Attribute Recognition by Joint Recurrent Learning of Context and Correlation

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, wei.li}@qmul.ac.uk eddy@visionsemantics.com

Abstract

Recognising semantic pedestrian attributes in surveillance images is a challenging task for computer vision, particularly when the imaging quality is poor with complex background clutter and uncontrolled viewing conditions, and the number of labelled training data is small. In this work, we formulate a Joint Recurrent Learning (JRL) model for exploring attribute context and correlation in order to improve attribute recognition given small sized training data with poor quality images. The JRL model learns jointly pedestrian attribute correlations in a pedestrian image and in particular their sequential ordering dependencies (latent high-order correlation) in an end-to-end encoder/decoder recurrent network. We demonstrate the performance advantage and robustness of the JRL model over a wide range of state-of-the-art deep models for pedestrian attribute recognition, multi-label image classification, and multi-person image annotation on two largest pedestrian attribute benchmarks PETA and RAP.

1. Introduction

Pedestrian attributes, e.g. age, gender, and hair style are humanly searchable semantic descriptions and can be used as soft-biometrics in visual surveillance, with applications in person re-identification [20, 26, 32], face verification [18], and human identification [35]. An advantage of attributes as semantic descriptions over low-level visual features is their robustness against viewpoint changes and viewing condition diversity. However, it is inherently challenging to automatically recognise pedestrian attributes from real-world surveillance images because: (1) The imaging quality is poor, in low resolution and subject to motion blur (Fig. 1); (2) Attributes may undergo significant appearance changes and situate at different spatial locations in an image; (3) Labelled attribute data from surveillance images are difficult to collect and only available in small numbers. These factors render learning a pedestrian attribute model very difficult. Early attribute recognition methods mainly rely on hand-crafted features like colour and texture [20, 26, 15, 7]. Recently, deep learning based attribute models have started to gain attraction [21, 33, 10, 9], due to deep model's capacity for learning more expressive representations *when large scale data is available* [17, 2, 39]. However, large scale training data is not available for pedestrian attributes. The two largest pedestrian attribute benchmark datasets PETA [7] and RAP [22] contain only $9,500$ and $33,268$ training images, much smaller than the popular ILSVRC (1.2 million) [36] and MS COCO ($165,482$) datasets [25]. Deep learning of pedestrian attributes is further compounded by degraded fine-grained details due to poor image quality, low resolution and complex appearance variations in surveillance scenes.

Figure 1. Pedestrian attribute recognition in poor surveillance images is inherently challenging due to ambiguous visual appearance from low resolution and large variations in human pose and viewing conditions, e.g. illumination, background clutter, occlusion.

To address these difficulties, one idea is to discover the interdependency and correlation among attributes [3, 21, 50, 47, 48], e.g. two attributes "female" and "skirt" are likely to co-occur in a person image. This correlation provides an inference constraint complementary to visual appearance recognition. Another idea is to explore visual context as an extra information source to assist attribute recognition [23, 11]. For instance, different people may share similar attributes in the same scene, e.g. most skiers wear sunglasses. However, these two schemes are mostly studied independently by existing methods.

In this work, we explore *both* modelling inter-person image context and learning intra-person attribute correlation in a unified framework. To this end, we formulate a Joint Recurrent Learning (**JRL**) of attribute correlation and context. We introduce a novel Recurrent Neural Network (RNN) encoder-decoder architecture specifically designed for *sequential* pedestrian attribute prediction jointly guided by both *intra-person attribute* and *inter-person similarity* context awareness. This RNN based model explores explicitly a *sequential* prediction constraint that differs significantly from the existing CNN based *concurrent* prediction policy [23, 11, 21]. We argue that this approach enables us to exploit more latent and richer higher-order dependency among attributes, therefore better mitigating the small training data challenge. Our approach is motivated by natural language sentence prediction which models inter-word relations [45, 27]. Importantly, two information sources (intra-person attribute correlations and inter-person image similarities) are *simultaneously* modelled to learn person-centric inter-region correlation to compensate poor (or missing) visual details. This provides the model with a more robust embedding given poor surveillance images and learns more accurate intra-person attribute correlations. Crucially, we do not assume people in the same scene share common attributes [23], nor assuming person body-part detection [11]. Because people appearances in surveillance scenes are without a common theme, and person body-part detection in low resolution images under poor lighting is inconsistent, resulting in many poor detections.

More specifically, our approach considers the *sequence-to-sequence mapping* framework (with a paired encoder and decoder) [5, 42, 4]. To explore a sequence prediction model, we convert any given person image into a region sequence (Fig. 2(b)) and a set of attributes into an ordered list (Fig. 2(c)). An *encoder* maps a *fixed-length* image region sequence into a continuous feature vector. The recurrent step is to encode sequentially localised person spatial context and to propagate inter-region contextual information. We call this *intra-person attribute context* modelling. Moreover, we also incorporate *inter-person similarity context* (Fig. 2(a)). That is, we identify visually similar exemplar images in the training set, encode them so to be combined with the encoded image by similarity max-pooling. This fused encoding feature representation is used to initialise a decoder. The *decoder* transforms the image feature vector from the encoder to a *variable-length* attribute sequence as output. This joint sequence-to-sequence encoding and decoding process enables a low- to high-order attribute correlation learning in a unified model. As attributes are weakly-labelled at the image level without fine-grained localisation, we further exploit a data-driven attention mechanism [1] to identify attribute sensitive image regions and to guide the decoder to those image regions for feature extraction.

The **contributions** of this work are: **(1)** We propose a Joint Recurrent Learning (JRL) approach to pedestrian attribute correlation and context learning in a unified model. This is in contrast to existing methods that separate the two learning problems thus suboptimal [23, 11, 21, 41]. **(2)** We formulate a novel end-to-end encoder-decoder architecture capable of jointly learning image level context and attribute level sequential correlation. To our best knowledge, this is the first attempt of formulating *pedestrian attribute recognition as a sequential prediction problem* designed to cope with poor imagery data with missing details. **(3)** We provide extensive comparative evaluations on the two largest pedestrian attribute benchmarks (PETA [7] and RAP [22]) against 7 contemporary models including 5 pedestrian attribute models (SVM [19], MRFr2 [8], ACN [41], DeepSAR and DeepMAR [21]), a multi-label image classification model (Semantically Regularised CNN-RNN [27]), and a multi-person image annotation model (Contextual CNN-RNN [24]). The proposed JRL model yields superior performance compared to these methods.

2. Related Work

Pedestrian Attribute Recognition. Semantic pedestrian attributes have been extensively exploited for person identification [15] and re-identification [20, 40, 32]. Earlier methods typically model multiple attributes independently and train a separate classifier for each attribute (e.g. SVM or AdaBoost) based on hand-crafted features such as colour and texture histograms [20, 51, 19, 7]. Inter-attribute correlation was considered as complementary information to compensate noisy visual appearance for improving prediction performance, e.g. graph model based methods allow to capture attribute co-occurrence likelihoods by using conditional random field or Markov random field to estimate the final joint label probability [8, 3, 38]. However, these methods are prohibitively expensive to compute when dealing with a large set of attributes, due to the huge number of model parameters on pairwise relations. Deep CNN models have been exploited for joint multi-attribute feature and classifier learning [52, 21, 41, 50, 23, 9], and shown to benefit from learning attribute co-occurrence dependency. However, they do not explore modelling high-order attribute sequential correlations. Other schemes also exploited contextual information [11, 23], but making too strong assumptions about image qualities to be applicable to surveillance data. Attribute correlation and contexting are often treated independently by existing methods. This work aims to explore jointly their complementary benefits in improving attribute recognition when only small sized and poor quality training data is available.

Multi-Label Image Classification. Pedestrian attribute recognition is a Multi-Label Image Classification (MLIC)

problem [28, 13]. Sequential multi-label prediction has been explored before [46, 27]. These methods are based on a CNN-RNN model design, whilst our JRL model has a CNN-RNN-RNN architecture. Crucially, these existing MLIC models assume (1) the availability of large scale labelled training data (2) with sufficiently good image quality, e.g. $165,482$ carefully selected Flickr photo images in the MS COCO dataset [25]. Both assumptions are invalid for pedestrian attribute recognition in surveillance images. A very recent multi-person image annotation method advances this sequential MLIC paradigm by incorporating additional inter-person social relations and scene context [24]. This method exploits specifically context among family members and friend-centric photo images of high-resolution, but not scalable to open-world surveillance scenes of poor image data. Moreover, strong attribute-level labels are required [24], whilst pedestrian attributes are mostly weakly-labelled at the image level. In contrast, the proposed JLR model is designed specially to combat the challenges of attribute recognition in low-resolution poor quality images with weakly-labelled data in small training data size. Learning image region sequence correlation has been exploited for face recognition [37] and person re-identification [43]. Their problem settings are different from this work: here we aim to exploit image level context for enhancing sequential attribute correlation learning in a multi-label classification setting, whist both [37] and [43] consider a single-label image classification problem.

3. Joint Recurrent Learning of Attributes

To establish a deep model for recognising pedestrian attributes in inherently ambiguous surveillance images, we assume n labelled training images $\mathcal{I}_{\text{tr}} = \{\boldsymbol{I}_i\}_{i=1}^n$ available, with the attribute labels as $\mathcal{A}_{\text{tr}} = \{\boldsymbol{a}_i\}_{i=1}^n$. Each image-level label annotation $\boldsymbol{a}_i = [a_{(i,1)}, \ldots, a_{(i,n_{\text{attr}})}]$ is a binary vector, defined over n_{attr} pre-defined attributes with 0 and 1 indicating the absence and presence of the corresponding attribute, respectively. Intrinsically, this is a *multi-label* recognition problem since the n_{attr} pedestrian attribute categories may co-exist in a single image. It is necessary to learn *attribute sequential correlations (high-order) in similar imagery context* in order to overcome the limitations in training data due to poor image quality, weak labelling, and small training size. To that end, we formulate a Joint Recurrent Learning (JRL) of both attribute context and correlation in an end-to-end sequential deep model.

3.1. Network Architecture Design

An overview of the proposed JRL architecture is depicted in Fig. 2. We consider the RNN encoder-decoder framework as our base model because of its powerful capability in learning sequence data and modelling the translation between different types of sequences [42, 6, 5]. Specifically, RNN is a neural network consisting an internal hidden state $\boldsymbol{h} \in \mathbb{R}^d$ and operating on a variable-length input sequence $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots)$. At each time step t, the RNN takes sequentially an element $\boldsymbol{x}_t$ of X and then updates its hidden state $\boldsymbol{h}_t$ as

$$\boldsymbol{h}_t = \phi_\theta(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) \tag{1}$$

where ϕ_θ denotes the non-linear activation function parameterised by θ. To capture the long range dependency of attribute-attribute, region-region, and attribute-region, we adopt the LSTM [14] as recurrent neuron for both encoder and decoder RNN. Also, LSTM is effective to handle the common gradient vanishing and exploding problems in training RNN [31]. Particularly, at each time step t, the LSTM updates using the input $\boldsymbol{x}_t$ and the LSTM previous status $\boldsymbol{h}_{t-1} \in \mathbb{R}^d$, and $\boldsymbol{c}_{t-1} \in \mathbb{R}^d$ as:

$$\begin{aligned}
\boldsymbol{f}_t &= \text{sigmoid}(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{i}_t &= \text{sigmoid}(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{o}_t &= \text{sigmoid}(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{g}_t &= \tanh(\boldsymbol{W}_{gx}\boldsymbol{x}_t + \boldsymbol{W}_{gh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_g) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{2}$$

where $\text{sigmoid}(\cdot)$ refers to the logistic sigmoid function, $\tanh(\cdot)$ the hyperbolic tangent function, the operator $\odot$ the element-wise vector product. The LSTM contains four multiplicative gating units: forget gate $\boldsymbol{f} \in \mathbb{R}^d$, input gate $\boldsymbol{i} \in \mathbb{R}^d$, output gate $\boldsymbol{o} \in \mathbb{R}^d$, input modulation gate $\boldsymbol{g} \in \mathbb{R}^d$, with matrix $\boldsymbol{W}$s and vector $\boldsymbol{b}$s the corresponding to-be-learned parameters. The memory cell $\boldsymbol{c}_t$ depends on (1) the previous memory cell $\boldsymbol{c}_{t-1}$, modulated by $\boldsymbol{f}_t$, and (2) the input modulation gate, modulated by $\boldsymbol{i}_t$. As such, the LSTM learns to forget its previous memory and exploit its current input selectively. Similarly, the output gate $\boldsymbol{o}$ learns how to transfer the memory cell $\boldsymbol{c}_t$ to the hidden state $\boldsymbol{h}_t$. Collectively, these gates learn to effectively modulate the behaviour of input signal propagation through the recurrent hidden states for helping capture complex and long-term dynamics/dependency in sequence data.

(I) Intra-Person Attribute Context. We model the intra-person attribute context within each person image $\boldsymbol{I}$ by the encoder LSTM. This is achieved by mapping recurrently each input image into a fixed-length feature vector (Fig. 2 (b)). Specifically, for allowing sequential modelling of the person image $\boldsymbol{I}$, we first divide it into m (empirically set $m = 6$ in our experiment) horizontal strip regions and form a region sequence $S = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m)$ in top-bottom order. Then the encoder reads each image region sequentially, and the hidden state $\boldsymbol{h}^{\text{en}}$ of the encoder LSTM updates according to Eqn. (2). Once reaching the end of this region sequence, the hidden state $\boldsymbol{h}_m^{\text{en}}$ of the encoder can be considered as the summary representation $\boldsymbol{z} = \boldsymbol{h}_m^{\text{en}}$ of the entire

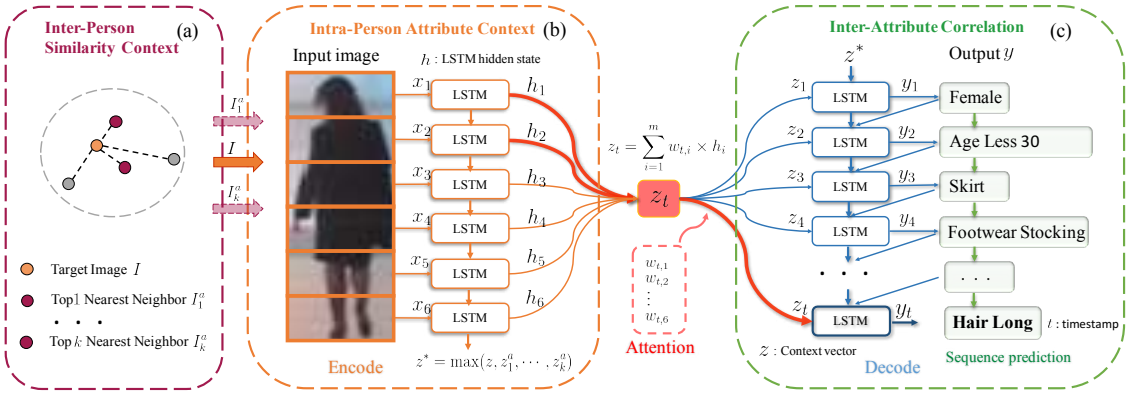

Figure 2. An overview of the proposed Joint Recurrent Learning (JRL) of attribute context and correlation.

sequence or the person image. We call $\boldsymbol{z}$ as *context vector*. Importantly, this allows to selectively extract and encode the spatial dependency between different body parts whilst also propagate relevant localised topological contextual information through the recurrent network, thanks to the capability of LSTM in modelling the long short term relationships between sequence elements.

(II) Inter-Person Similarity Context. To compensate appearance ambiguity and poor image quality in a target image, we explore auxiliary information from visually similar exemplar training images to provide a inter-person similarity context constraint. Specifically, we search top-k exemplars $\{\boldsymbol{I}_i^a\}_{i=1}^k$ that are visually similar to the image $\boldsymbol{I}$ from the training pool. For each exemplar $\boldsymbol{I}_i^a$, we compute its own context vector $\boldsymbol{z}_i^a$ using the same encoding process as that for the image $\boldsymbol{I}$. Then, we ensemble all the context vector representations of auxiliary images as the inter-person context to $\boldsymbol{z}$ as follows:

$$\boldsymbol{z}^* = \max(\boldsymbol{z}, \boldsymbol{z}_1^a, \cdots, \boldsymbol{z}_k^a) \tag{3}$$

where $\max(\cdot)$ defines the element-wise maximum operation over all input feature vectors of both the input image and top-k exemplars. While the averaging based ensemble may be more conservative and reducing the likelihood of introducing additional noisy information, we found empirically that maximum-pooling based ensemble is more effective. The rational for this inter-person similarity context compensation is that missing or corrupted local information in the input image cannot be easily recovered in the decoding process whilst newly introduced localised noise can be largely suppressed by optimising the decoder.

Image Representation and Similarity Search. As input to the LSTM encoder, we utilise a deep CNN initialised by ImageNet (e.g. the AlexNet [17]), then fine-tune the CNN on the pedestrian attribute training data to better represent pedestrian images by its deep feature vectors. Specifically, for a given person image, we decompose the activations of the 5$^{\text{th}}$ convolutional layer into m horizontal regions, each of which is pooled into a vector by directly concatenating all dimensions. Moreover, we use the FC$_7$ layer's output as the feature space for top-k exemplar similarity search using L2 distance metric.

(III) Inter-Attribute Correlation. We construct a decoder LSTM to model the latent high-order attribute correlation subject to jointly learning a multi-attribute prediction classifier. Specifically, given the encoded context vectors $\boldsymbol{z}$ and $\boldsymbol{z}^*$, the decoder LSTM aims to model a sequential recurrent attribute correlation within both *intra-person attribute context* ($\boldsymbol{z}$) and *inter-person similarity context* ($\boldsymbol{z}^*$) and to generate its variable-length output as a predicted sequence of attributes $\boldsymbol{y}_t$ over time t (Fig. 2(c)). This is desired since the co-occurring attribute number varies among individual images. An attribute label sequence of a person image is generated from a fixed order list of all attributes (Sec. 3.2). We initialise the decoder hidden state $\boldsymbol{h}_1^{\text{de}}$ with the improved encoder context vector: $\boldsymbol{h}_1^{\text{de}} = \boldsymbol{z}^*$. This is to incorporate the *inter-person similarity context* into the decoding process. Compared to the encoder counterpart, $\boldsymbol{h}_t^{\text{de}}$ and $\boldsymbol{y}_t$ are additionally conditioned on the previous output $\boldsymbol{y}_{t-1}$ (initialised $\boldsymbol{y}_0 = \boldsymbol{0}$, i.e. the "start" token). In essence, it is this sequential recurrent feedback connection that enables our model to mine the varying high-orders of attribute-attribute dependency – longer prediction, higher-order attribute correlation modelled. Formally, rather than by Eqn. (1), $\boldsymbol{h}_t^{\text{de}}$ is updated via:

$$\boldsymbol{h}_t^{\text{de}} = \phi_\theta(\boldsymbol{h}_{t-1}^{\text{de}}, \boldsymbol{y}_{t-1}, \boldsymbol{z}). \tag{4}$$

In case of LSTM, the particular update formulation is:

$$\begin{aligned}
\boldsymbol{f}_t &= \text{sigmoid}(\boldsymbol{W}_{fz}\boldsymbol{z} + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1}^{\text{de}} + \boldsymbol{W}_{fy}\boldsymbol{y}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{i}_t &= \text{sigmoid}(\boldsymbol{W}_{iz}\boldsymbol{z} + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1}^{\text{de}} + \boldsymbol{W}_{iy}\boldsymbol{y}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{o}_t &= \text{sigmoid}(\boldsymbol{W}_{oz}\boldsymbol{z} + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1}^{\text{de}} + \boldsymbol{W}_{oy}\boldsymbol{y}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{g}_t &= \tanh(\boldsymbol{W}_{gz}\boldsymbol{z} + \boldsymbol{W}_{gh}\boldsymbol{h}_{t-1}^{\text{de}} + \boldsymbol{W}_{gy}\boldsymbol{y}_{t-1} + \boldsymbol{b}_g) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{5}$$

This is similar to Eqn. (2) except the extra dependence on the previous prediction $\boldsymbol{y}_{t-1}$ and corresponding param-

eters[1]. To predict the attribute, we first compute the conditional probability over all attributes and a "stop" signal as:

$$\begin{aligned} \boldsymbol{p}(\{y_{t,i}=1\}_{i=1}^{n_{\text{attr}}+1}) &= \phi_y(\boldsymbol{h}_{t-1}^{\text{de}}, \boldsymbol{y}_{t-1}, \boldsymbol{z}) \\ &= \boldsymbol{W}_y \boldsymbol{o}_t + \boldsymbol{b}_y \end{aligned} \tag{6}$$

where $\boldsymbol{W}_y \in \mathbb{R}^{(n_{\text{attr}}+1)\times d}$ and $\boldsymbol{b}_y \in \mathbb{R}^{(n_{\text{attr}}+1)}$ are model parameters, and $\boldsymbol{o}_t \in \mathbb{R}^d$ can be obtained by Eqn. (5). Then, we predict the current attribute $\boldsymbol{y}_t$ as:

$$i^* = \text{argmax}_{i\in[1,\ldots,n_{\text{attr}}+1]}(y_{t,i}), \tag{7}$$

i.e. the i^*-th bit of $\boldsymbol{y}_t$ is 1 whilst all the others are 0.

Recurrent Attribute Attention. Appearance attribute patterns in real-world person images can vary complexly and significantly. By summarising a person image into a single fixed-length context vector $\boldsymbol{z}$ with the encoder, a large amount of semantic information, (e.g. spatial distribution) may be not well encoded due to the limited representation capacity of context vector [5]. To overcome this limitation, we propose to improve our JRL model by incorporating the attention mechanism [1, 4] so as to automatically identify and focus on the most relevant parts of the input region sequence when predicting the current attribute to improve the correlation modelling and finally the prediction performance. This is essentially an explicit sequence-to-sequence alignment mechanism. We achieve this by imposing a structure into the encoder output and then reformulating the attribute decoding algorithm accordingly.

Specifically, we first allow the encoder to output a structured representation, a set of summary vectors as:

$$H^{\text{en}} = (\boldsymbol{h}_1^{\text{en}}, \ldots, \boldsymbol{h}_i^{\text{en}}, ..., \boldsymbol{h}_m^{\text{en}}), \quad \boldsymbol{h}_i^{\text{en}} \in \mathbb{R}^d \tag{8}$$

for an input image region sequence $S = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m)$ with m the number of all time steps or the input region sequence length. Clearly, $\boldsymbol{h}_i^{\text{en}}$ represents the context representation of the i-th (top-down order) spatial region of the input image. The aim of our attribute attention is to identify an optimised weighting distribution $\boldsymbol{w}_t = (w_{t,1}, \cdots, w_{t,i}, \cdots, w_{t,m})$ over $(\boldsymbol{h}_1^{\text{en}}, \cdots, \boldsymbol{h}_i^{\text{en}}, \cdots, \boldsymbol{h}_m^{\text{en}})$ at each time step t when the decoder is predicting the attribute, as:

$$w_{t,i} = \frac{\exp(\alpha_{t,i})}{\sum_{i=1}^{m}\exp(\alpha_{t,i})}, \text{with } \alpha_{t,i} = \phi_{\text{att}}(\boldsymbol{h}_{t-1}^{\text{de}}, \boldsymbol{h}_i^{\text{en}}) \tag{9}$$

where ϕ_{att} defines the attention function realised with a feed forward neural network in our approach as in [44]. Once obtaining the attention weighting score $\boldsymbol{w}_t$, we utilise it to compute the step-wise context representation $\boldsymbol{z}_t$ by

$$\boldsymbol{z}_t = \sum_{i=1}^{m} w_{t,i} \times \boldsymbol{h}_i^{\text{en}} \tag{10}$$

[1]We do not use the encoded $\boldsymbol{z}^*$ with inter-person similarity context as the decoder input to avoid possible divergence from the exemplar images.

The final prediction $\boldsymbol{y}_t$ can be obtained similarly by computing $\boldsymbol{h}_t^{\text{de}}$ with Eqn. (4), and applying Eqns. (6) and (7).

Note that the context representation $\boldsymbol{z}_t$ utilised in attention-aware attribute decoding is varying over time t due to the difference in the spatial attention distribution $\boldsymbol{w}_t$. In contrast, in case of no attention, $\boldsymbol{z}$ is constant during the whole decoding process. Implicitly, the current $\boldsymbol{w}_t$ is conditioned on $\boldsymbol{w}_{t-1}$ (the attention used at time $t-1$) due to the dependence of $\boldsymbol{h}_{t-1}$ on $\boldsymbol{z}_{t-1}$ (Eqn. (4)) and of $\boldsymbol{z}_{t-1}$ on $\boldsymbol{w}_{t-1}$ (Eqn. (10)), therefore allowing to optimise the underlying correlation in per-step attention selection for sequential attribute decoding during model training.

Attribute Embedding. To incorporate the previous attribute prediction as recurrent feedback on the next prediction, we need a proper attribute representation. One common way is the simple one-hot vector. Alternatively, word embedding [30] has been shown as a favourable text representation by learning a lookup table optimised for specific annotation error and thus more semantically meaningful. Therefore, we adopt the latter scheme in our attribute decoder.

3.2. Model Training and Inference

To train the JRL model, we need to determine attribute selection order. However, pedestrian attributes are naturally unordered without a fixed ordering, similar to generic multi-label image classification [27, 46] and dissimilar to image caption [45]. To address this problem, one can either randomly fix an order [24] or define some occurrence frequency based orders, e.g. rare first (rarer attributes are placed earlier so that they can be promoted) [27], or frequent first (more frequent attributes appear earlier with the intuition that easier ones can be predicted before harder ones) [46]. In our model training, we explore the ensemble idea [34] to incorporate the complementary benefits of all these different orders and thus capture more high-order correlation between attributes in context. We consider that using an *order ensemble* is critical for pedestrian attribute modelling because: (1) Small sized training data makes poor model learning for most attribute classes; (2) Given significant attribute appearance change in surveillance data, the optimal sequential attribute correlation can vary significantly between different pedestrian images, with no single universally optimal sequential order. Thus, we employ an ensemble of 10 attribute orders: rare first, frequent first, top-down and bottom-up (for encoding body topological structure information), global-local and local-global (for interacting coarse and fine grained attributes), and 4 random orders (for incorporating randomness).

Model Training. For each attribute order in the ensemble, we train an order-specific JRL model. We learn any JRL model end-to-end by back-propagation through time [49] so as to jointly optimise the encoder and decoder LSTM. We use the cross-entropy loss on the softmax score subject

Figure 3. Example images. **Left**: PETA [7]; **Right**: RAP [22].

to training attribute labels. To avoid noise back propagation from the RNN to CNN, we do not train the CNN image feature representation network together with the JRL RNN encoder-decoder. Each JRL model is optimised against per-image attribute sequences without duplication. Therefore, repeated prediction is inherently penalised and discouraged.

Model Inference. Given a test image, each trained JRL model gives a multi-attribute prediction. We generate a set of 10 predictions per test image given 10 order-specific models. To infer the final prediction, we adopt the majority voting scheme [29].

4. Experiments

Datasets. For evaluations, we used the two largest pedestrian attribute datasets publically available (Fig.3): **(1)** The PEdesTrian Attribute (*PETA*) dataset [7] consists of 19,000 person images collected from 10 small-scale person datasets. Each image is labelled with 65 attributes (61 binary + 4 multi-valued). Following the same protocol per [8, 21], we randomly divided the whole dataset into three non-overlapping partitions: 9500 for model training, 1900 for verification, and 7600 for model evaluation. **(2)** The Richly Annotated Pedestrian (*RAP*) attribute dataset [22] has in total 41,585 images drawn from 26 indoor surveillance cameras. Each image is labelled with 72 attributes (69 binary + 3 multiple valued). We adopted the same data split as in [22]: 33,268 images for training and the remaining 8,317 for test. We evaluated the same 51 binary attributes per [22] for a fair comparison. For both datasets, we converted multi-valued attributes into binary attributes as in [8, 21, 22]. Both datasets pose significant challenges to pedestrian attribute recognition under difficult illumination with low resolution, occlusion and background clutter.

Performance Metrics. We use four metrics to evaluate attribute recognition performance. **(1)** Class-centric: For each attribute class, we compute the classification accuracy of positive and negative samples respectively, average them to obtain an Average Precision (AP) for this attribute, then take the mean of AP over all attributes (mAP^{cls}) as the metric [8]. **(2)** Instance-centric: We measure per instance (image) attribute prediction precision and recall. This measure additionally considers the inter-attribute correlation, in contrast to mAP that assumes independence between attributes [22]. Specifically, we compute the precision and recall of predicted attributes against the groundtruth for each test image, and then take the mean of the two measures over all test images to yield mean Precision (mPrc^{ins}) and mean Recall (mRcl^{ins}) rates. We also compute the F1 score (F1^{ins}) based on mPrc^{ins} and mRcl^{ins} as a more comprehensive metric.

Competitors. We compared our model JRL against 7 contemporary and state-of-the-art models. They include (I) two conventional discriminative attribute methods: **(1)** We adopted CNN features (FC_7 output of the AlexNet) with the SVM attribute model [19], replacing its original Ensemble of Localized Features (ELF) [12, 19]; **(2)** MRFr2 [8] is a graph based attribute recognition method that exploits the context of neighbouring images by Markov random field for mining the visual appearance proximity relations between different images to support attribute reasoning; (II) Three deep learning attribute recognition methods: **(3)** Attributes Convolutional Network (ACN) [41] trains jointly a CNN model for all attributes, and sharing weights and transfer knowledge among different attributes; **(4)** DeepSAR [21] is a deep model that that treats attribute classes individually by training multiple attribute-specific AlexNet models [17]; **(5)** DeepMAR [21] , unlike DeepSAR, considers additionally inter-attribute correlation by jointly learning all attributes in a single AlexNet model [17], so to capture the concurrent attribute relationships, similar to [41]; (III) One multi-person image annotation recurrent model: **(6)** Contextual CNN-RNN (CTX CNN-RNN) [24] is a CNN-RNN based sequential prediction model designed to encode the scene context and inter-person social relations for modelling multiple people in an image[2]; (IV) One generic multi-label image classification model: **(7)** Semantically Regularised CNN-RNN (SR CNN-RNN) [27] is a state-of-the-art multi-label image classification model that exploits the groundtruth attribute labels for strongly supervised deep learning and richer image embedding.

Implementation Details. The hidden state for both the encoder LSTM and the decoder LSTM of the JRL model has 512 units (neurons). We set empirically the learning rate as 0.0001 with AdamOptimizer [16], and the dropout rate as 0.5. By default, we adopted the AlexNet (same as DeepMAR) as the network architecture for image embedding, and top-2 exemplars were selected for inter-person context.

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

Tables 1 and 2 show evaluations on PETA and RAP respectively. It is evident that the proposed JRL model

[2]In our weakly supervised setting for attribute recognition, we have no attribute fine-grained location labelling. So we feed the whole image CNN features at each recurrent decoding step for both training and test.

Table 1. Evaluation on PETA [7], 1st/2nd best results in red/blue.

Method \ Metric	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
MRFr2[8]	75.60	-	-	-
ELF+SVM [19]	75.21	49.45	74.24	59.36
CNN+SVM [22]	76.65	51.33	75.14	61.00
ACN [41]	81.15	84.06	81.26	82.64
DeepSAR [21]	81.30	-	-	-
DeepMAR [21]	82.60	83.68	83.14	83.41
CTX CNN-RNN [24]	80.13	79.68	80.24	79.68
SR CNN-RNN [27]	82.83	82.54	82.76	82.65
JRL	85.67	86.03	85.34	85.42

Table 2. Evaluation on RAP [22], 1st/2nd best results in red/blue.

Method \ Metric	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
MRFr2[8]	-	-	-	-
ELF+SVM [19]	69.94	32.84	71.18	44.95
CNN+SVM [22]	72.28	35.75	71.78	47.73
ACN [41]	69.66	80.12	72.26	75.98
DeepSAR [21]	-	-	-	-
DeepMAR [21]	73.79	74.92	76.21	75.56
CTX CNN-RNN [24]	70.13	71.03	71.20	70.23
SR CNN-RNN [27]	74.21	75.11	76.52	75.83
JRL	77.81	78.11	78.98	78.58

achieves the best accuracy on PETA given by all four evaluation metrics, and on RAP the best accuracy given by three metrics except mPrc^{ins} coming second best (JRL 78.11% vs. ACN 80.12%). This implies that ACN is conservative in prediction, i.e. predicting only very confident attributes. More significantly, JRL outperforms in mAP^{cls} the state-of-the-art attribute model DeepMAR [21] and multi-label image annotation model SR CNN-RNN [27] respectively by 3.07% and 2.84% on PETA, 4.02% and 3.60% on RAP. Similar margins are observed with other performance metrics, except with the mPrc^{ins} ACN [41] achieves the best score (80.12% vs. 78.11% by JRL) on RAP, but with a much lower mRcl^{ins} (72.26% vs. 78.98% by JRL) also a lower overall F1^{ins} (75.98% vs. 78.58% by JRL). This shows clearly the benefit of the proposed correlation and context joint recurrent learning approach to predicting ambiguous pedestrian attributes in poor quality surveillance images. This is mainly due to JRL's capacity to maximise and exploit the complementary effect of correlation and context on sparsely labelled weak annotations, through an end-to-end encoder/decoder learning.

Robustness Against Training Label Sparsity. In addition to the overall performance comparisons, we further conducted a model scalability evaluation against the training data size to better understand model robustness to small sized (difficult-to-collect) attribute labels. To that end, we randomly removed varying ratios (25∼75%) of the whole training data set with the test data remaining unchanged on both PETA and RAP. We compared the JRL model with the best two competitors: DeepMAR [21] and SR CNN-RNN [27]. It is evident from Table 3 that JRL is more robust than

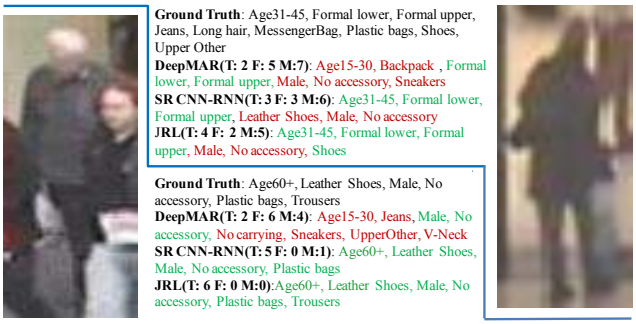

Figure 4. Qualitative evaluation of pedestrian attribute recognition on PETA [7], with wrong predictions in red, true in green.

both DeepMAR and SR CNN-RNN against training data sparsity. When training data decreased from 100% to 25%, the mAP^{cls} performance drop of JRL is 3.64% and 3.55% on PETA and RAP respectively. This compares favourably against DeepMAR (6.23% and 5.73%) and SR CNN-RNN (6.24% and 5.85%). This validate the advantages and potentials of our proposed JRL model in handling sparse training data situations by effectively maximising the joint benefits of modelling attribute context and correlation end-to-end from only limited labelled data.

Table 3. Model robustness vs. training data size (TDS) in %.

Dataset	TDS (%)	Model \ Metric	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
PETA [8]	100	DeepMAR [21]	82.60	83.68	83.14	83.41
		SR CNN-RNN[27]	82.83	82.54	82.76	82.65
		JRL	**85.67**	**86.03**	**85.34**	**85.42**
	75	DeepMAR [21]	80.83	81.02	81.73	81.37
		SR CNN-RNN[27]	81.06	81.11	81.66	81.21
		JRL	**84.45**	**84.86**	**84.23**	**84.07**
	50	DeepMAR [21]	79.16	80.66	80.39	80.52
		SR CNN-RNN[27]	79.09	80.40	80.13	80.06
		JRL	**83.42**	**84.16**	**82.39**	**82.46**
	25	DeepMAR [21]	76.37	79.12	77.93	78.52
		SR CNN-RNN[27]	76.59	79.23	78.12	78.39
		JRL	**82.03**	**83.16**	**81.01**	**81.51**
RAP [22]	100	DeepMAR [21]	73.79	74.92	76.21	75.56
		SR CNN-RNN[27]	74.21	75.11	76.52	75.83
		JRL	**77.81**	**78.11**	**78.98**	**78.58**
	75	DeepMAR [21]	71.38	72.40	74.62	73.49
		SR CNN-RNN[27]	71.96	72.33	74.73	73.64
		JRL	**76.69**	**77.34**	**77.76**	**77.36**
	50	DeepMAR [21]	70.01	71.52	72.53	72.06
		SR CNN-RNN[27]	70.53	71.96	72.77	72.36
		JRL	**75.51**	**76.31**	**76.69**	**76.64**
	25	DeepMAR [21]	68.06	70.33	69.86	70.08
		SR CNN-RNN[27]	68.36	70.67	70.39	70.67
		JRL	**74.26**	**75.16**	**75.21**	**75.34**

4.2. Further Analysis and Discussions

(1) Benefit of intra-person attribute context. We evaluated explicitly the benefit of modelling intra-person attribute context by the encoder LSTM. For that, we built a stripped-down JRL model by removing the encoder and

directly using the CNN FC features for the decoder input. Table 4 shows the difference on both PETA and RAP, improving mAP^{cls} by 2.22% and 2.62% respectively, similarly under the other metrics.

Table 4. Benefit of intra-person Attribute Context (AC).

Dataset	Metric / Method	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
PETA [8]	JRL (No AC)	83.45	83.96	83.97	83.89
	JRL	**85.67**	**86.03**	**85.34**	**85.42**
RAP [22]	JRL (No AC)	75.19	75.55	76.93	75.97
	JRL	**77.81**	**78.11**	**78.98**	**78.58**

(2) Effect of inter-person similarity context. We also evaluated explicitly the benefit of exploiting auxiliary exemplar images as inter-person context. For that, we excluded them in both model training and inference stages. Table 5 shows that this context modelling brings 0.65% and 0.87% boost in mAP^{cls} on PETA and RAP respectively.

Table 5. Effect of inter-person Similarity Context (SC).

Dataset	Metric / Method	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
PETA [8]	JRL(No SC)	85.02	85.27	84.36	84.86
	JRL	**85.67**	**86.03**	**85.34**	**85.42**
RAP [22]	JRL(No SC)	76.94	77.39	78.23	77.92
	JRL	**77.81**	**78.11**	**78.98**	**78.58**

(3) Effects of model ensemble. We evaluated the benefit of exploiting attribute order ensemble. We compared the average results of all 10 orders. Table 6 shows that the ensemble of *distinct sequential orders* improves significantly the model performance, improving mAP^{cls} by 3.54% and 3.07% on PETA and RAP when compared to the average. This validates our ensemble design intuition for modelling ambiguous attributes in poor surveillance images from sparsely labelled training data.

Table 6. Effects of the model ensemble.

Dataset	Metric / Method	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
PETA [8]	*Average*	82.13	82.55	82.12	82.02
	Ensemble	**85.67**	**86.03**	**85.34**	**85.42**
RAP [22]	*Average*	74.74	75.08	74.96	74.62
	Ensemble	**77.81**	**78.11**	**78.98**	**78.58**

(4) Effects of recurrent attribute attention. Table 7 shows that the data-driven alignment between image region sequence and attribute label sequence is beneficial, giving 1.64% and 1.85% in mAP^{cls} boost on PETA and RAP.

Table 7. Effects of recurrent attribute attention.

Dataset	Metric / Method	mAP^{cls}	mPrc^{ins}	mRcl^{ins}	F1^{ins}
PETA [8]	JRL(No Attention)	84.03	84.92	84.19	84.24
	JRL	**85.67**	**86.03**	**85.34**	**85.42**
RAP [22]	JRL(No Attention)	75.96	76.89	77.49	77.13
	JRL	**77.81**	**78.11**	**78.98**	**78.58**

(5) Qualitative analysis on the effect of attribute correlations. We examined more carefully the effect of attribute correlations on the JRL model performance. Fig. 5 shows two examples. The person on the left with a fashionable bag and legging/shoes was predicted reliably by JRL to be "young", whilst the "hair long" is less obvious and "skirt" almost invisible but both predicted correctly by JRL invoking the relevant ordering context. In contrast, a non-sequence prediction model DeepMAR [21] missed both "hair long" and "skirt". This is because that JRL benefited from identifying the relevant sequential "age-hair-skirt" ordering as inter-attribute correlation context for attribute prediction. When getting the wrong ordering context, JRL missed "skirt". The person on the right wears "skirt", clearly visible, with both "hair long" and "age" unclear. The JRL model again benefited from invoking the useful "skirt-hair-age" ordering context for attribute prediction. When given the wrong ordering, JRL makes the mistake on "age" prediction. DeepMAR missed both "skirt" and "hiar long", and predicted "age" incorrectly.

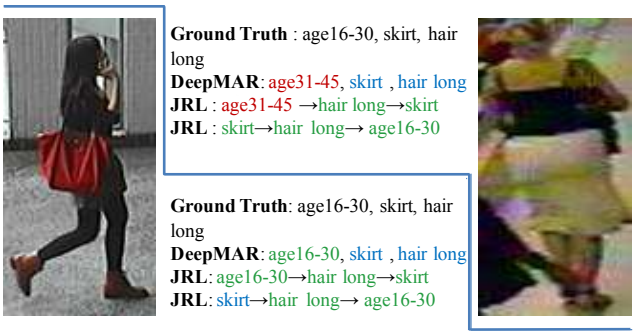

Figure 5. Qualitative analysis of latent attribute correlation, with wrong predictions in red, true in green and missed predictions in blue. The two examples are from PETA [7].

5. Conclusion

In this work, we presented a novel deep Joint Recurrent Learning (JRL) model for exploring attribute context and correlation in deep learning of pedestrian attributes given low quality surveillance images and small sized training data. Our JRL method outperforms a wide range of state-of-the-art pedestrian attribute and multi-label classification methods. Extensive experiments demonstrate the advantages and superiority of joint learning high-order (sequential) inter-attribute correlation on two pedestrian attribute benchmarks. Moreover, the JRL model is shown to be more robust than state-of-the-art deep models when trained with small sized training data, thus more scalable to real-world applications with limited annotation budget available.

Acknowledgements

This work was partially supported by the China Scholarship Council, Vision Semantics Ltd., and Royal Society Newton Advanced Fellowship Programme (NA150459).

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2014. 2, 5

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1

[3] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 1, 2

[4] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015. 2, 5

[5] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv e-prints*, 2014. 2, 3, 5

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. 3

[7] Y. DENG, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACMMM*, 2014. 1, 2, 6, 7, 8

[8] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Learning to recognize pedestrian attribute. *arXiv e-prints*, 2015. 2, 6, 7, 8

[9] Q. Dong, S. Gong, and X. Zhu. Class rectification hard mining for imbalanced deep learning. In *ICCV*, 2017. 1, 2

[10] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *IEEE Winter Conference on Applications of Computer Vision*, 2017. 1

[11] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *ICCV*, 2015. 1, 2

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

[13] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009. 3

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[15] E. S. Jaha and M. S. Nixon. Soft biometrics for subject identification using clothing attributes. In *IEEE International Joint Conference on Biometrics*, 2014. 1, 2

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 4, 6

[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1

[19] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014. 2, 6, 7

[20] R. Layne, T. M. Hospedales, S. Gong, et al. Person re-identification by attributes. In *BMVC*, 2012. 1, 2

[21] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, 2015. 1, 2, 6, 7, 8

[22] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv e-prints*, 2016. 1, 2, 6, 7, 8

[23] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 2016. 1, 2

[24] Y. Li, G. Lin, B. Zhuang, L. Liu, C. Shen, and A. v. d. Hengel. Sequential person recognition in photo albums with a recurrent network. In *CVPR*, 2017. 2, 3, 5, 6, 7

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 3

[26] C. Liu, S. Gong, C. Loy, and X. Lin. Person re-identification: What features are important? In *First International Workshop on Re-Identification, European Conference on Computer Vision*, 2012. 1

[27] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. Semantic regularisation for recurrent image annotation. In *CVPR*, 2017. 2, 3, 5, 6, 7

[28] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008. 3

[29] K. O. May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica: Journal of the Econometric Society*, pages 680–684, 1952. 6

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv e-prints*, 2013. 5

[31] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013. 3

[32] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang. Joint learning of semantic and latent attributes. In *ECCV*, 2016. 1, 2

[33] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[34] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, 2011. 5

[35] D. A. Reid, M. S. Nixon, and S. V. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1216–1228, 2014. 1

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[37] F. Samaria and F. Fallside. *Face identification and feature extraction using hidden markov models*. Citeseer, 1993. 3

[38] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015. 2

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

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

[41] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015. 2, 6, 7

[42] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. 2, 3

[43] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 3

[44] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. Grammar as a foreign language. In *NIPS*, 2015. 5

[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*. 2, 5

[46] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2016. 3, 5

[47] J. Wang, X. Zhu, and S. Gong. Video semantic clustering with sparse and incomplete tags. In *AAAI Conference on Artificial Intelligence*, 2016. 1

[48] J. Wang, X. Zhu, and S. Gong. Discovering visual concept structure with sparse and incomplete tags. *Artificial Intelligence*, 250:16–36, 2017. 1

[49] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990. 5

[50] J. Zhu, S. Liao, Z. Lei, and S. Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 2016. 1, 2

[51] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *Workshop of IEEE International Conference on Computer Vision*, 2013. 2

[52] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics*, 2015. 2