# Progressive Learning for Person Re-Identification with One Example

Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian and Yi Yang

***Abstract*—In this paper, we focus on the one-example person re-identification (re-ID) task, where *each identity has only one labeled example along with many unlabeled examples*. We propose a progressive framework which gradually exploits the unlabeled data for person re-ID. In this framework, we iteratively (1) update the Convolutional Neural Network (CNN) model and (2) estimate pseudo labels for the unlabeled data. We split the training data into three parts, *i.e.*, labeled data, pseudo-labeled data, and index-labeled data. Initially, the re-ID model is trained using the labeled data. For the subsequent model training, we update the CNN model by the joint training on the three data parts. The proposed joint training method can optimize the model by both the data with labels (or pseudo labels) and the data without any reliable labels. For the label estimation step, instead of using a static sampling strategy, we propose a progressive sampling strategy to increase the number of the selected pseudo-labeled candidates step by step. We select a few candidates with most reliable pseudo labels from unlabeled examples as the pseudo-labeled data, and keep the rest as index-labeled data by assigning them with the data indexes. During iterations, the index-labeled data are dynamically transferred to pseudo-labeled data. Notably, the rank-1 accuracy of our method outperforms the state-of-the-art method by 21.6 points (absolute, *i.e.*, 62.8% vs. 41.2%) on MARS, and 16.6 points on DukeMTMC-VideoReID. Extended to the few-example setting, our approach with only 20% labeled data surprisingly achieves comparable performance to the supervised state-of-the-art method with 100% labeled data.**

***Index Terms*—Person Re-Identification, semi-supervised learning, few-example learning**

## I. Introduction

PERSON re-identification (re-ID) aims at spotting the person-of-interest from non-overlapping camera views. In recent years, deep convolutional neural networks (CNN) has led to impressive successes in the field of re-ID [1], [2], [3]. Most existing re-ID methods, in particular deep learning models, adopt the supervised learning approach. These methods rely on the full annotations, *i.e.*, the identity labels of all the training data from multiple cross-view cameras. However, it is labor intensive to annotate large-scale data. The intensive human labor may limit re-ID applications, especially when there are many cameras.

Recently, there are a few semi-supervised person re-ID methods [4], [5], [6]. We focus on the one-example setting, in which only one example is available for each identity. The setting is more challenging but provides a much more cost-effective solution to the real world re-ID problems, where cross camera annotation requires substantial annotation efforts.

Most existing methods [6], [7] employ a static strategy to determine the quantity of selected pseudo-labeled data for further training. The samples with confidence higher than a *pre-defined* threshold are then selected for the subsequent training. During iterations, these algorithms select a fixed size of pseudo-labeled training data from beginning to end. However, in the initial stage, only a few label predictions are reliable due to the very few labeled examples as initialization. As the iteration goes, the model gets more robust, resulting in more accurate label predictions. Therefore, keeping the size of the selected data fixed would hinder the performance improvement.

We propose a progressive learning framework to better exploit the unlabeled data for person re-ID with limited exemplars. Initially, a CNN model is trained on the one-example labeled samples. We then generate the pseudo labels for all unlabeled samples, and select some reliable pseudo-labeled data for training according to the prediction confidence. Different from existing methods [5], [6], the selected subset is continuously enlarged during iterations according to a sampling strategy. At the initial stages, we only include the most reliable and easiest ones. In the subsequent stages, we gradually select a growing number of pseudo-labeled data to incorporate more difficult and diverse data.

Wu *et al.* [8] proposed a progressive sampling strategy and dissimilarity sampling criterion for one-example video-based re-ID, where each identity has a labeled tracklet (62 image frames on average). In the progressive sampling strategy, the data is split into two parts, labeled data and pseudo-labeled data which are selected to update the model. Different from [8], in this paper, we consider the one-shot image-based re-ID setting, where only one *image* instead of a *tracklet* is labeled for each identity. Given the very limited data, it is very hard to initialize the model. We therefore propose to leverage the unlabeled data in a self-supervised manner to help to learn a robust model. The previous pseudo labelling in [8] utilizes part of the unlabelled data, but overlooks a large number of unlabelled data (unselected data) whose pseudo labels are not reliable. These human images also preserve the data distribution information due to the large amount at initial iterations. We propose to use the remaining unselected data (which is unreliable during pseudo labeling) on top of the pseudo-labeled data. We label these data by their indexes and design the exclusive loss to optimize the CNN model by these *index-labeled* data. Different from the widely used identity

(Corresponding author: Yi Yang.)

Y. Wu, and Y. Yang are with SUSTech-UTS Joint Centre of CIS, Southern University of Science and Technology, and the Centre for Artificial Intelligence, University of Technology Sydney (e-mail: yu.wu-3@student.uts.edu.au; yi.yang@uts.edu.au). Y. Lin, X. Dong, Y. Yan, W. Bian are with the Centre for Artificial Intelligence, University of Technology Sydney (e-mail: yutian.lin@student.uts.edu.au; xuanyi.dong@student.uts.edu.au; yan.yan-3@student.uts.edu.au; wei.bian@uts.edu.au).

cross-entropy loss, the target of the exclusive loss is to repel any two unlabelled data way from each other, which optimizes the model by learning to distinguish the difference of all the input images (or tracklets).

Specifically, we split training data into three parts, labeled data, selected pseudo-labeled data and index-labeled data. We then propose a joint learning method to simultaneously train the CNN model on all the data splits. Figure **??** shows the three different data sources in our framework. For the labeled data and selected pseudo-labeled data, we apply an identity classifier on their CNN features and further optimize the model by comparing identity predictions and (pseudo-) labels. For those index-labeled data, we use the exclusive loss to optimize the model without any identity labels. The classification loss pulls representations of the same identity data close to each other, while the exclusive loss pushes representations of all the index-labeled samples away from each other. We observe in our experiments that our method can effectively uncover data distribution and generate a robust model even at initial stages. We also extend our method to the few-example setting. On the MARS dataset, our method reduces 80% annotation cost with only a 4.3% rank-1 accuracy drop.

Our contributions are summarized as follows:

- We propose a progressive method for one-example person re-ID to better exploit the unlabeled data. This method adopts a dynamic sampling strategy that we start with reliable samples and gradually includes diverse ones, which significantly makes the model robust.
- We apply a distance-based sampling criterion for label estimation and apply candidates selection to remarkably improve the performance of label estimation.
- We propose a joint learning method to simultaneously train the CNN model on the labeled, pseudo-labeled and index-labeled data.
- Our method achieves surprisingly superior performance on the one-example setting, outperforming the state-of-the-art by 21.6 points (absolute) on MARS and 16.6 points (absolute) on DukeMTMC-VideoReID.
- Our approach can be readily extended into the few-example setting (with 20% labeled data). It achieves comparable rank-1 performance (76.5%) compared to the state-of-the-art performance (79.80%) in the supervised setting (with 100% labeled data).

## II. Related Works

### A. Supervised person re-ID

In recent years, deep learning based methods have been proved to be effective in many computer vision applications [9], [10], [11]. To address the re-ID problem, CNN models [12], [1], [13], [14], [15] are used for the person re-ID task and have obtained impressive performance. A branch of works [10], [16] use the siamese model which takes image pairs or triplets as input. Li *et al.* [12] train the network with pairs of pedestrian images, where the verification model with the patch-matching layer is adopted. Another branch of researches adopts identity classification models. Zheng *et al.* [2] propose an identity discriminative embedding (IDE) that directly use a conventional fine-tuning approach on the Market-1501 dataset and obtain competitive results. Xiao *et al.* [1] train a classification model from multiple domains and propose a domain guided dropout. Zheng *et al.* [17] adopt a joint classification and verification model and use two pairs of images for training. In this work, we train our classification model in the same way as in [2].

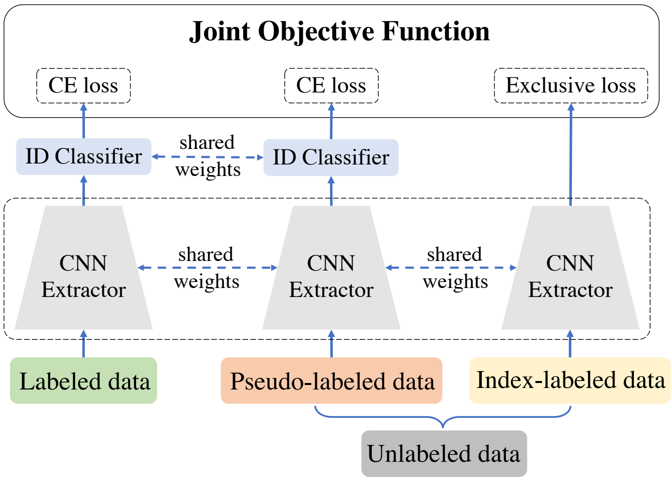


Fig. 1. The proposed joint training procedure for one-example person re-ID. "CE loss" denotes the Cross-Entropy loss. We select a few unlabeled data with reliable pseudo labels as the pseudo-labeled data. On the labeled and pseudo-labeled data, we optimize the CNN model by an ID classifier and the Cross-Entropy loss. The remaining unlabeled data are taken as the index-labeled data, where no reliable pseudo label is available. Without any identity labels, the exclusive loss can still help to learn a robust model.

### B. Semi-supervised learning and Progressive Paradigm

Semi-supervised learning [18], [19], [20], [21], [22] takes advantages from both labeled and unlabeled data to solve the given task. Some semi-supervised approaches [23] use graph representations in recent years. Kipf *et al.* [23] encode the graph structure directly using a neural network model and train on a supervised target for all nodes with labels. Most recently, with the great success of the Generative Adversarial Network (GAN) [24], many researchers adopt semi-supervised learning to explore images generated by GAN [25], [3], [26]. Salimans *et al.* [25] present a variety of new architectural features and training procedures that apply to the generative adversarial networks framework.

Curriculum Learning (CL) is proposed in [27], which progressively obtains knowledge from easy to hard samples in a pre-defined scheme. Kumar *et al.* [28] 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 [29], [30]. We are inspired by these progressive algorithms. Compared with the existing SPL and CL algorithms, we incorporate the retrieval measures (the distance in feature space) into the learning mechanism, which well fits the evaluation metric for person re-ID. We also introduce the joint learning method to explore all of the labeled and unlabeled data.

### C. Semi-Supervised person re-ID

In this paper, we follow the semi-supervised person re-ID setting as in [31], which assumes that few data from the training set are labeled, and the rest of the training set is unlabeled. In [31], a novel semi-supervised region metric learning method is proposed, that estimates positive neighbors to generate positive regions and learn a discriminative region-to-point metric. There are some works that focus on the few-example video-based re-ID task. Ye *et al.* [6] propose a dynamic graph matching (DGM) method, which iteratively updates the image graph matching and the label estimation to learn a better feature. Liu *et al.* [5] 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. Even though [5], [6] claim that they are *unsupervised* methods, they are actually *one-example* methods in experiments, because both of them require at least one labeled tracklet for each identity. To be more rigorous, we consider this problem as a one-example task.

### D. Unsupervised domain adaptation for person re-ID

Recently, some cross-domain transfer learning methods [7], [32], [33] focus on the unsupervised domain adaptation re-ID task, where information from an external source dataset is utilized. Peng *et al.* [32] propose to learn a discriminative representation for target domain based on asymmetric multi-task dictionary learning. Fan *et al.* [7] propose a progressive method for domain adaptation, where the K-means clustering and the IDE [2] network pre-trained on the source dataset is updated iteratively. Different from these methods, our work focuses on the one-example setting that we do not require the images and identity annotations from an additional dataset.

## III. The Progressive Model

We first introduce the framework overview of the proposed method in Section III-A, and the preliminaries in Section III-B. Then we illustrate the two key parts of our method, *i.e.*, the joint learning method in Section III-C and the label estimation in Section III-D. Lastly, we present the overall progressive iteration strategy in Section III-E.

### A. Framework Overview

As shown in Figure 2, our method updates the model by the following two steps iteratively: 1. train the CNN model by the joint learning on the *labeled* data, *pseudo-labeled* data, and *index-labeled* data; 2. select a few reliable pseudo-labeled candidates from unlabeled data according to a prediction reliability criterion. Specifically, in the first iteration, all the unlabeled data have no pseudo labels. During iterations, we continuously enlarge the set of selected pseudo-labeled candidates. The remaining unlabeled data are taken as the index-labeled data by labeling them with the data index. Our joint learning method then progressively learns a robust and stable model on the three data splits. In the next iteration, the pseudo labels are assigned to the unlabeled candidates by the identity labels of their nearest labeled neighbors in the feature space. The distance between them is considered as the dissimilarity cost, which is the measure of reliability for the pseudo label.

### B. Preliminaries

We first introduce the necessary notations for the one-example re-ID task. Let $x$ and $y$ denote the pedestrian visual data and the identity label, respectively. The visual data $x$ can be either a person image for image-based re-ID, or a tracklet (a series of person images) for video-based re-ID. For the training in the one-example re-ID task, we have the labeled data set $\mathcal{L} = \{(x_1, y_1), ..., (x_{n_l}, y_{n_l})\}$ and the unlabeled data set $\mathcal{U} = \{x_{n_l+1}, ..., x_{n_l+n_u}\}$. Usually, these data are utilized in an identity classification way to train the re-ID model $\phi(\boldsymbol{\theta}, \cdot)$. For the evaluation stage, the trained CNN model $\phi$ is used to embed both query data and gallery data 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.*., $||\phi(\boldsymbol{\theta}; x_q) - \phi(\boldsymbol{\theta}; x_g)||$, where $x_q$ and $x_g$ denote the query data and the gallery data, respectively. To exploit abundant unlabeled data, we predict the pseudo label $\hat{y}_i$ for each unlabeled data $x_i \in \mathcal{U}$ and select a few reliable ones for the identity classification learning. We denote $\mathcal{S}^t$ and $\mathcal{M}^t$ as the pseudo-labeled dataset and index-labeled data set at $t$-th step, respectively.

### C. The Joint Learning Method

We first introduce the model updating step. At the $t$-th iteration, we have three kinds of the data source for training, *i.e.*, the labeled data $\mathcal{L}$, the selected pseudo-labeled data $\mathcal{S}^t$ and the remaining index-labeled data $\mathcal{M}^t$. We utilize the labeled data $\mathcal{L}$ and the pseudo-labeled data $\mathcal{S}^t$ by the identity classification learning with their (pseudo-) labels. For the index-labeled data $\mathcal{M}^t$, their pseudo labels are not-yet-reliable and may harm the model training. Therefore, we utilize the exclusive loss to optimize the CNN model on $\mathcal{M}^t$.

**The Exclusive Loss** aims at learning a discriminative embedding on $\mathcal{M}^t$ without any identity labels. In general, we optimize the CNN model by learning to distinguish samples, rather than identities. To push each index-labeled data $x_i \in \mathcal{M}^t$ away from the other data $x_j \in \mathcal{M}^t, i \neq j$ in the feature space, we have the following target for unsupervised feature learning:

$$\max_{\boldsymbol{\theta}} \sum_{\substack{x_i, x_j \in \mathcal{M}^t \\ x_i \neq x_j}} ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x_j)||, \qquad (1)$$

where $|| \cdot ||$ denotes the Euclidean distance.

To solve Eq. (1) in an efficient way, we have the following approximation. Let $v_i = \tilde{\phi}(\boldsymbol{\theta}; x_i)$ be the L2-normalized feature embedding for the data $x_i$, *i.e.*, $||v_i|| = 1$. Since $||v_i - v_j||^2 = 2 - v_i^{\mathrm{T}} v_j$, maximizing the Euclidean distance between data $x_i$ and $x_j$ is equivalent to minimize the cosine similarity $v_i^{\mathrm{T}} v_j$. Therefore, Eq. (1) can be approximately optimized by a softmax-like loss:

$$\ell_e(\boldsymbol{V}; \tilde{\phi}(\boldsymbol{\theta}; x_i)) = -\log \frac{\exp({v_i}^{\mathrm{T}} \tilde{\phi}(\boldsymbol{\theta}; x_i)/\tau)}{\sum_{j=1}^{|\mathcal{M}^t|} \exp({v_j}^{\mathrm{T}} \tilde{\phi}(\boldsymbol{\theta}; x_i)/\tau)}, \qquad (2)$$

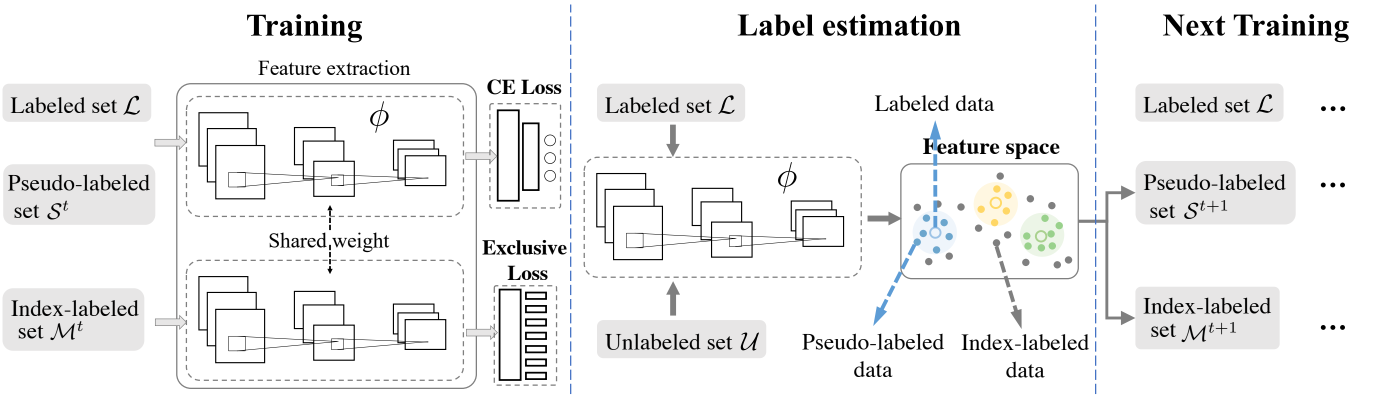

Training
Label estimation
Next Training
Feature extraction
Labeled set $\mathcal{L}$
Pseudo-labeled set $\mathcal{S}^t$
Index-labeled set $\mathcal{M}^t$
$\phi$
Shared weight
CE Loss
Exclusive Loss
Labeled set $\mathcal{L}$
Unlabeled set $\mathcal{U}$
Labeled data
Feature space
Pseudo-labeled data
Index-labeled data
Labeled set $\mathcal{L}$
Pseudo-labeled set $\mathcal{S}^{t+1}$
Index-labeled set $\mathcal{M}^{t+1}$
...

Fig. 2. Overview of the proposed iterative framework. In each iteration, (1) we train the CNN model by the joint learning on the *labeled* data, *pseudo-labeled* data, and *index-labeled* data. We utilize the labeled data and the pseudo-labeled data by an identity classification learning with their (pseudo-) labels. On the index-labeled data where no reliable pseudo label is available, we apply the exclusive loss directly on the images to optimize the model without any identity label. (2) In the label estimation step, we select a few reliable pseudo-labeled candidates from unlabeled data $\mathcal{U}$ to the selected set $\mathcal{S}$ according to the distance in feature space. Nodes with different colors in the feature space box denote different identity samples.

where $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{M}^t| \times n_\phi}$ is a lookup table that stores the features of each index-labeled data $x_i$. $\tau$ is a temperature parameter that controls the concentration level of the distribution. A higher temperature $\tau$ leads to a softer probability distribution. Inspired by [34], we adopt the lookup table $\boldsymbol{V}$ to avoid the exhaustive computation of extracting features from all data at each training step. In the forward operation, we compute cosine similarities between data $x_i$ and all the other data by $\boldsymbol{V}^T\tilde{\phi}(\boldsymbol{\theta}; x_i)$. During backward, we update the $i$-th column of the table $\boldsymbol{V}$ by $v_i \leftarrow \frac{1}{2}(v_i + \tilde{\phi}(\boldsymbol{\theta}; x_i))$ and then L2-normalize $v_i$ to a unit vector.

The exclusive loss is a self-supervised auxiliary loss to learn discriminative representations from the unlabeled data. During the model optimization, to achieve the target that any two unlabelled samples are repelled away, the exclusive loss forces the model to learn to distinguish the difference of input images. Therefore, the learned representation is expected to focus more on details of an input identity. During this procedure, our model accesses more samples. Although these samples have neither ground truth labels nor reliable pseudo labels, they can still provide some weak supervision information by exploiting the differences between human images.

**The Joint objective Function.** There are three data parts, *i.e.*, the labeled data, pseudo-labeled data, and the index-labeled data. We jointly optimize our model on all data parts. On the labeled dataset $\mathcal{L}$ where we have the ground truth identity labels, we follow recent works [7], [35], [36] to train the re-ID model. We have the following objective function:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \sum_{i=1}^{n_l} \ell_{\text{CE}}(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{3}$$

where $f(\boldsymbol{w}; \cdot)$ is an identity classifier, parameterized by $\boldsymbol{w}$, to classify the embedded feature $\phi(\boldsymbol{\theta}; x_i) \in \mathbb{R}^{n_\phi}$ into a $k$-dimension confidence estimation, in which $k$ is the number of *identities*. $\ell_{\text{CE}}$ denotes the cross-entropy loss on the identity label prediction $f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)) \in \mathbb{R}^k$ and its ground truth identity label $y_i$. Similarly, we can optimize the model on the pseudo-labeled data set $\mathcal{S}$ by

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \sum_{i=n_l+1}^{n_l+n_u} s_i \ell_{\text{CE}}(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), \hat{y}_i), \tag{4}$$

where $s_i \in \{0, 1\}$ is the selection indicator for the unlabeled sample $x_i$, which is generated from previous label estimation step. $s_i$ determines whether we should select pseudo-labeled data $(x_i, \hat{y}_i)$ for identity classification training. We will discuss it later in Section III-D.

Considering the three data splits, we design the following objective function for the model training at $t$-th iteration:

$$\begin{aligned} \min_{\boldsymbol{\theta},\boldsymbol{w}} \lambda \sum_{i=1}^{n_l} \ell_{\text{CE}}(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), y_i) + \\ \lambda \sum_{i=n_l+1}^{n_l+n_u} s_i^{t-1} \ell_{\text{CE}}(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), \hat{y}_i) + \\ (1-\lambda) \sum_{i=n_l+1}^{n_l+n_u} (1 - s_i^{t-1}) \ell_e(\boldsymbol{V}; \tilde{\phi}(\boldsymbol{\theta}; x_i)), \end{aligned} \tag{5}$$

where $\lambda$ is a hyper-parameter to adjust the contribution of the identity classification loss $\ell_{\text{CE}}$ and the exclusive loss $\ell_e$. The Eq. (5) consists of three loss parts. The first part is the ID classification loss on the labeled set $\mathcal{L}$. The second one is the ID classification loss on the selected pseudo-labeled set $\mathcal{S}^t$. The last one is the exclusive loss on the index-labeled set $\mathcal{M}^t$.

### D. The Effective Sampling Criterion

The label estimation step is crucial to obtain the appropriately selected candidates $\mathcal{S}^t$ to exploit the unlabeled data. The previous works sample the unlabeled data from confident to uncertain ones according to the classification loss [37]. However, the loss from classification prediction does not well fit the retrieval evaluation. Moreover, the classifier may easily over-fit the one-example labeled data. Thus it may be not robust in predicting identities. To address this problem, we propose an effective sampling criterion, which takes the distance in the feature space as a measure of pseudo label reliability. For the label estimation on unlabeled data, we adopt the Nearest Neighbors (NN) classifier instead of the learned identity classification. The NN classifier assigns the pseudo label for each unlabeled data by its nearest labeled neighbor in feature space. And the distance between them is regarded as the confidence of label estimation. For the candidates selection, we select a few top reliable pseudo-labeled data according to their label estimation confidence.

More formally, we estimate the pseudo label for each unlabeled data $x_i \in \mathcal{U}$ by:

$$x^*, y^* = \arg \min_{(x_l, y_l) \in \mathcal{L}} ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x_l)||, \tag{6}$$

$$d(\boldsymbol{\theta}; x_i) = ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x^*)||, \tag{7}$$

$$\hat{y}_i = y^*, \tag{8}$$

where $d(\boldsymbol{\theta}; x_i)$ is the **dissimilarity cost** of label estimation. To select candidates, at the iteration step $t$, we sample the pseudo-labeled candidates into training by setting the selection indicators as follows:

$$\boldsymbol{s}^t = \arg \min_{||\boldsymbol{s}^t||_0 = m_t} \sum_{i=n_l+1}^{n_l+n_u} s_i d(\boldsymbol{\theta}; x_i), \tag{9}$$

where $m_t$ denotes the size of selected pseudo-labeled set and $\boldsymbol{s}^t$ is the vertical concatenation of all $s_i$. Eq. (9) selects the top $m_t$ nearest unlabeled data for all the labeled data at the iteration step $t$.

### *E. The overall iteration strategy*

We iteratively train the CNN model and then estimate labels for unlabeled data. At each iteration, we first optimize the model by Eq. (5). Then we estimate labels for unlabeled data by Eq. (8) and select some reliable ones by applying the trained model on Eq. (9). Since the initial labeled data are too few to depict the detailed underlying distribution, it is irrational to incorporate excessive pseudo-labeled data in training at the initial iteration.

We propose a dynamic sampling strategy to ensure the reliability of selected pseudo-labeled samples. It starts with a small proportion of pseudo-labeled data at the beginning stages and then incorporates more diverse samples in the following stages. We start our framework by setting $m_0 = 0$ and $\mathcal{M}^0 = \mathcal{U}$, *i.e.*, optimizing the model by (1) identity classification training on labeled data $\mathcal{L}$ and (2) unsupervised training by exclusive loss on all the unlabeled data. In later iterations, we progressively increase the size of selected pseudo-labeled candidates set $|\mathcal{S}^t|$. At iteration step $t$, we enlarge the size of sampled pseudo-labeled data by set $m_t = m_{t-1} + p \cdot n_u$, where $p \in (0, 1)$ is the **enlarging factor** which indicates the speed of enlarging the candidates set during iterations. As described in Algorithm 1, we evaluate the model $\phi(\boldsymbol{\theta}; \cdot)$ on the validation set at each iteration step and output the best model. In the one-example experiment, we take another person re-ID training set as the validation set.

**How to find a proper enlarging factor $p$ for real-life applications?** The enlarging factor controls the speed of enlarging the reliable pseudo-labeled candidates set during iterations. Smaller enlarging factor indicates lower enlarging speed, therefore, more iteration steps and training time. In the real-life application, this factor is a trade-off between efficiency and accuracy. 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.

**Algorithm 1** The proposed framework

```
Require: Labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$, enlarging factor $p \in (0,1)$, initialized CNN model $\boldsymbol{\theta}_0$.
Ensure: 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 $m_{t+1} \le |\mathcal{U}|$ do
    $t \leftarrow t + 1$
    Update the model $(\boldsymbol{\theta}_t, \boldsymbol{w}_t)$ on $\mathcal{L}$, $\mathcal{S}^t$ and $\mathcal{M}^t$ via Eq. (5).
    Estimate pseudo labels for $\mathcal{U}$ via Eq. (8)
    Generate the selection indicators $\boldsymbol{s}_t$ via Eq. (9)
    Update the sampling size: $m_{t+1} \leftarrow m_t + p \cdot n_u$
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
```

## IV. Experimental Analysis

### *A. Datasets and Settings*

We evaluate our method on four large-scale re-ID datasets.

**Market-1501** [2] contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing.

**DukeMTMC-reID** [17] is a re-ID dataset derived from the DukeMTMC dataset [38]. It contains 16,522 training images of 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images.

**MARS** [39] 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.

**DukeMTMC-VideoReID** [8] is a subset of DukeMTMC [38] for video-based person re-ID. DukeMTMC-VideoReID consists of 702 identities for training, 702 identities for test, and 408 identities as distractors. In total there are 2,196 videos for training and 2,636 videos for test. Each video contains images sampled every 12 frames. During test, a video for each ID is used as the query, and the remaining videos are placed in the gallery.

**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 precision across all queries. We report the Rank-1, Rank-5, Rank-10, 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-example experiments, we use the same protocol as [5]. In all datasets, we randomly choose an image/trackelt from Camera 1 for each identity as initialization. If there is no data recorded by Camera 1 for one identity, we randomly select a sample from the next camera to make sure each identity has one sample for initialization. Note that as discussed in Section II-C, [5], [6] are using the same one-example setting in experiments.

**Implementation Details.** We adopt ResNet-50 with the last classification layer removed as our feature embedding model $\phi$ to conduct all the experiments. We initialize it by the ImageNet [40] pre-trained model. To optimize the model by the (pseudo-) label loss, we append an additional fully-connected layer with batch normalization and a classification layer on the top of the CNN feature extractor. For the exclusive loss, we process the unlabeled feature by a fully-connected layer with batch normalization, followed by a L2-normalization operation. Following [34], the temperature scalar $\tau$ in Eq. (2) is set to 0.1. We set $\lambda$ in Eq. (5) to be 0.8 for all the experiments. In each model updating step, the stochastic gradient descent (SGD) with momentum 0.5 and weight decay 0.0005 is used to optimize the parameters for 70 epochs with batch size 16. The overall learning rate is initialized to 0.1. In the last 15 epochs, to stabilize the model training, we change the learning rate to 0.01 and set $\lambda = 1$. For the experiments on video-based re-ID datasets, we simply add a temporal average pooling layer on the CNN extractor, where we element-wisely average features of all frame within a tracklet.

### B. Comparison with the State-of-the-Art Methods

There are two recent works designed for one-example video-based person re-ID, *i.e.*, DGM [6] and Stepwise [5]. Note that although [5], [6] claim them as *unsupervised* methods, they are actually *one-example* methods in experiments, because they require at least one labeled tracklet for each identity. We compare our method to them on the one-example video-based re-ID task. 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 ResNet-50 as ours. The re-ID performance of our method on the four large-scale re-ID datasets are summarized in Table I and Table II. With only one labeled example for each identity, our method achieves surprising performance on both image-based and video-based re-ID task.

Moreover, we compare our method to two baseline methods, *i.e.*, the Baseline (one-example) and Baseline (supervised), which are our initial model and the upper bound model (100% data are labeled), respectively. Baseline (one-example) takes only the one-example labeled data as the training set and do not exploit the unlabeled data. Baseline (supervised) is conducted on the fully supervised setting that all data are labeled and adopted in training. Specifically, we achieve 29.8, 32.4, 26.6 and 33.3 points of rank-1 accuracy improvement over the Baseline (one-example) on Market-1501,DukeMTMC-reID, MARS and DukeMTMC-VideoReID, respectively. The superior performances on the four large-scale datasets validate the effectiveness of our proposed method.

### C. Ablation studies

We conduct ablation studies on the two key parts of our methods, *i.e.*, the joint learning method and the dissimilarity criterion, as shown in Table III and Figure 3. All experiments share the same training parameters and initial labeled images.

**The effectiveness of the joint learning method.** We compare our method to the model trained without the joint learning method, denoted as "Ours w/o J" in Table III. The "Ours w/o J" model is only optimized by the identity classification loss on the labeled and selected pseudo-labeled data, as proposed in the preliminary version [8]. The comparison results on the two datasets prove the effectiveness of the joint learning method. Compared to the great improvement on the image-based task (Market-1501), the improvement of the video-based task (MARS) is relatively small. The main reason is that the one-example initial model in the video-based re-ID task is much more robust compared to the image-based one, since an initial tracklet contains many images (62 frames on average on MARS) of the same identity. It can be seen from the accuracy difference of the initial label predictions on all the unlabeled data, *i.e.*, 30.0% on MARS while 11.9% on Market-1501. Exploiting the unlabeled data with a relatively robust model may not benefit the feature learning a lot.

**The effectiveness of the sampling criteria.** As mentioned in Section III-D, some previous works such as SPL take the classification loss as the criterion. The label estimation accuracy and re-ID performances of sampling by classification loss and by dissimilarity cost are illustrated in Figure 3 and Table III. We observe the huge performance gaps in Figure 3 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 III shows the evaluation performance differences between the two criteria with different enlarging factors. "Ours w/o D" denotes the method with classification loss as the criterion. With the same enlarging factor, the criterion of sampling by dissimilarity cost always leads to superior performances.

### D. Algorithm Analysis

**Analysis over iterations.** Figure 3 illustrates the label estimation performance and re-ID performance over iterations. Since we only collect a few most reliable unlabeled samples as pseudo-labeled data, the precision score of label estimation is relatively high at the beginning. As iteration goes, we adopt

TABLE I

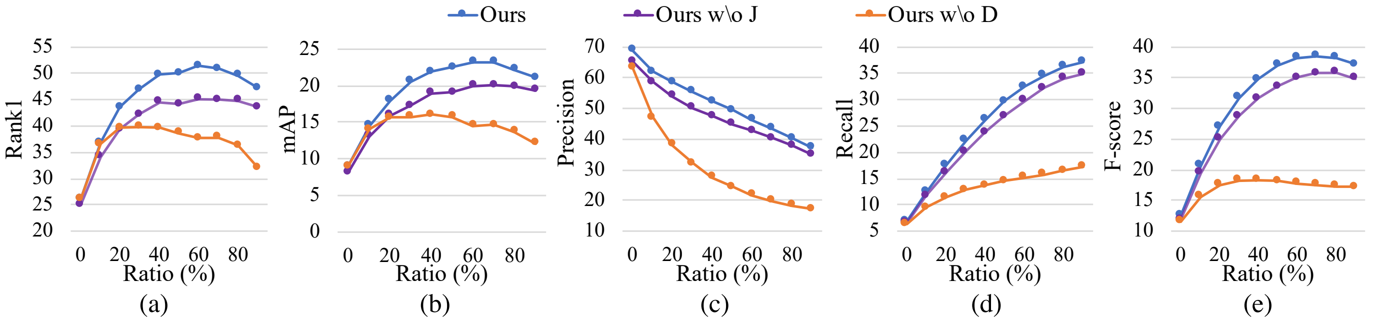
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON TWO IMAGE-BASED LARGE-SCALE RE-ID DATASETS. BASELINE (ONE-EXAMPLE) IS THE INITIAL MODEL TRAINED ON ONE-EXAMPLE LABELED DATA. BASELINE (SUPERVISED) SHOWS THE UPPER BOUND PERFORMANCE WHERE 100% TRAINING DATA ARE LABELED. $p$ IS THE ENLARGING FACTOR THAT INDICATES THE ENLARGING SPEED OF THE SELECTED PSEUDO-LABELED SUBSET.

| Settings | Methods | Market-1501 | | | | | DukeMTMC-reID | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | rank-1 | rank-5 | rank-10 | rank-20 | mAP | rank-1 | rank-5 | rank-10 | rank-20 | mAP |
| Supervised | Baseline [35] | 83.1 | 92.5 | 95.0 | 96.9 | 63.7 | 71.0 | 83.2 | 87.3 | 89.9 | 49.3 |
| One-Example | Baseline [35] | 26.0 | 41.4 | 49.2 | 59.6 | 9.0 | 16.4 | 27.9 | 32.8 | 39.0 | 6.8 |
| | Ours ($p = 0.30$) | 35.5 | 52.8 | 60.5 | 68.6 | 13.4 | 23.3 | 35.7 | 42.2 | 48.0 | 11.1 |
| | Ours ($p = 0.20$) | 41.4 | 59.6 | 66.4 | 73.5 | 17.4 | 30.0 | 43.4 | 49.2 | 54.8 | 15.1 |
| | Ours ($p = 0.15$) | 44.8 | 61.8 | 69.1 | 76.1 | 19.2 | 35.1 | 49.1 | 54.3 | 60.0 | 18.2 |
| | Ours ($p = 0.10$) | 51.5 | 66.8 | 73.6 | 79.6 | 23.2 | 40.5 | 53.9 | 60.2 | 65.5 | 21.8 |
| | Ours ($p = 0.05$) | **55.8** | **72.3** | **78.4** | **83.5** | **26.2** | **48.8** | **63.4** | **68.4** | **73.1** | **28.5** |

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON TWO VIDEO-BASED LARGE-SCALE RE-ID DATASETS. BASELINE (ONE-EXAMPLE) IS THE INITIAL MODEL TRAINED ON ONE-EXAMPLE LABELED DATA. BASELINE (SUPERVISED) SHOWS THE UPPER BOUND PERFORMANCE WHERE 100% TRAINING DATA ARE LABELED. $p$ IS THE ENLARGING FACTOR THAT INDICATES THE ENLARGING SPEED OF THE SELECTED PSEUDO-LABELED SUBSET.

| Settings | Methods | MARS | | | | | DukeMTMC-VideoReID | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | rank-1 | rank-5 | rank-10 | rank-20 | mAP | rank-1 | rank-5 | rank-10 | rank-20 | mAP |
| Supervised | Baseline [35] | 80.8 | 92.1 | 94.6 | 96.1 | 63.7 | 83.6 | 94.6 | 96.9 | 97.6 | 78.3 |
| One-Example | Baseline [35] | 36.2 | 50.2 | 57.2 | 61.9 | 15.5 | 39.6 | 56.8 | 62.5 | 67.0 | 33.3 |
| | DGM+IDE [6] | 36.8 | 54.0 | 59.6 | 68.5 | 16.9 | 42.4 | 57.9 | 63.8 | 69.3 | 33.6 |
| | Stepwise [5] | 41.2 | 55.6 | 62.2 | 66.8 | 19.7 | 56.3 | 70.4 | 74.6 | 79.2 | 46.8 |
| | Ours ($p = 0.30$) | 44.5 | 58.7 | 65.7 | 70.6 | 22.1 | 66.1 | 79.8 | 84.9 | 88.3 | 56.3 |
| | Ours ($p = 0.20$) | 49.6 | 64.5 | 69.8 | 74.4 | 27.2 | 69.1 | 81.2 | 85.6 | 89.6 | 59.6 |
| | Ours ($p = 0.15$) | 52.7 | 66.3 | 71.9 | 76.4 | 29.9 | 69.3 | 81.4 | 85.9 | 89.2 | 59.5 |
| | Ours ($p = 0.10$) | 57.9 | 70.3 | 75.2 | 79.3 | 34.9 | 71.0 | 83.8 | 87.4 | 90.3 | 61.9 |
| | Ours ($p = 0.05$) | **62.8** | **75.2** | **80.4** | **83.8** | **42.6** | **72.9** | **84.3** | **88.3** | **91.4** | **63.3** |



(a) (b) (c) (d) (e)

Fig. 3. Ablation studies on Market-1501. We validate the effectiveness of the two parts of our method, *i.e.*, the joint learning method (denoted as J) and the dissimilarity cost (denoted as D). The enlarging factor $p$ is set to 0.1 in the comparison. (a) and (b): Rank-1 accuracy and mAP on the evaluation set during iterations. (c), (d) and (e): Precision, recall, and F-score of the label prediction of selected pseudo-labeled candidates during iterations. The x-axis stands for the percentage of selected data over entire unlabeled data. Each solid point indicates an iteration step.

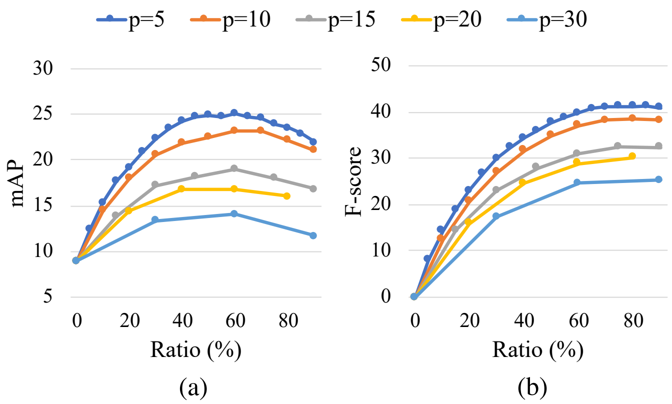
more unlabeled data into the pseudo-labeled set, resulting in a continuous precision drop of the label estimation. However, in this procedure, the recall score of label estimation gradually increases as more correctly estimated pseudo-labeled data are used. The overall label estimation performance, the F-score, appears a rapid increase at the first several iterations and a slight performance drop in the last iterations. Interestingly, the re-ID evaluation performances, *i.e.*, Rank-1 and mAP scores, show a similar curve with F-score, which indicates the label estimation quality is the key factor for the one-example task.

**Analysis on the enlarging factor.** The enlarging factor $p$ is a key parameter in our framework. It controls the speed of enlarging pseudo-labeled candidates set during iterations. Smaller enlarging factor indicates lower enlarging speed, therefore, more iteration steps and training time. The results of different enlarging factors can be found in Figure 4. As we can see, in experiments, a smaller enlarging factor always yields better performance. It is consistent with human intuition since each enlarging step is more cautious and thus the label estimation is more accurate. We could also find that the gaps among the five curves are relatively small in the first several iterations and gradually enlarge in the later iterations, which shows the estimation errors are accumulated during iterations.

**Qualitative Analysis.** We visualize our selected pseudo-labeled samples for an identity during iterations in Figure 5. As shown in the left, the initial labeled sample is captured from the front view of the pedestrian, wearing a black shirt and yellow pants. At the beginning stages (iteration 1 to 4), the selected samples are very similar samples that are also captured from the front view of the pedestrian. The

TABLE III
ABLATION STUDIES W.R.T. RANK-1 AND MAP ON MARKET-1501 (IMAGE-BASED) AND MARS (VIDEO-BASED). "OURS W/O D" DENOTES THE MODEL WITHOUT THE DISSIMILARITY COST, *i.e.*, USING CLASSIFICATION LOSS AS THE CRITERION. NOTE THAT "OURS W/O D" IS OPTIMIZED ON ALL THE THREE DATA PARTS. "OURS W/O J" INDICATES THE MODEL WITHOUT THE JOINT LEARNING METHOD, *i.e.*, ONLY OPTIMIZED BY THE IDENTITY CLASSIFICATION TRAINING ON LABELED AND SELECTED PSEUDO-LABELED DATA. "OURS" IS THE FULL MODEL.

| Enlarging factor | Methods | Market-1501 | | MARS | |
|---|---|---|---|---|---|
| | | rank-1 | mAP | rank-1 | mAP |
| $p = 0.30$ | Ours w/o D | 35.2 | 13.2 | 42.0 | 20.3 |
| | Ours w/o J | 28.9 | 10.5 | 42.8 | 21.1 |
| | Ours | 35.5 | 13.4 | 44.5 | 22.1 |
| $p = 0.20$ | Ours w/o D | 36.5 | 13.7 | 45.5 | 23.5 |
| | Ours w/o J | 36.2 | 14.0 | 48.7 | 26.6 |
| | Ours | 41.4 | 17.4 | 49.6 | 27.2 |
| $p = 0.10$ | Ours w/o D | 39.8 | 16.1 | 46.4 | 24.1 |
| | Ours w/o J | 45.1 | 20.1 | 57.6 | 34.7 |
| | Ours | 51.5 | 23.2 | 57.9 | 34.9 |
| $p = 0.05$ | Ours w/o D | 40.3 | 16.2 | 48.1 | 25.2 |
| | Ours w/o J | 49.8 | 22.5 | 62.6 | 42.4 |
| | Ours | 55.8 | 26.2 | 62.8 | 42.6 |



(a)

(b)

Fig. 4. Comparison with the different value of enlarging factor $p$ on Market-1501. (a) mAP of person re-ID on the evaluation set with different enlarging factors. (b) F-score of the label prediction of selected candidates with different enlarging factors. The x-axis stands for the ratio of selected data from the entire unlabeled set. Each solid point indicates an iteration step.

above samples are relatively easy for the model to distinguish. In iteration 5, samples in the side and back views of the pedestrian are selected into the pseudo-labeled set. In later iterations, some samples from other pedestrians are selected for this identity by mistake, indicated by the red box in the figure. Although these are error cases, the selected data are very similar to the initial labeled sample that they share almost the same cloth appearance. It's clear that the samples are selected from easy to hard, from similar to diverse. There are also two back-view samples missed for this identity, *i.e.*, assigned to other identities by mistake.

### E. Evaluation on the Few-example Setting

Our method can be easily extended to the few-example re-ID task by annotating more labeled data for initialization. We report the few-example performances on the Market-1501 dataset (see Table V) and the MARS dataset (see Table IV). The performances of our method in different ratios of labeled data are reported. On the Marker-1501 dataset, our method outperforms the state-of-the-art method SPACO [30] by a large margin. On the MARS dataset, when using 20% labeled training data, our method achieves 76.5% rank-1 and 60.3% mAP, which is close to the state-of-the-art supervised methods with 100% labeled tracklets (upper bound). Although it costs more annotation effort for the few-example task comparing to the one-example task, it can easily achieve the competitive results to supervised performance with less labeled data.

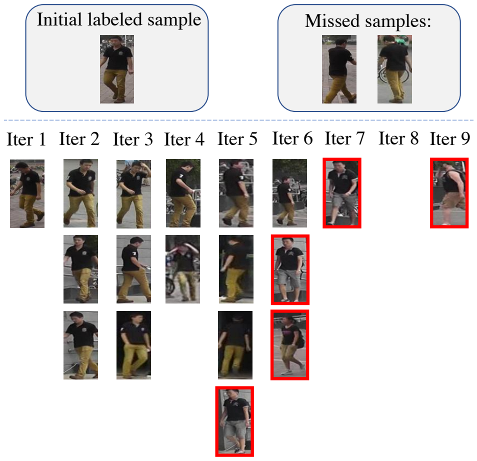


Fig. 5. The selected pseudo-labeled samples for an identity example on Market-1501. We use the enlarging factor $p$ of 0.1. Error estimated samples are in red rectangles. All the samples selected in former iterations will be selected afterward. We only show the new samples of each iteration. For this identity, two samples are missed, and five false samples are selected. The selected samples are easy and reliable at the beginning and then more difficult and diverse at a later stage.

TABLE IV
COMPARISON TO THE STATE-OF-THE-ART ON MARS. THESE WORKS ARE SUPERVISED METHODS, AND THE OTHER PERFORMANCE WE REPORTED ARE IN THE FEW-EXAMPLE SETTING. THE NUMBER IN THE BRACKET INDICATES THE PERCENTAGE OF USED LABELED TRAINING DATA.

| Settings | Methods | rank-1 | rank-5 | rank-20 | mAP |
|---|---|---|---|---|---|
| Supervised | MSCAN[41] | 71.8 | 86.6 | 93.1 | 56.0 |
| | K-reciprocal[35] | 73.9 | - | - | 68.4 |
| | IDTriplet[16] | 79.8 | 91.4 | - | 67.7 |
| Semi-supervised | Ours (10%) | 72.2 | 84.8 | 91.8 | 54.2 |
| | Ours (20%) | 76.5 | 88.4 | 93.3 | 60.3 |
| | Ours (40%) | 79.2 | 91.1 | 95.6 | 65.5 |
| | Ours (60%) | 80.6 | 91.6 | 95.7 | 66.8 |

TABLE V
COMPARISON TO THE STATE-OF-THE-ART ON MARKET-1501. THE NUMBER IN THE BRACKET INDICATES THE PERCENTAGE OF USED LABELED TRAINING DATA.

| Settings | Methods | rank-1 | rank-5 | rank-20 | mAP |
|---|---|---|---|---|---|
| Supervised | IDE [2] | 72.5 | - | - | 46.0 |
| | K-reciprocal[35] | 77.1 | - | - | 63.6 |
| | Siamese [17] | 79.5 | 90.9 | - | 59.9 |
| | GAN [3] | 83.9 | - | - | 66.0 |
| | IDTriplet[16] | 84.9 | 94.2 | - | 69.1 |
| Semi-supervised | SPACO (20%) [30] | 68.3 | - | - | - |
| | Ours (5%) | 70.1 | 84.2 | 92.1 | 43.6 |
| | Ours (10%) | 80.7 | 90.4 | 95.8 | 58.3 |
| | Ours (20%) | 82.5 | 92.4 | 97.2 | 63.6 |

## V. Conclusion

We focus on the one-example person re-ID task, where only one labeled example is labeled for each identity. We propose a progressive training framework and a joint training method. In the progressive training framework, we iteratively train the CNN model and estimate pseudo labels for the unlabeled data. For the label estimation step, we propose a progressive sampling strategy to enlarge the pseudo-labeled data set. For the model training, our proposed joint training method can effectively exploit the labeled data, the selected pseudo-labeled data, and the index-labeled data. The promising performance improvement proves the effectiveness of our method.

## References

[1] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1249 – 1258. 1, 2

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1116 – 1124. 1, 2, 3, 5, 8

[3] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *The IEEE International Conference on Computer Vision*, 2017, pp. 3774 – 3782. 1, 2, 8

[4] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1571 – 1580. 1

[5] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *IEEE International Conference on Computer Vision*, 2017, pp. 2448 – 2457. 1, 3, 6, 7

[6] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *IEEE International Conference on Computer Vision*, 2017, pp. 5152 – 5160. 1, 3, 6, 7

[7] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, 2018. 1, 3, 4

[8] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5177 – 5186. 1, 5, 6

[9] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 409–421, 2017. 2

[10] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018. 2

[11] X. Dong, Y. Yan, M. Tan, Y. Yang, and I. W. Tsang, "Late fusion via subspace search with consistency preservation," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 1, pp. 518–528, Jan 2019. 2

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152 – 159. 2

[13] Y.-G. Lee, S.-C. Chen, J.-N. Hwang, and Y.-P. Hung, "An ensemble of invariant features for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 470–483, 2017. 2

[14] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3472–3483, 2018. 2

[15] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence*, 2019. 2

[16] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. 2, 8

[17] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, p. 13, 2017. 2, 5, 8

[18] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, March 2018. 2

[19] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589. 2

[20] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 3546–3554. 2

[21] H. Ma and W. Liu, "A progressive search paradigm for the internet of things," *IEEE MultiMedia*, vol. 25, no. 1, pp. 76–86, Jan 2018. 2

[22] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, doi:10.1109/TPAMI.2018.2844853. 2

[23] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017. 2

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680. 2

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242. 2

[26] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 379–388. 2

[27] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *International Conference on Machine Learning*, 2009, pp. 41–48. 2

[28] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197. 2

[29] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Advances in Neural Information Processing Systems*, 2014, pp. 2078–2086. 2

[30] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong, "Self-paced co-training," in *International Conference on Machine Learning*, 2017, pp. 2275–2284. 2, 8

[31] J. Li, A. J. Ma, and P. C. Yuen, "Semi-supervised region metric learning for person re-identification," *International Journal of Computer Vision*, pp. 1–20, 2018. 3

[32] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1306–1315. 3

[33] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176–1190, 2019. 3

[34] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3376–3385. 4, 6

[35] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3652 – 3661. 4, 7, 8

[36] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017. 4

[37] X. Dong, D. Meng, F. Ma, and Y. Yang, "A dual-network progressive approach to weakly supervised object detection," in *ACM on Multimedia Conference*, 2017, pp. 279–287. 4

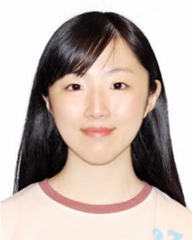
[38] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*, 2016, pp. 17–35. 5

[39] 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 *European Conference on Computer Vision*, 2016, pp. 868–884. 5

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105. 6

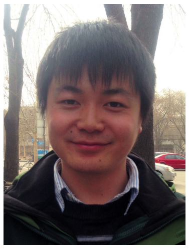
[41] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7398 – 7407. 8

**Yu Wu** received the B.E. degree from Shanghai Jiao Tong University, China, in 2015. He is currently a Ph.D. candidate in the Center for Artificial Intelligence, University of Technology Sydney, Australia. His research interests are vision language and person re-identification.



**Yutian Lin** received the B.E. degree from Zhejiang University, China, in 2016. She is currently a Ph.D. student in the Center for Artificial Intelligence, University of Technology Sydney, Australia. Her research interests are person re-identification and deep learning.

**Xuanyi Dong** received the B.E. degree in Computer Science and Technology from Beihang University, Beijing, China, in 2016. He is currently a Ph.D. student in the Center of Artificial Intelligence, University of Technology Sydney, Australia. His research interests include deep learning and its application to computer vision, especially self-supervised learning and AutoML.



**Yan Yan** obtained the Ph.D. degree at the Centre for Artificial Intelligence, University of Technology Sydney, Australia, in 2018. He received the B.E. degree in Computer Science from Tianjin University, Tianjin, China, in 2013. His current research interest includes machine learning and computer vision.

**Wei Bian** received the B.Eng. degree in electronic engineering, the B.Sc. degree in applied mathematics in 2005, and the M.Eng. degree in electronic engineering in 2007 from the Harbin institute of Technology, Harbin, China, and the Ph.D. degree in computer science from the University of Technology Sydney, Ultimo, NSW, Australia, in 2012. Currently, he is a lecturer with the University of Technology Sydney. His research interests include machine learning and computer vision.

**Yi Yang** received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.