Weakly Supervised Person Re-Identification

Jingke Meng[1,3] , Sheng Wu[1] , and Wei-Shi Zheng[1,2*]

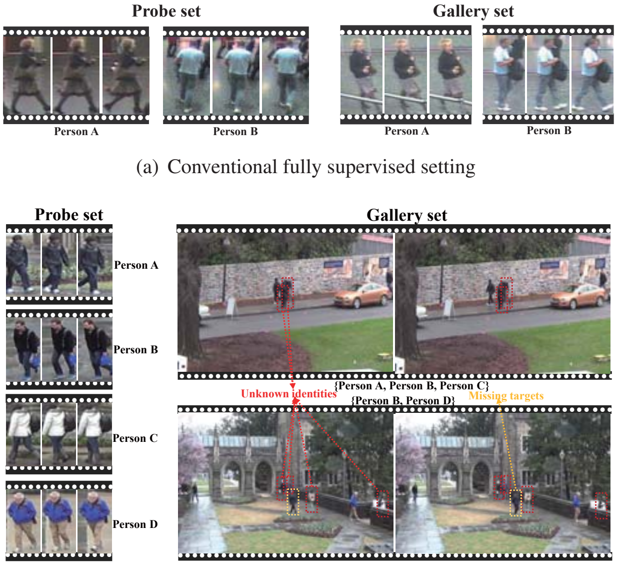
[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[3] Accuvision Technology Co. Ltd, China
mengjke@mail2.sysu.edu.cn, wush43@mail2.sysu.edu.cn, wszheng@ieee.org

Abstract

In the conventional person re-id setting, it is assumed that the labeled images are the person images within the bounding box for each individual; this labeling across multiple nonoverlapping camera views from raw video surveillance is costly and time-consuming. To overcome this difficulty, we consider weakly supervised person re-id modeling. The weak setting refers to matching a target person with an untrimmed gallery video where we only know that the identity appears in the video without the requirement of annotating the identity in any frame of the video during the training procedure. Hence, for a video, there could be multiple video-level labels. We cast this weakly supervised person re-id challenge into a multi-instance multi-label learning (MIML) problem. In particular, we develop a Cross-View MIML (CV-MIML) method that is able to explore potential intraclass person images from all the camera views by incorporating the intra-bag alignment and the cross-view bag alignment. Finally, the CV-MIML method is embedded into an existing deep neural network for developing the Deep Cross-View MIML (Deep CV-MIML) model. We have performed extensive experiments to show the feasibility of the proposed weakly supervised setting and verify the effectiveness of our method compared to related methods on four weakly labeled datasets.

1. Introduction

Given an image from a set of probe images, the objective of person re-identification (re-id) is to identify the same person across a set of gallery images from nonoverlapping camera views. The changes in illumination, camera viewpoint, background and occlusions lead to considerable visual ambiguity and appearance variation and make person re-id a challenging problem. Several representative methods [33, 32, 45, 20] have been developed to solve this problem.



(a) Conventional fully supervised setting

(b) Proposed weakly supervised setting

Figure 1. Comparison of two settings. (a) Conventional fully supervised setting: image sequences in the probe and gallery set are manually trimmed and labeled from video surveillance in a frame-by-frame manner. (b) Proposed weakly supervised setting: the untrimmed videos in the gallery set are tagged by multiple video-level labels, while the specific label of each individual is absent from the labeling process.

While numerous methods have been developed for fully supervised person re-id, conventionally, it is assumed that for model training, 1) the images in the probe set and gallery set are manually trimmed and labeled from raw video surveillance (probably with the assistance of detection) frame-by-frame (as shown in Figure 1(a)), and 2) all training samples are of the target to be matched, and no outliers exist. Although such precise annotations could eliminate the difficulty of learning robust person re-id models, they require strong supervision, which makes the entire learning process difficult to adapt to large-scale person re-id in a more practical and challenging scenario.

[*] Corresponding author

Instead of relying on costly labeling/annotations, we wish to investigate the person re-id modeling in a weakly supervised setting. This setting assumes that annotators only need to take a rough glance at the raw videos to determine which identities appear in such videos, and they do not need to annotate the identity in any frame of the video. That is, only the video-level label indicating the presence of the identity is given, while the ground-truth regarding in which frame and which bounding box in a frame the identity is present is not provided. In such a setting, the labeling cost of person re-id can be greatly reduced compared to the conventional fully supervised setting. We call this setting ***weakly supervised person re-id***.

More specifically, as shown in Figure 1(b), the first row of a video clip in the gallery set is annotated with a set of video-level labels {Person A, Person B, Person C} indicating that Person A, Person B and Person C have appeared in this video clip, but there is no additional prior knowledge that precisely indicates which individual is Person A, Person B or Person C. Hence, these labels are weak. Note that it is possible that some labels for a video are missing because the annotators fail to recognize (*e.g*., pedestrians framed by yellow dotted lines in Figure 1(b)). It is also practically possible that unknown identities appear in the untrimmed video clips (*e.g*., pedestrians framed by red dotted lines in Figure 1(b)). Overall, the videos in the gallery set are untrimmed and tagged with the multiple video-level weak labels in this weakly supervised setting. Based on this setting, we aim to find in the gallery the raw videos where the target person appears, given a probe set of images from nonoverlapping camera views.

To solve the problem of weakly supervised person re-id, we consider every video clip in the gallery set as a bag; each bag contains multiple instances of the person images detected in each raw video clip and associates with multiple bag-level labels. For the probe set, it contains the target individuals to be searched for in the gallery; thus, each input is a set of manually trimmed images of the target person. For convenience, we also regard the probe input as a bag. We consider the whole weakly supervised person re-id problem as a multi-instance multi-label learning (MIML) problem and develop a *Cross-View MIML (CV-MIML) method*. Compared to existing MIML algorithms [3, 2, 16, 15, 26, 46, 10], our CV-MIML is able to exploit similar instances within a bag for intra-bag alignment and mine potential matched instances between bags that are captured across camera views through embedding distribution prototype into MIML, which is called the cross-view bag alignment in our modeling. Finally, we embed this CV-MIML method into a deep neural network to form an end-to-end deep cross-view multi-label multi-instance learning (Deep CV-MIML) model.

To the best of our knowledge, this paper is the first to propose and study the weakly supervised problem in person re-id. We have performed comprehensive experiments on four datasets with one genuine dataset and three simulated datasets. Since existing person re-id methods do not suit the weakly supervised setting, we compare the proposed method to other state-of-the-art MIML methods and several state-of-the-art one-shot, unsupervised and sully supervised person re-id methods. The results demonstrate the feasibility of the weakly supervised person re-id method and show that the proposed Deep CV-MIML model is a superior approach to solving the problem.

2. Related Work

2.1. Person Re-identification

Most studies of person re-id are supervised [43, 31, 7, 33, 32, 45, 20, 35, 28, 5] and require annotating each person in the video precisely (e.g., indicating the frame and the position in the frame within the video). It is impractical to extend to the above person re-id methods in a more practical and challenging scenario due to the expensive cost of the labeling process. So we propose the weakly supervised setting for person re-id which only requires video-level weak labels.

Recently, several unsupervised learning methods have been developed to learn person re-id models [40, 9, 24, 23, 39, 19, 4]. The general idea of these methods is to explore unlabeled data progressively by alternately assigning pseudo-labels to unlabeled data and updating the model according to these pseudo-labeled data. The unsupervised learning process can be easily adapted to large-scale person re-id since the unlabeled data can be accessed without manual operations. However, the performance of these unsupervised methods is limited because the visual ambiguity and appearance variations are not easy to address due to the lack of clear supervised information.

In the weakly supervised setting, the gallery set is composed of the raw videos, which is closely related to the person search [34] that aims to search for the target person from the whole raw images. However, in the setting of the person search, the manually annotated bounding boxes for the gallery set are required to train the model in a fully supervised manner, which is much more supervised than our weakly supervised setting.

2.2. Multi-Instance Multi-Label Learning

In general, an object of interest has its inherent structure and it can be represented as a bag of instances with multiple labels associated on the bag level. Multi-Instance Multi-Label learning (MIML) [44] provides a framework for handling this kind of problems. Due to the limitation of the current person re-id methods in the weakly supervised setting, we adopt the MIML formulation to solve our weakly supervised re-id problem. During the past few years, many related algorithms have been investigated and developed for MIML problems [3, 2, 16]. The MIML formulation has also been applied in many practical vision domains, such as image annotation [36, 25] and classification tasks [37, 6, 38, 41].

While it is possible to apply existing MIML to our problem, there still exist several intractable issues that may not be readily resolved because of the following: 1) the existing MIML methods ignore mining the intra-bag variation between similar instances belonging to the same person; 2) previous approaches are based on the idea that highly relevant labels mean sharing common instances among the corresponding classes, but the class labels are independent from each other in person re-id; and 3) most MIML methods are not able to mine potential matched instances between bags effectively when applied to person re-id for cross-view matching. The proposed Deep Cross-View MIML model for the person re-id can overcome the above limitations by exploiting similar instances within a bag for intra-bag alignment and mining potential matched instances across camera views simultaneously.

3. The Proposed Approach

In this section, we formally introduce the weakly supervised person re-id setting and then introduce the Deep CV-MIML model for addressing this problem.

3.1. Problem Statement and Notation

In the weakly supervised person re-id setting, our goal is to find the videos that the target person appears in, given a probe set of images from nonoverlapping camera views. Suppose that we have C known identities from V camera views and that every known identity appears in at least two camera views. Since some unknown identities (*e.g.*, pedestrians framed by red dotted lines in Figure 1(b)) would appear in the untrimmed videos, these unknown identities can be affiliated to a novel class; we define an extra 0-class to represent it. For simplicity, we denote the overall number of classes by $\tilde{C} = C + 1$.

In our learning, given $N_{\mathcal{X}}$ videos, the training set $\mathcal{X}$ consists of two distinct parts: the probe set $\mathcal{X}_p$ and the gallery set $\mathcal{X}_g$. The videos in the gallery set are untrimmed and tagged with the multiple video-level weak labels that indicate the presence of individuals as shown in Figure 1(b); the person images within a raw video in the gallery set are automatically detected in advance. Note that even though the person images are detected during this stage, the specific label of each individual is still unknown.

We consider every raw video as a bag; each bag contains multiple instances of the person images detected in each video. For the probe set, each query is composed of a set of detected images of the same person. For convenience, we also regard each query in the probe set as a bag. More specifically, the training set can be denoted by $\mathcal{X} = \{\mathcal{X}_p, \mathcal{X}_g\}$, where the probe set is $\mathcal{X}_p = \{(X_b, \mathbf{y}_b, v_b)\}_{b=1}^{N_p}$ and the gallery set is $\mathcal{X}_g = \{(X_b, \mathbf{y}_b, v_b)\}_{b=1}^{N_g}$, $N_{\mathcal{X}} = N_p + N_g$. For the bags (videos) in the probe set, each bag X_b containing the same person images is labeled by $\mathbf{y}_b$ under the v_b-th camera view, where $v_b \in \{1, 2, ..., V\}$, and $\mathbf{y}_b = [y_b^0, y_b^1, ..., y_b^C] \in \{0, 1\}^{\tilde{C}}$ is a label vector containing $\tilde{C}$ class labels, in which $y_b^c = 1$ if the c-th label is tagged for X_b, and $y_b^c = 0$ otherwise. In contrast to the conventional person re-id, for the bags (videos) in the gallery set, $y_b^c = 1$ denotes that the c-th identity appears in this bag (video), while $y_b^c = 0$ denotes uncertainty of whether the c-th identity has appeared in this video. Moreover, the bag X_b consists of n_b instances $\mathbf{x}_{b,1}$, $\mathbf{x}_{b,2}$, ..., $\mathbf{x}_{b,i}$, ..., $\mathbf{x}_{b,n_b}$, where $\mathbf{x}_{b,i} = f_e(\mathbf{I}_{b,i}; \theta) \in \mathbb{R}^d$ is the feature vector extracted from the corresponding person image $\mathbf{I}_{b,i}$, and $f_e(\cdot; \theta)$ is a learnable feature extractor.

3.2. Cross-View MIML for Person Re-id

We cast the weakly supervised person re-id as the problem of multi-instance multi-label learning (MIML) and present the cross-view multi-instance multi-label (CV-MIML) learning method to solve this problem.

3.2.1 Weakly Supervised Person Re-id by MIML

For the task of weakly supervised classification, we formulate a MIML classifier for our weakly supervised person re-id. With this classifier $f_c(\cdot; W)$, the high-dimensional input $\mathbf{x}_{b,i} \in \mathbb{R}^d$ can be transformed into a $\tilde{C}$-dimensional vector $\tilde{\mathbf{y}}_{b,i} = f_c(\mathbf{x}_{b,i}; W) \in \mathbb{R}^{\tilde{C}}$ that can be interpreted as a label distribution, embedding the similarities among all classes.

For the probe set $\mathcal{X}_p$, all instances $\{\mathbf{x}_{b,i}\}_{i=1}^{n_b}$ in each bag X_b are tagged with the same label $\mathbf{y}_b$. For the gallery set $\mathcal{X}_g$, all instances $\{\mathbf{x}_{b,i}\}_{i=1}^{n_b}$ in each bag X_b share the same weak video-level label $\mathbf{y}_b$. The softmax classifier cannot be directly applied to the instances in the gallery set because the specific label of each instance is absent. Therefore, the instances from the probe set $\mathcal{X}_p$ and gallery set $\mathcal{X}_g$ are separately processed by the following two procedures to learn a classification model.

On the one hand, we expect the estimated label distribution to eventually approximate the true one; thus, the classification loss for these instances in the probe set can be written as follows:

$$\mathcal{L}_p = \frac{1}{N_p} \sum_{X_b \in \mathcal{X}_p} \sum_{i \in \{1, \cdots, n_b\}} \sum_{c \in \{0, \cdots, C\}} (-y_b^c \log \tilde{y}_{b,i}^c), \tag{1}$$

where y_b^c denotes the ground truth video-level labels of bag X_b at the c-th entry, $\tilde{y}_{b,i}^c$ is the c-th estimated probability

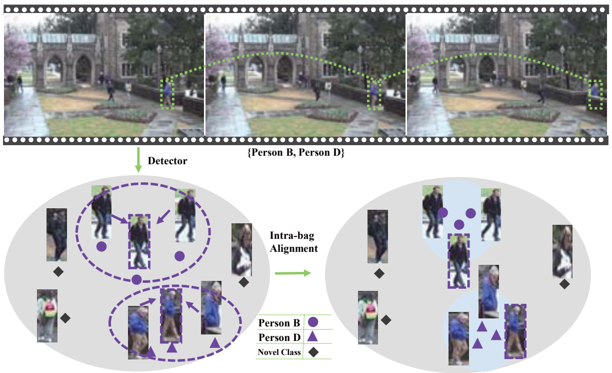


Figure 2. Illustration of the intra-bag alignment. The instances in the rectangular with dotted purple lines are the seed instances corresponding to the Person B and Person D, respectively. Then two groups (*e.g.*, framed by the purple oval dotted lines) are formed around these two seed instances. In the intra-bag alignment process, the label distributions of instances belonging to the same group are aligned such that these instances can be compact between each other in the learned feature space.

of the i-th instance in bag X_b, and N_p indicates the overall number of instances involved in the loss calculation.

On the other hand, we further expect that our classifier can fully exploit the weak labels to learn a more robust re-id model. More specifically, for any tagged label c in bag X_b, we select an instance with the largest prior probability w.r.t the c-th class as the *seed instance* $\mathbf{x}_{b,q_c}$, where the index q_c can be defined by

$$q_c = \text{argmax}_{i \in \{1,2,\cdots,n_b\}} \{\tilde{y}^c_{b,i}\}. \tag{2}$$

Then we force the estimated label of the seed instance approximate to the corresponding tagged video-level label. Accordingly, we define the classification loss for the gallery set as follows:

$$\mathcal{L}_g = \frac{1}{N_g} \sum_{X_b \in \mathcal{X}_g} \sum_{c \in \{0,\cdots,C\}} (-y^c_b \log \max\{\tilde{y}^c_{b,1}, \tilde{y}^c_{b,2}, ..., \tilde{y}^c_{b,n_b}\}), \tag{3}$$

where the operation $\max\{\tilde{y}^c_{b,1}, \tilde{y}^c_{b,2}, ..., \tilde{y}^c_{b,n_b}\}$ is used to select the largest prior probability of the seed instance $\mathbf{x}_{b,q_c}$. In such a case, the classification model can be leveraged to infer the prior probability of each instance in the bag.

Combining the two classification losses for the probe set (Eq.(1)) and the gallery set (Eq.(3)), we obtain the following MIML classification loss:

$$\mathcal{L}_C = \mathcal{L}_p + \mathcal{L}_g. \tag{4}$$

3.2.2 Intra-bag Alignment

Since individuals often appear in a video across several consecutive frames (*e.g.*, green dotted lines in Fig. 2), there will be a set of instances, probably of the same person, in a bag in the weakly labeled gallery set. These instances are expected to be merged into a group such that the instances belonging to the same group should be close to each other in the learned feature space. However, the MIML classifier cannot achieve this agglomeration and the classifier only processes the instance with the largest prior probability w.r.t the corresponding classes, which we call the *seed instance*.

To this end, we expect that the set of instances probably of the same person can be gathered around the seed instance $\mathbf{x}_{b,q_c}$ that has the largest prior probability with respect to the c-th class in the bag X_b. Then, we form a group that contains the instances gathered around the seed instance $\mathbf{x}_{b,q_c}$ by $\mathcal{G}_{b,c} = \{p | \mathbf{x}_{b,p} \in \mathcal{N}_{q_c} \text{ and } \tilde{y}^c_{b,p} \geq \gamma \tilde{y}^c_{b,q_c}\}$. In this group, the selected instances should be among the K-nearest neighbors $\mathcal{N}_{q_c}$ in the feature space around the seed instance $\mathbf{x}_{b,q_c}$. Additionally, the prior probability corresponding to the c-th class of these instances should be no less than $\gamma \tilde{y}^c_{b,q_c}$, where $\tilde{y}^c_{b,q_c}$ is the prior probability of the corresponding seed instance. Here, $\gamma \in (0, 1)$ is a relaxation parameter. Then, the intra-bag alignment loss can be defined as follows:

$$\mathcal{L}_{IA} = \frac{1}{N_{IA}} \sum_{X_b \in \mathcal{X}_g} \sum_{c \in \{0,\cdots,C\}} \sum_{p \in \mathcal{G}_{b,c}} y^c_b D_{KL}(\tilde{\mathbf{y}}_{b,p} \| \tilde{\mathbf{y}}_{b,q_c}), \tag{5}$$

$$D_{KL}(\tilde{\mathbf{y}}_{b,p} \| \tilde{\mathbf{y}}_{b,q_c}) = \sum_{c \in \{0,\cdots,C\}} \tilde{y}^c_{b,p} (\log \tilde{y}^c_{b,p} - \log \tilde{y}^c_{b,q_c}). \tag{6}$$

The intra-bag alignment loss in Eq.(5) is designated to evaluate the discrepancy of the label distribution between the instances within the group $\mathcal{G}_{b,c}$ and the corresponding seed instance $\mathbf{x}_{b,q_c}$. The discrepancy between two label distributions is defined by the Kullback-Leibler divergence depicted in Eq.(6). As illustrated in Figure 2, by minimizing the intra-bag alignment loss, the features of the same group can become closer to each other due to the alignment between potential instances of the same class in a bag.

3.2.3 Cross-view Bag Alignment

The intra-bag alignment term mainly considers the person images that appear in the same bag. We further expect to mine potential matched images of the same person between bags not only from the same camera view but also from non-overlapping camera views. In the meantime, all instances belonging to the same person should form a compact cluster in the learned feature space. For this purpose, we introduce a distribution prototype for each class, and then all the potential matched images of the same person from all the camera views are expected to be aligned to the corresponding distribution prototype. Formally, the distribution prototype of the c-th class at the current epoch t is denoted

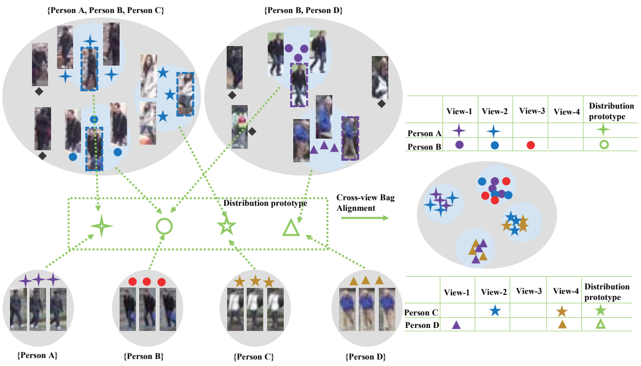


Figure 3. Illustration of cross-view bag alignment. The potential matched instances of the same person between bags from all the camera views are denoted by the same shape. The different camera views are represented by different colors. By performing cross-view bag alignment, the label distributions of these instances belonging to the same person are aligned w.r.t. the corresponding distribution prototype such that their features can be compact between each other in the learned feature space.

by $\hat{\mathbf{p}}_c^t$ that can be calculated by

$$\mathbf{p}_c^t = \frac{1}{|\mathcal{V}_c|}\sum_{v\in\mathcal{V}_c}\left(\frac{1}{|\mathcal{I}_{c,v}|}\sum_{i\in\mathcal{I}_{c,v}}\tilde{\mathbf{y}}_i\right), \tag{7}$$

$$\hat{\mathbf{p}}_c^t = \alpha\hat{\mathbf{p}}_c^{t-1} + (1-\alpha)\mathbf{p}_c^t, \tag{8}$$

where $\mathcal{V}_c$ is the collection of all the camera views, $\mathcal{I}_{c,v}$ is the set of instance indexes that belong to the c-th class under the v-th camera view, and α is a smoothing hyperparameter that controls the weight of the historical distribution prototype $\hat{\mathbf{p}}_c^{t-1}$ at the previous epoch $t-1$ when updating the distribution prototype at current epoch t.

After that, we alternate between the following two steps in the training stage: 1) calculate the distribution prototype at current epoch t for each class based on Eq. (7) and Eq. (8); 2) align the label distributions of instances belonging to the same person from all the camera views to the corresponding distribution prototype. Specifically, the *Cross-view Bag Alignment* is defined by

$$\mathcal{L}_{CA} = \frac{1}{N_{CA}}\sum_{X_b\in(\mathcal{X}_p\bigcup\mathcal{X}_g)}\sum_{c\in\{0,\cdots,C\}}\sum_{i\in\mathcal{I}_c} y_b^c D_{KL}(\tilde{\mathbf{y}}_{b,i}\|\hat{\mathbf{p}}_c^t), \tag{9}$$

where $\mathcal{I}_c$ is the collection of instance indexes from all the camera views for the c-th class, and $\hat{\mathbf{p}}_c^t$ is the distribution prototype of the c-th class at the current epoch t. As illustrated in Figure 3, $\mathcal{L}_{CA}$ is minimized to make potential instances of the same person from different bags captured from different camera views to gather together.

3.3. Deep Cross-view MIML Model

Summarizing the above analysis, we obtain the ***Cross-view Multi-label Multi-Instance learning (CV-MIML)*** method described below:

$$\mathcal{L}_{CV-MIML} = \mathcal{L}_C + \delta(\mathcal{L}_{IA} + \mathcal{L}_{CA} + \mathcal{L}_E), \tag{10}$$

where δ controls the weight and contribution of $\mathcal{L}_{IA}$, $\mathcal{L}_{CA}$ and $\mathcal{L}_E$ to the whole CV-MIML loss. By incorporating the intra-bag alignment and the cross-view bag alignment, the label distributions of intraclass instances are aligned not only within the same video (bag) but also between videos (bags) across camera views, so that the intra-class instances can be compact between each other in the learned feature space. Here, $\mathcal{L}_E$ is an *entropy regularization* term. In the learning process, we expect that each instance can be ideally partitioned into a certain class (i.e., the known classes or a novel class). For a weakly labeled bag in the gallery set, there may exist a certain number of instances far away from all the data groups that are formed in the intra-bag and cross-view bag alignment process. We call these instances *outlier instances*. This designation indicates that these outlier instances probably do not approach any of the known identity classes. To alleviate the effect of these outlier instances, we design an entropy regularization term as follows:

$$\mathcal{L}_E = \frac{1}{N_E}\sum_{X_b\in\mathcal{X}_g}\sum_{i\in\{1,\cdots,n_b\}}\sum_{c\in\{0,\cdots,C\}}(-\tilde{y}_{b,i}^c\log\tilde{y}_{b,i}^c). \tag{11}$$

Reducing the entropy in Eq.(11) is to facilitate the outlier instances to be affiliated to a certain class. We now embed the proposed CV-MIML method into a deep neural network to form an end-to-end framework of the *Deep CV-MIML* model that can learn coherent features and a robust MIML classifier simultaneously.

3.4. Implementation Details

To implement our proposed model, we adopt Resnet-50 [13] as our basic CNN for feature extraction. The fully-connected layer in Resnet-50 is replaced by our MIML classifier. All input images are resized to 256×128. The values of hyperparameters γ, K and α are set by cross validation on the validation set. The parameter δ in Eq.(10) is designed as a function of t that varies with time. Specifically, we let $\delta = w(t)$; the value of $w(t)\subseteq[0,1]$ initially increases with time and then reaches saturation and remains at the maximum value [18], which helps enhance the reliability of the model used in deep neural networks. The bounding boxes we used were automatically generated by the Mask R-CNN algorithm [12] in advance for the genuine WL-DukeMTMC-REID dataset. Indeed, many false positive bounding boxes are detected. To exclude these distractors, each bounding box is assigned a confidence score that indicates the possibility of that bounding box belonging to any of known classes. We set a threshold for excluding the samples with confidence scores below the threshold. The

Dataset	# Camera views	# Identities (training/testing)	# Training BBoxes (probe/gallery)	# Testing BBoxes (probe/gallery)
WL-DukeMTMC-REID	8	880/1695	60,267/923,879	116,128/904,066
WL-PRID2011	2	100/100	11,201/8,191	12,129/8,512
WL-iLIDS-VID	2	150/150	9,731/11,278	12,129/8,512
WL-MARS	6	631/630	38,324/460,236	36,988/472,978

Table 1. Detailed information of the one genuine and three new simulated datasets for the weakly supervised person re-id.

confidence score is obtained from a deep network that is pretrained on the probe set.

3.5. Testing

In the testing phase, the probe set and gallery set are formed in the same manner as the training set. Accordingly, our goal is to find the raw videos where the target person appears in the weakly supervised setting. Specifically, for a bag X_p in the testing probe set, the feature of this bag $\mathbf{x}_p$ is the average pooling of features over all image frames in this sequence. Then, the distance between the bag $\mathbf{x}_p$ in the testing probe set and the bag $\mathbf{x}_q$ in the testing gallery set is

$$D(p,q) = \min\{d(\mathbf{x}_p, \mathbf{x}_{q,1}), d(\mathbf{x}_p, \mathbf{x}_{q,2}), ..., d(\mathbf{x}_p, \mathbf{x}_{q,n_p})\} \tag{12}$$

where d is the Euclidean distance operator.

4. Experiments

4.1. Datasets and Settings

The experiments were carried out on one genuine dataset WL-DukeMTMC-REID and three simulated datasets WL-PRID 2011, WL-iLIDS-VID and WL-MARS. The probe set contained all the target individuals to search for in the gallery set, and every known identity had trimmed image sequences in the probe set for all datasets. The remainder of the videos formed the gallery set. The four datasets were constructed as follows.

WL-DukeMTMC-REID For the genuine WL-DukeMTMC-REID dataset, a set of raw videos DukeMTMC [27] is available. DukeMTMC is a multi-camera dataset recorded outdoors at the Duke University campus with 8 synchronized cameras. The WL-DukeMTMC-REID dataset was constructed from the first 50-minute raw HD videos. We split the raw videos into halves; the training set and testing set both have 25-minute raw videos. There are 880 and 1,695 identities appearing in at least two camera views in the training and testing sets. To form the gallery set for the WL-DukeMTMC-REID dataset, we first randomly split the raw video into short video clips, with each clip comprising between 20 and 120 raw frames. Afterwards, we applied Mask-RCNN [12] to these video clips to detect individuals. Note that even though we obtain the bounding boxes, the specific label of each individual is still unknown for the gallery set. The details of this dataset is shown in Table 1.

For the three simulated datasets WL-PRID 2011, WL-iLIDS-VID and WL-MARS, the raw videos of these datasets are unavailable, so we formed the simulated datasets as follows. First, we randomly selected one trimmed image sequence for every known identity to form the probe set, and the rest of videos were used to form the gallery set. Then, 3 $\sim$ 8 sequences were randomly selected to form a weakly labeled bag, where only bag-level labels were available, and the specific label of each individual was unknown. In this way, we converted three existing video-based person re-id datasets PRID 2011 [14], iLIDS-VID [30] and MARS [42] to WL-PRID 2011, WL-iLIDS-VID and WL-MARS, respectively, for weakly supervised person re-id. The details of these new datasets are shown in Table 1.

4.2. Evaluation Protocol

To evaluate the performance of our method, the widely used cumulative matching characteristics (CMC) curve and mean average precision (mAP) are used for quantitative measurement.

4.3. Evaluation of the Deep CV-MIML Model

In our modeling of Deep CV-MIML, we introduce 1) the intra-bag alignment term, 2) the cross-view bag alignment term, and 3) an entropy regularization to eliminate outlier instances. To evaluate the efficiency of the each component, we adopt the MIML classifier (Eq. (4)) as the baseline method and conduct ”baseline with IA”, ”baseline with CA” and ”baseline with entropy” for comparison to prove the effectiveness of all proposed components separately. The results are reported in Table 2.

Comparing the CV-MIML method to the baseline MIML classifier in Table 2, it is clear that our CV-MIML method is very effective in handling the weakly supervised person re-id problem. By simultaneously minimizing the intra-bag alignment and cross-view bag alignment loss functions, the same identities from the same camera view and nonoverlapping camera views could be more coherent with each other. These results represent a notable improvement in the rank-1 matching accuracy, e.g., 10.79%, 5.00%, 18.67% and 13.41% improvements were observed on the WL-DukeMTMC-REID, WL-PRID 2011, WL-iLIDS-VID and WL-MARS datasets, respectively. Considering mAP, we also obtain 8$\sim$14% improvement on these four weakly labeled re-id datasets.

Moreover, as reported in Table 2, the ablation study indicates that adopting the intra-bag alignment term will lead to a significant rise of the model performance because the intra-bag alignment term facilitates forming a coherent clustered structure for instances of the same identity. Additionally, including the cross-view bag alignment term would also notably increase the performance of CV-MIML (with approximately 5%, 1%, 11% and 10% rise of rank-1 matching accuracy on the WL-DukeMTMC-REID, WL-PRID 2011,

WL-DukeMTMC-REID	r=1	r=5	r=10	r=20	mAP
CV-MIML	**78.05**	**90.50**	**93.75**	**95.99**	**59.53**
baseline + IA	74.69	88.50	92.15	94.81	56.97
baseline+CA	72.92	87.96	92.04	94.75	55.30
baseline+entropy	70.56	85.90	90.15	92.68	53.05
baseline	67.26	84.90	89.50	92.68	50.96

WL-PRID2011	r=1	r=5	r=10	r=20	mAP
CV-MIML	**72.00**	**89.00**	95.00	**99.00**	**70.78**
baseline+IA	69.00	**89.00**	93.00	98.00	65.89
baseline+CA	68.00	87.00	**96.00**	98.00	63.72
baseline+entropy	70.00	**89.00**	**96.00**	**99.00**	67.32
baseline	67.00	86.00	95.00	97.00	62.87

WL-iLIDS-VID	r=1	r=5	r=10	r=20	mAP
CV-MIML	**60.00**	80.00	87.33	**96.67**	**56.01**
baseline+IA	55.33	**80.67**	**89.33**	95.33	53.78
baseline+CA	52.67	78.00	88.00	95.33	50.58
baseline+entropy	44.67	69.33	81.33	92.67	44.99
baseline	41.33	70.00	83.33	94.67	42.26

WL-MARS	r=1	r=5	r=10	r=20	mAP
CV-MIML	**66.88**	**82.02**	**87.22**	**91.48**	**55.16**
baseline+IA	62.15	80.44	85.80	89.75	50.27
baseline+CA	63.09	79.97	84.23	88.96	50.61
baseline+entropy	60.88	79.34	85.49	89.43	49.13
baseline	53.47	71.77	79.02	85.49	40.31

Table 2. Ablation study of the proposed CV-MIML method. The matching accuracy values (%) at rank(r) = 1, 5, 10, 20 and mAP are shown on the four datasets. The best results are shown in black boldface font.

WL-iLIDS-VID and WL-MARS datasets, respectively) because the cross-view bag alignment is useful for making the features of the same identities from nonoverlapping camera views aligned to each other in the feature space.

Finally, Table 2 indicates that the entropy regularization term also plays a significant role in our CV-MIML model, as with it, the effect of outlier instances can be eliminated, thus boosting the performance of our model.

4.4. Comparison with State-of-the-Art MIML Methods

In Table 3, we report the comparison of our method to existing state-of-the-art MIML learning methods DeepMIML [10] and MIMLfast [16]. The DeepMIML [10] method is an end-to-end deep neural network that integrates the instance representation learning process into the MIML learning. For a fair comparison, we reimplemented this method using the same CNN structure and the same training process. The MIMLfast [16] approach is a conventional two-stage framework that first requires extracting the image features and then learns a discriminative representation. In this study, we extracted the features from a Resnet-50 CNN that was pretrained on the 3DPeS [1], CUHK01 [21], CUHK03 [22], Shinpuhkan [17] and VIPeR [11] person re-id datasets and then performed the MIML learning based on the MIMLfast method.

The comparison shows that the proposed Deep CV-MIML model outperformed the existing MIML methods. The proposed Deep CV-MIML model clearly outperformed the second-best method DeepMIML on the four datasets. Specifically, the extra gain of the rank-1 matching accuracy between the Deep CV-MIML network and the DeepMIML method is 12.68%, 5.00%, 16.00% and 19.72% on the WL-DukeMTMC-REID, WL-PRID 2011, WL-iLIDS-VID and WL-MARS datasets, respectively. Moreover, comparing the proposed method to the Deep MIML method, the mAP matching gain on all datasets can reach 11.51%, 8.98%, 12.52% and 18.57% on the WL-DukeMTMC-REID, WL-PRID 2011, WL-iLIDS-VID and WL-MARS datasets, respectively. These results indicate the advantage of our Deep CV-MIML model in handling the weakly supervised person re-id problem. The better performance is mainly due to the newly designed intra-bag alignment term and cross-view bag alignment term. With these terms, the features of the same individual obtained from the same camera view and across nonoverlapping camera views can be more coherent, while the functions of these two terms are not considered in MIMLfast and DeepMIML.

WL-DukeMTMC-REID	r=1	r=5	r=10	r=20	mAP
MIMLfast[16]	13.63	44.66	55.69	64.78	10.05
DeepMIML[10]	65.37	82.30	86.90	90.68	48.02
Deep CV-MIML	**78.05**	**90.50**	**93.75**	**95.99**	**59.53**

WL-PRID2011	r=1	r=5	r=10	r=20	mAP
MIMLfast[16]	29.00	56.00	72.00	87.00	31.66
DeepMIML[10]	67.00	**90.00**	94.00	99.00	61.80
Deep CV-MIML	**72.00**	89.00	**95.00**	**99.00**	**70.78**

WL-iLIDS-VID	r=1	r=5	r=10	r=20	mAP
MIMLfast[16]	28.00	58.67	69.33	78.67	27.42
DeepMIML[10]	44.00	70.00	81.33	89.33	43.49
Deep CV-MIML	**60.00**	**80.00**	**87.33**	**96.67**	**56.01**

WL-MARS	r=1	r=5	r=10	r=20	mAP
MIMLfast[16]	20.50	37.22	43.06	52.05	11.39
DeepMIML[10]	47.16	70.19	76.18	81.07	36.59
Deep CV-MIML	**66.88**	**82.02**	**87.22**	**91.48**	**55.16**

Table 3. Comparison with state-of-the-art MIML methods. The best results are in black boldface font.

4.5. Comparison with Related Re-id Methods

As existing supervised person re-id methods could not be applied to our weakly supervised setting directly, we compare our method to unsupervised person re-id methods, such as CAMEL [40], PUL [8] and the one-shot person re-id method called EUG [32]. Among the listed methods, the CAMEL method is a conventional two-stage framework that first requires extracting the image features and then learns an asymmetric representation. PUL and EUG are progressive methods that alternate between assigning the pseudo-labels to the tracklets and training the CNN model according to these pseudo-labeled data samples. To further demonstrate the effectiveness of our method, we also compared with a state-of-the-art fully supervised approach PCB[29]. The results are reported in Table 4. Compared to unsupervised or one-shot methods, the performance of these methods is consistently unsatisfactory in comparison to that of the proposed Deep CV-MIML model. The table can also tell us that the performance of our model (Deep CV-MIML) is comparable to the fully supervised model PCB on the WL-DukeMTMC-REID and WL-MARS datasets.

4.6. Hyperparameter Analysis

There are four hyperparameters involved in our CV-MIML formulation. The trade-off parameter δ is used to

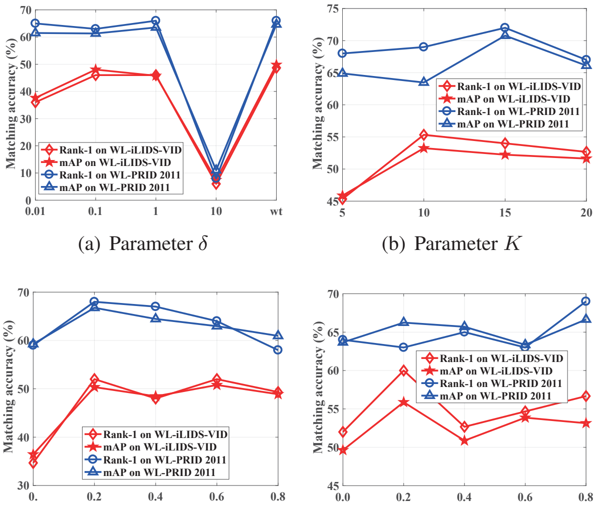
WL-DukeMTMC-REID	r=1	r=5	r=10	r=20	mAP
CAMEL [40]	0.53	0.77	1.18	3.24	0.90
PUL[8]	-	-	-	-	-
EUG[32]	35.93	50.74	59.41	66.96	21.94
Deep CV-MIML	78.05	90.50	93.75	95.99	59.53
PCB[29]	79.82	90.38	93.45	96.17	62.09
WL-PRID2011	r=1	r=5	r=10	r=20	mAP
CAMEL [40]	2.00	11.00	20.00	44.00	4.59
PUL[8]	32.00	58.00	71.00	85.00	35.28
EUG[32]	55.00	83.00	93.00	97.00	53.26
Deep CV-MIML	72.00	89.00	95.00	99.00	70.78
PCB[29]	88.00	97.00	99.00	99.00	87.35
WL-iLIDS-VID	r=1	r=5	r=10	r=20	mAP
CAMEL [40]	4.67	16.00	26.67	43.33	6.26
PUL[8]	20.00	44.00	59.33	76.00	22.56
EUG[32]	26.67	60.67	72.00	86.67	29.86
Deep CV-MIML	60.00	80.00	87.33	96.67	56.01
PCB[29]	72.00	89.33	92.67	96.00	69.87
WL-MARS	r=1	r=5	r=10	r=20	mAP
CAMEL [40]	0.32	1.10	2.52	5.52	0.56
PUL[8]	-	-	-	-	-
EUG[32]	25.87	39.59	46.21	55.21	15.63
Deep CV-MIML	66.88	82.02	87.22	91.48	55.16
PCB[29]	68.14	84.07	86.28	90.54	54.18

Table 4. Comparison with related re-id methods. The $1^{st}/2^{nd}$ best results are indicated in red/blue.

balance the weight of $\mathcal{L}_{IA}$, $\mathcal{L}_{CA}$ and $\mathcal{L}_E$ with respect to the overall CV-MIML loss in Eq. (10). During training, we consider $\delta = w(t)$, a time-dependent function of time t. To verify the advantage of this approach, we compared the performance to that of a fixed value of δ, where $\delta = 0.01, 0.1, 1, 10$ to investigate the impact of δ on the overall performance on the WL-PRID 2011 and WL-iLIDS-VID datasets. As shown in Figure 4(a), the time-dependent setting is preferable. The reason is that the reliability of the intra-bag alignment and cross-view bag alignment process is tightly related to the confidence of the re-id model by the seed instances selection and the distribution prototype calculation. Additionally, the confidence of the re-id model is fairly low in the beginning and then steadily increases during the training procedure. Similarly, the weight parameter $\delta \in [0, 1]$ initially increases during the early training stage, subsequently reaching saturation at approximately the maximum value 1 once the model has been sufficiently trained.

The group formed in the intra-bag alignment process is closely related to parameters K and γ. Parameter K represents selecting the K-nearest neighbors in the feature space, and parameter γ controls the number of instances corresponding to those with the largest prior probabilities that should be shared with the same weak label. The impacts of K and γ are reported in Figure 4(b) and Figure 4(c). The results suggest that the best performance can be reached on both datasets if $\gamma = 0.2$ and K is approximately 15.

The impact of α is presented in Figure 4(d). Parameter α controls the impact of the historical distribution prototype when calculating the distribution prototype for the current epoch in Eq. (8). The figure suggests that the performance with and without historical information in the calculation of the distribution prototype is distinct. Specifically, the worst performance is observed if $\alpha = 0$, i.e., involving the historical information that eliminates the bias of the current output is useful for the calculation of the distribution prototype.



(a) Parameter δ

(b) Parameter K

(c) Parameter γ

(d) Parameter α

Figure 4. Performance illustrations for the Deep CV-MIML model with different hyperparameters.

5. Conclusion

We aim to remove the need for costly labeling efforts for conventional person re-id by considering weakly supervised person re-id modeling. In this weakly supervised setting, no specific annotations of individuals inside gallery videos are necessary; the only requirement is the indication of whether or not a person appears in a given video. In such a setting, one can search for individuals and the videos that they appear in, given a (set of) probe person image(s). We cast the weakly supervised person re-id problem as a multi-instance-multi-label (MIML) problem. We develop a cross-view MIML (CV-MIML) method, which is able to mine potential intraclass variation in a bag and potential cross-view change between instances of the same person across bags from all camera views. Finally, CV-MIML is optimized by being embedded in a deep neural network. The experimental results have verified the feasibility of weakly supervised modeling for person re-id and have also shown the effectiveness of the proposed CV-MIML models.

Acknowledgment

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(61522115), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), and the Royal Society Newton Advanced Fellowship (NA150459).

References

[1] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *J-HGBU Workshop*, pages 59–64, 2011. 7

[2] Forrest Briggs, Xiaoli Z Fern, and Raviv Raich. Context-aware MIML instance annotation. In *ICDM*, pages 41–50, 2013. 2, 3

[3] Forrest Briggs, Xiaoli Z Fern, Raviv Raich, and Qi Lou. Instance annotation for multi-instance multi-label learning. *TKDD*, 7(3):1–14, 2013. 2, 3

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. In *BMVC*, 2018. 2

[5] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 40(2):392–408, 2018. 2

[6] Zenghai Chen, Zheru Chi, Hong Fu, and Dagan Feng. Multi-instance multi-label image classification: A neural approach. *Neurocomputing*, 99:298–306, 2013. 3

[7] De Cheng, Yihong Gong, Xiaojun Chang, Weiwei Shi, Alexander Hauptmann, and Nanning Zheng. Deep feature learning via structured graph Laplacian embedding for person re-identification. *PR*, 82:94–104, 2018. 2

[8] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: clustering and fine-tuning. *TOMM*, 14(4):83, 2018. 7, 8

[9] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*, 2017. 2

[10] Ji Feng and Zhi-Hua Zhou. Deep MIML network. In *AAAI*, pages 1884–1890, 2017. 2, 7

[11] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshop*, pages 1–7, 2007. 7

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 5, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[14] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 6

[15] Yu-Feng Li Ju-Hua Hu and Yuan Jiang Zhi-Hua Zhou. Towards discovering what patterns trigger what labels. In *AAAI*, pages 1012–1018, 2012. 2

[16] Sheng-Jun Huang, Wei Gao, and Zhi-Hua Zhou. Fast multi-instance multi-label learning. In *AAAI*, pages 1868–1874, 2014. 2, 3, 7

[17] Yasutomo Kawanishi, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *FCV Workshop*, 2014. 7

[18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 5

[19] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *CVPR*, 2018. 2

[20] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 1, 2

[21] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 7

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 7

[23] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, pages 2448–2457, 2017. 2

[24] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *PR*, 65:197–210, 2017. 2

[25] Cam-Tu Nguyen, De-Chuan Zhan, and Zhi-Hua Zhou. Multi-modal image annotation with multi-instance multi-label LDA. In *IJCAI*, pages 1558–1564, 2013. 3

[26] Anh T Pham, Raviv Raich, Xiaoli Z Fern, Jesús Pérez Arriaga, et al. Multi-instance multi-label learning in the presence of novel class instances. In *ICML*, pages 2427–2435, 2015. 2

[27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop*, pages 17–35, 2016. 6

[28] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, pages 1179–1188, 2018. 2

[29] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 7, 8

[30] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 6

[31] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Deep adaptive feature embedding with local sample distributions for person re-identification. *PR*, 73:275–288, 2018. 2

[32] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, pages 5177–5186, 2018. 1, 2, 7, 8

[33] Yang Wu, Jie Qiu, Jun Takamatsu, and Tsukasa Ogasawara. Temporal-enhanced convolutional network for person re-identification. In *AAAI*, pages 7412–7419, 2018. 1, 2

[34] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2, 2016. 2

[35] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*, 2018. 2

[36] Xin-Shun Xu, Xiangyang Xue, and Zhi-Hua Zhou. Ensemble multi-instance multi-label learning approach for video annotation task. In *ACM MM*, pages 1153–1156, 2011. 3

[37] Oksana Yakhnenko and Vasant Honavar. Multi-instance multi-label learning for image classification with large vocabularies. In *BMVC*, pages 1–12, 2011. 3

[38] Hao Yang, Joey Tianyi Zhou, Jianfei Cai, and Yew Soon Ong. MIML-FCN+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In *CVPR*, pages 5996–6004, 2017. 3

[39] Mang Ye, Andy Jinhua Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, pages 5152–5160, 2017. 2

[40] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017. 2, 7, 8

[41] Zheng-Jun Zha, Xian-Sheng Hua, Tao Mei, Jingdong Wang, Guo-Jun Qi, and Zengfu Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, pages 1–8, 2008. 3

[42] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016. 6

[43] Sanping Zhou, Jinjun Wang, Deyu Meng, Xiaomeng Xin, Yubing Li, Yihong Gong, and Nanning Zheng. Deep self-paced learning for person re-identification. *PR*, 76:739–751, 2018. 2

[44] Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. Multi-instance multi-label learning. *AI*, 176(1):2291–2320, 2012. 2

[45] Xiaoke Zhu, Xiao-Yuan Jing, Xinge You, Xinyu Zhang, and Taiping Zhang. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. *TIP*, 27(11):5683 – 5695, 2018. 1, 2

[46] Yue Zhu, Kai Ming Ting, and Zhi-Hua Zhou. Discover multiple novel labels in multi-instance multi-label learning. In *AAAI*, pages 2977–2984, 2017. 2