Person Re-Identification Combining Deep Features and Attribute Detection

Gregory Watson and Abhir Bhalerao

Department of Computer Science, University of Warwick, CV4 7AL
{g.a.watson, abhir.bhalerao}@warwick.ac.uk

Abstract

Attributes-Based Re-Identification is a way of identifying individuals when presented with multiple pictures taken under varying conditions. The method typically builds a classifier to detect the presence of certain appearance characteristics in an image, and creates feature descriptors based on the output of the classifier. We improve attribute detection through spatial segregation of a person's limbs using a skeleton prediction method. After a skeleton has been predicted, it is used to crop the image into three parts - top, middle and bottom. We then pass these images to an attribute prediction network to generate robust feature descriptors. We evaluate the performance of our method on the VIPeR, PRID2011 and i-LIDS data sets, comparing our results against the state-of-the-art to demonstrate competitive overall matching performance.

1 Introduction

Person Re-Identification (Re-ID) is the process of matching different images of people, taken from separate, non-overlapping cameras. Its applications include, but are not limited to, surveillance, tracking and security. Automated methods for Re-ID reduce the need for manual search through large amounts of data looking for a particular person, however, these methods [1–3] typically struggle with variations in illumination, pose, background and occlusion, as shown in Figure 1.

Traditional approaches to Re-ID typically exploit low-level features such as colour or texture histograms, due to how easy they are to obtain and compare. However, such features can be heavily influenced by variations in visual characteristics, such as background, illumination and pose. A series of unaligned person images may cause corresponding feature regions to not represent the same part of the person, causing problems during matching. Furthermore, the low resolution of most Re-ID cameras render biometrics such as facial recognition infeasible. Whilst automated Re-ID methods traditionally use low-level appearance features to describe an image, such as colour and texture, humans instead rely on attribute descriptions such as *short hair*, *long sleeves* and *jeans* to describe a person. Compared to low-level features such as colour and texture, the attribute's appearance features are significantly more invariant to

illumination and pose variation. For example, a *red shirt* will still be considered a *red shirt* even when significant visual variation is present.

Similar to the work by Li, Chen et al. [4] and Yi, Lei and Li [5], we create an attribute detection network which takes as input a Re-ID image and the three sub-images which represent certain regions of a persons body. We use four ResNet-50-based [6] deep CNNs, and concatenate their outputs and then use this as input to a fully-connected layer with n nodes, where n is the number of attributes being predicted. The contributions of this paper are as follows: The first is an attribute detection network which learns a mapping between the four input images and an attribute vector. We train our attribute detection network and evaluate on a separate set of data sets, proving our networks ability to generalise. An additional contribution is the combination of our deep attribute features with state-of-the-art hand-crafted appearance features [3] to improve matching rates further. We evaluate our method on the VIPeR [7], PRID2011 [8] and i-LIDS [9, 10] data sets and show competitive performance against other methods.

2 Related Work

Pose variation can lead to corresponding regions in different Re-ID images representing different parts of a person's body. As such, feature extraction between two corresponding image regions can be significantly different even if both images represent the same person. To overcome this, several techniques have been proposed [1–3]: Farenzena, Bazzani et al. [2] created Symmetry-Driven Accumulation of Local Features (SDALF), which divides each person image into three parts - the head, torso and legs, and finds a vertical axis of symmetry that best divides the appearance on each side of it. This allows the authors to isolate the foreground from the background and extract features more representative of the person. Liao, Hu et al. [3] developed Local Maximal Occurrence (LOMO), where each image is divided into patches, extracting a HSV and SILTP [11] histogram from each. In order to prevent the negative effects of pose variation, for each row of patches, the final feature descriptor is built by choosing the highest value in each histogram bin. In [1] this is extended further by building a model to predict the skeleton of a person based on their image. Once a skeleton is predicted, it is then used to create a binary mask, weighting each patch by the percentage of the patch that is considered foreground. This ensures features are more representative of the person rather than the background, and compared to the original unweighted features, matching rates are shown to improve.

Whilst automated Re-ID methods traditionally use appearance features to describe an image, humans would instead rely on attribute descriptions such as *short hair*, *long sleeves* and *jeans* to describe a person. Compared to low-level features such as colour and texture, the attribute's appearance features are significantly more invariant to illumination and pose variation. Some of the earliest work using attributes for Re-ID was carried out by Layne, Hospedales et al. [16], where the authors start by defining a set of fifteen attributes - *shorts*, *skirt*, *sandals*, *backpack*, *jeans*, *logo*, *v-neck*, *open-outerwear*, *stripes*, *sunglasses*, *headphones*, *long-hair*, *short-hair*, *gender* and *carryingobject*. As some of these attributes will only be present on certain areas of the person's body, such as jeans only occurring on the lower-half, the authors divide the person image into six equal sized stripes. From each stripe, a 464-

Figure 1: Examples of various images from the 3DPeS [12, 13], VIPeR [7] and QMUL GRID [14, 15] data sets. Each column represents a single identity. All images have been rescaled to the same resolution.

dimensional feature vector is extracted, consisting of RGB, HSV and YCbCr colour values and Gabor and Schmid texture filters. After building these low-level feature descriptors, an SVM is trained to carry out attribute detection. However, the dimensionality of attribute features is typically small in size, leading to very similar feature descriptors for different people. The authors prevent this by fusing their attribute descriptor with SDALF features [2].

In recent years, methods are utilizing deep CNNs for attribute detection. Su, Zhang et al. [17] proposed a framework with three stages. The first stage involves training a deep CNN to predict a series of attributes, using a 'sigmoid cross-entropy loss layer' to learn 105 attributes obtained from the PETA data set [18]. The second stage fine-tunes the model using the MOTChallenge [19] data set, whilst also incorporating the ID labels for each image. Triplets are produced which consist of an anchor image, an image with the same identity as it, and an image with a different identity to it. Given these triplets, the authors use triplet loss to force features extracted from individuals of the same identity to be more similar than those with a different identity. The final stage combines all previously used data sets and performs fine-tuning. The results show superior generalisation to other methods and state-of-the-art matching rates. The authors then extend their solution [20], which divides the 105 attributes from the PETA data set into a set of types, such as *age*, *gender* and *hair style*. This allows them to enforce only a single positive attribute for each type. The output of the

model is altered from a feature vector of length 105, to a collection of K attributes belonging to C types, $A = \{A^1, A^2, ..., A^C\}$. For a given type, c, $A^c = \{a_1^c, a_2^c, ..., a_{K^c}^c\}$, denoting the label of each attribute present within type c, where $a \in \{0, 1\}$. By enforcing only a single positive attribute per type, the system removes nonsensical combinations such as labeling the presence of both *short hair* and *long hair*. This binary attribute feature is then used as the final feature vector, increasing matching rates compared to their previous work.

Khamis, Kuo et al. [21] combine both traditional appearance features and attribute features. The authors extract appearance information, and follow by learning a distance metric which projects images of the same identity closer than images of those with a different identity. Afterwards, the projected subspace is augmented using semantic attribute information, increasing its invariance to pose and illumination variation. Their proposed method then jointly optimizes both the ranking loss and attribute classification loss. The authors demonstrate that their method of combining the two feature types outperforms using an individual feature type. Ye, Zhou, and Dong [22] propose a body parts-based also combining both LOMO [3] hand-crafted features and attribute features. Attribute features are learnt by using a LIBSVM [23] to generate an attribute classifier for each attribute. A Sample-Specific SVM (SSSVM) [24] is utilised to weight each body part according to each parts contribution to Re-ID matching. The weighted distances between corresponding parts of different images are then fused to form the final distance between two images, which demonstrates high performance compared to other state-of-the-art methods. In [25] the authors extend traditional attribute methods by utilising video sequences to improve Re-ID matching rates. Features extracted from single frames are divided into groups of sub-features, which each correspond to specific attributes, and are then weighted according to the confidence of the attribute prediction. Finally, the features across the set of video frames are aggregated across the temporal dimension to produce the final feature vector.

3 Method

In this section, we describe our approach to first predict the skeleton and relative widths of people, and uses this information to divide an image into three parts. The original image and the divided sub-images are then used to train an attribute prediction model, which is combined with deep feature extraction to perform matching.

3.1 Deep Foreground Appearance Modelling

We use the work proposed in [1,26] to learn a regression between image appearance information and skeleton keypoints using a deep CNN. This method has been shown to be able to predict accurate skeletons even with the inherent low resolution of Re-ID images. Thus, we first apply data augmentation to all images in the training and validation sets by creating additional images and skeletons by inducing small rotations, translations and reflections in the y-axis. Each skeleton consists of a series of keypoints, representing the head, torso and limbs. In total, there are 15 (x, y) keypoints representing the central axis of the limbs, plus an additional 14 (x, y) keypoints representing the widths of the limbs, located perpendicular to the end of the limb's central axis. Figure 2 shows examples of images and their

Figure 2: Four examples of images from the 3DPeS [12, 13] data set and their corresponding ground-truth skeletons. The skeleton can be seen marked in red, whereas the limb widths are marked in green.

corresponding ground-truth skeletons.

The CNN takes the person images as input, and outputs the skeleton keypoints. We take advantage of transfer learning by passing the images through the ResNet-50 architecture [6] with pre-trained weights, and therefore resize all images to the required 224×224 resolution. To adapt the ResNet-50 model to our task, we remove the fully-connected layers and replace them with a fully-connected layer of size 1024 and an output layer of size 58.

3.2 Deep Attribute Prediction

As can be seen in Figure 3, once the skeleton of each image has been predicted, we divide each person image into three sections. The top part consists of the head and shoulders, the middle part of the torso and arms, whilst the bottom part consists of the legs. To allow for skeleton prediction errors, we extend the bounding box around these areas by 15% in each dimension. We resize the four images to a resolution of 224×224 pixels, and apply the standard ResNet-50 preprocessing algorithms [6].

(a) (b) (c)

Figure 3: An example of how the foreground modelling method separates each person into three sections: top, middle and bottom. We create a bounding box around each section, and then add padding of 15% the width and height, to account for any errors in skeleton prediction. The original image and the three sections are then passed through our Attribute Prediction model. (a) The original input image; (b) The original image with the skeleton and parts separation overlayed; (c) The individual body parts images.

Each attribute vector is defined as a binary vector indicating the presence or absence of an attribute. We use four identical ResNet-50-based [6] networks with the pre-trained weights, with each sub-network taking one of four images as input - the whole image, as well as the three cropped images produced with the aid of the skeleton prediction model. We remove the dense layers from ResNet-50, and replace with our own fully-connected layer of size 512, with a sigmoidal activation function and a dropout of 0.5. We then concatenate the output of each sub-network to create a 2048-dimensional feature vector, and finally append a fully-connected layer of size n, where n is the number of attributes to be predicted. The 2048-dimensional vector forms our final deep attribute feature. The architecture for our attribute prediction model is shown in Figure 4.

4 Results and Discussion

In the following section, we discuss in detail the data sets and other evaluation settings used when training and testing our models.

4.1 Training

We evaluate on three public data sets, whilst we train on a separate set of data sets. For training the skeleton prediction model, we use the 3DPeS [12, 13] and QMUL GRID [14, 15]

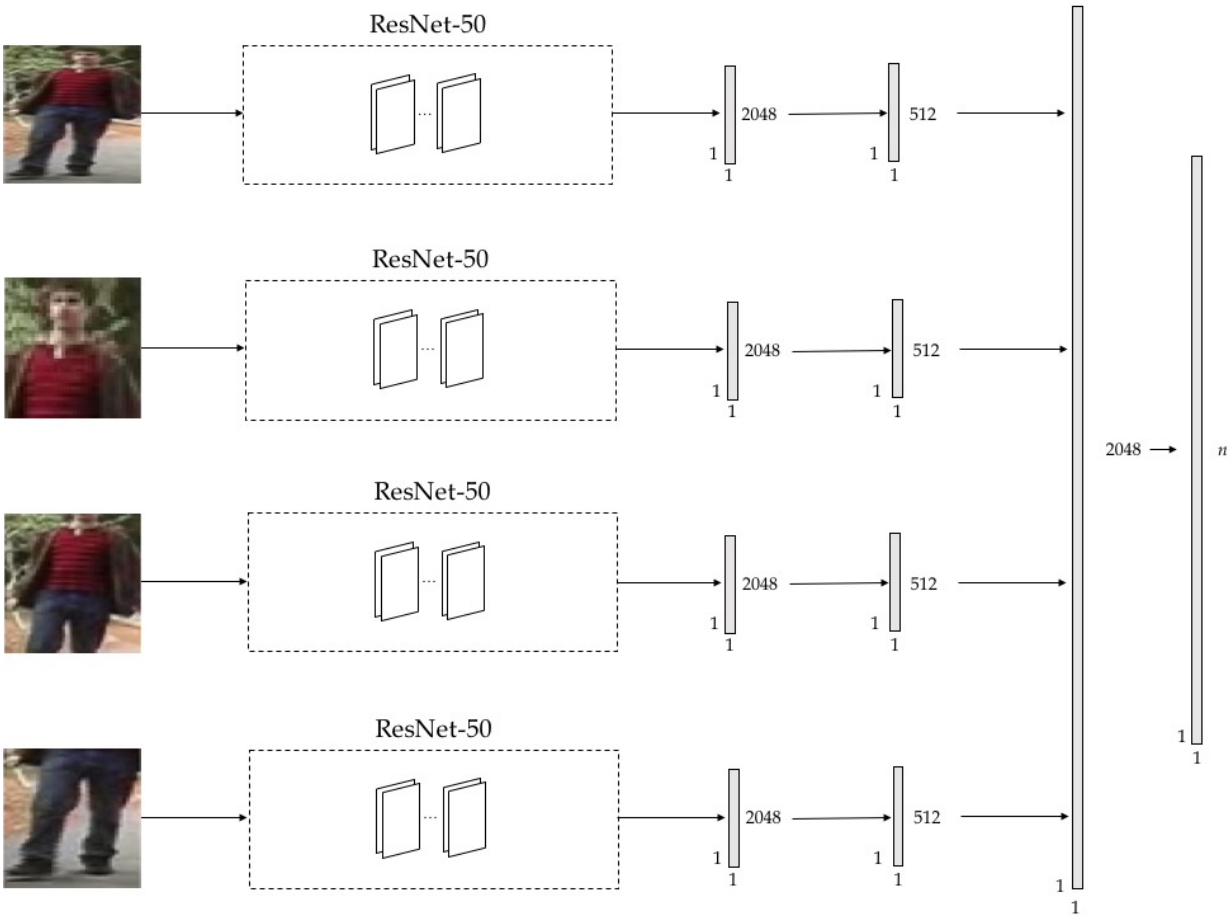
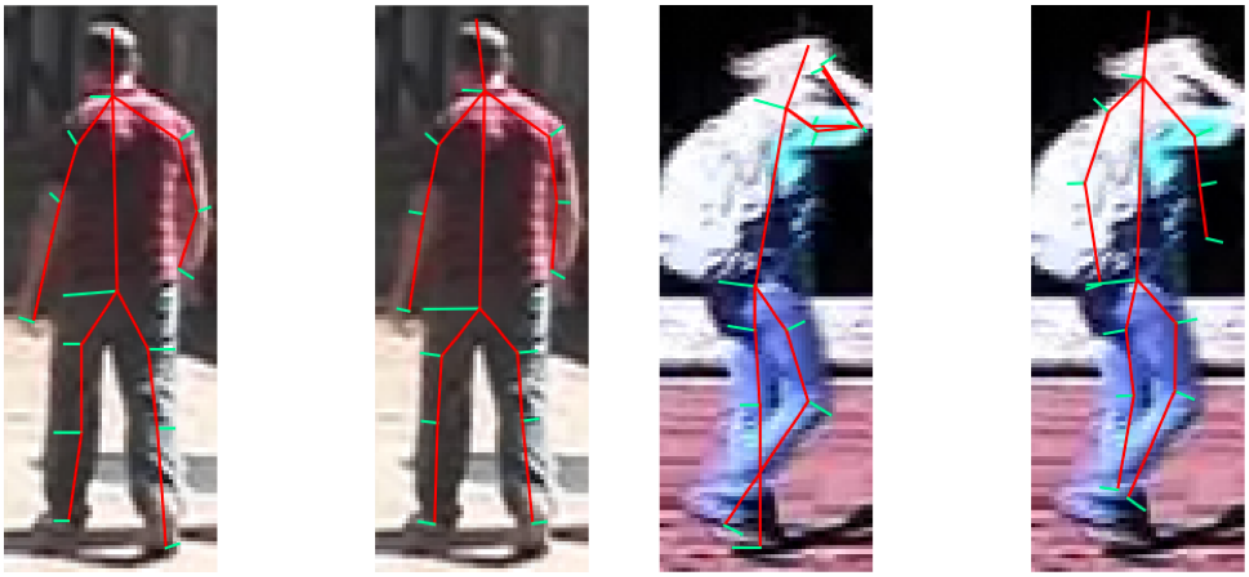


Figure 4: The network architecture of our deep attribute prediction model. We first split the image into four parts - the original, the top, middle and bottom, and pass each image through an identical ResNet-50 [6] network architecture. We remove ResNet-50's fully connected layers and replace with our own - we add a fully connected layer of size 512 to each individual ResNet-50 model, and concatenate. Finally, we append a fully connected layer of size n, where n is the number of attributes being predicted.



(a) (b)

Figure 5: Examples of person images and their ground-truth and predicted skeleton from the VIPeR [7] data set: (a) The image with the minimum root-mean-square error (RMSE) of 2.12 pixels; (b) The image with the maximum RMSE of 15.77 pixels. The average RMSE across the entire VIPeR data set is 5.67 pixels.

data sets:

- 3DPeS [12, 13] contains multiple images of 193 pedestrians taken in varying poses and illumination conditions. The images have a variety of different sizes and shapes. Most images are cropped to the person, but there are a minority of images taken in low sunlight which are cropped to the person and their shadow.

- QMUL GRID [14, 15] contains 250 image pairs taken in an underground transport station. There are also an additional 775 person images which do not share an identity with any of the other images. Images in this data set have a variety of different sizes and shapes, and have numerous occlusion and pose variations.

For the 3DPeS data set, we take approximately 80% of all identities as training and 20% as validation. For the QMUL GRID data set, we first separate the images into those which form an image pair, and those which do not, before taking 80% of identities from each set to form the training set, and the remaining 20% to form the validation set. This ensures that the 80/20 split not only relates to identities, but also to images. We first train the final two fully-connected layers only with a batch size of 32 for 15 epochs, followed by training the network from ResNet-50's third stage onwards, also using a batch size of 32 for 15 epochs. We use the RMSProp [27] optimizer with a learning rate of 0.001, and use a mean squared error loss. Two comparisons between ground-truth skeletons and the corresponding predicted skeletons can be seen in Figure 5.

For the attribute model, we train on a subset of the PETA [18] data set. The PETA data set consists of the VIPeR [7], 3DPeS [12, 13], CAVIAR4REID [28], CUHK [29–31], QMUL GRID [14,15], i-LIDS [9,10], MIT [32], PRID2011 [8], SARC3D [12] and TownCentre [33–35] data sets. In addition to the 3DPeS [12, 13] and QMUL GRID [14, 15] data sets, we select the following for training our attribute model:

- CAVIAR4REID [28] contains multiple images of 72 people taken inside a shopping centre. The data set has severe amounts of pose and illumination variation, but mainly from variation in resolution.
- CUHK [29–31] is a combination of various CUHK Re-ID data sets. Images in this data set are clear and have a consistent resolution, but many images contain occlusion and pose variation.
- MIT [32] consists of 888 images of people taken at ground level. Images are clear and have a fixed, consistent resolution. Occlusion is generally absent and illumination levels are mostly constant. All images are taken at ground level of either the front or back of the person.
- SARC3D [12] consists of 200 images of fifty people, with one image taken of the front, back, left and right view of each person. The data set has significant variation in pose and illumination, as well as large differences in image resolution and clarity.
- TownCentre [33–35] is a large data set consisting of 6967 images of 222 people. The data set contains occlusion, as well as pose and illumination variation. Images are of a variety of different sizes and resolutions, with image clarity being inconsistent between images.

We allocate approximately 80% of the identities in each data set to the training set, with the remaining 20% to the validation set. For each identity, the PETA data set provides information on the presence of 105 attributes, such as whether or not a person is carrying a backpack, the colour of their upper body, and the length of their hair. From this information, we select the fifty most common attributes and produce a binary vector. We first train our attribute model from our appended 512-dimensional fully-connected layer onwards for 5 epochs with a batch size of 16, followed by all layers from ResNet-50s third-stage onwards for an additional 30 epochs, also with a batch size of 16. We use the Adam optimizer [36,37] with a learning rate of 10^{-5}, and a binary cross entropy (BCE) loss. Examples of correctly (true-positive) and incorrectly (false-positive) classified images for two attributes can be seen in Figure 6. Examples of attribute accuracy on the VIPeR [7] data set can be seen in Table 1.

4.2 Testing and Re-Identification

We evaluate our method, which we name *Deep Features & Attribute Detection* (DFAD) by calculating the rank-n score for each data set: the percentage of probe images with their positive match found in n or fewer guesses. We evaluate on the following three public data sets:

(a)

(b)

Figure 6: Two examples of attribute prediction results by our model. All images in (a) are predicted to be wearing red on their upper body, whilst all images in (b) are predicted to be carrying a backpack. Images correctly classified (true-positive) have a green border, whilst those incorrectly classified (false-positive) have a red border. We report the predicted probability of the presence of each attribute below each image.

Attribute	Accuracy (%)	Attribute	Accuracy (%)
lowerBodySuits	99.8	carryingPlasticBags	96.5
footwearStocking	98.4	lowerBodyCasual	96.4
upperBodyPlaid	97.8	lowerBodyShortSkirt	96.2
upperBodyFormal	97.5	upperBodyRed	95.8
lowerBodyFormal	96.9	upperBodyCasual	95.4

Attribute	Accuracy (%)	Attribute	Accuracy (%)
footwearSneaker	66.5	footwearWhite	60.0
personalLess30	64.5	accessoryNothing	57.2
lowerBodyBlack	63.8	footwearShoes	57.1
lowerBodyGrey	63.7	footwearBlack	55.1
carryingNothing	62.7	upperBodyOther	53.4

Table 1: Attribute detection accuracy on the VIPeR [7] data set. The best and worst attributes detection accuracies are shown.

- VIPeR [7] is one of the main data sets used for the Re-ID problem, and contains 632 image pairs. Images contain strong variations in pose and illumination, but are cropped to the pedestrian and have a consistent resolution of 128×48 pixels.
- PRID2011 [8] consists of images from two cameras, with 385 people appearing in the first camera, and 749 appearing in the other; 200 people appear in both. All images are cropped to the person, with the resolution being consistent, yet the data set has strong variations in pose and poor image clarity. Most images in this data set have a strong blue tint.
- i-LIDS [9, 10] contains multiple images of 120 pedestrians, taken at a busy airport. Image quality is generally poor and occlusion is frequent in this data set. This images in this data set also contain a large variation in pose and image size.

For each data set, we use the features extracted from the penultimate, 2048-dimensional layer of our attribute prediction network as our feature vector. We apply ℓ_2-normalization to all vectors prior to matching. Rather than calculating the Euclidean or cosine distance between feature vectors, we learn a distance metric between the features by using the Cross-view Quadratic Discriminant Analysis (XQDA) distance metric learning technique [3]. By combining more traditional distance metric learning methods with dimensionality reduction, XQDA allows us to find a subspace of features such that the distance between features with matching identities is minimised. Whilst we do not use any images from VIPeR, PRID2011 or i-LIDS to train the skeleton or attribute prediction networks, we do allocate a set of images from these data sets to train the XQDA distance metric when testing on these data sets.

For the VIPeR data set, we randomly select 316 identities for testing, with the other 316 identities being used for training the XQDA distance metric. For PRID2011, we select 100 from the 200 identities that are present in both cameras to form our testing set, with the

other 100 being used for training the XQDA distance metric. The remaining 549 identities only present in the second camera are added to the testing gallery set. Similarly to [38], for the i-LIDS data set, we separate the data set into 69 training identities and 50 testing identities. For all data sets, we use the single-shot approach and carry out our evaluation ten times, averaging to produce the final result. To evaluate the effect and contribution of our deep attribute features when combined with traditional hand-crafted features, we extract LOMO [3] features from the original (i.e. not parts-based) images and concatenate to our deep attribute features.

From Table 2, we can see that our attribute model performs competitively against other state-of-the-art methods. On the VIPeR data set [7], we can see that by combining the deep attribute features with LOMO [3] features, we can achieve a 7.3% increase in the rank-1 rate vs. using the LOMO features alone. When using only the deep attribute features, our proposed method performs similar to the closest attribute-only method, ACSM [39], in rank-1 score, but performs significantly better at higher ranks. We can see similar results on the PRID2011 data set, where we see an improvement of 8.7% when compared to using the LOMO features alone. For the i-LIDS [9,10] data set, our deep attributes + LOMO method achieves a 8.9% increase when compared to only using the LOMO [3] features, whilst out deep attributes only method achieves an increase of 24.1% versus the closest attribute-only method.

4.3 Experimentation with different number of parts-based images

We perform further evaluation by experimenting with the number of parts-based images used within our attribute detection network. For this purpose, we extract attribute features using a different combination of the original and parts-based images as input. We train three networks, one which takes the original and three-parts based images as input, a second which takes as input only the three parts-based images, and a final network which takes only the original image as input. We evaluate on the VIPeR [7] data set, and compare the results in Table 3.

From Table 3, we observe that when using attribute features in combination with LOMO features (DFAD (+ LOMO + XQDA)), the highest results are obtained when using the original images in combination with the three-parts based images. However, we also observe that training the network using only the three parts-based images produces rank-n scores only slightly lower than those obtained when training using the original and three parts-based images. When training on the original images only, we observe that the rank-n scores are significantly lower than in other experiments. Similarly, when using only the attribute features (DFAD (+ XQDA)), we also observe an increase in rank-n score when using the network trained using the original and three parts-based images. Under this scenario, the increase seen when using the original and three parts-based images is significantly larger across all rank-n scores, demonstrating the importance of using all four images as input to produce robust attribute features which can achieve high matching rates.

	VIPeR			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA)	45.7	**76.0**	85.2	94.2
DFAD (+ XQDA)	16.3	38.4	52.4	66.6
WSMTAL (+ XQDA) [20]	**47.1**	71.5	80.3	88.2
BPBPR [22]	44.7	-	84.5	92.1
DLDAFN [40]	44.1	72.6	81.7	91.5
AFSB (+ LOMO + XQDA) [41]	43.9	-	**86.6**	**94.6**
CVSP (+ LOMO) [42]	43.0	73.0	84.2	92.8
FT-CNN (Comb. + Multi) (+ XQDA) [43]	42.5	72.0	83.0	92.0
MTL-LOREA [44]	42.3	72.2	81.6	89.6
LOMO (+ XQDA) [3]	38.4	69.4	80.5	91.5
JLAC [21]	29.5	60.3	76.0	87.3
SCAKR (Kernel + Attributes) [38]	28.0	57.1	70.8	83.7
SCAKR (Kernel only) [38]	26.3	54.7	68.4	81.7
ACSM [39]	16.4	34.3	45.2	-
AFSB (+ XQDA) [41]	13.4	-	72.5	93.3
SCAKR (Attributes only) [38]	10.1	24.4	35.3	48.8

	PRID2011			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA)	**32.9**	**55.7**	**67.7**	**79.4**
DFAD (+ XQDA)	13.2	24.8	32.5	45.6
BPBRP [22]	28.2	-	61.0	70.4
WSMTAL (+ XQDA) [20]	24.4	52.3	62.5	74.2
LOMO (+ XQDA) [3]	24.2	48.2	59.3	71.3
MTL-LOREA [44]	18.0	37.4	50.1	66.6
RF+MA+AC [45]	6.5	22.0	32.5	47.6

	i-LIDS			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA)	**57.3**	**85.0**	**92.8**	**97.4**
DFAD (+ XQDA)	45.8	76.4	87.1	94.8
LOMO (+ XQDA) [3]	48.4	76.4	87.1	95.3
SCAKR (Kernel + Attributes) [38]	44.1	64.9	76.3	89.2
SCAKR (Kernel only) [38]	42.7	62.0	74.6	86.7
SCAKR (Attributes only) [38]	21.7	41.3	56.8	77.0

Table 2: Matching results on the VIPeR [7], PRID2011 [8] and i-LIDS [9,10] data sets. Most of the results in this table use attribute data.

4.4 Weighted Binary Cross Entropy

As the prevalence of each attribute can vary significantly from attribute-to-attribute, several methods have been proposed which attempt to mitigate the negative effects of class imbal-

	VIPeR			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA) (Original + Three Parts-based images)	**45.7**	**76.0**	**85.2**	**94.2**
DFAD (+ LOMO + XQDA) (Three Parts-based images only)	45.2	74.0	84.8	94.0
DFAD (+ LOMO + XQDA) (Original images only)	39.8	70.6	83.2	92.4
DFAD (+ XQDA) (Original + Three Parts-based images)	16.3	38.4	52.4	66.6
DFAD (+ XQDA) (Three Parts-based images only)	13.7	33.1	44.7	58.7
DFAD (+ XQDA) (Original images only)	8.8	23.5	35.7	50.1

Table 3: Results on the VIPeR [7] data set utilising different combinations of the original and parts-based images. Models are trained with BCE loss. The best results are highlighted in bold.

ance [46–51]. For our work, we perform additional experimentation by using a Weighted Binary Cross Entropy loss function (WBCE), replacing the BCE loss function used in prior experiments. Let t_i^j represent the i^{th} attribute of the j^{th} person. For each attribute i, we calculate the ratios of positive to negative instances by:

$$pos_i = \frac{1}{p}\sum_{j=0}^{p-1} t_i^j, \tag{1}$$

$$neg_i = 1 - \frac{1}{p}\sum_{j=0}^{p-1} t_i^j, \tag{2}$$

where p is the number of attribute vectors in the training set. We can use these ratios to calculate a weight w_i for each attribute, used to weight the cost of a positive error relative to a negative error such that:

$$w_i = \frac{neg_i}{pos_i}, \tag{3}$$

Inspired by the implementations used by Tensorflow [52] and Tensorpack [53], we calculate the Weighted Binary Cross Entropy, $loss$, as:

$$loss = (\mathbf{1} - \mathbf{z})\mathbf{r} + \mathbf{m}(log(\mathbf{1} + exp(-abs(\mathbf{r}))) + max(-\mathbf{r}, 0)), \tag{4}$$

which outputs a vector containing the component-wise weighted logistic losses, where $\mathbf{z}$ is the ground-truth attribute vector, $\mathbf{r}$ is the predicted attribute vector and $\mathbf{m}$ is equal to $(\mathbf{1} + (\mathbf{w} - \mathbf{1})\mathbf{z})$. As $\mathbf{z}$ is a binary vector representing the presence of absence of a set of I attributes:

	VIPeR			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA) (WBCE)	**47.2**	**76.1**	**86.7**	**94.7**
DFAD (+ LOMO + XQDA) (BCE)	45.7	76.0	85.2	94.2
DFAD (+ XQDA) (WBCE)	17.0	41.1	54.7	69.0
DFAD (+ XQDA) (BCE)	16.3	38.4	52.4	66.6

Table 4: Results on the VIPeR [7] data set utilising WBCE loss and BCE loss. The best results are highlighted in bold.

	PRID2011			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA) (WBCE)	32.8	**56.4**	**68.1**	78.3
DFAD (+ LOMO + XQDA) (BCE)	**32.9**	55.7	67.7	**79.4**
DFAD (+ XQDA) (WBCE)	13.6	28.9	38.8	50.0
DFAD (+ XQDA) (BCE)	13.2	24.8	32.5	45.6

Table 5: Results on the PRID2011 [8] data set utilising WBCE loss and BCE loss. The best results are highlighted in bold.

$$m_i = \begin{cases} w_i, & \text{if } z_i = 1 \\ 1, & \text{if } z_i = 0 \end{cases} \tag{5}$$

The final loss value is then calculated by weighting each component-wise loss value by its corresponding positive ratio, and calculating the mean:

$$\hat{loss} = \frac{1}{I}\sum_{i=1}^{I}(loss_i \times pos_i). \tag{6}$$

We compare the performance of using the WBCE loss with the previous used BCE loss, and show results in Tables 4-6.

These results demonstrate that both BCE and WBCE loss are able to achieve high matching rates. However, neither WBCE or BCE loss perform significantly better than the other, both showing similar performance, only outperforming the other at certain ranks on

	i-LIDS			
	r=1	r=5	r=10	r=20
DFAD (+ LOMO + XQDA) (WBCE)	**58.5**	83.9	92.5	**97.5**
DFAD (+ LOMO + XQDA) (BCE)	57.3	**85.0**	**92.8**	97.4
DFAD (+ XQDA) (WBCE)	43.9	77.3	86.9	95.0
DFAD (+ XQDA) (BCE)	45.8	76.4	87.1	94.8

Table 6: Results on the i-LIDS [9, 10] data set utilising WBCE loss and BCE loss. The best results are highlighted in bold.

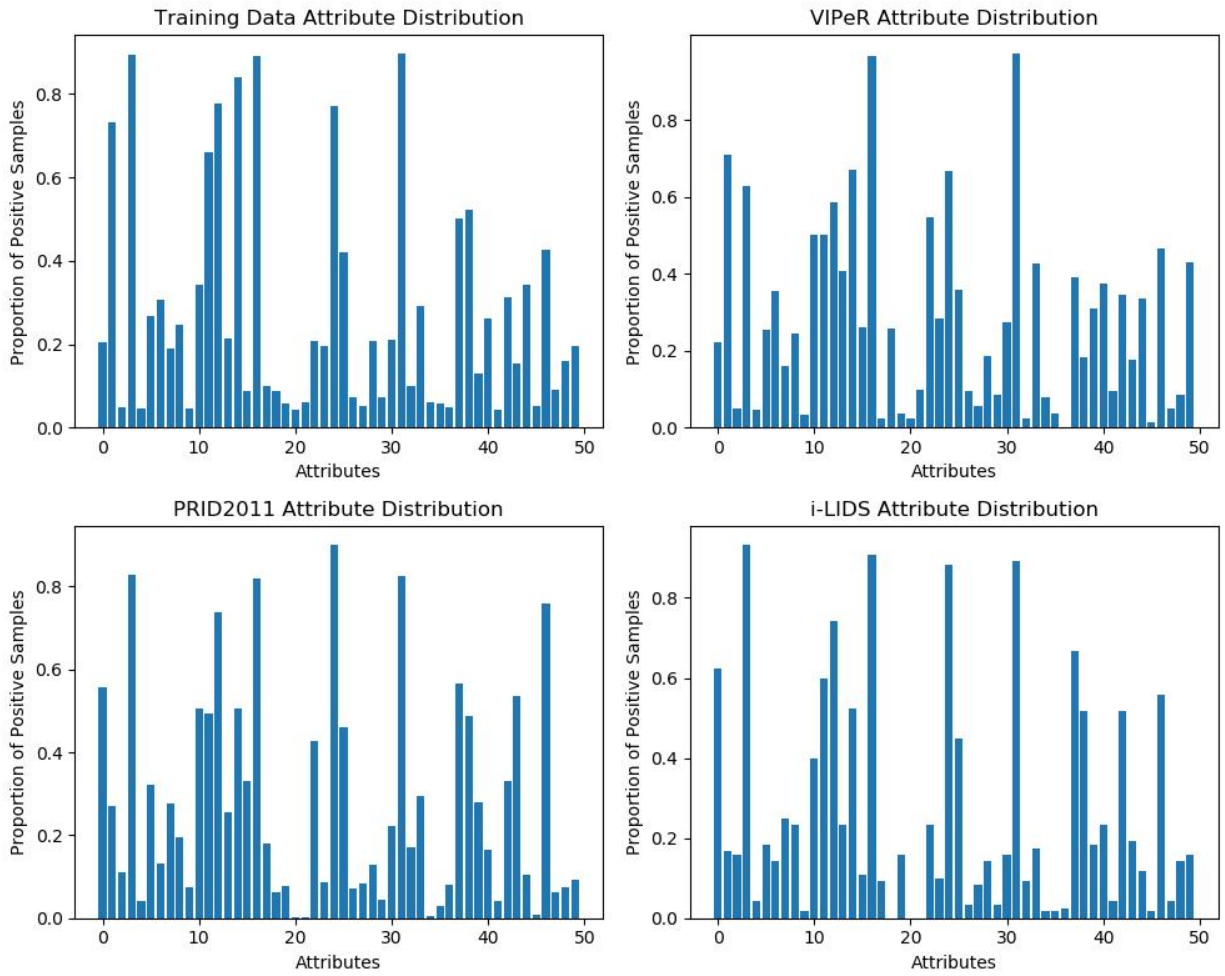


Figure 7: The distribution of attributes on the data sets used to train the attribute model, versus the three data sets used to evaluate the attribute model.

certain data sets. The greatest increase is seen on the VIPeR [7] data set, where all rank-n scores are higher when utilising WBCE loss. A significant increase in rank-1 score is also observed when evaluating on the i-LIDS [9, 10] data set. To investigate why these increases only occur on certain data sets, we compare the distribution of attributes from the training data set with those used to evaluate our approach, which can be seen in Figure 7.

In order to measure the distances between the attribute distribution of the training data set and each evaluation data set, we calculate the cosine distance, which can be seen in Table 7. From Table 7, we observe that the attribute distribution of the VIPeR [7] data set is most similar to that of the training data set. As shown in Table 4, the VIPeR [7] data set also had the largest increase in rank-n score when using the WBCE loss, versus using the standard BCE loss. The i-LIDS [9, 10] data set has the second-most similar attribute distribution to the training data set, and also demonstrated an increased matching score when using WBCE loss at rank-1 and rank-20, whilst having a higher matching score at rank-5 and rank-10 when using the BCE loss. Finally, Table 7 shows that the data set with the largest difference between its attribute distribution and that of the training data

VIPeR	PRID2011	i-LIDS
0.049	0.076	0.061

Table 7: The cosine distance between the attribute distribution of the training data set and each evaluation data set.

set is PRID2011 [8], which produced better rank-1 and rank-20 scores when using BCE, whilst only showing marginally improved rank-5 and rank-10 scores when using WBCE. From this experimentation, we have demonstrated that data sets with more similar attribute distribution to the training data set are more likely to benefit when using a weighted loss function, where weights are derived from the prevalence of attributes within a training data set.

5 Conclusions

In this paper, we proposed using attribute information to construct a feature for person re-identification. We first trained a deep CNN to learn a mapping between a series of person images and their corresponding skeletons. We then used this model to predict the skeleton of a series of unseen person images. Using this skeleton information, we separated a person image into three parts - top, middle and bottom. The original image and three parts were then used to train an attribute prediction deep CNN, and the penultimate layer of this network was used as our feature vector. Prior to performing matching, the extracted features were then weighted using the XQDA distance metric learning technique.

We demonstrated that using a deep attribute feature vector computed with the aid of spatial information can produce a significantly improved matching result. When considering only our attribute feature combined with the XQDA distance metric learning technique, we can see that our results improve on other attributes-only methods. Given that some attributes are only present in a particular part of the image, such as shoes only being found on a person's feet, we believe that the increase in results can be credited to the spatial separation of body parts and extracted features. As an additional experiment, we concatenated the LOMO [3] feature to our deep attribute feature vector, and improved matching rates further. We experimented on the VIPeR [7], PRID2011 [8] and i-LIDS [9, 10] data sets, and achieved a 7.3%, 8.7% and 8.9% increase in rank-1 rate respectively versus using only the LOMO features.

Although we have demonstrated that a significant increase in matching results can be obtained by incorporating spatial information, we observed that the skeleton prediction model struggled with people in unusual poses, such as with hands-raised (see Figure 5), due to lack of training data with these poses. Such inaccuracies in the predicted skeleton will lead to incorrect body part segmentation and thus lead to greater attribute prediction errors. Future work will focus on alternate methods for foreground modelling, such as incorporating a greater variety of person poses into the training set to improve the ability to more accurately predict these poses.

In addition, we have demonstrated that the combined skeleton and attribute models generalise well between data sets. Whilst we evaluated on the VIPeR [7], PRID2011 [8]

and i-LIDS [9, 10] data sets, none of these contributed training images to either the skeleton or the attribute models. Generalisation is very important within Re-ID due to the large variation in pose, illumination and resolution between different data sets, as well as due to the need for large amounts of data required for training a deep CNN. In particular, we note that the PRID2011 data set has a significant blue tint which is not present in the other data sets. However, even though no images from the PRID2011 data set were used for training the skeleton or attribute models, the features extracted from our attribute model were still able to perform competitively against other state-of-the-art methods. We believe that this demonstrates that the use of attributes can help to overcome the issues that more traditional methods face when dealing with the variation present in Re-ID images.

An extension of this work is to incorporate Zero-Shot Identification [54], which deals with the situation where the testing set may contain novel classes not present in the training set. An example of this within the context of attributes would be the presence of a highly distinctive, unseen attribute within a person image. The incorporation of such methods that could deal with this issue might greatly improve the accuracy of attributes and have the potential to increase Re-ID matching rates significantly. Furthermore, more recent data sets often include short video sequences, and as such an extension of this work may be to improve attribute prediction performance and Re-ID matching rates by utilising video sequences to produce more robust space-time features.

Acknowledgements

The authors gratefully acknowledge funding by the UK Engineering and Physical Sciences Research Council (grant no. EP/L016400/1), the EPSRC Centre for Doctoral Training in Urban Science.

References

[1] Gregory Watson and Abhir Bhalerao. Person re-identification using partial least squares appearance modelling. In *International Conference on Image Analysis and Processing*, pages 25–36. Springer, 2017.

[2] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[3] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[4] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017.

[5] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007.

[8] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[9] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, volume 2, 2009.

[10] UK Home Office. i-lids multiple camera tracking scenario definition. 2008.

[11] Shengcai Liao, Guoying Zhao, Vili Kellokumpu, Matti Pietikäinen, and Stan Z Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1301–1306. IEEE, 2010.

[12] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *Proceedings of the 16th International Conference on Image Analysis and Processing*, pages 197–206, Ravenna, Italy, September 2011.

[13] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects*, pages 59–64, Scottsdale, Arizona, USA, November 2011.

[14] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010.

[15] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision*, pages 391–401. Springer, 2012.

[16] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *Bmvc*, volume 2, page 8, 2012.

[17] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*, pages 475–491. Springer, 2016.

[18] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM, 2014.

[19] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[20] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, 75:77–89, 2018.

[21] Sameh Khamis, Cheng-Hao Kuo, Vivek K Singh, Vinay D Shet, and Larry S Davis. Joint learning for attribute-consistent person re-identification. In *European Conference on Computer Vision*, pages 134–146. Springer, 2014.

[22] Xin Ye, Wen-yuan Zhou, and Lu-an Dong. Body part-based person re-identification integrating semantic attributes. *Neural Processing Letters*, 49(3):1111–1124, 2019.

[23] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[24] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. Sample-specific svm learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2016.

[25] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4913–4922, 2019.

[26] Gregory Watson and Abhir Bhalerao. Person reidentification using deep foreground appearance modeling. *Journal of Electronic Imaging*, 27(5):051215, 2018.

[27] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[28] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Bmvc*, volume 1, page 6. Citeseer, 2011.

[29] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[30] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[31] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[32] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[33] Ben Benfold and Ian Reid. Colour invariant head pose classification in low resolution video. In *Proceedings of the 19th British Machine Vision Conference*, September 2008.

[34] Ben Benfold and Ian Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, September 2009.

[35] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464, June 2011.

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[38] Husheng Dong, Chunping Liu, Yi Ji, Zhaohui Wang, and Shengrong Gong. Fusion of spatially constrained attributes with kernelized ranking for person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015.

[39] Jun Liu, Chao Liang, Mang Ye, Zheng Wang, Yang Yang, Zhen Han, and Kaimin Sun. Person re-identification via attribute confidence and saliency. In *Pacific Rim Conference on Multimedia*, pages 591–600. Springer, 2015.

[40] Lin Wu, Chunhua Shen, and Anton Van Den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017.

[41] Le An, Xiaojing Chen, Shuang Liu, Yinjie Lei, and Songfan Yang. Integrating appearance features and soft biometrics for person re-identification. *Multimedia Tools and Applications*, 76(9):12117–12131, 2017.

[42] Ju Dai, Ying Zhang, Huchuan Lu, and Hongyu Wang. Cross-view semantic projection learning for person re-identification. *Pattern Recognition*, 75:63–76, 2018.

[43] Tetsu Matsukawa and Einoshin Suzuki. Person re-identification using cnn features learned from combination of attributes. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2428–2433. IEEE, 2016.

[44] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3739–3747, 2015.

[45] Chi Su, Shiliang Zhang, Fan Yang, Guangxiao Zhang, Qi Tian, Wen Gao, and Larry S Davis. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping. *Pattern Recognition*, 66:4–15, 2017.

[46] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.

[47] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[48] Shiven Sharma, Colin Bellinger, Bartosz Krawczyk, Osmar Zaiane, and Nathalie Japkowicz. Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 447–456. IEEE, 2018.

[49] Qiong Wu, Pingyang Dai, Peixian Chen, and Yuyu Huang. Deep adversarial data augmentation with attribute guided for person re-identification. *Signal, Image and Video Processing*, pages 1–8, 2019.

[50] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[51] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[52] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[53] Yuxin Wu et al. Tensorpack. `https://github.com/tensorpack/`, 2016.

[54] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng. Zero-shot person re-identification via cross-view consistency. *IEEE Transactions on Multimedia*, 18(2):260–272, 2016.