Weakly Supervised Discriminative Feature Learning with State Information for Person Identification

Hong-Xing Yu[1] and Wei-Shi Zheng[1,2,3*]

[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen 518005, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
xKoven@gmail.com, wszheng@ieee.org

Abstract

Unsupervised learning of identity-discriminative visual feature is appealing in real-world tasks where manual labelling is costly. However, the images of an identity can be visually discrepant when images are taken under different states, *e.g. different camera views and poses. This visual discrepancy leads to great difficulty in unsupervised discriminative learning. Fortunately, in real-world tasks we could often know the states without human annotation, e.g. we can easily have the camera view labels in person re-identification and facial pose labels in face recognition. In this work we propose utilizing the state information as weak supervision to address the visual discrepancy caused by different states. We formulate a simple pseudo label model and utilize the state information in an attempt to refine the assigned pseudo labels by the weakly supervised decision boundary rectification and weakly supervised feature drift regularization. We evaluate our model on unsupervised person re-identification and pose-invariant face recognition. Despite the simplicity of our method, it could outperform the state-of-the-art results on Duke-reID, MultiPIE and CFP datasets with a standard ResNet-50 backbone. We also find our model could perform comparably with the standard supervised fine-tuning results on the three datasets. Code is available at* `https://github.com/KovenYu/state-information`.

1. Introduction

While deep discriminative feature learning has shown great success in many vision tasks, it depends highly on the manually labelled large-scale visual data. This limits its scalability to real-world tasks where the labelling is costly and tedious, e.g. person re-identification [76, 53] and unconstrained pose-invariant face recognition [73]. Thus, learning identity-discriminative features without manual labels has drawn increasing attention due to its promise to address the scalability problem [65, 18, 67, 66, 61].
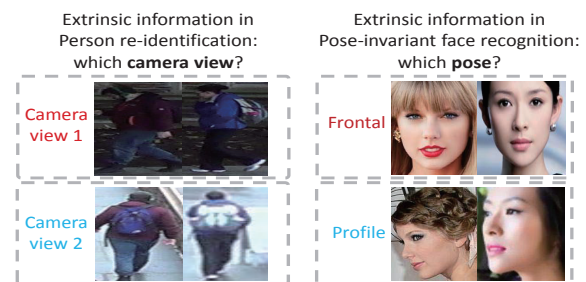

Figure 1. Examples of the state information. A pair of images in each column are of the same individual (We do *not* assume to know any pairing; this figure is only for demonstrating that different states induce visual discrepancy. We only assume to know the camera/pose label of each image but not the pairing).

However, the images of an identity can be drastically different when they are taken under different states such as different poses and camera views. For example, we observe great visual discrepancy in the images of the same pedestrian under different camera views in a surveillance scenario (See Figure 1). Such visual discrepancy caused by the different states induces great difficulty in unsupervised discriminative learning. Fortunately, in real-world discriminative tasks, we can often have some *state information* without human annotation effort. For instance, in person re-identification, it is straightforward to know from which camera view an unlabeled image is taken [65, 67, 66, 12], and in face recognition the pose and facial expression can be estimated by off-the-shelf estimators [68, 48] (See Figure 1). We aim to exploit the state information as weak supervision to address the visual discrepancy in unsupervised discriminative learning. We refer to our task as the weakly supervised discriminative feature learning.

In this work, we propose a novel pseudo label model for weakly supervised discriminative feature learning. We assign every unlabeled image example to a surrogate class (i.e. artificially created pseudo class) which is expected to represent an unknown identity in the unlabelled training set,

---

and we construct the surrogate classification as a simple basic model. However the unsupervised assignment is often incorrect, because the image features of the same identity are distorted due to the aforementioned visual discrepancy. When the visual discrepancy is moderate, in the feature space, an unlabeled example "slips away" from the correct decision region and crosses the decision boundary to the decision region of a nearby surrogate class (See the middle part in Figure 2). We refer to this effect as the feature distortion. We develop the *weakly supervised decision boundary rectification* to address this problem. The idea is to rectify the decision boundary to encourage the unlabeled example back to the correct decision region.

When the feature distortion is significant, however, the unlabeled example can be pushed far away from the correct decision region. Fortunately, the feature distortion caused by a state often follows a specific distortion pattern (e.g., extremely dark illumination in Figure 1 may suppress most visual features). Collectively, this causes a specific global feature drift (See the right part in Figure 2). Therefore, we alleviate the significant feature distortion to a moderate level (so that it can be addressed by the decision boundary rectification) by countering the global-scale feature drifting. Specifically, we achieve this by introducing the *weakly supervised feature drift regularization*.

We evaluate our model on two tasks, i.e. unsupervised person re-identification and pose-invariant face recognition. We find that our model could perform comparably with the standard supervised learning on DukeMTMC-reID [77], Multi-PIE [24] and CFP [47] datasets. We also find our model could outperform the state-of-the-art unsupervised models on DukeMTMC-reID and supervised models on Multi-PIE and CFP. To our best knowledge, this is the first work to develop a weakly supervised discriminative learning model that can successfully apply to different tasks, leveraging different kinds of state information.

2. Related Work

Learning with state information. State information has been explored separately in identification tasks. In person re-identification (RE-ID), several works leveraged the camera view label to help learn view-invariant features and distance metrics [34, 12, 7, 32, 81]. In face recognition, the pose label was also used to learn pose-invariant models [74, 72, 29, 84, 62, 50, 63]. Specifically, [66] and [42] visualized the feature embedding to illustrate the feature distortion problem nicely for person re-identification and face recognition, respectively. However, most existing methods are based on supervised learning, and thus the prohibitive labelling cost could largely limit their scalability. Therefore, *unsupervised RE-ID* [66, 65, 31, 19, 56, 61, 59] and *cross-domain transfer learning RE-ID* [67, 13, 78, 79, 52, 80, 11, 71, 20] have been attracting increasing attention. These methods typically incorporate the camera view labels to learn the camera

view-specific feature transforms [65, 66], to learn the soft multilabels [67], to provide associations between the video RE-ID tracklets [31], or to generate augmentation images [78, 79, 80]. Our work is different from the cross-domain transfer learning RE-ID methods in that we do not need any labeled data in the training stage. As for the unsupervised RE-ID methods, the most related works are [65, 66] where Yu et.al. proposed the asymmetric clustering in which the camera view labels were leveraged to learn a set of view-specific projections. However, they need to learn as many projections as the camera views via solving the costly eigen problem, which limits their scalability. In contrast we learn a generalizable feature for all kinds of states (camera views).

Weakly supervised learning. Our method is to iteratively refine pseudo labels with the state information which is regarded as weak supervision. The state information serves to guide the pseudo label assignments as well as to improve the feature invariance against distractive states.

In literatures, weak supervision is a broadly used term. Typical weak supervision [82] includes image-level coarse labels for finer tasks like detection [5, 6] and segmentation [55, 40]. Another line of research that is more related to our work is utilizing large-scale inaccurate labels (typically collected online [1] or from a database like Instagram [36] or Flickr [28]) to learn general features. Different from existing works, our objective is to learn identity-discriminative features that are directly applicable to identification tasks without supervised fine-tuning.

Unsupervised deep learning. Beyond certain vision applications, general unsupervised deep learning is a long-standing problem in vision community. The typical lines of research include clustering based methods [9, 60, 2, 17] which discovered cluster structures in the unlabelled data and utilized the cluster labels, and the generation based methods which learned low-dimensional features that were effective for generative discrimination [44, 16, 23] or reconstruction [51, 30, 3].

Recently, self-supervised learning, a promising paradigm of unsupervised learning, has been quite popular. Self-supervised methods typically construct some pretext tasks where the supervision comes from the data. Typical pretext tasks include predicting relative patch positions [15], predicting future patches [39], solving jigsaw puzzles [37, 38], image inpainting [41], image colorization [69, 70] and predicting image rotation [22]. By solving the pretext tasks, they aimed to learn features that were useful for downstream real-world tasks.

Our goal is different from these works. Since they aim to learn useful features for various downstream tasks, they were designed to be downstream task-agnostic, and required supervised fine-tuning for them. In contrast, we actually focus on the "fine-tuning" step, with a goal to reduce the need of manual labeling.

3. Weakly supervised Discriminative Learning with State Information

Let $\mathcal{U} = \{u_i\}_{i=1}^{N}$ denote the unlabelled training set, where u_i is an unlabelled image example. We also know the state $s_i \in \{1, \cdots, J\}$, e.g., the illumination of u_i is dark, normal or bright. Our goal is to learn a deep network f to extract identity-discriminative feature which is denoted by $x = f(u; \theta)$. A straightforward idea is to assume that in the feature space every x belongs to a surrogate class which is modelled by a surrogate classifier μ. A surrogate class is expected to model a potential unknown identity in the unlabeled training set. The discriminative learning can be done by a surrogate classification:

$$\min_{\theta, \{\mu_k\}} L_{surr} = -\Sigma_x \log \frac{\exp(x^{\mathrm{T}}\mu_{\hat{y}})}{\Sigma_{k=1}^{K} \exp(x^{\mathrm{T}}\mu_k)}, \qquad (1)$$

where $\hat{y}$ denotes the surrogate class label of x, and K denotes the number of surrogate classes. An intuitive method for surrogate class assignment is:

$$\hat{y} = \arg\max_{k} \exp(x^{\mathrm{T}}\mu_k). \qquad (2)$$

However, the visual discrepancy caused by the state leads to incorrect assignments. When the feature distortion is moderate, wrong assignments happen locally, i.e., x wrongly crosses the decision boundary into a nearby surrogate class' decision region. We develop the *Weakly supervised Decision Boundary Rectification* (WDBR) to address it. As for the *significant* feature distortion, however, it is extremely challenging as x is pushed far away from the correct decision region. To deal with it, we introduce the *Weakly supervised Feature Drift Regularization* to alleviate the significant feature distortion down to a moderate level that WDBR can address. We show an overview illustration in Figure 2.

3.1. Weakly supervised decision boundary rectification (WDBR)

We first consider the moderate visual feature distortion. It "nudges" an image feature x to wrongly cross the decision boundary into a nearby surrogate class. For example, two persons wearing dark clothes are even harder to distinguish when they both appear in a dark camera view. Thus, these person images are assigned to the same surrogate class (see Figure 2 for illustration). In this case, a direct observation is that most members of the surrogate class is taken from the same dark camera view (i.e. the same state). Therefore, we quantify the extent to which a surrogate class is dominated by a state. We push the decision boundary toward a highly dominated surrogate class or even nullify it, in an attempt to correct these local boundary-crossing wrong assignments.

We quantify the extent by the *Maximum Predominance Index* (MPI). The MPI is defined as the proportion of the most common state in a surrogate class. Formally, the MPI of the k-th surrogate class R_k is defined by:

$$R_k = \frac{\max_j |\mathcal{M}_k \cap \mathcal{Q}_j|}{|\mathcal{M}_k|} \in [0, 1], \qquad (3)$$

where the denominator is the number of members in a surrogate class, formulated by the cardinality of the *member set* of the k-th surrogate class $\mathcal{M}_k$:

$$\mathcal{M}_k = \{x_i | \hat{y}_i = k\}, \qquad (4)$$

and the numerator is the number of presences of the most common state in $\mathcal{M}_k$. We formulate it by the intersection of $\mathcal{M}_k$ and the *state subset* corresponding to the j-th state $\mathcal{Q}_j$:

$$\mathcal{Q}_j = \{x_i | s_i = j\}. \qquad (5)$$

Note that the member set $\mathcal{M}_k$ is dynamically updated, as the surrogate class assignment (Eq. (2)) is on-the-fly along with the learning, and is improved upon better learned features.

As analyzed above, a higher R_k indicates that it is more likely that some examples have wrongly crossed the decision boundary into the surrogate class μ_k due to the feature distortion. Hence, we shrink that surrogate class' decision boundary to purge the potential boundary-crossing examples from its decision region. Specifically, we develop the weakly supervised rectified assignment:

$$\hat{y} = \arg\max_{k} p(k) \exp(x^{\mathrm{T}}\mu_k), \qquad (6)$$

where $p(k)$ is the rectifier function that is monotonically decreasing with R_k:

$$p(k) = \frac{1}{1 + \exp(a \cdot (R_k - b))} \in [0, 1], \qquad (7)$$

where $a \geq 0$ is the rectification strength and $b \in [0, 1]$ is the rectification threshold. We typically set $b = 0.95$. In particular, we consider $a = \infty$, and thus we have:

$$p(k) = \begin{cases} 1, & \text{if } R_k \leq b \\ 0, & \text{otherwise} \end{cases} \qquad (8)$$

This means that when the MPI exceeds the threshold b we nullify it by shrinking its decision boundary to a single point. We show a plot of $p(k)$ in Figure 3(a).

For any two neighboring surrogate classes μ_1 and μ_2, the decision boundary is (where we leave the derivation to the supplementary material):

$$(\mu_1 - \mu_2)^{\mathrm{T}}x + \log \frac{p(1)}{p(2)} = 0. \qquad (9)$$

Discussion. To have a better understanding of the WDBR, let us first consider the hard rectifier function. When a surrogate class' MPI exceeds the threshold b (typically we set $b = 0.95$), the decision region vanishes, and no example would be assigned to the surrogate class (i.e., it is completely nullified). Therefore, WDBR prevents the unsupervised learning from being misled by those severely affected surrogate classes. For example, if over 95% person images assigned to a surrogate class are from the same dark camera view, it is highly likely this is simply because it is too dark to distinguish them, rather than because they are the same
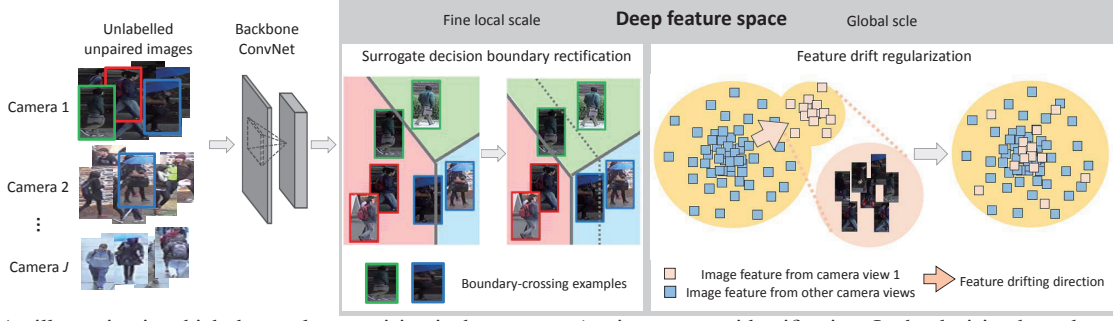
Figure 2. An illustration in which the weak supervision is the *camera view* in person re-identification. In the decision boundary rectification, each colored bounding box denotes an identity (the identity labels are *unknown*; we use them here only for illustrative purpose). Each colored region denotes the decision region of a surrogate class. In the feature drift regularization, we illustrate the visually dominant state by the dark camera view 1. Best viewed in color.
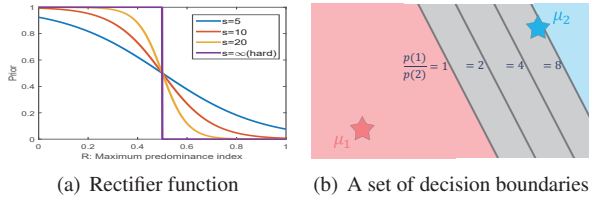

(a) Rectifier function   (b) A set of decision boundaries

Figure 3. (a) The rectifier function $p(k)$ with different strength a. (b) Decision boundaries with different $\frac{p(1)}{p(2)}$ values.

person. Thus, WDBR nullifies this poorly formed surrogate class.

When we use the soft rectifier function, the WDBR does not directly nullify the surrogate class that exceeds the threshold, but favors the surrogate class which has lower MPI (because they are less likely to have the boundary-cross problem) by moving the decision boundary. This can be seen from Figure 3(b) where we plot a set of decision boundaries in the two-class case. In some sense, the soft WDBR favors the state-balanced surrogate classes. This property may further improve the unsupervised learning, especially if the unlabelled training set is indeed state-balanced for most identities. However, if we do not have such a prior knowledge of state balance, using hard rectifier can be more desirable, because hard rectifier does not favor state-balanced surrogate classes. We will discuss more about this property upon real cases in Sec. 4.2.

In the supplementary material, we theoretically justify our model by showing that the rectified assignment is the maximum a posteriori optimal estimation of $\hat{y}$. However, the WDBR is a local mechanism, i.e. WDBR deals with the moderate feature distortion that nudges examples to slip in nearby surrogate classes. Its effectiveness might be limited when the feature distortion is significant.

3.2. Weakly supervised feature drift regularization

A visually dominant state may cause a *significant* feature distortion that pushes an example far away from the correct surrogate class. This problem is extremely difficult

to address by only considering a few surrogate classes in a local neighborhood. Nevertheless, such a significant feature distortion is likely to follow a specific pattern. For example, the extremely low illumination may suppress all kinds of visual features: dim colors, indistinguishable textures, etc. Collectively, we can capture the significant feature distortion pattern in a global scale. In other words, such a state-specific feature distortion would cause many exmaples x in the state subset to drift toward a specific direction (see Figure 2 for illustration). We capture this by the state sub-distribution and introduce the Weakly supervised Feature Drift Regularization (WFDR) to address it and complement the WDBR.

In particular, we define the *state sub-distribution* as $\mathbb{P}(\mathcal{Q}_j)$, which is the distribution over the state subset $\mathcal{Q}_j$ defined in Eq. (5). For example, all the unlabeled person images captured from a dark camera view. We further denote the distribution over the whole unlabelled training set as $\mathbb{P}(\mathcal{X})$, where $\mathcal{X} = f(\mathcal{U})$. Apparently, the state-specific feature distortion would lead to a specific sub-distributional drift, i.e., $\mathbb{P}(\mathcal{Q}_j)$ drifts away from $\mathbb{P}(\mathcal{X})$. For example, all person images from a dark camera view may be extremely low-valued in many feature dimensions, and this forms a specific distributional characteristic. Our idea is straightforward: we counter this "collective drifting force" by aligning the state sub-distribution $\mathbb{P}(\mathcal{Q}_j)$ with the overall total distribution $\mathbb{P}(\mathcal{X})$ to suppress the significant feature distortion. We formulate this idea as the Weakly supervised Feature Drift Regularization (WFDR):

$$\min_{\theta} L_{drift} = \Sigma_j d(\mathbb{P}(\mathcal{Q}_j), \mathbb{P}(\mathcal{X})), \qquad (10)$$

where $d(\cdot, \cdot)$ is a distributional distance. In our implementation we adopt the simplified 2-Wasserstein distance [4, 26] as $d(\cdot, \cdot)$ due to its simplicity and computational ease. In particular, it is given by:

$$d(\mathbb{P}(\mathcal{Q}_j), \mathbb{P}(\mathcal{X})) = ||m_j - m||_2^2 + ||\sigma_j - \sigma||_2^2, \qquad (11)$$

where m_j/σ_j is the mean/standard deviation feature vector over $\mathcal{Q}_j$. Similarly, m/σ is the mean/standard deviation feature vector over the whole unlabelled training set $\mathcal{X}$.

Ideally, WFDR alleviates the significant feature distortion down to a mild level (i.e., x is regularized into the correct

decision region) or a moderate level (i.e., x is regularized into the neighborhood of the correct surrogate class) that the WDBR can address. Thus, it is mutually complementary to the WDBR. We note that the WFDR is mathematically akin to the soft multilabel learning loss in [67], but they serve for different purposes. The soft multilabel learning loss is to align the cross-view *associations* between unlabeled target images and labeled source images, while we aim to align the feature distributions of unlabeled images and we do not need a source dataset.

Finally, the loss function of our model is:

$$\min_{\theta, \{\mu_k\}} L = L_{surr} + \lambda L_{drift}, \tag{12}$$

where $\lambda > 0$ is a hyperparameter to balance the two terms.

In our implementation we used the standard ResNet-50 [25] as our backbone network. We trained our model for approximately 1,600 iterations with batchsize 384, momentum 0.9 and weight decay 0.005. We followed [25] to use SGD, set the learning rate to 0.001, and divided the learning rate by 10 after 1,000/1,400 iterations. We used a single SGD optimizer for both θ and $\{\mu_k\}_{k=1}^{K}$. Training costed less than two hours by using 4 Titan X GPUs. We initialized the surrogate classifiers $\{\mu_k\}_{k=1}^{K}$ by performing standard K-means clustering on the initial feature space and using the cluster centroids. For further details please refer to the supplementary. We also summarize our method in an algorithm in the supplementary material[1].

4. Experiments

4.1. Datasets

We evaluated our model on two real-world discriminative tasks with state information, i.e. person re-identification (RE-ID) [76] and pose-invariant face recognition (PIFR) [27, 14]. In RE-ID which aims to match person images across non-overlapping camera views, the state information is the camera view label, as illustrated in Figure 4(a) and 4(b). Note that each camera view has its specific conditions including illumination, viewpoint and occlusion (e.g. Figure 4(a) and 4(b)). In PIFR, which aims to identify faces across different poses, the state information is the pose, as illustrated in Figure 4(c). We note that on both tasks the training identities are completely different from the testing identities. Hence, these tasks are suitable to evaluate the discriminability and generalisability of learned feature.

Person re-identification (RE-ID). We evaluated on Market-1501 [75] and DukeMTMC-reID [77, 45]. Market-1501 contains 32,668 person images of 1,501 identities. Each person is taken images from at least 2 out of 6 disjoint camera views. We followed the standard evaluation protocol where the training set had 750 identities and testing set had the other

---

[1] Code can be found at `https://github.com/KovenYu/state-information`

751 identities [75]. The performance was measured by the cumulative accuracy and the mean average precision (MAP) [75]. DukeMTMC-reID contains 36,411 person images of 1,404 identities. Images of each person were taken from at least 2 out of 8 disjoint camera views. We followed the standard protocol which was similar to the Market-1501 [45]. We followed [67] to pretrain the network with standard softmax loss on the MSMT17 dataset [54] in which the scenario and identity pool were completely different from Market-1501 and DukeMTMC-reID. It should be pointed out that in fine-grained discriminative tasks like RE-ID and PIFR, the pretraining is important for unsupervised models because the class-discriminative visual clues are not general but highly task-dependent [21, 18, 52, 66], and therefore some extent of field-specific knowledge is necessary for successful unsupervised learning. We resized the images to 384×128. In the unsupervised setting, the precise number of training classes (persons) P (i.e. 750/700 for Market-1501/DukeMTMC-reID) should be unknown. Since our method was able to automatically discard excessive surrogate classes, an "upper bound" estimation could be reasonable. We set $K = 2000$ for both datasets.

Pose-invariant face recognition (PIFR). We mainly evaluated on the large dataset Multi-PIE [24]. Multi-PIE contains 754,200 images of 337 subjects taken with up to 20 illuminations, 6 expressions and 15 poses [24]. For Multi-PIE, most experiments followed the widely-used setting [84] which used all 337 subjects with neutral expression and 9 poses interpolated between $-60°$ and $60°$. The training set contained the first 200 persons, and the testing set contained the remaining 137 persons. When testing, one image per identity with the frontal view was put into the gallery set and all the other images into the query set. The performance was measured by the top-1 recognition rate. We detected and cropped the face images by MTCNN [68], resized the cropped images to 224×224, and we adopted the pretrained model weights provided by [8]. Similarly to the unsupervised RE-ID setting, we simply set $K = 500$. We also evaluated on an unconstrained dataset CFP [47]. The in-the-wild CFP dataset contains 500 subjects with 10 frontal and 4 profile images for each subject. We adopted the more challenging frontal-profile verification setting [47]. We followed the official protocol [47]. to report the mean accuracy, equal error rate (EER) and area under curve (AUC).

In the unsupervised RE-ID task, the camera view labels were naturally available [67, 66]. In PIFR we used groundtruth pose labels for better analysis. In the supplementary material we showed the simulation results when we used the estimated pose labels. The performance did not drop until the correctly estimated pose labels were less than 60%. In practice the facial pose is continuous and we need to discretize it to produce the pose labels. In our preliminary experiments on Multi-PIE we found that merging the pose labels into coarse-grained groups did not affect the

(a) Market-1501   (b) DukeMTMC-reID
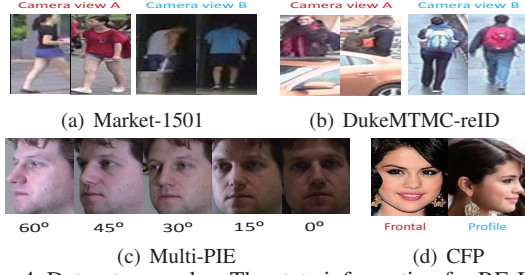
(c) Multi-PIE   (d) CFP

Figure 4. Dataset examples. The state information for RE-ID and PIFR is camera view labels and pose labels, respectively.

Table 1. Model evaluation on the person re-identification (%). Please refer to Sec. 4.2 for description of the compared methods.

Methods	DukeMTMC-reID			Market-1501		
	rank-1	rank-5	mAP	rank-1	rank-5	mAP
Supervised fine-tune	75.0	85.0	57.2	85.9	95.2	66.8
Pretrained	43.1	59.2	28.8	46.2	64.4	24.6
K-means as labels	37.3	52.1	25.2	47.3	63.1	25.6
Basic model	56.8	71.1	39.6	55.8	72.2	31.5
Basic + WDBR (hard)	69.4	80.5	50.2	60.3	73.4	34.5
Basic + WDBR (soft)	63.6	77.2	45.4	60.0	75.6	34.3
Basic + WFDR	67.7	79.4	47.5	67.4	82.3	39.4
Full model (hard)	**72.1**	**83.5**	**53.8**	**74.0**	**87.4**	**47.9**
Full model (soft)	70.3	81.7	50.0	70.7	85.2	43.4

Table 2. Model evaluation on the Multi-PIE (%). We report the mean results averaged over 5 runs. Please refer to Sec. 4.2 for description of the compared methods.

Methods	avg	0°	±15°	±30°	±45°	±60°
Supervise fine-tune	98.2	99.7	99.4	98.8	98.1	95.7
Pretrained	88.7	98.5	97.5	93.7	89.7	71.2
K-means as labels	81.0	95.7	94.6	89.1	76.7	56.0
Basic model	54.5	91.2	86.5	60.0	34.5	18.8
Basic + WDBR (hard)	91.7	98.9	98.7	97.5	91.2	75.9
Basic + WDBR (soft)	97.0	99.1	98.9	98.3	96.8	93.0
Basic + WFDR	95.7	98.4	98.1	97.0	95.5	91.0
Full model (hard)	95.7	98.3	98.1	97.0	95.3	91.1
Full model (soft)	**97.1**	**99.1**	**98.9**	**98.3**	**96.8**	**93.1**

Table 3. Model evaluation on CFP (%). Please refer to Sec. 4.2 for description of the compared methods.

Methods	Accuracy	EER	AUC
Supervised fine-tune	95.50(0.98)	4.74(1.05)	98.82(0.50)
Pretrained	92.90(1.37)	7.40(1.37)	97.75(0.73)
Basic model	93.57(1.32)	6.89(1.51)	97.55(0.93)
Full model (soft)	**95.49**(0.70)	**4.74**(0.72)	**98.83**(0.29)

performance significantly. Therefore, for fair comparison to other methods, we followed the conventional setting to use the default pose labels. We set $\lambda = 10$ and $b = 0.95$ for all datasets except Multi-PIE which has more continual poses and thus we decreased to $\lambda = 1$, $b = 0.5$. We evaluated both soft version $a = 5$ and hard version $a = \infty$. We provide evaluations and analysis for K, λ, a and b in the supplementary material.

4.2. Model evaluation and analysis

We decomposed our model for analysis. To ground the performance, we provided the standard supervised fine-tuning results (i.e. replacing our proposed loss with softmax loss with groundtruth class labels, and keeping other settings the same) which could be seen as an upper bound. As an unsupervised baseline, we used K-means cluster labels

(i.e. we performed K-means once on the pretrained feature space to obtain the cluster labels, and used the cluster labels instead of the groundtruth labels for fine-tuning) and we denote this as "K-means as labels". We also ablated both WDBR and WFDR from our full model to obtain a "Basic model". The key difference between "K-means as labels" and "Basic model" is that the former uses fixed cluster labels while the latter dynamically infers pseudo labels every batch along with model training. We show the results in Table 1, 2 and 3. On CFP the observation was similar to Multi-PIE and we show the most significant results only.

Comparable performance to standard supervised fine-tuning. Compared to the standard supervised fine-tuning, we found that our model could perform *comparably with the supervised results* in both the person re-identification task on DukeMTMC-reID and the face recognition task on Multi-PIE and CFP. The overall effectiveness was clear when we ground the performance by both the supervised results and the pretrained baseline results. For example, on DukeMTMC-reID, the supervised learning improved the pretrained network by 31.9% in rank-1 accuracy, while our model improved it by 29.0%, leaving only a gap of 2.9%. On Multi-PIE our model achieved 97.1% average recognition rate which was very closed to the supervised result 98.2%. On CFP our model even achieved approximately the same performance as supervised fine-tuning, probably because the small training set (6300 images) favored a regularization. We also notice that significant performances held both when the initial pretrained backbone network was *weak* (e.g. in RE-ID the initial rank-1 accuracy performance was below 50%) and when the initial backbone network was *strong* (i.e. in PIFR the initial recognition accuracy performance was over 80%). These comparisons verified the effectiveness of our model.

Soft vs. hard decision boundary rectification. We found that the soft rectification performed better on PIFR benchmarks while hard rectification excelled at RE-ID. We assumed that a key reason was that on the RE-ID datasets, different persons' images were *unbalanced*, i.e., some IDs appeared only in two camera views while some may appear in up to six camera views. For example, for a person A who appeared in 2 camera views, the MPI R_A was at least 1/2, while this lower bound was 1/6 for another person who appeared in 6 camera views. Thus the soft rectifier may unfairly favor the surrogate class corresponding to the person appearing in more camera views. While the hard rectifier does not favor state-balance: it only nullified highly likely incorrect surrogate classes with very high MPI. Therefore, the hard rectification could be more robust to the state imbalance. On the other hand, for Multi-PIE and CFP where the classes were balanced, soft rectification would fine-tune the decision boundary to a better position, and thus achieved better results. *Hence, in this paper we used the hard WDBR*

Table 4. Comparison to the state-of-the-art unsupervised RE-ID (upper) and domain adaptation RE-ID (middle) models.

Methods	Reference	DukeMTMC-reID		Market-1501	
		rank-1	mAP	rank-1	mAP
CAMEL [65]	ICCV'17	40.3	19.8	54.5	26.3
PUL [18]	ToMM'18	30.0	16.4	45.5	20.5
DECAMEL [66]	TPAMI'18	-	-	60.2	32.4
Distill [57]	CVPR'19	48.4	29.4	61.5	33.5
Wu et.al. [58]	ICCV'19	59.3	37.8	65.4	35.5
HHL [78]	ECCV'18	46.9	27.2	62.2	31.4
ECN [79]	CVPR'19	63.3	40.4	75.1	43.0
MAR [67]	CVPR'19	67.1	48.0	67.7	40.0
UCDA-CCE [43]	ICCV'19	55.4	36.7	64.3	34.5
PDA-Net [33]	ICCV'19	63.2	45.1	**75.2**	47.6
DeepCluster [9]	ECCV'18	40.2	26.7	48.0	26.1
ours	This work	**72.1**	**53.8**	74.0	**47.9**

Table 5. Comparison to the state-of-the-art **supervised** PIFR models on the Multi-PIE dataset.

Methods	avg	0°	±15°	±30°	±45°	±60°
FIP [83]	72.9	94.3	90.7	80.7	64.1	45.9
MVP [84]	79.3	95.7	92.8	83.7	72.9	60.1
CPI [62]	83.3	**99.5**	95.0	88.5	79.9	61.9
DRGAN [50]	90.2	97.0	94.0	90.1	86.2	83.2
FFGAN [64]	91.6	95.7	94.6	92.5	89.7	85.2
p-CNN [63]	93.5	95.4	95.2	94.3	93.0	90.3
DeepCluster [9]	86.4	96.7	96.6	93.3	84.8	65.6
ours	**97.1**	99.1	**98.9**	**98.3**	**96.8**	**93.1**

Table 6. Comparison to the state-of-the-art **supervised** PIFR models on the CFP dataset. Format: mean(standard deviation).

Methods	Accuracy	EER	AUC
Deep Features [47]	84.91(1.82)	14.97(1.98)	93.00(1.55)
Triplet Embedding [46]	89.17(2.35)	8.85(0.99)	97.00(0.53)
Chen et.al. [10]	91.97(1.70)	8.00(1.68)	97.70(0.82)
PIM [73]	93.10(1.01)	7.69(1.29)	97.65(0.62)
DRGAN [50]	93.41(1.17)	6.45(0.16)	97.96(0.06)
p-CNN [63]	94.39(1.17)	5.94(0.11)	98.36(0.05)
Human	94.57(1.10)	5.02(1.07)	**98.92**(0.46)
DeepCluster [9]	91.30(1.58)	8.86(1.70)	96.77(0.96)
ours	**95.49**(0.70)	**4.74**(0.72)	98.83(0.29)

for RE-ID and the soft WDBR for PIFR.

Complementary nature of WDBR and WFDR. Comparing the basic model (our model without WDBR or WFDR) to basic model with either WDBR or WFDR, the performance was consistently improved. With both WDBR and WFDR, the performance was further improved. This showed that the fine local-scale WDBR and the global-scale WFDR were complementarily effective.

We noticed that on Multi-PIE this complementary nature was less significant, as using WDBR alone could achieve similar results to the full model. This may be due to the *continual* nature of the pose variation on Multi-PIE: the variation from 0° to 60° is in a locally connected *manifold* [49], with 15°/30°/45° in between. Therefore, it was easier for our local mechanism to gradually "connect" some 0°/15° surrogate classes with some 15°/30° surrogate classes to finally have a global aligning effect. In contrast, in RE-ID this manifold nature is less apparent since it lacks evidence of inherent relations between each pair of camera views.

4.3. Comparison to the state-of-the-art methods

We further demonstrated the effectiveness of our model by comparing to the state-of-the-art methods on both tasks. It should be pointed out that for RE-ID and PIFR where

the goal was to solve the real-world problem, there were no standards on architecture: different works used different networks and different pretraining data. Thus we simply kept using the standard ResNet-50 without task-specific improvements and using the public pretraining data. For a fairer comparison, we also compared our method with a recent unsupervised deep learning method DeepCluster [9], which also uses a discriminative classification loss. We used the same architecture and pretraining as for our method. We show the results in Table 4, 5 and 6.

Superior performance across tasks and benchmarks. Compared to the reported results, our method could achieve the state-of-the-art performances. On unsupervised RE-ID task, our method achieved a 5.0%/5.8% absolute improvement in rank-1 accuracy/MAP on DukeMTMC-reID, compared to the recent state-of-the-art RE-ID model MAR [67], which used exactly the same architecture and pretraining data as ours. Although a few recent domain adaptation methods [33] achieve comparable performances to our method, it is worth noting that they rely on labeled source data for discriminative learning, while we do *not* use labeled data and our method can generalize to different tasks instead of specifically modeling a single RE-ID task. We note that many of the compared recent state-of-the-art RE-ID methods also exploited the camera view labels [78, 79, 67, 65, 66, 58, 43]. For instance, the domain adaptation RE-ID models HHL [78]/ECN [79] leveraged the camera view labels to synthesize cross-view person images for training data augmentation [78, 79], and MAR [67] used view labels to learn the view-invariant soft multilabels. On the pose-invariant face recognition task, our model outperformed the state-of-the-art supervised results on both Multi-PIE and the CFP benchmarks. We also note that most compared PIFR models exploited both the identity labels and the pose labels.

Our model also outperformed the DeepCluster [9] significantly on all the four benchmarks. A major reason should be that some discriminative visual clues (e.g. fine-grained clothes pattern) of persons (/faces) were "overpowered" by the camera view (/pose) induced feature distortion. Without appropriate mechanisms to address this problem, the clustering might be misled by the feature distortion to produce inferior cluster separations. In contrast, our model addressed this problem via the weakly supervised decision boundary rectification and the feature drift regularization.

4.4. Visualization

To provide visual insights of the problems our model tried to address, we show the t-SNE embedding [35] of the learned features in Figure 5. Note that the shown identities were *unseen* during training, and thus the characteristic reflected in the qualitative results were generalisable.

Addressing intra-identity visual feature discrepancy. Let us compare Figure 5(a) and Figure 5(b) for an illustration. Figure 5(a) illustrated that the same person (see the

(a) RE-ID, basic model   (b) RE-ID, full model

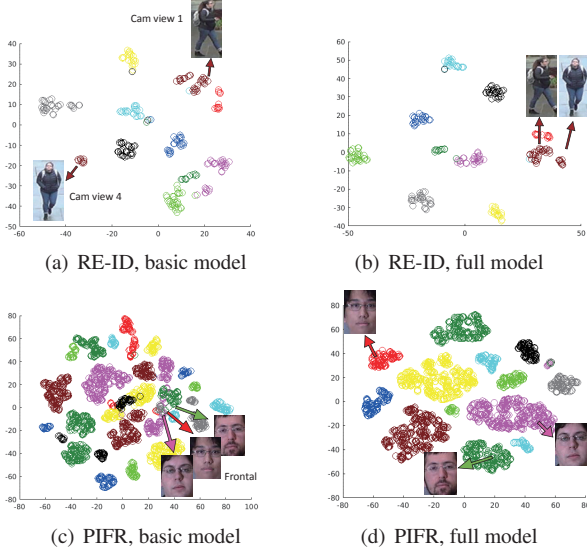(c) PIFR, basic model   (d) PIFR, full model

Figure 5. The t-SNE embedding of 10 identities on the DukeMTMC-reID and Multi-PIE datasets. Each color represents an identity randomly chosen from the unseen *testing* set. Best viewed in screen and zoom in.

Table 7. Multiple kinds of state information on Multi-PIE. We report mean results with 5 runs (%).

Methods	avg	0°	±15°	±30°	±45°	±60°
Supervised fine-tune	96.6	98.9	98.6	97.6	96.0	93.1
Pretrained	83.8	95.6	94.2	89.3	81.5	64.6
w/ only pose labels	94.6	96.5	96.2	95.9	94.9	90.6
w/ only illumination labels	30.3	73.3	65.8	24.9	7.0	2.2
w/ only expression labels	43.5	80.6	75.2	51.5	26.2	2.6
w/ all three kinds of labels	**95.9**	**98.3**	**98.2**	**97.2**	**95.7**	**91.1**

4.5. Multiple kinds of state information

Our method is easy to extend to incorporate multiple kinds of state information. We experimented on Multi-PIE with the state information being expression, illumination and pose. We used all 6 expressions, 20 illuminations and 9 poses. We decomposed the rectifier by $p(k) = p_p(k) \cdot p_i(k) \cdot p_e(k)$, where the subscripts $p/i/e$ stand for pose/illumination/expression, respectively. We also accordingly use three equally-weighted feature drift regularization terms in the loss function. We used hard WDBR to have a regular shape of rectifier function. We show the results in Table 7. Exploiting pose labels produced much better results than illumination and expression, indicating that pose was the most distractive on Multi-PIE. Exploiting all three kinds of state information further improved the performance to 95.9%, which was closed to the supervised result 96.6%. This comparison showed that our model could be further improved when more valuable state information was available.

5. Conclusion and Discussion

In this work we proposed a novel psuedo label method with state information. We found that some proper state information could help address the visual discrepancy caused by those distractive states. Specifically, we investigate the state information in person re-identification and face recognition and found the camera view labels and pose labels to be effective. Our results indicate that it is reasonable to make use of the free state information in unsupervised person re-identification and face recognition. Since the weakly supervised feature drift regularization (WFDR) is a simple loss term which is model-free, it can be plugged into other different methods than our proposed pseudo label method.

However, we should point out that our method works with the state information that corresponds to the visually distractive states. As for more general state information, it still remains an open problem to effectively utilize it.

6. Acknowledgement

highlighted brown points) appeared differently in two camera views, which had different viewpoints and backgrounds. This visual discrepancy caused significant feature distortion of the identity. Apparently, it was extremely difficult to address this problem if other effective mechanisms were not provided besides feature similarity. From Figure 5(b) we observed that when equiped with the WDBR and WFDR, the feature distortion was significantly alleviated. This observation indicated that the our model leveraged the state information to effectively alleviate the intra-identity visual discrepancy for better discriminative feature learning.

Addressing inter-identity visual feature entanglement. In a more complex case shown in Figure 5(c), we observed that some visually similar frontal face images (males with eye glasses) were entangled with each other in the feature space learned by the basic model. In particular, some magenta, red and dark green points highly overlapped with each other. This demonstrated that if we simply used feature similarity, it was also extremely difficult to address the inter-identity visual feature entanglement. Nevertheless, as shown in Figure 5(d) our full model could address this problem with the WDBR and WFDR. The learned feature space was much desirable, and the inter-identity overlapping points were now distant from each other. In other words, our model could leverage the state information to help the unsupervised learning via alleviating the inter-identity visual feature entanglement.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016.

[2] Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, and Bjorn Ommer. Cliquecnn: Deep unsupervised exemplar learning. In *NIPS*, 2016.

[3] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007.

[4] David Berthelot, Thomas Schumm, and Luke Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[5] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, 2015.

[6] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[7] Alessandro Borgia, Yang Hua, Elyor Kodirov, and Neil M Robertson. Cross-view discriminative feature learning for person re-identification. *TIP*, 2018.

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[10] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, 2013.

[11] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*, 2019.

[12] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2018.

[13] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018.

[14] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 2016.

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[16] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[17] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*, 2014.

[18] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018.

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

[20] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[21] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014.

[24] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *IVC*, 2010.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[26] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *TPAMI*, 2018.

[27] Fu Jie Huang, Zhihua Zhou, Hong-Jiang Zhang, and Tsuhan Chen. Pose invariant face recognition. In *FG*, 2000.

[28] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016.

[29] Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, 2014.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[32] Sheng Li, Ming Shao, and Yun Fu. Person re-identification by cross-view multi-level dictionary learning. *TPAMI*, 2018.

[33] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019.

[34] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[36] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[38] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *CVPR*, 2018.

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[40] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *CVPR*, 2015.

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[42] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017.

[43] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *ICCV*, 2019.

[44] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[45] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on Benchmarking Multi-Target Tracking*, 2016.

[46] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, 2016.

[47] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016.

[48] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 2009.

[49] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[50] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[51] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010.

[52] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *CVPR*, 2018.

[53] Xiaojuan Wang, Wei Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *TCSVT*, 2015.

[54] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[55] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 2016.

[56] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *CVPR*, 2019.

[57] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *CVPR*, 2019.

[58] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *ICCV*, 2019.

[59] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z. Li. Unsupervised graph association for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[60] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, 2016.

[61] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, 2019.

[62] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015.

[63] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *TIP*, 2018.

[64] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

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

[66] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *TPAMI (DOI 10.1109/TPAMI.2018.2886878)*, 2018.

[67] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jianhuang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019.

[68] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 2016.

[69] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[70] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017.

[71] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, 2019.

[72] Yizhe Zhang, Ming Shao, Edward K Wong, and Yun Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *CVPR*, 2013.

[73] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *CVPR*, 2018.

[74] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen,

Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *CVPR*, 2018.

[75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[76] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. In *arXiv preprint arXiv:1610.02984*, 2016.

[77] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[78] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, 2018.

[79] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.

[80] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *arXiv preprint arXiv:1908.00485*, 2019.

[81] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[82] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 2017.

[83] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *ICCV*, 2013.

[84] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, 2014.