A Convolutional Baseline for Person Re-Identification Using Vision and Language Descriptions

Ammarah Farooq, Muhammad Awais, Fei Yan, Josef Kittler, Ali Akbari, and Syed Safwan Khalid

***Abstract*—Classical person re-identification approaches assume that a person of interest has appeared across different cameras and can be queried by one of the existing images. However, in real-world surveillance scenarios, frequently no visual information will be available about the queried person. In such scenarios, a natural language description of the person by a witness will provide the only source of information for retrieval. In this work, person re-identification using both vision and language information is addressed under all possible gallery and query scenarios. A two stream deep convolutional neural network framework supervised by cross entropy loss is presented. The weights connecting the second last layer to the last layer with class probabilities, i.e., logits of softmax layer are shared in both networks. Canonical Correlation Analysis is performed to enhance the correlation between the two modalities in a joint latent embedding space. To investigate the benefits of the proposed approach, a new testing protocol under a multi modal ReID setting is proposed for the test split of the CUHK-PEDES and CUHK-SYSU benchmarks. The experimental results verify the merits of the proposed system. The learnt visual representations are more robust and perform 22% better during retrieval as compared to a single modality system. The retrieval with a multi modal query greatly enhances the re-identification capability of the system quantitatively as well as qualitatively.**

***Index Terms*—Cross modal retrieval, vision, language, person ReID, person search.**

I. Introduction

In recent years, the task of person re-identification (ReID) has gained considerable attention in the research community. It aims at recognising a person across non-overlapping camera views. The problem incorporates many underlying computer vision challenges such as illumination changes, occlusion, background changes, pose variations and varying camera resolution. It has significant applications in security and video surveillance, making it even more challenging to work with low resolution CCTV footage. The increasing public safety demands and large networks of installed surveillance cameras are making it difficult to rely solely on manual practice of tracking and spotting a person across various cameras. A typical ReID system takes an image query and searches for the corresponding person in the pool of gallery images (Figure 1) or videos. Thus, this scenario assumes that a visual example of a person identity is always available as a query. A huge amount of literature and benchmarking datasets

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/dstl/MURI project EP/R018456/1.

Fig. 1. A typical person ReID task takes vision based query only. The proposed system works with vision only, language only as well as both modalities as query.

are available for vision based re-identification [1]–[11]. Conventional person ReID approaches mainly focus on hand-crafted visual features [1]–[3] and learning discriminative metrics [4]–[6]. Inspired by the success of convolutional neural networks (CNNs) in large scale visual classification [12], modern approaches [7]–[11] are mainly relying on CNN based feature representations and attention learning mechanisms.

However, in many practical cases, no visual information for the person of interest is available for matching. In such cases, language description of the person plays an important role to gather useful clues to assist in identification. Such cases are more relevant to the open-set scenarios. For example, we may have a description of a missing person, a description of a criminal by witnesses and shortlisting suspects using appearance cues, where query identity may or may not be present in the available gallery data. Presumably, like images, textual data also contain unique representations to identify a person. Recently, Yan *et al.* [13] and Li *et al.* [14] pioneered the language based person retrieval, where the former one specifically dealt with person ReID. However, this aspect of person ReID has not yet been explored in more detail by the research community.

Attributes based person retrieval [15]–[17] offers an alternative to a free-form description based approaches. Such approaches are being used by police where semantic attributes related to clothes, gender, appearance etc. are assigned according to the information given by witnesses and then the search is performed by comparing with the meta-data of the gallery database. However, the use of free-form natural

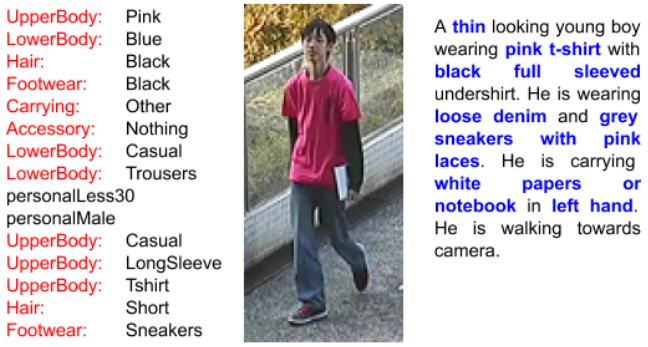

Fig. 2. An example of attribute based annotation from PETA dataset vs natural language description

language descriptions offers much more flexibility and robustness as compared to these predefined attributes. Attributes have limited capability of describing persons appearance. For instance, the PETA dataset [18] defined 61 binary and 4 multiclass person attributes, while there are hundreds of words for describing a person′s appearance. On the other hand, even with a constrained set of attributes, labelling them for a large-scale person image dataset is cumbersome and requires more mindfulness from annotators.

In contrast to attributes, unique details can be captured by natural language descriptions as shown in Figure 2. For example, "grey sneakers with pink laces" compared to only "black" and "sneakers" is more distinctive and accurate. Furthermore, descriptions can include indications of a degree of certitude or ambiguities to preserve as much information as possible ("white papers or notebook"). Another important aspect is the expansion of the set of allowed assignments for attributes ("pink t-shirt with black undershirt" compared to "pink t-shirt"). Natural language descriptions can describe all such attributes with any possible word and in any possible length. Another limitation of the attribute based person retrieval is the sensitivity of the retrieval results. We show in section V-E that even flipping a few attributes, while using the ground truth set for retrieval, leads to significant performance drop in the results. In practice at test time these attributes are generated automatically, using a machine learning algorithm which will have some error. The sensitive nature of the attribute based person retrieval will have significant impact on the performance in the presence of automatically generated attributes.

With the recent advances in the field of computer vision and natural language processing, joint modelling of vision and natural language is finding applications in image caption generation [19]–[21], bidirectional image-text retrieval [22]–[25], visual question answering [21], [26], text to image generation [27], [28] and language assisted visual navigation [29]. Hence, there is a strong basis to assess the merit of the combined vision and language based techniques in the context of person ReID.

In this work, person ReID using both person images (vision) and corresponding descriptions (language) is addressed. The proposed framework is based on two deep residual CNNs jointly optimised with cross entropy loss to embed the two modalities into a joint feature space. A well-known statistical technique, canonical correlation analysis (CCA) is adopted to further enhance the cross modal retrieval capability. The experimental results on a large scale datasets in various practical scenarios are reported. The scenarios include retrieving from vision to vision, language to vision, vision-language to vision and other combined cross modal scenarios. It extends our conference version paper [13] in the following aspects:

- We propose to change the language network to deep residual network having similar number of layers as the vision branch.
- We propose a joint optimisation strategy based on cross entropy loss for the two networks along with Canonical Correlation analysis (CCA) to model the embedding space for cross modal retrieval.
- We propose new testing protocols under the cross modal ReID setting for two large-scale datasets namely CUHK-PEDES and CUHK-SYSU, as compared to Viper and CUHK03 data, to verify the robustness of the proposed approach on a large number of identities coming from diverse domains.
- The proposed approach is directly compared with state-of-the-art algorithms for person search on the CUHK-PEDES benchmark.
- Extensive experiments have been performed with large scale data under various practical scenarios: vision to vision, language to vision, language to language and vision language to vision.

The rest of the paper is organised as follows. Section II presents a review of existing ReID methods and cross modal retrieval research. Methodology Section III describes the proposed joint vision-language CNN framework in detail. The implementation details and evaluation protocols are provided in Section IV. In Section V, a discussion on various issues, including training strategies, the experimental results, the merits of natural language descriptions versus attributes is presented along with the qualitative results. Finally, Section VI concludes the paper.

II. Related Work

A. Vision based Person Re-identification

Vision based person ReID has gained notable attention in all of its aspects. Research has focused mainly on two issues: robust feature extraction and distance metric learning. The early approaches were focused on improving hand-crafted features such as SIFT, LBP, histograms in colour spaces [2] and the hierarchical Gaussian descriptor [1]. Following the success of convolutional neural networks, recent approaches attempt to learn deep representations along with background noise removal [11], using human pose masks [8], [30], [31] to attend solely people in images and minimize the effect of occlusion. Other techniques [7], [10], [32]–[35] focus on fusing global features from the whole image and body part based local features.

In terms of distance metric learning, Liao *et al.* [4] proposed cross-view quadratic discriminant analysis (XQDA) that be-

came a popular metric for person ReID. Given features from two views of a person, a low dimensional target space is learnt using eigen decomposition along with a distance function to measure similarity in the subspace. Another measure was proposed by Zhong *et al.* [36], namely re-ranking to use a k-reciprocal nearest neighbours encoding to re-rank the retrieved results using the Jaccard distance to enhance the retrieval performance.

One important work to mention is Chen *et al.* [9] who used language as an additional supervision signal in the training phase to learn robust visual features. A global-local strategy was adopted to semantically align the visual features according to text. A global association was established according to the identity labels, while the local association was based upon the implicit correspondences between image regions and noun phrases. However, the retrieval was performed using only vision features.

B. Person Search

The task of person search shares the idea of using natural language descriptions for retrieving a person/pedestrian image. However, unlike person ReID, it does not impose constraints of across camera/pose retrieval. The initial work by Li *et al.* [14] introduced a large scale person search (CUHK-PEDES) data. They proposed a CNN-RNN based network to learn word by word affinity of a sentence with image features. The final affinity was obtained by applying a word level gated neural attention mechanism. In a follow-up work, Li *et al.* [37] used a two stage CNN-LSTM network to first learn the identity aware cross modal embedding space using cross modal cross entropy (CMCE) loss. In the next stage, image regions corresponding to each word were spatially attended followed by a latent alignment of the encoded features to enhance the robustness against sentence structure variations. In order to achieve better local matching between two modalities, Chen *et al.* [38] used a patch-word matching and learnt an adaptive threshold for each word to enhance or suppress its effect on the final text-image affinity. A major improvement in performance was achieved by the work of Zheng *et al.* [23]. They proposed a dual CNN architecture to extract image-text features and used each image-sentence pair as a single class to train the model. Finally, they fine-tuned the networks combined with ranking loss to improve the discriminative ability in a joint embedding space. In a recent work, Jing *et al.* [39] used body pose confidence maps to attend visual features of pedestrians and applied an adaptive similarity based attention mechanism to consider only a description-related part of the image for the final affinity scores.

C. Image-Text Retrieval

Cross modal person ReID is a fine-grained application of learning a joint image-text embedding for person retrieval in a bi-directional manner. Learning a cross modal embedding has many applications in image captioning [19]–[21], visual question answering [21], [26] and image-text retrieval [22]–[25]. Recently, in [24], the authors used a stacked cross attention mechanism to learn the semantic alignment between the objects in image and the corresponding words in the sentence. For each salient object in the image, similarity based attention was computed over the words in the sentence. A relevance score between the corresponding image patch and attended sentence was then computed and used to calculate the final Log-SumExp pooling (LSE) score for matching. A similar attention stacking is applied from text to image and the triplet ranking loss was used as a final objective for optimization. Another work [25], incorporated image-to-text and text-to-image generative models into the conventional cross modal feature embedding. For a given text-input, a GAN was trained to generate an image conditioned on text and similarly, a caption sentence was obtained for the query image. A final representation of each modality is a combination of original features and the features from generated space. Both of these works demonstrated a clear performance gain on MSCOCO and Flicker benchmarks. All of these approaches inspire to push the boundary of cross modal matching in the application of person ReID.

III. Methodology

This section introduces a detailed formulation of the cross modal re-identification system. The cross modal ReID framework, presented in Figure 3, consists of two stream convolutional neural network with identity level classification. This unified framework jointly optimises vision and language modalities to learn a highly distinctive combined feature space. Features obtained from this learnt space are maximally correlated by applying the canonical correlation technique. At the retrieval stage, gallery entities are ranked with respect to their feature similarity with the query feature. Each of the components is discussed in the subsequent sub-sections.

A. Network Architecture

1) Deep Vision Net: In Figure 3, the upper branch constitutes vision CNN which is based on the ResNet-50 [40] model. Unlike the original architecture, the proposed network contains two fully-connected (FC) layers before the classifier layer. Batch normalisation is applied after both FC layers and a dropout is applied before the last layer. The network takes input image of size 224×224 and generates a 2048 dimensional feature vector f_{img}. All images are mean normalized, randomly cropped and horizontally flipped (50% probability) before passing to the network.

2) Deep Language Net: A CNN based language model has been adopted for textual feature learning. The lower branch of the framework in Figure 3 corresponds to the language CNN. It is also a 50 layered deep ResNet [40] modified to deal with one dimensional textual input. The first layer is a special word embedding layer which maps each word to a vector space. The rest of the network is similar to the vision net except that the convolution filters are of size 1×3 instead of 3×3. For each word in a sentence, filter would produce response by looking at the two neighbouring words as well. For a given textual description, the 2048 dimensional feature vector f_{txt} is obtained directly after global average pooling. It is used as a sentence representation at retrieval time.

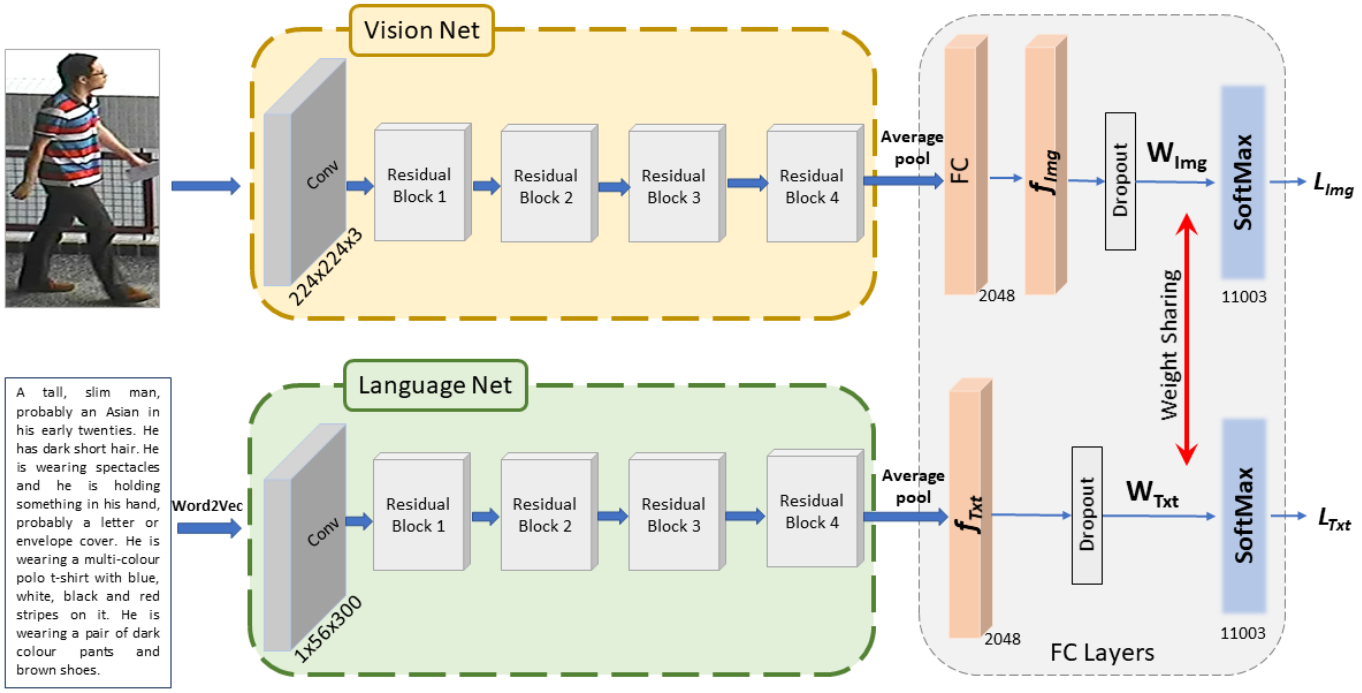

Fig. 3. Two stream CNN framework for cross modal person ReID. Feature representations are learnt based on ResNet-50 [40] model with filter size of 3×3 and 1×3 for image and textual inputs respectively. $\boldsymbol{W_{img}} = \boldsymbol{W_{txt}} = \boldsymbol{W_{joint}}$ are shared between two networks during training.

Text pre-processing and data augmentation: The first step of natural language processing is to map the text corpus into a vector space. Each word in the dictionary is converted to a fixed length vector by an embedding model. In this work, word2vec models pre-trained on Google News corpus (3 billion words) have been used to prepare the dictionary for person descriptions and vector space conversion. Word vectors for words that share common contexts are induced to lie closer to each other in the vector space by using these models. Given a sentence description S containing n words, each word provides a unique index into the corpus. Thus the sentence is converted to a $n \times 300$ matrix where 300 is the output dimension of the word2vec embedding model. As the length of descriptions varies, we have fixed it to the max limit of 56 words to serve as a fixed size input to the CNN. Extra zeros have been appended at the end of shorter sentences.

Appropriate data augmentation techniques can also be applied to the textual data to create diversity in the training dataset. Similar to random image cropping, a random word dropping operation has been adopted. The number of words to drop is chosen dynamically. Inspired by [23], position shifting for appended zeros is also applied. Instead of keeping all zeros at the end of sentence, their positions are rotated randomly to the beginning of sentence as well.

B. Objective Function

Under the category-level supervision, CNNs have shown incredible multi-class discriminative ability [12]. Zheng *et al.* [41] proposed to view person re-identification as a multi-class classification problem and used a transfer learning approach to learn a ReID model, which is termed as identity discriminative embedding (IDE). Strong inter-identity intra-modal representations can be learnt by using this approach. However, for cross modal person Re-ID, given a person image and description, the aim is to learn a joint feature embedding space for retrieval along with such intra-modal inter-identity discriminative representations. Feature representations of a person image and corresponding descriptions are expected to be closer in this joint space. Similarly, for negative pairs, representations should be further away in the joint cross modal space and in the same modality as well.

The proposed framework is based on identity based supervised learning for both CNNs, where all images of a person and corresponding descriptions are considered as a single class. Hence, two softmax losses L_{img} and L_{txt} are employed, corresponding to each CNN. Specifically, given a batch of N image-text pairs belonging to I person IDs, the image ID loss is the average of cross-entropy loss for all images in the batch and is given as:

$$L_{img} = -\frac{1}{N}\Sigma_{n=1}^{N} log\left(\frac{exp((\boldsymbol{w_{img}})_{i(n)}^{T}\ f_{img})}{\sum_{j=1}^{I} exp((\boldsymbol{w_{img}})_{j}^{T}\ f_{img})}\right), \quad (1)$$

similarly, for text ID loss:

$$L_{txt} = -\frac{1}{N}\Sigma_{n=1}^{N} log\left(\frac{exp((\boldsymbol{w_{txt}})_{i(n)}^{T}\ f_{txt})}{\sum_{j=1}^{I} exp((\boldsymbol{w_{txt}})_{j}^{T}\ f_{txt})}\right) \quad (2)$$

where $i(n)$ represents the target class ID of the $n-th$ pair and $(\boldsymbol{w_{img}})_j$ and $(\boldsymbol{w_{txt}})_j$ are the classifier weights for $j-th$ class. Let $\boldsymbol{W_{img}}$ and $\boldsymbol{W_{txt}}$ are the classifier weight matrices with weight of each class j in a row. To enforce the learnt representations to be in common space, the weights of the final classifier layer are shared across the image and

text CNN imposing $\boldsymbol{W_{img} = W_{txt}}$ and further denoted as $\boldsymbol{W_{joint}}$. In essence, this joint training approach leads to learning a classifier (weights) for each identity (class) reflecting the information conveyed by both modalities simultaneously and pushing the corresponding f_{img} and f_{txt} pairs closer. The optimization is performed over the combined objective function defined as:

$$Loss = \lambda_1 L_{img} + \lambda_2 L_{txt}$$

where λ_1 and λ_2 set the relative weights for the two losses. During back-propagation, $\boldsymbol{W_{joint}}$ is updated with respect to the gradients from both modalities by taking the average of the gradients. The classifier weights can also be thought of as a template with which the feature vectors f_{img} and f_{txt} align themselves. In the case of separate weight matrices for each modality, the features vectors f_{img} and f_{txt} will be aligned to their respective weight matrices $\boldsymbol{W_{img}}$ and $\boldsymbol{W_{txt}}$. In contrast, for the case of training with shared weights both feature vectors f_{img} and f_{txt} will try to align themselves with the joint weight matrix $\boldsymbol{W_{joint}}$, hence, bringing them closer in the same feature space.

C. Canonical Correlation Analysis

Canonical correlation analysis (CCA) [42], [43] is a well-known statistical technique to find a space in which two sets of random variables are maximally correlated in the form of their linear combinations. Let $X = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_m}) \in R^{d_x \times m}$ and $Y = (\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_m}) \in R^{d_y \times m}$ are the two sets of random vectors with their co-variance matrices denoted as Σ_{xx}, Σ_{yy} and cross co-variance as Σ_{xy}. The aim of CCA is to explain the correlation structure of X and Y in terms of linear combinations $\mathbf{w_x^T}X$ and $\mathbf{w_y^T}Y$. Concretely, for cross modal ReID, we are seeking those combinations $(\mathbf{w_x^*}, \mathbf{w_y^*})$ which maximise the correlation between the two modalities:

$$\begin{aligned} \rho_{max} &= \text{corr}(\mathbf{w_x^{*T}}X, \mathbf{w_y^{*T}}Y) \\ &= \frac{\text{cov}(\mathbf{w_x^{*T}}X, \mathbf{w_y^{*T}}Y)}{\sqrt{Var(\mathbf{w_x^{*T}}X) \times Var(\mathbf{w_y^{*T}}Y)}} \\ &= \frac{\mathbf{w_x^{*T}}\Sigma_{xy}\mathbf{w_y^*}}{\sqrt{\mathbf{w_x^{*T}}\Sigma_{xx}\mathbf{w_x^*}\mathbf{w_y^{*T}}\Sigma_{yy}\mathbf{w_y^*}}} \end{aligned} \tag{3}$$

By substituting $\mathbf{w_x} = \Sigma_{xx}^{-1/2}\mathbf{u}$ and $\mathbf{w_y} = \Sigma_{yy}^{-1/2}\mathbf{v}$, the above equation becomes

$$\rho_{max} = \frac{\mathbf{u}^T\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{u}\mathbf{v}^T\mathbf{v}}} \tag{4}$$

In practice, it is assured that Σ_{xx} and Σ_{yy} are non-singular by using regularization. In that case, a singular value decomposition can be used to solve eq. 4 as

$$\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2} = U\Lambda V \tag{5}$$

where $\mathbf{u}$ and $\mathbf{v}$ in eq. 4 correspond to the first left and right singular vectors in U and V respectively with top singular value in Λ equal to ρ_{max} which gives the maximum correlation between $\mathbf{w_x^T}X$ and $\mathbf{w_y^T}Y$ and this optimum is attained at $(W_x^*, W_y^*) = (\Sigma_{xx}^{-1/2}U, \Sigma_{yy}^{-1/2}V)$.

In the proposed framework for cross modal ReID, the 2048 dimensional feature vectors f_{img} and f_{txt} from the two modalities serve as two sets of random variables which are jointly correlated in the (W_{img}^*, W_{txt}^*) space by CCA.

IV. Experiments

A. Datasets

1) CUHK-PEDES Data: Currently, the CUHK person description (CUHK-PEDES) data [44], introduced for the person search task, is the only large scale available data with natural language descriptions. It contains 40,206 images by combining several ReID datasets, for 13003 identities. Each image is annotated by two sentence descriptions giving about 80,440 descriptions in total. The data has been divided into predefined train/val/test splits.There are 3,074 test images with 6,156 descriptions for 1000 identities, 3,078 validation images with 6,158 descriptions and 34,054 training images with 68,126 descriptions for 11003 identities.

2) Cross Modal ReID Data: Since there is no available data specifically designed for cross modal ReID and to do a fair evaluation under the person ReID constraints, a new evaluation protocol on the test and validation split of CUHK-PEDES has been proposed in this work. The data has been carefully separated across poses and across cameras so that gallery and query images are disjoint (Figure 4). Person IDs having no change in pose have been discarded, resulting in the total of 1594 final distinct IDs. The gallery set contains 2935 images and 5870 descriptions. The query set contains 2127 images and 4264 descriptions. This annotated data is referred to as the cross modal ReID (crossRe-ID) data in the rest of the paper. All the results are reported for the single-shot single query (SS-SQ) scenario.

3) CUHK-SYSU Data: This dataset has been introduced for the joint person detection and identification task [45]. Each identity has at least two images and the descriptions are available from CUHK-PEDES data. The training set contains 15080 images and 30160 descriptions for 5532 identities. The test set includes 8341 images and 16690 descriptions for 2900 identities. A careful annotation has been performed to split poses and viewpoints as much as possible on the test set resulting in 3271 query images and 5070 gallery images. For this dataset, the number of test IDs has been preserved without discarding any image. In most cases where the pose or viewpoint were the same, there were still other challenges like occlusions and severe blur or low resolution. All the results are reported for the SS-SQ scenario.

B. Implementation Details

The network training has been performed using a stochastic gradient descent (SGD) algorithm with momentum fixed to 0.9. For both CNNs, dropout is set to 0.75 indicating 75% randomly switched off nodes. The number of FC layers in both networks are empirically chosen to maximise performance. The maximum sentence length is set to 56 following [23].

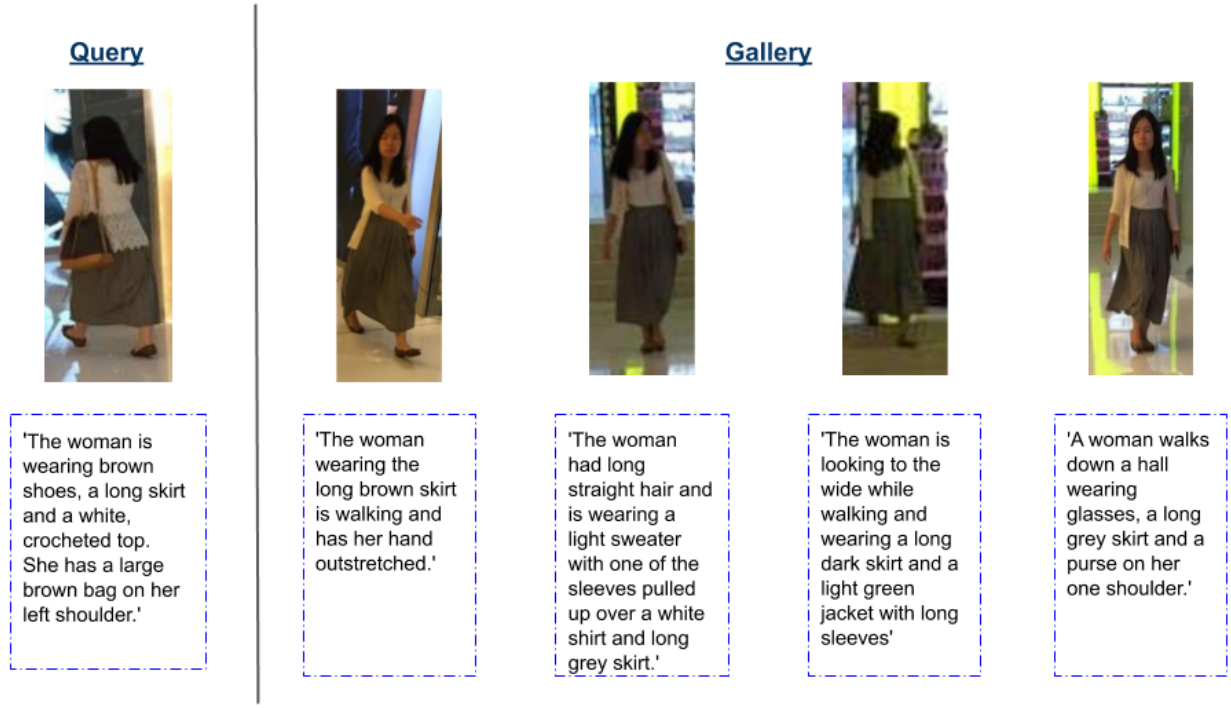

Fig. 4. Example of query/gallery pair from annotated cross modal ReID data. Descriptions are annotated per image instead of per identity. Different views have different descriptions and distinctive details.

The average sentence length is about 19 words. The input to the language net is 1×56 vector where each entry indicates the index to the word in the dictionary. For the CUHK-PEDES data, dictionary size is 7263 words. The word2vec initialized embedding layer then generates $1 \times 56 \times 300$ dimensional data as input to the first convolution layer. The embedding layer parameters are also tuned during the training to increase the robustness.

The network has been trained for 300 epochs with a batch size 64 and a learning rate of 0.01, which decreases by a multiplicative factor of 0.1 after every 100 epochs. λ_1 and λ_2 are set to 1 for an equal contribution of both modalities. In the next section, the training strategies are discussed, based on the network initialisation choices and the number of learning stages. The implementation has been realised in the PyTorch framework [46].

C. Evaluation

Following the ReID literature, three evaluation metrics; recall@K (Rank@1,5,10), mean average precision (mAP) and median rank (medR) have been adopted. At the test time, image feature f_{img} and language feature f_{txt} are obtained independently from the two CNNs. The retrieval is based on cosine similarity between the query and the gallery features which can lie in the range [-1,1].

$$S(f_{Query}, f_{Gallery}) = \left(\frac{f_{Query}}{\parallel f_{Query} \parallel} \right)^T \frac{f_{Gallery}}{\parallel f_{Gallery} \parallel}$$

In the reported results, a query and gallery pair is denoted as Query $\times$ Gallery. Table I indicates the gallery and query features used in each retrieval scenario.

TABLE I
GALLERY AND QUERY FEATURES USED FOR EVALUATION. W^*_{img} AND W^*_{txt} REPRESENT CCA PROJECTED SPACES. [A,B] REPRESENTS A CONCATENATION OF THE TWO VECTORS.

Retrieval Case	Query Features	Gallery Features	Size
V X V	f_{img}	f_{img}	2048
L X L	f_{txt}	f_{txt}	2048
L x V	$W^*_{txt} \times f_{txt}$	$W^*_{img} \times f_{img}$	2048
VL X V	$[f_{img}, W^*_{txt} \times f_{txt}]$	$[f_{img}, W^*_{img} \times f_{img}]$	4096
VL X VL	$[f_{img}, f_{txt}]$	$[f_{img}, f_{txt}]$	4096

V. RESULTS AND DISCUSSION

A. Training Strategies

To avoid compromising one modality with another and gain better insight into the joint training of a multi-modal system, the retrieval performance of different training strategies defined by various initialisation of the weights and learning schemes has been analysed. The strategies are presented in Table II. The weight parameters of the vision CNN and language CNN up to the average pool operation are denoted by W_{Vision} and W_{Lang} respectively. Similarly, the weights for all fully connected layers, including the shared classifier layer, are denoted by W_{FC}.We have also trained the vision and language CNNs separately on the CUHK-PEDES data with 11003 train IDs and used these weights for the initialisation of W_{Vision} and W_{Lang}. These weights are denoted by "Person Init" in the table. For vision, we also used a pre-trained ImageNet model in Strategy 4. In most cases, the networks are trained in two stages, freezing a set of parameters in the first stage and tuning all the parameters together in the second stage.

TABLE II
WEIGHTS INITIALISATION AND LEARNING STRATEGIES. W_{FC} ARE INITIALISED TO RANDOM VALUES IN FIRST TRAINING STAGE AND ARE TRAINED IN ALL STAGES.

Strategy	Weights	Person Init.	ImageNet Init.	Learnable in Stage 1
1	W_{Vision}	✓	✗	✗
	W_{Lang}	✓		✗
2	W_{Vision}	✓	✗	✓
	W_{Lang}	✓		✓
3	W_{Vision}	✓	✗	✗
	W_{Lang}	✓		✓
4	W_{Vision}	✗	✓	✗
	W_{Lang}	✗		✓
5	W_{Vision}	✗	✗	✓
	W_{Lang}	✓		✗

TABLE III
COMPARISON OF TRAINING STRATEGIES ON CROSSRE-ID DATA

Rank@1	V × V	L × L	L × V
St1: pretrained + tune last layers	71.2	18.38	18.69
St2: pretrained + tune entirely	73.5	15.4	16.56
St3: pretrained (2 stage training)	71.0	14.17	15.62
St4: ImageNet + Lang. scratch	82.05	14.93	24.6
St5: Vision scratch + pretrained lang.	71.5	19.44	18.88

Table III compares the performance for the above mentioned strategies. It is interesting to note that ImageNet initialisation (St4) offers better starting point and a wider learning surface as compared to the person specific weights for vision. It is also evident from St3 and St4 where the learning policy is identical. Another observation in this regard is that vision trained from scratch (St5) is similar to that trained from person specific weights (St1,2,3). The choice of initial weights is greatly affecting the results. However, for language, the performance is affected by the number of learning stages. Strategies 1 and 5 performed better with one learning stage for the text CNN as compared to Strategies 3 and 4 where the performance is similar regardless of the initialisation choice. The cross modal retrieval performance is directly proportional to vision; being stronger and less ambiguous compared to language. These results suggest that Strategy 4 be adopted for further evaluation.

B. Results on Cross Re-ID

In this section, cross modal re-identification performance obtained using the proposed testing protocol is discussed. Since, there is no previous work available for this task, the results are compared with the separate-training technique proposed by Yan *et al.* [13] in two ways. First, the language network is kept the same as [13] with one convolution layer after word embedding and two FC layers. This version is referred as "Yan *et al.* [13]" in the results. Second, the language network is changed to the proposed deep residual network. It is referred as "separately train" in the results. In both versions, the vision and language CNNs are trained separately ($\boldsymbol{W_{img} \neq W_{txt}}$) and CCA is applied for joint embedding. The results for the first learning stage of the joint optimisation, where vision CNN is frozen to ImageNet and other network parts of the system are trained from scratch, are also mentioned as Stage 1. Rank@1 comparison is presented in Table IV and the detail quantitative results for each query-gallery scenario are presented in Table V.

TABLE IV
RANK@1 RETRIEVAL PERFORMANCE ON CROSSRE-ID DATA

Rank@1 (%)	Yan *et al.* [13]	Separately Train + CCA	Jointly Train Stage 1	Jointly Train Stage 2 + CCA
V × V	-	59.91	28.48	82.05
L × L	5.39	37.32	14.8	14.93
L × V	9.22	13.6	7.46	27.9
VL × V	63.6	65.87	37.7	84.7
VL × VL	58.84	68.0	42.09	80.86
VL × VL	-	-	-	84.06

The results in Table IV validate the scientific hypothesis behind learning the classifier layer jointly. The weights for each ID are forced to achieve maximum alignment between the two modalities. Clearly, the proposed approach outperforms the separate training solutions by a large margin in all cases. Specially, in the cross modal L × V scenario, the gain is about 14%. In all other cases, language has helped in learning more distinctive visual features. By using both modalities, the rank@1 performance has further increased by 2.65%. It indicates that language provides auxiliary information to better define the query ID. Referring to Table V, Rank@10 suggests that in all scenarios where both modalities are available as a query, the results have the correct match for person in top ten ranks across pose variations with more than 96% accuracy. However, one striking observation in these results is nearly 4% decrease in VL × VL performance as compared to VL × V. This contrasts with the separate training and Stage 1 for the joint training cases where the increase has been observed. The reason is clear from the severe performance decline of the language modality, while the vision modality appears to become more robust through joint optimisation. We noticed that the language acts as a supervisory signal for vision. Moreover, the vision is a stronger modality, hence after joint training, the shared weights are more influenced by vision and more aligned with it, resulting ina better performance of VxV. On the other hand the language, being weaker modality, has less influence on the weights and hence, the performance of LxL drops. The last row of the table corresponds to the performance when we use the jointly trained vision model with the separately trained language model. As expected, a better language model improved the results to 84% without any joint embedding between the two features. These results offer an interesting direction for further investigation.

The first two columns of the Table IV support the idea of incorporating a deeper language model. The proposed deep language network achieved higher accuracy as compared to the simple model [13]. However, for VL × V, it is interesting to note that even the weaker language model helps in retrieval compared to the vision only scenario.

With regard to the L × V scenario, note that the above mentioned results correspond to a query sentence description for one view and gallery images corresponding to a different view. From one view to another, the description of person can change greatly and the set of discriminative words to identify the subject across the views may become very small. In such

TABLE V
DETAILED RETRIEVAL PERFORMANCE ON CROSSRE-ID DATA IN TERMS OF RANK@K, MAP AND MEDR

Model	Rank@1 (%)	Rank@5 (%)	Rank@10 (%)	mAP (%)	medR
V × V					
Separately Train + CCA	59.91	80.5	85.7	64.45	1
Jointly Train + CCA	82.05	94.3	96.8	84.75	1
L × V					
Separately Train + CCA	13.6	32.99	43.04	18.5	15
Jointly Train + CCA	27.9	50.6	60.7	33.4	5
VL × V					
Separately Train + CCA	65.87	84.19	88.9	64.8	1
Jointly Train + CCA	84.7	95.0	97.1	84.1	1
VL × VL					
Separately Train + CCA	68.0	84.7	89.58	71.8	1
Jointly Train + CCA	80.86	94.16	96.6	83.85	1

TABLE VI
TEXT-TO-IMAGE RETRIEVAL PERFORMANCE ON CROSSRE-ID DATA WITHIN POSE AND ACROSS POSE

L x V Rank (%)	Separately Train + CCA	Jointly Train Stage 1	Jointly Train stage 2 + CCA
Across Pose	13.4	7.46	27.9
Within Pose	17.6	10.8	40.46

a challenging case, across pose cross modal retrieval accuracy of 27.9% is quite encouraging. In Table VI, the results are reported for both, within and across pose language to vision retrieval. The joint training not only increased the across pose performance, compared to separately trained networks, but the within pose retrieval is enhanced by a massive 23%. In comparison to across pose scenario which is more challenging, within pose has higher performance. For example, the performance level for within pose is 4% better for separately trained networks and 13% for jointly trained networks. These results broadly support the intuition of cross modal feature alignment with joint optimisation.

C. Results on Person Search

The proposed approach has been evaluated on the person search task under L × V scenario. The quantitative and qualitative results on the test set of CUHK-PEDES are reported in Table VII and Figure 5, respectively. The results are compared with the current state-of-the-art DPC approach [23] and other techniques. Note that Zheng *et al.* [23] has also used a two stage training procedure by including a ranking loss along with ID level losses in the second stage. The proposed joint learning with CCA and identity loss only has achieved competitive results with 2% boost in rank@1 accuracy.

TABLE VII
RETRIEVAL PERFORMANCE ON THE CUHK-PEDES DATA

Model	Rank@1 (%)	Rank@5 (%)	Rank@10 (%)	mAP (%)	medR
Yan et al. [13]	17.5	35.25	46.34	16.65	13
Separately Train + CCA	19.6	38.88	50.2	18.2	10
GNA-RNN [44]	19.05	-	53.64	-	-
IATV [47]	25.94	-	60.48	-	-
GDA [9]	43.58	66.93	76.26	-	-
DPC [23]	44.4	66.26	75.07	-	2
Jointly train + CCA (proposed)	46.44	67.9	76.3	41.3	2

D. Results on CUHK-SYSU

The results under the re-identification setting on CUHK-SYSU are presented in Table VIII. As mentioned earlier, the test data is annotated under the strict ReID setting and the challenging single shot setting. Again, vision achieved nearly a 20% gain by learning jointly with language. Rank@1 of 77.82% and 80.5% mean average precision is achieved under the multi modal scenario. The results obtained are quite competitive and set a baseline for future work. The choice of this data for cross modal ReID is mainly due to its size, number of IDs and the availability of corresponding language descriptions.

TABLE VIII
DETAILED RETRIEVAL PERFORMANCE ON CUHK-SYSU DATA IN TERMS OF RANK@K, MAP AND MEDR

Model	Rank@1 (%)	Rank@5 (%)	Rank@10 (%)	mAP (%)	medR
V × V					
Separately Train + CCA	55.03	71.41	77.89	58.95	1
Jointly Train + CCA	74.13	87.2	90.48	77.16	1
L × V					
Yan *et al.* [13]	3.03	11.06	17.51	5.48	89
Separately Train + CCA	2.31	9.13	15.06	4.57	94
Jointly Train + CCA	11.37	28.41	38.06	15.78	24
VL × V					
Yan *et al.* [13]	58.72	74.82	79.44	57.01	1
Separately Train + CCA	58.72	74.79	79.44	57.0	1
Jointly Train + CCA	77.68	89.2	92.03	75.8	1
VL × VL					
Yan *et al.* [13]	57.77	77.17	82.13	62.13	1
Separately Train + CCA	59.65	76.27	81.34	63.3	1
Jointly Train + CCA	77.82	89.31	92.41	80.4	1

E. Natural Language vs Attributes

The task of person re-identification using descriptions closely resembles the attributes based person retrieval. In this section, we show, with the help of experiments, that using unstructured language descriptions as compared to attributes enhances retrieval robustness. For this purpose, we choose the Market-1501 dataset which has been manually annotated with 27 appearance based attributes [48]. There are 8 colours for upper-body clothing, 9 colours for lower body clothing, 1 for age with four possible values and 9 other binary attributes including gender, hair length, carrying backpack, wearing hat etc. For evaluation, we use 750 test set identities. First we extract their corresponding language descriptions from CUHK-PEDES which contains around four images per identity. We then select from this subset 750 images from one camera as the gallery and 750 images from another camera as a query set. Finally, we concatenate the corresponding attributes or descriptions for retrieval.

Since the attributes are annotated at the identity level, there exists no notion of viewpoint. When we concatenate ideal attributes for gallery and query, 93.46% rank@1 is achieved. However, in practical systems, it is highly unlikely to obtain all attributes correctly on an unseen query. If attributes from one viewpoint are provided for search, then the system may overlook other important attributes which may be visible from other viewpoint and tries to match according to the given set only. In Table IX we investigate the sensitivity

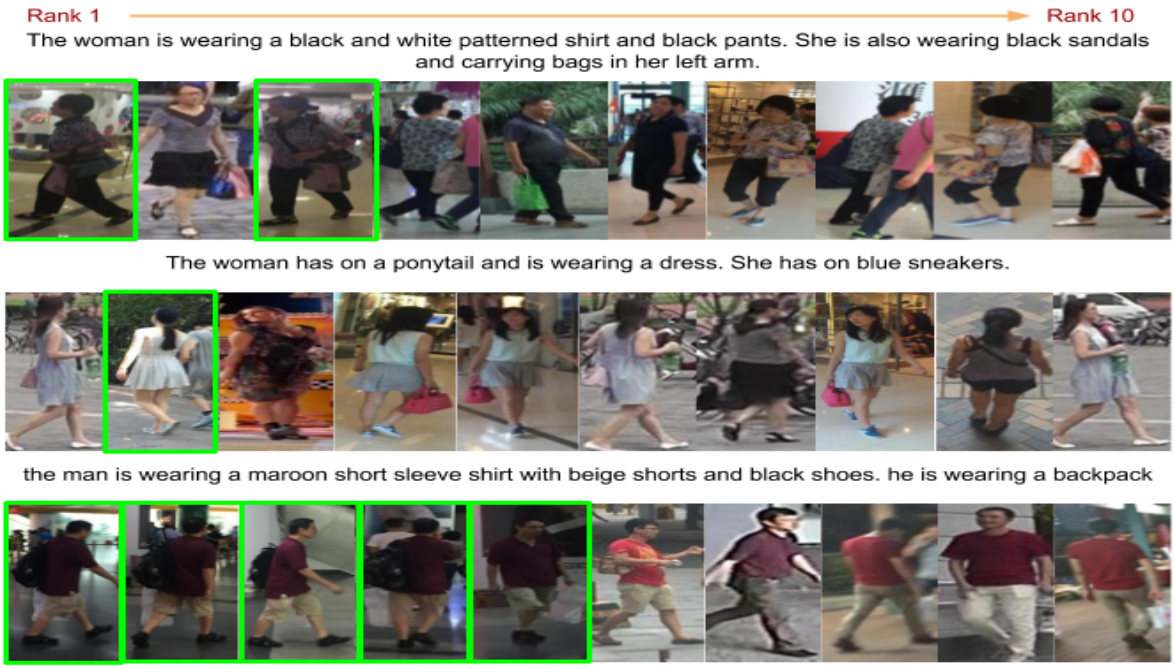

Fig. 5. Qualitative results on CUHK-PEDES in the LXV scenario. Correct matches are shown by green boxes.

of attribute based retrieval to missing attributes. Note that if we change just one attribute in query (N=1), we observe 4.5% decrease. By changing two attributes, the performance we achieve drops again and becomes comparable to VL $\times$ VL. The accuracy decreases by 7.2% compared to sentence descriptions and by 20% from ideal attributes just by flipping three random attributes. Practical attribute prediction systems can not produce 100% accurate attribute features for unseen data and consequently the retrieval will always be subject to attribute flipping . The results reported in the table are obtained with ground truth attributes and still a significant decrease in the performance has been observed. With the predicted attribute features, retrieval will always suffer from viewpoint variance. On the other hand, the proposed joint training system is trained on image level annotations with many different words and sentences to describe one person. In such a case, the system tries to learn different semantic concepts present in the images instead of memorising individual identities.

F. Discussion on Language to Language Performance

As observed in Table IV, vision features become more discriminative when trained with language descriptions.In contrast language only performs poorly, compared to the separately trained system. One possible reason could be that as the joint optimisation forces the two modalities to lie closer in the embedding space, language tends to become more specific to match with its corresponding image because the weights $\boldsymbol{W}_{joint}$ serve as a common template for alignment. These weights are more influenced by stronger modality and thus force the weaker modality to follow along. As a result, language becomes more sensitive compared to the vision modality. An evidence that can be qualitatively observed is

TABLE IX
RANK@K PERFORMANCE ON THE MARKET-1501 DATA WITH VISION AND ATTRIBUTES

	Rank @ 1 (%)	Rank @ 5 (%)	Rank @ 10 (%)	mAP (%)	medR
V $\times$ V	74.66	92.66	96.13	78.70	1
VL $\times$ VL	81.60	95.46	97.46	84.64	1
VA $\times$ VA	93.46	99.4	100	94.89	1
VA $\times$ VA (N = 1)	88.93	99.2	100	91.2	1
VA $\times$ VA (N = 2)	82.93	97.86	99.3	86.25	1
VA $\times$ VA (N = 3)	74.40	95.73	98.53	78.85	1
VA $\times$ VA (N = 4)	62.12	90.26	94.9	67.98	1
VA $\times$ VA (N = 5)	50.53	78.80	88.26	56.87	1

that after joint training the overall text classification (word-to-word) accuracy has been improved as shown in Table X. Since the ReID evaluation criterion is based on identity, the L $\times$ L retrieval exhibits a decrease as the same set of words can describe two different persons. For future investigation, it will be interesting to examine adjective-noun association, impact of non-informative frequent words and the organisation of the sentences for learning feature representations.

G. Qualitative Results

Figure 6 presents qualitative results on the crossRe-ID dataset. The results are shown for both separate and joint modelling. In the first two examples, there is no correct match for the given query for V $\times$ V case with separate modelling. However, it can be seen that with the help of language, correct matches are observed in top five retrieved images. Specifically, if we look at the second example, top match for the vision only (separate training) is influenced by the background of the image. On the other hand, textual description served as an attention to focus on the person only features. Thus, the

Fig. 6. Qualitative results on crossRe-ID dataset. Correct matches are shown by green boxes. For each sample, first two rows show results for separately trained model and last two rows show results for the jointly trained model. For both models, language has refined the feature representations and boosted the retrieval performance.

TABLE X
QUALITATIVE RESULTS FOR L × L USING JOINTLY TRAINED MODEL. SAMPLES PROVIDED ARE CORRECTLY RETRIEVED BY SEPARATELY TRAINED MODEL. THE HIGHLIGHTED PHRASES SHOW THE IDENTICAL WORDS IN THE QUERY AND RETRIEVED SENTENCE.

Query	Ground Truth	Retrieved
'A man wearing a black jacket with long sleeves, a pink shirt, a pair of dark colored pants and a pair of black shoes.'	'I am watching a man walking by, he is wearing all black with a pink dress shirt, and a long coat, carrying a newspaper, and looking around as if he's somewhat lost.'	'A woman wearing a pink shirt, a pair of black pants and a pair of black shoes.'
'The lady wears a white and black shirt white and black pants with grey and white sneakers she walks inside the building and carries a black back pack.'	'The woman is wearing a black and white patterned outfit, and has a large pack on her back.'	'The man wears a blue and white stripped shirt black pants with black and white sneakers he carries a black back pack.'
'A pedestrian walking to the right with a black bag on their back,grey pants, and red shoes. '	'The girl has her hair tied back and has a back pack on. She has her back to the onlooker and has dark pants on. Her pack has a touch of red on it and her sneakers are red as well.'	'This person is walking. They have on a red shirt and light shorts. they have a purse or bag over their right shoulder and are carrying something in their right hand.'
'The woman is wearing a white dress and pink shoes. She is carrying a black bag and is looking down at her phone.'	'white dress asain women pink shoes black long hair wearing backward backpack and glasses black back pack and gold phone'	'The woman is talking on the phone and she is wearing a white dress with white shoes. She also is carrying a grey and black shoulder bag'

top five matches contains a white dress and black hair. In case of joint training, it is evident that the corresponding vision features become stronger and use of language feature further pushes the correct match towards the top ranks. The performance of the proposed system has also been tested practically, including using query images from CCTV cameras which are not part of the dataset. This exercise has helped in gaining insight on the challenges in practical applications. We have observed that with image only, we are able to get closer to main concepts like gender of person and colours of the clothes. By adding language description to image, we obtain refined results capturing more details.

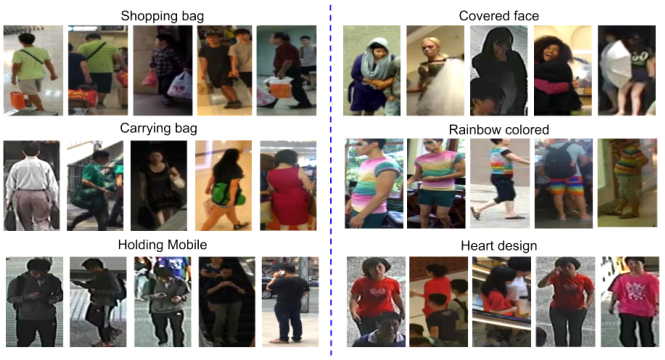

Fig. 7. Retrieved images corresponding to different semantic concepts

It is also interesting to examine whether the individual semantic concepts in a language description are implicitly represented by the individual visual units in f_{img}. Some examples of the images obtained for various semantic concepts are in Figure 7. Instead of the full description, we just passed the desired phrases as input to the language CNN. The results in the figure confirm the presence of a specific visual unit. This kind of search also assists in database shortlisting for a query, for example "covered face". One important aspect for future studies from these observations is to check the impact of less frequent words.

VI. CONCLUSION

Person re-identification by jointly modelling vision and language reflects better practical scenarios in security applications. In this paper, an integrated baseline framework has been proposed, based on the joint optimisation of the two modalities. We have investigated various training strategies to develop deep understanding of the proposed joint-embedding learning and evaluated the performance on all the possible query-gallery combinations. The proposed joint embedding learning and the two stage training protocol achieved superior results as compared to the separate training strategy, outperforming it by 22% in V×V, 11% in L×V, 18% in VL×V and 12% in VL×VL on crossRe-ID data. These results lay the groundwork for future research in cross modal Re-ID. The proposed method is quite promising as it is competitive to current state-of-the-art for the person search task. Evidently, language provides complementary information and facilitates learning enriched representations for person images. Using free-form descriptions has provided a significant edge over attribute based retrieval in terms of relaxed annotation constraints, incorporating more unique details and overall more powerful language modelling.

REFERENCES

[1] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1363–1372.

[2] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3318–3325.

[3] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2528–2535.

[4] S. Liao, Y. Hu, X. Zhu, and S. 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*, 2015, pp. 2197–2206.

[5] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2666–2672.

[6] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2288–2295.

[7] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1077–1085.

[8] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960–3969.

[9] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 54–70.

[10] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2119–2128.

[11] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5794–5803.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[13] F. Yan, J. Kittler, and K. Mikolajczyk, "Person re-identification with vision and language," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2136–2141.

[14] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[15] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *2009 workshop on applications of computer vision (WACV)*. IEEE, 2009, pp. 1–8.

[16] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. Springer, 2014, pp. 93–117.

[17] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognition*, vol. 75, pp. 77–89, 2018.

[18] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 789–792.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[20] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[21] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[22] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3441–3450.

[23] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen, "Dual-path convolutional image-text embedding with instance loss," *arXiv preprint arXiv:1711.05535*, 2017.

[24] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[25] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7181–7189.

[26] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[27] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5907–5915.

[28] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[29] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 538–12 547.

[30] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[31] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1249–1258.

[32] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[33] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[34] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, 2019.

[35] D. Li, X. Chen, Z. Zhang, and K. 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*, 2017, pp. 384–393.

[36] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.

[37] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1890–1899.

[38] T. Chen, C. Xu, and J. Luo, "Improving text-based person search by spatial matching and adaptive threshold," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1879–1887.

[39] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Cascade attention network for person search: Both image and text-image similarity selection," *arXiv preprint arXiv:1809.08440*, 2018.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1367–1376.

[42] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[43] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[44] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[45] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, vol. 2, p. 2, 2016.

[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[47] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1890–1899.

[48] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, 2019.