Rewarded Semi-Supervised Re-Identification on Identities Rarely Crossing Camera Views

Ancong Wu, Wenhang Ge, Wei-Shi Zheng

Abstract—Semi-supervised person re-identification (Re-ID) is an important approach for alleviating annotation costs when learning to match person images across camera views. Most existing works assume that training data contains abundant identities crossing camera views. However, this assumption is not true in many real-world applications, especially when images are captured in nonadjacent scenes for Re-ID in wider areas, where the identities rarely cross camera views. In this work, we operate semi-supervised Re-ID under a relaxed assumption of identities rarely crossing camera views, which is still largely ignored in existing methods. Since the identities rarely cross camera views, the underlying sample relations across camera views become much more uncertain, and deteriorate the noise accumulation problem in many advanced Re-ID methods that apply pseudo labeling for associating visually similar samples. To quantify such uncertainty, we parameterize the probabilistic relations between samples in a relation discovery objective for pseudo label training. Then, we introduce reward quantified by identification performance on a few labeled data to guide learning dynamic relations between samples for reducing uncertainty. Our strategy is called the Rewarded Relation Discovery (R^2D), of which the rewarded learning paradigm is under-explored in existing pseudo labeling methods. To further reduce the uncertainty in sample relations, we perform multiple relation discovery objectives learning to discover probabilistic relations based on different prior knowledge of intra-camera affinity and cross-camera style variation, and fuse the complementary knowledge of different probabilistic relations by similarity distillation. To better evaluate semi-supervised Re-ID on identities rarely crossing camera views, we collect a new real-world dataset called REID-CBD, and perform simulation on benchmark datasets. Experiment results show that our method outperforms a wide range of semi-supervised and unsupervised learning methods. Project: https://github.com/wuancong/REID-CBD.

Index Terms—Person re-identification, semi-supervised person re-identification.

1 Introduction

Person re-identification (Re-ID) for matching person images across non-overlapping camera views in video surveillance has received substantial attention in recent years. Many existing deep models [1], [2], [3], [4], [5], [6], [7] have achieved high performance on benchmark datasets.

For reducing heavy annotation costs, semi-supervised learning [8], [9], [10], [11], [12] and unsupervised domain adaptation (UDA) [13], [14], [15], [16], [17], [18] have undergone fast development and the performance is increasingly close to that of supervised learning, and we focus on studying semi-supervised learning in this work. In these advanced Re-ID methods, explicitly or implicitly associating cross-camera positive pairs can effectively alleviate cross-camera scene variations such as lighting, view angle and background clutters for representation learning, which are the key challenges for Re-ID. An underlying assumption for these methods is that a large number of identities crossing camera views can be captured for training, e.g., data collected in adjacent scenes in small-scale surveillance systems. However, this assumption is violated in many cases of real-world training data collection. It is difficult to capture identities crossing some camera views, especially for nonadjacent scenes in large-scale surveillance systems.

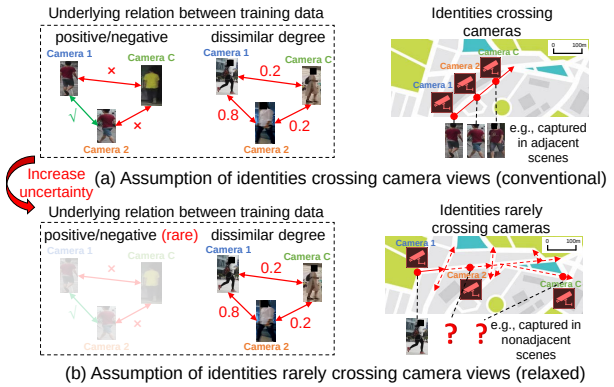

Fig. 1. Assumption of identities crossing camera views v.s. assumption of identities rarely crossing camera views. (a) An underlying assumption for conventional semi-supervised Re-ID methods is that there are abundant identities crossing camera views in unlabeled training data. For example, images captured in adjacent scenes usually satisfy this assumption. (b) Our relaxed assumption for semi-supervised Re-ID is that there are identities rarely crossing camera views for training. For example, when Re-ID is extended to more distant nonadjacent scenes, it is more difficult to capture underlying identities crossing camera views because of uncertain possible paths indicated by red dotted arrows. The underlying relations between training data are increasingly uncertain, since deterministic relations of cross-camera positive/negative pairs are rare, and the underlying relations are mainly modeled by probabilistic relations of dissimilar degrees.

To show some examples, we compare the statistics of multi-site dataset and single-site datasets in Table 1. Existing benchmark datasets MSMT17 [19] , Market-1501 [20] and

- *Ancong Wu is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. E-mail: wuancong@gmail.com*
- *Wenhang Ge is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. E-mail: gewh@mail2.sysu.edu.cn*
- *Wei-Shi Zheng is with the School of Computer Science and Engineering, Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, and Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou, China. E-mail: wszheng@ieee.org.*

(Corresponding author: Wei-Shi Zheng.)

TABLE 1
Comparison of average number of cameras passed by per identity (#$\text{camera}_{\text{pass}}$) between multi-site dataset Person30K and single-site datasets. In multiple disjoint sites, #$\text{camera}_{\text{pass}}$ is much smaller than the total number of cameras #$\text{camera}_{\text{total}}$.

Datasets	Person30K[1] [23]	MSMT17 [19]	Duke [21]	Market [20]
#site	89	1	1	1
Avg #camera passed by per identity (#$\text{camera}_{\text{pass}}$)	16.01	4.67	2.67	4.42
#$\text{camera}_{\text{total}}$	6,497	15	8	6
#$\text{camera}_{\text{pass}}$/#$\text{camera}_{\text{total}}$	0.01	0.31	0.33	0.74
#identity	30,000	4,101	1,812	1,501

many other datasets [1], [21], [22] were captured in local areas of only one site. As shown by "#$\text{camera}_{\text{pass}}$/#$\text{camera}_{\text{total}}$" in Table 1, each identity appears in 31%-74% of cameras on average. While collecting data in 89 disjoint sites in a large-scale dataset Person30K[1] [23], each identity only appears in 1% of cameras on average. Therefore, it is significant to operate under a relaxed assumption that identities crossing camera views are rare for training Re-ID model, which is still largely ignored by existing Re-ID methods.

For real-world applications in large-scale systems such as matching between camera views in distant nonadjacent scenes, it is necessary to explore the underlying relations between cross-camera sample pairs under the assumption of identities rarely crossing camera views, because annotating cross-camera positive pairs is difficult. We show the comparison between the conventional assumption of identities crossing camera views and our relaxed assumption of identities rarely crossing camera views for semi-supervised Re-ID in Figure 1. When there are only identities rarely crossing camera views (Figure 1 (b)), deterministic relations of cross-camera positive/negative pairs are rare and the underlying relations are mainly modeled by probabilistic relations of dissimilar degrees. This causes high uncertainty during the exploration of the underlying relations between cross-camera sample pairs. High uncertainty incurs more serious noise accumulation problem [24] for existing pseudo labeling methods for Re-ID [9], [11], [17], [18] that are effective for the cases with identities crossing camera views (Figure 1 (a)). Our experimental findings in Section 3.2 demonstrate that applying pseudo labeling for training data with identities rarely crossing camera views results in performance degradation caused by wrongly associating cross-camera negative pairs.

To overcome the problem of high uncertainty in cross-camera unpaired training data, we propose a rewarded pseudo label training strategy called ***R**ewarded **R**elation **D**iscovery* (R^2D). To quantify the uncertainty, we represent similar samples via clusters and parameterize the probabilistic relations between the unlabeled sample pairs by cluster relation matrix in the relation discovery objective for pseudo label training. To discover the probabilistic relations, we introduce reward for the relation discovery objective. We regard the identification performance on limited labeled data as reward, which is quantified by a few-shot validation objective. By maximizing the reward using bi-level optimization [25], the cluster relation matrix as the parameter of the relation discovery objective is dynamically updated during training. As a result, the high uncertainty is explicitly quantified by the learned probabilistic relations and reduced by minimizing the relation discovery objective.

Based on our rewarded relation discovery, we further propose multiple relation discovery objectives learning to investigate prior knowledge of intra-camera affinity and cross-camera style variation for two different relation discovery objectives, respectively. Then, we fuse the complementary knowledge learned by different relation discovery objectives in a single model by similarity knowledge distillation, in order to further reduce the uncertainty in the underlying sample relations.

To evaluate semi-supervised Re-ID on identities rarely crossing camera views, existing public benchmark datasets cannot be directly applied, since the images are manually selected and annotated to guarantee all identities are crossing camera views. We collect a more large-scale real-world dataset REID-CBD in nonadjacent scenes, which contains identities rarely crossing camera views. It can be publicly available after data masking.

2 Related Work

2.1 Person Re-identification

Current person re-identification methods mainly focus on supervised learning [1], [3], [4], [5], [6], [7], [26], [27], [28], [29], [30], [31] and unsupervised learning [13], [14], [15], [17], [18], [19], [32], [33], [34], [35], [36], [37]. When sufficient labeled data is available, deep supervised models [1], [2], [3], [4], [5], [6], [7], [28], [38], [39], [39] have achieved remarkable performance on benchmark datasets. To get rid of the dependence on labeled data, unsupervised domain adaptation (UDA) techniques [13], [14], [15], [17], [18], [19], [32], [33], [34], [35], [36], [37] are explored for unsupervised Re-ID. For learning Re-ID on identities rarely crossing camera views, the supervised methods are prone to overfitting limited cross-camera positive pairs and the unsupervised learning methods have difficulties to reduce high uncertainty in unlabeled cross-camera unpaired data.

Generally, identity annotations are available for training but limited due to budget in real-world applications. Semi-supervised learning is a realistic setting for this scenario. Semi-supervised learning methods [8], [9], [10], [11], [12], [13], [40] exploit limited labeled data and unlabeled data to learn Re-ID model. Pseudo label training [13] is an effective approach for exploiting the unlabeled data for cluster discrimination. Representative pseudo label training methods include PUL [13], MVC [9] and one-example [11]. MVC [9] applies multi-view clustering to obtain more reliable pseudo labels. Liu et al. [8] model representations of unlabeled data by dictionary learning. Chang et al. [12] propose a graph-based transductive hard mining method for hard triplets in unlabeled data. Hao et al. [41] propose a two-stream encoder-decoder structure for disentangled feature learning.

Pseudo labeling methods for unsupervised Re-ID can also be applied to exploit the unlabeled data in semi-supervised learning. Some training strategies are proposed to progressively refine and exploit the noisy pseudo labels, such as mutual mean-teaching (MMT) [17], self-paced contrastive learning [42], asymmetric metric learning [15], camera-aware proxy [34], pose disentanglement [43], group-aware label transfer [18] and online pseudo label gener-

1. So far, Person30K is not publicly available.

ation [35]. Some methods such as MAR [44] and PCSL [45] take soft labels into account to model probabilistic relations. MEB-Net [37] adopts mutual learning among multiple networks with different architectures to exploit their heterogeneity. MeanTeacher [46] and MMT [17] exploit logit-based knowledge distillation between different models to refine the noisy labels; our method exploits similarity-based knowledge distillation for fusing complementary knowledge learned by different relation discovery objectives. The above discussed methods including MEB-Net [37], MeanTeacher [46] and MMT [17] do not operate under the assumption of identities rarely crossing camera views, so that they cannot tackle the high uncertainty problem caused by identities rarely crossing camera views as our method. UNRN [47] measures uncertainty of each sample by soft multilabel agreement of mean teacher model and student model to alleviate the influence of noisy labels. For cross-camera sample pairs, UNRN [47] relies on smoothness assumption that neighboring input points are probably from the same class; while our method does not rely on this assumption. Since our method reduces uncertainty with reward from labeled data, the cluster relations learned by our method are able to improve validation performance on target domain, while UNRN [47] quantifies uncertainty based on empirical uncertainty estimation principle and the uncertainty may be incorrectly estimated without using the reward in our method. Compared with pseudo labeling in UNRN [47], our method additionally takes camera labels into account for intra-camera clustering, and thus can avoid wrong association of cross-camera negative pairs in the scenario with rare identities crossing camera views. DG-Net [48] synthesizes fake images to increase diversity of training data regardless of camera label. Our method takes the assumption of rare identities crossing camera views into account and synthesizes cross-camera images to provide knowledge of cross-camera image style variation to initialize the learnable cluster relations for reducing the uncertainty when the identities rarely cross camera views; while DG-Net [48] cannot provide such knowledge by the synthesized images and cannot learn cluster relations with reward as our method. Fu et al. [49] collect LUPerson dataset from Internet and propose contrastive learning method LUP [49] for pre-training model on unlabeled data with limited potential cross-camera positive pairs. Compared with this setting, we do not concern the pre-training problem for model initialization. Moreover, in the target scene for Re-ID deployment, the setting of LUPerson [49] still requires training data with cross-view positive pairs for each identity, while we assume that training data contains only limited cross-view positive pairs. When LUP [49] is applied for Re-ID on identities rarely crossing camera views, it cannot tackle the high uncertainty problem as our method.

Existing semi-supervised Re-ID methods as well as some unsupervised Re-ID methods rely on the smoothness assumption that neighboring input points are probably of the same identity, so that they associate cross-camera sample pairs that are close in the feature space to reduce the identity uncertainty. However, the smoothness assumption is invalid for cross-camera sample pairs when learning Re-ID on identities rarely crossing camera views in our study. This makes the relations between cross-camera sample pairs become highly uncertain, which is ignored by the pseudo labeling methods and incurs serious noise accumulation problem [24]. Rather than determining the sample relations based on the smoothness assumption, our method learn the relations between unlabeled sample pairs with reward from limited labeled data to quantify and reduce the high uncertainty.

2.2 General Semi-Supervised Learning Methods

In general, semi-supervised learning [50] leverages unlabeled data as well as limited labeled data for training models. A broad variety of semi-supervised learning methods have been proposed. Pseudo labeling [51], [52], [53] is a type of advanced semi-supervised learning method that assigns pseudo labels to unlabeled data for training. Co-training [54] is a type of disagreement-based method that trains multiple models to exploit unlabeled data. Graph-based methods [55], [56] construct similarity graph to propagate labels from labeled samples to unlabeled samples. Curriculum learning is combined with semi-supervised learning in FlexMatch [57] to leverage unlabeled data according to the learning status of the model. The common assumptions for semi-supervised learning methods include the smoothness assumption, low-density assumption or manifold assumption. Based on these assumptions, some regularizers are introduced, such as density regularization [58], Laplacian regularization [59] and manifold regularization [60]. Perturbation-based methods aim to learn representation by making predictions for the noisy and the clean inputs be similar, such as Ladder Network [61], pseudo-ensembles [62], mean teacher [63], and virtual adversarial training [64]. Self-supervised learning has also been explored for semi-supervised learning. Chen et al. [65] perform semi-supervised learning by three steps of self-supervised pre-training, fine-tuning and knowledge distillation.

Existing semi-supervised learning methods [50], [66] operate under the smoothness assumption that neighboring inputs are probably of the same class, which is invalid for the large amount of unlabeled cross-camera unpaired data for Re-ID on identities rarely crossing camera views. Our method does not rely on this assumption and considers reducing the high uncertainty of cross-camera unpaired data by using the identification performance on labeled data as reward to update the pseudo label training loss, which is an under-explored learning paradigm for pseudo labeling methods. Moreover, compared with previous semi-supervised learning settings, additional camera labels are specifically available for unlabeled training data in person re-identification, and more importantly the large discrepancy between different camera views is a challenging problem to be addressed. Our method can perform camera-specific modeling to address the high uncertainty problem caused by limited cross-camera positive pairs, but previous semi-supervised methods cannot be straightforwardly extended to effectively make use of the camera labels.

2.3 Training Objective Optimization

To optimize the training process of the model, bi-level optimization [25] has been widely applied for hyperparameter optimization [67], meta learning [68] and neural

architecture search [69], [70]. In bi-level optimization problem, the feasible region of the upper-level optimization problem is restricted by the solution of the lower-level optimization problem. Training objective optimization [71], [72], [73] is closely related to our proposed rewarded pseudo label training method. These methods generally regard the performance on a clean unbiased validation set as reward to guide optimizing the training objectives. MSLG [71] learns meta soft labels for each sample to correct the noisy labels. Ren et al. [72] learn to reweight samples for more robust deep learning. Pham et al. [73] propose meta pseudo labels (MPL) by using a teacher model to generate pseudo labels to help the student model to generalize.

These related training objective optimization methods are designed for closed-set image classification where all samples belong to known classes, so that they are not suitable for Re-ID without known identities in the unlabeled data. Our method discovers the underlying probabilistic relations between the cross-camera sample pairs of unknown identities by rewarded pseudo label training strategy.

3 Rewarded Semi-Supervised Re-ID

As introduced in Section 1, when there are rare identities crossing camera views under our relaxed assumption for semi-supervised Re-ID, there are hardly deterministic relations of cross-camera positive/negative pairs to be explored. This causes high uncertainty during the exploration of the underlying relations between cross-camera sample pairs. To quantify and reduce such uncertainty, we propose a rewarded semi-supervised Re-ID strategy called Rewarded Relation Discovery (R^2D) to discover the probabilistic relations for unlabeled data that can maximize the reward of identification performance on labeled data.

3.1 Problem Formulation for Semi-Supervised Learning on Identities Rarely Crossing Camera views

We assume that only a small amount of cross-camera positive pairs can be captured and labeled. The labeled data set is denoted by $\mathcal{D}_L = \{(\mathbf{I}_i^L, y_i, v_i^L)\}_{i=1}^{N_L}$, where $\mathbf{I}_i^L$ is person image, $y_i = 1, ..., C_L$ is the identity label and $v_i^L = 1, ..., V_{cam}$ is the camera view label that can be obtained without annotation. The number of identity C_L is 10 in our case. Images of the same identity are captured from at least two different cameras. Besides the small amount of labeled data, we can easily obtain a large amount of unlabeled data by pedestrian detection. The unlabeled data set is denoted by $\mathcal{D}_U = \{(\mathbf{I}_i^U, v_i^U)\}_{i=1}^{N_U}$, where $\mathbf{I}_i^U$ is person image and $v_i^U = 1, ..., V_{cam}$ is camera view label. Underlying cross-camera positive pairs hardly exist in unlabeled data, that is, $\mathbf{I}_k^U$ and $\mathbf{I}_l^U$ are unlikely from the same person if $v_k^U \neq v_l^U$.

Given labeled data set $\mathcal{D}_L$ and unlabeled data set $\mathcal{D}_U$, we aim to learn a model $F(\cdot;\mathbf{\Theta})$ for extracting feature $\mathbf{x}_i = F(\mathbf{I}_i;\mathbf{\Theta})$ for matching by similarity measurement. We do not rely on the assumption of existence of cross-camera positive pairs in unlabeled data.

3.2 Noise Accumulation Problem of Pseudo Labeling on Identities Rarely Crossing Camera Views

To discover the underlying relation between unlabeled samples for Re-ID, pseudo label training is popular in existing Re-ID methods for utilizing unlabeled data [11], [12], [15], [17]. For learning Re-ID on abundant identities crossing camera views, associating samples of similar visual appearances to reduce the identity uncertainty contributes to the success of these methods.

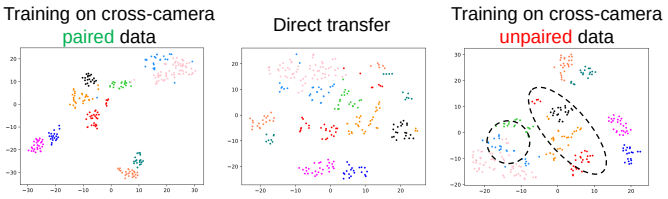

Fig. 2. Comparison between unsupervised learning by pseudo labeling method Sbase [47] on cross-camera paired data and cross-camera unpaired data. The feature distributions of randomly selected 8 identities on training set of Market-1501 dataset [20] indicated by different colors are shown by t-SNE [74]. Training on cross-camera unpaired data suffers from more severe noise accumulation problem than training on cross-camera paired data.

TABLE 2
Performances (%) of using pseudo labeling methods on subsets of cross-camera unpaired data and cross-camera paired data on Market-1501. "-intra" denotes intra-camera clustering.

Method	cross-camera unpaired			cross-camera paired		
	mAP	rank-1	rank-10	mAP	rank-1	rank-10
Sbase [47]	34.8	58.6	79.8	47.5	72.1	89.6
Sbase-intra [47]	35.9	59.6	81.2	41.1	64.1	83.4
MMT [17]	29.7	52.5	74.1	42.4	64.9	85.2
MMT-intra [17]	34.1	54.5	79.1	37.7	58.8	82.2
GLT [18]	34.7	60.7	80.1	39.6	66.0	83.8
GLT-intra [18]	36.2	61.7	81.7	35.4	59.7	78.9

However, for cross-camera unpaired data with identities rarely crossing camera views, the uncertainty of the underlying relations between cross-camera sample pairs becomes much higher, which challenges existing pseudo labeling methods for Re-ID. Generally, they discover the underlying relations and associate unlabeled sample pairs without reward from labeled data. They are prone to suffer from noise accumulation problem [24] caused by wrongly associating unlabeled cross-camera samples pairs of different identities.

To show the noise accumulation problem, we conducted an experiment to evaluate typical unsupervised pseudo-labeling-based Re-ID methods Sbase [47], MMT [17] and GLT [18] on Market-1501 dataset [20]. To simulate two different cases of learning Re-ID on identities crossing camera views and learning Re-ID on identities rarely crossing camera views, we modified the training set to two subsets of the same size that contain cross-camera paired data and cross-camera unpaired data, respectively. We also modified these methods by using intra-camera clustering to avoid wrong association of cross-camera negative samples for cross-camera unpaired data, denoted by "Sbase-intra", "MMT-intra", "GLT-intra". A ResNet-50 [75] was initialized by pretraining on MSMT17 dataset [19]. The experiment results are shown in Table 2. t-SNE [74] was applied to visualize the features of 8 randomly selected identities learned by Sbase [47] on these two subsets in Figure 2.

Compared with SBase, MMT and GLT, the intra-camera clustering versions perform better on cross-camera unpaired data and perform much worse on cross-camera paired data. Compared with SBase, MMT and GLT on cross-camera

TABLE 3
Definitions of important notations.

Notation	Definition
$\mathbf{\Theta}$, $\mathbf{p}_c$	Parameter of model F, the c-th prototype
$\mathbf{\Theta}^*$, $\mathbf{p}_c^*$	Optimal $\mathbf{\Theta}$, $\mathbf{p}_c$
$\mathbf{\Theta}'$, $\mathbf{p}_c'$	Approximated optimal $\mathbf{\Theta}$, $\mathbf{p}_c$
$\hat{\mathbf{R}}^{dyn}$ ($\hat{\mathbf{R}}^{dyn(0)}$)	Cluster relation matrix (initial value)
$\mathbf{x}_i^U$ ($\mathbf{x}_i^L$)	Feature of unlabeled (labeled) data

paired data, the performances of SBase-intra, MMT-intra and GLT-intra on cross-camera unpaired data are significantly degraded. The high uncertainty in cross-camera unpaired data is not well quantified and reduced in existing advanced pseudo labeling methods for Re-ID. Identities rarely crossing camera views lead to such high uncertainty in underlying relations of training data.

3.3 Rewarded Relation Discovery

For semi-supervised learning on identities rarely crossing camera views, the uncertainty of underlying sample relations is significantly increased as compared with learning Re-ID on abundant identities crossing camera views. How to quantify and reduce the uncertainty in cross-camera sample pairs is important for pseudo label training.

To quantify the high uncertainty of cross-camera sample relations, we parameterize the probabilistic relations between unlabeled samples in the relation discovery objective for pseudo label training and introduce reward quantified by identification performance on limited labeled data for this process. The probabilistic relations are learned to maximize the reward, so that minimization of the relation discovery objective can reduce the uncertainty of the underlying sample relations. The overview of rewarded pseudo label training is shown in Figure 3. Some important notations are defined in Table 3 for clarity.

Rewarded Pseudo Label Training. To guide relation discovery for pseudo label training on unlabeled data, we make use of a small amount of labeled data to provide reward of identification performance.

The rewarded pseudo label training strategy consists of two objectives: the relation discovery objective on unlabeled data for pseudo label training and a few-shot validation objective on limited labeled data for quantifying the identification performance as reward. The strategy is formulated by bi-level optimization [25] as:

$$\begin{aligned} &\min_{\mathbf{\Theta}_{RD}} \ \mathcal{L}_{val}(\mathcal{D}_L; \mathbf{\Theta}^*), \\ &s.t. \ \mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \ \mathcal{L}_{RD}(\mathcal{D}_U; \mathbf{\Theta}_{RD}, \mathbf{\Theta}), \end{aligned} \tag{1}$$

where $\mathbf{\Theta}_{RD}$ is the parameter of the relation discovery objective function $\mathcal{L}_{RD}$ on unlabeled data; $\mathcal{L}_{val}$ is the few-shot validation objective function on labeled data; $\mathbf{\Theta}$ is the parameter of the feature extractor F.

When minimizing the relation discovery objective function $\mathcal{L}_{RD}$, we expect to achieve maximization of the reward of identification performance on limited labeled data, i.e., minimization of the few-shot validation loss $\mathcal{L}_{val}$. By minimizing $\mathcal{L}_{val}$, the parameter $\mathbf{\Theta}_{RD}$ of the relation discovery objective function $\mathcal{L}_{RD}$ is learned to discover the probabilistic relations of cross-camera sample pairs. To realize this strategy, we propose *Rewarded Relation Discovery (R^2D)*, of which the pipeline is shown in Figure 3. The objective functions are formulated as follows:

$$\begin{aligned} &\min_{\hat{\mathbf{R}}^{dyn}} \ \mathcal{L}_{id}(\mathcal{D}_L; \mathbf{\Theta}^*, \{\mathbf{p}_c^*\}), \\ &s.t. \ (\mathbf{\Theta}^*, \{\mathbf{p}_c^*\}) = \arg\min_{\mathbf{\Theta}, \{\mathbf{p}_c\}} \ \mathcal{L}_{cd}^{dyn}(\mathcal{D}_U; \hat{\mathbf{R}}^{dyn}, \mathbf{\Theta}, \{\mathbf{p}_c\}), \end{aligned} \tag{2}$$

where $\hat{\mathbf{R}}^{dyn}$ is the cluster relation matrix that plays the role of $\mathbf{\Theta}_{RD}$ in Eq. (1) for parameterizing the probabilistic relations; $\mathbf{\Theta}$ is the parameter of feature extraction model F; $\{\mathbf{p}_c\}$ is the set of cluster prototypes; the dynamic cluster discrimination loss $\mathcal{L}_{cd}^{dyn}$ plays the role of the relation discovery objective $\mathcal{L}_{RD}$ in Eq. (1); the identification loss $\mathcal{L}_{id}$ plays the role of the few-shot validation objective $\mathcal{L}_{val}$ in Eq. (1). These losses and parameters are detailed below.

1) Relation Discovery Objective: To efficiently model the probabilistic relation between different samples, we first represent similar samples by clusters, and then learn the probabilistic cluster relations instead of the probabilistic instance relations. We design the relation discovery objective based on cluster discrimination. For unlabeled data set $\mathcal{D}^U$, we divide the unlabeled samples into C_U clusters by some clustering algorithm and use C_U prototypes $\{\mathbf{p}_c\}_{c=1}^{C_U}$ to represent each cluster. For labeled data set $\mathcal{D}^L$, we use C_L prototypes $\{\mathbf{p}_c\}_{c=C_U+1}^{C_U+C_L}$ to represent C_L identities. By merging the prototypes of unlabeled data and labeled data, there are totally $C = C_U + C_L$ prototypes denoted by $\{\mathbf{p}_c\}_{c=1}^{C}$. To quantify the probabilistic relation between sample $\mathbf{I}_i^U$ and the samples in data set $\mathcal{D}_U \cup \mathcal{D}_L$, we assign soft label $\hat{\mathbf{r}}_i \in \mathbb{R}^C$ to sample $\mathbf{I}_i^U$ as learning target to indicate the affinity between the unlabeled feature $\mathbf{x}_i^U$ and the prototypes $\{\mathbf{p}_c\}_{c=1}^{C}$.

We define a cluster relation matrix $\hat{\mathbf{R}}^{dyn} \in \mathbb{R}^{C_U \times C}$ as a learnable parameter. The non-diagonal elements represent the probabilistic relations between the unlabeled prototypes and all prototypes; the diagonal elements represent the degree of variation within a cluster. To embed prior knowledge of the clusters in $\hat{\mathbf{R}}^{dyn}$, its initial value $\hat{\mathbf{R}}^{dyn(0)}$ is

$$\hat{r}_{c_k,c_l}^{dyn(0)} = \begin{cases} \lambda & c_k = c_l \text{ and } c_k, c_l \in \mathcal{Y}_U, \\ (1-\lambda)/C_U & c_k \neq c_l \text{ and } c_k, c_l \in \mathcal{Y}_U, \\ 0 & c_k \in \mathcal{Y}_U \text{ and } c_l \in \mathcal{Y}_L, \end{cases} \tag{3}$$

where $\hat{r}_{c_k,c_l}^{dyn(0)}$ is the element in the c_k-th row and the c_l-th column of $\hat{\mathbf{R}}^{dyn(0)}$; $\hat{r}_{c_k,c_l}^{dyn(0)}$ indicates the target affinity between cluster c_k and cluster c_l; λ is a hyper-parameter of label smoothing for balancing the affinity of the corresponding cluster and other clusters, which is set as 0.8 empirically to allow uncertainty in the affinity between different clusters and intra-cluster variations; $\mathcal{Y}_U$ and $\mathcal{Y}_L$ denote the index sets of prototypes of unlabeled data and labeled data, respectively. We assume that the identities are non-overlapping in unlabeled data and labeled data, so that $\hat{r}_{c_k,c_l}^{dyn(0)}$ is 0 when $c_k \in \mathcal{Y}_U$ and $c_l \in \mathcal{Y}_L$.

For unlabeled sample $\mathbf{I}_i^U$ of which the cluster index is c_i, its pseudo soft label is $\hat{\mathbf{r}}_i = \hat{\mathbf{r}}_{c_i}^{dyn}$, the c_i-th row of the cluster relation matrix $\hat{\mathbf{R}}^{dyn}$. For pseudo label training, we minimize the dynamic cluster discrimination loss $\mathcal{L}_{cd}^{dyn}$ as follows:

$$\mathcal{L}_{cd}^{dyn}(\mathcal{D}_U; \hat{\mathbf{R}}^{dyn}, \mathbf{\Theta}, \{\mathbf{p}_c\}) = -\sum_{i=1}^{N_U}\sum_{c=1}^{C} \hat{r}_{c_i,c}^{dyn} \log \frac{\exp(\mathbf{p}_c^\top \mathbf{x}_i^U)}{\sum_{k=1}^{C} \exp(\mathbf{p}_k^\top \mathbf{x}_i^U)}, \tag{4}$$

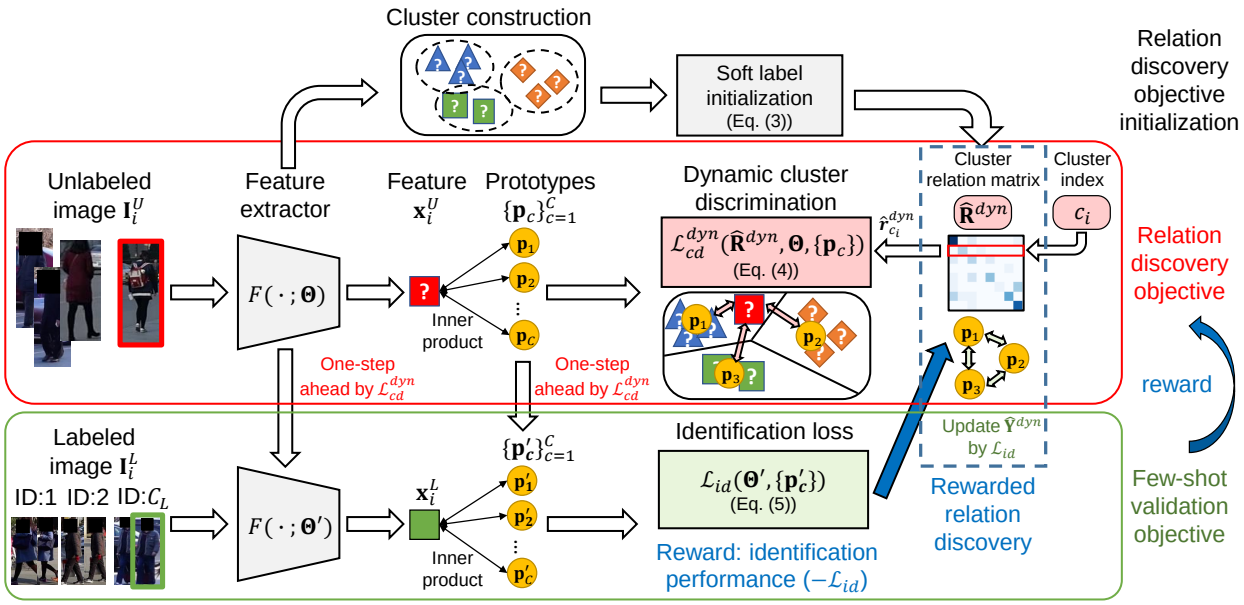

Fig. 3. Pipeline of Rewarded Relation Discovery (R^{2}D). It is a rewarded pseudo label training strategy formulated by bi-level optimization. First, clusters are constructed based on features of unlabeled data to initialize the relation discovery objective. Different cluster construction methods can embed different prior knowledge for the relation discovery objective. Then, the relation discovery objective of cluster discrimination on unlabeled data is regarded as the lower-level problem and the validation objective on limited labeled data is regarded as the upper-level problem. The cluster relation matrix $\hat{\mathbf{R}}^{dyn}$ explicitly quantifies uncertainty of sample relations. By updating it with bi-level optimization, the probabilistic relations between clusters can be discovered for dynamic cluster discrimination loss $\mathcal{L}_{cd}^{dyn}$ with the reward from identification loss $\mathcal{L}_{id}$ on labeled data. Minimization of $\mathcal{L}_{cd}^{dyn}$ can reduce the uncertainty in training data with identities rarely crossing camera views.

where $\mathbf{x}_i^U = F(\mathbf{I}_i^U;\boldsymbol{\Theta})$ is the feature extracted by model F; $\boldsymbol{\Theta}$ and $\mathbf{p}_c$ are parameters of the model; $\hat{\mathbf{R}}^{dyn}$ is the parameter of the relation discovery objective. During training, both the affinities between different clusters and the degrees of intra-cluster variations are learned in $\hat{\mathbf{R}}^{dyn}$.

2) Few-Shot Validation Objective: To learn $\hat{\mathbf{R}}^{dyn}$ in the relation discovery objective, we exploit a small amount of labeled data $\mathcal{D}_L$ to provide reward for learning $\hat{\mathbf{R}}^{dyn}$ by formulating a few-shot validation objective to quantify identification performance. Following popular representation learning methods for Re-ID [76], we jointly apply a cross entropy classification loss and a triplet loss on the labeled data $\mathcal{D}_L$ to measure identity discriminative degree more strictly. The identification loss $\mathcal{L}_{id}$ is formulated as:

$$\mathcal{L}_{id}(\mathcal{D}_L;\boldsymbol{\Theta},\{\mathbf{p}_c\}) = -\sum_{i=1}^{N_L}\sum_{c=1}^{C} y_{i,c}\log\frac{\exp(\mathbf{p}_c^\top \mathbf{x}_i^L)}{\sum_{k=1}^{C}\exp(\mathbf{p}_k^\top \mathbf{x}_i^L)} + \sum_{(a,p,n)\in\mathcal{I}_{tri}} \max(\left\|\mathbf{x}_a^L - \mathbf{x}_p^L\right\|_2 - \left\|\mathbf{x}_a^L - \mathbf{x}_n^L\right\|_2 + m, 0), \tag{5}$$

where the first term is cross entropy classification loss and the second term is triplet loss. In the cross entropy loss, $\mathbf{x}_i^L = F(\mathbf{I}_i^L;\boldsymbol{\Theta})$ is the feature of labeled sample $\mathbf{I}_i^L$; $y_{i,c}$ is the c-th element of the groundtruth one-hot label $\mathbf{y}_i \in \mathbb{R}^C$. In the triplet loss, m is the parameter of margin; $\mathcal{I}_{tri}$ is the index set of triplets. (a,p,n) denotes the indices of anchor sample, positive sample and negative sample, respectively.

Analysis. In the nested bi-level optimization problem [25], minimizing the dynamic cluster discrimination loss $\mathcal{L}_{cd}^{dyn}$ (the relation discovery objective) on unlabeled data $\mathcal{D}_U$ is regarded as the lower-level problem in the inner loop; minimizing the identification loss $\mathcal{L}_{id}$ (the few-shot validation objective) on labeled data $\mathcal{D}_L$ is regarded as the upper-level problem in the outer loop. The cluster relation matrix $\hat{\mathbf{R}}^{dyn}$ is the key parameter that connects the relation discovery objective and the few-shot validation objective. In the inner loop, it parameterizes the probabilistic cluster relations for pseudo label training by $\mathcal{L}_{cd}^{dyn}$. In the outer loop, it is optimized for discovering probabilistic relations by maximizing the reward of the identification performance (i.e., minimizing $\mathcal{L}_{id}$) on limited labeled data.

Comparison between conventional pseudo labeling methods and our proposed rewarded pseudo label training strategy is shown in Figure 4. Conventional pseudo labeling methods in Figure 4(a) performs pseudo label training without reward from labeled data, while our method in Figure 4(b) can learn the probabilistic relations parameterized by $\boldsymbol{\Theta}_{RD}$ with the reward from labeled data to quantify and reduce the high uncertainty in training data with identities rarely crossing camera views. To better understand uncertainty reduction, we quantify the uncertainty by entropy of classification probabilities and compare the entropy distributions in Section 2 of the supplementary material.

3.4 Bi-Level Optimization

To make the bi-level optimization problem [25] in Eq. (2) feasible, we relax the constraint of using the optimal $\boldsymbol{\Theta}^*$ and $\{\mathbf{p}_c^*\}$ in the inner loop following the gradient-based hyperparameter optimization methods [67]. The optimal $\boldsymbol{\Theta}^*$ and $\{\mathbf{p}_c^*\}$ are approximated by $\boldsymbol{\Theta}'$ and $\{\mathbf{p}'_c\}$ obtained by one-step gradient descent as follows:

$$\boldsymbol{\Theta}' = \boldsymbol{\Theta}^{(t)} - \alpha\frac{\partial\mathcal{L}_{cd}^{dyn}(\mathcal{D}_U;\hat{\mathbf{R}}^{dyn(t)},\boldsymbol{\Theta}^{(t)},\{\mathbf{p}_c^{(t)}\})}{\partial\boldsymbol{\Theta}^{(t)}}, \tag{6}$$

$$\mathbf{p}'_c = \mathbf{p}_c^{(t)} - \alpha\frac{\partial\mathcal{L}_{cd}^{dyn}(\mathcal{D}_U;\hat{\mathbf{R}}^{dyn(t)},\boldsymbol{\Theta}^{(t)},\{\mathbf{p}_c^{(t)}\})}{\partial\mathbf{p}_c^{(t)}}, \tag{7}$$

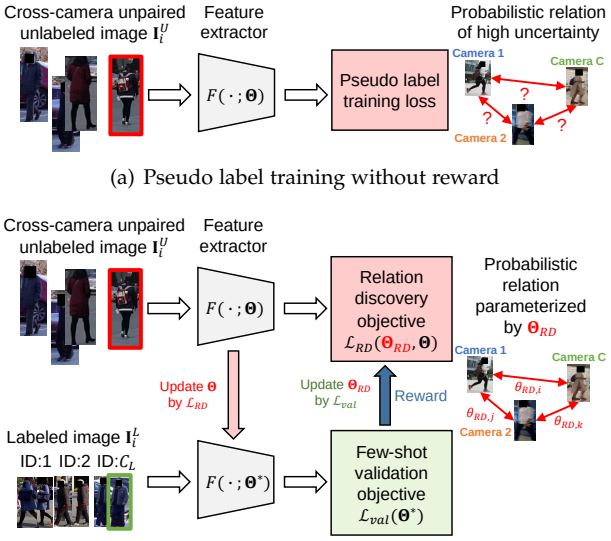

Fig. 4. Comparison between pseudo label training without and with reward. (a) Conventional pseudo labeling methods discover probabilistic relations and associate unlabeled sample pairs without reward from labeled data. The probabilistic relations are highly uncertain. (b) In our method, the probabilistic relations are parameterized by $\mathbf{\Theta}_{RD}$ in the relation discovery loss $\mathcal{L}_{RD}$ on unlabeled data. Then, the probabilistic relations are learned with the reward from labeled data.

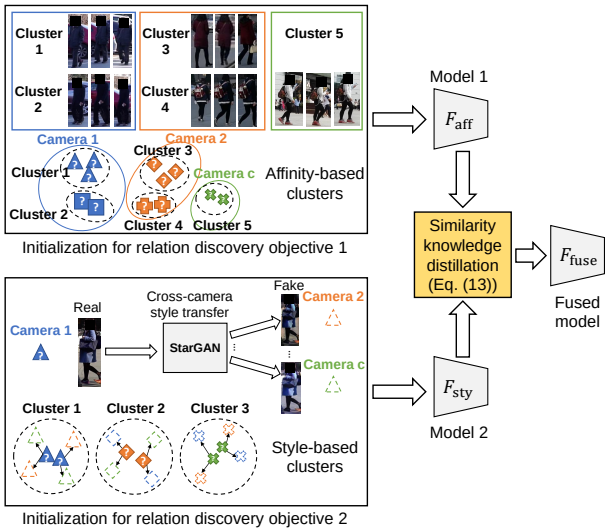

Fig. 5. Overview of multiple relation discovery objectives learning. Based on different initializations of affinity-based clusters and style-based clusters for the relation discovery objectives, model 1 F_{aff} and model 2 F_{sty} are learned by rewarded relation discovery (R^{2}D). The complementary knowledge of probabilistic relations learned by two different relation discovery objectives is fused by similarity knowledge distillation.

where α is the step size and the superscript (t) denotes that the parameter is in the t-th iteration of optimization.

By substituting $\mathbf{\Theta}'$ and $\{\mathbf{p}'_c\}$ into Eq. (2), the objective becomes minimizing $\mathcal{L}_{id}(\mathbf{\Theta}', \{\mathbf{p}'_c\})$. The cluster relation matrix $\hat{\mathbf{R}}^{dyn}$ is updated by

$$\hat{\mathbf{R}}^{dyn(t+1)} = \hat{\mathbf{R}}^{dyn(t)} - \gamma_R \frac{\partial \mathcal{L}_{id}(\mathcal{D}_L; \mathbf{\Theta}', \{\mathbf{p}'_c\})}{\partial \hat{\mathbf{R}}^{dyn(t)}}, \tag{8}$$

where γ_R is learning rate for cluster relation matrix $\hat{\mathbf{R}}^{dyn}$.

We use the updated relation discovery objective parameterized by $\hat{\mathbf{R}}^{dyn(t+1)}$ to learn the feature extraction model parameter $\mathbf{\Theta}$ and the prototype parameters $\{\mathbf{p}_c\}$ by

$$\mathbf{\Theta}^{(t+1)} = \mathbf{\Theta}^{(t)} - \gamma_\theta \frac{\partial \mathcal{L}_{cd}^{dyn}(\mathcal{D}_U; \hat{\mathbf{R}}^{dyn(t+1)}, \mathbf{\Theta}^{(t)}, \{\mathbf{p}_c^{(t)}\})}{\partial \mathbf{\Theta}^{(t)}}, \tag{9}$$

$$\mathbf{p}_c^{(t+1)} = \mathbf{p}_c^{(t)} - \gamma_p \frac{\partial \mathcal{L}_{cd}^{dyn}(\mathcal{D}_U; \hat{\mathbf{R}}^{dyn(t+1)}, \mathbf{\Theta}^{(t)}, \{\mathbf{p}_c^{(t)}\})}{\partial \mathbf{p}_c^{(t)}}, \tag{10}$$

where γ_θ and γ_p are learning rates for model parameter $\mathbf{\Theta}$ and prototypes $\{\mathbf{p}_c\}$, respectively.

In summary, the bi-level optimization problem in Eq. (2) is relaxed and solved by iterative update. Given parameters $\mathbf{\Theta}^{(t)}$, $\{\mathbf{p}_c^{(t)}\}$ and $\hat{\mathbf{R}}^{dyn(t)}$ of the t-th iteration, the update process from Eq. (6) to Eq. (10) is a loop for obtaining $\mathbf{\Theta}^{(t+1)}$, $\{\mathbf{p}_c^{(t+1)}\}$ and $\hat{\mathbf{R}}^{dyn(t+1)}$ in the next iteration. The pseudo code of training is shown in the supplementary material.

Finally, after the iteration of the maximum times t_{max}, the feature extractor $F(\cdot, \mathbf{\Theta}^{(t_{max})})$ is used for inference.

4 Multiple Relation Discovery Objectives Learning

In our proposed rewarded relation discovery strategy, the initialization of the cluster relation matrix $\hat{\mathbf{R}}^{dyn(0)}$ in the dynamic cluster discrimination loss $\mathcal{L}_{cd}^{dyn}$ embeds prior knowledge in the initial relation discovery objective. To benefit from different types of prior knowledge (intra-camera affinity and cross-camera style variation), we investigate affinity-based cluster discrimination and style-based cluster discrimination. Furthermore, we fuse the complementary knowledge of different probabilistic relations learned by different relation discovery objectives in a single model by similarity knowledge distillation, in order to further reduce the uncertainty of underlying sample relations. The overview is shown in Figure 5.

4.1 Affinity-based Cluster Construction

We assume that the initial Re-ID model $F(\cdot;\mathbf{\Theta}^{(0)})$ is pre-trained on a source domain. The affinity between samples determined by the pre-trained model contains prior knowledge for matching. Because of unseen cross-camera scene variations in the target domain, the affinities between cross-camera sample pairs are less reliable than the affinities between intra-camera sample pairs. Thus, we apply intra-camera clustering on unlabeled data.

For unlabeled data set $\mathcal{D}_U = \{(\mathbf{I}_i^U, v_i^U)\}_{i=1}^{N_U}$, a feature set $\mathcal{X}_U^{(0)} = \{\mathbf{x}_i^{U(0)}\}_{i=1}^{N_U}$ is extracted by the initial Re-ID model $F(\cdot;\mathbf{\Theta}^{(0)})$. Then, a pairwise distance matrix $\mathbf{D}$ is computed based on the features. To avoid associating cross-camera sample pairs, we convert $\mathbf{D}$ to an intra-camera distance matrix $\mathbf{D}^{intra}$ by replacing the distances between cross-camera pairs with the maximum distance. By applying DBSCAN [77], a popular clustering algorithm for Re-ID [78], the cluster index vector $\mathbf{c}^{intra}$ is determined by

$$\mathbf{c}^{intra} = \text{DBSCAN}(\mathbf{D}^{intra}), \tag{11}$$

where the i-th element c_i^{intra} in $\mathbf{c}^{intra}$ denotes the cluster index of unlabeled sample $\mathbf{I}_i^U$.

By substituting cluster index c_i^{intra} into c_i in $\mathcal{L}_{cd}^{dyn}$ (Eq. (4)), we embed the prior knowledge of intra-camera sample affinity for rewarded relation discovery, as shown in relation discovery objective 1 in Figure 5.

4.2 Style-based Cluster Construction

Affinity-based cluster construction embeds affinity-based prior knowledge for intra-camera sample pairs. Moreover, we expect that the learned features can be robust to cross-camera image style variations, such as lighting condition and background, but the identities crossing camera views are rare for aligning the features of cross-camera positive pairs. To take the place of the missing cross-camera positive pairs, we generate cross-camera fake samples by using an image-image translation model $T_{CamStyle}$ based on StarGAN [79] following the approach in HHL [14]. $T_{CamStyle}$ is trained on unlabeled data $\mathcal{D}_U$ with camera view labels to enable translation between any camera pairs. Given an image $\mathbf{I}_i^U$ and the target camera view label $v \in \{1, 2, ..., V_{cam}\}$, a fake image $\tilde{\mathbf{I}}_i^{U,v}$ of camera view v is generated by image-image translation as

$$\tilde{\mathbf{I}}_i^{U,v} = T_{CamStyle}(\mathbf{I}_i^U, v). \tag{12}$$

To generate fake images more efficiently, we select representative real unlabeled samples by clustering for image translation. Similar to affinity-based clusters in Section 4.1, we apply intra-camera clustering based on DBSCAN [77] for real unlabeled samples $\mathbf{I}_i^U$. For each cluster, we randomly select N_{real} samples denoted as $\mathbf{I}_1^{U,sel}, \mathbf{I}_2^{U,sel}, ..., \mathbf{I}_{N_{real}}^{U,sel}$ and remove the other samples in this cluster. Then, for each selected real sample $\mathbf{I}_i^{U,sel}$, we translate it to all camera views to obtain $\tilde{\mathbf{I}}_i^{U,1}, \tilde{\mathbf{I}}_i^{U,2}, ..., \tilde{\mathbf{I}}_i^{U,V_{cam}}$. We regard the real image $\mathbf{I}_i^{U,sel}$ and the fake images $\tilde{\mathbf{I}}_i^{U,1}, \tilde{\mathbf{I}}_i^{U,2}, ..., \tilde{\mathbf{I}}_i^{U,V_{cam}}$ as images of the same identity and assign the same cluster index to them.

By augmenting the real unlabeled training set $\mathcal{D}_U$ with the translated unlabeled samples, we embed the prior knowledge of cross-camera image style variations for rewarded relation discovery, as shown in relation discovery objective 2 in Figure 5.

4.3 Knowledge Fusion for Relation Discovery Objectives

We denote the model learned based on affinity-based clusters as F_{aff} and denote the model learned based on style-based clusters as F_{sty}. As shown in Figure 5, the intra-cluster variations for affinity-based clusters are mainly poses and orientations; while the intra-cluster variations for style-based clusters are mainly lighting and background. By discriminating these two types of clusters, F_{aff} and F_{sty} learn features that are robust to different types of variations and thus complementary to each other. We fuse the complementary knowledge of different probabilistic relations in a single model F_{fuse} to further reduce the uncertainty in underlying sample relations without increasing inference costs.

To achieve this, we regard F_{aff} and F_{sty} as two teacher models to learn a student model F_{fuse} by similarity knowledge distillation [40]. Given the unlabeled data set $\mathcal{D}_U$, we extract feature matrices $\mathbf{X}_a$, $\mathbf{X}_s$ and $\mathbf{X}_f$ by $F_{\text{aff}}(\cdot\,;\boldsymbol{\Theta}_a)$, $F_{\text{sty}}(\cdot\,;\boldsymbol{\Theta}_s)$ and $F_{\text{fuse}}(\cdot\,;\boldsymbol{\Theta}_f)$, respectively. In the feature matrices $\mathbf{X}_a$, $\mathbf{X}_s$ and $\mathbf{X}_f$, the vector in the i-th column denotes the feature vector normalized by $\ell 2$ norm for sample $\mathbf{I}_i^U$. The similarity knowledge distillation loss is

$$\min_{\boldsymbol{\Theta}_f} w_a \left\| \mathbf{X}_a^\top \mathbf{X}_a - \mathbf{X}_f^\top \mathbf{X}_f \right\|_1 + w_s \left\| \mathbf{X}_s^\top \mathbf{X}_s - \mathbf{X}_f^\top \mathbf{X}_f \right\|_1, \tag{13}$$

where $\boldsymbol{\Theta}_f$ is the parameter of the fused model $F_{\text{fuse}}(\cdot\,;\boldsymbol{\Theta}_f)$ and $\|\cdot\|_1$ is the entry-wise $\ell 1$ norm; w_a and w_s are trade-off parameters.

For training, the similarity knowledge distillation loss in Eq. (13) was applied on unlabeled data. The identification loss in Eq. (5) is applied on labeled data in addition to the distillation loss. In the inference stage, only the fused model F_{fuse} is used to extract features for matching.

TABLE 4
Comparisons between our constructed datasets REID-CBD, DukeMTMC-NA, MSMT17-NA for learning Re-ID on identities rarely crossing camera views and benchmark datasets DukeMTMC, MSMT17. "#site" denotes the number of sites. "#ID" denotes the number of identities. "#bbox" denotes the number of bounding boxes. "#cam" denotes the number of cameras. "UKN" denotes unknown. Compared with existing publicly available benchmark datasets, our dataset REID-CBD was captured in 6 sites and there were rare identities crossing camera views in the training set. DukeMTMC-NA and MSMT17-NA were simulated by removing samples of identities crossing camera views.

Dataset	#site	#cam	#cross-camera ID (train)	#ID (train)	#ID (test)	#bbox
REID-CBD	6	6	10	UKN	103*	146,510
DukeMTMC-NA	1	8	10	702	1,110	26,199
MSMT17-NA	1	15	10	1,041	3,060	100,777
DukeMTMC [21]	1	8	702	702	1,110	36,411
MSMT17 [19]	1	15	1,041	1,041	3,060	126,441

* 103 is the number of identities in query set. Since there are 113,031 distractors in gallery set, the total number of identities is unknown.

5 Constructing Datasets with Identities Rarely Crossing Camera Views

We have introduced a rewarded semi-supervised learning strategy for learning Re-ID on identities rarely crossing camera views, which is still an under-explored problem. Existing benchmark datasets Market-1501 [20], DukeMTMC [21] and MSMT17 [19] were captured in small-scale surveillance systems in a local area of one site. Most persons appear in at least two cameras in the training set. Person30K [23] was captured in multiple sites, but it does not contain identities crossing sites for evaluating Re-ID across sites and it is not public yet. Since existing publicly available benchmark datasets were captured only in one site and not suitable for evaluation of learning Re-ID on identities rarely crossing camera views, we construct new datasets in two ways: collecting a new real-world dataset in nonadjacent scenes and simulating this scenario on existing benchmark datasets. The comparisons between our constructed datasets and existing benchmark datasets are shown in Table 4.

5.1 Real-World Dataset REID-CBD

We collected a new multi-site dataset called REID-CBD captured in nonadjacent scenes in real-world scenario, which can be publicly available after data masking. We

TABLE 5
Splits of our constructed datasets REID-CBD, DukeMTMC-NA and MSMT17-NA. "#bbox" denotes the number of bounding boxes. "#ID" denotes the number of identities. "UKN" denotes unknown.

Dataset	Training set				Testing set			
	unlabeled		labeled		query		gallery	
	#bbox	#ID	#bbox	#ID	#bbox	#ID	#bbox	#ID
REID-CBD	24,000	UKN	131	10	4,647	103	117,732	UKN
DukeMTMC-NA	6,148	692	162	10	2,228	702	17,661	1,110
MSMT17-NA	6,607	1,031	350	10	11,659	3,060	82,161	3,060

Fig. 6. Samples of our REID-CBD dataset. In each row for a camera, every two images are of the same identity. Since the images were captured in different distant sites, identities crossing camera views hardly exist in the unlabeled training data. The locations of the cameras are shown on the map. To avoid privacy problem, the last step of data processing is masking the faces of all persons.

selected 6 nonadjacent crowded scenes as 6 sites in the central business district (CBD) of a city to capture videos by cellphone cameras in the day. The locations of cameras on the map and the examples in each camera are shown in Figure 6. Occlusion problem makes it difficult to automatically annotate intra-camera identities by pedestrian trackers. Different locations are distant to each other, so that most persons unlikely appear in different locations in the period of video recording. There are illumination variations, viewpoint change and occlusions in different scenes. The person images are approved to be shown in research papers with privacy licenses of the captured pedestrians.

Data Collection and Processing. We captured long image sequences in 6 nonadjacent scenes by cellphone cameras. To obtain person images as training data, we applied Mask R-CNN [80] to detect person images from all captured frames and selected the bounding boxes with high confidence to form the unlabeled training data set $\mathcal{D}_U$, in which there was hardly any underlying cross-camera positive pairs. For each scene, there were 4000 person images in average.

Besides unlabeled data, we annotated a small amount of cross-camera positive pairs as labeled training data $\mathcal{D}_L$. Since it is difficult to find cross-camera positive pairs in nonadjacent scenes, we arranged for 113 actors to appear in all 6 scenes and then captured and annotated their images. Bounding boxes and identities of all actors were annotated manually to guarantee the quality of labeled training set and the testing set. For each camera, about 13 images of each identity were annotated. For labeled training data set $\mathcal{D}_L$, we used images of 10 actors. For testing set, we used images of the other 103 actors. Moreover, we used Mask R-CNN to additionally detect 113,031 person bounding boxes from image sequences that were non-overlapping with those of training sets $\mathcal{D}_U$ and $\mathcal{D}_L$ as distractors in gallery images.

The detailed statistics of the collected dataset are shown in Table 4. The number of identities of unlabeled training data and that of the gallery distractors are unknown, since the images were detected automatically from raw videos without annotation.

Evaluation Protocol. Training data includes the unlabeled data set $\mathcal{D}_U$ automatically detected from raw image sequences and the small labeled data set $\mathcal{D}_L$ of 10 identities captured and annotated manually.

For testing, the images of 103 actors are split half-and-half for each identity to form the query image set and the gallery image set. The 113,031 images as distractors are added in gallery images for evaluation in real-world situation. Each query image is matched with all gallery images to evaluate the matching performance. The split is shown in Table 5.

5.2 Simulated Datasets

Besides constructing real-world dataset, we used two large benchmark datasets DukeMTMC[2] [21] and MSMT17 [19] to simulate the scenario of Re-ID in nonadjacent scenes. We randomly selected 10 identities in the training set and used their images captured in multiple cameras to form the labeled training set $\mathcal{D}_L$. For the other samples in the training set, we randomly selected images of only one camera for each identity to construct the unlabeled training set $\mathcal{D}_U$, in which there was no cross-camera positive pair. The testing data and protocols were kept the same as those of the original datasets. The modified datasets are denoted by DukeMTMC-NA and MSMT17-NA, of which the splits are shown in Table 5. Compared with simulated datasets, the number of samples in our REID-CBD is much larger.

6 Experiments

To show the effectiveness of our method for semi-supervised Re-ID on identities rarely crossing camera views, we conducted comparative evaluations of our method against a wide range of Re-ID methods and related representation learning methods on three datasets specifically constructed for Re-ID in nonadjacent scenes.

6.1 Experiment Settings

Datasets. The datasets include our real-world dataset REID-CBD and two simulated datasets DukeMTMC-NA and MSMT17-NA introduced in Section 5. In our default setting, for each dataset, the training set consists of a labeled data set $\mathcal{D}_L$ of 10 identities and an unlabeled data set $\mathcal{D}_U$ without underlying cross-camera positive pairs.

Additionally, to simulate nonadjacent scenes with different distances, we evaluated using unlabeled data with different ratios of underlying identities crossing camera views for training. Moreover, we considered different types of supervisions, such as increasing labeled identities and annotating intra-camera identities.

2. In our evaluation, we used the data on DukeMTMC for academic use without identifying or showing the person images.

Evaluation Protocol. By default, the experiments were carried out in semi-supervised setting. At the training stage, the labeled training set of Market-1501 [20] was used for pre-training to initialize the model, since it is challenging to learn from unlabeled cross-camera unpaired data without prior knowledge of Re-ID. Then, on the target domain (REID-CBD, DukeMTMC-NA or MSMT17-NA), a large amount of unlabeled data $\mathcal{D}_U$ and a small amount of labeled data $\mathcal{D}_L$ were used for training.

At the testing stage, for REID-CBD, we followed the testing split in Section 5; for DukeMTMC-NA or MSMT17-NA, we followed the original testing split of DukeMTMC [21] and MSMT17 [19]. The rank-k accuracy and mean average precision (mAP) were applied as performance metrics following the standard evaluations of Re-ID [20].

6.2 Implementation Details

Backbone Model. We applied ResNet-50 [75] as the backbone for our model and all compared methods. We replaced the global average pooling (GAP) layer with generalized mean pooling [7] and used batch normalization neck (BN-Neck) [76] after the generalized mean pooling layer. The stride size of the last residual layer was set as 1. The prototypes $\{\mathbf{p}_c\}$ in Section 3.3 were implemented by a fully connected layer (FC) followed by Softmax function. Our model took person images resized to 256×128 as input.

Hyperparameters. For the initial value of cluster relation matrix $\hat{r}_{c_k,c_l}^{dyn(0)}$ in Eq. (3), we set $\lambda = 0.8$. In the identification loss $\mathcal{L}_{id}$ in Eq. (5), we set the margin parameter $m = 0.3$ for the triplet loss term. For DBSCAN clustering [77] in affinity-based cluster construction and style-based cluster construction, the maximum intra-cluster distance between two samples ϵ was 0.6 and the minimum number of samples was 2. For style-based cluster discrimination, the number of real images selected in each cluster was $N_{real} = 2$. In the similarity knowledge distillation loss in Equation (13), we set $w_a = w_s = 1$.

At the training stage of rewarded relation discovery (R^2D), the step size α for optimum approximation in Eq. (6) and (7) was set as 0.1. The learning rate γ_R for cluster relation matrix in Eq. (8) was set as 0.06. The learning rates for model parameter $\mathbf{\Theta}$ and prototypes $\{\mathbf{p}_c\}$ were $\gamma_\theta = 6\times10^{-5}$ and $\gamma_p = 3 \times 10^{-3}$, respectively. For optimization, SGD [81] was applied, for which we set weight decay as 5×10^{-4} and momentum as 0.9. In each batch, $N_{BU} = 128$ images were randomly sampled from the unlabeled data set and $N_{BL} = 128$ in the labeled data set were used. The training process consisted of 120 epochs. The learning rates γ_θ and γ_p were multiplied by 0.1 in epoch 40 and epoch 90.

As for multiple relation discovery objectives learning, the image-image translation model $T_{CamStyle}$ was implemented by following the method in HHL [14]. For training of knowledge distillation, the weights of both the similarity knowledge distillation loss and the identification loss were set as 1. In each batch, 64 images were randomly sampled from the unlabeled data set and all images in the labeled data set were used. There were totally 60 epochs. The SGD optimizer with the same parameters as that for R^2D was used. The learning rate was initialized as 3×10^{-3} and multiplied by 0.1 in epoch 20 and epoch 40.

Our method was implemented on Pytorch and trained on 1 NVIDIA NVIDIA RTX A6000 GPU.

6.3 Compared Methods

To show the uncertainty reduction ability on training data with identities rarely crossing camera views, we compared our method with existing semi-supervised, unsupervised and supervised Re-ID methods, as well as some closely related representation learning methods including unsupervised domain adaptation, self-supervised learning, training objective optimization and clustering.

6.3.1 Re-ID Models

Unsupervised Re-ID. Since the performances of the state-of-the-art unsupervised Re-ID methods have been close to the performances of supervised learning, we compared our method with representative unsupervised Re-ID methods HHL [14], MMT [17], GLT [18], UNRN [47] and MEB-Net [37]. HHL [14] learns representation from generated images by camera style transfer, which is a representative generative adversarial method for Re-ID. MMT [17], GLT [18], UNRN [47] and MEB-Net [37] are advanced unsupervised Re-ID methods based on pseudo labeling. DG-Net [48] jointly learns image generation and discrimination for supervised learning and can be adapted for unsupervised Re-ID by replacing the groundtruth labels with pseudo labels. These unsupervised Re-ID models were applied on only unlabeled training data $\mathcal{D}_U$.

Semi-Supervised Re-ID. To utilize both unlabeled data $\mathcal{D}_U$ and limited labeled training data $\mathcal{D}_L$, we adapted the state-of-the-art unsupervised Re-ID methods to semi-supervised learning by additionally applying a cross entropy classification loss and a triplet loss for identifying labeled data as in the identification loss $\mathcal{L}_{id}$ in Eq. (5). The semi-supervised versions of unsupervised Re-ID methods HHL [14], MMT [17], GLT [18] UNRN [47], MEB-Net [37] and DG-Net [48] were denoted by HHL-semi, MMT-semi, GLT-semi, UNRN-semi, MEB-Net-semi, and DG-Net-semi, respectively.

Few-Shot Supervised Re-ID. To avoid negative transfer effect of unlabeled data, we evaluated using only the labeled data $\mathcal{D}_L$ for few-shot supervised learning. We compared with the state-of-the-art supervised Re-ID model AGW [7].

6.3.2 Representation Learning Methods

Semi-Supervised Learning and Domain Adaptation. For comparison with general unsupervised domain adaptation methods, we evaluated some representative methods MMD [82], CORAL [83], SpCL [42] and SBase [47]. MMD [82] and CORAL [83] are loss functions for moment matching. They were applied on our backbone model for camera-specific feature distribution alignment. SpCL [42] and SBase [47] are advanced pseudo labeling methods for domain adaptive object re-identification. Their semi-supervised versions MMD-semi, CORAL-semi, SpCL-semi and SBase-semi were adapted from the unsupervised versions in the same way as the semi-supervised Re-ID models in Section 6.3.1.

Self-Supervised Learning. As for self-supervised representation learning, we compared with advanced contrastive

TABLE 6
Comparisons with the state-of-the-art methods for person re-identification on unlabeled data $\mathcal{D}_U$ without identities crossing camera views and limited labeled data $\mathcal{D}_L$ of 10 identities crossing camera views. "R-k" denotes the rank-k accuracy (%) and "mAP" denotes mean average precision (%). The best performances are marked in **bold**. The second-best performances are marked by underline.

Methods		Reference	REID-CBD				DukeMTMC-NA				MSMT17-NA			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Few-shot supervised ($\mathcal{D}_L$)	AGW [7]	TPAMI'21	43.5	59.7	67.3	17.4	56.1	69.2	74.5	36.1	30.6	42.9	48.4	11.5
Unsupervised ($\mathcal{D}_U$)	HHL [14]	ECCV'18	38.2	55.2	63.6	13.0	58.7	70.9	75.2	37.9	24.5	34.7	40.0	8.5
	MMT [17]	ICLR'20	35.2	45.2	50.5	13.8	44.1	59.0	66.9	32.0	18.1	27.8	33.1	7.7
	GLT [18]	CVPR'21	24.1	30.7	34.5	6.5	47.3	63.1	69.8	34.1	17.1	26.9	32.4	7.1
	MMD [82]	JMLR'12	29.2	36.7	40.9	9.8	48.5	63.5	68.9	30.3	14.3	22.9	27.8	6.0
	CORAL [83]	ECCVW'16	29.2	36.7	40.9	9.8	44.3	58.9	65.3	32.4	13.8	21.9	26.8	5.7
	SpCL [42]	NeurIPS'20	27.6	34.2	37.5	8.0	46.1	61.4	67.7	27.0	18.3	29.4	34.8	6.0
	MEB-Net [37]	ECCV'20	30.5	39.8	45.0	9.9	44.8	59.1	66.2	31.6	16.7	26.6	31.9	7.0
	DG-Net [48]	CVPR'19	37.4	49.6	62.3	14.6	56.5	67.8	76.1	38.4	25.6	35.7	42.6	9.1
	UNRN [47]	AAAI'21	38.1	48.4	53.1	16.6	53.2	65.6	71.5	38.8	20.1	29.5	34.7	8.4
	SBase [47] (K-means)	AAAI'21	30.1	38.5	43.1	9.2	46.8	60.9	67.3	33.4	17.6	26.7	31.8	7.2
	SBase [47] (Mean shift)	AAAI'21	29.1	35.6	39.9	8.9	44.7	57.8	64.1	32.1	17.2	26.1	31.5	7.0
	SBase [47] (AP)	AAAI'21	29.6	36.3	39.8	9.4	45.6	59.3	65.4	32.2	17.5	26.3	31.8	7.1
	SBase [47] (DBSCAN)	AAAI'21	29.4	36.2	40.2	9.5	45.2	59.0	65.8	32.5	18.0	26.9	32.2	7.3
Semi-supervised ($\mathcal{D}_U + \mathcal{D}_L$)	HHL-semi [14]	ECCV'18	45.8	63.1	70.7	18.4	59.2	72.6	77.4	39.9	27.6	38.9	44.5	9.5
	MMT-semi [17]	ICLR'20	36.3	47.1	52.9	15.7	47.7	61.1	67.8	33.4	21.8	32.2	38.3	9.2
	GLT-semi [18]	CVPR'21	27.6	35.8	41.5	9.2	50.6	65.7	72.0	36.3	21.7	32.3	38.2	9.0
	MMD-semi [82]	JMLR'12	31.5	39.8	45.0	12.5	50.5	65.9	70.6	33.1	18.6	28.4	35.1	7.8
	CORAL-semi [83]	ECCVW'16	31.7	40.7	45.6	12.3	45.9	60.1	68.6	34.6	17.8	27.8	33.0	7.5
	SpCL-semi [42]	NeurIPS'20	30.2	37.3	41.1	9.8	47.1	63.5	68.9	29.6	18.5	27.8	33.1	7.8
	MEB-Net-semi [37]	ECCV'20	32.0	43.1	49.2	12.5	47.3	62.7	70.1	34.4	20.6	31.6	37.6	8.3
	DG-Net-semi [48]	CVPR'19	44.2	60.5	69.4	18.1	60.2	70.5	78.1	40.4	28.9	40.1	45.1	9.9
	UNRN-semi [47]	AAAI'21	38.9	49.7	55.6	17.9	55.6	67.8	74.2	41.3	22.8	33.4	39.0	9.7
	SBase-semi [47] (K-means)	AAAI'21	34.9	44.8	50.7	14.2	49.3	62.3	68.8	34.8	20.1	29.6	35.2	8.7
	SBase-semi [47] (Mean shift)	AAAI'21	33.5	43.3	49.1	13.7	46.2	60.1	67.5	33.8	21.4	31.4	36.7	8.5
	SBase-semi [47] (AP)	AAAI'21	35.1	43.9	51.7	14.1	47.1	61.6	68.7	34.5	20.5	30.3	35.8	8.8
	SBase-semi [47] (DBSCAN)	AAAI'21	35.4	46.8	52.7	14.3	47.0	61.2	68.1	34.2	21.8	31.9	37.4	8.9
Self-supervised + fine-tune ($\mathcal{D}_U + \mathcal{D}_L$)	LUP [49]	CVPR'21	32.4	46.5	53.7	10.4	44.3	57.9	63.2	26.0	15.1	24.6	29.5	4.9
	SimSiam [84]	CVPR'21	33.8	49.0	56.7	11.6	40.1	54.6	60.1	22.0	19.0	29.0	34.0	5.9
Training Objective Optimization ($\mathcal{D}_U + \mathcal{D}_L$)	Reweight [72]	ICML'18	37.7	51.9	58.3	15.3	53.0	67.5	73.3	36.1	27.6	39.6	45.5	10.8
	MSLG [71]	ICPR'21	33.1	45.4	50.4	10.4	51.4	65.7	71.9	35.1	24.0	34.7	39.7	8.9
	R^2D (ours)	-	**59.0**	**76.9**	**82.5**	**28.2**	**64.9**	**78.2**	**81.7**	**44.5**	**39.4**	**52.2**	**57.8**	**15.2**

learning methods LUP [49] and SimSiam [84]. LUP is specifically designed for Re-ID task and SimSiam is for general image classification. After unsupervised pre-training on unlabeled training set $\mathcal{D}_U$, the model was fine-tuned on the labeled training set $\mathcal{D}_L$.

Training Objective Optimization. Our method can be regarded as learning the objective function on a small labeled validation set by bi-level optimization, which is a type of training objective optimization method. We compared with two representative training objective optimization methods Reweight [72] and MSLG [71]. Based on bi-level optimization, Reweight [72] learns different weights for different samples and MSLG [71] learns meta soft labels for each sample. Since they are designed for supervised closed-set image classification, we adapted these methods for semi-supervised Re-ID by assigning pseudo labels obtained by clustering to unlabeled data.

Clustering Algorithms. To compare with representative clustering algorithms, we applied K-means [85], Mean shift [86], affinity propagation (AP) [87] and DBSCAN [77] for a pseudo-labeling-based strong baseline method SBase [47] and its semi-supervised version SBase-semi.

6.3.3 Implementation for the Compared Methods

Implementations of all Re-ID models were based on the codes released by the papers. ResNet-50 [75] was used as backbone for all compared methods. By default, the models were initialized by pre-training on Market-1501 [20] to provide prior knowledge for Re-ID. For self-supervised learning methods, the models were initialized by self-supervised training and then further trained on Market-1501 [20].

6.4 Model Comparison and Analysis

Comparative experiments on REID-CBD, DukeMTMC-NA and MSMT17-NA datasets are shown in Table 6.

Comparison with Semi-Supervised and Unsupervised Re-ID Models. The performance of our method is clearly better than all compared semi-supervised and unsupervised Re-ID models. Our method outperformed the second-best method on rank-1 accuracy by 13.2%, 4.7% and 8.8% on REID-CBD, DukeMTMC-NA and MSMT17-NA, respectively.

When training with identities rarely crossing camera views, the advanced pseudo-label-based unsupervised Re-ID methods MMT [17], GLT [18] and their semi-supervised versions failed and they were even worse than supervised learning by AGW [7] on limited labeled data $\mathcal{D}_L$. These results demonstrate that, these pseudo-label-based methods cannot effectively discover the underlying sample relations of cross-camera unpaired data with high uncertainty and suffer from noise accumulation problem [24].

HHL-semi [14] and DG-Net-semi [48] can improve the performance as compared with supervised learning on $\mathcal{D}_L$, which indicates that diversifying training data variations by GANs is effective for cross-camera unpaired data.

The above models either used affinity-based clusters (MMT [17] and GLT [18]) or used style-based clusters (HHL [14]), where the cluster relations are determined without reward from labeled data. In comparison, our method automatically discovers the probabilistic relations by the reward from few-shot validation objective on limited labeled data $\mathcal{D}_L$ and thus can reduce uncertainty more effectively.

Comparison with Few-shot Supervised Training on $\mathcal{D}_L$. Our method outperformed a strong supervised Re-ID base-

line method AGW [7], since it overfitted limited labeled data. This indicates that training data with identities rarely crossing camera views is beneficial to improving identification performance when it was exploited by our method.

Comparison with Self-Supervised Learning Models. Our method outperformed all compared self-supervised learning methods. For LUP [49] and SimSiam [84], the pre-training objectives are designed without reward from labeled data. In our method, the reward from validation objective can guide the relation discovery objective to discover probabilistic relations for better alleviation of the uncertainty in underlying sample relations.

Comparison with Training Objective Optimization. Our method clearly outperformed Reweight [72] and MSLG [71]. During training, Reweight [72] learns to adjust the weights of different samples and MSLG [71] learns the meta soft labels for each sample individually. As Reweight [72] and MSLG [71] are designed for supervised learning on training sets with biases or noises, they ignore the cluster relations of unlabeled data for unsupervised learning. In comparison, our method parameterizes and learns the probabilistic relations between different clusters to reduce the uncertainty of underlying sample relations.

Comparison with Domain Adaptation Models. Our method outperformed the compared domain adaptation methods. The distribution-level alignment by MMD [82] and CORAL [83] ignore the probabilistic relations between samples. As for pseudo-label-based approaches SpCL [42] and SBase [47], the result analysis is similar to the above analysis for unsupervised Re-ID methods.

Comparison with Clustering Algorithms. Compared with SBase-semi based on different clustering algorithms, our method outperformed all of them. These clustering algorithms cannot learn the probabilistic relations between different clusters as our method to quantify the uncertainty.

6.5 Further Evaluations

We evaluated and analyzed the effectiveness of the components and sensitivity of hyperparameters in our method. To simulate different real-world scenarios, we evaluated training on data with different ratios of underlying identities crossing camera views. Moreover, we evaluated using different types of supervisions, such as more labeled identities, intra-camera supervision. In addition, to better understand the effect of rewarded relation discovery, we visualized the differences of probabilistic relations between different clusters before and after training. We also provide analysis for running time and convergence of loss.

6.5.1 Ablation Study

We evaluated the effectiveness of the key components of our method in Table 7. Our model based on different relation discovery objectives are denoted by $\text{R}^2\text{D}_{\text{aff}}$ and $\text{R}^2\text{D}_{\text{sty}}$. The subscripts "aff" and "sty" denote affinity-based clusters and style-based clusters, respectively.

Comparison with Baseline Models. To evaluate the performance of the baseline methods, we trained our backbone model on a source dataset Market-1501 [20] (denoted by "Direct transfer") and then fine-tuned it on the small labeled data set $\mathcal{D}_L$ of a few identities (denoted by "Fine-tune on $\mathcal{D}_L$") on the target dataset. Compared with "direct transfer", fine-tuning on labeled data of only 10 identities can improve the performance, but the model suffered from overfitting problem. Compared with the baseline models "Fine-tune on $\mathcal{D}_L$", our method achieved significant improvements of over 15% rank-1 accuracy on REID-CBD and about 10% rank-1 accuracy on DukeMTMC-NA and MSMT17-NA, which indicates the effectiveness of rewarded relation discovery on unlabeled data in our method.

TABLE 7
Ablation study results of our method R^2D. The subscripts "aff" and "sty" denote the model learned based on affinity-based clusters and style-based clusters. "Cluster" denotes clustering-based pseudo label training on unlabeled data $\mathcal{D}_U$ without reward. "GT intra-cam ID" denotes groundtruth intra-camera identities. "Cluster + fine-tune" denotes fine-tuning on labeled data $\mathcal{D}_L$ based on the model learned by clustering-based pseudo label training without reward. "rand $\hat{\mathbf{R}}^{dyn(0)}$" denotes initializing cluster relation matrix $\hat{\mathbf{R}}^{dyn}$ randomly. "$\text{R}^2\text{D}_{\text{xxx}}$ + $\text{R}^2\text{D}_{\text{xxx}}$" denotes score fusion of two models. "Distill two $\text{R}^2\text{D}_{\text{xxx}}$" denotes fusion of two models by knowledge distillation in our method. "R-1" denotes the rank-1 accuracy (%) and "mAP" denotes mean average precision (%).

Methods	REID-CBD		DukeMTMC-NA		MSMT17-NA	
	R-1	mAP	R-1	mAP	R-1	mAP
Direct transfer (baseline)	28.7	7.1	51.9	33.1	21.2	7.4
Fine-tune on $\mathcal{D}_L$ (baseline)	43.1	17.1	53.5	35.6	29.9	11.2
$\text{Cluster}_{\text{aff}}$	29.4	9.5	51.8	34.5	16.8	6.1
$\text{Cluster}_{\text{aff}}$ (GT intra-cam ID)	-	-	53.6	36.3	19.6	7.1
$\text{Cluster}_{\text{aff}}$ + fine-tune	44.7	18.6	55.5	38.1	28.3	10.5
$\text{R}^2\text{D}_{\text{aff}}$ (rand $\hat{\mathbf{R}}^{dyn(0)}$)	41.9	14.3	51.8	33.8	24.8	9.3
$\text{R}^2\text{D}_{\text{aff}}$	50.4	23.3	61.5	41.4	37.8	14.4
$\text{Cluster}_{\text{sty}}$	38.2	13.0	58.7	37.9	24.5	8.5
$\text{Cluster}_{\text{sty}}$ + fine-tune	45.8	18.4	59.2	39.9	27.6	9.5
$\text{R}^2\text{D}_{\text{sty}}$ (rand $\hat{\mathbf{R}}^{dyn(0)}$)	43.2	16.1	55.1	35.5	26.6	9.6
$\text{R}^2\text{D}_{\text{sty}}$	55.0	24.5	62.4	41.2	38.1	14.4
$\text{R}^2\text{D}_{\text{aff}} + \text{R}^2\text{D}_{\text{aff}}$	51.2	23.5	61.7	41.8	37.9	14.3
$\text{R}^2\text{D}_{\text{sty}} + \text{R}^2\text{D}_{\text{sty}}$	55.3	24.8	62.6	41.3	38.3	14.4
$\text{R}^2\text{D}_{\text{aff}} + \text{R}^2\text{D}_{\text{sty}}$	54.7	26.4	62.8	42.7	38.6	14.7
Distill two $\text{R}^2\text{D}_{\text{aff}}$	52.7	24.8	62.6	42.1	38.2	14.6
Distill two $\text{R}^2\text{D}_{\text{sty}}$	56.0	25.7	63.1	41.8	38.6	14.7
R^2D (full model)	59.0	28.2	64.9	44.5	39.4	15.2

Effectiveness of Reward. To maximize the reward for R^2D, the cluster relation matrix $\hat{\mathbf{R}}^{dyn}$ is learned by bi-level optimization in Eq. (2) during training to update the relation discovery objective for minimizing the few-shot validation objective. To evaluate the effectiveness of using reward, we learned relation discovery objective and validation objective separately with fixed $\hat{\mathbf{R}}^{dyn}$. We trained models with fixed soft pseudo labels obtained by clustering on unlabeled data $\mathcal{D}_U$ (denoted by "$\text{Cluster}_{\text{aff}}$" and "$\text{Cluster}_{\text{sty}}$") and then fine-tuned the models on labeled data $\mathcal{D}_L$ (denoted by "$\text{Cluster}_{\text{aff}}$ + fine-tune" and "$\text{Cluster}_{\text{sty}}$ + fine-tune"). For "$\text{Cluster}_{\text{aff}}$", we additionally used the groundtruth intra-camera identities to replace the pseudo labels for evaluating the ideal case for clustering that each cluster contains samples of one identity. This is denoted by "$\text{Cluster}_{\text{aff}}$ (GT intra-cam ID)".

Compared with the baseline model "Direct transfer" and "Fine-tune on $\mathcal{D}_L$", the performance gains of rewarded relation discovery $\text{R}^2\text{D}_{\text{aff}}$ and $\text{R}^2\text{D}_{\text{sty}}$ were much more significant than clustering on unlabeled data without reward ("$\text{Cluster}_{\text{aff}}$", "$\text{Cluster}_{\text{aff}}$ (GT intra-cam ID)" and "$\text{Cluster}_{\text{sty}}$") as well as using clustering-based pseudo label training and few-shot validation objective separately without reward ("$\text{Cluster}_{\text{aff}}$ + fine-tune" and "$\text{Cluster}_{\text{sty}}$ + fine-tune"). Reward from labeled data can guide relation

TABLE 8
Effect of smoothness parameter λ for initialization of $\hat{\mathbf{R}}^{dyn}$ in Eq. (3) for R^2D_{aff} on rank-1 (R-1) accuracy and mAP (%) on REID-CBD.

λ	0.6	0.7	0.8	0.9	1
R-1	48.7	49.9	50.4	50.0	50.4
mAP	21.5	22.0	23.3	22.5	22.8

TABLE 9
Effect of using different ϵ in DBSCAN and different clustering algorithms for R^2D_{aff} on rank-1 (R-1) accuracy and mAP (%) on REID-CBD.

Methods	DBSCAN				K-means	Mean-shift	AP
	$\epsilon = 0.4$	$\epsilon = 0.5$	$\epsilon = 0.6$	$\epsilon = 0.7$			
R-1	49.9	50.7	50.4	48.8	49.6	49.3	50.1
mAP	22.5	22.9	23.3	21.5	22.9	22.6	23.1

TABLE 10
Effect of margin parameter m of triplet loss in Equation (5) for R^2D_{aff} on rank-1 (R-1) accuracy and mAP (%) on REID-CBD.

m	0.0	0.1	0.3	0.5	0.7	0.9
R-1	49.7	50.2	50.4	50.3	50.1	50.0
mAP	22.9	23.1	23.3	23.2	23.0	23.1

TABLE 11
Effect of weights w_a and w_s in distillation loss in Equation (13) for R^2D_{aff} on rank-1 (R-1) accuracy and mAP (%) on REID-CBD.

w_a	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
w_s	2.0	1.8	1.6	1.4	1.2	1.0	0.8	0.6	0.4	0.2	0.0
R-1	54.4	55.8	57.1	57.8	58.6	59.0	58.1	57.2	56.5	54.9	53.6
mAP	25.3	25.9	26.7	27.4	27.9	28.2	27.6	27.1	26.2	25.4	24.4

discovery on unlabeled data to better quantify and reduce the uncertainty of underlying sample relations.

Effectiveness of Cluster Relation Initialization. The initialization of cluster relation matrix $\hat{\mathbf{R}}^{dyn}$ is the key to embedding either affinity-based prior knowledge or style-based prior knowledge in the relation discovery objective. Instead of using Eq. (3), we initialized each row of $\hat{\mathbf{R}}^{dyn}$ by a random soft label obtained by applying Softmax function to a random vector. We denote the cases of random initialization of $\hat{\mathbf{R}}^{dyn}$ for relation discovery objectives based on affinity-based clusters and style-based clusters by "R^2D_{aff} (rand $\hat{\mathbf{R}}^{dyn(0)}$)" and "$R^2D_{sty}$ (rand $\hat{\mathbf{R}}^{dyn(0)}$)", respectively.

Compared with "R^2D_{aff}" and "R^2D_{sty}", random initialization of the cluster relation matrix significantly degraded the performance, which demonstrates that initialization of the cluster relation matrix can effectively embed prior knowledge in relation discovery objective for rewarded relation discovery.

Effectiveness of Knowledge Fusion for Relation Discovery Objectives. To show the effectiveness of similarity knowledge distillation for fusing the knowledge of two relation discovery objectives, we evaluated score fusion for models R^2D_{aff} and R^2D_{sty} by summing up the distances, which is denoted by "R^2D_{aff}+R^2D_{sty}". We also evaluated score fusion (R^2D_{aff} + R^2D_{aff}" and "R^2D_{sty} + R^2D_{sty}") and distillation-based fusion in our method ("Distill two R^2D_{aff}" and "Distill two R^2D_{sty}") for two models trained individually by the same relation discovery objective.

Compared with "Distill two R^2D_{aff}" and "Distill two R^2D_{sty}", our full model R^2D that fuses knowledge of different relation discovery objectives achieves better performance, which indicates that complementary knowledge is learned from affinity-based clusters and style-based clusters. Score fusion methods "R^2D_{xxx}+R^2D_{xxx}" only bring marginal improvement on single models, since late fusion ignores the relation between features of different samples; while our similarity knowledge distillation fuses the complementary knowledge of two models more effectively. We note that the style-based clusters provide more signal than the affinity-based clusters for Re-ID, because the style-based clusters contain variations of lighting and background that dominate cross-camera variations of Re-ID, while the affinity-based clusters contain pose and orientation variations within the same camera. For inference, the fused model is of the same size as a single model and thus more computationally efficient than model ensemble.

6.5.2 Parameter Analysis

We analyzed the effect of some key hyperparameters of our method R^2D_{aff} on REID-CBD dataset.

Smoothness Parameter λ for Initialization of $\hat{\mathbf{R}}^{dyn}$. For initialization of cluster relation matrix $\hat{\mathbf{R}}^{dyn}$, λ in Eq. (3) controls the smoothness of the initial soft labels. We varied λ from 0.6 to 1.0 with step size of 0.1. The results on REID-CBD are shown in Table 8. The performance variation is small and the best performance is achieved when $\lambda = 0.8$. When λ decreases, the soft labels become increasingly smooth and excessive uncertainty in the cluster relations degrades the effect of the embedded prior knowledge.

Parameter ϵ of DBSCAN Clustering and Different Clustering Algorithms. The maximum intra-cluster distance between two samples ϵ is the key parameter for DBSCAN clustering used in our relation discovery objectives. We varied ϵ from 0.4 to 0.7 for R^2D_{aff}. Moreover, as our method does not depend on specific clustering algorithm, we also evaluated cluster construction by K-means [85] ($k = 2000$), Mean shift [86], affinity propagation (AP) [87]. The results on REID-CBD are shown in Table 9. The results show that the performance of our method is insensitive to variation of ϵ in the range of [0.4,0.7] and different clustering algorithms, because the clustering results only determine the initialization of the relation discovery objective. The cluster relation matrix is further updated during training.

Margin Parameter m of Triplet Loss. We varied m in $\mathcal{L}_{id}$ in Equation (5) from 0 to 0.9. The results in Table 10 show that our method is not sensitive to the margin parameter.

Weights w_a and w_s of Distillation Loss. We varied w_a from 0 to 2 and set $w_s = 2 - w_a$ in Equation (13) to control the contributions of knowledge learned by R^2D_{aff} and R^2D_{sty}, respectively. As shown in Table 11, the best performance is achieved when $w_a = w_s = 1$, which indicates that using equal contributions for R^2D_{aff} and R^2D_{sty} to distill make them complement each other better.

Batch Size. In our training strategy, a batch consists of N_{BL} labeled samples and N_{BU} unlabeled samples. By default, we set $N_{BL} = 128$ and $N_{BU} = 128$ and the batch size is $N_{BL} + N_{BU} = 256$. We varied N_{BU} from 32 to 256 and varied N_{BL} from 32 to 128 for learning our full model R^2D on REID-CBD. The results are shown in Table 12. When using our default batch size 256 ($N_{BL} = N_{BU} = 128$), the best performance is achieved. Compared with the best performance, the performance degradation on mAP is fewer than 2% when varying the batch size. The results demon-

TABLE 12
Performances (%) of using different batch sizes for our R^{2}D on REID-CBD. N_{BU} is the number of unlabeled samples and N_{BL} is the number of labeled samples.

N_{BU}	32			64			128			256		
N_{BL}	32	64	128	32	64	128	32	64	128	32	64	128
Batch size	64	96	160	96	128	192	160	192	256	288	320	384
R-1	57.2	57.5	58.6	56.7	58.1	58.6	57.7	58.5	59.0	57.1	57.9	58.7
mAP	26.6	27.1	27.9	26.5	27.4	28.0	27.1	27.9	28.2	26.7	27.2	28.0

TABLE 13
Training on subsets with different ratios of identities crossing camera views by R^{2}D$_{\text{aff}}$ (affinity-based clusters), HHL-semi and GLT-semi. Performance metrics are rank-1 (R-1) accuracy and mAP (%).

Ratio of ID crossing camera		0	25%	50%	75%	100%
HHL-semi [14]	R-1	59.2	59.0	59.3	59.7	59.5
	mAP	39.9	40.1	40.2	40.4	40.1
GLT-semi [18]	R-1	50.6	55.3	58.8	61.1	63.6
	mAP	36.3	40.5	43.1	45.2	47.4
R^{2}D$_{\text{aff}}$ (ours)	R-1	61.5	62.7	68.0	69.5	71.1
	mAP	41.4	44.3	48.5	50.1	51.2

strate that our method is robust to batch size when N_{BL} is in [32, 128] and N_{BU} is in [32, 256].

6.5.3 Training on Data with Different Ratios of Underlying Identities Crossing Camera Views

For training data collection in nonadjacent scenes with different distances, the closer the distance between two nonadjacent scenes is, the more possible there exist underlying identities crossing camera views. The visually similar samples that probably contain underlying cross-camera positive pairs can be associated to reduce identity uncertainty by pseudo labeling methods. To simulate different real-world scenarios, we varied the ratios of identities crossing camera views from 0% to 100% on DukeMTMC [21]. We sampled different unlabeled training subsets of the same size by removing samples of identities crossing camera views for fair comparison between different ratios. Two competitive semi-supervised learning methods style-transfer-based HHL-semi [14] and pseudo-label-based GLT-semi [18] were compared with our method "R^{2}D$_{\text{aff}}$" (affinity-based clusters). The results are shown in Table 13.

When varying the ratios of identities crossing camera views in the subsets from 0% to 100%, our method consistently outperformed HHL-semi [14] and GLT-semi [18]. HHL-semi [14] cannot bring about improvement with the ratio increasing. GLT-semi [18] can reduce identity uncertainty more effectively when the ratio is high. However, when the ratio was 0, GLT-semi [18] was worse than the baseline model in Table 7 (Rank-1 53.5% and mAP 35.6%) due to noise accumulation problem [24]. In real-world applications, the ratio is uncertain to be high or low in the unlabeled data captured in unseen scenes. With reward from labeled data to guide relation discovery, our method can stably improve the baseline model with the ratio increasing.

6.5.4 Training with Different Types of Supervision

We evaluated using different types of supervisions, including more labeled identities and additional intra-camera identity labels on modified DukeMTMC-NA. We also evaluated unsupervised domain adaptation.

TABLE 14
Evaluation of using different numbers of labeled identities in training set C_L for our method R^{2}D$_{\text{aff}}$ (affinity-based clusters). The values in the brackets for "R^{2}D$_{\text{aff}}$ (ours)" are the gains compared with the baseline.

C_L		10	30	50	70
Fine-tune on $\mathcal{D}_L$ (baseline)	R-1	53.5	57.6	60.6	62.1
	mAP	35.6	37.8	40.2	42.0
R^{2}D$_{\text{aff}}$ (ours)	R-1	61.5 (+8.0)	66.0 (+8.4)	68.2 (+7.6)	71.1 (+9.0)
	mAP	41.4 (+5.8)	46.7 (+8.9)	49.3 (+9.1)	51.5 (+9.5)

TABLE 15
Evaluation of using additional intra-camera identity labels. The performances are indicated by rank-k (R-k) accuracy and mAP (%).

Methods	R-1	R-5	R-10	mAP
MCNL [88]	63.1	77.4	82.3	44.3
MCNL [88]+ R^{2}D (ours)	**67.1**	**80.4**	**85.1**	**48.3**

The Number of Identities C_L in Labeled Data $\mathcal{D}_L$. When more labeled data can be obtained for training, we compared with the baseline model of "Fine-tune on $\mathcal{D}_L$" to show the effectiveness of our method "R^{2}D$_{\text{aff}}$". We increased the number of labeled identities C_L from 10 to 70 with step size of 20 and evaluated our method on DukeMTMC-NA. The results are shown in Table 14. With more labeled identities, the improvement of our R^{2}D on the baseline model is increasingly significant from 5.8% mAP to 9.5% mAP, since the validation objective with stronger supervision can provide better guidance for discovering the probabilistic cluster relations for the relation discovery objective.

Additional Intra-Camera Supervision. Although identities crossing camera views are rare, intra-camera identity labels can still be annotated, so we considered additional intra-camera supervision for training. We compared with the state-of-the-art intra-camera supervised Re-ID loss MCNL [88] that is specifically designed for cross-camera unpaired data. For fair comparison in the implementation, our identification loss (Eq. (5)) was applied on the cross-camera paired labeled data $\mathcal{D}_L$ in addition to the MCNL loss based on the codes released by the authors. Our method R^{2}D was applied on the model pre-trained by MCNL. As intra-camera supervision is available, we used ImageNet for pre-training instead of Market-1501 [20].

The results on DukeMTMC-NA (with intra-camera supervision) are shown in Table 15. When combining our R^{2}D with MCNL, clear improvement can be achieved, since our method can learn to discover probabilistic cluster relations by the rewarded pseudo label training strategy to quantify and reduce uncertainty of underlying sample relations.

Unsupervised Domain Adaptation. Our method can also be applied to unsupervised domain adaptation (UDA) task. We compared with the state-of-the-art unsupervised Re-ID method UNRN [47] on REID-CBD, DukeMTMC-NA and MSMT17-NA (with rare cross-camera positive pairs) and the original DukeMTMC [21] and MSMT17 [19] (with cross-camera positive pairs for each identity).

To adapt our method for UDA task, we replace the labeled image set with pseudo labeled image set for computing the few-shot validation objective to provide reward for relation discovery objective. Since there exists limited cross-camera positive pairs, we construct the pseudo labeled image set by potential cross-camera positive pairs with high confidence level.

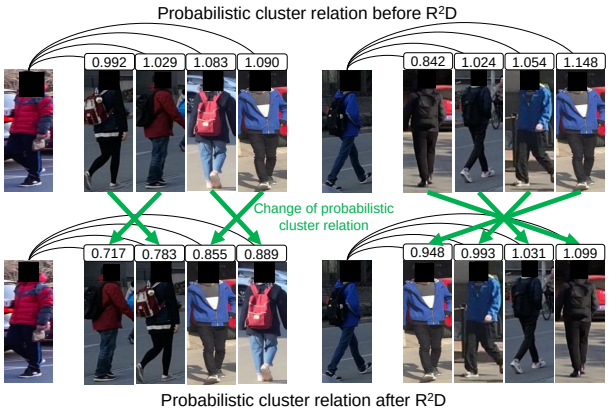

Fig. 7. Visualization of probabilistic relations between randomly selected clusters. Each cluster is represented by an image in it and the value in the rectangle is the Cosine distance between the prototype of the two clusters. After learning cluster relations by R^2D, the feature similarities between different clusters become more consistent with visual similarities of human perception and thus the uncertainty of underlying sample relations are better quantified.

TABLE 16
Comparisons with UNRN [47] in unsupervised domain adaptation (UDA) setting on REID-CBD, DukeMTMC-NA, MSMT17-NA (with rare cross-camera positive pairs) and DukeMTMC, MSMT17 (with cross-camera positive pairs for each identity).

Methods	REID-CBD		Duke-NA		MSMT17-NA		DukeMTMC		MSMT17	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
UNRN [47]	38.3	16.9	53.7	39.5	20.7	8.7	82.0	69.1	52.4	25.3
R^2D (ours)	57.5	26.7	62.6	42.3	38.5	14.6	82.8	70.4	53.3	25.8

The results of comparative evaluations are reported in Table 16. Our method outperforms UNRN [47] on all evaluated datasets. The improvements of our method are especially significant in the cases with rare cross-camera positive pairs on REID-CBD, DukeMTMC-NA and MSMT17-NA, because the uncertainty of sample relation is higher than that on the original DukeMTMC and MSMT17. Our method learns to reduce uncertainty with reward from labeled data; while UNRN [47] reduces uncertainty based on empirical uncertainty estimation principle of soft multilabel agreement, which may cause incorrect estimation without using reward in our method.

6.5.5 Visualization of Learned Probabilistic Relations

To understand the effect of learning cluster relations in rewarded relation discovery (R^2D), we visualize the images of 9 randomly selected clusters and the Cosine distances between the prototypes of corresponding clusters for R^2D_{aff} on MSMT17-NA dataset, as shown in Figure 7. Due to space limitation, more examples of probabilistic relation improvement and degradation in both normal cases and hard cases are shown in the supplementary material.

The cluster relations before and after using R^2D are shown in the first row and the second row, respectively. Intuitively, the distances between different clusters should reflect the degrees of visual differences. For example, the distance between two persons in tops of the same color and bottoms of the same color should be smaller than that between two persons in tops or bottoms of different colors. Before applying R^2D, such rankings of distances that are contradictory to human perception exist in both two groups of images in the first row. After learning the cluster relations by R^2D, as indicated by the arrows, the feature similarities between different clusters change and become more consistent with visual similarities of human perception. Thus, the uncertainty of underlying sample relations are better quantified by the learned probabilistic relations.

TABLE 17
Running time of each step of our method in training and testing on REID-CBD. The style transfer method is StarGAN [79].

Step	R^2D_{aff}	Style transfer [79]	R^2D_{sty}	R^2D_{aff},R^2D_{sty} knowledge fusion	Inference	Retrieval
Time	3.5h	12.0h	4.0h	1.7h	2min	2min

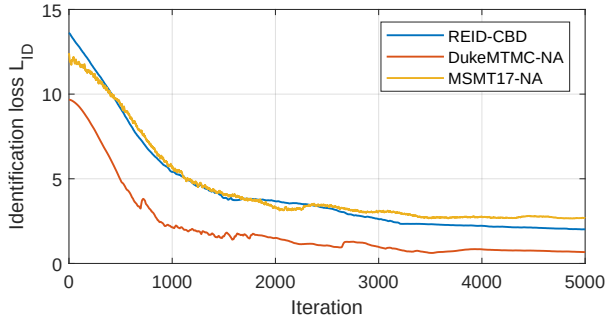

Fig. 8. Variations of the identification loss $\mathcal{L}_{ID}$ along with increasing iterations on REID-CBD, DukeMTMC-NA and MSMT17-NA.

6.5.6 Complexity Analysis and Running Time

We analyze the model size and computation costs for our method R^2D. ResNet-50 [75] was applied as backbone model for our method. The number of parameter is 25.6M and the dimension of feature is 2048. We use floating-point operations per second (FLOPs) to indicate computation cost. The computation cost for a forward pass is 2.7G FLOPs. For training rewarded relation discovery, an iteration contains three forward passes and three backward passes, of which the cost is 24.6G FLOPs. For similarity knowledge distillation, an iteration contains one forward pass and one backward pass, of which the cost is 8.2G FLOPs.

We evaluated the running time of each step of our method in training and testing on REID-CBD. The numbers of samples in training set and testing set are 24,131 and 122,379, respectively. The results are reported in Table 17. The style transfer method StarGAN [79] applied in the style-based cluster construction of our method took longer time than other steps. Since our method does not depend on specific style transfer methods, alternative methods can be applied to improve training efficiency.

6.5.7 Convergence of Loss

In the training phase, we show the variations of the identification loss $\mathcal{L}_{ID}$ along with increasing iterations on REID-CBD, DukeMTMC-NA and MSMT17-NA in Figure 8. Generally, the loss converges after 4000 iterations.

6.5.8 Evaluation on Vehicle Re-Identification

Our method is not limited to applications of person re-identification. We also evaluated on vehicle Re-ID dataset

TABLE 18
Performances (%) of the state-of-the-art methods and our method R^2D for vehicle re-identification in semi-supervised learning setting.

Methods	VeRi-776-NA				VeRi-776			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
SpCL [42]	25.1	37.1	45.0	8.5	68.1	77.8	82.3	29.7
PPLR [90]	26.6	34.8	42.6	9.6	80.3	84.2	85.7	38.5
SpCL-semi [42]	29.3	44.7	53.7	10.7	70.2	79.6	85.1	32.2
PPLR-semi [90]	34.9	44.8	52.6	14.1	82.5	86.5	88.1	39.1
R^2D (ours)	**50.4**	**65.4**	**72.2**	**18.0**	**83.2**	**87.6**	**89.4**	**39.5**

VeRi-776 [89]. Following the processing of simulating nonadjacent scenes for MSMT17-NA [19] in Section 5.2, we processed VeRi-776 to obtain the simulated dataset VeRi-776-NA by selecting samples in one random camera for each class. On both VeRi-776-NA and VeRi-776, we evaluated in semi-supervised setting as the person Re-ID setting in Section 6.1. For training, our model and the compared models were initialized by ImageNet pre-training. We compared with the state-of-the-art vehicle Re-ID methods SpCL [42], PPLR [90] and their semi-supervised versions SpCL-semi, PPLR-semi. The results in Table 18 show that, our method significantly outperformed the compared methods on VeRi-776-NA simulated for nonadjacent scenes and the performance of our method is comparable with the state-of-the-art performance on VeRi-776.

7 Conclusion

In this work, we study semi-supervised person re-identification on training data with identities rarely crossing camera views. Compared with existing Re-ID approaches especially for semi-supervised Re-ID that rely on abundant identities crossing camera views, we operate semi-supervised Re-ID under a relaxed assumption of identities rarely crossing camera views. To overcome the problem of high uncertainty in such cases, we propose Rewarded Relation Discovery (R^2D) to discover the underlying probabilistic relations by a rewarded pseudo label training strategy. In this strategy, we quantify the uncertainty by parameterizing the probabilistic relations in the relation discovery objective for pseudo label training. The reward quantified by the identification performance on limited labeled data is introduced for this objective. By maximizing the reward to learn probabilistic relations parameterized by cluster relation matrix, minimization of the relation discovery objective can reduce the uncertainty of underlying sample relations. Furthermore, we embed prior knowledge of intra-camera affinity and cross-camera style variation in different relation discovery objectives and further fuse the knowledge of different probabilistic relations by similarity knowledge distillation to further reduce the uncertainty of sample relations. Extensive evaluations for semi-supervised Re-ID on identities rarely crossing camera views were carried out on our new public real-world dataset REID-CBD captured in nonadjacent scenes and two simulated datasets DukeMTMC-NA and MSMT17-NA. The results show the effectiveness of our method as compared with a wide range of semi-supervised, unsupervised and self-supervised representation learning methods.

Acknowledgment

This work was supported partially by the National Science Foundation for Young Scientists of China (62106288), the Guangdong Basic and Applied Basic Research Foundation (2023A1515012974), the NSFC (U21A20471, U1911401, U1811461) and Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085).

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 (CVPR)*, Jun. 2014, pp. 152–159.

[2] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3908–3916.

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

[4] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *European Conference on Computer Vision (ECCV)*, Sep. 2018, pp. 480–496.

[5] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3760–3769.

[6] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 3183–3192.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 6, pp. 2872–2893, 2022.

[8] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 3550–3557.

[9] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, and N. Zheng, "Semi-supervised person re-identification using multi-view clustering," *Pattern Recognition*, vol. 88, pp. 285–297, 2019.

[10] J. Li, A. J. Ma, and P. C. Yuen, "Semi-supervised region metric learning for person re-identification," *International Journal of Computer Vision (IJCV)*, vol. 126, no. 8, pp. 855–874, 2018.

[11] 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 (TIP)*, vol. 28, no. 6, pp. 2872–2881, 2019.

[12] X. Chang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Transductive semi-supervised metric learning for person re-identification," *Pattern Recognition*, vol. 108, p. 107569, 2020.

[13] 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 (TOMM)*, vol. 14, no. 4, p. 83, 2018.

[14] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *European Conference on Computer Vision (ECCV)*, Sep. 2018, pp. 172–188.

[15] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 4, pp. 956–973, April 2020.

[16] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2020.

[17] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *International Conference on Learning Representations (ICLR)*, Apr. 2020.

[18] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha, "Group-aware label transfer for domain adaptive person re-identification." in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 5310–5319.

[19] 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 (CVPR)*, June 2018, pp. 79–88.

[20] 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 (ICCV)*, Dec. 2015, pp. 1116–1124.

[21] 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 (ICCV)*, Oct. 2017, pp. 3774–3782.

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

[23] Y. Bai, J. Jiao, W. Ce, J. Liu, Y. Lou, X. Feng, and L.-Y. Duan, "Person30k: A dual-meta generalization network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 2123–2132.

[24] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8.

[25] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[26] 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 (CVPR)*, Jun. 2015, pp. 2197–2206.

[27] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 3, pp. 653–668, 2013.

[28] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, October 2021, pp. 15 013–15 022.

[29] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Feature completion for occluded person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 9, pp. 4894–4912, 2022.

[30] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and C. X., "Iaunet: Global context-aware feature learning for person reidentification," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, vol. 32, no. 10, pp. 4460–4474, 2020.

[31] X. Wang, S. Li, M. Liu, Y. Wang, and A. K. Roy-Chowdhury, "Multi-expert adversarial attack detection in person re-identification using context inconsistency," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 077–15 087.

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

[33] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised $\ell 1$ graph learning," in *European Conference on Computer Vision (ECCV)*, Oct. 2016, pp. 178–195.

[34] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2021, pp. 2764–2772.

[35] Y. Zheng, S. Tang, G. Teng, Y. Ge, K. Liu, J. Qin, D. Qi, and D. Chen, "Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, October 2021, pp. 8371–8381.

[36] X. Wang, R. Panda, M. Liu, Y. Wang, and A. K. Roy-Chowdhury, "Exploiting global camera network constraints for unsupervised video person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 31, no. 10, pp. 4020–4030, 2020.

[37] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 594–611.

[38] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning generalisable omni-scale representations for person re-identification." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2021.

[39] Y. Li, H. Yao, L. Duan, H. Yao, and C. Xu, "Adaptive feature fusion via graph neural network for person re-identification," in *ACM International Conference on Multimedia (ACM MM)*, Oct. 2019, pp. 2115–2123.

[40] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 1187–1196.

[41] G. Hao, Y. Yang, X. Zhou, G. Wang, and Z. Lei, "Horizontal flipping assisted disentangled feature learning for semi-supervised person re-identification," in *Asian Conference on Computer Vision (ACCV)*, Nov. 2020, pp. 21–37.

[42] Y. Ge, F. Zhu, D. Chen, R. Zhao, and hongsheng Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Dec. 2020, pp. 11 309–11 321.

[43] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 7919–7929.

[44] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 2148–2157.

[45] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 30, no. 9, pp. 2815–2829, 2020.

[46] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems (NeurIPS)*, vol. 30, 2017.

[47] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2021.

[48] 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 (CVPR)*, 2019, pp. 2138–2147.

[49] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen, "Unsupervised pre-training for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 14 750–14 759.

[50] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[51] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *International Conference on Machine Learning (ICML)*, vol. 3, no. 2, Jun. 2013, p. 896.

[52] Q. Wang, W. Li, and L. Van Gool, "Semi-supervised learning by augmented distribution alignment," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1466–1475.

[53] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2019, p. 5049–5059.

[54] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of annual conference on Computational learning theory*, 1998, pp. 92–100.

[55] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *International Conference on Machine Learning workshop (ICML)*, vol. 3, 2003.

[56] J. Li, C. Xiong, and S. C. Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 9475–9484.

[57] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2021.

[58] Y. Y. Grandvalet, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems (NeurIPS)*, p. 529–536, Dec. 2005.

[59] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *International conference on Machine learning (ICML)*, 2005, pp. 824–831.

[60] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, "The manifold tangent classifier," in *Advances in neural information processing systems (NeurIPS)*, Dec. 2011, pp. 2294–2302.

[61] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NeurIPS)*, Dec. 2015, p. 3546–3554.

[62] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," in *Advances in neural information processing systems (NeurIPS)*, vol. 4, Dec. 2014, p. 3365–3373.

[63] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems (NeurIPS)*, Dec. 2017, p. 1195–1204.

[64] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 8, pp. 1979–1993, 2019.

[65] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, Dec. 2020, pp. 22 243–22 255.

[66] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," *Machine Learning*, vol. 109, no. 4, pp. 719–760, 2020.

[67] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *International Conference on Machine Learning (ICML)*, Jul. 2015, pp. 2113–2122.

[68] Y. Lee and S. Choi, "Gradient-based meta-learning with learned layerwise metric and subspace." in *International Conference on Machine Learning (ICML)*, Jul. 2018, pp. 2927–2936.

[69] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *International Conference on Learning Representations (ICLR)*, Oct. 2018.

[70] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *International Conference on Learning Representations (ICLR)*, May 2016.

[71] G. Algan and I. Ulusoy, "Meta soft label generation for noisy labels," in *International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 7142–7148.

[72] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International Conference on Machine Learning (ICML)*, Jul. 2018, pp. 4331–4340.

[73] H. Pham, Q. Xie, Z. Dai, and Q. V. Le, "Meta pseudo labels," *arXiv preprint arXiv:2003.10580*, 2020.

[74] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research (JMLR)*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[76] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 10, pp. 2597–2609, 2020.

[77] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in *International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

[78] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognition*, vol. 102, p. 107173, 2020.

[79] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 8789–8797.

[80] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 2, pp. 386–397, 2020.

[81] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *International Conference on Computational Statistics (COMPSTAT)*, Aug. 2010, pp. 177–186.

[82] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research (JMLR)*, vol. 13, no. Mar, pp. 723–773, 2012.

[83] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision Workshop (ECCVW)*, Aug. 2016, pp. 443–450.

[84] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 15 750–15 758.

[85] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 1, pp. 281–297, 1967.

[86] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 5, pp. 603–619, 2002.

[87] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[88] T. Zhang, L. Xie, L. Wei, Y. Zhang, B. Li, and Q. Tian, "Single camera training for person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 7, Feb. 2020, pp. 12 878–12 885.

[89] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *IEEE international conference on multimedia and expo (ICME)*. IEEE, 2016, pp. 1–6.

[90] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7308–7318.

Ancong Wu received the bachelor's degree in intelligence science and technology from Sun Yat-sen University in 2015 and received Ph.D. degree in information and communication engineering from Sun Yat-sen University in 2020. He is now a Postdoc in Sun Yat-sen University. His research interest is computer vision algorithms and the applications for person re-identification.

Wenhang Ge is now a Master student in Sun Yat-sen University. He received the bachelor's degree in school of Intelligent engineering from Sun Yat-sen University in 2020. His research interest are deep learning algorithms and computer vision tasks, such as the applications for person re-identification, self supervised learning and knowledge distillation.

Wei-Shi Zheng is now a full Professor with Sun Yat-sen University. Dr. Zheng received his Ph.D. degree in Applied Mathematics from Sun Yat-sen University in 2008. His research interests include person/object association and activity understanding in visual surveillance, and the related machine learning algorithm such as continuous learning and weakly supervised learning. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He has ever served as area chairs of CVPR, ICCV, BMVC and NeurIPS. He was a technical program chair of ICME 2022. He is an associate editor of IEEE TPAMI and the Pattern Recognition Journal. He is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of the Royal Society-Newton Advanced Fellowship of the United Kingdom.