[33] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):723–742, 2012. 3

[34] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. *ICCV*, 2017. 1, 2, 5, 6

[35] W. Zhang, S. Hu, and K. Liu. Learning compact appearance representation for video-based person re-identification. *arXiv preprint arXiv:1702.06294*, 2017. 1, 2

[36] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 5

[37] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 5

[38] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 3

[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 2

[40] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017. 2

[41] X. Zhu, X. Y. Jing, L. Yang, X. You, D. Chen, G. Gao, and Y. Wang. Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017. 2