# Unsupervised Person Re-identification by Soft Multilabel Learning

Hong-Xing Yu[1] , Wei-Shi Zheng[1,4*] ,
Ancong Wu[1] , Xiaowei Guo[2] , Shaogang Gong[3] , and Jian-Huang Lai[1]

[1]Sun Yat-sen University, China
[2]YouTu Lab, Tencent      [3]Queen Mary University of London, UK
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
xKoven@gmail.com, wszheng@ieee.org, wuancong@mail2.sysu.edu.cn, scorpioguo@tencent.com,
s.gong@qmul.ac.uk, stsljh@mail.sysu.edu.cn

## Abstract

*Although unsupervised person re-identification (RE-ID) has drawn increasing research attentions due to its potential to address the scalability problem of supervised RE-ID models, it is very challenging to learn discriminative information in the absence of pairwise labels across disjoint camera views. To overcome this problem, we propose a deep model for the soft multilabel learning for unsupervised RE-ID. The idea is to learn a soft multilabel (real-valued label likelihood vector) for each unlabeled person by comparing the unlabeled person with a set of known* reference *persons from an auxiliary domain. We propose the soft multilabel-guided hard negative mining to learn a discriminative embedding for the unlabeled target domain by exploring the* similarity consistency *of the visual features and the soft multilabels of unlabeled target pairs. Since most target pairs are cross-view pairs, we develop the cross-view consistent soft multilabel learning to achieve the learning goal that the soft multilabels are consistently good across different camera views. To enable effecient soft multilabel learning, we introduce the reference agent learning to represent each reference person by a reference agent in a joint embedding. We evaluate our unified deep model on Market-1501 and DukeMTMC-reID. Our model outperforms the state-of-the-art unsupervised RE-ID methods by clear margins. Code is available at* `https://github.com/KovenYu/MAR`*.*

## 1. Introduction

Existing person re-identification (RE-ID) works mostly focus on supervised learning [17, 20, 58, 63, 1, 45, 51, 43, 38, 41, 33]. However, they need substantial pairwise labeled data across every pair of camera views, limiting the scalability to large-scale applications where only unlabeled data
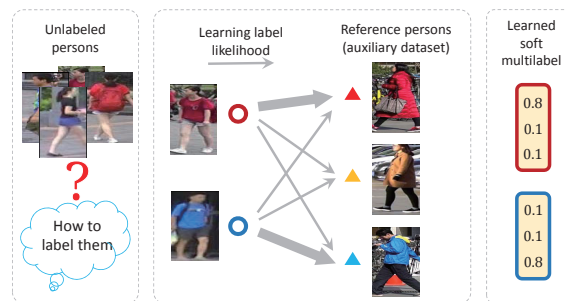
_____
*Corresponding author



Figure 1. Illustration of our soft multilabel concept. We learn a soft multilabel (real-valued label vector) for each unlabeled person by comparing to a set of known auxiliary reference persons (thicker arrowline indicates higher label likelihood). Best viewed in color.

is available due to the prohibitive manual efforts in exhaustively labeling the pairwise RE-ID data [49]. To address the scalability problem, some recent works focus on unsupervised RE-ID by clustering on the target unlabelled data [52, 53, 8] or transfering the knowledge from other labeled source dataset [29, 48, 7, 62]. However, the performance is still not satisfactory. The main reason is that, without the pairwise label as learning guidance, it is very challenging to discover the identity discriminative information due to the drastic cross-view intra-person appearance variation [52] and the high inter-person appearance similarity [60].

To address the problem of lacking pairwise label guidance in unsupervised RE-ID, in this work we propose a novel soft multilabel learning to mine the potential label information in the unlabeled RE-ID data. The main idea is, for every unlabeled person image in an unlabeled RE-ID dataset, we learn a soft multilabel (i.e. a real-valued label likelihood vector instead of a single pseudo label) by comparing this unlabeled person with a set of *reference persons* from an existing labeled auxiliary source dataset. Figure 1 illustrates this soft multilabel concept.

Based on this soft multilabel learning concept, we propose to mine the potential discriminative information by the

*soft multilabel-guided hard negative mining*, i.e. we leverage the soft multilabel to distinguish the visually similar but different unlabeled persons. In essence, the soft multilabel represents the unlabelled target person by the reference persons, and thus it encodes the *relative* comparative characteristic of the unlabeled person, which is from a different perspective than the *absolute* visual feature representation. Intuitively, a pair of images of the same person should be not only visually similar to each other (i.e. they should have similar absolute visual features), but also equally similar to any other reference person (i.e. they should also have similar relative comparative characteristics with respect to the reference persons). If this *similarity consistency* between the absolute visual representation and the relative soft multilabel representation is violated, i.e. the pair of images are visually similar but their comparative characteristics are dissimilar, it is probably a hard negative pair.

In the RE-ID context, most image pairs are *cross-view* pairs which consist of two person images captured by different camera views. Therefore, we propose to learn the soft multilabels that are consistently good across different camera views. We refer to this learning as the *cross-view consistent soft multilabel learning*. To enable the efficient soft multilabel learning which requires comparison between the unlabeled persons and the reference persons, we introduce the *reference agent learning* to represent each reference person by a reference agent which resides in a joint feature embedding with the unlabeled persons. Specifically, we develop a unified deep model named deep soft **m**ultil**a**bel **r**eference learning (MAR) which jointly formulates the soft multilabel-guided hard negative mining, the cross-view consistent soft multilabel learning and the reference agent learning.

We summarize our **contributions** as follows: (1). We address the unsupervised RE-ID problem by a novel soft multilabel reference learning method, in which we mine the potential label information *latent* in the unlabeled RE-ID data by exploiting the auxiliary source dataset for reference comparison. (2). We formulate a novel deep model named deep soft **m**ultil**a**bel **r**eference learning (MAR). MAR enables simultaneously the soft multilabel-guided hard negative mining, the cross-view consistent soft multilabel learning and the reference agent learning in a unified model. Experimental results on Market-1501 and DukeMTMC-reID show that our model outperforms the state-of-the-art unsupervised RE-ID methods by significant margins.

## 2. Related Work

**Unsupervised RE-ID**. Unsupervised RE-ID refers to that the target dataset is unlabelled but the auxiliary source dataset is not necessarily unlabelled [29, 8, 48, 7, 62]. Existing methods either transfer source label knowledge [29, 8, 48, 7, 62] or assuming strong prior knowledge (i.e. either assuming the target RE-ID data has specific cluster

structure [52, 53, 8] or assuming the hand-crafted features could be discriminative enough [52, 16, 15, 57, 9, 46]). Recently attempts have been made on exploiting video tracklet associations for unsupervised RE-ID [5, 19]. Another line of work focusing on reducing the labelling effort is to minimize the labelling budget on the target [32] which is complementary to the unsupervised RE-ID. The most related works are the clustering-based models [52, 53, 8], e.g. Yu et.al. [52, 53] proposed an asymmetric metric clustering to discover labels latent in the unlabelled target RE-ID data. The main difference is that the soft multilabel could leverage the auxiliary reference information other than visual feature similarity, while the pseudo label only encodes the feature similarity of an unlabelled pair. Hence, the soft multilabel could mine the potential label information that cannot be discovered by directly comparing the visual features.

Some unsupervised RE-ID works also proposed to use the labeled source dataset by the unsupervised domain adaptation [50, 7, 62, 48] to transfer the discriminative knowledge from the auxiliary source domain. Our model is different from them in that these models do not mine the discriminative information in the unlabeled target domain, which is very important because the transferred discriminative knowledge might be less effective in the target domain due to the domain shift [28] in discriminative visual clues.

**Unsupervised domain adaptation**. Our work is also closely related to the unsupervised domain adaptation (UDA) [23, 39, 40, 36, 25, 42, 10, 24], which also has a source dataset and an unlabeled target dataset. However, they are mostly based on the assumption that the classes are the same between both domains [3, 28, 6, 27], which does not hold in the RE-ID context where the persons (classes) in the source dataset are completely different from the persons in the target dataset, rendering these UDA models inapplicable to the unsupervised RE-ID [50, 7, 62, 48].

**Multilabel classification**. Our soft multilabel learning is conceptually different from the multilabel classification [54]. The multilabel in the multilabel classification [54] is a groundtruth binary vector indicating whether an instance *belongs to* a set of classes, while our soft multilabel is learned to represent an unlabeled target person by *other different* reference persons. Hence, existing multilabel classification models are for a different purpose and thus not suitable to model our idea.

**Zero-shot learning**. Zero-shot learning (ZSL) aims to recognize novel testing classes specified by semantic attributes but unseen during training [18, 31, 55, 14, 56]. Our soft multilabel reference learning is related to ZSL in that every unknown target person (unseen testing class) is represented by a set of known reference persons (attributes of training classes). However, the predefined semantic attributes are not available in unsupervised RE-ID. Nevertheless, the success

of ZSL models validates/justifies the effectiveness of representing an unknown class (person) with a set of different classes. A recent work also explores a similar idea by representing an unknown testing person in an ID regression space which is formed by the known training persons [47], but it requires substantial labeled persons from the target domain.

# 3. Deep Soft Multilabel Reference Learning

## 3.1. Problem formulation and Overview

We have an unlabeled target RE-ID dataset $\mathcal{X} = \{x_i\}_{i=1}^{N_u}$ where each $x_i$ is an unlabeled person image collected in the target visual surveillance scenario, and an auxiliary RE-ID dataset $\mathcal{Z} = \{z_i, w_i\}_{i=1}^{N_a}$ where each $z_i$ is a person image with its label $w_i = 1, \cdots, N_p$ where $N_p$ is the number of the reference persons. Note that the reference population is completely non-overlapping with the unlabeled target population since it is collected from a different surveillance scenario [50, 7, 62]. Our goal is to learn a soft multilabel function $l(\cdot)$ such that $y = l(x, \mathcal{Z}) \in (0, 1)^{N_p}$ where all dimensions add up to 1 and each dimension represents the label likelihood w.r.t. a reference person. Simultaneously, we aim to learn a discriminative deep feature embdding $f(\cdot)$ under the guidance of the soft multilabels for the RE-ID task. Specifically, we propose to leverage the soft multilabel for hard negative mining, i.e. for visually similar pairs we determine they are positive or hard negative by comparing their soft multilabels. We refer to this part as the *Soft multilabel-guided hard negative mining* (Sec. 3.2). In the RE-ID context, most pairs are *cross-view* pairs which consist of two person images captured by different camera views. Therefore, we aim to learn the soft multilabels that are consistently good across different camera views so that the soft multilabels of the cross-view images are comparable. We refer to this part as the *Cross-view consistent soft multilabel learning* (Sec. 3.3). To effeciently compare each unlabeled person $x$ to all the reference persons, we introduce the *reference agent learning* (Sec. 3.4), i.e. we learn a set of *reference agents* $\{a_i\}_{i=1}^{N_p}$ each of which represents a reference person in the shared *joint feature embedding* where both the unlabeled person $f(x)$ and the agents $\{a_i\}_{i=1}^{N_p}$ reside (so that they are comparable). Therefore, we could learn the soft multilabel $y$ for $x$ by comparing $f(x)$ with the reference agents $\{a_i\}_{i=1}^{N_p}$, i.e. the soft multilabel function is simplified to $y = l(f(x), \{a_i\}_{i=1}^{N_p})$.

We show an overall illustration of our model in Figure 2. In the following, we introduce our deep soft **m**ultil**a**bel **r**eference learning (MAR). We first introduce the soft multilabel-guided hard negative mining given the reference agents $\{a_i\}_{i=1}^{N_p}$ and the reference comparability between $f(x)$ and $\{a_i\}_{i=1}^{N_p}$. To facilitate learning the joint embedding, we enforce a unit norm constraint, i.e.

$||f(\cdot)||_2 = 1, ||a_i||_2 = 1, \forall i$, to learn a hypersphere embedding [44, 22]. Note that in the hypersphere embedding, the cosine similarity between a pair of features $f(x_i)$ and $f(x_j)$ is simplified to their inner product $f(x_i)^{\mathrm{T}} f(x_j)$, and so as for the reference agents.

## 3.2. Soft multilabel-guided hard negative mining

Let us start by defining the soft multilabel function. Since each entry/dimension of the soft multilabel $y$ represents the label likelihood that adds up to 1, we define our soft multilabel function as

$$y^{(k)} = l(f(x), \{a_i\}_{i=1}^{N_p})^{(k)} = \frac{\exp(a_k^{\mathrm{T}} f(x))}{\Sigma_i \exp(a_i^{\mathrm{T}} f(x))} \quad (1)$$

where $y^{(k)}$ is the $k$-th entry of $y$.

It has been shown extensively that mining hard negatives is more important in learning a discriminative embedding than naively learning from all visual samples [13, 37, 26, 35, 34]. We explore a soft multilabel-guided hard negative mining, which focuses on the pairs of visually similar but different persons and aims to distinguish them with the guidance of their soft multilabels. Given that the soft multilabel encodes relative comparative characteristics, we explore a representation consistency: Besides the similar *absolute* visual features, images of the same person should also have similar *relative* comparative characteristics (i.e. equally similar to any other reference person). Specifically, we make the following assumption in our model formulation:

**Assumption 1.** *If a pair of unlabeled person images $x_i, x_j$ has high feature similarity $f(x_i)^{\mathrm{T}} f(x_j)$, we call the pair a* similar pair. *If a similar pair has highly similar comparative characteristics, it is probably a positive pair. Otherwise, it is probably a hard negative pair.*

For the similarity measure of the comparative characteristics encoded in the pair of soft multilabels, we propose the *soft multilabel agreement $A(\cdot, \cdot)$*, defined by:

$$A(y_i, y_j) = y_i \wedge y_j = \Sigma_k \min(y_i^{(k)}, y_j^{(k)}) = 1 - \frac{||y_i - y_j||_1}{2}, \quad (2)$$

which is based on the well-defined L1 distance. Intuitively, the soft multilabel agreement is an analog to the voting by the reference persons: Every reference person $k$ gives his/her conservative agreement $\min(y_i^{(k)}, y_j^{(k)})$ on believing the pair to be positive (the more similar/related a reference person is to the unlabeled pair, the more important is his/her word), and the soft multilabel agreement is cumulated from all the reference persons. The soft multilabel agreement is defined based on L1 distance to treat fairly the agreement of every reference person by taking the absolute value.

Now, we mine the hard negative pairs by considering both the feature similarity and soft multilabel agreement according to Assumption 1. We formulate the soft multilabel-guided hard negative mining with a mining ratio $p$: We define the similar pairs in Assumption 1 as the $pM$ pairs
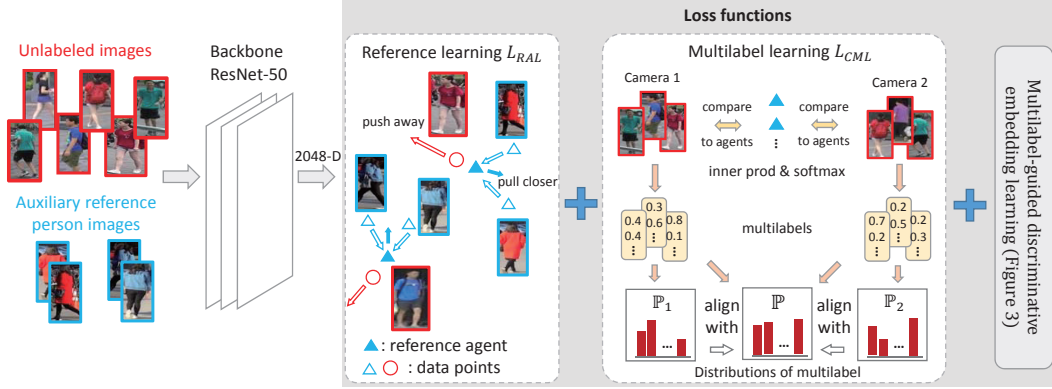
Figure 2. An illustration of our model MAR. We learn the soft multilabel by comparing each target unlabeled person image $f(x)$ (red circle) to a set of auxiliary reference persons represented by a set of reference agents $\{a_i\}$ (blue triangles, learnable parameters) in the feature embedding. The soft multilabel judges whether a similar pair is positive or hard negative for discriminative embedding learning (Sec. 3.2). The soft multilabel learning and the reference learning are elaborated in Sec. 3.3 and Sec. 3.4, respectively. Best viewed in color.

that have highest feature similarities among all the $M = N_u \times (N_u - 1)/2$ pairs within the unlabeled target dataset $\mathcal{X}$. For a similar pair $(x_i, x_j)$, if it is also among the top $pM$ pairs that have the highest soft multilabel agreements, we assign $(i, j)$ to the positive set $\mathcal{P}$, otherwise we assign it to the hard negative set $\mathcal{N}$ (see Figure 3). Formally, we construct:

$$\mathcal{P} = \{(i,j)|f(x_i)^{\mathrm{T}}f(x_j) \geq S, A(y_i, y_j) \geq T\}$$
$$\mathcal{N} = \{(k,l)|f(x_k)^{\mathrm{T}}f(x_l) \geq S, A(y_k, y_l) < T\} \quad (3)$$

where $S$ is the cosine similarity (inner product) of the $pM$-th pair after sorting all $M$ pairs in an descending order according to the feature similarity (i.e. $S$ is a similarity threshold), and $T$ is the similarly defined threshold value for the soft multilabel agreement. Then we formulate the soft Multilabel-guided Discriminative embedding Learning by:

$$L_{MDL} = -\log \frac{\overline{P}}{\overline{P} + \overline{N}}, \quad (4)$$

where

$$\overline{P} = \frac{1}{|\mathcal{P}|}\Sigma_{(i,j)\in\mathcal{P}} \exp(-||f(z_i) - f(z_j)||_2^2),$$
$$\overline{N} = \frac{1}{|\mathcal{N}|}\Sigma_{(k,l)\in\mathcal{N}} \exp(-||f(z_k) - f(z_l)||_2^2).$$

By minimizing $L_{MDL}$, we are learning a discriminative feature embedding using the mined positive/hard negative pairs. Note that the construction of $\mathcal{P}$ and $\mathcal{N}$ is dynamic during model training, and we construct them within every batch with the up-to-date feature embedding during model learning (in this case, we simply replace $M$ by $M_{batch} = N_{batch} \times (N_{batch} - 1)/2$ where $N_{batch}$ is the number of unlabeled images in a mini-batch).

### 3.3. Cross-view consistent soft multilabel learning

Given the soft multilabel-guided hard negative mining, we notice that most pairs in the RE-ID problem context are
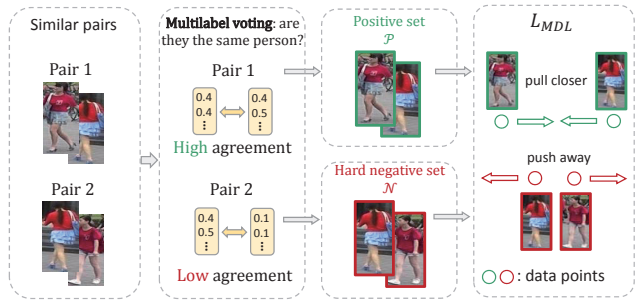


Figure 3. Illustration of the soft multilabel-guided hard negative mining. Best viewed in color.

the *cross-view* pairs which consist of two person images captured by different camera views [52]. Therefore, the soft multilabel should be consistently good across different camera views to be cross-view comparable. From a distributional perspective, given the reference persons and the unlabeled target dataset $\mathcal{X}$ which is collected in a given target domain, the distribution of the *comparative characteristic* should *only* depend on the distribution of the person appearance in the target domain and be independent of its camera views. For example, if the target domain is a cold open-air market where customers tend to wear dark clothes, the soft multilabels should have higher label likelihood in the entries which are corresponding to those reference persons who also wear dark, no matter in which target camera view. In other words, the distribution of the soft multilabel in every camera view should be consistent with the distribution of the target domain. Based on the above analysis, we introduce a Cross-view consistent soft Multilabel Learning loss[1]:

$$L_{CML} = \Sigma_v d(\mathbb{P}_v(y), \mathbb{P}(y))^2 \quad (5)$$

where $\mathbb{P}(y)$ is the soft multilabel distribution in the dataset $\mathcal{X}$, $\mathbb{P}_v(y)$ is the soft multilabel distribution in the $v$-th camera

---

[1] For conciseness we omit all the averaging divisions for the outer summations in our losses.

view in $\mathcal{X}$, and $d(\cdot, \cdot)$ is the distance between two distributions. We could use any distributional distance, e.g. the KL divergence [11] and the Wasserstein distance [2]. Since we empirically observe that the soft multilabel approximately follows a log-normal distribution, in this work we adopt the simplified 2-Wasserstein distance [4, 12] which gives a very simple form (please refer to the supplementary material for the observations of the log-normal distribution and the derivation of the simplified 2-Wasserstein distance):

$$L_{CML} = \Sigma_v ||\mu_v - \mu||_2^2 + ||\sigma_v - \sigma||_2^2 \qquad (6)$$

where $\mu/\sigma$ is the mean/std vector of the log-soft multilabels, $\mu_v/\sigma_v$ is the mean/std vector of the log-soft multilabels in the $v$-th camera view. The form of $L_{CML}$ in Eq. (6) is computationally cheap and easy-to-compute within a batch. We note that the camera view label is naturally available in the unsupervised RE-ID setting [52, 62], i.e. it is typically known from which camera an image is captured.

## 3.4. Reference agent learning

A reference agent serves to represent a unique reference person in the feature embedding like a compact "feature summarizer". Therefore, the reference agents should be mutually discriminated from each other while each of them should be representative of all the corresponding person images. Considering that the reference agents are compared within the soft multilabel function $l(\cdot)$, we formulate the Agent Learning loss as:

$$L_{AL} = \Sigma_k - \log l(f(z_k), \{a_i\})^{(w_k)} = \Sigma_k - \log \frac{\exp(a_{w_k}^{\mathrm{T}} f(z_k))}{\Sigma_j \exp(a_j^{\mathrm{T}} f(z_k))} \qquad (7)$$

where $z_k$ is the $k$-th person image in the auxiliary dataset with its label $w_k$.

By minimizing $L_{AL}$, we not only learn discriminatively the reference agents, but also endow the feature embedding with basic discriminative power for the soft multilabel-guided hard negative mining. Moreover, it reinforces implicitly the validity of the soft multilabel function $l(\cdot)$. Specifically, in the above $L_{AL}$, the soft multilabel function learns to assign a reference person image $f(z_k)$ with a soft multilabel $\hat{y}_k = l(f(z_k), \{a_i\}_{i=1}^{N_p})$ by comparing $f(z_k)$ to all agents, with the learning goal that $\hat{y}_k$ should have minimal cross-entropy with (i.e. similar enough to) the ideal one-hot label $\hat{w}_k = [0, \cdots, 0, 1, 0, \cdots, 0]$ which could produce the *ideal soft multilabel agreement*, i.e. $A(\hat{w}_i, \hat{w}_j) = 1$ if $z_i$ and $z_j$ are the same person and $A(\hat{w}_i, \hat{w}_j) = 0$ otherwise. However, this $L_{AL}$ is minimized for the auxiliary dataset. To further improve the validity of the soft multilabel function for the unlabeled target dataset (i.e. the *reference comparability* between $f(x)$ and $\{a_i\}$), we propose to learn a joint embedding as follows.

**Joint embedding learning for reference comparability**. A major challenge in achieving the reference comparability



Market-1501    DukeMTMC-reID    MSMT17 (Auxiliary dataset)

Figure 4. Dataset examples.

is the domain shift [28], which is caused by different person appearance distributions between the two independent domains. To address this challenge, we propose to mine the *cross-domain* hard negative pairs (i.e. the pair consisting of an unlabeled person $f(x)$ and an auxiliary reference person $a_i$) to rectify the cross-domain distributional misalignment. Intuitively, for each reference person $a_i$, we search for the unlabeled persons $f(x)$ that are visually similar to $a_i$. For a joint feature embedding where the discriminative distributions are well aligned, $a_i$ and $f(x)$ should be discriminative enough to each other despite their high visual similarity. Based on the above discussion, we propose the Reference agent-based Joint embedding learning loss[2]:

$$L_{RJ} = \Sigma_i \Sigma_{j \in \mathcal{M}_i} \Sigma_{k \in \mathcal{W}_i} [m - ||a_i - f(x_j)||_2^2]_+ + ||a_i - f(z_k)||_2^2 \qquad (8)$$

where $\mathcal{M}_i = \{j| ||a_i - f(x_j)||_2^2 < m\}$ denotes the mined data associated with the $i$-th agent $a_i$, $m = 1$ is the agent-based margin which has been theoretically justified in [44] with its recommaned value 1, $[\cdot]_+$ is the hinge function, and $\mathcal{W}_i = \{k|w_k = i\}$. The center-pulling term $||a_i - f(z_k)||_2^2$ reinforces the representativeness of the reference agents to improve the validity that $a_i$ represents a reference person in the cross-domain pairs $(a_i, f(x_j))$.

We formulate the Reference Agent Learning by:

$$L_{RAL} = L_{AL} + \beta L_{RJ} \qquad (9)$$

where $\beta$ balances the loss magnitudes.

## 3.5. Model training and testing

To summarize, the loss objective of our deep soft **m**ulti**la**bel **r**eference learning (MAR) is formulated by:

$$L_{MAR} = L_{MDL} + \lambda_1 L_{CML} + \lambda_2 L_{RAL} \qquad (10)$$

where $\lambda_1$ and $\lambda_2$ are hyperparameters to control the relative importance of the cross-view consistent soft multilabel learning and the reference agent learning, respectively. We train our model end to end by the Stochastic Gradient Descent (SGD). For testing, we compute the cosine feature similarity of each probe(query)-gallery pair, and obtain the ranking list of the probe image against the gallery images.

# 4. Experiments

## 4.1. Datasets

**Evaluation benchmarks**. We evaluate our model in two widely used large RE-ID benchmarks Market-1501 [59] and

---

[2]For brevity we omit the negative auxiliary term (i.e. $w_k \neq i$) which is for a balanced learning in both domains, as our focus is to rectify the cross-domain distribution misalignment.

DukeMTMC-reID [61, 30]. The Market-1501 dataset has 32,668 person images of 1,501 identities. There are in total 6 camera views. The Duke dataset has 36,411 person images of 1,404 identities. There are in total 8 camera views. We show example images in Figure 4. We follow the standard protocol [59, 61] where the training set contains half of the identities, and the testing set contains the other half. We do not use any label of the target datasets during training. The evaluation metrics are the Rank-1/Rank-5 matching accuracy and the mean average precision (MAP) [59].

**Auxiliary dataset**. Essentially the soft multilabel represents an unlabeled person by a set of reference persons, and therefore a high appearance diversity of the reference population would enhance the validity and capacity of the soft multilabel. Hence, we adopt the MSMT17 [50] RE-ID dataset as the auxiliary dataset, which has more identities (i.e. 4,101 identities) than any other RE-ID dataset and which is collected along several days instead of a single day (different weathers could lead to different dressing styles). There are in total 126,441 person images in the MSMT17 dataset. Adopting the MSMT17 as auxiliary dataset enables us to evaluate how various numbers of reference persons (including when there are only a small number of reference persons) affect our model learning in Sec. 4.6.

### 4.2. Implementation details

We set batch size $B = 368$, half of which randomly samples unlabeled images $x$ and the other half randomly samples $z$. Since optimizing entropy-based loss $L_{AL}$ with the unit norm constraint has convergence issue [44, 37], we follow the training method in [44], i.e. we first pretrain the network using only $L_{AL}$ (without enforcing the unit norm constraint) to endow the basic discriminative power with the embedding and to determine the directions of the reference agents in the hypersphere embedding [44], then we enforce the constraint to start our model learning and multiply the constrained inner products by the average inner product value in the pretraining. We set $\lambda_1 = 0.0002$ which controls the relative importance of soft multilabel learning and $\lambda_2 = 50$ which controls the relative importance of agent reference learning. We show an evaluation on $\lambda_1$ and $\lambda_2$ in Sec. 4.6. We set the mining ratio $p$ to 5‰and set $\beta = 0.2$. Training is on four Titan X GPUs and the total time is about 10 hours. We leave the evaluations on $p/\beta$ and further details in the supplementary material due to space limitation.

### 4.3. Comparison to the state of the art

We compare our model with the state-of-the-art unsupervised RE-ID models including: **(1)** the *hand-crafted feature representation* based models LOMO [20], BoW [59], DIC [16], ISR [21] and UDML [29]; **(2)** the *pseudo label learning* based models CAMEL [52], DECAMEL [53] and PUL [8]; and **(3)** the *unsupervised domain adaptation* based mod-

Table 1. Comparison to the state-of-the-art unsupervised results in the Market-1501 dataset. **Red** indicates the best and **Blue** the second best. Measured by %.

| Methods | Reference | Market-1501 | | |
|---|---|---|---|---|
| | | rank-1 | rank-5 | mAP |
| LOMO [20] | CVPR'15 | 27.2 | 41.6 | 8.0 |
| BoW [59] | ICCV'15 | 35.8 | 52.4 | 14.8 |
| DIC [16] | BMVC'15 | 50.2 | 68.8 | 22.7 |
| ISR [21] | TPAMI'15 | 40.3 | 62.2 | 14.3 |
| UDML [29] | CVPR'16 | 34.5 | 52.6 | 12.4 |
| CAMEL [52] | ICCV'17 | 54.5 | 73.1 | 26.3 |
| PUL [8] | ToMM'18 | 45.5 | 60.7 | 20.5 |
| TJ-AIDL [48] | CVPR'18 | 58.2 | 74.8 | 26.5 |
| PTGAN [50] | CVPR'18 | 38.6 | 57.3 | 15.7 |
| SPGAN [7] | CVPR'18 | 51.5 | 70.1 | 27.1 |
| HHL [62] | ECCV'18 | **62.2** | **78.8** | 31.4 |
| DECAMEL [53] | TPAMI'19 | 60.2 | 76.0 | **32.4** |
| MAR | This work | **67.7** | **81.9** | **40.0** |

Table 2. Comparison to the state-of-the-art unsupervised results in the DukeMTMC-reID dataset. Measured by %.

| Methods | Reference | DukeMTMC-reID | | |
|---|---|---|---|---|
| | | rank-1 | rank-5 | mAP |
| LOMO [20] | CVPR'15 | 12.3 | 21.3 | 4.8 |
| BoW [59] | ICCV'15 | 17.1 | 28.8 | 8.3 |
| UDML [29] | CVPR'16 | 18.5 | 31.4 | 7.3 |
| CAMEL [52] | ICCV'17 | 40.3 | 57.6 | 19.8 |
| PUL [8] | ToMM'18 | 30.0 | 43.4 | 16.4 |
| TJ-AIDL [48] | CVPR'18 | 44.3 | 59.6 | 23.0 |
| PTGAN [50] | CVPR'18 | 27.4 | 43.6 | 13.5 |
| SPGAN [7] | CVPR'18 | 41.1 | 56.6 | 22.3 |
| HHL [62] | ECCV'18 | **46.9** | **61.0** | **27.2** |
| MAR | This work | **67.1** | **79.8** | **48.0** |

els TJ-AIDL [48], PTGAN [50], SPGAN [7] and HHL [62]. We show the results in Table 1 and Table 2.

From Table 1 and Table 2 we observe that our model could significantly outperform the state-of-the-art methods. Specifically, our model achieves an improvement over the current state of the art (HHL in ECCV'18) by **20.2%/20.8%** on Rank-1 accuracy/MAP in the DukeMTMC-reID dataset and by **5.5%/8.6%** in the Market-1501 dataset. This observation validates the effectiveness of MAR.

**Comparison to the hand-crafted feature representation based models**. The performance gaps are most significant when comparing our model to the hand-crafted feature based models [20, 59, 16, 21, 29]. The main reason is that these early works are mostly based on heuristic design, and thus they could not learn optimal discriminative features.

**Comparison to the pseudo label learning based models**. Our model significantly outperforms the pseudo label learning based unsupervised RE-ID models [52, 8]. A key reason is that our soft multilabel reference learning could exploit the auxiliary reference information to mine the potential discriminative information that is hardly detectable when directly comparing the visual features of a pair of visually similar persons. In contrast, the pseudo label learning based models assign the pseudo label by direct comparison of the visual features (e.g. via K-means clustering [52, 8]), rendering them blind to the potential discriminative information.

**Comparison to the unsupervised domain adaptation based models**. Compared to the unsupervised domain

Table 3. Ablation study. Please refer to the text in Sec. 4.4.

| Methods | Market-1501 | | | |
|---|---|---|---|---|
| | rank-1 | rank-5 | rank-10 | mAP |
| Pretrained (source only) | 46.2 | 64.4 | 71.3 | 24.6 |
| Baseline (feature-guided) | 44.4 | 62.5 | 69.8 | 21.5 |
| MAR w/o $L_{CML}$ | 60.0 | 75.9 | 81.9 | 34.6 |
| MAR w/o $L_{CML}\&L_{RAL}$ | 53.9 | 71.5 | 77.7 | 28.2 |
| MAR w/o $L_{RAL}$ | 59.2 | 76.4 | 82.3 | 30.8 |
| MAR | **67.7** | **81.9** | **87.3** | **40.0** |
| Methods | DukeMTMC-reID | | | |
| | rank-1 | rank-5 | rank-10 | mAP |
| Pretrained (source only) | 43.1 | 59.2 | 65.7 | 28.8 |
| Baseline (feature-guided) | 50.0 | 66.4 | 71.7 | 31.7 |
| MAR w/o $L_{CML}$ | 63.2 | 77.2 | 82.5 | 44.9 |
| MAR w/o $L_{CML}\&L_{RAL}$ | 60.1 | 73.0 | 78.4 | 40.4 |
| MAR w/o $L_{RAL}$ | 57.9 | 72.6 | 77.8 | 37.1 |
| MAR | **67.1** | **79.8** | **84.2** | **48.0** |

adaptation based RE-ID models [50, 7, 62, 48], our model achieves superior performances. A key reason is that these models only focus on transfering/adapting the discriminative knowledge from the source domain but ignore the discriminative label information mining in the unlabeled target domain. The discriminative knowledge in the source domain could be less effective in the target domain even after adaptation, because the discriminative clues can be drastically different. In contrast, our model mines the discriminative information in the unlabeled target data, which contributes direct effectiveness to the target RE-ID task.

## 4.4. Ablation study

We perform an ablation study to demonstrate (1) the effectiveness of the soft multilabel guidance and (2) the indispensability of the cross-view consistent soft multilabel learning and the reference agent learning to MAR. For (1), we adopt the *pretrained* model (i.e. only trained by $L_{AL}$ using the auxiliary source MSMT17 dataset to have basic discriminative power, as mentioned in Sec. 4.2). We also adopt a *baseline* model that is feature similarity-guided instead of soft multilabel-guided. Specifically, after the same pretraining procedure, we replace the soft multilabel agreement with the feature similarity, i.e. in the hard negative mining we partition the mined similar pairs into two halves by a threshold of feature similarity rather than soft multilabel agreement, and thus regard the high/low similarity half as positive set $\mathcal{P}$/hard negative set $\mathcal{N}$. For (2), we remove the $L_{CML}$ or $L_{RAL}$. We show the results in Table 3.

**Effectiveness of the soft multilabel-guided hard negative mining**. Comparing MAR to the pretrained model where the soft multilabel-guided hard negative mining is missing, we observe that MAR significantly improves the pretrained model (e.g. on Market-1501/DukeMTMC-reID, MAR improves the pretrained model by 21.5%/24.0% on Rank-1 accuracy). This is because the pretrained model is only discriminatively trained on the auxiliary source dataset without mining the discriminative information in the unlabeled target dataset, so that it is not discriminative enough on the target dataset. This comparison demonstrates the effectiveness of the soft

multilabel-guided hard negative mining.

**Effectiveness of the soft multilabel agreement guidance**. Comparing MAR to the baseline model, we observe that MAR also significantly outperforms the similarity-guided hard negative mining baseline model. (e.g. on Market-1501/DukeMTMC-reID, MAR outperforms the similarity-guided hard negative mining baseline by 23.3%/17.1% on Rank-1 accuracy). Furthermore, even when the soft multilabel learning and reference agent learning losses are missing (i.e. "MAR w/o $L_{CML}\&L_{RAL}$" where the soft multilabel is much worse than MAR), the soft multilabel-guided model still outperforms the similarity-guided model by 14.8%/7.9% on Rank-1 accuracy on Market-1501/DukeMTMC. These demonstrate the effectiveness of the soft multilabel guidance.

**Indispensability of the soft multilabel learning and the reference agent learning**. When the cross-view consistent soft multilabel learning loss is absent, the performances drastically drop (e.g. drop by 7.7%/5.4% on Rank-1 accuracy/MAP in the Market-1501 dataset). This is mainly because optimizing $L_{CML}$ improves the soft multilabel comparability of the cross-view pairs [52], giving more accurate judgement in the positive/hard negative pairs. Hence, the cross-view consistent soft multilabel learning is indispensable in MAR. When the reference agent learning loss is also absent, the performances further drop (e.g. drop by 13.8%/11.8% on Rank-1/MAP in the Market-1501 dataset). This is because in the absence of the reference agent learning, the soft multilabel is learned via comparing to the less valid reference agents (only pretrained). This observation validates the importance of the reference agent learning.

## 4.5. Visual results and insight

To demonstrate how the proposed soft multilabel reference learning works, in Figure 5 we show the similar target pairs with the lowest soft multilabel agreements (i.e. the mined soft multilabel-guided hard negative pairs) mined by our trained model. We make the following observations:

(1) For an unlabeled person image $x$, the maximal entries (label likelihood) of the learned soft multilabel are corresponding to the reference persons that are highly visually similar to $x$, i.e. the soft multilabel represents an unlabeled person mainly by visually similar reference persons.

(2) For a pair of visually similar but unlabeled person images, the soft multilabel reference learning works by discovering potential fine-grained discriminative clues. For example, in the upper-right pair in Figure 5, the two men are dressed similarly. A potential fine-grained discriminative clue is whether they have a backpack. For the man taking a backpack, the soft multilabel reference learning assigns maximal label likelihood to two reference persons who also take backpacks, while for the other man the two reference persons do not take backpacks, either. As a result, the soft multilabel agreement is very low, giving a judgement that

Figure 5. Visual results of the soft multilabel-guided hard negative mining. Each pair surrounded by the red box is the similar pair mined by our model with the lowest soft multilabel agreements, and the images on their right are the reference persons corresponding to the first/second maximal soft multilabel entries. The first row is from the Market-1501 and the second from DukeMTMC-reID. We highlight the discovered fine-grained discriminative clues in the bottom text for each pair. Please view in the screen and zoom in.

this is a hard negative pair. We highlight the discovered fine-grained discriminative clues in the bottom of every pair.

These observations lead us to conclude that the soft multilabel reference learning distinguishes visually similar persons by giving high label likelihood to different reference persons to produce a low soft multilabel agreement.

### 4.6. Further evaluations

**Various numbers of reference persons**. We evaluate how the number of reference persons affect our model learning. In particular, we vary the number by using only the first $N_u$ reference persons (except that we keep all data used in $L_{AL}$ to guarantee that the basic discriminative power is not changed). We show the results in Figure 6.

From Figure 6(a) we observe that: (1) Empirically, the performances become stable when the number of reference persons are larger than 1,500, which is approximately *two times of* the number of the training persons in both datasets (750/700 training persons in Market-1501/DukeMTMC-reID). This indicates that MAR does not necessarily require a very large reference population but a median size, e.g. two times of the training persons. (2) When there are only a few reference persons (e.g. 100), the performances drop drastically due to the poor soft multilabel representation capacity of the small reference population. In other words, this indicates that MAR could not be well learned using a very small auxiliary dataset.
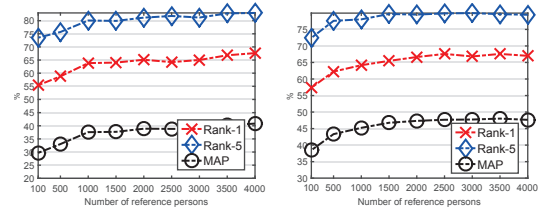


(a) Market-1501      (b) DukeMTMC-reID

Figure 6. Evaluation on different numbers of reference persons.
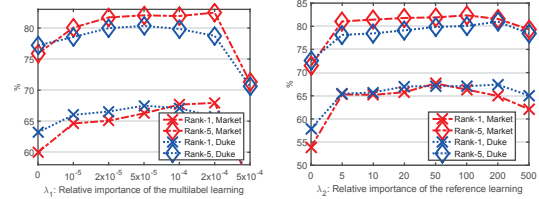


(a) $\lambda_1$      (b) $\lambda_2$

Figure 7. Evaluation on important hyperparameters. For (a) we fix $\lambda_2 = 50$ and for (b) we fix $\lambda_1 = 0.0002$.

**Hyperparameter evaluations**. We evaluate how $\lambda_1$ (which controls the relative importance of the soft multilabel learning) and $\lambda_2$ (the relative importance of the reference agent learning) affect our model learning. We show the results in Figure 7. From Figure 7 we observe that our model learning is stable within a wide range for both hyperparameters (e.g. $2 \times 10^{-5} < \lambda_1 < 5 \times 10^{-4}$ and $10 < \lambda_2 < 200$), although both of them should not be too large to overemphasize the soft multilabel/reference agent learning.

## 5. Conclusion

In this work we demonstrate the effectiveness of utilizing auxiliary source RE-ID data for mining the potential label information latent in the unlabeled target RE-ID data. Specifically, we propose MAR which enables simultaneously the soft multilabel-guided hard negative mining, the cross-view consistent soft multilabel learning and the reference agent learning in a unified model. In MAR, we leverage the soft multilabel for mining the latent discriminative information that cannot be discovered by direct comparison of the *absolute* visual features in the unlabeled RE-ID data. To enable the soft multilabel-guided hard negative mining in MAR, we simultaneously optimize the cross-view consistent soft multilabel learning and the reference agent learning. Experimental results in two benchmarks validate the effectiveness of the proposed MAR and each learning component of MAR.

## Acknowledgement

# References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 2010.

[4] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

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

[6] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[7] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018.

[8] 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)*, 2018.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[10] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *ICML*, 2015.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, 2014.

[12] R. He, X. Wu, Z. Sun, and T. Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *TPAMI*, 2018.

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

[14] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015.

[15] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised $l1$ graph learning. In *ECCV*, 2016.

[16] E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015.

[17] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[19] M. Li, X. Zhu, and S. Gong. Unsupervised person re-identification by deep learning tracklet association. *ECCV*, 2018.

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[21] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *TPAMI*, 2015.

[22] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[23] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. *ICML*, 2015.

[24] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. *ICML*, 2017.

[25] P. Morerio, J. Cavazza, and V. Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *ICLR*, 2018.

[26] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

[27] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011.

[28] S. J. Pan, Q. Yang, et al. A survey on transfer learning. *TKDE*, 2010.

[29] 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 *CVPR*, 2016.

[30] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on Benchmarking Multi-Target Tracking*, 2016.

[31] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[32] S. Roy, S. Paul, N. E. Young, and A. K. Roy-Chowdhury. Exploiting transitivity for learning person re-identification models on a budget. In *CVPR*, 2018.

[33] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018.

[34] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[35] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016.

[36] R. Shu, H. H. Bui, H. Narui, and S. Ermon. A dirt-t approach to unsupervised domain adaptation. *ICLR*, 2018.

[37] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

[38] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[39] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[40] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCVW*, 2016.

[41] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. *ECCV*, 2018.

[42] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[43] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[44] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: l 2 hypersphere embedding for face verification. In *ACMMM*, 2017.

[45] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[46] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014.

[47] H. Wang, X. Zhu, S. Gong, and T. Xiang. Person re-identification in identity regression space. *IJCV*, 2018.

[48] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *CVPR*, 2018.

[49] X. Wang, W. S. Zheng, X. Li, and J. Zhang. Cross-scenario transfer person reidentification. *TCSVT*, 2015.

[50] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[51] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[52] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[53] H.-X. Yu, A. Wu, and W.-S. Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *TPAMI (DOI 10.1109/TPAMI.2018.2886878)*, 2019.

[54] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *TKDE*, 2014.

[55] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[56] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.

[57] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[58] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by saliency learning. *TPAMI*, 2017.

[59] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[60] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013.

[61] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

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

[63] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.