Weakly Supervised Person Re-ID: Differentiable Graphical Learning and A New Benchmark

Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, and Liang Lin

Abstract—Person re-identification (Re-ID) benefits greatly from the accurate annotations of existing datasets (e.g., CUHK03 [1] and Market-1501 [2]), which are quite expensive because each image in these datasets has to be assigned with a proper label. In this work, we ease the annotation of Re-ID by replacing the accurate annotation with inaccurate annotation, i.e., we group the images into bags in terms of time and assign a bag-level label for each bag. This greatly reduces the annotation effort and leads to the creation of a large-scale Re-ID benchmark called SYSU-30k. The new benchmark contains $30k$ categories of persons, which is about 20 times larger than CUHK03 ($1.3k$ categories) and Market-1501 ($1.5k$ categories), and 30 times larger the ImageNet ($1k$ categories). It sums up to $29,606,918$ images. Learning a Re-ID model with bag-level annotation is called the weakly supervised Re-ID problem. To solve this problem, we introduce a differentiable graphical model to capture the dependencies from all images in a bag and generate a reliable pseudo label for each person image. The pseudo label is further used to supervise the learning of the Re-ID model. When compared with the fully supervised Re-ID models, our method achieves the state-of-the-art performance on SYSU-30k and other datasets. The code, dataset, and pretrained model will be available at https://github.com/wanggrun/SYSU-30k.

Index Terms—Weakly Supervised Learning, Person Re-identification, Deep Learning, Differentiable Graphical Learning

I. INTRODUCTION

PERSON re-identification (Re-ID) has been extensively studied in recent years [3]–[7], which refers to the problem of recognizing persons across cameras. Solving the Re-ID problem has many applications in video surveillance for public safety. Existing attempts mainly focus on learning to extract robust and discriminative representations [8], [9], and learning matching functions or metrics [8], [10]–[12] in a supervised manner. In the past four years, deep learning [13], [14] has been introduced to the Re-ID community and has achieved promising results.

However, a crucial bottleneck in building deep-learning-based models is that they typically require strongly annotated images during training. In the context of Re-ID, strong annotation refers to assigning a clear category label (i.e., person ID) for each person image, which is very expensive because it is difficult for annotators to remember persons who are strangers to the annotators, particularly when the crowd is massive. Moreover, due to the wide range of human activities, many

The first two authors contribute equally and share the first authorship. The authors are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, P. R. China. Email: wanggrun@mail2.sysu.edu.cn.; wanggc3@mail2.sysu.edu.cn; stsljh@mail.sysu.edu.cn; linliang@ieee.org; Corresponding author: Liang Lin.
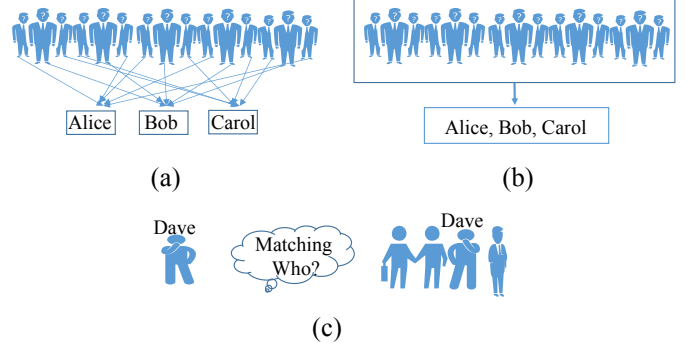

Fig. 1: Problem definition for the weakly supervised Re-ID. (a) is an example of strong annotation while (b) is an example of a weak annotation. During testing, there is no difference between the fully and weakly supervised Re-ID problems, i.e., they both aim at finding the best-matching image for a given person image, as shown in (c).

images must be annotated in a short amount of time (see Fig. 1 (a)).

An alternative way to create a Re-ID benchmark is to replace image-level annotations with bag-level annotations. Suppose that there is a short video containing many person images; we do not need to know who is in each image. A cast of characters is enough. Here, the clear ID of each image is called the image-level label (Fig. 1 (a)), and the cast of characters is called the bag-level label (Fig. 1 (b)). Based on our experience, collecting bag-level annotations is approximately three times faster/cheaper than collecting image-level annotations. Once the dataset has been collected, the goal is to train a weakly supervised Re-ID model that is as powerful as the fully supervised one. We call this the *weakly supervised Re-ID problem*.

Formally, with strong supervision, the supervised learning task is to learn $f : X \rightarrow Y$ from a training set $\{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_m, y_m)\}$, where $x_i \in X$ is a person image and $y_i \in Y$ is its exact person ID. By contrast, the weakly supervised learning task here is to learn $f : B \rightarrow L$ from a training set $\{(b_1, l_1), \cdots, (b_j, l_j), \cdots, (b_n, l_n)\}$, where $b_j \in B$ is a bag of person images, i.e., $b_j = \{x_{j1}, x_{j2}, x_{j3}, \cdots, x_{jp}\}$; and $l_j \in L$ is its bag-level label, i.e., $l_j = \{y_{j1}, y_{j2}, \cdots, y_{jq}\}$. Note that the mappings between $\{x_{j1}, x_{j2}, x_{j3}, \cdots, x_{jp}\}$ and $\{y_{j1}, y_{j2}, \cdots, y_{jq}\}$ are unknown. Furthermore, it is not necessary for the labels in $\{y_{j1}, y_{j2}, \cdots, y_{jp}\}$ to be accurate; i.e., they may be insufficient, redundant, or even incorrect. During testing, there is no difference between fully and weakly supervised Re-ID problems (see Fig. 1 (c)).
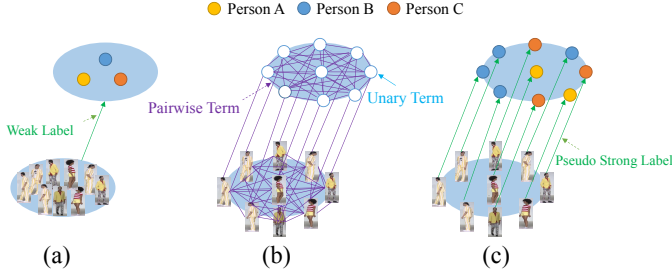
Fig. 2: An illustration of the proposed method for weakly supervised Re-ID. (a) shows a bag of images and their bag-level label. (b) represents the process of differentiable graphical learning. Using graphical modeling, we can obtain the pseudo image-level label for each image, as shown in (c).

Solving the weakly supervised Re-ID problem is challenging. Because without the help of strongly labeled data, it is rather difficult to model the dramatic variances across camera views, such as the variances in illumination and occlusion conditions, which makes it very challenging to learn a discriminative representation. Existing Re-ID approaches cannot solve the weakly supervised Re-ID problem. Regardless of whether they are designed for computing either cross-view-invariant features or distance metrics [1], [3], [4], [15]–[19], the existing models all assume that a strong annotation of each person image is available. This is also reflected in the existing benchmarking Re-ID datasets, most of which consist of a precise person category label for each image. None of them are designed to train a weakly supervised model.

Although the weak annotation lacks detailed clues for directly recognizing each person image, they usually contain global dependencies among images, which are very useful to model the variances of images across camera views. Hence, the weak annotations are as powerful as the strong annotations. Specifically, we introduce a differentiable graphical model to address the weakly supervised Re-ID problem, which includes several steps. **First**, the person images are fed into the DNNs in term of bags (Fig. 2 (a)) to obtain the rough categorization probabilities. These categorization probabilities are modeled as the unary terms in a discriminative undirected probabilistic graphical model; see Fig. 2 (b). **Second**, we further model the relations between person images as the binary terms in the graph by considering their feature similarity, their apparent similarity, and their index in different bags (representing the spatiotemporal information); see Fig. 2 (b). Note that both the unary term and the pairwise term are formulated as probabilities. These two terms are summed to form the refined categorization probability. **Third**, we maximize the refined categorization probabilities and obtain the pseudo-image-level label for each image. **Fourth**, we use the generated pseudo labels to supervise the learning of the deep Re-ID model. Note that different from traditional non-differentiable graphical models (e.g., CRFs), our proposed model is differentiable and thus can be integrated into DNNs, which is optimized by using stochastic gradient descent (SGD). All of the above steps are trained in an end-to-end fashion. We summarize the **Contributions** of this work in the following three aspects.

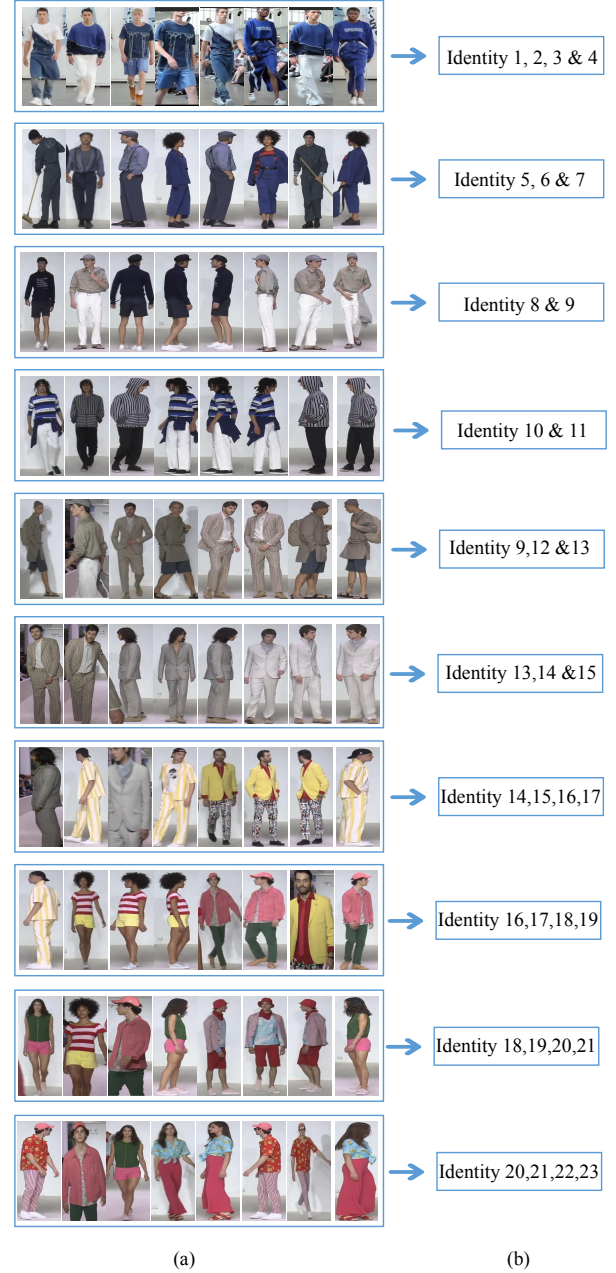**1)** We take the first step to define the unexplored weakly

Fig. 3: Examples in our SYSU-$30k$ dataset. (a) are person images in terms of bag and (b) are their bag-level weak annotations.

supervised Re-ID problem by replacing the image-level annotations in conventional Re-ID systems with bag-level annotations. This new problem is worth exploring because it significantly reduces the labor of annotation and offers the potential to obtain large-scale training data.

2) Since existing benchmarks largely ignore this weakly supervised Re-ID problem, we contribute a newly dedicated dataset called the SYSU-$30k$ for facilitating further research on the Re-ID problem. SYSU-$30k$ contains $30k$ categories of persons, which is about 20 times larger than CUHK03 ($1.3k$ categories) and Market-1501 ($1.5k$ categories), and 30 times larger than ImageNet ($1k$ categories). SYSU-$30k$ contains 29,606,918 images. Moreover, SYSU-$30k$ provides not only a large platform for the weakly supervised Re-ID problem but also a more challenging test set that is consistent with the

realistic setting for standard evaluation. Fig. 3 shows some samples from the SYSU-30k dataset.

3) We introduce a differentiable graphical model to tackle the unreliable annotation dilemma in the weakly supervised Re-ID problem. When compared with the fully supervised Re-ID models, our method achieves the state-of-the-art performance on SYSU-30k and other datasets.

The remainder of this work is organized as follows. Section II provides a brief review of the related work. Then, we introduce the annotation of SYSU-30k in Section III and follow with the weakly supervised Re-ID model in Sections IV. The experimental results and comparisons are presented in Section V. Section VI concludes the work and presents some outlooks for future work.

II. RELATED WORKS

Re-ID has been widely investigated in the literature. Most recent works can be categorized into three groups: (1) extracting invariant and discriminant features [1], [3]–[5], [17], [20]–[24], (2) learning a robust metric or subspace for matching [4], [9], [13], [15], [25], [26], and (3) joint learning of the above two methods [27]–[29]. Recently, there are many works on the generalization of Re-ID, such as video-based Re-ID [30], image-to-video Re-ID [31], spatio-temporal Re-ID [32], partial/occluded Re-ID [33], [34], and natural language Re-ID [35]. However, all these methods assume that the training labels are strong. They are thus ineffective for solving the weakly supervised learning problem in our scenario.

Another approach that is free from the prohibitively high cost of manual labeling is unsupervised learning Re-ID [36]–[41]. These methods either use local saliency matching [40], [41] or resort to clustering models [36]. However, without the help of labeled data, it is difficult to model the dramatic variances across camera views, e.g., representation learning and metric learning. Therefore, it is difficult for these pipelines to achieve high accuracies [42]–[46]. In contrast, the proposed weakly supervised Re-ID problem has a good solution. Note that compared to unsupervised Re-ID, the annotation effort of weakly supervised Re-ID is also very inexpensive.

Beyond Re-ID, although training deep models with weak annotations is a challenging problem, it has been partially investigated in the literature, such as image classification [47], [48], semantic segmentation [49]–[51], object detection [47], [52], [53] tasks. Take semantic segmentation as an example; it has exploited the advantages of weak annotation, including bounding box label [54], image-level label [49], scribble label [55] and language label [56], [57]. Our method is related to them in that our model is also based on the generation of a pseudo label. However, the weakly supervised Re-ID problem has two unique characteristics that distinguish it from other weakly supervised learning tasks. (1) We cannot find a representative image for a permanent ID because people will change their clothes at short intervals. The same person wearing different clothes may be regarded as two different persons. This results in thousands of millions of person IDs. Therefore, the label for a weakly supervised Re-ID sample is fuzzier than other tasks. (2) The entropy of the weakly supervised Re-ID problem is larger than other tasks. In the weakly

supervised segmentation task, pixels in images share certain motion of rigidity and stability, increasing the correction rate of prediction. Whereas in the case of the weakly supervised Re-ID task, persons in video bags are more unordered and irregular. Due to the above two reasons, it is considerably more challenging to re-identify a person in a weakly supervised scenario.

Apart from our model, there have been some uncertain label learning models, among which the one-shot/one-example Re-ID [58], [59] is the most related to ours. The main differences between their methods and ours are two-fold. First, in one-shot Re-ID, at least one accurate label for each person category is still in desire. While in our weakly supervised Re-ID, no accurate label is needed. Second, there are bag-level labels as constraints to guide the estimation of the pseudo labels in our method, ensuring that our generated pseudo labels to be more reliable than those generated by one-shot Re-ID. Besides, [60] also proposes to cope with the uncertain-label Re-ID problem using multiple-instance multiple-label learning. However, similar to [59], at least one accurate label for each person category is still in a desire to form the probe set in [60]. Note that mathematically, [58]–[60] are all semi-supervised Re-ID but NOT weakly supervised Re-ID.

To address the weakly supervised Re-ID problem, we propose to generate the pseudo label for each image by introducing a differentiable graphical learning [61], which is inspired by the advances in semantic image segmentation [62], [63]. Recently, one classical graphical model, i.e., conditional random field or CRF, has also been introduced to Re-ID problem for deep similarity [64]. However, our method differs from [64] in two aspects. First, like all existing methods, [64] uses CRF as a post-processing tool to refine the predictions in fully supervised learning, while our method fully exploits the supervision-independent property of graphical learning [62] to generate pseudo labels for our weakly supervised Re-ID learning. Second, different from traditional non-differentiable graphical models and [64], our proposed model directly formulates the graphical learning as an additional loss, which is differentiable to the neural network parameters and thus can be optimized by using stochastic gradient descent (SGD).

Another problem that is very related to our problem is person search [35], [70], which aims to fuse the processes of person detection and Re-ID. There are two significant differences between weakly supervised Re-ID and person search. First, the weakly supervised Re-ID only focuses on visual matching, which is reasonable because current human detectors are competent enough to detect persons. Second, the weakly supervised Re-ID problem enjoys the inexpensive efforts of weak annotation, while the person search still needs a strong annotation for each person image.

III. SYSU-30k DATASET

Data Collection. No weakly supervised Re-ID dataset is publicly available. To fill this gap, we contribute a new Re-ID dataset named SYSU-30k in the wild to facilitate studies. We download many short program videos from the Internet. TV programs are considered as our video source for two

TABLE I: A comparison of different Re-ID benchmarks. **Categories:** Each person identity is a category. **Scene:** whether the video is taken indoors or outdoors. **Annotation:** whether image-level labels are provided. **Images:** the person images which are obtained by using a human detector to detect the video frames. Actually, the person images in this work refer to the bounding boxes.

(a) Comparision with existing Re-ID datasets.

Dataset	CUHK03 [1]	Market-1501 [2]	Duke [65]	MSMT17 [66]	CUHK01 [67]	PRID [68]	VIPeR [4]	CAVIAR [69]	SYSU-30k
Categories	1,467	1,501	1,812	4,101	971	934	632	72	**30,508**
Scene	Indoor	Outdoor	Outdoor	Indoor, Outdoor	Indoor	Outdoor	Outdoor	Indoor	**Indoor, Outdoor**
Annotation	Strong	Strong	Strong	Strong	Strong	Strong	Strong	Strong	**Weak**
Cameras	2	6	8	15	10	2	2	2	**Countless**
Images	28,192	32,668	36,411	126,441	3,884	1,134	1,264	610	**29,606,918**

(b) Comparison with ImageNet-1k

Dataset	ImageNet-1k	SYSU-30k
Categories	1,000	**30,508**
Images	1,280,000	**29,606,918**
Annotation	Strong	**Weak**

reasons. **First**, the pedestrians in a TV program video are often cross-view and cross-camera because **1)** the scenes in TV program videos are generally recorded by many cameras for post-processing and **2)** the cameras in a program are generally movable for following shots. Therefore, identifying the pedestrians in a TV program video is exactly a Re-ID problem in the wild. **Second**, the number of pedestrians in a program is suitable for annotation, i.e., neither too many nor too few. On average, each video contains 30.5 pedestrians walking around.

Our final raw video set contains 1,000 videos. The annotators are then asked to annotate the persons in the video in a weak fashion. In particular, each video is divided into 84,924 bags of arbitrary length. Then, the annotators record the pedestrians identity for each bag. YOLO-v2 [71] is utilized for pedestrian bounding box detection. Three annotators review the detected bounding boxes and annotate person category labels for 20 days. Finally, 29,606,918 ($\approx 30M$) bounding boxes of 30,508 ($\approx 30k$) person categories are annotated. We then select 2,198 identities as the test set, leaving the rest as the training set. There is no overlap between the training set and the test set.

Dataset Statistics. SYSU-30k contains 29,606,918 person images with 30,508 categories in total, which is further divided into 84,930 bags (only for training set). Fig. 4 (a) summarizes the number of bags with respect to the number of images per bag, showing that each bag has 2,885 images on average. This histogram reveals the person image distribution of these bags in the real world without any manual cleaning and refinement. Each bag is provided with an annotation of bag-level labels.

Comparison with Existing Re-ID Benchmarks. We compare SYSU-30k with existing Re-ID datasets, including CUHK03 [1], Market-1501 [2], Duke [65], MSMT17 [66], CUHK01 [67], PRID [68], VIPeR [4], and CAVIAR [69]. Fig. 4 (c) and (d) plots the person categories and the number of images, respectively, indicating that SYSU-30k is much larger than existing datasets. To evaluate the performance of the weakly supervised Re-ID approach, we randomly choose 2,198 person categories from SYSU-30k as the test set. These person categories are not utilized in training. We annotate an accurate person ID for each person image. We also compare the test set of SYSU-30k with existing Re-ID datasets. From

Fig. 4 (b) and (c), we can observe that the test set of SYSU-30k is more challenging than those of the competitors in terms of both the image number and person categories. Thanks to the above annotation fashion, the SYSU-30k test set can adequately reflect the real world setting and is consequently more challenging than existing Re-ID datasets. Therefore, SYSU-30k is not only a large benchmark for the weakly supervised Re-ID problem but is also a significant standard platform for evaluating existing fully-supervised Re-ID methods in the wild.

A further comparison of SYSU-30k with existing Re-ID benchmarks is shown in Table I (a), including categories, scene, annotation, cameras, and image numbers (bounding boxes). After the comparison, we summarize the new features in SYSU-30k in the following aspects. First, SYSU-30k is the first weakly annotated dataset for Re-ID. Second, SYSU-30k is the largest Re-ID dataset in terms of both person categories and image number. Third, SYSU-30k is more challenging due to many cameras, realistic indoor and outdoor scenes, and occasionally incorrect annotations. Four, the test set of SYSU-30k is not only suitable for the weakly supervised Re-ID problem but is also a significant standard platform to evaluate existing fully supervised Re-ID methods in the wild. Fig. 3 shows some training samples in the SYSU-30k dataset, and Fig. 5 shows some testing samples.

Comparison with ImageNet-1k. Beyond the Re-ID family, we also compare SYSU-30k with the well-known ImageNet-1k benchmark for general image recognition. As shown in Table I (b), SYSU-30k has several appealing advantages over ImageNet-1k. First, SYSU-30k has more object categories than ImageNet-1k, i.e., 30k vs 1k. Second, SYSU-30k has a greater number of images by 1-2 orders of magnitude than ImageNet-1k. Third, SYSU-30k saves annotation due to the effective weak annotation.

Evaluation Protocol. The evaluation protocol of SYSU-30k is similar to that of the previous datasets [2]. As SYSU-30k dataset is quite large, we do not need to repeat random partitioning the dataset into a training set and test set for ten times [1]. Instead, we fix the train/test partitioning. In the test set, each person category will have one probe, resulting in 1,000 probes. As the scalability is of most importance for the practicability of Re-ID systems, we propose to challenge
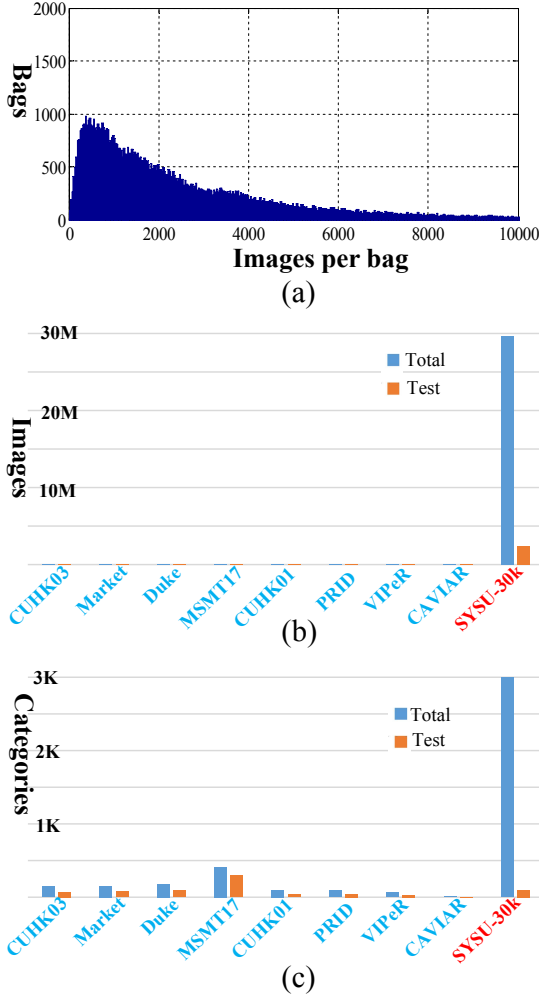
Fig. 4: The statistics of the SYSU-30k. (a) summarizes the number of the bags with respect to the number of the images per bag. (b) and (c) compare SYSU-30k with the existing datasets in terms of image number and person categories, respectively, for both the whole dataset and the test set.

the scalability of a Re-ID model by providing a gallery set containing a vast volume of distractors for validation. Specifically, for each probe, there is only one matching person image as the correct answer in the gallery, while there are 478,730 mismatching person images as the wrong answer in the gallery. Thus, the evaluation protocol is to search for a needle in the ocean. This is consistent with the practicability of Re-ID tasks because the police usually need to search a massive amount of videos for a criminal. Then, given a query image sequence, all gallery items are assigned a similarity score. We then rank the gallery according to their similarity to the query, based on which we calculate the CMC metric which represents the expectation of the true match being found within the first n ranks, following [2].

IV. WEAKLY SUPERVISED RE-ID MODEL

We aim at learning a Re-ID model using weak supervision by exploiting the dependencies among the person images. We first discuss the supervision in the traditional supervised Re-ID and the weakly supervised Re-ID (Section IV-A), then present a solution for the weakly supervised Re-ID using differentiable
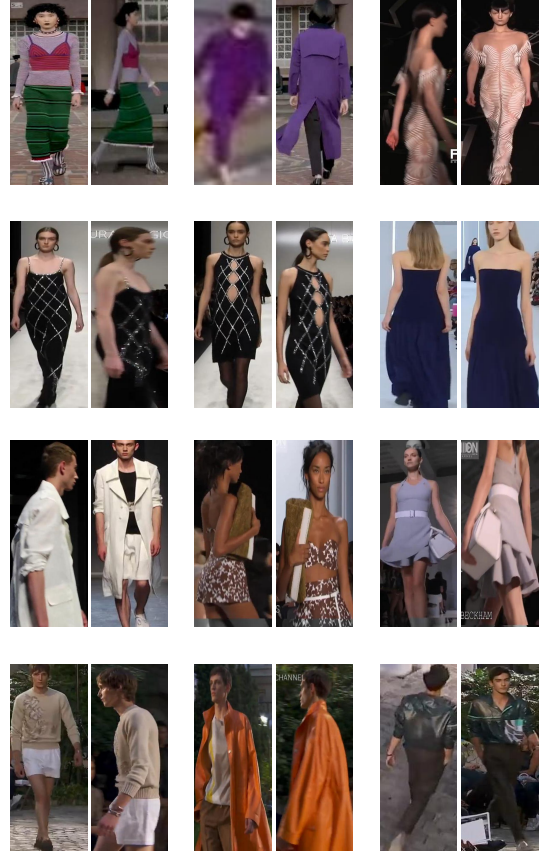

Fig. 5: Examples in the test set of SYSU-30k. Each pair represents a pair of images belonging to the same person category, but taken by different cameras. **Left:** query images; **Right:** gallery images.

graphical modeling (Section IV-B). The network architecture and implementation details are presented in Section IV-C and Section IV-D. Next, the computational complexity of our method is presented in Section IV-E. Finally, we discuss the relationship of our work to previous works in Section IV-F

A. From Supervised Re-ID to Weakly Supervised Re-ID

The training data for DNNs are usually organized in batches, which allows us to organize several bags of person images in a batch. Each bag has a flexible number of images. Hence, abundant inter-image relations and dependencies can be fully exploited to discover useful supervision information.

Let b denote a bag containing p images, i.e., $b = \{x_1, x_2, \cdots, x_j, \cdots, x_p\}$; $y = \{y_1, y_2, \cdots, y_j, \cdots, y_p\}$ are the image-level labels; while l denotes the bag-level label. In a fully supervised Re-ID problem, the image-level labels y are known. The goal of fully supervised learning is to learn the model by minimizing the error between the category prediction and the image-level label for each person image.

On the contrary, in a weakly supervised Re-ID problem, although the bag-level label l is provided, the image-level labels y are unknown. One possible solution is to estimate a pseudo image-level label $\hat{y}$ for each person image. Intuitively,

we can first obtain an image-level label in the form of a probabilistic vector (denoted as $\mathbf{Y}$) for each image from the bag-level label. Suppose l contains n categories of person, and in total there are m person categories in the training set. Then the preliminary image-level for each person image x_j can be deduced as the following:

$$
\mathbf{Y}_j = \begin{pmatrix} \mathbf{Y}_j^1 \\ \vdots \\ \mathbf{Y}_j^k \\ \vdots \\ \mathbf{Y}_j^m \end{pmatrix}, where \quad \mathbf{Y}_j^k = \begin{cases} \frac{1}{n}, & if \quad k \in l \\ 0, & otherwise \end{cases},
$$
(1)

Eqn. 1 reveals the restricting role of a bag-level label. Therefore, in the following we refer to Eqn. 1 as **bag constraint** for simplification. By fully exploiting the bag constraint and the dependencies among the images in a bag, we can further deduce the final pseudo-image-level labels $\hat{y}$ from the preliminary image-level labels $\mathbf{Y}$. Then, $\hat{y}$ are leveraged to supervised the learning of the model in the same manner as the fully supervised learning.

A Re-ID problem is different from an image classification problem because the training set and the test set in a Re-ID problem do not share the person categories. As a result, the similarity between the probe images and the gallery images must be measured. Let x_i be a probe image and x_j be a gallery image. The similarity of x_i and x_j is measured by calculating the Euclidean distance between the features of x_i and x_j learned by the DNNs.

B. Weakly Supervised Re-ID with Differentiable Graphical Learning

In this section, we will discuss the mechanism and formulation of using differentiable graphical learning to generate pseudo-image-level labels for the person images.

Graphically Modeling Re-ID. Our graph is a directed graph in which each node represents a person image x_i in a bag, and each edge represents the relation between person images, as illustrated in Fig. 6. Assigning a label y_i to each node x_i will have a cost. For example, imposing the labels *'Person 1', 'Person 2', and 'Person 3'* to x_1, x_2 and x_3 leads to an energy cost of $E(y_1 = 1; y_1 = 2; y_3 = 3 \mid x_1; x_2; x_3)$, which is abbreviated as $E(y_1; y_2; y_3 \mid x_1; x_2; x_3)$ or $E(y|x)$ for notation simplification. Let i denote an image index with respect to a bag. Formally, the energy function of our graph is defined as

$$
E(y|x) = \underbrace{\sum_{\forall i \in U} \Phi(y_i|x_i)}_{\text{unary term}} + \underbrace{\sum_{\forall i,j \in V} \Psi(y_i, y_j|x_i; x_j)}_{\text{pairwise term}}, \quad (2)
$$

where U and V denote a set of nodes and edges, respectively. $\Phi(y_i|x_i)$ is the unary term measuring the cost of assigning label y_i to a person image x_i. For instance, if an images belongs to the first category rather than the second one, we should have $\Phi(y_i = 1 \mid x_i) < \Phi(y_i = 2 \mid x_i)$. Moreover, $\Psi(y_i, y_j|x_i; x_j)$ is the pairwise term that measures the penalty of assigning labels to a pair of person images
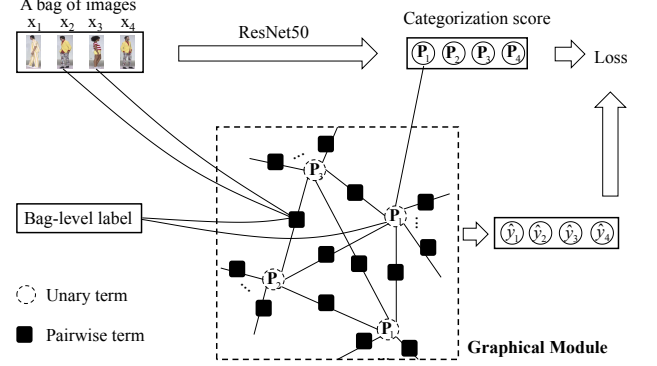

Fig. 6: Graphical model for the generation of pseudo image-level labels in a bag of person images. The unary terms are estimated by the deep neural networks, while the pairwise terms are obtained by considering the similarity of features, the raw image appearance, and the bag-level label.

(x_i, x_i), respectively. Mathematically, graphical modeling is employed to smooth noisy (uncertain) person ID prediction. The unary term in Eqn. 2 performs the prediction based on sole nodes. While the pairwise term in Eqn. 2 couples different nodes, favoring same-label assignments of nodes that are bag index proximal and similar in appearance. In summary, Eqn. 2 is to clean up the spurious predictions of classifiers learned in a weakly supervised manner.

Unary Term. Intuitively, the unary terms represent per-image classifications. The unary term in Eqn. 2 is typically defined as

$$
\Phi(y_i|x_i) = -\log(\mathbb{Y}_i[y_i]), \text{ where } \mathbb{Y} = \underbrace{\mathbf{Y}_i}_{\text{bag constraint}} \odot \underbrace{\mathbf{P}_i}_{\text{DNN output}},
$$
(3)

where $\mathbf{P}_i$ is the label assignment probability for the person image x_i as computed by a DNN. $\mathbf{Y}_i$ is the preliminary image-level label defined in Eqn. 1, indicating the estimation is subjected to the bag-level label, as illustrated in Fig. 6. Here, $\odot$ denotes element-wise product, and $[\cdot]$ denotes vector indexing.

The maximum a posteriori (MAP) labeling is good enough to be a candidate pseudo label due to the capacity of the DNNs. However, as the output of the unary classifier for each image is produced independently from the outputs of the classifiers for the other images, the unary term alone is generally noisy and inconsistent. Interactions between pairwise terms are required.

Pairwise Term. The pairwise terms represent a set of smoothness constraints. As in [61], we use the following expression:

$$
\Psi(y_i, y_j|x_i; x_j) = \underbrace{\zeta(y_i, y_j)}_{\text{label compatibility}} \underbrace{\mathbf{Y}_i[y_i]\mathbf{Y}_j[y_j]}_{\text{bag constraint}} \underbrace{\exp\left(-\frac{\|I_i - I_j\|^2}{2\sigma^2}\right)}_{\text{appearance similarity}}
$$
(4)

where a Gaussian kernel depending on RGB colors that measure the appearance similarity is used. The hyper parameter σ control the scale of the Gaussian kernels. The kernel forces person images with similar color and deep features to have the same labels. Similar to the unary term, the pairwise terms

are also bounded by the bag-level annotations $\mathbf{Y}_i$ and $\mathbf{Y}_j$, enabling more reliable estimations. The pairwise terms are widely known to improve accuracy, indicating that they can provide nontrivial knowledge (e.g., structural context dependencies) that is not captured by the unary term. A simple label compatibility function $\zeta(y_i, y_j) \in \{0, 1\}$ in Eqn. 4 is given by the Potts model, i.e.,

$$\zeta(y_i, y_j) = \begin{cases} 0, \text{if } y_i = y_j \\ 1, \text{otherwise} \end{cases}, \qquad (5)$$

which introduces a penalty for similar images that are assigned different labels. While the simple model in Eqn. 2 works well in practice, it is non-differentiable and thus is incompatible with DNNs. We can instead learn a differential version of Eqn. 2 that takes the deep model into account, as described in the following.

Bag Constraint. As mentioned above, both the unary and pairwise terms are constrained by the bag-level annotations $\mathbf{Y}_i$ and $\mathbf{Y}_j$. In fact, the bag-level annotation contains extra knowledge that helps to improve the estimation. For example, if the estimator mismatches a person image to a category that is not in the bag-level annotation, the estimation is undoubtedly supposed to be incorrect. Then, the estimation will be corrected by matching the person image to the category in the bag-level annotation with the most significant prediction score. Furthermore, if some person categories in the weak annotation are absent in the prediction, the proposed method will encourage a portion of the person images to be assigned to such categories to improve the performance. In this way, knowledge of the weakly labeled data can be fully exploited.

Specifically, given a bag of images and their bag-level label, we refine the DNN predictions by element-wise multiplication of $\mathbf{P}$ by the bag-level weak annotation $\mathbf{P}$, which is shown in the unary term in Eqn. 3. Similarly, we also impose $\mathbf{Y}_i$ and $\mathbf{Y}_j$ in the pairwise term in Eqn. 4.

One may argue that it is difficult to achieve perfect performance using bag-level labels because the mapping from input vectors to output vectors is ambiguous. However, there is a natural smoothness assumption in videos that could be ignored: person IDs in bags change slowly within a short time, e.g., an image-level label y_i in bag b_T could also be in bag b_{T+1}. A large amount of bags with overlapping IDs naturally exist in a video and thus partially disclose the underlying mapping from input vectors to output vectors, which sheds light on the competitive performance of the weakly supervised Re-ID. As a special example, if b_T contains $\{y_i, y_j\}$ and b_{T+1} contains $\{y_j, y_k\}$, then the two bags share $\{y_j\}$. In this case, our method can easily know which image in b_T belongs to $\{y_i\}$ and which image in b_{T+1} belongs to $\{y_k\}$.

Deduction of Pseudo Image-level Labels. By minimizing the Gibbs energy of Eqn. 2, we can obtain the pseudo image-level label for each person image, i.e.,

$$\hat{y}_i = \underset{y_i \in \{1,2,3,\cdots,m\}}{\arg\max} E(y_i|x_i), \qquad (6)$$

where $\{1, 2, 3, \cdots, m\}$ denotes all the person categories in the training set. Here $\hat{y}_i$ is the final pseudo image-level label
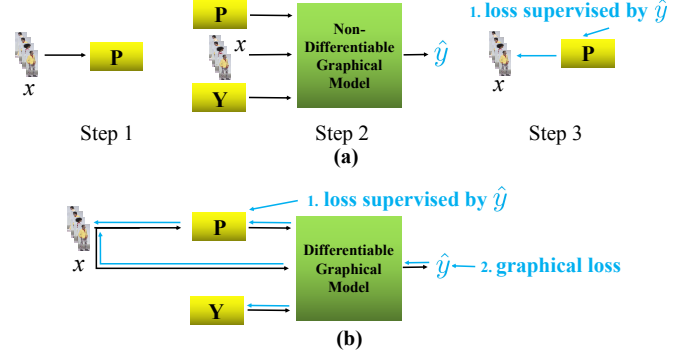

Fig. 7: Differentiable graphical modeling with deep neural networks, where $x, \mathbf{Y}, \mathbf{P}, \hat{y}$ denote the input images, bag-level label, preliminary categorization and refined categorization, respectively. (a) is the stepwise graphical modeling for the weakly supervised Re-ID model, while (b) is our proposed end-to-end differentiable graphical model. The implementation of our differentiable graphical model consists of two losses, i.e., an unsupervised loss for pseudo label estimation and a loss supervised by the pseudo labels. Here black lines denote forward-propagation, while blue lines denote back-propagation.

generated by our approach. Once such labels are generated, they can be used to update the network parameters as if they were authentic ground-truth labels.

Differentiablizing the Graphical Learning. The above weakly supervised Re-ID model is not end-to-end. Because we must first use an external graphical learning solver to obtain the pseudo labels and then use another solver to train the DNNs under the supervision of the pseudo labels (see Fig. 7 (a)). To enable an end-to-end optimization, we propose to make our graphical learning differentiable and compatible with DNNs (see Fig. 7 (b)).

We first investigate the mechanism of a non-differentiable graphical model. As is illustrated in Fig. 7 (a), a non-differentiable graphical model consists of three steps. **First**, the preliminary categorization score $\tilde{y}$ is obtained through a DNN. **Second**, the Gibbs energy in Eqn. 2 is minimized by appropriately (optimally) re-assigning labels to the images, subject to the apparent similarity, the preliminary categorization scores, and the bag constraint. **Third**, the re-assigned labels are considered as the pseudo labels and are used to supervise the learning of the Re-ID model.

Assigning labels in the second step listed above is non-differential, which makes the graphical model incompatible with the DNN. To fill this gap, a relaxation form of Eqn. 2 is in desire. Specifically, Eqn. 2 is rewritten as:

$$\mathbf{L}_{graph}(x) = \underbrace{\sum_{\forall i \in U} \hat{\Phi}(x_i)}_{\text{unary term}} + \underbrace{\sum_{\forall i,j \in V} \hat{\Psi}(x_i, x_j)}_{\text{pairwise term}}, \qquad (7)$$

Where we use a continuous version of $\hat{\Phi}$ and $\hat{\Psi}$ to approximate the discrete Φ and Ψ. Formally, $\hat{\Phi}$ and $\hat{\Psi}$ are defined as:

$$\hat{\Phi}(x_i) = -\log(\underset{k \in \{1,2,3,\cdot,m\}}{\arg\max} \mathbf{Y}_i[k] \odot \mathbf{P}_i[k]), \qquad (8)$$

$$\hat{\Psi}(x_i, x_j) = -\exp\left(-\frac{\|I_i - I_j\|^2}{2\sigma^2}\right)(\mathbf{Y}_i\mathbf{P}_i)^{\mathbf{T}}\log(\mathbf{Y}_j\mathbf{P}_j). \qquad (9)$$
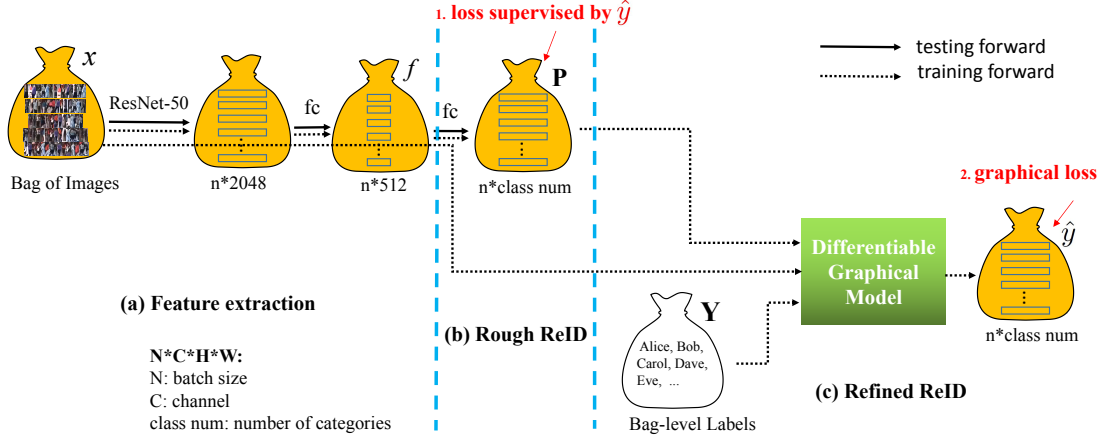
Fig. 8: Diagram of our approach, which consists of three main stages, i.e., feature extraction, rough Re-ID, and refined Re-ID by the differentiable graphical module. The solid black flow denotes the testing stage, while the black dotted flow denotes the training stage. For simplification, the back-propagation flow is omitted. The loss function is marked with a red arrow.

The differences between Eqn. 3 and Eqn. 8 are summarized as follows: **1)** The y_i in $\Phi(y_i)$ is replaced with the x_i in $\hat{\Phi}(x_i)$. In the non-differentiable graphical model, all possible ys are feed into the energy function. The y which leads to lowest energy will be considered as the optimal solution. Differently, in a differentiable graphical model, we feed the images x into the DNN and obtain the prediction y. **2)** We use an $\arg\max$ function to obtain the prediction, which is consistent with the nature of DNNs, i.e., during the testing phase, we directly obtain the prediction from the output of the DNN without the graphical losses. Besides, there is one more difference between Eqn. 4 and Eqn. 9. We use an cross-entropy term $-(\mathbf{Y}_i\mathbf{P}_i)^{\mathbf{T}} \log(\mathbf{Y}_j\mathbf{P}_j)$ to approximate the non-differential term $\zeta(y_i, y_j)\mathbf{Y}_i\mathbf{Y}_j$ in Eqn. 4.

C. Overall Neural Network Architecture

The network architectures for training and testing are shown in Fig. 8, where the black dotted lines denote training flow, and the solid black lines denote inference flow. It is noteworthy that we only perform graphical modeling in the training stage for two reasons. **First**, the graphical module is introduced to generate pseudo labels to supervise the training, which requires a bag-level label as a constraint. However, there is no bag-level label in the testing stage. **Second**, due to the specificity of the Re-ID problem, the images in the inference stage are not organized in the form of a bag. For example, only a query image and a set of gallery images are provided in inference, requiring the Re-ID systems to calculate the similarity between them. As a result, there is no bag-level dependency among the testing images to exploit. Thus, performing graphical modeling may be infeasible in the inference stage.

The implementation of our weakly supervised Re-ID model consists of three main modules, including **(a)** a feature embedding module built upon a ResNet-50 network followed by two fully connected layers, **(b)** a rough Re-ID module using a fully connected layer as the classifier, and **(c)** a refined Re-ID module that considers both the rough results and bag-

level weak annotation to perform graphical modeling. These modules are shown in Fig. 8.

Feature Embedding Module. Many current best-performing Re-ID models use multi-scale features as feature embeddings [72], which guarantees a robust feature representation and thus boosts the performance. However, in this work, our focus is the mechanism of the weakly supervised Re-ID model alone, rather than other tricks. Therefore, we simply take the ResNet-50 [73] as the backbone without any feature pyramid [72]. Our feature embedding is similar to [74]. Specifically, the last layer of the original ResNet-50 is discarded, and two new, fully connected layers are added. The first has 512 units, followed by a batch normalization [75], a Leaky ReLU [76] and a dropout [77]. This module is shown in Fig. 8 (a).

Rough Re-ID Module. To investigate the behavior of the weakly supervised Re-ID alone, we use the standard softmax classifier rather than triplet similarity [74] for rough Re-ID. Specifically, our model has another fully connected layer at the top of the feature embedding module, which has the same units as the person categorization numbers (denoted as 'class num' in Fig. 8). Then, a softmax cross-entropy loss is employed. The person categorization score (e.g., y in Fig. 8) is considered as the rough Re-ID estimation, indicating the possibility of a person ID being present in a bag b. This module is shown in Fig. 8 (b).

Refined Re-ID Module. Here, we aim to estimate a pseudo-image-level label for each person image by refining the previous estimation results. The refinement benefits from three aspects:

1) *Rough Re-ID score.* As mentioned above, the rough Re-ID module returns a preliminary categorization result, which can be considered as a baseline for further improvement.

2) *Appearance.* Although the rough Re-ID score is taken into consideration, it is a high-level abstraction of the images that lack details. As compensation, we propose to include more low-level information (i.e., RGB appear-

ance) in our refinement.

3) *Bag constraint.* Finally, we consider the bag-level labels. Intuitively, we eliminate any possibility of assigning a person image with a person category that is absent in the bag-level annotation. By contrast, we encourage a person image to be assigned with a person category that is present in the bag-level annotation.

The above three aspects are fed into our graphical modeling, as shown in Eqn. 8 and 9. Once such labels are generated, they can be used to update the network parameters as if they were authentic ground-truth labels. The diagram of our approach is presented in Fig. 8.

D. Optimization and Implementation Details

The optimization of our model is a joint process of estimating pseudo labels and solving the DNN model supervised by the pseudo labels. Once the pseudo labels are obtained, the weakly supervised Re-ID problem becomes a fully supervised one. Specifically, given the pseudo person IDs, we can compute the gradient of the overall losses with respect to the DNN parameters. With the back-propagation algorithm, the gradients from the loss propagate backward through all layers of the DCNN. Thus, all parameters of our weakly supervised model can be learned in an end-to-end manner.

Loss Function. The optimization object of our approach consists of two loss functions, including a graphical modeling loss $\mathbf{L}_{graph}$ and a classification/re-identification $\mathbf{L}_{cls}$, as illustrated in Fig. 7 (b), where the back-propagation is represented with blue lines. $\mathbf{L}_{cls}$ can be a simple softmax cross-entropy loss with the pseudo label $\hat{y}$ as supervision.

$$\mathbf{L}_{cls} = -\sum_{i=1}^{n}(g_{\text{one hot}}(\hat{y}_i))^{\mathbf{T}}\log(\mathbf{P}_i), \qquad (10)$$

where $g_{\text{one hot}}(\hat{y}_i)$ denotes a function that transforms $\hat{y}_i$ into a one-hot vector, and n denotes the image number in a bag. Here $\mathbf{P}_i$ denotes the preliminary categorization probability, which is the logits normalized by the softmax function, i.e.,

$$\mathbf{P}_i^k = \frac{\exp(z_k)}{\sum_{k=1}^{m}\exp(z_k)}, \qquad (11)$$

where m denotes the person category number in the training set and z is the output logit.

The combination of these two loss functions is a simple linear combination with predefined loss weights. In our implementation, the loss weights are set as 1 : 0.5. We have the total loss $\mathbf{L}$:

$$\mathbf{L} = w_{cls}\mathbf{L}_{cls} + w_{graph}\mathbf{L}_{graph}, \qquad (12)$$

where w_{cls} and w_{graph} denote the two loss weights.

Bag organization. In our implementation, an image batch b contains images of n person categories, and each person category has k images. With the image bags in each batch, we can perform graphical modeling to capture the image dependencies in a bag, thus enabling the weakly supervised learning.

Other Implementations Details. As mentioned in Section IV-C, our approach employs ResNet-50 as the backbone, where the parameters are initialized by classifying one-thousand image classes in ImageNet. The other parameters are initialized by sampling from a normal distribution. For SGD, we use a minibatch of 90 images and an initial learning rate of 0.01 (0.1 for the fully connected layer), multiplying the learning rate by 0.1 after a fixed number of iterations. We use the momentum of 0.9 and a weight decay of 0.0005. The training on SYSU-30k takes approximately ten days on a single GPU (i.e., NVIDIA TITAN X). During training, all of the images are resized to 288×144 and cropped to 256×128 at the center with a small random perturbation. Random mirroring is also adopted in our experiments.

E. Computational Complexity

We provide more discussion of computational complexity on our weakly supervised Re-ID model. The extra introduced time cost of our method is negligible for two reasons. **1)** In the training phase, the extra introduced time cost only relates to the estimation of pseudo labels, which is a graphical learning module. Conventionally, graphical learning needs many iterations to find the solution, and thus, the process is time-consuming. However, our approach formulates the differentiable graphical learning as an inherent part of the DNN. Therefore, in each training step of the DNN, there is only one iteration of inference in our graphical module, which is consistent with the back-propagation algorithm. This makes our graphical module very effective. In the experimental section, we will show that our training brings an additional time cost of only $0.004\times$. **2)** In the testing phase, there is no extra time cost when the pseudo label estimation component is disabled.

F. Relationship to Previous Works

In the following, we compare our weakly supervised Re-ID with previous works on Re-ID with uncertain labels, including the unsupervised Re-ID and the semi-supervised Re-ID. In general, we see that our weakly supervised Re-ID not only possess cheap annotation effort but also achieves high identification accuracy. The details are presented below.

Unsupervised Re-ID. To get rid of the prohibitively high cost of manual labeling, unsupervised learning Re-ID [36]–[41] proposes to use either local saliency matching [40], [41] models or clustering models [36]. However, without the help of labeled data, it is difficult to model the dramatic variances across camera views, e.g., representation learning and metric learning. Therefore, it is difficult for these pipelines to obtain high accuracies [42]–[46]. In contrast, our weakly supervised Re-ID problem has a good solution. Note that compared to unsupervised Re-ID, the annotation effort of our weakly supervised Re-ID is also very inexpensive.

Semi-supervised Re-ID. One-shot/one-example [58], [59] propose to reduce the annotation effort by annotating only one example for each person ID. The main differences between their methods and ours are two-fold. First, in one-shot Re-ID, at least one accurate label for each person category is

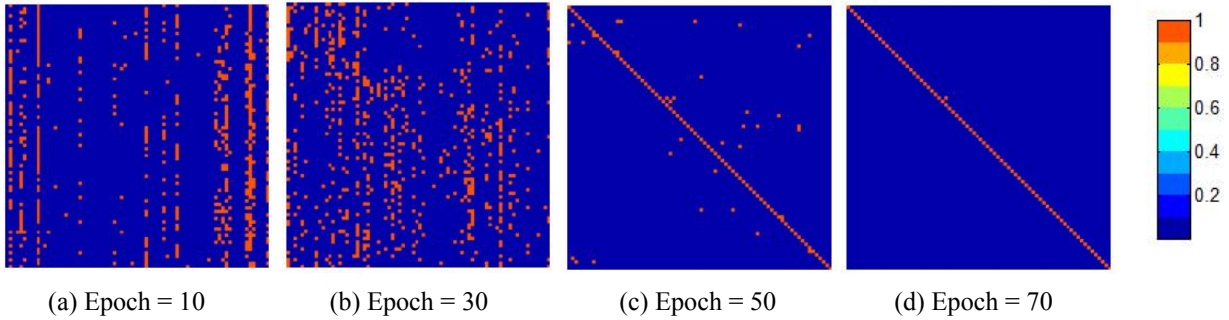(a) Epoch = 10   (b) Epoch = 30   (c) Epoch = 50   (d) Epoch = 70

Fig. 9: The effectiveness of our differentiable graphical learning module. Here we show the errors between the rough predictions and the weak annotations in the form of a confusion matrix containing 76×76 grids. Each grid indicates a bag of 10 categories, with a total sum of 760 categories, which is approximately equivalent to the person categories in the full training set (i.e., 767 categories).

in desire. While in our weakly supervised Re-ID, no accurate label is needed. Second, there is bag-level label as a constraint to estimate the pseudo labels in our method, ensuring that our generated pseudo labels to be more reliable than those generated by one-shot Re-ID.

We would also like to acknowledge the contribution of previous work [60] that matches a target person image with a bag-level gallery video using multiple-instance multiple-label learning. However, similar to [59], at least one accurate label (of the target person) for each person category is still in a desire to form the probe set in [60]. Hence, mathematically, [60] still belongs to semi-supervised Re-ID but NOT weakly supervised Re-ID.

Experimentally, Section V-B2 and V-B3 will compare the performance of our weakly supervised Re-ID with previous works.

V. EXPERIMENTS

We evaluate the weakly supervised Re-ID approach in two aspects. Section V-A conducts an extensive ablation study, including the effectiveness of the differentiable graphical learning module, the scalability of our method, the impact of bag diversity, and the compatibility with fully supervised learning tricks. Section V-B compares the Re-ID accuracy with state-of-the-art methods and analyzes the computational complexity.

Two Simulated Datasets In addition to the proposed SYSU-$30k$ dataset, another two simulated datasets are introduced to evaluate the effectiveness of our method by adjusting the existing datasets. Specifically, we replace the strong annotations on the training set of the PRID2011 [68] and CUHK03 [1] datasets with weak annotations while their test sets are kept unchanged. For a fair comparison (e.g., using the same images for both fully and weakly supervised learning), we generate the weak annotations from the strong annotations. This includes two steps. **1)** Each bag is simulated by randomly selecting several images and packaging them. **2)** Then, the weak labels are easily obtained by summarizing the strong annotations, e.g., four image-level labels {Alice, Bob, Alice, Carol} are summarized as a bag-level label {Alice, Bob, Carol}. We denote n categories/bag when a bag contains

n person categories. Note that unless otherwise stated, our weakly supervised learning setting is two categories/bag.

Originally, PRID2011 [68] contains 200 person categories appearing in at least two camera views and is further randomly divided into training/test sets following the general settings [31], i.e., both having 100 categories. The CUHK03 dataset [1] is one of the largest databases for Re-ID. This dataset contains 14,096 images of 1,467 pedestrians collected from 5 different pairs of camera views. Each person is observed by two disjointed camera views, which have an average of 4.8 images in each view. We follow the **new** standard setting [78] of using CUHK03 [1], and a training set (including 767 persons) is obtained without overlap. For the training sets of both the PRID2011 and CUHK03 benchmarks, person categories are further packed into bags, and bag-level labels are extracted from the image-level labels. This enables us to examine the proposed weakly supervised Re-ID problem. Note that the test sets of the two datasets are the same as the original ones, as the definition states that during testing, there is no difference between fully and weakly supervised Re-ID problems (Fig. 1 (c)).

Market-1501. In addition to PRID2011 and CUHK03, we also compare our method with existing state-of-the-art methods on Market-1501 dataset [2]. Market-1501 is another widely-used large-scale Re-ID benchmark, which contains 32,668 images of 1,501 pedestrians captured from 6 different cameras. The dataset is split into two parts: 12,936 images with 751 identities for training and 19,732 images with 750 identities for testing. In testing, 3,368 hand-drawn images with 750 identities are used as probe set to identify the true identities on the testing set.

Evaluation Metric For PRID2011, CUHK03, and Market-1501, the test set is further divided into a gallery set of images and a probe set. We use the cumulative matching characteristic (CMC) [79] as the evaluation metric. For SYSU-$30k$, the evaluation metric is described in Section III.

A. Ablation Study

First, we present ablation studies to reveal the benefits of each main component of our method.

TABLE II: Ablation studies of the proposed weakly supervised Re-ID method. **random:** Each bag contains random categories of persons, which reflects the real-world state. **reranking:** see [78], one of the effective tricks frequently used in fully supervised Re-ID problems. ⋆ **fully supervised:** when each bag contains only one category, the weakly supervised Re-ID problem degrades into a fully supervised Re-ID problem. ∗**full training set:** the overall training set of CUHK03 contains 767 person categories.

(a) Impact of Bag Diversity on PRID2011

categories / bag	Rank-1	Rank-5	Rank-10
1 (⋆ **fully supervised**)	71.8	91.2	95.9
2	68.0	87.5	94.8
3	66.1	86.4	92.3
10	49.5	73.9	82.2
random	69.3	89.0	94.0

(b) Impact of Bag Diversity on CUHK03

categories / bag	Rank-1	Rank-5	Rank-10
1 (⋆ **fully supervised**)	67.5	88.2	91.8
2	61.0	82.0	87.0
3	59.4	80.7	86.7
10	55.2	79.3	84.5
random	60.6	81.6	87.0

(c) Fully supervised learning Tricks on PRID2011

method	Rank-1	Rank-5	Rank-10
fully supervised alone	48.9	79.6	88.8
weaky supervised alone	39.9	71.2	83.3
fully supervised + **reranking**	71.8	91.2	95.9
weakly supervised + **reranking**	68.0	87.5	94.8

(d) Fully supervised learning Tricks on CUHK03

method	Rank-1	Rank-5	Rank-10
fully supervised alone	52.1	77.9	85.6
weaky supervised alone	44.0	70.6	79.7
fully supervised + **reranking**	67.5	88.2	91.8
weakly supervised + **reranking**	61.0	82.0	87.0

(e) Scalability of our method on CUHK03

categories	Rank-1	Rank-5	Rank-10
67	16.3	34.7	44.9
367	43.6	67.0	75.5
767 (∗**full training set**)	61.0	82.0	87.0

1) Effectiveness of the graphical learning Module: First, we investigate the effectiveness of the *refinement* operation. As discussed in Section IV-C, the graphical learning module plays the role of refining the person categorization results by correcting the errors between the rough Re-ID predictions and the weak annotations, which forms the basis of generating pseudo-image-level labels. During the training, we visualize the person categorization errors between the rough predictions and the weak annotations in Fig. 9. This experiment is conducted on CUHK03 using the setting of 10 categories/bag.

Fig. 9 displays the errors between the rough predictions and the weak annotations in the form of a confusion matrix containing 76×76 grids. Each grid indicates a bag of 10 categories, totally summing up to 760 categories, which approximates the number of person categories in the full training set (i.e., 767 categories). We have two major observations from Fig. 9. **First**, there is a significant gap between the rough predictions and the weak annotations (see 9 (a) or (b)), indicating that the rough Re-ID results are still not competent for generating pseudo-image-level labels. More importantly, the gap between the rough predictions and the bag-level annotations is non-negligible. This result indicates that it is necessary to refine the person categorization results by correcting the errors between the rough predictions and the bag-level weak annotations.

Second, the gap between the rough predictions and the weak annotations becomes smaller as the training iteration increases (from 10 epochs in 9(a) to 70 epochs in 9 (d)). Specifically, when the training finishes, the gap between the ground truth becomes significantly small. This result indicates that the problem is adequately addressed by the differentiable graphical learning module, which provides extra knowledge for the learning of the Re-ID model.

2) Scalability of Our Approach: We have shown that a Re-ID model can be learned with weakly labeled data. Next, we investigate whether increasing the amount of weakly labeled data will improve the performance of weakly supervised

learning. The entire CUHK03 training set is randomly partitioned into three subsets containing 67, 300, and 300 person categories, respectively. We evaluate the scalability of our approach by gradually adding one subset in training. The rank-1 accuracy is reported in Table II (e). For example, the first model is trained with the first 67 person categories, and the number of person categories is increased to 367 categories in the second model. The third model is trained with the full CUHK03 training set (i.e., 767 categories).

Table II (e) shows that the accuracies increase when we increase the scale of training data in CUHK03. For instance, our approach trained with full training data achieves the best performance and outperforms the other two models by 44.7% and 17.4%, respectively.

3) Impact of Bag Diversity: Intuitively, if a bag contains more person categories, it is more challenging to learn a weakly supervised Re-ID model because of the increase in entropy. Next, we investigate the performance with respect to such bag internal diversity. We conduct experiments on PRID2011 and CUHK03.

PRID2011. In Table II (a) and Fig. 10 (a), we compare five options, i.e., each bag containing 1, 2, 3, 4, or a random number of person categories, respectively. In particular, when each bag has only one person category, the weakly supervised Re-ID problem degrades into a fully supervised one.

We have three major observations from Table II (a) and Fig. 10 (a). **First**, the model that is trained with weakly labeled samples, achieves comparable accuracies to those trained with strongly labeled data. For example, in Table II (a), the rank-1 accuracies of the fully and weakly supervised learning are 71.8% and 68.0%, respectively. This result is very significant as we know that a weak annotation costs tens of times less money and time than a strong annotation. More importantly, the top 10 accuracies are almost the same, i.e., 95.9% *vs* 94.8%, indicating that our approach adequately addresses the weakly supervised scenario.

Second, the accuracy of the weakly supervised methods
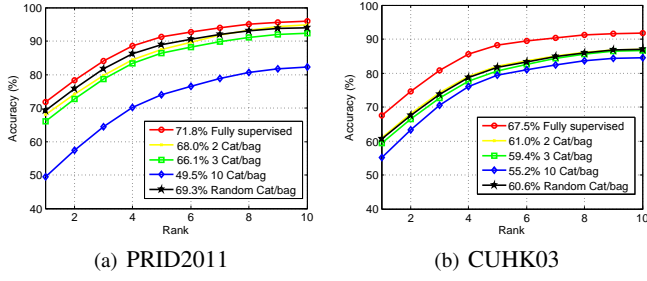
(a) PRID2011   (b) CUHK03

Fig. 10: Analysis on different bag diversities. **Cat/bag:** the number of person categories in each bag. **Random Cat/bag:** each bag contains random person categories, which reflects the real-world state. **Fully supervised:** each bag contains only one person category. In this case, the weakly supervised problem degrades into a fully supervised one.

(a) Standard PRID2011 test set.   (b) SYSU-30k

Fig. 11: Comparison with state-of-the-art methods. **Weakly sup.:** the proposed weakly supervised Re-ID approach. *** Fully Sup.:** each bag contains only one person category. In this case, the weakly supervised problem degrades into a fully supervised one. We thus consider the latter as the baseline of our weakly supervised Re-ID approach.

gradually decreases as the number of categories in each bag increases. In particular, the rank-1 accuracy of our approach drops by 18.5% when increasing the number of categories per bag from 2 to 10. We argue that the increase in uncertainty causes this optimization difficulty. When the category number within a bag increases, the uncertainty in the label assignment also increases. This means that the probability of adequately assigning an image-level label to each person image decreases, making the problem more challenging.

Third, it is noteworthy that the random version has appealing performance (69.3% *vs* 71.8% compared with the baseline), as shown in the last line of Table II (a). Specifically, the random version refers to each bag containing random categories of person, which reflects the real-world state. The good performance suggests that solving a weakly supervised Re-ID problem is feasible in reality.

CUHK03. A similar phenomenon can also be observed in the CUHK03 benchmark. In Table II (b) and Fig. 10 (b), we also compare the five settings consistent with those for the PRID2011 dataset. Table II (b) and Fig. 10 (b) show the behaviors of the weakly supervised methods. First, the model trained with weakly annotated data achieves comparable accuracy to those trained with fully annotated data (61.0% vs. 67.5%). Second, our approach suffers from an increased number of categories per bag, suggesting that such an increase in uncertainty is a fundamental problem.

4) Compatibility with Fully Supervised Learning Tricks: Intuitively, a weakly supervised Re-ID problem is likely to be upper bounded by fully supervised learning with all annotations. Next, we investigate the performance of our approach with respect to models with different fully-supervised learning capacities.

PRID2011. We first evaluate two different fully supervised learning baseline models. Both share the same architectures, as described in Section IV-C, except that the first one is a naked CNN framework, while the second one employs a reranking post-process (denoted as '**+reranking**' in Table II (c)). Table II (c) shows the top 1, 5, and ten accuracies of the fully supervised learning results, which form the baseline of this section.
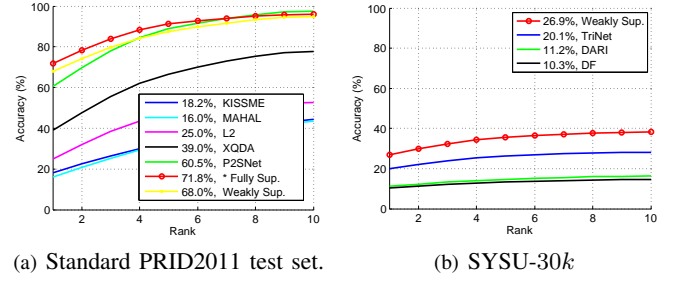
Next, we evaluate the weakly supervised learning scenario. The setting is similar to the above fully supervised setting, except that all of the image-level annotations are replaced with bag-level annotations in the training set. In this scenario, we present a horizontal comparison and a vertical comparison.

In the horizontal comparison, we focus on the performance gap between fully and weakly supervised learning. Once again, we observed that the rank-1 accuracy of using weak annotation approaches that of using strong annotation in both options.

In the vertical comparison, we compare the two weakly supervised learnings built on different baselines. The results are summarized in Table II (c). A finding of this experiment can be observed: weakly supervised learning with a stronger baseline ('weakly supervised **+ reranking**') yields better performance. For example, in the weak annotation setting, "weakly supervised **+ reranking**" yields 68.0%, compared to 39.9% obtained by "weakly supervised", a relative improvement of 70.4%. This comparison verifies the compatibility of our method with existing frameworks; i.e., the existing trick (e.g., reranking) used to improve the fully supervised learning can also be applied to the weakly supervised Re-ID.

CUHK03. Similar observations can also be obtained on CUHK03 in Table II (d). The approach with reranking [78] achieves better accuracies than that without re-reranking in both fully supervised learning (67.5% vs. 52.1%) and weakly supervised learning (61.0% vs. 44.0%), once again proving that the existing trick to improve the fully supervised learning can also be applied to the weakly supervised Re-ID.

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

In this section, we compare our weakly supervised approach with the best-performing fully supervised / semi-supervised / unsupervised methods.

1) Accuracy on PRID2011: In Table III (a) and Fig. 11 (a), we compare the results of our model with the current best model results. Note that although our method was trained in the weakly supervised scenario, we still evaluate it in the same setting as conventional methods do. This leaves our approach at a disadvantage. Five representative image-to-image Re-ID models are used as the competing methods:

the KISSME distance learning method [15], MAHAL, L2, and XQDA [8], and P2SNet [31]. For KISSME, MAHAL, L2, and XQDA, deep features [80] are utilized to represent an image of a person. For P2SNet, we train the model based on the image-to-video setting but sample one frame from each video to formulate the image-to-image setting. The above settings are all consistent with the traditional settings, e.g., [31]. Our method achieves excellent performance, even surpassing the state-of-the-art fully-supervised methods. Accurately, it achieves a rank-1 accuracy of 68.0%. We also observe that this result surpasses all of the above competitive methods, such as KISSME, MAHAL, L2, and XQDA, even if they are trained with all available strong annotations.

2) Accuracy on CUHK03: Our weakly supervised Re-ID is compared with state-of-the-art methods in two groups, including the traditional fully-supervised Re-ID and the unsupervised Re-Id.

Fully supervised Re-ID. In Table III (b), we compare our method with the current best models. Eleven representative state-of-the-art methods are used as competing methods, including BOW+XQDA [2], PUL [81], LOMO+XQDA [8], IDE(R) [82], IDE+DaF [83], IDE+XQ+reranking [78], PAN, DPFL [84], and newly proposed methods such as SVDNet [85] and TriNets [86]. All settings of the above methods are consistent with the common training setting. Our approach achieves very competitive accuracy. For example, our approach achieves a rank-1 accuracy of 61.0%. We also highlight that this result surpasses many of the current competitive methods, such as BOW+XQDA [2], PUL [81], LOMO+XQDA [8], IDE [82], IDE+DaF [83], IDE+XQ+reranking [78], PAN, DPFL [84] and SVDNet [85], which are trained with all available strong annotations. This result once again verifies the effectiveness of our method.

To validate the superiority of our weakly supervised Re-ID over previous annotation-saving Re-ID works, we further compare with unsupervised Re-ID methods.

Unsupervised Re-ID. In Table III (c), we compare our method with the unsupervised Re-ID models. Three representative state-of-the-art methods are used as competing methods, including CAMEL [36], PatchNet [87], and PAUL [87]. The results in Table III (c) show that our weakly supervised Re-ID problem has obtained significant gain over unsupervised Re-ID methods. For instance, our method outperforms the best-performing model PAUL [87] by a large margin (i.e., 8.7%). Note that compared to unsupervised Re-ID, the annotation effort of our weakly supervised Re-ID is also very inexpensive. These results verify the effectiveness of our method again.

3) Accuracy on Market-1501: Our weakly supervised Re-ID is compared with state-of-the-art methods in three groups, including the traditional fully-supervised Re-ID, the unsupervised Re-Id, and the semi-supervised Re-ID.

Fully supervised Re-ID. In Table III (c), we compare our method with the fully supervised Re-ID models. Twelve representative state-of-the-art methods are used as competing methods, including MSCAN [88], DF [89], SSM [90], SVDNet [85], GAN [65], PDF [91], TriNet [74], TriNet + Era. + reranking [86], PCB [72], VPM [92], JDGL [93], and AANet [94]. All settings of the above methods are consistent

TABLE III: Comparison with state-of-the-art methods. **Weakly supervised:** the proposed weakly supervised Re-ID approach. *** Fully supervised:** each bag contains only one person category. In this case, the weakly supervised problem degrades into a fully supervised one. We thus consider the latter as the baseline of our weakly supervised Re-ID approach. ‡**CUHK03:** pretrained on CUHK03. ‡**Market-1501:** pretrained on Market-1501. **reranking:** see [78].

(a) Comparison on standard PRID2011 test set

Supervision	Method	Rank-1	Rank-5	Rank-10
Fully	KISSME [15]	18.2	33.2	44.5
	MAHAL	16.0	32.5	43.6
	L2	25.0	46.6	52.8
	XQDA [8]	39.0	66.6	77.8
	P2SNet [31]	60.5	88.9	97.5
	*Fully supervised	71.8	91.2	95.9
Weakly	**Weakly supervised**	68.0	87.5	94.8

(b) Comparison on standard CUHK03 test set.

Supervision	Method	Rank-1
Fully	BOW+XQDA [2]	6.4
	PUL [81]	9.1
	LOMO+XQDA [8]	12.8
	IDE(R) [82]	21.3
	IDE+DaF [83]	26.4
	IDE+XQ.+reranking [78]	34.7
	PAN	36.3
	DPFL [84]	40.7
	SVDNet [85]	41.5
	TriNet+Era. [86]	55.5
	TriNet+Era.+reranking [86]	64.4
	*Fully supervised	67.5
Unsupervised	CAMEL [36]	31.9
	PatchNet [87]	45.4
	PAUL [87]	52.3
Weakly	**Weakly supervised**	61.0

(b) Comparison on standard Market-1501 test set.

Supervision	Method	Rank-1
Fully	MSCAN [88]	80.31
	DF [89]	81.0
	SSM [90]	82.21
	SVDNet [85]	82.3
	GAN [65]	83.97
	PDF [91]	84.14
	TriNet [74]	84.92
	TriNet+Era.+reranking [86]	85.45
	PCB [72]	93.4
	VPM [92]	93.0
	JDGL [93]	94.8
	AANet [94]	92.4
	*Fully supervised	94.2
Unsupervised	CAMEL [36]	54.5
	TAUDL [95]	63.7
	UTAL [45]	69.2
	UDA [42]	75.8
	MAR [43]	67.7
	DECAMEL [96]	60.2
	ECN [97]	75.1
	PAUL [87]	68.5
	HHL [98]	62.2
	Distilled [99]	61.5
	PUL [44]	50.9
Semi	SPACO [100]	68.3
	HHL [98]	54.4
	Distilled [99]	63.9
	One Example [58]	70.1
	Many Examples [58]	82.5
Weakly	**Weakly supervised**	88.6

(c) Comparison on SYSU-30k

Supervision	Method	Rank-1
Fully	DARI [29], ‡CUHK03	11.2
	DF [13], ‡CUHK03	10.3
	TriNet [86], ‡CUHK03	20.1
Weakly	**Weakly supervised**	26.9

with the common training setting. Our approach achieves very competitive accuracy. For example, our approach achieves a rank-1 accuracy of 88.6%. We also highlight that this result surpasses many of the current competitive methods, such as MSCAN [88], DF [89], SSM [90], SVDNet [85], GAN [65], PDF [91], TriNet [74], and TriNet + Era. + reranking [86], which are trained with all available strong annotations. This result once again verifies the effectiveness of our method.

To validate the superiority of our weakly supervised Re-ID over previous annotation-saving Re-ID works, we further compare with unsupervised and semi-supervised Re-ID.

Unsupervised Re-ID. In Table III (c), we compare our method with the unsupervised Re-ID models. Eleven representative state-of-the-art methods are used as competing methods, including CAMEL [36], TAUDL [95], UTAL [45], UDA [42], MAR [43], DECAMEL [96], ECN [97], PAUL [87], HHL [98], Distilled [99], and PUL [44]. The results in Table III (c) show that our weakly supervised Re-ID problem has obtained significant gain over unsupervised Re-ID methods. For instance, our method outperforms the best-performing model UDA [42] by a large margin (i.e., 12.8%). Note that compared to unsupervised Re-ID, the annotation effort of our weakly supervised Re-ID is also very inexpensive. These results verify the effectiveness of our method again.

Semi-supervised Re-ID. In Table III (c), we compare our method with the semi-supervised Re-ID models. Five representative state-of-the-art methods are used as competing methods, including SPACO [100], HHL [98], Distilled [99], One Example [58], and Many Examples [58]. The results show that our weakly supervised Re-ID problem has obtained significant gain over semi-supervised Re-ID methods. For instance, our method outperforms the best-performing model "ManyExamples" [58] by a large margin (i.e., 6.1%). Note that compared to semi-supervised Re-ID, the annotation effort of our weakly supervised Re-ID is also very inexpensive. These results verify the effectiveness of our method again.

4) Accuracy on SYSU-$30k$: As SYSU-$30k$ is the only weakly supervised Re-ID dataset and our method is the only weakly supervised Re-ID method, we propose to compare the conventional fully supervised Re-ID models with our weakly supervised method by using transfer learning. Specifically, three representative fully supervised Re-ID models including DARI [29], DF [13], and TriNet [86] are first trained on CUHK03. Then, they are used to performed cross-dataset evaluation on the test set of SYSU-$30k$. In contrast, our weakly supervised Re-ID is trained on the training set of the SYSU-$30k$ with weak annotations and then is tested on the test set of SYSU-$30k$.

Table III (c) and Fig. 11 (b) show the results of the comparisons. We can observe that our approach can achieve state-of-the-art performance (26.9% vs. 20.1%), even though our method is trained in a weakly supervised manner while the competitors are trained with full supervision. The success may be attributed to two reasons. First, our model is quite effective due to the graphical modeling that generates reliable pseudo labels as compensation for the absence of strong labels. Second, the large-scale SYSU-$30k$ dataset provides rich knowledge that improves the capacity of our model, even

TABLE IV: Computational complexity of weakly and fully supervised Re-ID. **secs / 100 images:** the time of forward-passing 100 images in the testing stage or the cycle of a forward-backward passing in the training stage when the batch size is 100.

	weakly (secs / 100 images)	fully (secs / 100 image)
Testing	0.0559	0.0559
Training	0.2453	0.2448

though SYSU-$30k$ is annotated weakly.

In summary, the comparisons provide a promising conclusion, i.e., learning a Re-ID model using less annotation effort is possible.

5) Computational Complexity: Table IV compares the computational time of Re-ID in the context of weak supervision to that in the context of full supervision in terms of time cost per 100 images. For a fair comparison, both methods are individually trained on the same desktop with 1 Titan-x GPU. As shown in the table, the weakly and fully supervised Re-ID methods have similar computational costs. Specifically, in the testing phase, both methods share the same computational costs. Even in the training phase, our method only performs $0.002\times$ slower than the fully supervised Re-ID (0.2453 vs. 0.2448 seconds per 100 images using TITAN X.).

VI. CONCLUSION

We have considered a new and more realistic Re-ID problem challenge: the weakly supervised Re-ID problem. To address this new problem, we proposed a graphical model to capture the dependencies among images in each weakly annotated bag, which are specifically designed to address the weakly supervised Re-ID problem. We further propose a weakly annotated Re-ID dataset (i.e., SYSU-$30k$) to facilitate future research, which is currently the largest Re-ID benchmark. Future work will include building more effective weakly supervised Re-ID models.

REFERENCES

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

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2360–2367.

[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[5] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[6] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2528–2535.

[7] ——, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

[8] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[9] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*. IEEE, 2011, pp. 649–656.

[10] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European conference on computer vision*. Springer, 2014, pp. 536–551.

[11] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6, pp. 379–390, 2014.

[12] G. Wang, J. Lai, Z. Xie, and X. Xie, "Discovering underlying person structure pattern with relative local distance for person re-identification," *arXiv preprint arXiv:1901.10100*, 2019.

[13] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[14] G. Wang, X. Xie, J. Lai, and J. Zhuo, "Deep growing learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2812–2820.

[15] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.

[16] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[17] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 413–422.

[18] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2666–2672.

[19] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[20] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[21] G. Wang, P. Luo, X. Wang, L. Lin *et al.*, "Kalman normalization: Normalizing internal representations across network layers," in *Advances in Neural Information Processing Systems*, 2018, pp. 21–31.

[22] Y. Li, G. Wang, L. Lin, and H. Chang, "A deep joint learning approach for age invariant face verification," in *CCF Chinese Conference on Computer Vision*. Springer, 2015, pp. 296–305.

[23] W. Liang, G. Wang, J. Lai, and J. Zhu, "M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification," *arXiv preprint arXiv:1811.03768*, 2018.

[24] G. Wang, K. Wang, and L. Lin, "Adaptively connected neural networks." 2019.

[25] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *BMVC*, vol. 2, no. 5, 2010, p. 6.

[26] Y. Li, G. Wang, L. Nie, Q. Wang, and W. Tan, "Distance metric optimization driven convolutional neural network for age invariant face recognition," *Pattern Recognition*, vol. 75, pp. 51–62, 2018.

[27] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1089–1102, 2017.

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

[29] G. Wang, L. Lin, S. Ding, Y. Li, and Q. Wang, "Dari: Distance metric and representation integration for person verification." in *AAAI*, 2016, pp. 3611–3617.

[30] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1345–1353.

[31] G. Wang, J. Lai, and X. Xie, "P2snet: Can an image match a video for person re-identification in an end-to-end way?" *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[32] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," 2019.

[33] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4678–4686.

[34] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," *arXiv preprint arXiv:1804.02792*, 2018.

[35] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," *arXiv preprint arXiv:1702.05729*, 2017.

[36] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *IEEE International Conference on Computer Vision*, 2017.

[37] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification." in *BMVC*, vol. 3, 2015, p. 8.

[38] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1629–1642, 2015.

[39] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1306–1315.

[40] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," 2014.

[41] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 356–370, 2017.

[42] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice." *arXiv: Computer Vision and Pattern Recognition*, 2018.

[43] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, and J. Lai, "Unsupervised person re-identification by soft multilabel learning." *CVPR*, 2019.

[44] 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, p. 83, 2018.

[45] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[46] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification." *BMVC*, p. 48, 2018.

[47] D. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. V. Der Maaten, "Exploring the limits of weakly supervised pretraining," *european conference on computer vision*, pp. 185–201, 2018.

[48] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Feifei, "The unreasonable effectiveness of noisy data for fine-grained recognition," *european conference on computer vision*, pp. 301–320, 2016.

[49] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation," *arXiv preprint arXiv:1502.02734*, 2015.

[50] P. Luo, G. Wang, L. Lin, and X. Wang, "Deep dual learning for semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2718–2726.

[51] G. Wang, P. Luo, L. Lin, and X. Wang, "Learning object interactions and descriptions for semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5859–5867.

[52] Z. Yan, J. Liang, W. Pan, J. Li, and C. Zhang, "Weakly-and semi-supervised object detection with expectation-maximization algorithm," *arXiv preprint arXiv:1702.08740*, 2017.

[53] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2017.

[54] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.

[55] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[56] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European Conference on Computer Vision*. Springer, 2016, pp. 108–124.

[57] L. Lin, G. Wang, R. Zhang, R. Zhang, X. Liang, and W. Zuo, "Deep structured scene parsing by learning with image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2276–2284.

[58] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, 2019.

[59] 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," pp. 5177–5186, 2018.

[60] J. Meng, S. Wu, and W. Zheng, "Weakly supervised person re-identification." *arXiv: Computer Vision and Pattern Recognition*, 2019.

[61] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[62] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[63] Q. Huang, M. Han, B. Wu, and S. Ioffe, "A hierarchical conditional random field model for labeling and segmenting images of street scenes," pp. 1953–1960, 2011.

[64] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *European Conference on Computer Vision*. Springer, 2018, pp. 508–526.

[65] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, vol. 3, 2017.

[66] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018, pp. 79–88.

[67] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 31–44.

[68] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

[69] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *Bmvc*, vol. 1, no. 2. Citeseer, 2011, p. 6.

[70] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, 2016.

[71] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *computer vision and pattern recognition*, pp. 6517–6525, 2017.

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

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[75] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[76] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[77] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

[79] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, no. 5. Citeseer, 2007, pp. 1–7.

[80] 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*. Springer, 2016, pp. 868–884.

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

[82] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[83] R. Yu, Z. Zhou, S. Bai, and X. Bai, "Divide and fuse: A re-ranking approach for person re-identification," *arXiv preprint arXiv:1708.04169*, 2017.

[84] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2590–2600.

[85] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," *arXiv preprint arXiv:1703.05693*, 2017.

[86] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[87] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[88] 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, pp. 384–393.

[89] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification." in *ICCV*, 2017, pp. 3239–3248.

[90] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, 2017.

[91] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*. IEEE, 2017, pp. 3980–3989.

[92] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification." *CVPR*, 2019.

[93] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification." *CVPR*, 2019.

[94] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[95] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," *ECCV*, pp. 772–788, 2018.

[96] H. Yu, A. Wu, and W. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[97] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification." *CVPR*, 2019.

[98] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *ECCV*, 2018.

[99] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[100] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong, "Self-paced co-training," pp. 2275–2284, 2017.

Guangrun Wang is currently a Ph.D. candidate in the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He received the B.E. degree from Sun Yat-sen University in 2014. From 2015 to 2017, he was a visiting scholar with the Department of Information Engineering, The Chinese University of Hong Kong. His research interests include machine learning, computer vision. He is the recipient of the 2018 Pattern Recognition Best Paper Award and one ESI Highly Cited Paper.

Guangcong Wang is pursuing a Ph.D. degree in the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He received the B.E. degree in communication engineering from Jilin University (JLU), Changchun, China, in 2015. His research interests are computer vision and machine learning. He has published several works on person re-identification, weakly supervised learning, semi-supervised learning, and deep learning.
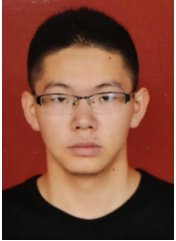
Xujie Zhang is currently an undergraduate student in the School of Data and Computer Science, Sun Yat-sen University (SYSU), Guangzhou, China. He majors in computer science. His research interest is computer vision and machine learning. Currently, he aims at developing algorithms for person re-identification, especially weakly supervised person re-identification.

Jianhuang Lai received the Ph.D. degree in mathematics in 1999 from Sun Yat-sen University, China. He joined Sun Yat-sen University in 1989 as an assistant professor, where he is currently a Professor of the School of Data and Computer Science. His current research interests are in the areas of digital image processing, pattern recognition, and applications. He has published over 200 scientific papers in the academic journals and conferences on image processing and pattern recognition.

Liang Lin (M'09, SM'15) a full Professor of Sun Yat-sen University. From 2008 to 2010, he was a Post-Doctoral Fellow at University of California, Los Angeles. He has authorized and co-authorized on more than 100 papers in top-tier academic journals and conferences. He has been serving as an associate editor of IEEE Trans. Human-Machine Systems. He served as Area Chairs for numerous conferences such as CVPR, and ICCV. He was the recipient of Google Faculty Award in 2012, Best Paper Diamond Award in IEEE ICME 2017, and Hong Kong Scholars Award in 2014, and the Annual Pattern Recognition Best Paper Award. He is a Fellow of IET.