Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification

Jingya Wang[1] Xiatian Zhu[2] Shaogang Gong[1] Wei Li[1]
Queen Mary University of London[1] Vision Semantics Ltd.[2]
{jingya.wang, s.gong, w.li}@qmul.ac.uk eddy@visionsemantics.com

Abstract

Most existing person re-identification (re-id) methods require supervised model learning from a separate large set of pairwise labelled training data for every single camera pair. This significantly limits their scalability and usability in real-world large scale deployments with the need for performing re-id across many camera views. To address this scalability problem, we develop a novel deep learning method for transferring the labelled information of an existing dataset to a new unseen (unlabelled) target domain for person re-id without any supervised learning in the target domain. Specifically, we introduce an Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) for simultaneously learning an attribute-semantic and identity-discriminative feature representation space transferrable to any new (unseen) target domain for re-id tasks without the need for collecting new labelled training data from the target domain (i.e. unsupervised learning in the target domain). Extensive comparative evaluations validate the superiority of this new TJ-AIDL model for unsupervised person re-id over a wide range of state-of-the-art methods on four challenging benchmarks including VIPeR, PRID, Market-1501, and DukeMTMC-ReID.

1. Introduction

Person re-identification (re-id) aims at matching people across non-overlapping camera views distributed at distinct locations. Most existing re-id studies follow the *supervised* learning paradigm such as optimising pairwise matching distance metrics [23, 58, 63, 55, 60, 52, 56, 8] or deep learning methods [29, 50, 57, 48, 6, 30, 31, 5]. They assume the availability of a large number of *manually* labelled matching pairs for each pair of camera views for learning a feature representation or a matching distance function optimised for that camera pair. However, this leads to a poor scalability in practical re-id deployments, because such scale manual labelling is not only prohibitively expensive to collect in the real-world as there are a quadratic number of camera pairs, but also implausible in many cases, e.g. there may not exist sufficient training people reappearing in every pair of camera views. This scalability limitation severely reduces the usability of existing supervised re-id methods.

One generic solution to large scale re-id in real-world deployment is designing *unsupervised* models. While a few unsupervised methods have been developed [13, 9, 21, 20, 34, 51, 61], they typically offer weaker re-id performances when compared to the supervised counterparts. This makes them less useful in practice. One main reason is that without labelled data across views, unsupervised methods lack the necessary knowledge on how visual appearance of identical objects changes cross-views due to different view angles, background and illumination. Another solution is to exploit simultaneously **(1)** unlabelled data from a target domain and **(2)** existing labelled datasets from some training source domains. Specifically, the idea is to learn a feature representation that contains some view-invariant information about people appearance learned from labelled source data, transfer and adapt it to a target domain by using only unlabelled target data for re-id matching in the target domain. As the target dataset has no label, this is regarded as an *unsupervised learning* problem.

There are a few studies on exploiting unlabelled target data for unsupervised re-id modelling using either identity or attribute label, or both from source datasets [40, 59, 47]. However, they generally offer weaker re-id performance due to either domain sensitive hand-crafted features or a lack of an effective knowledge transfer learning algorithm between attribute and identity discriminative features. It is very challenging to address this cross-domain and multi-task (between attribute and identity) transfer learning problem in a principled way due to three co-occurring uncertainties: (1) Source and target domains have unknown camera viewing conditions; (2) The identity/class population between source and target domains are non-overlapping therefore presents a more challenging open-set recognition problem, as compared to the closed-set assumption made by most existing transfer learning models [39]; (3) Joint ex-

2575-7075/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPR.2018.00242

ploitation of attribute and identity labels suffers from the heteroscedasticity (a mixture of different knowledge granularity and characteristics) learning problem [11].

In this work, we consider unsupervised person re-id by sharing the source domain knowledge through attributes learned from labelled source data and transfering such knowledge to unlabelled target data by a joint attribute-identity transfer learning across domains. We make three **contributions**: **(I)** We formulate a novel idea of *heterogeneous multi-task joint deep learning* of attribute and identity discrimination for unsupervised person re-id. To our best knowledge, this is the first attempt at *joint deep learning of auxiliary attribute and identity labels* for solving the unsupervised person re-id problem cross-domains. **(II)** We propose a *Transferable Joint Attribute-Identity Deep Learning* (TJ-AIDL) to simultaneously learn *global* identity and *local* attribute information from labelled source domain person images through an Identity Inferred Attribute (IIA) space for maximising the joint learning effectiveness between identity and attribute. This IIA is designed specially to address the notorious heteroscedasticity challenge from which the common space multi-task joint learning often suffers. Importantly, the IIA interacts concurrently with both the attribute and identity learning tasks inter-dependently without breaking the end-to-end model learning process. **(III)** We introduce an attribute consistency scheme for performing TJ-AIDL model unsupervised adaptation on the unlabelled target data to further enhance its discriminative compatibility towards each target domain re-id task at hand. Extensive evaluations demonstrate the superiority of the proposed TJ-AIDL model over a wide range of state-of-the-art re-id models on four challenging benchmarks VIPeR [14], PRID [17], Market-1501 [62], and DukeMTMC-ReID [64].

2. Related work

Person Re-ID Most existing re-id models are based on *supervised* learning for every camera pair on a separate set of labelled training data [23, 58, 63, 55, 60, 52, 56, 8, 29, 50, 57, 48, 6, 30, 7, 31, 5]. They suffer from poor scalability in realistic re-id deployments where no such a large training set is available for each single camera pair. To solve this scalability issue, unsupervised methods based on hand-crafted features [13, 9, 21, 20, 34, 51, 61, 51] can be chosen for deployment. However, they usually yield much weaker performance than supervised models therefore practically not very useful. While a balance between scalability and matching accuracy can be achieved by semi-supervised learning, existing methods [35, 53] still demand a fairly large set of pairwise labels which is again not scalable.

Recently, unsupervised re-id by cross-domain transfer learning has been developed to exploit labelled data from source datasets by extracting transferable identity-discriminative information to an unlabelled target dataset [40, 59, 47]. However, these methods have a few limitations that restrict their generalisation: (1) Relying on hand-crafted features without the deep learning capability of automatically learning stronger representations from training data [40]; (2) Using a pre-learned deep model on labelled source data but lacking an effective domain adaptation mechanism [59]; (3) Independently exploiting identity and attribute label supervision in model learning therefore ignoring their interaction and compatibility [47]. Data synthesis [2, 10] has also been proposed as a solution for addressing limited data, although it suffers from undesirable person appearance distortion and restricted source selection. The proposed TJ-AIDL method addresses these limitations of existing methods in a unified deep joint learning model. Moreover, our method goes beyond the common multi-task joint learning design by introducing a more transferable mechanism for discriminatively optimising both attribute and identity learning in a shared end-to-end process. Our experiments show that the proposed method significantly outperforms existing models even by using less supervision in the source domain.

Attribute for Re-ID Visual semantic attributes [54, 36] have been exploited as a mid-level feature representation for cross-view re-id [26, 24, 25, 47, 46, 40]. However, such semantic coefficient representations are less powerfull for identity discrimination than conventional feature vectors. The reasons are: (1) Attribute coefficient representations are usually of low dimensions (tens vs. thousands for typical low-level feature representations) [62, 32, 14, 63]; (2) Consistently predicting individually all the attributes is a difficult task when the labelled training data is sparse and person images have low quality as mostly in person re-id datasets, that is, inter-attribute discrimination can be weak on typical person re-id images. To overcome these problems, we explore attributes in our TJ-AIDL model by introducing a mechanism to extract identity discriminative attribute information through co-learning both attribute and identity labelled data jointly. Moreover, we uniquely employ the attribute space for unsupervised domain adaptation.

3. A Joint Attribute-Identity Space

Problem Definition For person re-id by attribute (semantic) based unsupervised domain adaptation, we have a *supervised* source dataset (domain) $\{(\boldsymbol{I}_i^s, y_i^s, \boldsymbol{a}_i^s)\}_{i=1}^{N^s}$ consisting of N^s person bounding box images $\boldsymbol{I}^s$, the corresponding *identity* $y^s \in \mathcal{Y} = \{1, \cdots, N_{\text{id}}^s\}$ (i.e. a total N_{id}^s different persons), and identity-level *binary attribute* $\boldsymbol{a}_s \in \mathcal{R}^{m\times 1}$ (i.e. a total m different attributes) labels. We also assume a set $\{\boldsymbol{I}_i^t\}_{i=1}^{N^t}$ of N^t unlabelled target training data, which can be used for model domain adaptation. The **objective** is to develop an unsupervised domain adaptation approach to learning the optimal feature representation by

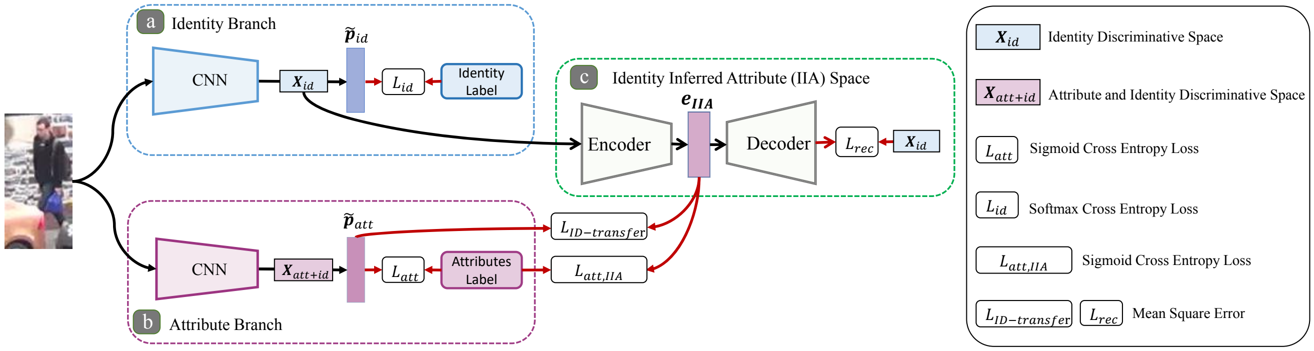

a
Identity Branch
CNN
X_{id}
$\tilde{p}_{id}$
L_{id}
Identity Label
c
Identity Inferred Attribute (IIA) Space
e_{IIA}
Encoder
Decoder
L_{rec}
X_{id}
$\tilde{p}_{att}$
$L_{ID-transfer}$
CNN
X_{att+id}
L_{att}
Attributes Label
$L_{att,IIA}$
b
Attribute Branch
X_{id} Identity Discriminative Space
X_{att+id} Attribute and Identity Discriminative Space
L_{att} Sigmoid Cross Entropy Loss
L_{id} Softmax Cross Entropy Loss
$L_{att,IIA}$ Sigmoid Cross Entropy Loss
$L_{ID-transfer}$ L_{rec} Mean Square Error

Figure 1. An overview of the proposed Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL).

transferring the supervised identity and attribute knowledge of the source domain to person re-id in a target domain with only *unlabelled* data of entirely a different pool of identity classes. Note that: **(1)** Unlike identity label, attribute detection is a *multi-label* recognition problem since the m attribute categories co-exist in every single person image. **(2)** These two types of label supervision lie at quite different levels: Most attributes are *localised* to image regions, even though the location information is not provided in the annotation; While person identity labels are at the *holistic* image-level. It is a non-trivial learning task since this is not only a multi-label learning problem – joint learning of mutually correlated attribute labels, but also a *heterogeneous multi-task joint learning* problem – inter-dependently learning a person re-id representation space by joint holistic identity and local attribute supervision.

In this work, we present a novel *Transferable Joint Attribute-Identity Deep Learning* (TJ-AIDL) approach to establishing an identity-discriminative and attribute-sensitive (i.e. dually-semantic) feature representation space optimal for person re-id on the labelled target domain without any identity and attribute labels provided. We avoid simply combining re-id and attribute feature vectors in deep model design to gain their complementary advantages, which may suffer from the heteroscedasticity problem [11] and finally results in sub-optimal results. Instead, we assign them into two separate branches for simultaneously learning individual discriminative features subject to the corresponding label supervision concurrently. Importantly, we design a progressive knowledge fusion mechanism by introducing an *Identity Inferred Attribute* (IIA) regularisation space for more smoothly transferring the global identity information into the local attribute feature representation space. It is also the proposed IIA space that provides an opportunity that allows for adapting the learned model to the target domain where no identity and attribute labels are available. As such, the proposed TJ-AIDL largely addresses the joint learning challenges of heterogeneous identity and attribute label information sources in a shared representational space in a more challenging cross-domain context.

3.1. Transferable Joint Deep Learning

Model Overview We consider a multi-branch network architecture for our heterogeneously supervised multi-task learning. The rational of this multi-branch composition is to maintain a sufficient independence of each supervision learning tasks for avoiding their potentially negative mutual influence due to their semantic discrepancy. An overview of the proposed Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) method is depicted in Fig. 1. The TJ-AIDL contains two branches: (1) *Identity Branch*: which aims to extract the re-id sensitive information from the available identity class labels in the source domain (Figure 1(a)). (2) *Attribute Branch*: which aims to extract the semantic knowledge from the attribute labels (also from the source domain) (Figure 1(b)). To establish a channel for knowledge fusion, we introduce the *Identity Inferred Attribute* (IIA) space (Figure 1(c)) designed for transferring the re-id discriminative information from the Identity Branch to the Attribute Branch where two-source information is synergistically integrated in a *smoother* manner. That is, once the TJ-AIDL is trained, the feature representations extracted from the Attribute Branch can be directly exploited for re-id deployment.

For unsupervised person re-id by cross-domain knowledge transfer and target data adaptation, we conduct the model training of our proposed TJ-AIDL in two steps: **(I)** *Attribute-Identity Transferable Joint Learning*: This is supervised by the source labelled training data; **(II)** *Unsupervised Target Domain Adaptation*: This is performed on the target unlabelled training data. We describe more details for each component of our TJ-AIDL in two training steps.

3.1.1 Attribute-Identity Transferable Joint Learning

Identity and Attribute Branches For building an efficient yet strong deep re-id model, we choose the lightweight MobileNet as the CNN architecture[1] for both identity and attribute branches. For training the identity branch (Fig. 1(a)), we use the softmax Cross Entropy loss function defined as:

$$L_{\text{id}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \log \left(p_{\text{id}}(\boldsymbol{I}_i^s, y_i^s) \right) \tag{1}$$

where $p_{\text{id}}(\boldsymbol{I}_i^s, y_i^s)$ specifies the predicted probability on the groundtruth class y_i^s of $\boldsymbol{I}_i^s$, and n_{bs} denotes the batch size.

Given that the attribute branch (Fig. 1(b)) is a multi-label classification learning task, we instead use the Sigmoid Cross Entropy loss function to generate the training signal by considering all m attribute classes:

$$L_{\text{att}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{m} \Big(a_{i,j} \log \left(p_{\text{att}}(\boldsymbol{I}_i, j) \right) + (1 - a_{i,j}) \log \left(1 - p_{\text{att}}(\boldsymbol{I}_i, j) \right) \Big) \tag{2}$$

where $a_{i,j}$ and $p_{\text{att}}(\boldsymbol{I}_i, j)$ define the groundtruth label and the predicted classification probability on the j-th attribute class of the training image $\boldsymbol{I}_i$, i.e. $\boldsymbol{a}_i = [a_{i,1}, \cdots, a_{i,m}]$ and $\boldsymbol{p}_{\text{att},i} = [p_{\text{att}}(\boldsymbol{I}_i, 1), \cdots, p_{\text{att}}(\boldsymbol{I}_i, m)]$.

By *independently* training the two branches using the above designs, we only allows for optimising their respective features without exploiting their complementary effect for maximising the compatibility. A common approach is to build a multi-task joint learning network which *directly* subjects a *shared* feature representation to both identity loss (Eq. (1)) and attribute loss (Eq. (2)) concurrently in model training. Instead, we present an alternative progressive scheme for more effective multi-source knowledge fusion as described below (see evaluations in Sec. 4.2).

Identity Inferred Attribute Space We introduce an intermediate Identity Inferred Attribute (IIA) Space for achieving the knowledge fusion learning on attribute and identity labels in a softer manner (Fig. 1(c)). The IIA space is jointly learned with the two branches while being exploited to perform information transfer and fusion from the identity branch to the attribute branch simultaneously. This scheme allows for both consistent and cumulative knowledge fusion in the whole training course.

More specifically, we build the IIA space in the encoder-decoder (auto-encoder) framework due to that: (1) It has a strong capability of capturing the most important information of a given target task (represented by the input data) via a concise feature vector representation; (2) More importantly, such a concise feature representation facilitates the inter-task information transfer whilst still preserving sufficient updating freedom space to every individual learning task [4, 43]. We call this sub-model *IIA encoder-decoder*. In our context, we want to extract and compress essential identity information into the IIA space for facilitating fusion. We therefore exploit the identity features (Fig. 1(a)) as the input of IIA encoder and also the groundtruth of IIA decoder (i.e. reconstruction unsupervised learning). Once the input is given, this model itself can be learned based on the reconstruction loss (Mean Square Error (MSE)):

$$L_{\text{rec}} = \|\boldsymbol{x}_{\text{id}} - f_{\text{IIA}}(\boldsymbol{x}_{\text{id}})\|^2 \tag{3}$$

where $\boldsymbol{x}_{\text{id}}$ represents the identity feature of a training image and $f_{\text{IIA}}()$ the mapping function of IIA encoder-decoder. By this unsupervised learning manner, we are able to obtain a latent feature embedding $\boldsymbol{e}_{\text{IIA}}$ with important identity information encoded. To transfer the identity information across branches, we need a corresponding low dimensional matchable space in the attribute counterpart, which however is not available.

To address the above problem, we propose to align the IIA embedding $\boldsymbol{e}_{\text{IIA}}$ with the prediction distribution over all m attribute classes, in spirit of knowledge distillation [16]. As such, we naturally set m as the dimension of $\boldsymbol{e}_{\text{IIA}}$ for easing alignment and cross-branch knowledge transfer without the need for an additional transformation.

More formally, we conduct the identity knowledge transfer via imposing an MSE based identity transfer loss:

$$L_{\text{ID-transfer}} = \|\boldsymbol{e}_{\text{IIA}} - \tilde{\boldsymbol{p}}_{\text{att}}\|^2 \tag{4}$$

where $\tilde{\boldsymbol{p}}_{\text{att}}$ is logits from the attribute branch. Considering that the $\boldsymbol{e}_{\text{IIA}}$ is derived in an unsupervised manner which may be over further way from the attribute prediction counterpart and hence giving a harder alignment task, we add similarly a sigmoid Cross Entropy loss to the learning of $\boldsymbol{e}_{\text{IIA}}$ by exploiting it as a pseudo attribute prediction, as

$$L_{\text{attr, IIA}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{m} \Big(a_{i,j} \log \left(p_{\text{IIA}}(\boldsymbol{I}_i, j) \right) + (1 - a_{i,j}) \log \left(1 - p_{\text{IIA}}(\boldsymbol{I}_i, j) \right) \Big) \tag{5}$$

where $p_{\text{IIA}}(\boldsymbol{I}_i, j)$ is the probability predicted based on $\boldsymbol{e}_{\text{IIA}}$ by the sigmoid function. Finally, we formulate the overall IIA loss function by incorporating the above components by weighted summation as:

$$L_{\text{IIA}} = L_{\text{attr, IIA}} + \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{ID-transfer}} \tag{6}$$

where λ_1 and λ_2 are scale normalisation parameters to ensure all three loss quantities are of a similar scale in value.

Impact of IIA on Identity and Attribute Branches The introduction of IIA imposes different influence on the two branches in model training. Since IIA is established on the identity features, no change is imposed into the learning of

[1] This selection is independent of our model design and others can be readily applied, e.g. ResNet [15], Inception [49] and VggNet [45].

this branch. For the attribute branch, however, an additional learning constraint is created for identity knowledge transfer. We therefore reformulate its supervised learning loss function by incorporating Eq. (4) as:

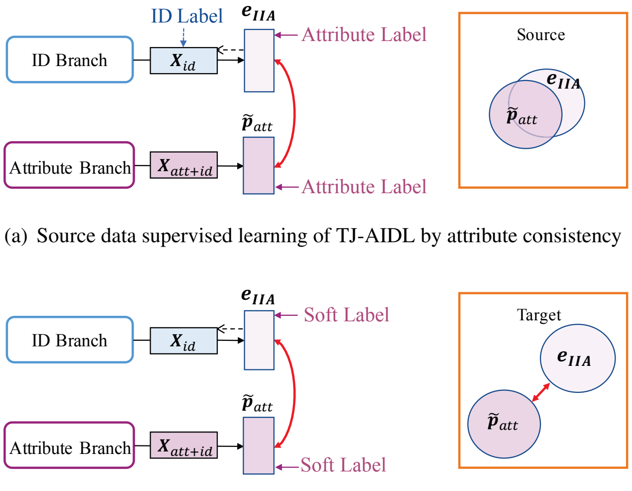
$$L_{\text{att-total}} = L_{\text{att}} + \lambda_2 L_{\text{ID-transfer}} \tag{7}$$

Remarks The IIA component aims at creating an interactive learning mechanism between the identity and attribute branches in a more transferable way. This significantly differs from the straightforward joint learning approach which suffers from the underlying multi-source information incompatible problem. We summarise the main information flow in model joint training: (1) The identity branch learns to extract identity discriminative information; (2) The IIA component then transfers the identity information to the attribute branch; (3) The attribute branch learns to extract attribute discriminative knowledge whilst simultaneously incorporating/fusing identity sensitive information. However, the TJ-AIDL model learned on the labelled source data is still not optimal for re-id in a typically *unlabelled* target domain due to the inevitable presence of domain shift in real-world deployment scenarios. This leads to the necessity of model unsupervised domain adaptation, as detailed below.

3.1.2 Unsupervised Target Domain Adaptation

We want to adapt a learned TJ-AIDL model to fit the unlabelled target domain data. To that end, we exploit the attribute consistency principle by treating the prediction of attribute branch and the embedding of IIA component as different attribute perspectives from different domains. This idea is based on the observation that, a well fitted TJ-AIDL model is supposed to have small discrepancy between the two different attribute perspectives, for example, the one trained on the source domain (Fig. 2(a)). In other words, their consistency degree suggests how well the model fits a given domain. This also partially shares the spirit of the cyclic consistency mechanism [44].

Specifically, our objective is to adapt the attribute branch since it is used in re-id deployment. Hence, we can ignore the updating of the identity branch. We design the following adaptation algorithm: **(1)** We deploy the TJ-AIDL model learned on the source domain on unlabelled target person images to obtain the attribute prediction $\boldsymbol{p}_{\text{att},t}$ from the attribute branch. **(2)** We then utilise the soft label $\boldsymbol{p}_{\text{att},t}$ as the pseudo groundtruth to update both the attribute branch and IIA component for reducing attribute discrepancy between domains (Fig. 2(b)). Intuitively, this soft attribute label is needed since we need to prevent the model drifting overly by maintaining the most attribute discriminative power obtained from the source domain. **(3)** We adapt the model on the target training data until convergence.



(a) Source data supervised learning of TJ-AIDL by attribute consistency

(b) Taget domain adaptation of TJ-AIDL by attribute consistency

Figure 2. An illustration of the attribute consistency maximisation idea for unsupervised target domain adaptation. Given a TJ-AIDL model trained on the source domain, it has more attribute consistency (a) on the source domain, (b) but less on the unseen target domain. See more details in the main text.

3.2. Model Optimisation and Deployment

Optimisation Our TJ-AIDL model can be trained using the standard Stochastic Gradient Descent algorithm in end-to-end manner. We summarise the training process in Alg. 1.

Deployment Given a TJ-AIDL model trained on a labelled source domain and adapted on the unlabelled target domain, we obtain a 1,024-D deep feature representation from the attribute branch (Fig. 1(b)). This feature vector is not only attribute semantic but also identity discriminative. Hence, we deploy this 1,024-D deep feature for person re-id deployment by the L_2 distance in the target domain.

4. Experiments

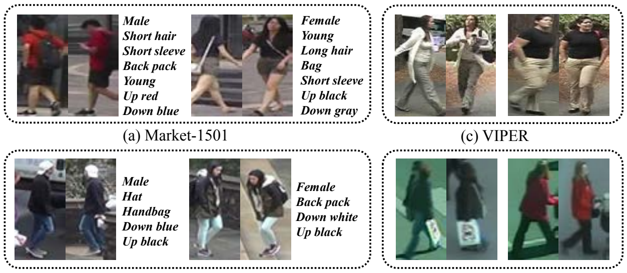
Datasets and Evaluation Protocol We choose four widely adopted person re-id benchmarks for experimental evaluations (Fig. 3). We adopt the standard supervised re-id data split settings and only use the test data for model evaluation whilst the training part is ignored.

(1) The ***Market-1501*** dataset [62] contains 32,668 images of 1,501 pedestrians, each of which was captured by at most six cameras at a university campus. All of the images were cropped by a pedestrian detector and therefore presenting more challenges to re-id models due to more background clutters and the misalignment problem. *Evaluation Protocol*: We used the standard training/test split (750/751) and evaluated on single-query evaluation settings [62].

(2) The ***DukeMTMC-ReID*** dataset [64] contains $2 \sim 426$ images per person captured by 8 non-overlapping camera views. This dataset was constructed from the multi-camera

Algorithm 1 Learning the TJ-AIDL model.

```
Input: $N^s$ labelled source $\{(\boldsymbol{I}_i^s, y_i^s, \boldsymbol{a}_i^s)\}_{i=1}^{N^s}$ and $N^t$ unlabelled target $\{\boldsymbol{I}_i^t\}_{i=1}^{N^t}$ training data;
Output: TJ-AIDL re-id model;
Step I: Transferable Joint Learning (Sec. 3.1.1)
for $t = 1$ to max-iteration do
    Sampling a batch of labelled source data;
    Identity branch evaluation (samples feed-forward);
    Attribute branch evaluation;
    Updating the identity branch (Eq. (1));
    Updating the IIA encoder-decoder (Eq. (6));
    Updating the attribute branch (Eq. (7));
end for
Step II: Unsupervised Target Domain Adaptation (Sec. 3.1.2)
for $t = 1$ to max-iteration do
    Sampling a batch of unlabelled target training data;
    Attribute branch evaluation to obtain the soft labels;
    Updating the IIA encoder-decoder (Eq. (6));
    Updating the attribute branch (Eq. (7)).
end for
```



(a) Market-1501 (c) VIPER

(b) DukeMCMT-ReID (d) PRID

Figure 3. Example of person images and attribute labels. Each pair represents two images of the same person.

Table 1. Unsupervised re-id performance evaluation. **Metric**: Rank-1 and mAP (%). The 1st/2nd best results are in red and blue. TJ-AIDL$^{\text{Duke}}$ / TJ-AIDL$^{\text{Market}}$: Our TJ-AIDL using DukeMCMT-ReID and Market-1501 as the labelled source, respectively.

Dataset	VIPeR	PRID	Market-1501		DukeMCMT	
Metric (%)	R1	R1	R1	mAP	R1	mAP
SDALF[13]	19.9	16.3	-	-	-	-
DLLR [21]	29.6	21.1	-	-	-	-
CPS [9]	22.0	-	-	-	-	-
GL [20]	33.5	25.0	-	-	-	-
GTS [51]	25.2	-	-	-	-	-
SDC[61]	25.8	-	-	-	-	-
ISR [34]	27.0	17.0	40.3	14.3	-	-
Dic[22]	29.9	-	50.2	22.7	-	-
RKSL[53]	25.8	-	34.0	11.0	-	-
SAE[27]	20.7	-	42.4	16.2	-	-
AML[42]	23.1	-	44.7	18.4	-	-
UsNCA [42]	24.3	-	45.2	18.9	-	-
CAMEL [59]	30.9	-	**54.5**	**26.3**	-	-
PUL [12]	-	-	44.7	20.1	**30.4**	**16.4**
kLFDA_N [58]	15.9	9.1	-	-	-	-
SADA+kLFDA [58]	15.2	8.7	-	-	-	-
AdaRSVM [37]	10.9	4.9	-	-	-	-
UDML [40]	31.5	24.2	-	-	-	-
SSDAL [47]	**37.9**	20.1	39.4	19.6	-	-
TJ-AIDL$^{\text{Duke}}$	35.1	**34.8**	**58.2**	**26.5**	N/A	N/A
TJ-AIDL$^{\text{Market}}$	**38.5**	**26.8**	N/A	N/A	**44.3**	**23.0**

tracking dataset DukeMTMC by random selection of manually labelled tracklet bounding boxes. *Evaluation Protocol*: We followed [64] by splitting all 1,404 person identities into two halves 702/702 for model training and test, respectively and testing re-id tasks in the single-query setting.

(3) The ***VIPeR*** dataset [14] has 632 identities each with two images captured from two camera views in different scenarios of illumination, postures and viewpoints. This dataset is also featured with low resolution therefore giving rise to an extremely challenging re-id task. *Evaluation Protocol*: We randomly split the whole population into two halves as training/test sets. We repeat 10 times of random split and report the average result.

(4) The ***PRID*** dataset [17] consists of person images from two camera views: View A captures 385 people, whilst View B contains 749 people. Only 200 people appear in both views. *Evaluation Protocol*: We use the single shot version in our experiments as [60]. In each data split, 100 people with one image from each view are randomly chosen from the 200 present in both camera views as the training set, while the remaining 100 of View A are used as the probe set, and the remaining 649 of View B are used as gallery. Experiments are repeated over 10 random splits.

For performance metric, we use the cumulative matching characteristic (CMC) and mean Average Precision (mAP).

Attribute Annotation In our evaluations, we use either *Market-1501* [62] or *DukeMTMC-ReID)* dataset [33] as the source domain, since they provide both identity and attribute labels (Fig. 3). Specifically, there are 27/23 classes of attributes labelled for Market-1501 / DukeMTMC-ReID [33]. To ensure the unsupervised re-id property, we do not test the Market-1501 when it is used as the source domain. This similarly applies to DukeMTMC-ReID.

Implementation Details We realised the TJ-AIDL model in the Tensorflow framework [1]. The IIA encoder is designed as a 3-FC-layers network with their output dimensions as 512/128/m (m is the number of attribute labels). A network of a mirror structure is used in the IIA decoder. We fixed both λ_1 and λ_2 to 10 (Eq. (6)) by scale alignment. We pre-trained the MobileNet on ImageNet for both identity and attribute branches. We used the Adam optimiser [19] with a learning rate of 0.002 and the default momentum terms $\beta_1 = 0.5$, $\beta_2 = 0.999$. We set the mini-batch size to 8. We started with training the identity branch by 100,000 iterations on the source identity labels and then the whole model by 20,000 iterations for both transferable joint learning on the labelled source data and unsupervised domain adaptation on the unlabelled target data.

4.1. Comparisons to the State-Of-The-Arts

We compare 19 models in three categories of existing unsupervised re-id methods: **(1)** Hand-crafted feature based methods *without* transfer learning: SDALF[13] and CPS [9], those features are designed to be view invariant. Dictionary Learning based methods DLLR [21], graph-learning-based model GL [20], sparse representation learning methods ISR [34], salience-learning-based GTS [51] and SDC[61]. **(2)** Source identity knowledge transfer learning based methods: Dic [22], RKSL [53], SAE [27], AML [42], UsNCA [42], CAMEL [59]. **(3)** Source identity and attribute knowledge based transfer methods: kLFDA_N [58] SADA+kLFDA [58] AdaRSVM [37] UDML [40].

Table 1 shows that: **(1)** Our method outperforms clearly all existing state-of-the-art models, improving the Rank-1 by 0.6% (38.5-37.9), 9.8% (34.8-25.0), 3.7% (58.2-54.5), 13.9% (44.3-30.4) over the best alternative method on VIPeR/PRID/Market-1501/DukeMCMT-ReID, respectively. This suggests the overall performance advantages of the proposed TJ-AIDL in the capability of multi-source (attribute and identity) information extraction and fusion for cross-domain unsupervised re-id matching. **(2)** When compared to the existing methods of 1$^{\text{st}}$ category (non-learning based) the performance margins are even much larger, e.g. the Rank-1 boost is 8.9% (38.5-29.6), 9.8% (34.8-25.0), 17.9% (58.2-40.3) on VIPeR/PRID/Market-1501, respectively. This indicates the importance of learning from labelled source supervision in cross-domain re-id scenarios, since hand-crafted features are not sufficiently generalisable across different domains with varying camera view conditions. **(3)** When comparing the methods between 2$^{\text{nd}}$ (identity transfer) and 3$^{\text{rd}}$ (identity and attribute joint transfer) category, it is interestingly found that the latter is not necessarily superior over the former. This means that using more supervision in cross-domain transfer learning is non-trivial particularly when the label property is heterogeneous such as identity and attribute. This also indirectly suggest the model design advantages of our TJ-AIDL in exploiting the diverse knowledge in different types of label data for the more challenging cross-domain re-id tasks in the unlabelled target scenario typical in real-world deployments.

Finally, it is worth noting that the performance advantages by our TJ-AIDL are achieved using much less supervision data of lower diversity from only *one* source domain (16,522 images of 702 identities/classes on DukeMCMT-ReID, or 12,936 images of 751 identities on Market-1501) than strong existing competitors. For example, the methods of 2$^{\text{nd}}$ category utilise 7 different person re-id datasets with high domain varieties (CUHK03[29], CUHK01[28], PRID, VIPeR, 3DPeS[3], i-LIDS[41], Shinpuhkan[18]) including a total of 44,685 images and 3,791 identities; The UDML [40] exploits three different source domains including 46,966 images of 3,246 identities for test on VIPeR (target), and 47,096 images of 3,493 identities for test on PRID (target). The SSDAL [47] benefits from 10 diverse datasets consisting in 19,000 images of 8,705 person identities and another 20,000 images of 1,221 person tracklets.

4.2. Comparisons to Alternative Fusion Methods

We compare the TJ-AIDL with two multi-source fusion methods: **(a)** Independent Supervision: Independently train a deep CNN model for either attribute or identity label in the source domain and use the concatenated feature vectors of the two models for re-id matching in the target domain. **(b)** Joint Supervision: A seminal multi-task joint learning CNN framework subjecting the identity and attribute supervision to a shared feature representation in the end-to-end model training. For re-id deployment on the target domain, we use the multi-supervision shared feature representation.

Table 2 shows that: (1) The TJ-AIDL outperforms both alternative fusion methods. This suggests a clear advantage of our method in exploiting and fusing multiple supervision for cross-domain re-id in an unsupervised manner. (2) Our method achieves more performance gain over the competitors on the transfer from Market-1501 (source) to DukeMTMC-ReID (target) than the opposite transfer. This is expected and reasonable because relative to Market-1501, person images from DukeMTMC-ReID have more changes in image resolution and background clutter due to wider camera views and more complex scene layout, which means the source information itself from Market-1501 is insufficient to generalise the target DukeMTMC-ReID setting and therefore leading to a higher need for domain adaptation. Our model strongly and naturally meets this deployment requirement. For the opposite transfer from DukeMTMC-ReID to Market-1501, our model gives less performance gain since there is a lower need for domain adaptation.

4.3. Further Analysis and Discussions

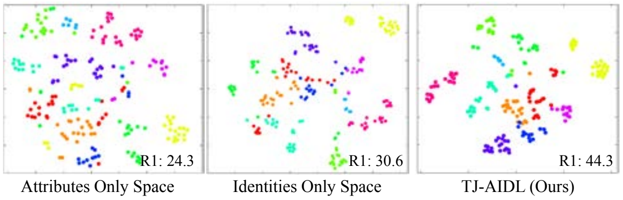
Effect of Joint Attribute and Identity Features We evaluated the effect of joint attribute and identity features by comparing their individual re-id performances against that of the joint feature. We obtain their individual model by training a MobileCNN using either identity or attribute label only. Table 3 shows feature representation learned by only one supervision is significantly inferior that that by our TJ-AIDL. For instance, the TJ-AIDL feature outperforms ID Only by 13.7%(44.3-30.6) in Rank-1 and 8.4% (23.0-14.6) in mAP on DukeMCMT-ReID (target); by 6.6% (58.2-51.6) in Rank-1 and 4.9%(26.5-21.6) in mAP on Market-1501 (target). These validate the complementary effect of jointly learning attribute and identity information and importantly strong capability of our model in maximising this latent information in a more transferable context. We also plot three feature distributions of 10 randomly selected test identities of DukeMTMC-ReID (transferred from Market-1501). Fig-

Table 2. Comparing different multi-source fusion methods.

Source → Target	Market-1501 → DukeMCMT-ReID					DukeMCMT-ReID → Market-1501				
Metric (%)	Rank1	Rank5	Rank10	Rank20	mAP	Rank1	Rank5	Rank10	Rank20	mAP
Independent Supervision	33.8	49.5	56.0	63.8	16.9	54.9	72.9	79.3	85.2	24.5
Joint Supervision	37.9	52.1	58.6	65.3	20.6	53.4	71.2	78.1	83.3	21.9
TJ-AIDL	**44.3**	**59.6**	**65.0**	**70.0**	**23.0**	**58.2**	**74.8**	**81.1**	**86.5**	**26.5**

Table 3. Complementary of identity-discriminative and attribute-sensitive features learned by the proposed TJ-AIDL.

Source → Target	Market-1501 → DukeMCMT-ReID				
Metric (%)	Rank1	Rank5	Rank10	Rank20	mAP
Attribute Only	24.3	38.3	45.7	53.0	10.0
ID Only	30.6	44.9	50.5	59.3	14.6
Attribute + ID (**Full**)	**44.3**	**59.6**	**65.0**	**70.0**	**23.0**
Source → Target	DukeMCMT-ReID → Market-1501				
Attribute Only	38.0	59.2	67.6	75.7	13.6
ID Only	51.6	69.8	76.6	81.6	21.6
Attribute + ID (**Full**)	**58.2**	**74.8**	**81.1**	**86.5**	**26.5**



Attributes Only Space
Identities Only Space
TJ-AIDL (Ours)

Figure 4. Feature distributions of 10 random test identities in three transferred feature spaces (Market-1501 → DukeMCMT-ReID) visualised by t-SNE [38]. Colour coded identity classes.

ure 4 shows that: (1) Neither transferring the knowledge of attributes or identities alone can form per-identity compact clusters; (2) By our TJ-AID that transfers attributes and identities jointly, the feature distributions of 10 test identities are much more separated.

Table 4. Effect of the target domain adaptation in TJ-AIDL.

Source → Target	Market-1501 → DukeMCMT-ReID				
Metric (%)	Rank1	Rank5	Rank10	Rank20	mAP
w/o Adaptation	39.6	55.5	62.2	67.5	22.0
w Adaptation	**44.3**	**59.6**	**65.0**	**70.0**	**23.0**
Source → Target	DukeMCMT-ReID → Market-1501				
w/o Adaptation	57.1	74.4	80.4	85.7	26.2
w Adaptation	**58.2**	**74.8**	**81.1**	**86.5**	**26.5**

Effect of Target Domain Adaptation We evaluated the effect of the attribute consistency driven domain adaptation on unlabelled target training data. Table 4 shows that this adaptation clearly improves the re-id performance for the transfer of DukeMCMT-ReID → Market-1501 (1.1% Rank-1 boost) and more significantly for the case of Market-1501 → DukeMCMT-ReID (4.7% Rank-1 boost). This shares a similar observation and underlying reason as in Table 2, validating the benefit of our method in varying cross-domain model adaptation in improving the model compatibility when deployed to a new target scenario.

5. Conclusion

We presented a novel Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) for more discriminative joint learning of the identity and attribute supervision from an auxiliary domain in order to particularly address the scalable unsupervised person re-identification problem in the context of heterogeneous multi-task joint learning and domain transfer learning. In contrast to most existing re-id methods that either ignore the scalability issue in re-id or exploit a straightforward yet sub-optimal multi-task joint learning of multi-supervision, the proposed model is capable of transferring and integrating multiple heterogeneous supervision and maximising their latent compatibility for optimal person re-id in a progressive and more transferable means. This is achieved by introducing an Identity Inferred Attribute space for interactive attribute and identity discriminative learning in a two-branches CNN architecture. Moreover, we introduce an attribute consistency maximisation mechanism to further discriminatively adapt a learned TJ-AIDL model to fit any given target re-id deployment without the need for additional data labelling and hence very scalable to real-world applications. Extensive evaluations were conducted on four re-id benchmarks to validate the advantages of the proposed TJ-AIDL model over a wide range of state-of-the-art methods on different re-id task scenarios with various challenges. We also compared the TJ-AIDL model with popular multi-supervision fusion methods and provided detailed component analysis with insights into the performance gain of our model design.

Acknowledgements

This work was partially supported by the China Scholarship Council, Vision Semantics Ltd, Royal Society Newton Advanced Fellowship Programme (NA150459), and InnovateUK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. 6

[2] L. An, Z. Qin, X. Chen, and S. Yang. Multi-level common space learning for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 2

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011. 7

[4] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 1988. 4

[5] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 1, 2

[6] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017. 1, 2

[7] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCV Workshop*, 2017. 2

[8] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE TPAMI*, 2017. 1, 2

[9] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 1, 2, 6, 7

[10] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 2

[11] R. Duin and M. Loog. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *IEEE TPAMI*, 2004. 2, 3

[12] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv*, 2017. 6

[13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 2, 6, 7

[14] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2, 6

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv*, 2015. 4

[17] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 2, 6

[18] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, volume 5, page 6, 2014. 7

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 6

[20] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised ℓ_1 graph learning. In *ECCV*, 2016. 1, 2, 6, 7

[21] E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015. 1, 2, 6, 7

[22] E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015. 6, 7

[23] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 2

[24] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. 2014. 2

[25] R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014. 2

[26] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, volume 2, page 8, 2012. 2

[27] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area v2. In *NIPS*, 2008. 6, 7

[28] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 7

[29] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 7

[30] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *IJCAI*, 2017. 1, 2

[31] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1, 2

[32] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2

[33] Y. Lin, L. Zheng, and W. Y. a. Y. Y. Zheng, Zhedong and. Improving person re-identification by attribute and identity learning. *arXiv*, 2017. 6

[34] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE TPAMI*, 2015. 1, 2, 6, 7

[35] 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. 2

[36] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 2

[37] A. J. Ma, J. Li, P. C. Yuen, and P. Li. Cross-domain person reidentification using domain adaptation ranking svms. *IEEE TIP*, 2015. 6, 7

[38] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[39] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 2010. 1

[40] 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. 1, 2, 6, 7

[41] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 7

[42] C. Qin, S. Song, G. Huang, and L. Zhu. Unsupervised neighborhood component analysis for clustering. *Neurocomputing*, 2015. 6, 7

[43] A. Rannen, R. Aljundi, and M. B. B. T. Tuytelaars. Encoder based lifelong learning. In *CVPR*, 2017. 4

[44] O. Sener, H. O. Song, A. Saxena, and S. Savarese. Learning transferrable representations for unsupervised domain adaptation. In *NIPS*, 2016. 5

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[46] 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*, 2015. 2

[47] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016. 1, 2, 6, 7

[48] A. Subramaniam, M. Chatterjee, and A. Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, 2016. 1, 2

[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4

[50] 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. 1, 2

[51] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 1, 2, 6, 7

[52] H. Wang, S. Gong, X. Zhu, and T. Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016. 1, 2

[53] H. Wang, X. Zhu, T. Xiang, and S. Gong. Towards unsupervised open-set person re-identification. In *ICIP*, 2016. 2, 6, 7

[54] J. Wang, X. Zhu, S. Gong, and W. Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, 2017. 2

[55] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014. 1, 2

[56] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE TPAMI*, 2016. 1, 2

[57] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1, 2

[58] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 1, 2, 6, 7

[59] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017. 1, 2, 6, 7

[60] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 1, 2, 6

[61] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *IEEE TPAMI*, 2017. 1, 2, 6, 7

[62] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 5, 6

[63] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE TPAMI*, 2013. 1, 2

[64] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 2, 5, 6