Joint Semantic and Latent Attribute Modelling for Cross-Class Transfer Learning

Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Tiejun Huang

Abstract—A number of vision problems such as zero-shot learning and person re-identification can be considered as cross-class transfer learning problems. As mid-level semantic properties shared cross different object classes, attributes have been studied extensively for knowledge transfer across classes. Most previous attribute learning methods focus only on human-defined/nameable semantic attributes, whilst ignoring the fact there also exist undefined/latent shareable visual properties, or latent attributes. These latent attributes can be either discriminative or non-discriminative parts depending on whether they can contribute to an object recognition task. In this work, we argue that learning the latent attributes jointly with user-defined semantic attributes not only leads to better representation but also helps semantic attribute prediction. A novel dictionary learning model is proposed which decomposes the dictionary space into three parts corresponding to semantic, latent discriminative and latent background attributes respectively. Such a joint attribute learning model is then extended by following a multi-task transfer learning framework to address a more challenging unsupervised domain adaptation problem, where annotations are only available on an auxiliary dataset and the target dataset is completely unlabelled. Extensive experiments show that the proposed models, though being linear and thus extremely efficient to compute, produce state-of-the-art results on both zero-shot learning and person re-identification.

Index Terms—Attribute Learning, Dictionary Learning, Multi-task Learning, Zero-shot Learning, Person Re-identification, Transfer Learning

---◆---

1 INTRODUCTION

A recent endeavour of computer vision research is to scale the visual recognition problem to large number of classes. This is made possible by the emergence of large-scale datasets such as ImageNet [1] and the advances in deep learning techniques [2], [3].

However, scalability remains an issue. This is because most existing recognition models are based on supervised learning and require sufficient training samples to be collected and annotated for each class. However, beyond daily objects, collecting image samples for rare and fine-grained object classes is difficult even with the modern image search engines. Zero-shot learning (ZSL) [4] and person re-identification (Re-ID) [5] are two such tasks at object category and instance level respectively. More concretely, ZSL aims at recognizing images from testing/unseen classes by learning from a set of different training/seen classes. While the objective of person Re-ID is to match people from one camera view (probe) to another camera view (gallery) under the condition that the persons in the probe and galley do not appear in the training data. Considering the people's identities as classes, the main challenge of both tasks is that the training classes and testing classes do not have

any intersection. Hence, the key problem is how to transfer knowledge learned from a set of labelled (seen) training classes to another set of testing classes without any training samples, namely the cross-class transfer learning problem.

One of the most popular approaches to cross-class transfer learning is attribute learning. Attributes are properties of visual objects that can be shared across different object classes. They can thus act as a bridge between the training/seen classes and testing/unseen classes for knowledge transfer. Most existing works define attributes as semantic, i.e., nameable properties which are annotated based on a user-defined ontology [4], [6], [7], [8], [9], [10]. Once being modelled from the seen classes, they can be used to recognise unseen classes at both the class and instance levels. For the ZSL problem, each unseen class is given an attribute "prototype" [11] and then an instance can be recognised by comparing the prototypes with the predicted attributes. In contrast, for the person Re-ID, the goal of attribute modelling is to compute a mid-level representation from low-level features [12].

Earlier attribute learning works [4], [13] learn a binary attribute classifier for each attribute separately and independently, whilst ignoring the existence of correlations among them, e.g., "female" and "long-hair" are correlated. This has been rectified by recent approaches [6], [7], [8], [9], [10], [14], [15] which jointly learn multiple attributes together with the object class labels in order to exploit their correlations. However, these joint modelling approaches focus on the user-defined semantic attributes only, whilst ignoring the

P. Peng, Y. Tian and T. Huang are with the National Engineering Laboratory for Video Technology, School of EE&CS, Peking University, Beijing, China and Cooperative Medianet Innovation Center, China; T. Xiang is with School of EECS, Queen Mary University of London, UK; M. Pontil is with Istituto Italiano di Tecnologia and University College London, UK. Y. Wang is with School of Information and Electronics, Beijing Institute of Technology, Beijing, China. Corresponding to Y. Tian (email: yhtian@pku.edu.cn).

Fig. 1: Some examples of user-defined semantic attributes and latent attributes. The user-defined semantic attributes are shown in the first row: (a) black jacket and (b) jeans. While those in the second row are two discriminative latent attributes learned by the proposed method: (c) white shirt with an open darker jacket and (d) white logo on the chest of a top. We can see that these latent attributes are often semantically meaningful and interpretable, but in a more subtle way, and may have been ignored by human annotators. It is clear that person A and person B cannot be distinguished by only two user-defined semantic attributes. But when complemented by latent attributes, it becomes easier.

factors that (1) semantic attributes are often not exhaustively defined; and (2) there are also other shareable but not nameable/semantic properties. These properties are termed *latent attributes* and have been studied as alternatives to the semantic attributes [7], [16], [17], [18], [19], [20], [21]. In this work, we argue that semantic and latent attributes are complementary to each other and jointly explain away the visual data; they thus should be jointly modelled.

Jointly learning semantic and latent attributes is useful for both more accurately predicting semantic attributes for ZSL and learning a more discriminative mid-level representation for instance-level object recognition. This is due to two reasons: First, these semantic and latent attributes can be discriminative thus useful for object recognition. For example, Fig. 1 shows that a limited list of user-defined semantic attributes are often inadequate for instance-level object recognition, in this case attribute-based person Re-ID [22]. However, when a set of complementary, interpretable and discriminative latent attributes are learned to augment the user-defined semantic attributes, recognition can be made easier. Second, even if predicting the user-defined attributes is the only goal, discovering and learning these latent attributes is still useful – it makes sure that shareable properties irrelevant to the user-defined attributes are accounted for in the model rather than acting as a distractor to corrupt the learned semantic attribute predictor.

In this work, two types of latent attributes are considered: those that are correlated to class labels thus potentially useful for object recognition, and
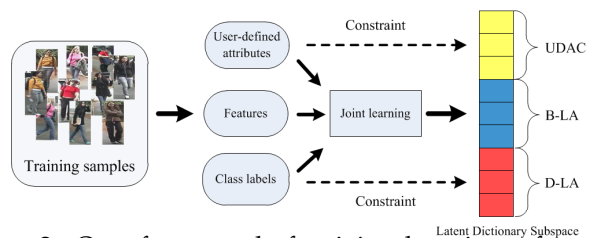

Fig. 2: Our framework for joint learning of user-defined-attribute-correlated (UDAC), discriminative latent attribute (D-LA), and background latent attribute (B-LA) dictionary subspace. Class labels are person identities in the Re-ID problem.

those that are not. The former is called discriminative latent attributes (D-LAs), and the latter background latent attributes (B-LAs) which could literally be background that may appear in any object classes. The primary goal of learning D-LAs is to compute discriminative representation useful for recognition. However, when learned jointly with user-defined semantic attributes, they often correspond to complementary and interpretable visual properties (see Fig. 1). Although B-LAs are useless for the targeted recognition task, they have to be modelled so as to avoid corrupting the other types of useful attributes. This makes our approach fundamentally different from the existing latent attribute modelling approaches which are only interested in discovering discriminative latent attributes [7], [16], [17], [18], [19], [20], [21].

To jointly learn both types of latent attributes as well as semantic attributes together with their correlations with the class labels, we propose a novel dictionary learning model. Given the features of training samples, the learned dictionary subspace is decomposed into three parts: (1) the user-defined-attribute-correlated (UDAC) dictionary subspace part which is correlated to the user-defined attribute annotations, (2) the D-LA dictionary subspace part that is subject to the class label correlation constraint to make sure that it is discriminative, and (3) the B-LA dictionary subspace part that only helps data reconstruction and is subject to no constraint. Note that since a dictionary learning model aims to reconstruct the original signal using all dictionary atoms together, it naturally enforces the learned three different types of attributes to be complementary to each others. Figure 2 illustrates the proposed dictionary learning framework.

Cross-class transfer learning aims to reduce or remove the need for both data collection and annotation. No sample collection for the testing classes is necessary; however, data from the training classes still need to be annotated in the form of class labels and semantic attributes. The joint attribute learning model depicted in Fig. 2 can be trained without semantic attributes (i.e., class labels only). Nevertheless for instance recognition problems such as Re-ID, there exist multiple domains and labelling the training sample class labels for each domain is labor-intensive [23],

[24], [25], [26]. It is thus much desirable if attribute learning can be carried out in an unsupervised manner, i.e., neither attributes nor class labels are required for the training samples.

This can be achieved by unsupervised domain adaptation or cross-dataset transfer learning [27], [28], [29], [30], [31]. More specifically, we assume that the target domain/dataset consists of a set of unlabelled training class samples and a set of test samples of different classes. In addition, a set of auxiliary domain datasets are available which are annotated with class labels and (optionally) semantic attributes. The task is to transfer knowledge from both the labelled auxiliary datasets and the unlabelled target training set to the target test set. To this end, assume all data are represented by a fixed/pre-computed features and we extend the proposed joint attribute learning model by following an asymmetric multi-task learning framework, where each dataset corresponds to a task. We assume that the UDAC and some D-LA dictionary atoms are dataset-independent, thus shared across different datasets, whilst the B-LA atoms are dataset-specific and unique to each dataset. Moreover, different decompositions of dictionaries are introduced for the auxiliary and target datasets respectively to reflect the fact that our multi-task learning model is asymmetric, i.e., it only aims to benefit the target task. The key strength of our model, which also distinguishes it from existing multi-task learning methods [32], is that it is able to learn from unlabelled target data. This is because our model is based on dictionary learning, which is originally designed for unsupervised learning and can thus be naturally reformulated for unsupervised transfer learning.

2 RELATED WORK

Learning latent attributes The idea of discovering and modelling latent attributes has been exploited before [7], [16], [17], [18], [19], [20], [21]. However, in theses studies, latent attributes are not learned jointly and thus are not necessarily complementary to user-defined attributes. There exist a few exceptions which learn discriminative latent attributes and user-defined ones jointly [15], [33], [34]. In [15], the discovered latent attributes are correlated with user-defined attributes and used to predict the latter. In contrast, our model aims to learn the latent attributes that are complementary to the user-defined ones, thus can be combined together for richer representation. Similar to ours, the method in [33] learns complementary user-defined and latent attributes jointly. However, it requires training samples of test classes; whilst our model does not need these training samples, making it more scalable. The method in [34] uses a generative topic model, and it is ideal for utilising human prior knowledge about the problem domain but weak on learning a discriminative representation for object recognition. In contrast, our dictionary learning based

model is discriminative and as a linear subspace model is easier to compute, compared with [34] which has to make approximations via variational inference to make it tractable.

Zero-shot learning Existing zero-shot learning (ZSL) methods differ in the semantic spaces used to embed the seen and unseen classes. Most methods use semantic attributes [9], [4], [35] and word vectors [36], [37], [38], or a combination of both [39], [40], [41]. In this work, we show that with the attribute space only, state-of-the-art ZSL performance can be achieved. Given a semantic embedding space, existing ZSL methods fall into two types: semantic embedding (SE) based methods and semantic relatedness (SR) based methods [40]. The SE-based methods first map the input image representations to the semantic space and then determine the class labels in the space by searching for the nearest class prototypes [4], [8], [9], [10], [39], [42], [43], [44]. While the SR-based methods first learn to measure the visual similarity between a testing unseen class sample and the seen classes, which is then compared with the semantic similarity computed as the class prototype distance between an unseen class and all seen classes [38], [40], [45], [46]. Similar to a recent work [47], our model fuses the SE- and SR-based approaches seamlessly by modelling both semantic and latent attributes. Specifically, SE-based ZSL is performed with the learned semantic attributes and SR-based ZSL with the D-LA; the final result is obtained by score-level fusion.

Attribute-based person Re-ID Semantic attributes have been exploited as a mid-level representation for Re-ID [22], [48], [49], [50]. However, none of these methods is competitive on benchmark datasets. This is because (1) the user-defined attribute representations have very low dimensions (dozens vs. tens of thousands, e.g., in [51]); and (2) no latent attributes are exploited. Recently, user-defined attributes and low-level features are modelled jointly in [52] for Re-ID. However, the user-defined attributes are predicted independently and no latent attributes are used. In contrast, our model not only models both types of attributes, but also is flexible in that discriminative latent attributes can still be learned when no user-defined attribute annotations are available. Another relevant work is [12] which deploys a generative model to transfer attribute annotations from auxiliary data (fashion clothing) to the target data (surveillance video). Again, as a generative model, it is weak in learning discriminative representation.

Unsupervised domain adaptation for person Re-ID Recently, unsupervised cross-dataset transfer learning or domain adaptation has been attempted for Re-ID in the hope that labelled data from auxiliary datasets/domains can provide transferable identity-discriminative information for a target dataset. Note that this problem is very different from the same-dataset unsupervised cross-class problems in some

early works [53], [54]. When both the dataset/domain and the identities are different, the transfer learning problem considered in this work is much harder. Among the existing cross-dataset transfer learning works, [23] adopted an SVM multi-kernel learning transfer strategy, and both [24] and [25] employed multi-task metric learning models. All theses works are supervised and need labelled data in the target dataset. As far as we know, the only existing unsupervised cross-dataset transfer learning model for Re-ID is [26], which utilised cross-domain ranking SVMs. However, an SVM-based model can not learn from completely unlabelled data. As a result, their target dataset is not exactly unlabelled because negative data are given for the target dataset. Therefore, strictly speaking, the model in [26] is a weakly-supervised rather than an unsupervised model. In contrast, our model is completely *unsupervised* without requiring any labelled data from the target dataset.

Dictionary learning Beyond attribute learning, dictionary learning [55], [56] has been studied extensively. Originally designed for unsupervised learning, it has been extended to supervised learning tasks such as face recognition [57] and person Re-ID [58], [59], [60]. Our model is related to these models in that discriminative latent attributes are learnt through the dictionary subspace. However, only our model is able to additionally learn user-defined attributes and background latent attributes for better representation.

Contributions Our contributions are summarised as follows: (1) A unified framework for learning both user-defined semantic attributes and discriminative latent attributes is proposed. (2) We further develop a novel dictionary learning model which decomposes the learned dictionary subspace into three parts corresponding to the semantic, discriminative latent as well as background latent attributes respectively. An efficient optimisation algorithm is also formulated. (3) The proposed method is extended to a novel asymmetric multi-task learning framework to address the unsupervised cross-dataset transfer learning problem when the target dataset is completely unlabelled. Extensive experiments are carried out on benchmark ZSL and person Re-ID datasets. The results show that our method generates state-of-the-art results on both tasks. Preliminary versions of this work on modelling semantic and latent attributes jointly in a same domain dataset [61] and learning a discriminative representation by unsupervised domain adaption [62] have been presented. Compared with them, this study presents a unified framework that tackles both problems and thus yields better performance.

3 METHODOLOGY

3.1 Joint attributes modelling

Assume that a set of training data are given which are labelled with some user-defined (semantic) attributes[1] and (training) classes. We aim to model user-defined semantic and latent attributes jointly so that they can be predicted on test samples which belong to different (test) classes. Adopting a dictionary learning model, the learned dictionary is decomposed into three parts corresponding to three different types of attributes (see Fig. 2): (1) D^u corresponding to the user-defined-attribute-correlated (UDAC) sub-dictionary, (2) D^d corresponding to the discriminative latent attributes (D-LA) sub-dictionary which is correlated to the class labels and (3) D^b corresponding to the background latent attributes (B-LA) sub-dictionary which captures all the residual information in the training data. D^b is uncorrelated to either user-defined attributes or class labels and thus is learned without any supervision.

Formally, let $Y \in \mathbb{R}^{m \times n}$ be a data matrix for n training samples, where each column y_i corresponds to an m-dim feature vector representing the ith data. A is a $p \times n$ matrix where each column $a_i \in \{0,1\}^p$ indicates the absence or presence of all p binary user-defined attributes[2]. The proposed joint attributes model is formulated as a regularised dictionary learning model:

$$
\begin{aligned}
&\left[D^u, D^d, D^b, W\right] = \arg\min \\
&\left\| Y - D^u X^u - D^d X^d \right\|_F^2 + \left\| Y - D^u X^u - D^d X^d - D^b X^b \right\|_F^2 \\
&+ \alpha \sum_{i,j=1}^{n} m_{i,j} \left\| x_i^d - x_j^d \right\|^2 + \beta^2 \left\| X^u - WA \right\|_F^2 \\
&s.t. \ \left\| d_i^u \right\|_2^2 \le 1, \ \left\| d_i^d \right\|_2^2 \le 1 \ \left\| d_i^b \right\|_2^2 \le 1, \ \left\| w_i \right\|_2^2 \le 1 \ \forall i,
\end{aligned}
\tag{1}
$$

where matrices X^u, X^d and X^b are codes/coefficients corresponding to sub-dictionaries D^u, D^d and D^b respectively; W represents the linear transformation between the codes obtained using D^u and the user-defined attribute annotation matrix A; d_i^u, d_i^d, d_i^b and w_i are the ith columns of D^u, D^d, D^b and W respectively; x_i^d is the ith column of X^d; α and β are free parameters controlling the strengths of two regularisation terms to be explained later; $||.||_F$ denotes the Frobenious norm of a matrix; $m_{i,j}$ is an element of an affinity matrix M indicating the class-relationships (same/different class) among different training samples. Specifically, $m_{i,j} = 1$ if x_i^d and x_j^d are of same class, and $m_{i,j} = 0$ otherwise.

There are four terms of three categories in the cost function which are now explained in detail:
(1) The first two terms are reconstruction errors that make sure the learned dictionaries can encode the data matrix Y well. Note that the two reconstruction

---

1. We will show later that the requirement on the availability of user-defined attributes can be removed.
2. Continuous attribute vectors can also be used here.

error terms are stepwise ordered. Specifically, the minimisation of the first reconstruction error term enables D^u and D^d to encode Y as much as possible, while the minimisation of the second reconstruction error term enables D^b to encode and align the residual part of Y that cannot be coded using D^u and D^d. This stepwise formulation is important to prevent the background latent attribute dictionary D^b from dominating the reconstruction error and consequently leading to trivial solutions for D^u and D^d.

(2) The third term can be rewritten as

$$\sum_{i,j=1}^{n} m_{i,j} \left\| x_i^d - x_j^d \right\|^2 = \mathrm{Tr}(X^d L X^{d\prime}), \qquad (2)$$

where $L = Q - M$ and Q is a diagonal matrix whose diagonal elements are the sums of the row elements of M. It is thus a graph Laplacian regularisation term dictating that the projections of columns of Y in the D-LA subspace, i.e., X^d, are close to each other if the corresponding data points belong to the same class. This term is thus to make the D-LA subspace, parametrised by D^d, to be discriminative (class-dependent).

(3) The last term is the constraint for learning the U-DAC subspace part. Note that we attempt to establish a linear transformation W between the projection in that subspace, X^u, and user-defined attribute annotations A, rather than simply setting them to be equal, i.e., $X^u = A$. This is because each learnt dictionary atom is additive, that is, each data point y_i is computed/reconstructed as a weighted sum of these atoms. Thus directly setting each user-defined attribute as a dictionary atom is inappropriate – the relationship between different user-defined attributes is complicated. For example, in Fig. 1(a), summing "long hair" and "trousers" to explain away the person's appearance makes sense but adding them with "female" makes less sense. Thus introducing the linear transformation overcomes this problem and also implicitly performs attribute selection.

Unlike conventional dictionary learning for sparse coding, our model has no l_1-norm sparsity penalty term on the code matrices X^u, X^d and X^b. Empirically, we find that less-sparse codes contain richer information thus more suitable to be used as representation for recognition. Moreover, removing these l_1-norm terms leads to a simpler optimisation problem.

Note that the proposed method can still work without the user-defined attribute annotations A in the training data. In this case, D^u, W and X^u will be dropped and only D-LA and B-LA are learned:

$$\left[D^u, D^b \right] = \arg\min \left\| Y - D^d X^d \right\|_F^2 +$$
$$\left\| Y - D^d X^d - D^b X^b \right\|_F^2 + \alpha \sum_{i,j=1}^{n} m_{i,j} \left\| x_i^d - x_j^d \right\|^2, \quad (3)$$
$$s.t. \ \left\| d_i^d \right\|_2^2 \leq 1 \ \left\| d_i^b \right\|_2^2 \leq 1, \ \forall i.$$

3.2 Unsupervised domain adaptation

The proposed joint attribute modelling approach can work without user-defined attribute annotation, but the training samples' class labels have to be provided which still hinders its scalability for certain cross-class transfer learning tasks such as person Re-ID. Here we consider an extension of the model for a more challenging unsupervised domain adaptation setting. Under this setting, the training samples are completely labelled. However, we assume a set of auxiliary datasets/domains exist, which are labelled and related to the target dataset/domain consisting of the unlabelled training data and test data of different classes. Here the auxiliary and target datasets should share user-defined attributes if annotated and some D-LAs. This condition is easily satisfiable, e.g., in person Re-ID, different domains are different camera networks which contain different person identities but can be described using a same set of attributes.

We adopt a multi-task learning approach because it is widely used for cross-datasets and additive models such as dictionary learning are naturally suited [63], although no attempt has been made for unsupervised domain adaptation. Here we consider learning the attribute model for each dataset as a task. We wish to learn all tasks jointly so that they can benefit each other. Importantly, since we are only concerned with the target dataset, the multi-task model is asymmetric and biased towards the target dataset.
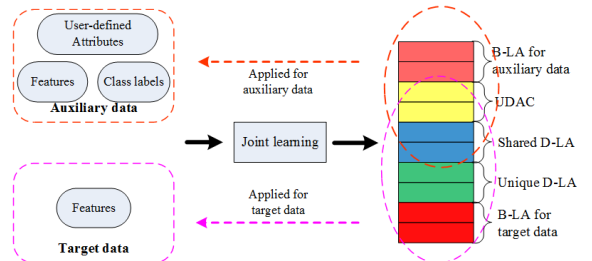

Fig. 3: A schematic illustration of the proposed multi-task dictionary learning model.

Formally, assume $Y_t \in \mathbb{R}^{m \times n_t}$ is a data matrix with each column $y_{t,i}$ corresponding to an m-dim feature vector in the dataset t ($t = 1, \cdots, T$) consisting of n_t samples. Let T be the index for the target dataset and the others are auxiliary datasets. Here the dictionary is decomposed into two types of sub-dictionaries: those shared between tasks/datasets, i.e., UDAC and D-LA, and the task-specific ones that capture dataset-unique aspects, i.e., B-LA. Moreover, the decomposition should be different for the auxiliary and target datasets to reflect that we only care about the target one. Based on these considerations, four types of sub-dictionaries are modelled: (1) D^u corresponding to the UDAC sub-dictionary which is shared by all datasets, (2) D^{ds} corresponding to the D-LA sub-dictionary which is also shared by all datasets, (3) D^{du} corresponding to the D-LA sub-dictionary which is unique

to the target dataset, and (4) D_t^b corresponding to the B-LA which is task-specific. Note that the auxiliary and target datasets are treated differently: For the target dataset, an additional dictionary D^{du} is needed to account for D-LA unique to the target dataset, making the decomposition biased towards the target dataset. Different components of our model and their relationships are illustrated Fig. 3 and the formulation of the model is:

$$\left[D^u, D^{ds}, D^{du}, D_1^b, \cdots, D_T^b, W \right] = \arg\min$$

$$\sum_{t=1}^{T-1} F_{data}^A(D^u, D^{ds}, D_t^b) + F_{D-LA}^A(X_t^{ds}) + F_{UDAC}^A(X_t^u, W)$$
$$+ F_{data}^T(D^u, D^{ds}, D^{du}, D_T^b) + F_{D-LA}^T(X_T^{ds}, X_T^{du}) + F_{UDAC}^T(X_T^u),$$

$$s.t. \ \|d_i^u\|_2^2 \le 1, \ \left\|d_i^{ds}\right\|_2^2 \le 1, \ \left\|d_i^{du}\right\|_2^2 \le 1, \ \left\|d_{t,i}^b\right\|_2^2 \le 1,$$
$$\|w_i\|_2^2 \le 1, \ \forall i, t. \tag{4}$$

Similar to (1), matrices X_t^u, X_t^{ds}, X_T^{du} and X_t^b are codes/coefficients corresponding to dictionaries D^u, D^{ds}, D^{du} and D_t^b respectively. $F_{data}^A(\cdot)$ and $F_{data}^T(\cdot)$ are the data reconstruction terms for the auxiliary and target datasets respectively and they are defined as:

$$F_{data}^A(D^u, D^{ds}, D_t^b) =$$
$$\left\| Y_t - D^u X_t^u - D^{ds} X_t^{ds} \right\|_F^2 + \left\| Y_t - D^u X_t^u - D^{ds} X_t^{ds} - D_t^b X_t^b \right\|_F^2,$$
$$F_{data}^T(D^u, D^{ds}, D^{du}, D_T^b) =$$
$$\left\| Y_T - D^u X_T^u - D^{ds} X_T^{ds} - D^{du} X_T^{du} \right\|_F^2$$
$$+ \left\| Y_T - D^u X_T^u - D^{ds} X_T^{ds} - D^{du} X_T^{du} - D_T^b X_T^b \right\|_F^2. \tag{5}$$

$F_{D-LA}^A(\cdot)$ and $F_{D-LA}^T(\cdot)$ are the constraints of D-LA for the auxiliary and target datasets respectively:

$$F_{D-LA}^A(X_t^{ds}) = \alpha \sum_{i,j=1}^{N_t} m_{t,i,j} \left\| x_{t,i}^{ds} - x_{t,j}^{ds} \right\|^2,$$
$$F_{D-LA}^T(X_T^{ds}, X_T^{du}) = \tag{6}$$
$$\alpha \sum_{i,j=1}^{N_T} m_{T,i,j} (\left\| x_{T,i}^{ds} - x_{T,j}^{ds} \right\|^2 + \left\| x_{T,i}^{du} - x_{T,j}^{du} \right\|^2).$$

M_t is an affinity matrix among different training samples in dataset t. Specifically, for the labelled auxiliary datasets, $m_{t,i,j} = 1$ if $y_{t,i}$ and $y_{t,j}$ are of the same class and $m_{t,i,j} = 0$ otherwise. For the unlabelled target dataset, M_T is initialised as a zero matrix and updated iteratively as explained later.

$F_{UDAC}^A(\cdot)$ and $F_{UDAC}^T(\cdot)$ are the constraints of U-DAC for the auxiliary and target datasets respectively. Since the user-defined attribute annotations are only available in the auxiliary datasets, we utilise the affinity matrix in F_{UDAC}^T, as follows:

$$F_{UDAC}^A(X_t^u, W) = \beta^2 \|X_t^u - WA_t\|_F^2,$$
$$F_{UDAC}^T(X_T^u) = \alpha \sum_{i,j=1}^{N_T} m_{T,i,j} \left\| x_{T,i}^u - x_{T,j}^u \right\|^2. \tag{7}$$

Note that the proposed method can again work without the user-defined attribute annotations A_t for

the auxiliary datasets. In this case, D^u, W and X_t^u will be dropped and only D^{ds}, D^{du} and D^b are learned.

3.3 Optimisation

Since (1) can be solved as a special case of (4) when $T = 1$, we only detail how the the optimisation problem in (4) is solved. Substituting (5), (6) and (7) to (4), the solution is achieved by solving the following subproblems iteratively:

(1) Computing X_t^{ds}, X_T^u and X_T^{du} Given fixed D^u, D^{ds}, D^{du}, D_t^b and X_t^b, the coding problem of the task t ($t = 1, \cdots, T$) becomes:

$$\min \left\| \tilde{Y}_t - \tilde{D}\tilde{X}_t \right\|_F^2 + \alpha \text{Tr}(\tilde{X}_t L_t \tilde{X}_t'), \tag{8}$$

where for the target dataset:

$$\tilde{Y}_T = \left[\begin{array}{c} Y_T \\ Y_T - D_T^b X_T^b \end{array} \right], \tilde{D} = \left[\begin{array}{c} D^u, D^{ds}, D^{du} \\ D^u, D^{ds}, D^{du} \end{array} \right], \tilde{X}_T = \left[\begin{array}{c} X_T^u \\ X_T^{ds} \\ X_T^{du} \end{array} \right],$$

and for the auxiliary dataset:

$$\tilde{Y}_t = \left[\begin{array}{c} Y_t - D^u X_t^u \\ Y_t - D^u X_t^u - D_t^b X_t^b \end{array} \right], \tilde{D} = \left[\begin{array}{c} D^{ds} \\ D^{ds} \end{array} \right], \tilde{X}_t = \left[X_t^{ds} \right].$$

L_t is calculated as similar as (2). Let the derivative of (8) equal to 0 and the analytical solution of $\tilde{x}_{t,i}$ (the ith column of $\tilde{X}_t$) can be obtained as:

$$\tilde{x}_{t,i} = \left(\tilde{D}'\tilde{D} + 2\lambda l_{t,i,i} I \right)^{-1} \left(\tilde{D}'\tilde{y}_{t,i} - 2\alpha \sum_{k \neq i} \tilde{y}_{t,k} l_{k,i} \right),$$

where $l_{t,k,i}$ is the (k, i) element of L_t, I is the identity matrix and $\tilde{x}_{t,i}$ is the ith column of $\tilde{X}_t$.
(2) Computing X_t^u, ($t = 1, \cdots, T-1$) Fix other terms, X_t^u is solved as:

$$\min \left\| \tilde{Y}_t - \tilde{D}X_t^u \right\|_F^2, \tag{9}$$

where

$$\tilde{Y}_t = \left[\begin{array}{c} Y_t - D^{ds} X_t^{ds} \\ Y_t - D^{ds} X_t^{ds} - D_t^b X_t^b \\ \beta W A_t \end{array} \right], \tilde{D} = \left[\begin{array}{c} D^u \\ D^u \\ \beta I \end{array} \right].$$

I is the identity matrix. Let the derivative of (9) equal to 0 and the analytical solution of X_t^u is:

$$X_t^u = \left(\tilde{D}'\tilde{D} \right)^{-1} \tilde{D}'\tilde{Y}_t. \tag{10}$$

(3) Computing X_t^b Fix other terms, X_t^b is solved as:

For the target dataset:

$$\min \left\| Y_T - D^u X_T^u - D^{ds} X_T^{ds} - D^{du} X_T^{du} - D_T^b X_T^b \right\|_F^2,$$
and for the auxiliary datasets: $\qquad (11)$
$$\min \left\| Y_t - D^u X_t^u - D^{ds} X_t^{ds} - D_t^b X_t^b \right\|_F^2.$$

Let the derivative of (11) equal to 0 and the analytical solution of X_t^b can be obtained as:

For the target dataset:

$$X_T^b = \left(D_T^{b'} D_T^b \right)^{-1} D_T^{b'} \left(Y_T - D^u X_T^u - D^{ds} X_T^{ds} - D^{du} X_T^{du} \right),$$

and for the auxiliary datasets:

$$X_t^b = \left(D_t^{b'} D_t^b \right)^{-1} D_t^{b'} \left(Y_t - D^u X_t^u - D^{ds} X_t^{ds} \right).$$

(4) Updating sub-dictionaries When other terms are given, D^u is optimised as:

$$\min \|\mathcal{Y} - D^u \mathcal{X}\|_F^2, \quad s.t. \ \|d_i^u\|_2^2 \le 1, (\forall i), \tag{12}$$

where

$$\mathcal{Y} = [Y_1 - D^{ds}X_1^{ds}, ..., Y_{T-1} - D^{ds}X_{T-1}^{ds}, Y_1 - D^{ds}X_1^{ds} - D_1^b X_1^b, ..., Y_{T-1} - D^{ds}X_{T-1}^{ds} - D_{T-1}^b X_{T-1}^b, Y_T - D^{ds}X_T^{ds} - D_T^{du}X_T^{du}, Y_T - D^{ds}X_T^{ds} - D_T^{du}X_T^{du} - D_T^b X_T^b]$$
$$\mathcal{X} = [X_1^u, ..., X_{T-1}^u, X_1^u, ..., X_{T-1}^u, X_T^u, X_T^u]. \tag{13}$$

(12) can be optimised by the Lagrange dual, and the analytical solution of D^u can be computed as $D^u = (\mathcal{Y}\mathcal{X}')(\mathcal{X}\mathcal{X}' + \Lambda)^{-1}$, where Λ is a diagonal matrix constructed from all the dual variables.

Similar to D^u, D^{ds} is optimised by:

$$\min \left\|\mathcal{Y} - D^{ds} \mathcal{X}\right\|_F^2, \quad s.t. \ \left\|d_i^{ds}\right\|_2^2 \le 1, (\forall i), \tag{14}$$

where

$$\mathcal{Y} = [Y_1 - D^u X_1^u, ..., Y_{T-1} - D^u X_{T-1}^u, Y_1 - D^u X_1^u - D_1^b X_1^b, ..., Y_{T-1} - D^u X_{T-1}^u - D_{T-1}^b X_{T-1}^b, Y_T - D^u X_T^u - D_T^{du}X_T^{du}, Y_T - D^u X_T^u - D_T^{du}X_T^{du} - D_T^b X_T^b]$$
$$\mathcal{X} = [X_1^{ds}, ..., X_{T-1}^{ds}, X_1^{ds}, ..., X_{T-1}^{ds}, X_T^{ds}, X_T^{ds}]. \tag{15}$$

Then, for the target dataset, fix other terms and D^{du} can be updated by:

$$\min \left\|\mathcal{Y}_T - D^{du} \mathcal{X}_T\right\|_F^2, \quad s.t. \ \left\|d_i^{du}\right\|_2^2 \le 1, (\forall i), \tag{16}$$

where

$$\mathcal{Y}_T = \left[Y_T - D^u X_T^u - D^{ds}X_T^{ds}, Y_T - D^u X_T^u - D^{ds}X_T^{ds} - D_T^b X_T^b\right],$$
$$\mathcal{X}_T = \left[X_T^{du}, X_T^{du}, X_T^{du}\right]. \tag{17}$$

At last, the objective function to solve D_t^b $(t = 1, ..., T)$ is:

$$\min \left\|\mathcal{Y}_t - D_t^b X_t^b\right\|_F^2, \quad s.t. \ \left\|d_{t,i}^b\right\|_2^2 \le 1 (\forall i), \tag{18}$$

where

For the target dataset:
$$\mathcal{Y}_T = Y_T - D^u X_T^u - D^{ds}X_T^{ds} - D^{du}X_T^{du},$$
and for the auxiliary datasets:
$$\mathcal{Y}_t = Y_t - D^u X_t^u - D^{ds}X_t^{ds}. \tag{19}$$

(14), (16) and (18) can be solved similarly as (12).

(5) Updating W Similar to the dictionary updating in Step 3, fix other variables and solve W by:

$$\min \|\mathcal{X} - W\mathcal{A}\|_F^2, \quad s.t. \ \|w_i\|_2^2 \le 1 (\forall i). \tag{20}$$

where

$$\mathcal{X} = [X_1^u, ..., X_{T-1}^u], \mathcal{A} = [A_1, ..., A_{T-1}]. \tag{21}$$

The analytical solution of W is given by: $W = (\mathcal{X}\mathcal{A}')(\mathcal{A}\mathcal{A}' + \Lambda)^{-1}$, where Λ is a diagonal matrix constructed from all the dual variables.

Alg. 1 summarizes our algorithm. It converges after a few (< 50) iterations in our experiments.

Algorithm 1: Unsupervised domain adaptation algorithm for attribute learning

Input: Y_t; initialize D^u, D^{ds} D^{du} and D_t^b randomly; $X_t^b \to \mathbf{0}$;
Output: D^u, D^{ds}, D^{du}, D_t^b, D^u and W $(t = 1, ..., T)$.
while *Non-convergence* **do**
  for $t = 1 \to T$ **do**
    if *Auxiliary dataset* **then**
      Compute code X_t^{ds} by (8).
      Compute code X_t^u by (9).
    if *Target dataset* **then**
      Compute code X_T^{ds}, X_T^u and X_T^{du} by (8).
    Compute code X_t^b by(11).
  Update D^u by (12).
  Update D^{ds} by (14).
  for $t = 1 \to T$ **do**
    if *Auxiliary datasets* **then**
      Update D_t^b by (18).
    if *Target dataset* **then**
      Update D^{du} by (16).
      Update D_t^b by (18).

Iterative Updating M_T After running Alg. 1, each training sample $y_{T,i}$ from the target dataset can be estimated (using (22) to be detailed below) as a code vector. With this code, we can measure the similarity between each pair of target data samples and recompute M_T. This matrix now captures the soft relationship among the training samples in the target dataset. Specifically, if $x_{T,j}$ is among the k-nearest-neighbors of $x_{T,i}$ and $x_{T,i}$ is also among the k-nearest-neighbors of $x_{T,j}$, $m_{T,i,j} = \frac{x_{T,i} \cdot x_{T,j}}{\|x_{T,i}\|\|x_{T,j}\|}$; otherwise, $m_{T,i,j} = 0$ ($k = 1$ in our experiments and we obtain similar results when $k < 5$). With the updated M_T, we re-run Alg. 1 to enter the next iteration. This iterative procedure stops when the value of cost function (4) converges (i.e., the difference divided by the value over consecutive iterations is smaller than 0.01), and the number of iterations is typically less than 5 in our experiments.

3.4 Attribute prediction

Once the various sub-dictionaries are learned, each test image y from an unseen test class can be encoded as $[x^u, x^d, x^b]$ via D^u, D^d and D^b respectively. The encoding problem under the unsupervised domain adaptation can be formulated as:

$$
\begin{aligned}
&\left[x^u, x^{ds}, x^{du}, x^b\right] \\
&= \arg\min \left\|y - D^u x^u - D^{ds}x^{ds} - D^{du}x^{du} - D_T^b x^b\right\|_2^2 \\
&\quad + \gamma \left(\|x^u\|_2^2 + \left\|x^{ds}\right\|_2^2 + \left\|x^{du}\right\|_2^2 + \left\|x^b\right\|_2^2\right),
\end{aligned} \tag{22}
$$

where x^u, x^{ds}, x^{du} and x^b are the projections of y using the UDAC, shared D-LA, and target dataset unique D-LA and B-LA sub-dictionaries respectively, and γ is a weight for the regularisation terms. (22) can be solved easily with a linear system. Under the

supervised setting (1), D^{du} and x^{du} will be dropped, and D^{ds} and x^{ds} will be replaced by D^{ds} and x^{ds} correspondingly.

After we obtain x^u, the user-defined attribute vector a can be predicted via the linear constraint W:

$$a = \arg\min \|x^u - Wa\|_2^2 + \gamma \|a\|_2^2. \tag{23}$$

4 APPLICATIONS

4.1 Zero-shot learning

It is a supervised cross-class transfer learning problem. Let $A^s = \{a_k^s\}_{k=1}^{K_s}$ and $A^u = \{a_k^u\}_{k=1}^{K_u}$ be semantic attribute prototypes of K_s seen classes and K_u unseen classes respectively. They can be obtained through human annotations following an attribute ontology. $L^u = \{l_k^u\}_{k=1}^{K_s}$ is the D-LA prototypes of K_s seen classes. Specifically, we first calculate the D-LA code x^d of all samples from the seen class k by (22), and then use the mean as the D-LA prototype of the seen class k.

For a test image y from an unseen class, it is represented by its semantic attribute vector a and D-LA vector x^d estimated using (23) and (22). The class label of y can be assigned using either of them:

Semantic attributes based ZSL In this semantic embedding (SE) based approach, y is recognised by comparing the prototypes $\{A_k^u\}_{k=1}^{K_u}$ with the predicted semantic attributes a by nearest neighbour search:

$$k^* = \arg\max_k \frac{a \cdot a_k^u}{\|a\|_2 \|a_k^u\|_2}. \tag{24}$$

D-LA based ZSL In this semantic relatedness (SR) based approach, we first measure the semantic relatedness between each unseen class and all the seen classes by attempting to reconstruct the unseen class prototype using all the seen class prototypes, resulting in a SR vector r_k^u for the k-th unseen class. Specifically, this SR vector is obtained by solving the following least square regression problem:

$$r_k^u = \arg\min \|a_k^u - A^s r_k^u\|_2^2 + \gamma \|r_k^u\|_2^2. \tag{25}$$

Meanwhile, the D-LA x^d of y can be used to measure the visual similarity between y and all the seen classes. The resultant visual similarity vector is obtained by solving another regression problem:

$$x^r = \arg\min \left\|x^d - L^u x^r\right\|_2^2 + \gamma \|x^r\|_2^2. \tag{26}$$

Finally, y can be recognised by comparing the visual similarity vector x^r and semantic relatedness vector r_k^u of each unseen class and assigning the class label to the one with smallest distance:

$$k^* = \arg\max_k \frac{x^r \cdot r_k^u}{\|x^r\|_2 \|r_k^u\|_2}. \tag{27}$$

Since the semantic attributes and D-LA are complementary to each other, these two approaches are combined by score-level fusion as our final approach:

$$k^* = \arg\max_k \left(\frac{a \cdot a_k^u}{\|a\|_2 \|a_k^u\|_2} + \frac{x^r \cdot r_k^u}{\|x^r\|_2 \|r_k^u\|_2} \right). \tag{28}$$

Approaches	AwA	CUB
ALE [9]	43.5	18.0
UMF [8]	48.6	18.2
CSHAP [10]	45.6	17.5
SC [47]	72.9	54.5
DAP* [4]	57.5	-
SSE* [42]	76.3	30.4
SJE* [39]	73.9	51.7
JLSE* [38]	80.5	42.1
MSS* [44]	-	56.5
LatEm* [43]	76.1	47.4
Ours	**82.9**	**57.1**

TABLE 1: Comparative results on zero-shot Learning (recognition accuracy in %). "*" means that the same VGG features are used.

4.2 Person Re-ID

For Re-iD, each test sample can be represented as the predicted semantic attributes a and D-LA x^d. Simply treating the two types of attributes as features, person Re-ID could be performed by summing the cosine distance of a and x^d respectively between the attribute vectors of a probe sample and a gallery one.

Different from ZSL, the proposed method can still work for Re-ID without user-defined attribute annotations A in the training data, based on (3). In this case, the test sample y is represented only by its D-LA x^d only. Another difference is that unsupervised domain adaptation can be performed for Re-ID, making our method applicable to unlabelled datasets/domains.

5 EXPERIMENTS

5.1 Zero-shot learning

Datasets and settings Two widely-used benchmark datasets are chosen in this experiment. **AwA** is composed of 30,475 images from 50 animal classes and each class is annotated with 85 user-defined attributes. Following the default split [4], we divide the dataset into two parts: 40 classes (24,295 images) as the seen classes for training and the remaining 10 classes (6,180 images) as unseen classes for testing. **CUB** contains 11,788 images of 200 bird classes and each class is annotated with 312 attributes. Again we use the same data split as in most other ZSL works (i.e., 150 seen classes for training and the rest 50 unseen target classes for testing). For both datasets, the 4,096-dim VGG deep features provided by [42] are used to compare other methods fairly. As for the parameter setting, the sizes of D^u, D^d and D^b are set to 400, 400 and 100 respectively for both datasets. The other free parameters, α and β in (1) and γ in (22), are set as 30, 1.5 and 3 respectively for both datasets. These parameters are all obtained using cross-validation.

Baselines We compare our method with 10 state-of-art methods. They fall into two groups. Methods in the first group project the input image representation to a semantic space, and the projections are then compared with the class labels embedded in the same space (class prototypes) for recognition. These include DAP [4], ALE [9], CSHAP [10] and UMF [8]. The

Approaches	AwA	CUB
ALE [9]	65.7	60.3
CSHAP [10]	74.3	68.7
DAP* [4]	72.8	61.8
IAP* [4]	72.1	-
TBOS* [15]	70.5	68.1
Ours	**75.8**	**78.3**

TABLE 2: Comparative results on predicting user-defined attributes.

second group of methods learn a joint embedding space for both the visual feature and semantic spaces, including SSE [42], SJE [39], JLSE [38], SC [47], LatEm [43], and MSS [44]. Most of these compared methods use the user-defined attributes as the semantic space. The only two exceptions are SJE [39] and LatEm [43], which additionally use other spaces obtained by natural langauge processing such as Word2Vec [64].

Comparative results The results are shown Table 1. All baselines use the same train-test split and most use the same features as ours. The results show that our method achieves the best performance on both datasets even though a number of the compared methods utilise additional information. For example, MSS [44] utilises visual parts annotations. SJE [39] and LatEm [43] combine the user-defined attributes with other semantic spaces. Futhermore, JLSE [38] is under the transductive setting which requires the access to the full test dataset for model adaptation.

Ablation study One of the key reasons for the superior ZSL performance is the more accurate attribute prediction. To verfiy that, the user-defined attribute prediction accuracy is reported in Table 2, measured using mean area under ROC curve (mAUC). It clearly shows that the proposed method achieves state-of-the-art performance on both datasets for attribute prediction. To find out which part of the proposed model contributes to the good performance for both attribute prediction and ZSL, we examine the contributions of the key components in our model (see (1)), including: (1) two types of latent attributes: D-LA (D^d) and B-LA (D^b) are learned together with the user-defined attributes, and (2) instead of learning the user-defined attributes directly as a sub-dictionary, we model a linear transformation (W) from the user-defined attributes A to the UDAC dictionary subspace (D^u). Table 3 compares our full model (Ours_full) with various striped-down versions. The results show that all these components contribute positively to the final performance of the model. Further evaluations on the complementarity of the semantic and latent attributes can be seen in the Supplementary Material.

	ZSL		A_Pre	
	AWA	CUB	AWA	CUB
Without D^d	81.6	50.7	73.6	74.4
Without D^b	80.2	48.9	71.2	72.6
Without W	74.4	42.6	68.9	69.8
Ours_full	**82.9**	**57.1**	**75.8**	**78.3**

TABLE 3: Evaluation on the contributions of different components to ZSL and attribute prediction (A_Pre).

5.2 Person Re-ID

5.2.1 Supervised person Re-ID

Datasets Four widely-used person re-ID benchmark datasets are chosen. **VIPeR** [65] contains 1,264 images of 632 individuals from two distinct camera views (two images per individual) featured with large viewpoint changes and varying illumination conditions. All individuals are randomly divided into two equal-sized subsets for training and testing respectively with no overlapping in identity. This random partition process is repeated 10 times, and the averaged performance is reported. For fair comparison, we use the same data splits as in [66]. **PRID** [67] consists of images extracted from two surveillance cameras. Camera view A contains 385 individuals, camera view B contains 749 individuals, with 200 of them appearing in both the two views. The single shot version and data splits of the dataset are used in our experiments as in [66]. In each data split, 100 people with one image from each view are randomly chosen from the 200 persons present in both camera views as the training set, while the remaining 100 persons of View A are used as the probe set, and the remaining 649 persons of View B are used as gallery. Experiments are repeated over the 10 splits. **CUHK03** [68] contains 13,164 images of 1,360 identities, captured from six cameras with each person only appearing in two views. It provides pedestrian bounding boxes manually labelled by human and automatically detected by a deformable-part-model (DPM) detector. We report results on both cases. The 20 training/test splits provided in [68] are used under the single-shot setting as in [51]. **Market-1501** [69] contains 32,668 detected person images of 1,501 identities. Each identity is captured by six cameras at most, and two cameras at least. We use training and test sets provided in [69], under both the single-query and multi-query evaluation settings.

Attribute annotation The training sets of all datasets have labels indicating the identities of the people. In addition, a total of 105 binary user-defined attributes have been annotated for VIPeR and PRID in [14]. We remove the user-defined attributes which do not appear in each dataset, and the numbers of the remaining attributes are 85 and 56 for VIPeR and PRID respectively. Note that attribution annotation is unavailable on CUHK03 and Market-1501. As mentioned in Sec. 3, our model works with and without the user-defined attributes. For fair comparison with the existing methods which do not use attribute annotations, we report results of our model both with and without user-defined attributes.

Features and evaluation metric Both hand-crafted low-level features and deep CNN features are considered for fair comparison with various published results. For hand-crafted feature, the 26,960-dim Local

Maximal Occurrence (LOMO) features [51] are used. In addition, deep features are used in our experiment. Specifically, we extract the 1024-dim CNN feature from the "pool5" layer in the GoogLeNet model [3] which is pre-trained on ImageNet dataset [70] and then fine-tuned on "non-target" person Re-ID datasets. For example, when extracting features on VIPeR dataset, we fine-tune the model on PRID, CUHK03 and Market datasets[3]. For evaluation, we compute Cumulated Matching Characteristics (CMC) curves. Due to space constraint as well as for easier comparison with published results, we only report the cumulated matching accuracy at selected ranks in tables rather than depicting the actual CMC curves. The only exception is the Market-1501 dataset. Since there are on average 14.8 cross-camera ground-truth matches for each query, we additionally use mean average precision (mAP) as in [69].

Parameter settings The sizes of D^u, D^d and D^b are all set to 300. We found that the performance of our model is insensitive to the dictionary size when it is between 200 to 400. The other free parameters in our model, α and β in (1) and γ in (22), are obtained using four-fold cross-validation.

Baselines Twelve state-of-the-art Re-ID methods are selected for comparison. They fall into four categories: (1) Distance metric leaning based methods: RPLM [73], Mid-level Filter [74], LADF [75], and Similarity Learning [76]; (2) Kernel-based Discriminative subspace learning methods: MFA [77], kLFDA [77], kCCA [66], XQDA [51], and MLAPG [78]; (3) Deep learning based: Improved Deep [79], Wang et al [80], FT-JSTL+DGD [81], Zhang et al [72], TCP [82] and Gated S-CNN [71]; (4) Attribute-based method: aMTL-LORAE [52], which is the most relevant to ours as it also utilises the user-defined attributes. Note that aMTL-LORAE requires multiple images of each person for training, hence they apply data augmentation to generate more training samples on VIPeR and utilises the multi-shot setting of PRID rather than the single-shot one adopted by most other methods including ours. Furthermore, aMTL-LORAE cannot work without user-defined attributes. For fair comparison, we use the same features and the same training-test splits for the compared methods whenever possible (i.e. when the code is available we use the same features as ours). Three versions of our models are evaluated: (1) "Ours_U" which means only user-defined attributes are used, (2)"Ours_L" which means only latent attributes are learned as representation without requiring user-defined attribute annotation, and (3) "Ours_All" which means both the user-defined and latent attributes are used. The used feature type is indicated in bracket, e.g., "Ours_L(L)" means low-level hand-crafted features are used.

3. It is a common practice to pretrain a deep CNN on auxiliary Re-ID datasets [71], [72].

(a) VIPeR

Rank	1	5	10	20
RPLM [73]	27.0	55.3	69.0	83.0
Mid-level [74]	29.1	52.3	65.9	79.9
Similarity Learning [76]	36.8	70.4	83.7	91.7
LADF [75]	30.2	64.7	78.9	90.4
kCCA* [66]	30.2	62.7	76.0	86.8
MFA* [77]	39.7	69.2	80.5	89.0
kLFDA* [77]	38.6	69.2	80.4	89.2
XQDA* [51]	40.0	-	80.5	91.1
MLAPG* [78]	40.7	-	82.3	92.4
NFST* [83]	**42.3**	71.5	**82.9**	**92.1**
Ours_L(L)	41.5	**73.4**	82.6	90.5
aMTL-LORAE [52]	42.3	72.2	81.6	89.6
Ours_U(L)	17.1	53.8	65.3	75.6
Ours_All(L)	44.3	74.6	84.2	92.5
Improved Deep [79]	34.8	63.6	75.6	84.5
Wang et al [80]	35.8	66.5	82.5	-
FT-JSTL+DGD [81]	38.6	-	-	-
Zhang et al [72]	43.0	-	87.3	94.8
TCP [82]	47.8	74.7	84.8	91.1
Ours_L(CNN)	**49.8**	**77.6**	**90.1**	**95.9**

(b) PRID

Rank	1	5	10	20
RPLM [73]	15.0	32.0	42.0	54.0
kCCA* [66]	14.3	37.4	47.6	62.5
MFA* [77]	22.3	45.6	57.2	68.2
kLFDA* [77]	22.4	46.5	58.1	68.6
NFST* [83]	**29.8**	**52.9**	**66.0**	**76.5**
Ours_L(L)	28.9	50.3	63.7	74.0
aMTL-LORAE [52]	18.0	37.4	50.1	66.6
Ours_U(L)	14.5	33.6	45.2	61.3
Ours_All(L)	**31.8**	**54.3**	**67.5**	**77.4**
Ours_L(CNN)	**37.5**	**64.1**	**72.6**	**81.2**

(c) CUHK03 (Manual)

Rank	1	5	10	20
XQDA* [51]	52.2	82.2	92.1	96.3
NFST* [83]	58.9	85.6	92.5	96.3
Ours_L(L)	60.8	87.5	93.3	97.2
Improved Deep [79]	54.7	86.5	93.9	98.1
Wang et al [80]	52.2	82.5	-	-
FT-JSTL+DGD [81]	75.3	-	-	-
Zhang et al [72]	57.0	85.0	93.5	97.0
Ours_L(CNN)	**77.5**	**92.4**	**96.5**	**99.2**

(d) CUHK03 (Detected)

Rank	1	5	10	20
XQDA* [51]	46.3	78.9	88.6	94.3
NFST* [83]	53.7	83.0	93.0	94.8
Ours_L(L)	**58.4**	**85.7**	**93.2**	**96.3**
Improved Deep [79]	45.0	76.0	83.5	93.2
Zhang et al [72]	51.2	80.5	90.0	94.5
Ours_L(CNN)	**64.2**	**89.1**	**93.4**	**96.1**

(e) Market-1501

Query	singleQ		multiQ	
Evaluation metrics	Rank-1	mAP	Rank-1	mAP
MFA* [77]	45.7	18.2	-	-
kLFDA* [77]	51.4	24.4	52.7	27.4
XQDA* [51]	43.8	22.2	54.1	28.4
NFST* [83]	55.4	29.9	68.0	41.9
Ours_L(L)	**61.1**	**32.3**	**73.2**	**45.8**
Gated S-CNN [71]	**65.8**	39.5	76.0	48.4
Ours_L(CNN)	65.7	**43.1**	**76.5**	**51.4**

TABLE 4: Supervised Re-ID results. '*' means that the same features are used. '-' means no reported result is available.

Comparative results From the results shown in Table 4, we have the following observations: (1) Even without using the additional attribute annotation, our method Ours_L(L) achieve better or comparable performance in comparison against the state-of-the-art alternatives using hand-crafted features. (2) Ours_All outperforms Ours_L on all datasets. This indicates that the learned user-defined attributes and discriminative latent attributes are indeed complementary to each other. (3) If only user-defined attributes are used to represent a person, Ours_U has a much weaker performance than Ours_L, suggesting that the user-defined attributes alone are limited in representing a person discriminatively. (4) Compared to the alternative attribute-based Re-ID model aMTL-LORAE, our model (Ours_All) is clearly better. In particular, our method outperforms aMTL-LORAE by a large margin even when they used more training data on PRID. In addition, aMTL-LORAE can only be applied when there are user-defined attribute annotations, whilst our model is not restricted by that. (5) Our model can be considered as a discriminative subspace learning based method. Compared with the alternative subspace learning based methods such as kLFDA [77] and XQDA [51], our model's performance is clearly superior. (6) With deep CNN features as input, the results are much better than low-level features. It indicates that our method can benefit from stronger deep features. Also, our model with the CNN feature outperforms all the recent deep Re-ID models even without additional attribute annotations (Ours_L(CNN)) on all four datasets. These results suggest that even an end-to-end deep Re-ID model can benefit from projecting the deep features into a lower dimensional latent discriminative attribute space using our model. Some visualisation of the learned D-LAs and an ablation study on the contributions of different model components can be found in the Supplementary Material.

5.2.2 Unsupervised person Re-ID

Under this setting, the training data from the target dataset is completely unlabelled. We assume there exist a number of auxiliary datasets labelled with identities and (optionally) user-defined attributes. A discriminative feature representation is then learned using both the auxiliary and target datasets using our asymmetric multi-task learning model (Sec. 3.2).

Datasets and settings The same four datasets (VIPeR, PRID, CUHK03 (manual) and Market) are used but the settings are different. Specifically, one dataset in turn is chosen as the target dataset and the other three are used as the auxiliary datasets. All the persons' identities in the auxiliary datasets are labelled and used for model training, while the target dataset is split into the training and test sets as in the supervised setting above. To compare with the reported results with other methods [84], [85] fairly, we utilise the 5132-dim hand-crafted feature [66].

Parameter settings The sizes of UDAC D^U, B-LA, shared and unique D-LA sub-dictionaries are all set to 200 in all experiments. We found that the model's performance is insensitive to the different dictionary sizes. Other parameters (α and β in (4)) are set automatically using three-fold cross-validation with one of the three auxiliary datasets as the validation set and the other two as the training set.

Baselines The compared methods can be categorised into two groups: (1) *Single-task methods*. Without transfer learning, the training data of these unsupervised methods are only the unlabelled data from the target dataset. Some state-of-the-art unsupervised Re-ID methods are selected for comparison, including the hand-crafted-feature-based method SDALF [86], the saliency-learning-based eSDC [87], the graphical-model-based GTS [88], the sparse-representation-classification-based ISR [84] and the dictionary-learning-based DLLR [85]. We also report results of the single-task version of our model by removing all auxiliary data related terms in (4), denoted as Ours_S. (2) *Multi-task methods*. There are few multi-task learning methods, or unsupervised transfer learning methods in general, available for the unsupervised setting. AdaRSVM [26] is the only unsupervised cross-data transfer learning work for person Re-ID that we are aware of. However, AdaRSVM assumes the availability of negative pairs in the target dataset, thus using more supervision than our method. We also use the subspace alignment based unsupervised domain adaptation method SA_DA [89] to align the data distributions of the auxiliary and target datasets first. Then a supervised Re-ID model, kLFDA [77], is trained on the labelled source datasets and applied to the aligned target dataset. This method is denoted as SA_DA+kFLDA. Note that as an unsupervised domain adaptation method, SA_DA assumes that the auxiliary and target domains have the same classes, which is invalid for cross-dataset transfer learning. In addition, we compare with a naive transfer approach, by learning kFLDA on the auxiliary datasets first and applying it directly to the target dataset without any model adaptation. It is denoted as kLFDA_N. Adversarial [28] is the deep unsupervised domain alignment model using gradient reversal and adversarial learning.

Comparative results Table 5 reports the results. From these results, it is evident that: (1) Compared with existing unsupervised methods including SDALF, eSDC, GTS and ISR, our model is significantly better. This shows that transfer learning indeed helps for unsupervised Re-ID. (2) The difference between the results of "Ours_S" and "Ours" models shows exactly how much the target dataset has benefited from the auxiliary datasets using our unsupervised asymmetric multi-task transfer learning method. (3) The results of kLFDA_N is very poor, showing that

(a) VIPeR

Rank	1	5	10	20
SDALF [86]	19.9	38.9	49.4	65.7
eSDC [87]	26.7	50.7	62.4	76.4
GTS [88]	25.1	50.0	62.5	75.8
ISR* [84]	27.0	49.8	61.2	73.0
DLLR* [85]	29.6	54.8	64.8	77.3
Ours_S	26.9	50.6	62.3	71.2
kLFDA_N [77]	15.9	42.4	50.0	60.7
SA_DA [89] + kLFDA [77]	15.2	41.4	49.8	58.7
AdaRSVM [26]	10.9	23.7	33.1	44.5
Adversarial [28]	22.8	38.6	50.3	63.9
Ours_L	**31.5**	**57.2**	**68.1**	**78.2**
Ours_All	**34.6**	**60.1**	**69.5**	**79.7**

(b) PRID

Rank	1	5	10	20
SDALF [86]	16.3	29.6	38.0	48.7
ISR* [84]	17.0	34.4	42.0	54.3
DLILR* [85]	21.1	43.7	55.8	64.8
Ours_S	14.1	28.5	37.4	46.3
kLFDA_N [77]	9.1	27.3	35.0	46.7
SA_DA [89] + kLFDA [77]	8.7	26.4	34.8	46.2
AdaRSVM [26]	4.9	13.1	18.4	26.3
Ours_L	**24.6**	**47.8**	**59.2**	**68.5**
Ours_All	**25.6**	**47.9**	**58.5**	**68.1**

TABLE 5: Unsupervised Re-ID results on VIPeR and PRID. '-' means no reported result is available.

the knowledge learned from the labelled auxiliary datasets cannot be directly used to help match the target data. This is due to the drastically different viewing conditions and illumination changes across views in the target dataset compared to those in the auxiliary datasets. A naive transfer learning approach such as kLFDA_N would not be able to cope with the domain shift/difference of this magnitude. (4) Importantly, it is noted that when an existing unsupervised domain-adaptation based transfer learning model is applied to alleviate the domain shift problem (SA_DA+kLFDA), the result is even worse. This is not surprising as the existing unsupervised domain adaptation methods are designed under the assumption that the auxiliary and target domains have the same recognition task (i.e. having the same set of classes) – an invalid assumption for our unsupervised Re-ID problem as different datasets contain different person identities. Also, the experimental results of Adversarial [28] suggest that the domain adaptation problem for Re-ID poses unique challenges that cannot be tackled by simple domain alignment. (5) The results of the only existing cross-dataset unsupervised Re-ID method AdaRSVM is actually the worst. Note that since its code is not available, these are the reported results in [26]. Since different feature representation and two instead of three source datasets were used, this comparison is only indicative. However, by examining some additional results (see Supplementary Material), we can conclude that indeed AdaRSVM is able to transfer very little useful information from the source datasets even when they use more supervision on the target dataset than our model. (6) It is also noted that on the two small datasets (VIPeR and PRID), our unsupervised results are not far off the best

reported results using the existing *supervised* methods (see Table 4). In addition, our model is also extremely efficient to compute (see Supplementary Material for running cost details and additional results).

6 CONCLUSION

We have proposed a novel attribute learning model for addressing the cross-class transfer learning problem. Our model learns user-defined semantic attributes jointly with discriminative and background latent attributes. The model is based on dictionary learning. An efficient algorithm is then formulated to solve the resultant optimisation problem. This model is further extended to deal with the unsupervised domain adaptation problem whereby the need of labelling training class samples is removed. Extensive experiments show that the proposed method produces state-of-the-art results on both zero-shot learning and person re-identification. One of the limitations of the proposed model is that it takes a fixed (either hand-crafted or deep) feature representation as input. With a fixed feature input, the learned attribute based representation could be sub-optimal; thus an end-to-end joint feature and attribute learning approach would be more desirable. End-to-end dictionary learning has been attempted [90]. However, integrating the proposed multi-task dictionary learning model into an end-to-end framework is non-trivial. Part of the ongoing work is on formulating the proposed model as neural network layers in an end-to-end deep learning model, particularly under an unsupervised setting.

ACKNOWLEDGEMENT

REFERENCES

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, pp. 1097–1105, 2012.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, pp. 1–9, 2015.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE TPAMI*, vol. 36, pp. 453–465, March 2014.

[5] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[6] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *CVPR*, pp. 1629–1636, 2014.

[7] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *ECCV*, pp. 155–168, 2010.

[8] K. Liang, H. Chang, S. Shan, and X. Chen, "A unified multiplicative framework for attribute learning," in *ICCV*, pp. 2506–2514, Dec 2015.

[9] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *CVPR*, pp. 819–826, 2013.

[10] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, "Learning hypergraph-regularized attribute predictors," in *CVPR*, pp. 409–417, 2015.

[11] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE TPAMI*, vol. 37, no. 11, pp. 2332–2345, 2015.

[12] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *CVPR*, 2015.

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, pp. 365–372, 2009.

[14] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *ACM MM*, pp. 789–792, 2014.

[15] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, and X. Chen, "Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction," in *ICCV*, pp. 3819–3827, 2015.

[16] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S. F. Chang, "Designing category-level attributes for discriminative visual recognition," in *CVPR*, pp. 771–778, 2013.

[17] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," *Computer Science*, pp. 73–86, 2012.

[18] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *ECCV*, pp. 808–822, 2012.

[19] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *ECCV*, 2010.

[20] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discriminative binary codes," in *ECCV*, 2012.

[21] J. Feng, S. Jegelka, S. Yan, and T. Darrell, "Learning scalable discriminative dictionary with sample relatedness," in *CVPR*, pp. 1645–1652, 2014.

[22] R. Layne, T. M. Hospedales, and S. Gong, *Attributes-Based Re-identification*. Springer London, 2014.

[23] R. Layne, T. M. Hospedales, and S. Gong, "Domain transfer for person re-identification," in *ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, pp. 25–32, ACM, 2013.

[24] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE TIP*, vol. 23, no. 8, pp. 3656–3670, 2014.

[25] X. Wang, W.-S. Zheng, X. Li, and J. Zhang, "Cross-scenario transfer person re-identification," *IEEE TCSVT*, 2015.

[26] A. J. Ma, J. Li, P. C. Yuen, and P. Li, "Cross-domain person reidentification using domain adaptation ranking svms," *IEEE TIP*, vol. 24, no. 5, pp. 1599–1613, 2015.

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

[28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.

[29] X. Zhang, F. X. Yu, S. Chang, and S. Wang, "Deep transfer network: Unsupervised domain adaptation," *CoRR*, vol. abs/1503.00591, 2015.

[30] M. Long, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," *CoRR*, vol. abs/1602.04433, 2016.

[31] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, "Semi-supervised domain adaptation with instance constraints," in *CVPR*, pp. 668–675, 2013.

[32] A. Maurer, M. Pontil, and B. Romera-Paredes, "Sparse coding for multitask and transfer learning," in *ICML*, 2013.

[33] V. Sharmanska, N. Quadrianto, and C. H. Lampert, *Augmented Attribute Representations*. Springer Berlin Heidelberg, 2012.

[34] Y. Fu, T. M. Hospedales, X. Tao, and S. Gong, "Learning multimodal latent attributes.," *IEEE TPAMI*, vol. 36, no. 2, pp. 303–316, 2014.

[35] X. Wang and Q. Ji, "A unified probabilistic approach modeling relationships between attributes and objects," in *ICCV*, pp. 2120–2127, 2013.

[36] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, *et al.*, "Devise: A deep visual-semantic embedding model," in *NIPS*, pp. 2121–2129, 2013.

[37] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *NIPS*, pp. 935–943, 2013.

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

[39] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *CVPR*, pp. 2927–2936, June 2015.

[40] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *CVPR*, pp. 2635–2644, 2015.

[41] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *ECCV*, pp. 584–599, Springer, 2014.

[42] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *ICCV*, pp. 4166–4174, Dec 2015.

[43] Y. Xian, Z. Akata1, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *CVPR*, 2016.

[44] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multi-cue zero-shot learning with strong supervision," in *CVPR*, 2016.

[45] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *ICCV*, pp. 2584–2591, 2013.

[46] E. Gavves, T. Mensink, T. Tommasi, C. G. Snoek, and T. Tuytelaars, "Active transfer learning with zero-shot priors: Reusing past datasets for future tasks," in *ICCV*, pp. 2731–2739, 2015.

[47] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *CVPR*, 2016.

[48] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *ICCV*, pp. 402–412, 2012.

[49] T. Hospedales, R. Layne, and S. Gong, "Re-id: Hunting attributes in the wild," in *BMVC*, 2014.

[50] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," *BMVC*, 2012.

[51] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, pp. 2197–2206, 2015.

[52] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *ICCV*, pp. 3739–3747, Dec 2015.

[53] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.

[54] W. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification," in *CVPR*, 2012.

[55] K. Kenneth, M.Joseph, R. Bhaskar, E. Kjersti, L. Te-Won, and S. Terrence, "Dictionary learning algorithms for sparse representation.," *Neural Computing*, vol. 15, Feb. 2003.

[56] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, 2006.

[57] H. Guo, Z. Jiang, and L. S. Davis, "Discriminative dictionary learning with pairwise constraints.," in *ACCV*, 2014.

[58] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *ICCV*, pp. 3176–3183, IEEE, 2013.

[59] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification.," in *CVPR*, 2014.

[60] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015.

[61] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang, "Joint learning of semantic and latent attributes," in *ECCV*, 2016.

[62] 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.

[63] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[64] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *NIPS*, vol. 26, pp. 3111–3119, 2013.

[65] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International Workshop on PETS*, Citeseer, 2007.

[66] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *International Conference on Distributed Smart Cameras*, 2014.

[67] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, pp. 91–102, Springer, 2011.

[68] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, pp. 152–159, IEEE, 2014.

[69] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, pp. 1116–1124, Dec 2015.

[70] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.

[71] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, pp. 791–808, Springer, 2016.

[72] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *CVPR*, 2016.

[73] M. Hirzer, M. Roth, M. Koestinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification.," in *ECCV*, 2012.

[74] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification.," in *CVPR*, 2014.

[75] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, pp. 3610–3617, IEEE, 2013.

[76] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *CVPR*, pp. 1565–1573, 2015.

[77] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods.," in *ECCV*, 2014.

[78] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *ICCV*, December 2015.

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

[80] 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.

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

[82] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

[83] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identifiation," in *CVPR*, 2016.

[84] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking.," in *IEEE TPAMI*, 2013.

[85] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *BMVC*, vol. 3, p. 8, 2015.

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

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

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

[89] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013.

[90] K. Wang, L. Lin, W. Zuo, S. Gu, and L. Zhang, "Dictionary pair classifier driven convolutional neural networks for object detection," in *CVPR*, pp. 2138–2146, 2016.

Peixi Peng received the B.S. degree from Xi'an Jiaotong University in 2010, Xi'an, China. At present Peng is pursuing the Ph.D. degree in the School of Electronics Engineering and Computer Science at Peking University, China. His research interests include surveillance video analysis, multimedia learning.

Yonghong Tian is a full professor with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China. His research interests include machine learning, computer vision, and multimedia big data. He is the author or coauthor of over 140 technical articles in refereed journals and conferences and has owned more than 40 patents.
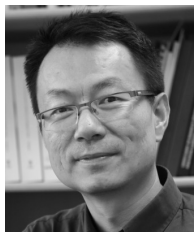
Tao Xiang received the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 120 papers in international journals and conferences.

Yaowei Wang, Ph.D., is currently an assistant professor at the Department of Electronics Engineering, Beijing Institute of Technology. He is also a guest assistant professor at NELVT, Peking University, China. His research interests include machine learning, multimedia content analysis and understanding. He is the author or coauthor of over 40 refereed journals and conference papers.

Massimiliano Pontil received a PhD in Physics from the University of Genova in 1999. He is Senior Researcher at Istituto Italiano di Tecnologia and Professor in the Department of Computer Science at University College London. His main research interests are in machine learning theory and algorithms.

Tiejun Huang , Ph.D, is a Professor with the School of Electronic Engineering and Computer Science, Head of Department of Computer Science, Peking University. His research areas include video coding and image understanding.