

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

Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, 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 individuals, which is about 20 times larger than CUHK03 (1.3k individuals) and Market-1501 (1.5k individuals), and 30 times larger than 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 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, Graphical Neural Networks, Visual Surveillance

## I. INTRODUCTION

PERSON re-identification (Re-ID) [3] has been extensively studied in recent years, which refers to the problem of recognizing persons across cameras. Solving the Re-ID problem has many applications in video surveillance for public safety. In the past years, deep learning has been introduced to Re-ID and has achieved promising results.

However, a crucial bottleneck in building deep models is that they typically require strongly annotated images during training. In the context of Re-ID, strong annotation refers to assigning a clear 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 images must be annotated within a rather short time (see Fig. 1 (a)).

The first two authors contribute equally and share the first authorship. G. Wang, G. Wang, X. Zhang, J. Lai and L. Lin 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; sts1jh@mail.sysu.edu.cn; linliang@ieee.org; Z. Yu is with Artificial Intelligent Key Laboratory of Yunnan province, Kunming University of Science and Technology, Kunming, 650500, China; Email: ztyu@hotmail.com. Corresponding author: Liang Lin.

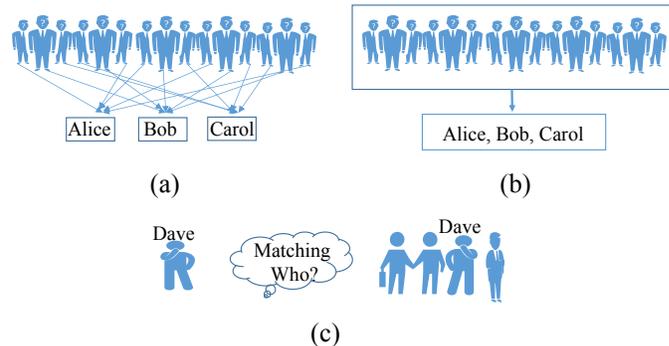


Fig. 1: Problem definition for the weakly supervised Re-ID. (a) is an example of strong annotation while (b) is an example of a weak annotation. During testing, there is no difference between the fully and weakly supervised Re-ID problems, i.e., they both aim at finding the best-matching image for a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

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 exact person 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, we train Re-ID models with the bag-level annotations. 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  $\{(x_1, y_1), \dots, (x_i, y_i), \dots\}$ , 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  $\{(b_1, l_1), \dots, (b_j, l_j), \dots\}$ , where  $b_j \in B$  is a bag of person images, i.e.,  $b_j = \{x_{j1}, x_{j2}, \dots, x_{jp}\}$ ; and  $l_j \in L$  is its bag-level label, i.e.,  $l_j = \{y_{j1}, y_{j2}, \dots, y_{jq}\}$ . Note that the mappings between  $\{x_{j1}, x_{j2}, \dots, x_{jp}\}$  and  $\{y_{j1}, y_{j2}, \dots, y_{jq}\}$  are unknown. During testing, there is no difference between fully and weakly supervised Re-ID problems (see Fig. 1 (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

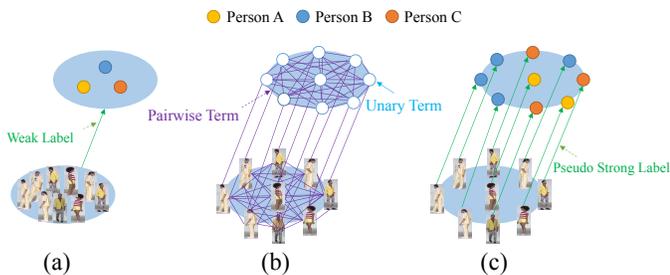


Fig. 2: An illustration of the proposed method for weakly supervised Re-ID. (a) shows a bag of images and their bag-level label. (b) presents the process of differentiable graphical learning. Using graphical modeling, we can obtain the pseudo image-level label for each image, as shown in (c).

they are designed for computing either cross-view-invariant features or distance metrics, 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 ID for each image. None of them are designed to train a weakly supervised model.

Although the weak annotations lack 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. By using the weak annotations, 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 terms of bags (Fig. 2 (a)) to obtain the rough categorization probabilities. These categorization probabilities are modeled as the unary terms in a probabilistic graphical model; see Fig. 2 (b). **Second**, we further model the relations between person images as the pairwise terms in a graph by considering their similarity in the features and appearance; see Fig. 2 (b). The unary term and the pairwise term 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 [4]), our 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 as follows.

1) We define a weakly supervised Re-ID problem by replacing the image-level annotations in traditional Re-ID 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 in Re-ID problems. SYSU-30k contains 30k individuals, which is about 20 times larger than CUHK03 (1.3k individuals) and Market-1501 (1.5k individuals), 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



Fig. 3: Examples in our SYSU-30k dataset. (a) training images in terms of bag; (b) their bag-level annotations; (c) test set.

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 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. Section III introduces the annotation of SYSU-30k, followed by the weakly supervised Re-ID model in Section IV. We also discuss

the relationship of our work with previous work in Section V. The experimental results are presented in Section VI. Section VII concludes the work and presents outlooks for future work.

## II. RELATED WORK

**Person Re-ID.** 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], (2) learning a robust metric or subspace for matching [3], [5], [6], and (3) joint learning of the above two methods [7], [8]. Recently, there are many works on the generalization of Re-ID, such as video-based and image-to-video Re-ID [9], spatio-temporal Re-ID [10], occluded Re-ID [11], and natural language Re-ID [12]. However, all these methods assume that the training labels are strong. They are thus ineffective for solving the weakly supervised Re-ID. Recently, the robustness of Re-ID are also examined by [13].

**Unsupervised Re-ID.** Another approach that is free from the prohibitively high cost of manual labeling is unsupervised learning Re-ID [14], [15]. These methods either use local saliency matching or resort to clustering models [14]. However, without the help of labeled data, it is difficult to model the dramatic variances across camera views in feature/metric learning. Therefore, it is difficult for these pipelines to achieve high accuracies [16]–[19]. In contrast, the proposed weakly supervised Re-ID problem has a better solution. Note that compared to unsupervised Re-ID, the annotation effort of weakly supervised Re-ID is also very inexpensive.

**Semi-supervised Re-ID.** Apart from our model, there have been some uncertain label learning models, among which the one-shot/one-example Re-ID [20], [21] 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 ID 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, [22] also proposes to cope with the uncertain-label Re-ID problem using multiple-instance multiple-label learning. However, similar to [21], at least one accurate label for each person ID is still in a desire to form the probe set in [22]. Therefore, mathematically, [20]–[22] are all semi-supervised Re-ID but NOT weakly supervised Re-ID.

**Weakly-supervised learning.** Beyond Re-ID, although training deep models with weak annotations is challenging, it has been partially investigated in the literature, such as in tasks of image classification [23], [24], semantic segmentation [25]–[28], object detection [29]. 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 weakly supervised segmentation tasks, pixels in an image share certain motion of rigidity and stability that benefits the prediction. Whereas in the weakly supervised Re-ID, persons in video bags are more unordered and irregular. Therefore, the weakly supervised Re-ID problem is considerably more challenging than other problems.

**Graphical learning.** To address the weakly supervised Re-ID problem, we propose to generate a pseudo label for each image by introducing differentiable graphical learning, which is inspired by the advances in image segmentation [30] and videos [31]. Recently, one classical graphical model, i.e., CRF, has been introduced to Re-ID for similarity learning [32]. However, our method differs from [32] in two aspects. First, like existing methods, [32] uses CRF as a post-processing tool to refine the prediction provided by fully supervised learning, while our method exploits the supervision-independent property of graphical learning [30] to generate pseudo labels for our weakly-supervised Re-ID learning. Second, different from traditional non-differentiable graphical models and [32], 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).

**Person search.** Another problem that is very related to our problem is person search [12], 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 reasons. **First**, the pedestrians in a TV program video are often cross-view and cross-camera because there are many movable cameras to capture the shots for post-processing. Re-identifying 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. 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 videos in a weak fashion. In particular, each video is divided into 84,924 bags of arbitrary length. Then, the annotators record the pedestrian’s identity for each bag. YOLO-v2 [38] 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

TABLE I: A comparison of different Re-ID benchmarks. **Categories:** We treat each person identity as 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) Comparison with existing Re-ID datasets.

Dataset	CUHK03 [1]	Market-1501 [2]	Duke [33]	MSMT17 [34]	CUHK01 [35]	PRID [36]	VIPeR [3]	CAVIAR [37]	SYSU-30k
Categories	1,467	1,501	1,812	4,101	971	934	632	72	<b>30,508</b>
Scene	Indoor	Outdoor	Outdoor	Indoor, Outdoor	Indoor	Outdoor	Outdoor	Indoor	<b>Indoor, Outdoor</b>
Annotation	Strong	Strong	Strong	Strong	Strong	Strong	Strong	Strong	<b>Weak</b>
Cameras	2	6	8	15	10	2	2	2	<b>Countless</b>
Images	28,192	32,668	36,411	126,441	3,884	1,134	1,264	610	<b>29,606,918</b>

(b) Comparison with ImageNet-1k

Dataset	ImageNet-1k	SYSU-30k
Categories	1,000	<b>30,508</b>
Images	1,280,000	<b>29,606,918</b>
Annotation	Strong	<b>Weak</b>

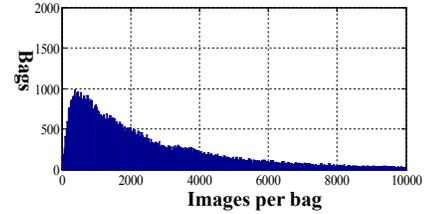
as the test set, leaving the rest as the training set. There is no overlap between the training set and the test set. Notably, we treat each person identity as a category in this paper.

We provide some samples of the SYSU-30k dataset in Fig. 3. As shown, our SYSU-30k dataset exhibits challenges of illumination variance (Row 2, 7, and 9), occlusion (Row 6 and 8), low resolution (Row 2 and 9), looking-downward cameras (Row 2, 5, 6, 8, and 9), and complicated backgrounds of the real scenes (Row 2-10).

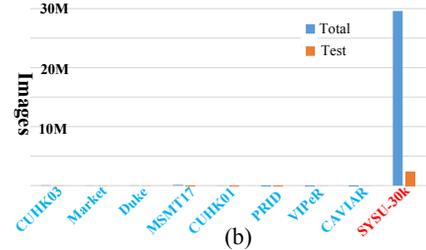
**Dataset statistics.** SYSU-30k contains 29,606,918 person images with 30,508 categories in total, which is further divided into 84,930 bags (only for training set). Fig. 4 (a) summarizes the number of bags with respect to the number of images per bag, showing that each bag has 2,885 images on average. 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 [33], MSMT17 [34], CUHK01 [35], PRID [36], VIPeR [3], and CAVIAR [37]. Fig. 4 (c) and (d) plots the person IDs 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 IDs. 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.

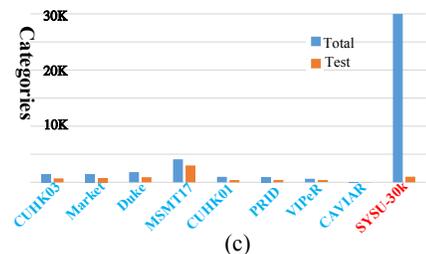
A further comparison of SYSU-30k with existing Re-ID benchmarks is shown in Table I (a), including categories, scene, annotation, cameras, and image numbers (bounding boxes). After the comparison, we summarize the new features in SYSU-30k in the following aspects. First, SYSU-30k is the first weakly 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



(a)



(b)



(c)

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 IDs for both the entire dataset and the test set.

due to many cameras, realistic indoor and outdoor scenes, and occasionally incorrect annotations. Fourth, 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.

**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 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]. Following [1], we fix the train/test partitioning. In the test set, we choose 1,000 images belonging to 1,000 different person IDs to form the query set. As the scalability is important for the practicability

of Re-ID systems, we propose to challenge 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, just like the police search a massive amount of videos for a criminal. Following [2], we use the rank-1 accuracy as the evaluation metric.

#### IV. WEAKLY SUPERVISED RE-ID MODEL

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

Let  $b = \{x_1, \dots, x_j, \dots, x_p\}$  denote a bag containing  $p$  images.  $y = \{y_1, \dots, y_j, \dots, y_p\}$  are the image-level labels and  $l$  is the bag-level label.

In the fully supervised Re-ID, the image-level labels  $y$  are known. The goal is to learn a model by minimizing the loss between the image-level labels and the predictions.

In contrast, in the weakly supervised Re-ID, the bag-level label  $l$  is provided, but the image-level labels  $y$  are unknown. Suppose a bag contains  $n$  person IDs and there are  $m$  IDs in the entire dataset. A preliminary image-level label  $\mathbf{Y}_j$  for each image  $x_j$  can be inferred from its bag-level label  $l$ :

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

Then,  $\mathbf{Y}$  can be used as a bag constraint to deduce a pseudo-image-level labels  $\hat{y}$ , which can be further used to supervise the model learning.

##### B. Weakly supervised Re-ID: differentiable graphical learning

In this section, we propose differentiable graphical learning to generate pseudo-image-level labels for the person images.

**Graphically modeling Re-ID.** In our graph, each node represents a person image  $x_i$  in a bag and each edge represents the relation between person images, as illustrated in Fig. 5. Here  $i$  is the image index in a bag. Assigning a label  $y_i$  to a node  $x_i$  has an energy cost. The energy cost  $E(y|x)$  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  $x_i$ .  $\Psi(y_i, y_j|x_i; x_j)$  is the pairwise term that measures the penalty of assigning labels to a pair of images  $(x_i, x_j)$ . Mathematically, graphical modeling is to smooth the uncertain prediction of person IDs. The unary term performs the prediction based on sole nodes. While the pairwise term smooths the prediction of multiple nodes by considering their appearance and features. In summary, Eq. (2) is to clean up the spurious predictions of classifiers learned in a weakly supervised manner.

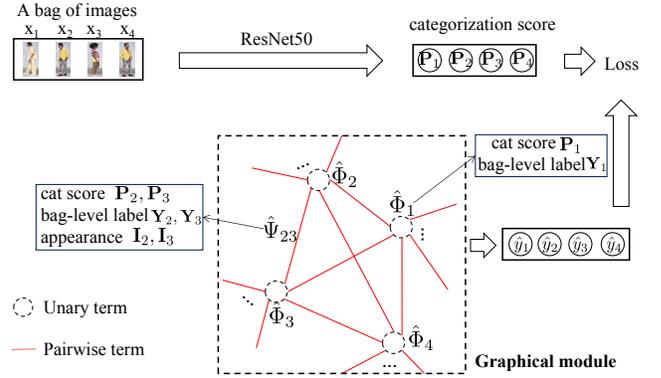


Fig. 5: Graphical model to generate pseudo image-level labels for person images. The unary terms are estimated by the deep networks, while the pairwise terms involve the similarity of features, the image appearance, and the bag-level label.

**Unary term.** The unary terms is typically defined as:

$$\Phi(y_i|x_i) = -\mathbf{P}_i^{y_i} \log(\mathbf{Y}_i^{y_i} \odot \mathbf{P}_i^{y_i}), \quad (3)$$

where  $\mathbf{P}_i$  is the categorization probability of a person image  $x_i$  outputted by a DNN.  $\odot$  denotes element-wise product.

As the unary term alone is generally noisy and inconsistent. Interactions between pairwise terms are required.

**Pairwise term.** The pairwise term is defined as:

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

where  $I_i$  and  $I_j$  are the low-level features and based on them, a Gaussian kernel is employed to measure their appearance similarity. The hyper-parameter  $\sigma$  controls the scale of the Gaussian kernel. This kernel forces the images with similar appearance 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  $\mathbf{Y}_j$ . The pairwise terms are widely known to 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 Eq. (4) is given by the Potts model, namely,

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

It introduces a penalty for similar images that are assigned different labels. Considering that Eq. (2) is non-differentiable, it is incompatible with DNNs. Thus, we will instead learn a differential version of Eq. (2) in a deep learning model.

**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 an ID that is not in the bag-level annotation, the estimation is undoubtedly considered as incorrect. Then, the estimation will be corrected by matching the image to the ID in the bag-level annotation with the most significant prediction score. Furthermore, if some IDs 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 IDs to improve the performance. In this way, knowledge of the weakly labeled data can be fully

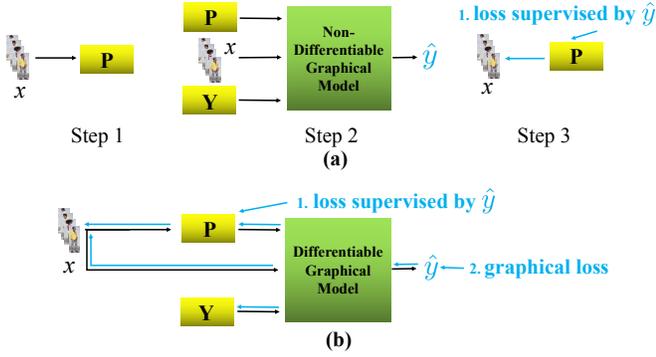


Fig. 6: Differentiable graphical model in DNNs, where  $x$ ,  $\mathbf{Y}$ ,  $\mathbf{P}$ ,  $\hat{y}$  denote the input images, bag-level label, preliminary categorization, and refined categorization, respectively. (a) is the stepwise graphical model, while (b) is our proposed end-to-end differentiable graphical model. Our model consists of two losses, i.e., an unsupervised loss for pseudo label generation and a loss supervised by the pseudo labels. Here, the black lines denote forward passing, while the blue lines denote back-propagation.

exploited. Given a bag of images and their bag-level label, we refine the DNN predictions by element-wisely multiplying  $\mathbf{P}$  with the bag-level weak annotation  $\mathbf{Y}$ . This is shown in the unary term in Eq. (3). Similarly, we also impose  $\mathbf{Y}_i$  and  $\mathbf{Y}_j$  on the pairwise term in Eq. (4).

Moreover, there is a natural smoothness in a video that could not be ignored. The person IDs in the adjacent bags change slowly within a short time; for instance, an image-level label  $y_i$  in bag  $b_T$  could also be in the bag  $b_{T+1}$ . A large amount of bags with overlapping IDs naturally exist in a video, which sheds light on the ability 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  $\{y_j\}$ . Meanwhile, one image in the first bag is similar to another image in the second bag. Hence, our model predicts these two images as  $y_i$ . Finally,  $y_i$  and  $y_k$  are assigned to the remaining images.

**Deducing pseudo image-level labels.** By minimizing the energy cost of Eq. (2), we can obtain the pseudo image-level label  $\hat{y}_i$  for the person image  $\hat{x}_i$ :

$$\hat{y}_i = \arg \max_{y_i \in \{1, \dots, m\}} E(y_i | x_i), \quad (6)$$

where  $\{1, \dots, m\}$  denote all the person IDs in the training set. Once such labels are generated, they are used to update the network parameters.

**Differentiable 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. 6 (a)). To enable an end-to-end optimization, we propose to make our graphical learning differentiable and compatible with DNNs (see Fig. 6 (b)).

We first investigate the mechanism of a non-differentiable graphical model. As illustrated in Fig. 6 (a), a non-differentiable graphical model consists of three steps. **First**, a preliminary categorization score  $\mathbf{P}$  is obtained through a DNN. **Second**, the energy cost in Eq. (2) is minimized by re-assigning labels to the images appropriately, subject to the appearance similarity, the preliminary categorization scores,

and the bag constraint. **Third**, the re-assigned labels are considered as the pseudo labels and used to supervise the learning of the Re-ID model.

The label reassignment in the second step claimed above is non-differentiable, which makes the graphical model incompatible with DNNs. To fill this gap, a relaxation form of Eq. (2) is desirable. With a continuous version of  $\hat{\Phi}$  and  $\hat{\Psi}$  to approximate the discrete  $\Phi$  and  $\Psi$ , we rewrite Eq. (2) as follows

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

where  $\hat{\Phi}$  and  $\hat{\Psi}$  are defined as:

$$\begin{cases} \hat{\Phi}(x_i) = - \sum_{j=1}^m [h(\arg \max_{k \in \{1, \dots, m\}} \mathbf{Y}_i^k \odot \mathbf{P}_i^k)]^j \log(\mathbf{P}_i^j), \\ \hat{\Psi}(x_i, x_j) = - \exp\left(-\frac{\|I_i - I_j\|^2}{2\sigma^2}\right) (\mathbf{Y}_i \mathbf{P}_i)^T \log(\mathbf{Y}_j \mathbf{P}_j). \end{cases} \quad (8)$$

$\arg \max$  returns the index of the largest element in a vector,  $h$  is a function that maps a scalar to one-hot vector, and the superscripts  $j$  and  $k$  mean indexing the  $j$  and  $k$  element of a vector, respectively. The differences between Eq. (3)-(4) and Eq. (8) are summarized as follows: **1)** the replacement of  $\Phi(y_i)$  with  $\hat{\Phi}(x_i)$  facilitates an end-to-end learning. Because, in a non-differentiable model,  $y_i$  is the input variable, while  $x_i$  is regarded as the input of DNNs in differentiable models. **2)** We use  $\arg \max$  to obtain the prediction, which is consistent with the nature of DNNs. Namely, during the testing phase, we directly obtain the prediction from the output of the DNN without the graphical losses. **3)** We use a differentiable term  $-(\mathbf{Y}_i \mathbf{P}_i)^T \log(\mathbf{Y}_j \mathbf{P}_j)$  to approximate the non-differential term  $\zeta(y_i, y_j) \mathbf{Y}_i \mathbf{Y}_j$  in Eq. (4).

### C. Network Architecture and Loss functions

The network architectures for training and testing are illustrated in Fig. 7, where the black dotted lines denote training flow, and the solid black lines denote inference flow.

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. It is noteworthy that we perform graphical modeling only in the training stage for two reasons. **First**, the graphical module is introduced to generate pseudo labels to supervise the model 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. 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. In the following, we will elaborate the three main modules.

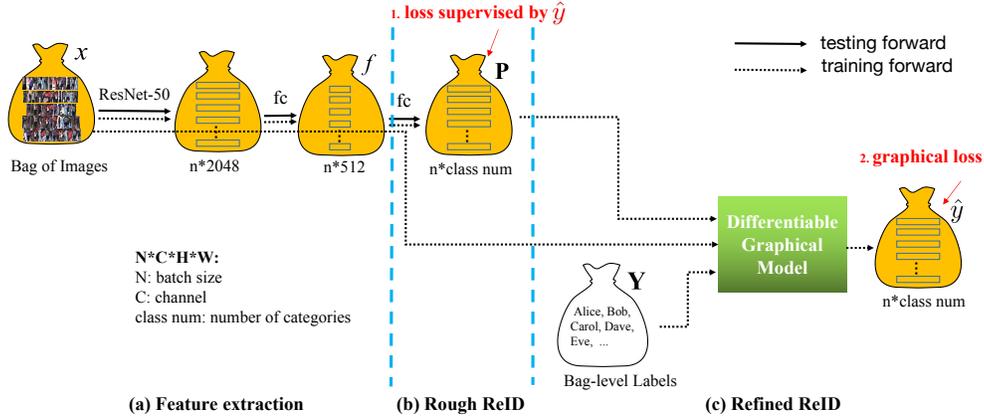


Fig. 7: Diagram of our approach. It mainly consists of three stages, i.e., feature extraction, rough Re-ID, and refined Re-ID. The solid black flow denotes the testing stage, while the black dotted flow denotes the training stage. For simplification, the back-propagation flow is omitted. The loss function is marked with a red arrow.

**Feature embedding module.** Many current best-performing Re-ID models use multi-scale features as feature embeddings [39], 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 [40] as the backbone without any feature pyramid [39]. Our feature embedding is similar to [41]. 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 [42], a Leaky ReLU [43] and a dropout [44]. This module is shown in Fig. 7 (a).

**Rough Re-ID module.** We utilize a softmax classifier for rough Re-ID. Specifically, our model has a fully connected layer at the top of the feature embedding module, which has the same number of units to that of person ID (i.e., ‘class num’ in Fig. 7). A softmax cross-entropy loss is employed for training. The derived person categorization score (e.g.,  $\mathbf{P}$  in Fig. 7) is considered as the rough Re-ID estimation, indicating the possibility of a person ID being present in a bag. This module is shown in Fig. 7 (b).

**Refined Re-ID module.** Here, we aim to generate a pseudo image-level label  $\hat{y}$  for each image by refining the previous estimation results. The refinement has the following inputs:

- 1) *Rough Re-ID score.* As mentioned above, the rough Re-ID module provides a preliminary categorization.
- 2) *Appearance.* Considering that rough Re-ID score is just a high-level abstraction of images, as compensation, we propose to integrate person appearance as low-level information for our refinement.
- 3) *Bag constraint.* Intuitively, our bag constraint eliminates any possibility of assigning a person image with a person ID that is absent in the bag-level annotation; on the contrast, it encourages a person image to be assigned with a person ID that is present in the bag-level annotation.

Accordingly, once the refined pseudo labels are generated, they are used to update the network weights as authentic ground truth:

$$\mathcal{L}_{cls} = - \sum_{i=1}^n (h(\hat{y}_i))^T \log(\mathbf{P}_i). \quad (9)$$

By combining Eq. (9) and Eq. (7), we have the final loss function:

$$\mathcal{L} = w_{cls} \mathcal{L}_{cls} + w_{graph} \mathcal{L}_{graph}, \quad (10)$$

where  $w_{cls}$  and  $w_{graph}$  weights two loss components, respectively. In our experiments, we search the loss weight in a grid of  $\{1:1, 1:0.5, 1:0.1\}$  and find that 1:0.5 has good results on the CUHK03 dataset.

#### D. Weakly-Supervised Triplet Loss

To further improve the performance of the Re-ID model, inspired by multiple granularity network (MGN) [45], we propose a weakly-supervised triplet loss and derive our weakly supervised MGN (W-MGN) for weakly supervised Re-ID. MGN are learned with a triplet loss given strong annotations, while our weakly-supervised triplet loss in W-MGN is developed to address the strict dependency on annotations.

Recall that a fully supervised triplet loss is defined as:

$$\mathcal{L}_{full\_triplet} = \sum_{k=1}^{tp} \left[ \Delta + \max_{j \neq k; j=1, \dots, tp; y_k = y_j} \|z_k - z_j\|_2^2 - \min_{j \neq k; j=1, \dots, tp; y_k \neq y_j} \|z_k - z_j\|_2^2 \right], \quad (11)$$

where  $t$  denotes the number of bags in a training batch,  $\|\cdot\|_2^2$  denotes a  $L_2$ -norm, and  $\Delta$  is a margin. The image-level labels  $y_k$  and  $y_j$  are given for the fully supervised triplet loss, but they are unavailable in the weakly supervised scenario.

To address this problem, we use  $\mathbf{Y}_k^T \mathbf{Y}_j > 0$  to approximate the constraint that  $y_k = y_j$  in Eq. (11), which means that the  $k$ -th and  $j$ -th samples belong to the same class. As  $\mathbf{Y}_k^T \mathbf{Y}_j > 0$  is a necessary but not sufficient condition of the constraint that  $y_k = y_j$ , we relax the max operation to a median operation. Similarly, we use  $\mathbf{Y}_k^T \mathbf{Y}_j = 0$  to approximate  $y_k \neq y_j$ . Accordingly, our weakly-supervised triplet loss is formulated as:

$$\mathcal{L}_{weak\_triplet} = \sum_{k=1}^{tp} \left[ \Delta + \text{median}_{j \neq k; j=1, \dots, tp; \mathbf{Y}_k^T \mathbf{Y}_j > 0} \|z_k - z_j\|_2^2 - \min_{j \neq k; j=1, \dots, tp; \mathbf{Y}_k^T \mathbf{Y}_j = 0} \|z_k - z_j\|_2^2 \right]. \quad (12)$$

Finally, our loss function is

$$\mathcal{L} = w_{cls} \mathcal{L}_{cls} + w_{graph} \mathcal{L}_{graph} + w_{triplet} \mathcal{L}_{weak\_triplet}, \quad (13)$$

where  $w_{triplet}$  is a weight of our triplet loss.

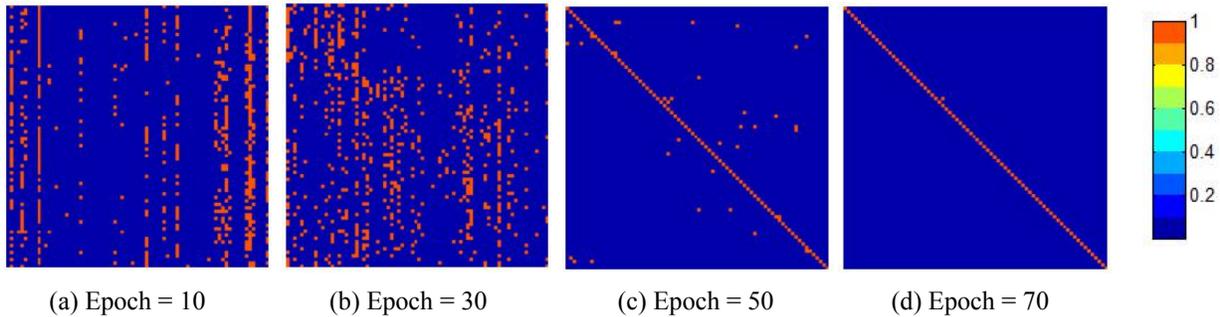


Fig. 8: The effectiveness of our differentiable graphical learning module. Here we show the errors between the rough predictions and the weak annotations in the form of a confusion matrix containing  $76 \times 76$  grids. Each grid indicates a bag of 10 categories, with a total sum of 760 categories, which is approximately equivalent to the person categories in the full training set (i.e., 767 categories).

### E. Computational Complexity

We discuss the computational cost of our weakly supervised Re-ID model. In the training phase, the extra time cost only relates to the generation of pseudo labels, which is a graphical learning module. Generally, graphical learning needs many iterations to search the optimal solution, and thus, the process is time-consuming. However, our approach formulates the differentiable graphical learning as a simple loss. This makes our graphical module very effective. In the experiment section, we will show that our training brings an additional time cost of only  $0.002\times$ . Particularly, in the testing phase, there is no extra time cost because the pseudo label generation component is disabled. In brief, this extra time cost of our method is negligible.

## V. 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/semi-supervised Re-ID. In general, our weakly supervised Re-ID possesses not only cheap annotation but also high accuracy. The details are presented below.

**Unsupervised Re-ID.** To get rid of the prohibitively high cost of manual labeling, unsupervised learning Re-ID proposes to use either local saliency matching models or clustering models [14]. However, without the help of labeled data, it is difficult to model the dramatic variances across camera views in representation/metric learning. Therefore, it is difficult for these pipelines to obtain high accuracies [16]–[19]. In contrast, our weakly supervised Re-ID problem has a better 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 [20], [21] 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 in desire. While in our weakly supervised Re-ID, no accurate label is needed. Second, there is a 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 [22] that matches a target person image with a

bag-level gallery video using multiple-instance multiple-label learning. However, similar to [21], at least one accurate label (of the target person) for each person category is still in a desire to form the probe set in [22]. Hence, mathematically, [22] still belongs to semi-supervised Re-ID but NOT weakly supervised Re-ID.

Section VI-C3 and VI-C1 will compare the accuracy of our weakly supervised Re-ID with previous works.

## VI. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate our weakly supervised Re-ID approach. Section VI-A presents the experimental settings. Section VI-B provides a comprehensive ablation study. Section VI-C presents comparisons of our approach with state-of-the-art methods and also provides the discussion on the computational cost.

### A. Experimental settings

1) *Datasets:* In addition to the proposed SYSU-30k dataset, another four 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 [36], CUHK03 [1], Market-1501 [2], and MSMT17 [34] with weak annotations while their test sets are kept unchanged, as the definition states that during testing, there is no difference between the fully and weakly supervised Re-ID (see Fig. 1 (c)). For a fair comparison (e.g., using the same images for both the fully and weakly supervised Re-ID), we generate the weak annotations from the strong annotations. This includes two steps. **First**, each bag is simulated by randomly selecting several images and packaging them. **Second**, the weak labels are 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  $n$  ID/bag when a bag contains  $n$  person IDs. Note that unless otherwise stated, our weakly supervised learning setting is two IDs/bag.

*PRID2011.* Originally, PRID2011 dataset contains 200 person IDs appearing in at least two camera views and is further randomly divided into training/test sets following the general settings [9], i.e., both having 100 IDs.

*CUHK03.* CUHK03 is a large-scale Re-ID, which contains 14,096 images of 1,467 IDs collected from 5 different pairs

TABLE II: Ablation studies of the proposed weakly supervised Re-ID method. **random**: Each bag contains random person IDs, which reflects the real-world state. **RK**: re-ranking, see [46], one of the effective tricks frequently used in fully supervised Re-ID problems. **\* fully supervised**: when each bag contains only one person ID, 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 IDs. **w/**: with. **w/o**: without.

(a) Impact of bag diversity on PRID2011				(b) Impact of bag diversity on CUHK03			
categories / bag	Rank-1	Rank-5	Rank-10	categories / bag	Rank-1	Rank-5	Rank-10
1 ( <b>* fully supervised</b> )	71.8	91.2	95.9	1 ( <b>* fully supervised</b> )	67.5	88.2	91.8
2	68.0	87.5	94.8	2	61.0	82.0	87.0
3	66.1	86.4	92.3	3	59.4	80.7	86.7
10	49.5	73.9	82.2	10	55.2	79.3	84.5
<b>random (5 on average)</b>	69.3	89.0	94.0	<b>random (5 on average)</b>	60.6	81.6	87.0

(c) Fully supervised learning tricks on PRID2011				(d) Fully supervised learning tricks on CUHK03			
method	Rank-1	Rank-5	Rank-10	method	Rank-1	Rank-5	Rank-10
fully supervised alone	48.9	79.6	88.8	fully supervised alone	52.1	77.9	85.6
weakly supervised alone	39.9	71.2	83.3	weakly supervised alone	44.0	70.6	79.7
fully supervised + <b>RK</b>	71.8	91.2	95.9	fully supervised + <b>RK</b>	67.5	88.2	91.8
weakly supervised + <b>RK</b>	68.0	87.5	94.8	weakly supervised + <b>RK</b>	61.0	82.0	87.0

(e) Scalability of our method on CUHK03				(f) Effectiveness of the graphical learning module on CUHK03			
categories	Rank-1	Rank-5	Rank-10	method	Rank-1	Rank-5	Rank-10
67	16.3	34.7	44.9	w/o graphical model	56.4	80.0	85.1
367	43.6	67.0	75.5	w/o pairwise term	59.2	80.9	86.7
767 ( <b>*full training set</b> )	61.0	82.0	87.0	w/ graphical model	61.0	82.0	87.0

of camera views [1]. Each ID is observed by two disjointed camera views. We follow the **new** standard protocol [46] of CUHK03, i.e., a training set including 767 IDs is obtained without overlap.

*Market-1501*. Market-1501 is another widely-used large-scale Re-ID benchmark, which contains 32,668 images of 1,501 IDs captured from 6 different cameras. The dataset is split into two parts: 12,936 images with 751 IDs for training and 19,732 images with 750 IDs for testing. In testing, 3,368 hand-drawn images with 750 IDs are used as probe set to identify the true IDs on the testing set.

*MSMT17*. MSMT17 is the current largest publicly available Re-ID dataset and there are 126,441 images in MSMT17 in total captured by 15 cameras. It has  $4k$  person IDs, namely, 7.5 times smaller than our SYSU-30k. We follow the standard protocol to split the training and testing set [34].

2) *Implementation details*: The parameters of the ResNet-50 backbone are initialized using ImageNet pre-training. 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. Training on SYSU-30k takes approximately ten days on a single GPU (i.e., NVIDIA TITAN X).

## B. Ablation Study

We first present ablation studies to reveal the benefits of each main component of our method.

1) *Effectiveness of the graphical learning module*: As aforementioned, the graphical learning module plays the role of refining the ID prediction by correcting the errors between the rough Re-ID predictions and the weak annotations, which forms the basis of generating pseudo-image-level labels. We visualize the errors between the rough predictions and the weak annotations in Fig. 8 during training. This experiment is conducted on CUHK03 using the setting of 10 IDs/bag.

Fig. 8 shows the errors between the rough predictions and the weak annotations in the form of a confusion matrix containing  $76 \times 76$  grids. Each grid indicates a bag of 10 IDs, totally summing up to 760 IDs, which approximates the number of person IDs in the full training set. We have two appealing observations from Fig. 8. **First**, there is a significant gap between the rough predictions and the weak annotations (see 8 (a) or (b)), indicating that the rough Re-ID results are still not competent for generating pseudo labels. Therefore, refining the ID prediction is necessary with our graphical learning module. **Second**, the gap between the rough predictions and the weak annotations becomes smaller as the training iteration increases (from 10 epochs in 8(a) to 70 epochs in 8 (d)). When the training model converges, the gap between the ground truth becomes significantly small, which indicates that the problem is well addressed by our graphical learning module.

To further demonstrate the effectiveness of the differential graphical models, we conduct two more empirical studies on CUHK03 and Market-1501 with and without the graphical model, respectively. The graphical module generates pseudo labels to further supervise the learning. Once it is removed, there will be no pseudo labels provided. To cope with this problem, we use the preliminary image-level labels in Eq. (1) as substitutes for pseudo labels, which is also a widely used strategy in existing weakly supervised learning. Table II (f) shows that without the graphical model, there is a significant performance drop on CUHK03, i.e., from 61.0% to 56.4%. Similarly, the performance of W-MGN drops significantly from 95.5% to 88.4% when the graphical model is removed, as shown in Table III (a). These comparisons clearly demonstrate the effectiveness of our graphical model.

2) *Effectiveness of the pairwise term*: As claimed above, the pairwise term is necessary for label smoothness in order to train better models. To validate the necessity of the pairwise term, we conduct two more experiments on CUHK03 and Market-1501 with and without the pairwise term, respectively. Without pairwise terms, a graphical model reduces to isolated

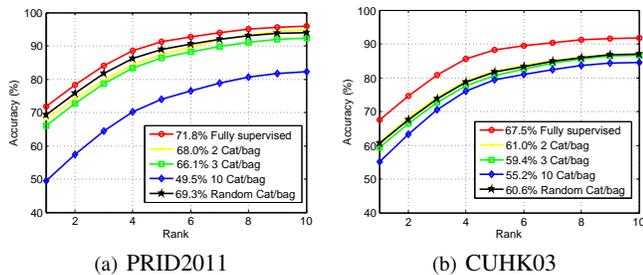


Fig. 9: Analysis on different bag diversities. **Cat/bag**: the number of person IDs in each bag. **Random Cat/bag**: each bag contains random number of person IDs, which reflects the real-world state. **Fully supervised**: each bag contains only one person ID. In this case, the weakly supervised problem degrades into a fully supervised one.

nodes. Experimental results are provided in Table II (f) and Table III (a). For example, on Market-1501 test set, the accuracy drops significantly from 95.5% to 94.0% without the pairwise term.

3) *Effectiveness of the weakly supervised triplet loss*: To validate the effectiveness, we conduct ablation studies on our weakly-supervised triplet loss on five benchmarks (i.e., Market-1501, CUHK03, PRID2011, MSMT17, and SYSU-30k), respectively. Table III (a)-(e) show that W-MGN with the weakly-supervised triplet loss outperforms that without the weakly-supervised triplet loss (i.e., “W-MGN w/o W-Tr”) by a large margin on all the five datasets, which verifies its effectiveness. For example, without the weakly-supervised triplet loss, W-MGN suffers a performance degradation from 95.5% to 92.9% on Market-1501.

4) *Effectiveness of stronger baselines*: To see how stronger baselines affect the performance of the weakly supervised Re-ID, we compare more strong baselines like Local CNN [47] and MGN [45] with our ResNet-50 baseline on the five datasets, i.e., Market-1501, CUHK03, PRID2011, MSMT17, and SYSU-30k. Table III (a)-(e) show that strong baselines for fully supervised Re-ID contribute to the overall performance of the weakly supervised method on all the five datasets. For example, the combination of our weakly supervised Re-ID and MGN (i.e., W-MGN) outperforms the combination of the weakly supervised Re-ID and ResNet-50 (i.e., W-Baseline) by 6.9% (95.5% vs. 88.6%).

5) *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 IDs, 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 IDs, and the number of person IDs is increased to 367 IDs in the second model. The third model is trained with the full CUHK03 training set (i.e., 767 IDs). 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.

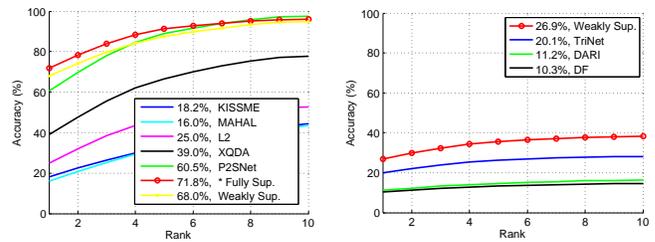


Fig. 10: Comparison with state-of-the-art methods. **Weakly sup.:** the proposed weakly supervised Re-ID approach. **\* Fully Sup.:** each bag contains only one person ID. In this case, the weakly supervised problem degrades into a fully supervised one.

6) *Impact of bag diversity*: Intuitively, if a bag contains more person IDs, it is more challenging to learn a weakly supervised Re-ID model because of the increase in uncertainty. Next, we investigate the performance with respect to such bag internal diversity. We conduct experiments on PRID2011 and CUHK03. In Table II (a)-(b) and Fig. 9 (a)-(b), we compare five options, i.e., each bag containing 1, 2, 3, 10, or a random number of person IDs, respectively. In particular, when each bag has only one person ID, the weakly supervised Re-ID problem degrades into a fully supervised one.

We have three major observations from Table II (a)-(b) and Fig. 9 (a)-(b). **First**, the models trained with weakly labeled samples achieves comparable accuracies to the models trained with strongly labeled data (e.g., 68.0% vs. 71.8% in Table II (a)). This result is quite important because a weak annotation costs much less money and time than a strong annotation.

**Second**, the accuracy of the weakly supervised methods gradually decreases as the number of IDs in each bag increases. In particular, the rank-1 accuracy of our approach drops by 18.5% when increasing the number of IDs per bag from 2 to 10 in Table II (a). We argue that the increase in uncertainty causes this optimization difficulty. When the IDs per bag increases, the uncertainty in the label assignment also increases, 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 a random number of person IDs (5 IDs on average), which reflects the real-world states. The high performance suggests that solving a weakly supervised Re-ID problem is feasible and appealing in reality.

7) *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 accuracy of our approach with respect to models with different fully-supervised learning capacities. Experiments are conducted on PRID2011 and CUHK03.

We first evaluate two different fully supervised learning baseline models with and without re-ranking post-process. Then, we evaluate them in the weakly supervised learning scenario. The setting is similar to the aforementioned fully supervised learning, except that all of the image-level annotations are replaced with bag-level annotations in the training

TABLE III: Comparison with state-of-the-art methods. **W-Baseline**: our weakly supervised Re-ID method. **Baseline**: each bag contains only one person ID. 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. **w/o Tri**: without triplet loss. **x(y)**: y is the number reported by the original paper; x is the result of our reproduction. **RK**: re-ranking. **‡**: pretrained on CUHK03. **w/**: with. **w/o**: without. **W-**: weakly supervised version of a method.

(a) Market-1501 test set.			(b) CUHK03 test set.			(d) MSMT17 test set.									
Supervision	Method	Rank-1	Supervision	Method	Rank-1	Supervision	Method	Rank-1							
Fully	MSCAN [48]	80.3	Fully	BOW+XQDA [2]	6.4	Fully	PDC [49]	58							
	DF [50]	81.0		PUL [18]	9.1		GLAD [51]	61.4							
	SSM [52]	82.2		LOMO+XQDA [53]	12.8		ABD-Net [54]	82.3							
	SVDNet [55]	82.3		IDE(R) [56]	21.3		GoogleNet [57]	47.6							
	GAN [33]	84.0		IDE+DaF [58]	26.4		IANet [59]	75.5							
	PDF [49]	84.1		IDE+XQ- <b>RK</b> [46]	34.7		SFT [60]	73.6							
	TriNet [41]	84.9		PAN	36.3		Auto-ReID [61]	78.2							
	TriNet+Era- <b>RK</b> [62]	85.5		DPFL [63]	40.7		MVP [64]	71.3							
	PCB [39]	93.4		TreeConv [65]	71.4		Verif-Identif [66]	60.5							
	VPM [67]	93.0		TriNet+Era. [62]	55.5		PCB [39]	68.2							
	JDGL [68]	94.8		ACNet [69]	64.8		JDGL [68]	77.2							
	AANet [70]	92.4		<b>Baseline</b>	67.5		ShuffleNet [71]	41.5							
	Local CNN [47]	95.9(97.0)		Local CNN [47]	69.6		Local CNN [47]	82.9							
	MGN [45]	95.8(96.6)		MGN [45]	70.4(68.0)		MGN [45]	83.1							
	MGN w/o Tri	93.4		MGN w/o Tri	67.6		MGN w/o Tri	81.0							
<b>Baseline</b>	94.2	CAMEL [14]	31.9	MobileNetV2 [72]	50.9										
Unsupervised	CAMEL [14]	54.5	Unsupervised	PatchNet [73]	45.4	Unsupervised	OSNet [74]	78.7							
	TAUDL [75]	63.7		PAUL [73]	52.3		PTGAN [34]	11.8							
	UTAL [19]	69.2		<b>W-Baseline</b>	61.0		SSG [76]	41.6							
	UDA [16]	75.8	W-Local CNN	67.6	TAUDL [77]		28.4								
	MAR [17]	67.7	W-MGN	69.8	UTAL [19]		31.4								
	DECAMEL [78]	60.2	W-MGN w/o W-Tri	67.2	ECN [79]		30.2								
	ECN [79]	75.1	(c) PRID2011 test set	Supervision	Method		Rank-1	Weakly	W-MGN	81.1					
	PAUL [73]	68.5							KISSME [5]	18.2	W-Local CNN	80.6			
	HHL [81]	62.2							MAHAL	16	W-MGN w/o W-Tri	77.9			
	Distilled [82]	61.5							L2	25	(e) SYSU-30k	Supervision	Method	Rank-1	
	Smoothing [83]	83.0							XQDA [53]	39					Fully
SPACO [84]	68.3	P2SNet [9]				60.5			‡DF [6]	10.3					
HHL [81]	54.4	<b>Baseline</b>				71.8			‡ <b>Baseline</b>	20.1					
Distilled [82]	63.9	MGN [45]				74.6			‡Local CNN [47]	23.0					
One Example [20]	70.1	MGN w/o Tri				72.5			‡MGN [45]	23.6					
Many Examples [20]	82.5	Local CNN [47]				74.2			‡MGN w/o Tri	21.5					
<b>W-Baseline</b>	88.6	W-Local CNN				71.6			<b>W-Baseline</b>	26.9					
W-Local CNN	95.7	<b>W-Baseline</b>	68.0	W-Local CNN	28.8										
W-MGN	95.5	W-MGN	72.7	W-MGN	29.5										
W-MGN w/o W-Tri	92.9	W-MGN w/o W-Tri	70.7	W-MGN w/o W-Tri	26.7										
W-MGN w/o graph	88.4	Weakly	Supervision	Method	Rank-1										
W-MGN w/o pair	94.0					KISSME [5]	18.2								
Semi	Supervision					Method	Rank-1	MAHAL	16						
								L2	25						
								XQDA [53]	39						
								P2SNet [9]	60.5						
								<b>Baseline</b>	71.8						
								MGN [45]	74.6						
								MGN w/o Tri	72.5						
								Local CNN [47]	74.2						
								W-Local CNN	71.6						
		<b>W-Baseline</b>	68.0												
		W-MGN	72.7												
W-MGN w/o W-Tri	68.4														
W-MGN w/o graph	88.4														
W-MGN w/o pair	94.0														

set.

It is observed that our rank-1 accuracy with weak annotations is close to that of using strong annotations, as shown in Table II (c)-(d). Moreover, weakly supervised learning with a stronger baseline (‘weakly supervised + **RK**’) yields better performance. For example, in the weak annotation setting, ‘‘weakly supervised + **RK**’’ yields 68.0% on PRID2011, 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; namely, existing tricks (e.g., re-ranking) for fully supervised learning could also be applied to the weakly supervised Re-ID.

### C. Comparison with the State-of-the-Arts

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

1) *Accuracy on Market-1501*: Our weakly supervised Re-ID is compared with state-of-the-art fully-supervised /unsupervised / semi-supervised methods.

**Fully supervised Re-ID.** We compare our method with the fully supervised Re-ID models. Fifteen representative state-of-the-art methods are used as comparison methods, including MSCAN [48], DF [50], SSM [52], SVDNet [55], GAN [33], PDF [49], TriNet [41], TriNet + Era. + reranking [62], PCB [39], VPM [67], JDGL [68], AANet [70], Local CNN [47],

and MGN [45]. Comparison results are provided in Table III (a). Our approach achieves very competitive accuracy. For example, our W-MGN and W-Local CNN achieve respectively a rank-1 accuracy of 95.5% and 95.7%, which surpass many of the compared fully-supervised methods. These results verify 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 our method with state-of-the-art unsupervised and semi-supervised Re-ID methods.

**Unsupervised Re-ID.** In Table III (a), we compare our method with 11 current best-performing models for unsupervised Re-ID, including CAMEL [14], TAUDL [75], UTAL [19], UDA [16], MAR [17], DECAMEL [78], ECN [79], PAUL [73], HHL [81], Distilled [82], and Smoothing [83]. The results in Table III (a) show that our weakly supervised Re-ID has obtained significant gain over unsupervised Re-ID methods. For instance, our W-MGN outperforms the best-performing model UDA [16] by a large margin (i.e., 19.7%). 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.

**Semi-supervised Re-ID.** In Table III (a), 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 [84], HHL [81], Distilled [82], One Example [20], and Many Examples [20]. 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” [20] by a large margin (i.e., 13%), indicating that as an annotation-saving method, our weakly Re-ID obtains higher accuracy than semi-supervised Re-ID.

2) *Accuracy on CUHK03*: Our weakly supervised Re-ID is compared with state-of-the-art methods in two groups on CUHK03, including the traditional fully-supervised Re-ID and the unsupervised Re-ID. The semi-supervised Re-ID is not compared here because existing works on semi-supervised Re-ID do not provide results on this dataset.

**Fully supervised Re-ID.** In Table III (b), we compare our method with the eleven current best models, including BOW+XQDA [2], PUL [18], LOMO+XQDA [53], IDE(R) [56], IDE+DaF [58], IDE+XQ+reranking [46], PAN, DPFL [63], and newly proposed methods such as SVDNet [55], TriNets [62], ACNet [69], Local CNN [47], MGN [45], and TreeConv [65]. Our W-MGN and W-Local CNN achieve a rank-1 accuracy of 69.8% and 67.6%, which surpass many of the compared fully-supervised methods. These results verify the effectiveness of our method.

**Unsupervised Re-ID.** In Table III (b), we compare our method with the unsupervised Re-ID models. Three representative state-of-the-art methods are used as competing methods, including CAMEL [14], PatchNet [73], and PAUL [73]. The results in Table III (b) show that our weakly supervised Re-ID problem has obtained significant gain over unsupervised Re-ID methods. For instance, our W-MGN outperforms the best-performing model PAUL [73] by a large margin (i.e., 17.5%). Given that our method also saves the annotation effort, we believe our weakly-supervised Re-ID balances well between annotation and accuracy.

3) *Accuracy on MSMT17*: Our weakly supervised Re-ID is compared with state-of-the-art methods fully-supervised / unsupervised Re-ID on MSMT17. Semi-supervised methods are not compared here because previous works on semi-supervised Re-ID do not provide results on this dataset.

**Fully supervised Re-ID.** We compare our method with the eleven current best models, including PDC [49], GLAD [51], ABD-Net [54], GoogleNet [57], IANet [59], SFT [60], Auto-ReID [61], MVP [64], Verif-Identif [66], PCB [39], JDGL [68], ShuffleNet [71], MobileNetV2 [72], OSNet [74], Local CNN [47], and MGN [45]. Comparison results are provided in Table III (d). Our W-MGN and W-Local CNN achieve a rank-1 accuracy of 81.1% and 80.6%, which surpass many of the compared fully-supervised methods. These results verify the effectiveness of our method.

**Unsupervised Re-ID.** We compare our method with the unsupervised Re-ID models, as shown in Table III (b). Five representative state-of-the-art methods are used as competing methods, including PTGAN [34], SSG [76], TAUDL [77], UTAL [19], ECN [79], and UGA [80]. The results in Table III (d) show that our weakly supervised Re-ID problem has obtained significant gain over unsupervised Re-ID methods. For instance, our W-MGN outperforms the best-performing model UGA [80] by a large margin (i.e., 81.1% vs. 49.5%). Given that our method also saves the annotation effort, we

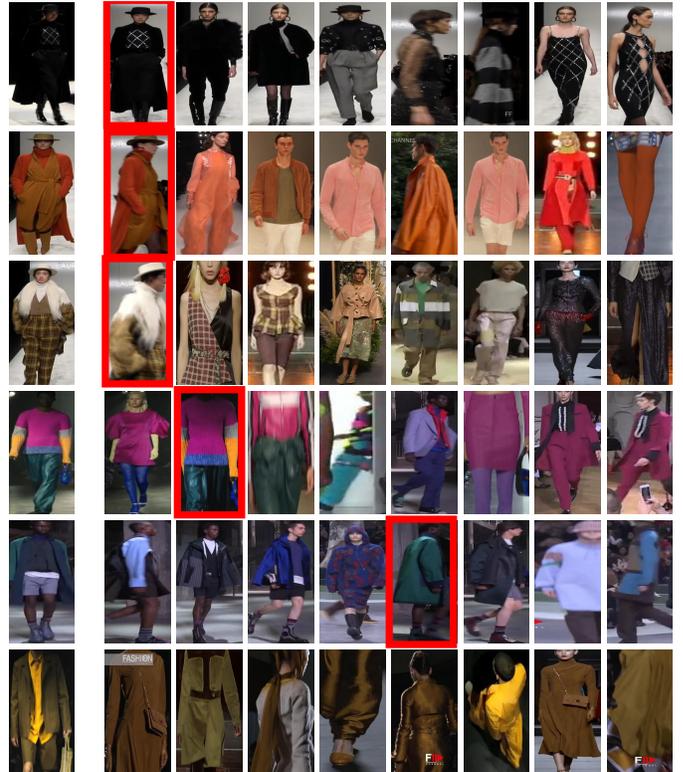


Fig. 11: Search examples of W-MGN on SYSU-30k dataset. Each row represents a ranking result with the first image being the query and the rest images being the returned list. The image with the red bounding box is the matched one.

believe our weakly supervised Re-ID balances well between annotation and accuracy.

4) *Accuracy on PRID2011*: In Table III (c) and Fig. 10 (a), we compare the results of our model with five current best models: the KISSME distance learning method [5], MAHAL, L2, and XQDA [53], P2SNet [9], Local CNN [47], and MGN [45]. For KISSME, MAHAL, L2, and XQDA, deep features [85] 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. Our W-MGN and W-Local CNN achieve a rank-1 accuracy of 72.7% and 71.6%, which surpass many of the compared fully-supervised methods. These results verify the effectiveness of our method.

5) *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 traditional fully supervised Re-ID models with our weakly supervised method by using transfer learning. Specifically, six representative fully supervised Re-ID models including DARI [7], DF [6], TriNet [62], Local CNN [47], MGN [45], and MGN without triplet loss 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 (e) and Fig. 10 (b) are the results of the comparisons. It is observed that our W-MGN achieves state-of-the-art performance (29.5%), even though it is trained in a

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

weakly supervised manner while the comparison methods 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 though SYSU-30k is annotated weakly.

We also qualitatively present some query examples of W-MGN for the SYSU-30k dataset in Fig. 11. Each row represents a ranking result with the first image being the query image and the rest being the result list. The matched one in the returned list is highlighted by a red bounding box. This figure exhibits the difficulty of this dataset. Actually, in the failed examples, the images ranked higher than the matched one often look more closer to the query image as in Row 46.

6) *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.).

## VII. CONCLUSION

We have considered a 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. 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. Extensive experiments have conducted on our SYSU-30k dataset and other four public Re-ID datasets and a superior performance is achieved with our proposed model, providing a promising and appealing conclusion that learning a Re-ID model with less annotation efforts is possible and feasible. Future work will include building automated models [86] for the weakly supervised Re-ID .

## ACKNOWLEDGEMENT

This work was supported in part by the State Key Development Program under Grant 2018YFC0830103, in part by the National Natural Science Foundation of China under Grant 61876224, U1811463, 61622214, and 61836012, and in part by Guangdong Natural Science Foundation under Grant 2017A030312006.

## REFERENCES

- [1] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.
- [2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 1116–1124.
- [3] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 262–275.
- [4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning*, 2001, pp. 282–289.
- [5] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.
- [6] 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.
- [7] G. Wang, L. Lin, S. Ding, Y. Li, and Q. Wang, "Dari: Distance metric and representation integration for person verification," in *Proceedings of The AAAI Conference on Artificial Intelligence*, 2016, pp. 3611–3617.
- [8] 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.
- [9] 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*, pp. 2777–2787, 2017.
- [10] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proceedings of The AAAI Conference on Artificial Intelligence*, 2019.
- [11] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2018, pp. 1–6.
- [12] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.
- [13] H. Wang, G. Wang, Y. Li, D. Zhang, and L. Lin, "Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [14] 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, pp. 994–1002.
- [15] W. Liang, G. Wang, J. Lai, and J. Zhu, "M2M-GAN: many-to-many generative adversarial transfer learning for person re-identification," *CoRR*, vol. abs/1811.03768, 2018.
- [16] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *CoRR*, vol. abs/1807.11334, 2018.
- [17] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, and J. Lai, "Unsupervised person re-identification by soft multilabel learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2148–2157.
- [18] 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.
- [19] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. pages 11, 2019.
- [20] 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.
- [21] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5177–5186.
- [22] J. Meng, S. Wu, and W. Zheng, "Weakly supervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 760–769.
- [23] D. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly

- supervised pretraining,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 185–201.
- [24] G. Wang, X. Xie, J. Lai, and J. Zhuo, “Deep growing learning,” in *IEEE International Conference on Computer Vision*, 2017, pp. 2812–2820.
- [25] G. Wang, P. Luo, L. Lin, and X. Wang, “Learning object interactions and descriptions for semantic image segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5859–5867.
- [26] P. Luo, G. Wang, L. Lin, and X. Wang, “Deep dual learning for semantic image segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2718–2726.
- [27] L. Lin, G. Wang, R. Zhang, R. Zhang, X. Liang, and W. Zuo, “Deep structured scene parsing by learning with image descriptions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2276–2284.
- [28] R. Zhang, L. Lin, G. Wang, M. Wang, and W. Zuo, “Hierarchical scene parsing by weakly supervised learning with image descriptions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 596–610, 2019.
- [29] 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.
- [30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deepplab: 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.
- [31] B. Wu, Y. Zhang, B. Hu, and Q. Ji, “Constrained clustering and its application to face clustering in videos,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3507–3514.
- [32] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, “Person re-identification with deep similarity-guided graph neural network,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2018, pp. 508–526.
- [33] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in *IEEE International Conference on Computer Vision*, 2017, pp. 3774–3782.
- [34] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 79–88.
- [35] W. Li, R. Zhao, and X. Wang, “Human reidentification with transferred metric learning,” in *Proceedings of the Asian Conference on Computer Vision*, 2012, pp. 31–44.
- [36] M. Hirzer, C. Belezni, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Scandinavian conference on Image analysis*, 2011, pp. 91–102.
- [37] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” in *Proceedings of The British Machine Vision Conference*, vol. 1, no. 2, 2011, p. 6.
- [38] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.
- [39] 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*, 2018, pp. 480–496.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [41] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *CoRR*, vol. abs/1703.07737, 2017.
- [42] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 448–456.
- [43] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 807–814.
- [44] 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.
- [45] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in *ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 274–282.
- [46] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3652–3661.
- [47] J. Yang, X. Shen, X. Tian, H. Li, J. Huang, and X. Hua, “Local convolutional neural networks for person re-identification,” in *ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 1074–1082.
- [48] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.
- [49] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Pose-driven deep convolutional model for person re-identification,” in *IEEE International Conference on Computer Vision*, 2017, pp. 3980–3989.
- [50] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in *IEEE International Conference on Computer Vision*, 2017, pp. 3239–3248.
- [51] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, “GLAD: global-local-alignment descriptor for pedestrian retrieval,” in *ACM Multimedia Conference on Multimedia Conference*, 2017, pp. 420–428.
- [52] S. Bai, X. Bai, and Q. Tian, “Scalable person re-identification on supervised smoothed manifold,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2530–2539.
- [53] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.
- [54] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, “Abd-net: Attentive but diverse person re-identification,” in *IEEE International Conference on Computer Vision*, 2019.
- [55] Y. Sun, L. Zheng, W. Deng, and S. Wang, “Svdnet for pedestrian retrieval,” in *IEEE International Conference on Computer Vision*, 2017, pp. 3820–3828.
- [56] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” *CoRR*, vol. abs/1610.02984, 2016.
- [57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [58] R. Yu, Z. Zhou, S. Bai, and X. Bai, “Divide and fuse: A re-ranking approach for person re-identification,” in *Proceedings of The British Machine Vision Conference*, 2017.
- [59] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “Interaction-and-aggregation network for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9317–9326.
- [60] C. Luo, Y. Chen, N. Wang, and Z. Zhang, “Spectral feature transformation for person re-identification,” in *IEEE International Conference on Computer Vision*, 2019.
- [61] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, “Auto-reid: Searching for a part-aware convnet for person re-identification,” in *IEEE International Conference on Computer Vision*, 2019.
- [62] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” *CoRR*, vol. abs/1708.04896, 2017.
- [63] Y. Chen, X. Zhu, and S. Gong, “Person re-identification by deep learning multi-scale representations,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2590–2600.
- [64] H. Sun, Z. Chen, S. Yan, and L. Xu, “Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification,” in *IEEE International Conference on Computer Vision*, 2019.
- [65] G. Wang, G. Wang, K. Wang, X. Liang, and L. Lin, “Grammatically recognizing images with tree convolution,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Diego, CA, USA, August 4-8, 2020. ACM*, 2020.
- [66] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned cnn embedding for person reidentification,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, pp. 13:1–13:20, 2018.
- [67] 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,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 393–402.
- [68] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147.
- [69] G. Wang, K. Wang, and L. Lin, “Adaptively connected neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1781–1790.

- [70] C. Tay, S. Roy, and K. Yap, "Aanet: Attribute attention network for person re-identifications," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7134–7143.
- [71] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.
- [72] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [73] Q. Yang, H. Yu, A. Wu, and W. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3633–3642.
- [74] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *IEEE International Conference on Computer Vision*, 2019.
- [75] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 772–788.
- [76] Y. Fu, Y. Wei, G. Wang, X. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *IEEE International Conference on Computer Vision*, 2019.
- [77] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 772–788.
- [78] 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.
- [79] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 598–607.
- [80] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li, "Unsupervised graph association for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [81] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 176–192.
- [82] A. Wu, W. Zheng, X. Guo, and J. Lai, "Distilled person re-identification: Towards a more scalable system," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1187–1196.
- [83] G. Wang, J. Lai, W. Liang, and G. Wang, "Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [84] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong, "Self-paced co-training," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 2275–2284.
- [85] 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 *Proceedings of the European Conference on Computer Vision*, 2016, pp. 868–884.
- [86] C. Li, L. Yuan, J. Peng, G. Wang, X. Liang, L. Lin, and X. Chang, "Block-wisely supervised neural architecture search with knowledge distillation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.



**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 academic journals and conferences on image processing and pattern recognition.



**Zhengtao Yu** received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language process, image processing, and machine learning.



**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 the 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 the 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.