# Transferring a Semantic Representation for Person Re-Identification and Search

Shi, Z; Yang, Y; Hospedales, T; XIANG, T; IEEE Conference on Computer Vision and Pattern Recognition

# Transferring a Semantic Representation for Person Re-Identification and Search

Zhiyuan Shi, Timothy M. Hospedales, Tao Xiang
Queen Mary, University of London, London E1 4NS, UK
{z.shi,t.hospedales,t.xiang}@qmul.ac.uk

## Abstract

*Learning semantic attributes for person re-identification and description-based person search has gained increasing interest due to attributes' great potential as a pose and view-invariant representation. However, existing attribute-centric approaches have thus far underperformed state-of-the-art conventional approaches. This is due to their non-scalable need for extensive domain (camera) specific annotation. In this paper we present a new semantic attribute learning approach for person re-identification and search. Our model is trained on existing fashion photography datasets – either weakly or strongly labelled. It can then be transferred and adapted to provide a powerful semantic description of surveillance person detections, without requiring any surveillance domain supervision. The resulting representation is useful for both unsupervised and supervised person re-identification, achieving state-of-the-art and near state-of-the-art performance respectively. Furthermore, as a semantic representation it allows description-based person search to be integrated within the same framework.*

## 1. Introduction

Person re-identification (re-id) and description-based search are crucial tasks in visual surveillance. They underpin many fundamental applications including multi-camera tracking, crowd analysis and forensic search. Both tasks aim to retrieve images of a specific person, but differ in the query used. Person re-identification queries using an image from a different view (e.g., in multi-camera tracking), while person search uses a textual person description (e.g., eyewitness description). Despite extensive research [20, 54], these tasks remain unsolved due to various challenges including the variability of viewpoints, illumination, pose, partial occlusion, low-resolution and motion-blur [19].

Attributes, as a mid-level semantic representation, could potentially address these shared challenges, as they are intrinsically invariant to these variabilities in viewing conditions. Attribute-centric approaches are inspired by the operating procedure of human experts, and revolve around the idea of learning detectors for person attributes (such as Blue-Jeans). In particular, a number of approaches are developed for person re-identification using an attribute-based representation [28, 45, 30]. It was demonstrated [28] that *if* the each person's attributes can be reliably detected, re-id is essentially solved (since a $K$ binary attribute descriptor can potentially disambiguate $2^K$ people). For person search, since most textual descriptions used for query are attributes, an attribute-based approach is often the only option.

Despite their hoped-for potentials, attribute-centric approaches to person re-identification have until now not achieved state-of-the-art performance compared to conventional alternatives focused on learning effective low-level features and matching models [54]. We argue that this is largely due to the difficulty of obtaining sufficient and sufficiently detailed annotations to learn robust and accurate attribute detectors. In particular, to achieve good attribute detection, per camera/dataset annotations need to be obtained, since attribute models typically do not generalise well across cameras (e.g. a blue shirt may look purple in a different camera view). This is exacerbated, because unlike annotation of person identity used in learning a re-id matching model, attributes require multiple labels per image. Moreover since most human attributes are associated with a specific body part, to learn accurate attribute detectors, ideally annotation needs to be done at the patch-level rather than the image-level. In short, an attribute-based approach is limited by the scale of its annotation requirements.

Although attribute-centric approaches to surveillance are limited by the lack of annotation, extensive person attribute annotations already exist in other domains, notably fashion image analysis [34, 59, 60, 58, 35]. In this line of work, many high resolution images have been annotated with clothing properties – sometimes strongly (per-pixel). However, learning attribute detectors from the fashion domain, and applying them directly to person re-id and search will fail, due to the severe domain shift problem – compared with surveillance data, the image characteristics are completely different (see Fig. 2). In particular, many of the challenges in surveillance are absent (e.g. illumination vari-

1

ability, occlusion, low-resolution, motion blur). These large and perfectly annotated person attribute datasets are thus useless, unless an attribute model learned from them can be successfully adapted and transferred to the surveillance domain.

In this paper we contribute a new framework that is capable of learning a semantic attribute model from existing fashion datasets, and adapting the resulting model to facilitate person re-identification and search in the surveillance domain. In contrast to most existing approaches to attribute detection [28, 45] which are based on discriminative modelling, we take a generative modelling approach based on the Indian Buffet Process (IBP) [23]. The generative formulation provides key advantages including: joint learning of all attributes; ability to naturally exploit weakly-annotated (image-level) training data; as well as unsupervised domain adaptation through Bayesian priors. Importantly a IBP-based model [11, 50, 41] provides the favourable property of combining attributes factorially in each local patch. This means that our model can differentiate potentially ambiguous situations such as Red-Shirt+Blue-Jeans versus Red-Jeans+Blue-Shirt (See Fig. 3). Moreover, with this representation, attribute combinations that were rare or unseen at training time can be recognised at test time so long as they are individually known (e.g. Shiny-Yellow-Jeans).

Our framework overcomes the significant problem of domain shift between fashion and surveillance data in an unsupervised way by Bayesian adaptation. It can exploit both strongly and weakly annotated source data during training, but is always able to produce a strong (patch-level) attribute prediction during testing. The resulting representation is highly person variant while being view-invariant, making it ideal for person re-id, where we obtain state-of-the-art results. Moreover, as the representation is semantic (nameable or describable by a human), we are able to unify description based person search within the same framework, where we also achieve state-of-the-art results.

## 2. Related Work

**Person Re-identification:** Person re-identification is now a very well studied topic, with [20, 54] providing good reviews. Existing approaches can be broadly grouped according to two sets of criteria: supervision and representation/matching. Unsupervised approaches [65, 10, 55] are more practical in that they do not require the use of per-target camera annotation (labelling people into matching pairs), while supervised approaches [62, 8, 56, 25, 64, 27] that use this information tend to perform better. Studies also either focus on building a good representation [10, 22, 65, 43, 32, 55, 62], or building a strong matching model [56, 36, 57, 33, 25, 64]. In this paper we focus on learning mid-level semantic representations for both unsupervised and unsupervised re-id. Zhao et al. [66] also at-

tempted to learn mid-level representations. However their method does not learn a *semantic* representation, so it cannot be used for description-based person search; and it relies on supervised learning. Ours can be used with or without supervision, but provides the biggest benefit in the unsupervised context. A few studies [38, 67] address person re-id based on transfer learning. However, they transfer matching models rather than representations, and between different surveillance datasets rather different domains.

Semantic attribute-based representations have also been studied for person re-identification. A key motivation is that it is hard for low-level representations (e.g., colour/texture histograms) to provide both view-invariance and person-variance. Early studies train individual attribute detectors independently on images weakly-annotated with multi-label attributes [28, 37, 29], followed by using the estimated attribute vector of a test image as its representation for matching. More recent studies have considered joint learning of attributes and inter-attribute correlation [45, 30, 24]. Nevertheless, these approaches do not reach state-of-the-art performance. This is partly due to the drawbacks of (1) requiring ample target domain attribute annotations, which are hard to obtain at scale for learning robust detectors; and (2) producing coarse/weak (image-level) rather than strong (patch-level) annotations – due to the lack of strongly annotated training data that is even harder to obtain. In contrast, by transferring an attribute representation learned on existing large annotated fashion datasets, we overcome these issues and obtain state-of-the-art results.

**Person Search:** Related to attribute-based person re-identification is description-based person search [52, 12, 28, 14, 48]. In this case detectors for aspects of person description (e.g., clothes, soft-biometrics) are trained. Person images are then queried by description rather than by another image as in re-id. Most methods however have limited accuracy due to: (i) training on, and producing weak (image-level) annotations; and (ii) training and testing each attribute detector independently rather than jointly. For these reasons, they are also limited in being able to make complex attribute-object queries such as "Black-Jeans + Blue-Shirt", which requires a strong joint segmentation and attribute model to disambiguate the associations between attributes.

Similar notions of attribute-based search have been exploited in a face image context [26]; and in a fashion/e-commerce context, where users search for clothing by description [58, 7, 34, 5]. Face image search however, is easier due to being more constrained and well aligned than surveillance images. Meanwhile in the fashion context, clean high resolution images are assumed; and crucially strong pixel-level annotations are typically used, which would be prohibitively costly to obtain for individual surveillance camera views. In this paper we bridge the gap between the clean and richly annotated fashion domain, and the noisy and

scarcely annotated surveillance domain, to produce a powerful semantic representation for person re-id and search.

**Domain Adaptation:** A key challenge for our strategy is the domain-shift between the fashion and surveillance domains. Addressing a change of domains with domain adaptation is well studied in computer vision [13, 47, 4] and beyond [42]. To avoid necessitating target domain annotation, our task requires unsupervised adaptation which is harder. Some off-the-shelf solutions exist [13, 18], but these underperform due to working blindly in the low-level feature space, disconnected to the semantics being modelled, i.e. attributes. In contrast, in the style of [39, 6, 4], we achieve domain adaptation by transferring the source domain attribute model as a prior when the target domain model is learned. This enables adaptation to the target domain, while exploiting the constraints provided by semantic attribute model.

**Attribute Learning:** Beyond person re-id and search, semantic attribute learning [15, 69] is well studied in computer vision, which is too broad to review here. However, we note that most attribute-learning studies have one or more simplifications compared to the problem we consider here: They consider within-domain rather than cross-domain attributes; or produce coarse image-level attributes rather than segmenting the region of an attribute; or they do not address representing multiple attributes simultaneously on a single patch (important for e.g. a detailed clothing representation including, e.g., category + texture + colour).

Most other semantic attribute detection studies take a discriminative approach to learning for maximum detection accuracy [52, 28, 48, 59, 7, 34, 26]. In contrast, our generative approach provides advantages in more naturally supporting weakly-supervised learning, and domain transfer through Bayesian priors. Related generative models include [55] which used unsupervised topic models to estimate saliency for re-id, and [16, 11] which addressed data driven attribute discovery and learning. In a more related work, [50] proposed a weakly supervised approach which models attribute and object associations using a generalised IBP model. However [50] (i) is designed for a single domain rather than for cross-domain use, and (ii) has no notion of spatial coherence which is important for correctly segmenting fine person attribute details – while we integrate a Markov Random Field (MRF) into the latent space of the IBP model for more accurate person segmentation.

**Contributions:** Overall, the contributions of our model are as follows: (i) We introduce a generative framework for person attribute learning that can be learned from strongly or weakly annotated data or a mix; (ii) We show how to perform domain adaptation from fashion to surveillance domain; (iii) We demonstrate that the resulting representation is useful for both unsupervised and supervised re-identification where it achieves state-of-the-art, and near state-of-the-art results respectively; and (iv) The same
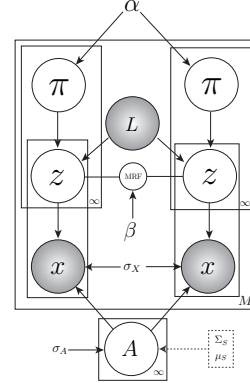


Figure 1: The graphical model for MRF-IBP. Shaded nodes are observed.

framework enables better description-based person search compared against the existing discriminative modelling approaches.

## 3. Semantic Representation Learning

Our model, termed MRF-IBP, works with super-pixel segmented person images. Each super-pixel is associated with a $K$-dimensional latent binary vector whose elements (factors) indicate what set of attribute properties are possessed by that patch, and which of a potentially infinite set of background clutter types are present. The first $K_s$ factors are associated with known annotations, while the subsequent entries are free to model unannotated aspects of the images. Our pipeline exploits two datasets: an annotated auxiliary dataset, and an unannotated target dataset:

**Auxiliary Training:** First we train on the auxiliary/source dataset using weak (image-level) or strong (patch-level) supervision. The supervision is a binary vector describing which attributes appear in the image/patch. For strong supervision annotations $L_j^{(i)}$ are given for the first $K_s$ factors of patches $j$ in image $i$, and the model learns from this each of the $K_s$ factors' appearance. The $K_s$ supervised factors also include foreground versus background patch annotation. For weak supervision, annotations $L^{(i)}$ are given for the first $K_s$ factors of each image, and the model solves the (more challenging task) of learning both factor appearance and infer which image-level factors occur in which patches.

**Target Adaptation:** We then use the learned parameters from the auxiliary set as a prior, and adapt it to the target dataset without any supervision. The representation learned here is then used for re-identification and person-search.

### 3.1. Model Formulation

Each image $i$ is represented as a bag of patches $\mathbf{X}^{(i)} = \{\mathbf{X}_{j\cdot}^{(i)}\}$, where $\boldsymbol{X}_{j\cdot}$ means the vector of row $j$ in matrix $\boldsymbol{X}$ and corresponds to a $D$-dimensional feature representation of each patch. Without supervision, the

generative process for each image is as follows:

For each latent factor $k \in 1 \ldots \infty$:

1. Draw an appearance distribution mean $\mathbf{A}_{k\cdot} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$.

For each image $i \in 1 \ldots M$:

1. Draw the binary indicator matrix $\mathbf{Z}$ describing the factor activation for every patch:

$$
\begin{aligned}
p(\mathbf{Z}^{(i)}|\alpha, \beta) \quad &\propto \quad \frac{\alpha^{K_+}}{\prod_{j=1}^{N_i} K_1^j!} \cdot \exp{-\alpha \sum_{j=1}^{N_i} \frac{1}{j}} \\
&\cdot \prod_{k=1}^{K_+} \frac{\left(N_i - m_k^{(i)}\right)!(m_k^{(i)}-1)!}{N_i!} \\
&\cdot \prod_{k=1}^{K_+} \exp\left(\sum_{j=1}^{N_i} \beta \sum_{j' \in N(j)} I(z_{jk}^{(i)} = z_{j'k}^{(i)})\right) \quad (1)
\end{aligned}
$$

2. For each super-pixel patch $j \in 1 \ldots N_i$: Sample patch appearance: $\boldsymbol{X}_{j\cdot}^{(i)} \sim \mathcal{N}(\boldsymbol{Z}_{j\cdot}^{(i)} \boldsymbol{A}, \sigma_X^2 \boldsymbol{I})$.

Notations: $\mathcal{N}$ corresponds to Normal distribution with the specified parameters; $\alpha$ is the prior expected sparsity of annotations; $\beta$ is the coupling strength of the inter-patch MRF; $\boldsymbol{\mu}_k$ and $\Sigma_A$ are the prior mean and covariance of each factor. $p(\boldsymbol{Z}^{(i)})$ in Eq. (1) corresponds to our MRF-IBP prior. It expresses the IBP sampling of an unbounded number of factors in each patch (first two lines), which are spatially correlated across patches by a Potts model MRF (last line) like [53]. Here $K_+ \geq K_s$ refers to the (inferred) number of active factors in the image, $K_1^j$ is the factor history [23], and $m_k^{(i)}$ is the number of times each factor $k$ is active.

Denote the set of hidden variables by $\boldsymbol{H} = \{\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(M)}, \boldsymbol{A}\}$, observed images by $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(M)}\}$, and model parameters by $\boldsymbol{\Theta} = \{\alpha, \beta, \sigma_X, \Sigma_k, \boldsymbol{\mu}_k\}$. Then the joint probability of the variables and data given the parameters is:

$$
\begin{aligned}
p(\boldsymbol{H}, \boldsymbol{X}|\Theta) \quad = \quad &\prod_{k=1}^{\infty} p(\mathbf{A}_{k\cdot}|\boldsymbol{\mu}_k, \Sigma_k) \\
&\cdot \prod_{i=1}^{M} p(\mathbf{Z}^{(i)}; \alpha, \beta) \prod_{j=1}^{N_i} p(\boldsymbol{X}_{j\cdot}^{(i)}|\boldsymbol{Z}_{j\cdot}^{(i)}, \boldsymbol{A}, \sigma_X) \quad (2)
\end{aligned}
$$

Learning in our model aims to compute the posterior $p(\boldsymbol{H}|\boldsymbol{X}, \boldsymbol{\Theta})$ for: discovering which factors (object/attributes) are active on each patch (inferring $\boldsymbol{Z}^{(i)}$), and learning the appearance of each factor (inferring $\mathbf{A}_{k\cdot}$).

## 3.2. Model learning from the Auxiliary Set

To learn our MRF-IBP model, we exploit Gibbs sampling for approximate inference inspired by [68]. For Gibbs sampling, we need to derive an update for each hidden variable conditional on the observations and all the other hidden variables.

**Unsupervised Factor Updates:** For all initialised factors $k$, we can sample the state of each latent factor $z_{jk}^{(i)}$ via:

$$
\begin{aligned}
p(z_{jk}^{(i)} = 1|\boldsymbol{Z}_{-jk}^{(i)}, \boldsymbol{X}_{j\cdot}^{(i)}) &\propto p(\boldsymbol{X}_{j\cdot}^{(i)}|\boldsymbol{Z}^{(i)}) P(z_{jk}^{(i)} = 1|\boldsymbol{Z}_{-jk}^{(i)}) \\
&= \frac{m_k^{(i)} - z_{jk}^{(i)}}{N_i} \cdot \exp \sum_{j' \in N(j)} \beta I(z_{jk}^{(i)} = z_{j'k}^{(i)}) \\
&\cdot \exp(-\frac{1}{2\sigma_X^2} \mathrm{tr}(\boldsymbol{X}_{j\cdot}^{(i)} - \boldsymbol{Z}_{j\cdot}^{(i)} \boldsymbol{A})^T (\boldsymbol{X}_{j\cdot}^{(i)} - \boldsymbol{Z}_{j\cdot}^{(i)} \boldsymbol{A})) \quad (3)
\end{aligned}
$$

where $\boldsymbol{Z}_{-jk}^{(i)}$ denotes the entries of $\boldsymbol{Z}^{(i)}$ other than $\boldsymbol{Z}_{jk}^{(i)}$. To sample new latent factors, Poisson$(\frac{\alpha}{N_i})$ is used as the expected number of new classes [23].

**Supervised Factor Updates:** Eq. (3) describes inference in the case where no supervision is available. If strong supervision $L_{jk}^{(i)}$ is available, Eq. (3) is replaced with $z_{jk}^{(i)} = L_{jk}^{(i)}$. If weak supervision $L_k^{(i)}$ is available Eq. (3) is replaced with $p(z_{jk}^{(i)} = 1|\boldsymbol{Z}_{-jk}^{(i)}, \boldsymbol{X}_{j\cdot}^{(i)}) \propto p(\boldsymbol{X}_{j\cdot}^{(i)}|\boldsymbol{Z}^{(i)}) P(z_{jk}^{(i)} = 1|\boldsymbol{Z}_{-jk}^{(i)}) \cdot L_k^{(i)}$

**Appearance Updates:** In order to sample factor appearance $\boldsymbol{A}$, we compute its Gaussian posterior $p(\boldsymbol{A}|\boldsymbol{X}, \boldsymbol{Z})$:

$$
\begin{aligned}
\mu_S &= (\tilde{\boldsymbol{Z}}^T \tilde{\boldsymbol{Z}} + \frac{\sigma_X^2}{\sigma_A^2} I)^{-1} \tilde{\boldsymbol{Z}}^T \tilde{\boldsymbol{X}} \\
\Sigma_S &= \sigma_X^2 (\tilde{\boldsymbol{Z}}^T \tilde{\boldsymbol{Z}} + \frac{\sigma_X^2}{\sigma_A^2} I)^{-1} \quad (4)
\end{aligned}
$$

where $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{X}}$ are the matrices that vertically concatenate the factor matrix and patch feature matrix for all images. Here $\mu_S$ is the $K_+ \times D$ matrix of appearance for each factor, and $\Sigma_S$ is the $K_+ \times K_+$ matrix of variance parameters for each factor appearance. Since this is the auxiliary set, we have assumed an uninformative prior, i.e., $\boldsymbol{A}_k \sim \mathcal{N}(0, \sigma_A)$.

## 3.3. Model Adaptation to the Target Set

In the target set, there is no supervision, so Eq. (3) is used to update the latent factors. The appearance updates however are changed to reflect the top-down influence from the learned auxiliary domain. Thus the target appearance $\mu_T$ is updated using the sufficient statistics from the source $\Sigma_S$ and $\mu_S$ (Eq. (4)) as the prior:

$$
\begin{aligned}
\mu_T &= \Sigma_T (\sigma_X^{-2} \tilde{\boldsymbol{Z}}^T \tilde{\boldsymbol{X}} + \Sigma_S^{-1} \mu_S) \\
\Sigma_T &= \sigma_X^2 (\tilde{\boldsymbol{Z}}^T \tilde{\boldsymbol{Z}} + \sigma_X^2 \Sigma_S^{-1})^{-1} \quad (5)
\end{aligned}
$$

| Method | | VIPeR | | | | CUHK01 | | | | PRID450S | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| single | SDC [65] | 25.1 | 44.9 | 56.3 | 70.9 | 15.1 | 25.4 | 31.8 | 40.9 | 23.7 | 38.4 | 46.1 | 58.5 |
| | GTS [55] | 25.2 | 50.0 | 62.5 | 75.6 | - | - | - | - | - | - | - | - |
| | SDALF [10] | 19.9 | 38.9 | 49.4 | 65.7 | 9.9 | 22.6 | 30.3 | 41.0 | 17.4 | 30.9 | 40.8 | 55.2 |
| | Our unsupervised | **27.7** | **55.3** | **68.3** | **79.7** | **23.3** | **35.8** | **46.6** | **60.7** | **28.5** | **48.9** | **59.6** | **71.3** |
| fused | SDC_*Final* (eSDC) [65] | 26.7 | 50.7 | 62.4 | 76.4 | 19.7 | 32.7 | 40.3 | 50.6 | 25.5 | 40.6 | 48.4 | 61.4 |
| | Our unsupervised_*Final* | **29.3** | **52.7** | **66.8** | **79.7** | **22.4** | **35.9** | **47.9** | **64.5** | **29.0** | **49.4** | **58.4** | **69.8** |

Table 1: Matching accuracy @ rank r (%): unsupervised learning approaches. '-' indicates no result was reported and no code is available for implementation. The best results for single-cue and fused-cue methods are highlighted in bold separately.

## 4. Semantic Representation Applications

After the semantic representation learning described previously, each target image $i$ is now described by a binary factor matrix $\boldsymbol{Z}^i$ containing the inferred factor vector $\{\mathbf{z}_j\}_{j=1}^{N_i}$ for each super pixel $j$. We could use this representation directly, but find it convenient to convert it into a fixed-size representation per image. We thus generate multiple heat-maps $M_k$ per image representing the $k$th factor activation. Similar to [56], we divide each image into 14 overlapping patches sampled on a $2 \times 7$ regular grid with a $32 \times 32$ window.[1] Each grid-patch is now represented by a $K$-dimensional attribute vector obtained by summing $\boldsymbol{z}_j$ for very pixel.

### 4.1. Person Re-identification

The proposed semantic person representation can be used for both unsupervised or supervised re-identification, according to whether a matching model is learned from the identity annotation of a given person re-id dataset.

**Unsupervised Matching:** Each image is represented by 14 patches each with a $K$ dimensional descriptor. The person match is now converted to a semantic patch matching problem. We adopt a patch matching algorithm TreeCANN [40] to efficiently compute the distance between images.

**Supervised Matching:** The 14 patch descriptors are concatenated to obtain an image-level descriptor. This is used as input to a recent metric learning algorithm kLFDA [56].

### 4.2. Person Search

Recall that K heat maps $M_k$ are generated from the K factors. The probability of factor $k$ appearing in an image can be obtained by $max(M_k)$. When querying two or more factors, there are two possible query semantics: To query the probability of two factors both appearing anywhere in an image without preference for co-location (e.g., Coat + Bag), we use $max(M_k) \cdot max(M_{k'})$. In contrast, to query two factors that should *simultaneously* appear in the same place (e.g., Blue-Jeans), we use $max(M_k \cdot M_{k'})$.

---

[1] Note that the overlapping area between two neighbouring patches depends on the size of the image.

## 5. Experiments

### 5.1. Datasets and Settings

**Auxiliary Datasets:** Two datasets are used as auxiliary sources. **Colourful-Fashion** [34] includes 2682 images. Pixel-level annotation is provided with 13 colour labels (e.g., brown, red) and 23 category labels (e.g., bag, T-shirt). Most of the images contain a single person with a relatively simple pose, against relatively clean background (see Fig. 2). **Clothing-Attribute** [7] includes 1,856 person images from social media sites, annotated with 26 attributes. Only image-level annotations are provided. However, it includes 6 texture attributes not included in Colourful-Fashion, so we include this auxiliary dataset mainly to enrich the representation with the 6 texture attributes.

**Target Datasets:** Four surveillance pedestrian datasets are used as target data. **VIPeR** [21] contains two views of 632 pedestrians. All images are normalised to $128 \times 48$. All images are also manually labeled by [48] with 22 attributes, named VIPeR-Tag dataset [48]. **CUHK01** [31] is captured with two camera views in a campus environment. It contains 971 persons, with two images each. Images are normalised to $160 \times 60$. **PRID450S** [46] is a recent and more realistic dataset, built on PRID 2011. It consists of 450 image pairs recorded from two static surveillance cameras. All images are normalised to $168 \times 80$. **PETA** [9] is a large-scale surveillance person attribute datatset that consists of 19000 images. Each image is labelled with 61 binary and 4 multi-class attributes, including colour, style etc.

**Features:** We divide the image into super-pixels using a recent segmentation algorithm [2]. We represent each super-pixel as a vector using following features: (1) Colour: We extract 3 dimensional colour descriptors from each pixel in both RGB and LAB colour space [17, 62]. We run k-means to obtain 150 code words for each colour space. Pixels are quantised to the nearest centres in the visual vocabulary. The resulting descriptor for each super-pixel is the normalised histogram over visual words. (2) SIFT: We compute 128 dimensional dense SIFT over a regular grid ($4 \times 4$ step size). Similar to Colour, we build a vocabulary of 300 words [34]. A histogram is built from quantised local words within each super-pixel. (3) Location: Following

| | Method | VIPeR | | | | CUHK01 | | | | PRID450S | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| single | MLF [66] | 29.1 | 52.3 | 65.9 | 79.9 | **34.3** | **55.1** | **65.0** | 74.9 | - | - | - | - |
| | KML [56] | 32.3 | 65.8 | 79.7 | 90.9 | 24.0 | 38.9 | 46.7 | 55.4 | 32.4 | 54.4 | 62.4 | 69.6 |
| | KISSME [25] | 19.6 | 48.0 | 62.2 | 77.0 | 8.4 | 25.1 | 38.7 | 50.2 | 26.5 | 47.8 | 57.6 | 68.5 |
| | SCNCD [62] | **33.7** | 62.7 | 74.8 | 85.0 | - | - | - | - | 41.5 | 66.6 | 75.9 | 84.4 |
| | FUSIA [29] | 19.1 | 55.3 | 73.5 | 84.8 | 9.8 | 32.4 | 49.8 | 60.1 | - | - | - | - |
| | Our supervised | 31.1 | **68.6** | **82.8** | **94.9** | 32.7 | 51.2 | 64.4 | **76.3** | **43.1** | **70.5** | **78.2** | **86.3** |
| fused | KML_*Final* [56] | 36.1 | 68.7 | 80.1 | 85.6 | - | - | - | - | - | - | - | - |
| | SCNCD_*Final* [62] | 37.8 | 68.6 | 81.0 | 90.5 | - | - | - | - | 41.6 | 68.9 | **79.4** | **87.8** |
| | MLF_*Final* [66] | **43.4** | **73.0** | 84.9 | 93.7 | - | - | - | - | - | - | - | - |
| | Our supervised_*Final* | 41.6 | 71.9 | **86.2** | **95.1** | 31.5 | 52.5 | **65.8** | 77.6 | **44.9** | **71.7** | 77.5 | 86.7 |

Table 2: Matching accuracy @ rank r (%): supervised learning approaches on re-id.

[34, 61], we consider a 2 dimensional coordinate of each super-pixel centroid as an absolute location feature. A relative location feature is defined by the distances between the centroid and each of 26 human key points generated by human pose estimation [61], giving a 106 dimensional location features. The final feature vector (706D) of each super-pixel is formed by concatenation of Colour (300D), SIFT (300D) and Location (106D). To compensate for the noise in the surveillance images, we also apply a rolling guidance filter [63] before generating super-pixels.

**Settings:** For training the auxiliary datasets, we use 60 supervised factors: 34 from Colourful-Fashion (12 colour + 22 category attributes), 6 (texture) from Clothing-Attribute, and 20 background factors (always off for foreground patches). Thus our model activates at least $K_+ \geq K_s = 60$ factors, although more may be used to explain un-annotated aspects of the data due to the use of IBP. We train by iterating Eqs. (3) and (4) for 2000 iterations. The supervision used varies across the strongly and weakly annotated auxiliary sets. Please see supplementary material for details. For transferring to the re-id datasets, we transfer the 60 auxiliary domain factors, and use $K_+ \geq 80$ by initialising a further 20 free factors randomly to accommodate new factors in the new domain. Any previously unseen unique aspects of the target domain can be modelled by these 20 factors. We adapt the learned model to the target data by iterating Eqs. (3) and (5) for 100 iterations. We then take the first $K = 80$ learned factors to produce an 80-dimensional patch representation (see Sec. 4) to be used in person re-id.

**Baselines:** In addition to comparing with start-of-the-art in person re-id and person search methods, we also consider alternative transfer methods that could potentially generate an analogous representation to our framework: **SVM/MI-SVM**: SVM (as in [28, 48]) and Multi-Instance SVM [1] are used to train patch-level attribute classifiers for strongly and weakly-labelled auxiliary data respectively. The learned SVMs can then be applied to estimate feature vectors for each target image patch similarly to our model. **DASA** [13]: An unsupervised domain adaptation methods to address domain shift by aligning the source and target subspaces.

**Computational Cost:** The complexity of our algorithm is $O(MN(K^2 + KD))$ for $M$ images with $N$ super pixels, $K$ factors, and $D$-dimensional patch features. We run our algorithm on a PC with Intel 3.47 GHz CPU and 12GB RAM. In practice this corresponds to 1 to 2 minutes for 1000 images per iteration, depending on the number of super-pixels.

## 5.2. Person Re-identification

We first evaluate person re-id performance against start-of-the-arts [56, 22, 25]. We randomly divide the dataset into two equal, non-overlapping subsets for training and testing. We use the widely used accuracy at rank $k$ of Cumulative Match Characteristic (CMC) curves to quantify performance. The results are obtained on VIPeR, CUHK01 and PRID450S datasets by averaging over 10 random splits. We distinguish using the suffix _*Final* the common practice of use of an ensemble of methods or features with score or feature level fusion.

**Unsupervised Matching:** We compare our model to recent state-of-the-art approaches under an unsupervised setting (i.e. no identity labels are used) including SDALF [10], eSDC [65], and GTS [55]. As shown in Table 1, our representation on its own significantly outperforms all other methods in all three datasets, and is not far off the most competitive supervised methods (Table 2). When fused with SDALF as in [65], performance improves further. See supplementary material for CMC curves and more comparisons.

**Supervised Matching:** Table 2 compares our method in a supervised matching context against recent state-of-the-art including: MLF [66], KML [56], KISSME [25], SCNCD [62], FUSIA [29]. It shows that our approach achieves comparable or better performance to state-of-the-art, especially at higher rank (i.e. r=5,10,20). In this setting our final result is obtained by fusing with kLFDA [56].

**Auxiliary Data:** Here we evaluate the effects of various auxiliary data sources and annotations. Our full framework is learned with fully-supervised (f-F) Colourful-Fashion and weakly-supervised (w-C) Clothing-Attribute datasets. Table 3 (on ViPeR) shows that: (i) the different annotations

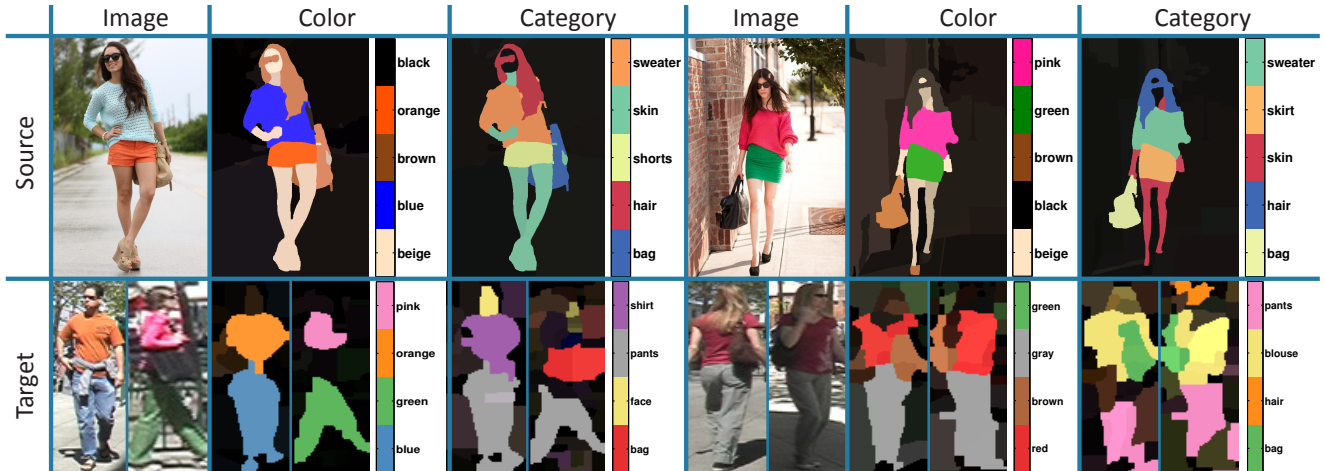| Image | Color | Category | Image | Color | Category |
|---|---|---|---|---|---|

Figure 2: Visualisation of our model output. Each patch is colour-coded to show the inferred dominant attribute of two types.

in the two auxiliary datasets are combined synergistically, and weakly-annotated data can be used effectively (f-F+w-C > f-F); and (ii) while capable of exploiting strong supervision where available, our framework does not critically rely on it (w-F+w-C close to f-F+w-C; w-F close to f-F).

| Auxiliary Data | Unsupervised | | | | Supervised | | | |
|---|---|---|---|---|---|---|---|---|
| | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| w-F | 18.3 | 38.3 | 49.5 | 62.9 | 26.2 | 58.2 | 71.1 | 83.4 |
| f-F | 25.4 | 51.4 | 63.9 | 75.3 | 29.4 | 64.9 | 78.8 | 91.7 |
| w-F + w-C | 22.4 | 43.6 | 57.1 | 67.3 | 28.3 | 62.2 | 75.8 | 88.5 |
| f-F + w-C | **27.7** | **55.3** | **68.3** | **79.7** | **31.1** | **68.6** | **82.8** | **94.9** |

Table 3: Effects of auxiliary data source and annotation.

**Contributions of Components:** To evaluate the contributions of each component of our framework, Table 4 summarises our model performance on ViPeR in 4 conditions: (1) Without MRF (NoMRF); (2) Direct transfer without adaptation (Eq. 5) (NoAdapt); (3) (1) & (2); (4) Solely unsupervised target domain learning (NoTransfer). The results show that each component (MRF modelling, transfer and adaptation) contributes to the final performance.

| Method | Unsupervised | | | | Supervised | | | |
|---|---|---|---|---|---|---|---|---|
| | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| NoMRF | 23.3 | 47.4 | 59.1 | 70.9 | 28.0 | 62.1 | 75.2 | 87.2 |
| NoAdapt | 19.2 | 39.6 | 50.2 | 61.9 | 21.8 | 49.2 | 60.9 | 73.8 |
| NoMrfAdapt | 17.7 | 36.2 | 45.5 | 54.8 | 20.2 | 46.0 | 57.6 | 70.9 |
| NoTransfer | 9.5 | 20.5 | 26.9 | 35.6 | 14.3 | 32.9 | 41.7 | 52.5 |
| Ours | **27.7** | **55.3** | **68.3** | **79.7** | **31.1** | **68.6** | **82.8** | **94.9** |

Table 4: Contribution of each model component

**Alternative Transfer Approaches:** Our model is compared against alternative SVM-based approaches. Table 5 reveals that: (i) While (MI)SVMs can in principle deal with weakly or strongly supervised representation learning, it clearly under-performs our approach, and (ii) Although conventional feature-level domain adaptation (DASA [13]) can improve the SVM performance, it is much less effective than our model-level adaptation.

| | Method | Unsupervised | | | | Supervised | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| w-F | MI-SVM | 8.0 | 17.8 | 24.4 | 34.4 | 15.6 | 36.2 | 46.5 | 59.9 |
| | DASA [13] | 12.2 | 25.8 | 33.9 | 43.7 | 17.1 | 39.2 | 49.4 | 61.5 |
| | Ours | **18.3** | **38.3** | **49.5** | **62.9** | **26.2** | **58.2** | **71.1** | **83.4** |
| f-F | SVM | 13.2 | 29.6 | 40.3 | 55.4 | 17.4 | 40.5 | 51.9 | 66.8 |
| | DASA [13] | 16.0 | 33.5 | 42.8 | 53.2 | 20.8 | 47.7 | 60.2 | 75.4 |
| | Ours | **25.4** | **51.4** | **63.9** | **75.3** | **29.4** | **64.9** | **78.8** | **91.7** |

Table 5: Comparing different transfer learning approaches

**What is learned:** Fig. 2 visualises after model learning how attributes are detected given a new image. We visualise the top 5 most confident colour and category factors/attributes for each image in the test set of Colourful-Fashion. Our model can almost perfectly recognise and localise the attributes (top row). As expected, the inferred attributes are much more noisy for the re-id data (bottom row). However, overall they are accurate (e.g. bags of different colours are detected), and crucially provide a much stronger representation than the even noisier low-level features.

### 5.3. Person Search

Although attribute-based query is a widely studied problem, there are few studies on person search [12, 52] in surveillance. To evaluate description-based surveillance person search, we conduct experiments on VIPeR-Tag [48] and PETA [9]. For both datasets, following [48], we randomly chose 50% of the data for training (not used in our transfer framework, but used in other baselines) and the remaining for testing, and repeat this procedure 10 times to obtain average results. Person Search is a retrieval task, so we evaluate the performance of each query with a precision-recall curve like [48, 51, 3].

In VIPeR-Tag, all 15 queries used in [48] are contained in our source data attribute list. We can thus directly compare with the results in [48]. The 15 queries are composed of a combination of an adjective (A) and a noun (N) (e.g. Red-Shirt). To ensure two query terms co-exist in the

|  | Red-Shirt | Blue-Trousers | Red-Shirt + Blue-Trousers |
|---|---|---|---|
| Our | | | |
| SVM | | | |

Figure 3: Person search qualitative results. The top ranked images for each query are shown. Red boxes are false detections.
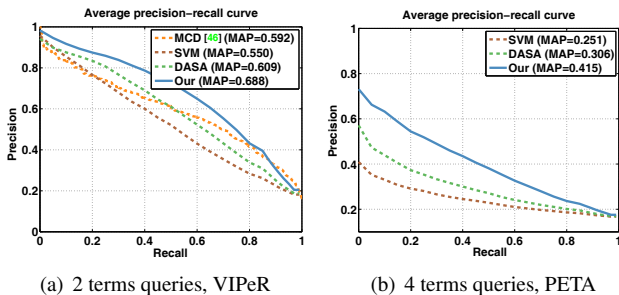


(a) 2 terms queries, VIPeR

(b) 4 terms queries, PETA

Figure 4: Person search: comparison with state-of-the-art.

same patch, we use $max(M_k \cdot M_{k'})$ to compute the score (see Sec. 4.2). Fig. 4(a) shows the average PR curves over all annotated queries, and it is clear that our method outperforms MCD [48] [2], SVM [28] and an unsupervised domain adaptation-based method DASA [13], even though no annotated VIPeR-Tag data is used for learning our model. Similar to [44], SVM scores have been calibrated by [49] before being fused to generate probability estimates of queries. More detailed results (including query-specific PR curves) are available in supplementary material.

In PETA, we consider a more challenging search task. Each query contains 4 terms of the form A-N+A-N (e.g. Red-Shirt+Blue-Trousers). We select all multi-class attribute labels of PETA [9], including 11 colour (A) and 4 categories (N). In total 44 A-N combinations are generated and any two of them can form a 4-term query. Like [44], we randomly generate 200 4-term queries to evaluate the methods. Note that as the query form is A-N+A-N, the two query strategies in Sec. 4.2 need to be combined to compute a score. Fig. 4(b) shows that our method outperforms alternatives by a larger margin in this more difficult query setting.

Our model has two important advantages over the compared existing methods: (1) In order to better detect conjunctive person attributes such as "Red-Shirt", many exist-

ing methods [48] train a single attribute classifier for each combination of interest. This is not scalable because there will always be rare combinations that have too few instances to train a reliable classifier for; or at test time a combination may be required that no classifier has been trained for. By representing person attributes factorially, our model has no problem searching for combinations of attributes unseen at train time. (2) Because attributes are represented conjunctively at the patch-level, we can make complex queries such as (Black-Jeans + Blue-Shirt). An existing method such as the SVM-based one in [28], which uses image-level predictions for each attribute independently, may be confused by "Blue-Jeans + Black-Shirt" as an alternative. This explains the larger performance margin on PETA. Fig. 3 gives some qualitative illustration of these advantages. For example, Fig. 3 shows that the SVM-based model in [28], learned on each attribute separately at the image level, wrongly detects a person with blue top when we query blue trousers. This limitation is more apparent for the more challenging "Red-Shirt+Blue-Trousers" query. In contrast, with patch-based joint attribute modelling, our model achieves much better results.

## 6. Conclusion

We have introduced a framework to generate semantic attribute representations of surveillance person images for re-id and search. Our framework exploits weakly and/or strongly annotated source data from other domains and transfers it with adaptation to obtain a good representation without any target domain annotation. The resulting patch-level semantic representation obtains competitive performance for supervised re-id, and state-of-the-art performance for unsupervised re-id – which is the more practically relevant problem contexts since camera specific identity annotation is not scalable. Moreover as a semantic representation it allows unification of re-id and person search within the same model.

---

[2] Various MCD versions are evaluated in [48]. We compare with MCD$_1$ which gives the best MAP.

# References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2003. 6

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011. 5

[3] A. Babenko, A. Slesarev, A. Chigorin, and Lempitsky. Neural codes for image retrieval. In *ECCV*, 2014. 7

[4] W. Bian, D. Tao, and Y. Rui. Cross-domain human action recognition. *Systems, Man, and Cybernetics, IEEE Transactions on*, 2012. 3

[5] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011. 2

[6] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *CVPR*, 2010. 3

[7] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 2, 3, 5

[8] A. Das, A. Chakraborty, and A. Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, 2014. 2

[9] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, 2014. 5, 7, 8

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2, 5, 6

[11] J. Feng, S. Jegelka, S. Yan, and T. Darrell. Learning scalable discriminative dictionary with sample relatedness. In *CVPR*, 2014. 2, 3

[12] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, 2014. 2, 7

[13] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 3, 6, 7, 8

[14] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: Retrieving people using their pose. In *CVPR*, 2009. 2

[15] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2008. 3

[16] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*, 2012. 3

[17] T. Gevers and A. Smeulders. Color based object recognition. *Pattern Recognition*, 1997. 5

[18] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 3

[19] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales. The re-identification challenge. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-Identification*. Springer London, 2014. 1

[20] S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors. *Person Re-Identification*. Springer, 2014. 1, 2

[21] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007. 5

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

[23] T. L. Griffiths and Z. Ghahramani. The indian buffet process: An introduction and review. *JMLR*, 2011. 2, 4

[24] S. Khamis, C.-H. Kuo, V. K. Singh, V. Shet, and L. S. Davis. Joint learning for attribute-consistent person re-identification. In *ECCV Workshop on Visual Surveillance and Re-Identification*, 2014. 2

[25] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2, 6

[26] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008. 2, 3

[27] C.-H. Kuo, S. Khamis, and V. Shet. Person re-identification using semantic color names and rankboost. In *WACV*, 2013. 2

[28] R. Layne, T. Hospedales, and S. Gong. *Person Re-identification*, chapter Attributes-based Re-identification. Springer, 2014. 1, 2, 3, 6, 8

[29] R. Layne, T. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014. 2, 6

[30] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan. Clothing attributes assisted person re-identification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2014. 1, 2

[31] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 5

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

[33] 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*, June 2013. 2

[34] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *Multimedia, IEEE Transactions on*, 2014. 1, 2, 3, 5, 6

[35] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012. 1

[36] 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

[37] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, 2012. 2

[38] A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013. 2

[39] A. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In *ICML*, 1998. 3

[40] I. Olonetsky and S. Avidan. Treecann - k-d tree coherence approximate nearest neighbor algorithm. In *ECCV*, 2012. 5

[41] K. Palla, D. Knowles, and Z. Ghahramani. Relational learning and network modelling using infinite latent attribute models. *TPAMI*, 2014. 2

[42] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 2010. 3

[43] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 2

[44] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013. 8

[45] J. Roth and X. Liu. On the exploration of joint attribute learning for person re-identification. In *ACCV*, 2014. 1, 2

[46] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*. Springer, 2014. 5

[47] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 3

[48] R. Satta, G. Fumera, and F. Roli. People search with textual queries about clothing appearance attributes. In *Person Re-Identification*. Springer, 2014. 2, 3, 5, 6, 7, 8

[49] W. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012. 8

[50] Z. Shi, Y. Yang, T. M. Hospedales, and T. Xiang. Weakly supervised learning of objects, attributes and their associations. In *ECCV*, 2014. 2, 3

[51] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 7

[52] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009. 2, 3, 7

[53] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007. 4

[54] R. Vezzani, D. Baltieri, and R. Cucchiara. People re-identification in surveillance and forensics: a survey. *ACM Computing Surveys*, Dec. 2013. 1, 2

[55] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 2, 3, 5, 6

[56] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 2, 5, 6

[57] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 2

[58] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 1, 2

[59] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1, 3

[60] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014. 1

[61] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *TPAMI*, 2013. 6

[62] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Li. Salient color names for person re-identification. In *ECCV*, 2014. 2, 5, 6

[63] Q. Zhang, X. Shen, L. Xu, and J. Jia. Rolling guidance filter. In *ECCV*, 2014. 6

[64] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 2

[65] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2, 5, 6

[66] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identfiation. In *CVPR*, 2014. 2, 6

[67] W.-S. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012. 2

[68] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *CVPR*, 2011. 4

[69] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 3