Improving Deep Visual Representation for Person Re-identification by Global and Local Image-language Association

Dapeng Chen[1][0000−0003−2490−1703], Hongsheng Li[†1][0000−0002−2664−7975], Xihui Liu[1][0000−0003−1831−9952], Yantao Shen[1][0000−0001−5413−2445], Jing Shao[2][0000−0003−3521−6744], Zejian Yuan[3], and Xiaogang Wang[1]

[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Research   [3]Xi'an Jiaotong University
{dpchen, hsli, xhliu, ytshen, xgwang}@ee.cuhk.edu.hk

Abstract. Person re-identification is an important task that requires learning discriminative visual features for distinguishing different person identities. Diverse auxiliary information has been utilized to improve the visual feature learning. In this paper, we propose to exploit natural language description as additional training supervisions for effective visual features. Compared with other auxiliary information, language can describe a specific person from more compact and semantic visual aspects, thus is complementary to the pixel-level image data. Our method not only learns better global visual feature with the supervision of the overall description but also enforces semantic consistencies between local visual and linguistic features, which is achieved by building global and local image-language associations. The global image-language association is established according to the identity labels, while the local association is based upon the implicit correspondences between image regions and noun phrases. Extensive experiments demonstrate the effectiveness of employing language as training supervisions with the two association schemes. Our method achieves state-of-the-art performance without utilizing any auxiliary information during testing and shows better performance than other joint embedding methods for the image-language association.

Keywords: Person re-identification, Local-global language association, Image-text correspondence

1 Introduction

Person re-identification (re-ID) is a critical task in intelligent video surveillance, aiming to associate the same people across different cameras. Encouraged by the remarkable success of deep Convolutional Neural Network (CNN) in image classification [23], the re-ID community has made great process by developing various networks, yielding quite effective visual representations [1, 12, 24, 28, 34,

---

[†] Hongsheng Li is the corresponding author.

Positive pair   Negative pair   Phrase-Region pair

The guy is wearing a white shirt that has black letters on it and a pair of black pants, He is also wearing blue sneakers.

The girl has long black hair with bangs. She is wearing knee-length blue shorts, a green and grey stripped shirt with something white in her hand.

Noun Phrase Extraction

long black hair with bangs

knee-length blue shorts

green and grey stripped shirt

something white in her hand

(a) Global Discriminative Association   (b) Local Image-language Association
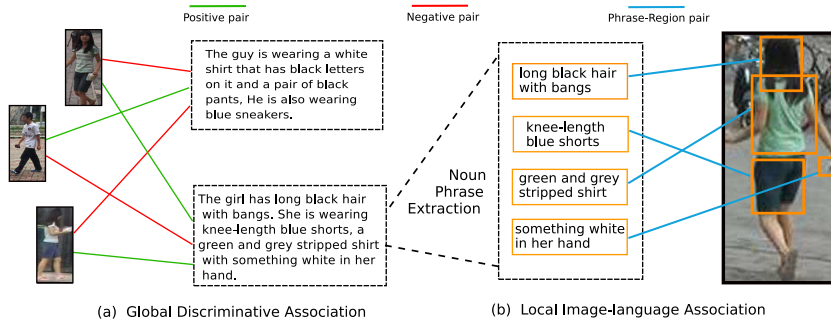
Fig. 1: Illustration of global and local image-language association in our framework. The global association is applied to the whole image and the language description, aiming to discriminate the matched image-language pairs from the unmatched ones. The local association aims to model the correspondences between the noun-phrases and images regions. The global and local image-language association is utilized to supervise the learning of person visual features.

36,45,50,52,67]. To further boost the identification accuracy, diverse auxiliary information has been incorporated in the deep neural networks, such as the camera ID information [32], human poses [59], person attributes [33,48], depth maps [7], and infrared person images [54]. These data are utilized as either the augmented information for an enhanced inter-image similarity estimation [32,54,59] or the training supervisions that can regularize the feature learning process [33,48]. Our work belongs to the latter category and proposes to use language descriptions as training supervisions to improve the person visual features. Compared with other types of auxiliary information, natural language provides a flexible and compact way of describing the salient visual aspects for distinguishing different persons. Previous efforts on language-based person re-ID [26] is about cross-modal image-text retrieval, aiming to search the target image from a gallery set by a text query. Instead, we are interested in how the language can help the image-to-image search when they are only utilized in the training stage. This task is non-trivial because it requires a detailed understanding of the content of images, language, and their cross-modal correspondences.

To exploit the semantic information conveyed in the language descriptions, we not only need to identify the final image representation but also propose to optimize the global and local association between the intermediate features and linguistic features. The global image-language association is learned from their ID labels. That is, the overall image feature and text feature should have high relevance for the same person, and have low relevance when they are from different persons (Fig. 1, left). The local image-language association is based on the implicit correspondences between image regions and noun phrases (Fig . 1, right). As in a coupled image-text pair, a noun phrase in the text usually describes a specific region in the image, thus the phrase feature is more related to some local visual features. We design a deep neural network to automatically asso-

ciate related phrases and local visual features via the attention mechanism, then aggregate these visual features to reconstruct the phrase. Reasoning such latent and inter-modal correspondence makes the feature embedding interpretable, can be employed as a regularization scheme for feature learning.

In summary, our contributions are three-fold: (1) We propose to use language description as training supervisions for learning more discriminative visual representation for person re-ID. This is different from existing text-image embedding methods aiming at cross-modal retrieval. (2) We provide two effective and complementary image-language association schemes, which utilize semantic, linguistic information to guide the learning of visual features in different granularities. (3) Extensive ablation studies validate the effectiveness and complementarity of the two association schemes. Our method achieves state-of-the-art performance on person re-ID and outperforms conventional cross-modal embedding methods.

2   Related Work

Early works on person re-ID concentrated on either feature extraction [17, 37, 53] or metric learning [9–11, 22, 38]. Recent methods mainly benefit from the advances of CNN architectures [26], which combine the above two aspects to produce robust and ID-discriminative image representation [1, 8, 28, 46, 50, 52]. Our work aims to further improve the deep visual representation by making use of language descriptions as training supervisions.

Diverse auxiliary information has been introduced to improve the visual feature representations for person re-ID. Several works [47, 59, 61] detected person pose landmarks to obtain the human body regions. They firstly decomposed the feature maps according to the regions, then fused them to create the well-aligned feature maps. Lin *et al.* utilized Camera ID information to assist inter-image similarity estimation [32] by keeping consistencies in a camera network. Also, different types of sensors such as depth cameras [7], or infrared [54] cameras have been employed in person re-ID to generate more reliable visual representations. For these methods, the auxiliary information is used in both training and testing stage, requiring an additional model or data acquisition device for algorithm deployment. Differently, human attributes usually serve as a kind of training supervisions. For example, Lin *et al.* [33] improved the interpretability of the intermediate feature maps by jointly optimizing the identification loss and attribute classification loss. Although attributes proves helpful for feature learning, they are quite difficult to obtain as people need to remember tens of attribute labels for annotations. They are also less flexible to describe diverse variations in human appearance.

Associating image and language helps establish correspondences for their inter-relations. It has attracted great attention in recent years because of its wide applications in image captioning [13, 20, 35, 51, 57], visual QA [4, 19, 30], and text-image retrieval [18, 41]. These cross-modal associations can be modeled by either generative methods or discriminative methods. Generative models utilize probabilistic models to capture the temporal or spatial dependencies within
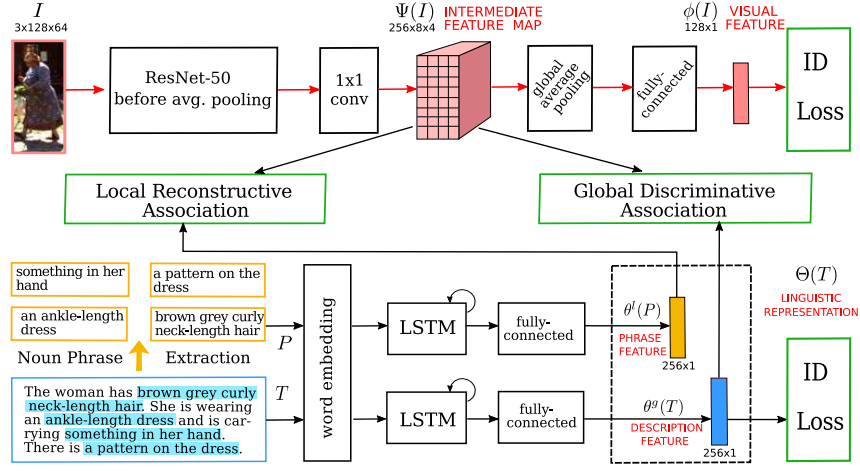
Fig. 2: Overall framework of our proposed approach. We employ the ResNet-50 as the backbone architecture. The produced intermediate feature $\Psi(I)$ is associated to the description feature $\theta^g(T)$ and the phrase feature $\theta^l(P)$ by global discriminative association and local reconstructive association, respectively.

the image or text [39, 51], and have popular applications like caption generation [3, 35, 43, 51, 57] and image generation [41, 42]. On the other hand, discriminative models have also been developed for image-text association. Karpathy and Fei-Fei [21] formulated a bidirectional ranking loss to associate the text and image fragments. Reed *et al.* [41] proposed deep symmetric structured joint embeddings, and enforced the embedding of matched image-text pair should be higher than those of unmatched pairs. Our method combines the merits of both discriminative and generative methods to build image-text association in different granularities, where the language descriptions act as training supervisions to improve visual representation.

3   Our Approach

We aim to exploit language descriptions of person images as the training supervisions in addition to the original ID labels for better visual representation. The visual representations are not only required to be discriminative for different persons but also need to keep consistencies with the linguistic representations. We, therefore, propose the global and local image-language association schemes. The global visual feature of one person should be more relevant to the language description features of the same person than those of a different person. Unlike existing cross-modal joint embedding methods, we do not require the visual and linguistic features to be mapped to a unified embedding space. Furthermore, based on the assumption that the image and language are spatially decomposable and temporally decomposable, we also try to find the mutual correspondences

between the features of the image regions and the noun-phrases. The overall framework is illustrated in Fig. 2.

3.1 Visual and Linguistic Representation

Given a dataset $\mathcal{D} = \{(I_n, T_n, l_n)\}_{n=1}^N$ containing N tuples, each tuple has an image I, a text description T, and an ID label l. To improve the learned visual feature $\phi(I)$, we build global and local correspondences between the intermediate visual feature maps $\Psi(I)$ and linguistic representation $\Theta(T)$.

The visual representation. The visual feature $\phi(I)$ and the intermediate feature map $\Psi(I)$ are obtained from standard convolutional neural network (CNN), which takes ResNet-50 as the backbone network. $\Psi(I)$ is the feature map obtained with the $1{\times}1$ convolution over the last residual-block. Suppose the $\Psi(I)$ has K bins, the feature vector at the kth bin is denoted by $\psi_k(I)$, then $\Psi(I)$ can be represented as $\Psi(I) = \{\psi_k(I)\}_{k=1}^K$. The objective visual feature vector $\phi(I)$ is linear projection from the average feature map $\bar{\psi}(I) = \frac{1}{K}\sum_{k=1}^K \psi_k(I)$:

$$\phi(I) = f_\phi(\Psi(I)) = \mathbf{W}_\phi \bar{\psi}(I) + \mathbf{b}_\phi. \tag{1}$$

We employ the ID loss over $\phi(I)$. Specifically, given N images belonging to I persons, the ID loss is the average negative log-likelihood of the feature maps being correctly classified to its ID:

$$\mathcal{L}_I = -\frac{1}{N}\sum_{n=1}^N\sum_{i=1}^I y_{i,n} \log\left(\frac{\exp(\mathbf{w}_i^\top \phi(I_n))}{\sum_{j=1}^I \exp(\mathbf{w}_j^\top \phi(I_n))}\right), \tag{2}$$

where $y_{i,n}$ is the index label with $y_{i,n} = 1$ if the nth image I_n belongs to the ith person and $y_{i,n} = 0$ otherwise. $\mathbf{w}_i$ are the classifier parameters associated with the ith person over the visual feature vectors.

The linguistic representation. $\Theta(T)$ contains two types of feature vectors as shown in Fig. 2. One is the global description feature $\theta^g(T)$ mapped from the whole text, the other is the local phrase feature $\theta^l(P)$ that encodes a distinctive noun phrase P cropped from the text T. The noun-phrase extraction procedure is demonstrated in Fig. 3, and the obtained phrases in T form the set $\mathcal{P}(T)$. Each word in text T or phrase P is firstly represented as a D dimensional one-hot vector, denoted by $\mathbf{o}_m \in \mathbb{R}^D$ for the mth word and D is the vocabulary size. Then the one-hot vector is projected to a word embedding: $\mathbf{e}_m = \mathbf{W}_e \mathbf{o}_m$.

Based on the embedding, we feed either a whole description or a short phrase to a long short-term memory network (LSTM) word by word, which has the following updating procedure: $\mathbf{h}_{m+1} = \text{LSTM}(\mathbf{e}_m, \mathbf{h}_m)$. The LSTM unit takes the current word embedding $\mathbf{e}_m$ and hidden state $\mathbf{h}_m$ as inputs, and outputs the hidden state of the next step $\mathbf{e}_{m+1}$. The hidden states at the final time step are effective summarization of the description T or phrase P, obtaining the description feature $\theta^g(T) = \mathbf{W}_g \mathbf{h}_F(T) + \mathbf{b}_g$ or the phrase feature $\theta^l(P) = \mathbf{W}_l \mathbf{h}_F(P) + \mathbf{b}_l$, where $\mathbf{h}_F(T)$ and $\mathbf{h}_F(P)$ are the final hidden states for text T and phrase P, respectively. Because T describes abundant person characteristics
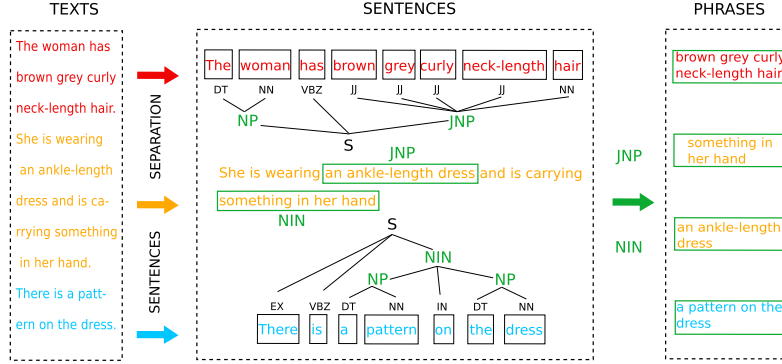
Fig. 3: The flowchart of extracting interested noun phrases from the text. We perform word-level tokenization and part-of-speech tagging, then extract noun phrases by chunking. As not all the phrases can have discriminative information, we are interested in two kinds of the phrases: (1) the noun phrase with adjectives (JJ), defined as JNP (2) the noun phrase consists of multiple nouns joined by preposition (IN).

throughout the body, $\theta^g(T)$ could describe a specific person. We therefore impose another ID loss to make $\theta^g(T)$ be separable for different persons,

$$\mathcal{L}_T = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} y_{i,n} \log \left(\frac{\exp(\mathbf{v}_i^\top \theta^g(T_n))}{\sum_{j=1}^{I} \exp(\mathbf{v}_j^\top \theta^g(T_n))} \right), \tag{3}$$

where $y_{i,n}$ is label defined similar to the one in Eqn. (2), and $\mathbf{v}_i$ indicates the classifier parameters associated with the ith person over the text feature.

3.2   Global Discriminative Image-language Association

The ID losses in the previous section only enforce the visual and linguistic feature be discriminative within each modality but do not establish image-language correspondences to enhance the visual feature. As the global description is usually related to multiple and diverse regions in the image, $\theta^g(T)$ can be associated to $\bar{\psi}(I)$ (Eqn. (1)) in a discriminative fashion. Specifically, $\bar{\psi}(I)$ and $\theta^g(T)$ firstly form a joint representation $\varphi(I,T)$: $\varphi(I,T) = \left(\bar{\psi}(I) - \theta^g(T) \right) \circ \left(\bar{\psi}(I) - \theta^g(T) \right)$, where $\circ$ denotes the Hadamard product. The joint representation is then projected into a scalar value within the range $(0,1)$ by:

$$s(I,T) = \frac{\exp(\mathbf{w}_s^\top \varphi(I,T) + b_s)}{1 + \exp(\mathbf{w}_s^\top \varphi(I,T) + b_s)}. \tag{4}$$

To build the relevance between $\bar{\psi}(I)$ and $\theta^g(T)$, we expect $s(I,T)$ to be 1 when I and T belong to the same person and to be 0 when they belong to different persons. We thus impose the binary cross-entropy loss over the scores:

$$\mathcal{L}_{dis} = -\frac{1}{\hat{N}} \sum_{i,j} \left[l_{i,j} \log \left(s(I_i, T_j) \right) + (1 - l_{i,j}) \log \left(1 - s(I_i, T_j) \right) \right], \tag{5}$$

where $\hat{N}$ is the number of sampled image-text pairs. $l_{i,j} = 1$ if I_i and T_j are describing a same person and $l_{i,j} = 0$ otherwise.

Discussion. Here, we draw a distinction between the proposed discriminative scheme and the bi-directional ranking [21, 41, 63], which is formulated by:

$$\mathcal{L}_{rank} = \frac{1}{\hat{N}} \sum_{i,j} \max(0, k_{i,j} - k_{i,i} + \alpha) + \max(0, k_{j,i} - k_{i,i} + \alpha), \qquad (6)$$

where $k_{i,j} = \bar{\psi}(I_i)^\top \theta^g(T_j)$. The loss stipulates that the cosine similarity $k_{i,i}$ for one image-text tuple should be higher than $k_{i,j}$ or $k_{j,i}$ for any $i \neq j$ by at least a margin of α. We highlight two main differences between the proposed $\mathcal{L}_{dis}$(Eqn. (5)) and $\mathcal{L}_{rank}$: (1) As $\mathcal{L}_{rank}$ is originally applied in the image-text retrieval task, it associates the image and text features by simply checking whether they are from the same tuple. Differently, $\mathcal{L}_{dis}$ is based on person ID, which is more reasonable as one description can well correspond to different images of the same person. (2) $\mathcal{L}_{rank}$ estimates the image-text relevance by cosine similarity, requiring $\bar{\psi}(I_i)$ and $\theta^g(T_j)$ lie in the same feature space. While the proposed method doesn't have such restriction.

3.3   Local Reconstructive Image-language Association

A phrase usually only describes one part of an image and could be contained in the descriptions of different persons. For this reason, a phrase is disjoint with the person ID, but can still build correspondences with a certain region in the image it describes. We therefore propose a reconstruction scheme. That is, the phrase feature $\theta^l(P)$ can select relevant feature vectors in visual feature map $\Psi(I_n)$ if $P \in \mathcal{P}(T_n)$, and the selected feature vectors are able to reconstruct the phrase P in turn.

Image feature aggregation. Suppose P is a phrase that describes a specific region in image I_n, we aim to estimate a vector $\hat{\psi}_P(I_n)$ that can reflect the features in the region. For this purpose, we compute $\hat{\psi}_P(I_n)$ by weighted aggregation of the feature vectors $\{\psi_k(I_n)\}_{k=1}^K$ in the feature map $\Psi(I_n)$: $\hat{\psi}_P(I_n) = \sum_{k=1}^K r_k(P, I_n)\psi_k(I_n)$, where $r_k(P, I_n)$ is the attention weight reflecting the relevance between the phrase P and the feature vector $\psi_k(I_n)$. It is estimated by an attention function $f_{att}(\psi_k(I_n), \theta^l(P))$, which first computes the the unnormalized weight $\bar{r}_k(P, I_n)$ with a linear projection over the joint representation of $\psi_k(I_n)$ and $r_k(P, I_n)$:$\bar{r}_k(P, I_n) = \mathbf{w}_{\bar{r}}^\top \big((\psi_k(I_n) - \theta^l(P)) \circ (\psi_k(I_n) - \theta^l(P)) \big) + b_{\bar{r}}$, then normalizes the values by using a softmax operation over all the K bins: $r_k(P, I_n) = \exp(\bar{r}_k(P, I_n))/\sum_{k=1}^K \exp(\bar{r}_k(P, I_n))$. In practice, the attention model is easy to overfit with limited training data. Besides, the spatially adjacent feature maps possibly represent one phrase, they are more reasonable to be merged. For these reasons, we reduce the training burden by average pooling the neighboring feature maps in $\Psi(I_n)$ before the weighted aggregation, which is also illustrated in Fig. 4.

Phrase reconstruction. To enforce the consistency between the aggregated feature map $\hat{\psi}_P(I_n)$ and the input phrase P, we build the conditional probability $p(P|\hat{\psi}_P(I_n))$ to reconstruct P with $\hat{\psi}_P(I_n)$. Since a phrase has a unbounded
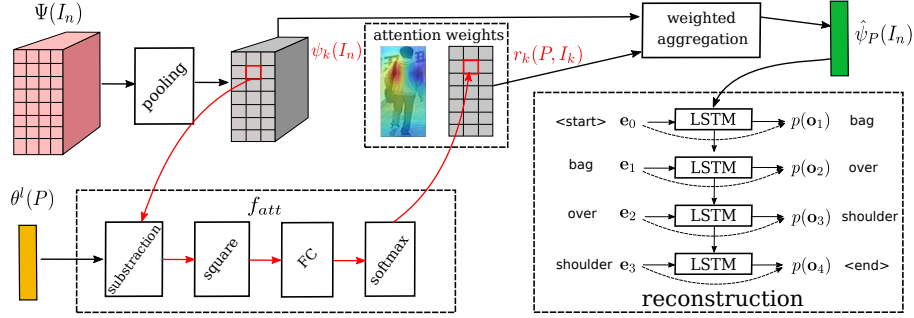
Fig. 4: The network structure for the local reconstructive image-language association. We first use the feature maps $\Psi(I_n)$ and the phrase feature $\theta^l(P)$ to compute the attention weights for the intermediate features at different locations, then perform weighted aggregation to obtain the visual feature $\hat{\psi}_P(I_n)$, and finally employ LSTM to reconstruct P with $\hat{\psi}_P(I_n)$.

length M, it is common to apply the chain rule to model the probability over $\{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_{M+1}\}$: $\log p(P|\hat{\psi}_P(I_n)) = \sum_{m=0}^{M} \log p(\mathbf{o}_{m+1}|\hat{\psi}_P(I_n), \hat{\mathbf{o}}_0, ..., \hat{\mathbf{o}}_m)$. More specifically, $\mathbf{o}_{m+1}(m = 0, ..., M)$ is the random variable over the one-hot vectors of the m-*th* word, and $\{\hat{\mathbf{o}}_0, ..., \hat{\mathbf{o}}_{M+1}\}$ are one-hot vectors of the ground truth words. Among them, $\hat{\mathbf{o}}_0, \hat{\mathbf{o}}_{M+1}$ are the one-hot vectors that designate the start and end of the phrase. Inspired by the task of image caption generation [51, 57], LSTM is employed to model $p(\mathbf{o}_{m+1}|\hat{\psi}_P(I_n), \hat{\mathbf{o}}_0, ..., \hat{\mathbf{o}}_m)$. More specifically, we initially feed $\hat{\psi}_P(I_n)$ to the LSTM, then feed the embedding of the current word to obtain the hidden state of the next word. The next word probability is computed from the hidden state $\mathbf{h}_{m+1}$ and the word embedding $\mathbf{e}_m$. The word probability can be formulated as: $p(\mathbf{o}_{m+1}|\hat{\psi}_P(I_n), \hat{\mathbf{o}}_0, ..., \hat{\mathbf{o}}_m) \propto \exp(\mathbf{W}_{oh}\mathbf{h}_{m+1} + \mathbf{W}_{oe}\mathbf{e}_m)$. The reconstruction loss is the sum of the negative log likelihood of the correct word at each step:

$$\mathcal{L}_{rec} = -\frac{1}{N} \sum_{n=1}^{N} \frac{1}{|\mathcal{P}(T_n)|} \sum_{P \in \mathcal{P}(T_n)} \log p(P|\hat{\psi}_P(I_n)). \qquad (7)$$

3.4   Training and Testing

The final loss function is a combination of the image ID loss, the text ID loss as well as the discriminative and reconstructive image-language association losses:

$$\mathcal{L} = \mathcal{L}_I + \lambda_T \mathcal{L}_T + \lambda_{dis} \mathcal{L}_{dis} + \lambda_{rec} \mathcal{L}_{rec}, \qquad (8)$$

where λ_T, λ_{dis} and λ_{rec} are balancing parameters. For network training, we adopt stochastic gradient descent (SGD) with an initial learning rate of 10^{-2}, which is further decayed to 10^{-3} after the 20th epoch. We organize the training batch as follows. The data tuple (I_n, T_n, d_n) is firstly transformed to $(I_n, T_n, \mathcal{P}(T_n), d_n)$.

Each batch contains the samples from 32 randomly selected persons, and each person has two randomly sampled tuples. For global discrimination, we form 32×4 positive image-description pairs by exploiting all the intra-tuple and inter-tuple image-description compositions, and sample 6 negative pairs for each image, yielding 64×6 negative pairs, keeping the pos/neg ratio to be 1:3. Meanwhile, the local reconstruction is performed within each tuple.

In testing, only image features are extracted, and no language descriptions are used. The distance between two image features are simply the Euclidean distance, $i.e.$, $d_{i,j} = \|\phi(I_i) - \phi(I_j)\|_2$. Person Re-ID is performed by ranking the distances between the probe image and gallery images in ascending order.

4   Experiments

We evaluate the proposed approach on three standard person re-ID datasets, whose language annotations can be fully or partially obtained from the CUHK-PEDES dataset [26]. Ablation studies are mainly conducted on Market-1501 [62] and CUHK-SYSU [56], which are convenient for extensive evaluation as with fixed training/testing splits. We also report the overall results on Market-1501, CUHK03 [28] and CUHK01 [27] to compare with the state-of-the-art approaches.

4.1   Experimental Setup

Datasets and Metrics.   To verify the utility of language descriptions in person re-ID, we augment four standard person re-ID datasets (Market-1501, CUHK03, CUHK01, and CUHK-SYSU) with language descriptions. The language descriptions are obtained from the CUHK-PEDES dataset, which is originally developed for cross-modal text-based person search and contains 40,206 images of 13,003 persons from five existing person re-ID datasets. Since persons in Market-1501 and CUHK03 have many similar samples, only four images of each person in this two datasets have language descriptions.

Among the four datasets, Market-1501, CUHK03 and CUHK01 follow the standard training and testing partitions. CUHK-SYSU is a new dataset used for joint detection and identification. According to separation in CUHK-PEDES, 15,080 images from 5,532 identities are used for training, 8,341 images from 2,900 persons are used for testing with 2,900 query images and 5,441 gallery images. Mean average precision (mAP) and CMC top-1, top-5, top-10 accuracies are adopted as the evaluation metrics.

Implementation details. All the person images are resized to 256×128. For data augmentation, random horizontal flipping and random cropping are adopted. We empirically set the dimensions of feature embeddings $\phi(I)$, $\theta^l(P)$ and $\theta^g(T)$ to be 256, and set the balancing parameters $\lambda_T = 0.1$, $\lambda_{dis} = 1$, $\lambda_{rec} = 1$, respectively. As some images in Market-1501 and CUHK03 do not have language descriptions, we employ the description of the same person (in the same camera if possible) for them to compose the data tuple (I_n, T_n, d_n). The ResNet-50 backbone is initialized by the parameters pre-trained on ImageNet [16].

Methods	Training Loss				
	$\mathcal{L}_I$	$\mathcal{L}_T$	$\mathcal{L}_{dis}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{rank}$
basel.	✓	✗	✗	✗	✗
basel. + rank	✓	✓	✗	✗	✓
basel. + GDA	✓	✓	✓	✗	✗
basel. + LRA	✓	✗	✗	✓	✗
proposed	✓	✓	✓	✓	✗

Table 1: The loss configurations for the baseline and other variants.

Methods	Market-1501				CUHK-SYSU			
	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
basel.	74.4	89.2	95.5	96.9	85.8	87.3	93.7	95.1
basel.+ rank[1] [21]	75.5	88.5	95.9	97.5	87.0	88.0	94.5	95.9
basel.+ rank[2] [63]	77.7	90.5	96.1	97.6	88.8	90.2	95.5	96.8
basel. + GDA	80.0	91.5	96.4	98.0	90.2	91.0	96.2	97.5
basel. + LRA	79.6	91.6	96.7	97.9	89.7	90.7	96.0	97.4
proposed	**81.8**	**93.3**	**97.4**	**98.5**	**91.4**	**92.0**	**96.7**	**97.9**

Table 2: Comparison of different association schemes upon our baseline method. Top-1.-5,-10 accuracies (%) and mAP(%) are reported.

Baseline and variants. The baseline is just the visual CNN that produces the feature map $\phi(I)$, indicated by the red lines in Fig. 2. We additionally build 4 variants on the baseline for ablation study. The loss configuration of them are displayed in Table. 1. Among them, *basel.* only imposes the ID loss to make $\phi(I)$ be separable for different persons. Both *basel.+rank* and *basel.+GDA* additionally impose the ID loss over the global description feature $\theta^g(T)$ but have different global image-language association schemes. *basel.+rank* employs the $\mathcal{L}_{rank}$ in Eqn.(6), while *basel.+GDA* utilizes the proposed $\mathcal{L}_{dis}$ in Eqn.(5). The variant *basel.+LRA* employs the reconstruction loss $\mathcal{L}_{rec}$ in Eqn.(7) to build the local association between the aggregated feature vector $\hat{\psi}_P(I_n)$ and the phrase feature $\theta^l(P)$. Our proposed method takes advantages of both global and local image-language association schemes.

4.2   The Effect of Global Discriminative Association (GDA)

Comparison with non-discriminative variants. We evaluate the effects of global discriminative image-language association by comparing the variants with and without using the description feature $\theta^g(T)$. Among them, *basel.+GDA* improves *basel.* by 5.6% and 4.4% in term of mAP on Market-1501 and CUHK-SYSU respectively (Table. 2), which shows that GDA can benefit the learning of visual representation. Furthermore, our proposed method yields better performance than *basel.+LRA*, indicating the effect of global discriminative association is complementary to that of the local reconstructive association.

Comparison with bi-directional ranking loss [21, 63]. $\mathcal{L}_{dis}$ in GDA aims to discriminate the matched image-text pairs from the unmatched ones. It has the similar functions with the bidirectional ranking loss $\mathcal{L}_{rank}$ (Eqn. (6)) for image-language cross-modal retrieval. We implement two types of ranking losses for comparison. The first one is more similar to the loss in [21], where a positive image-text pair is composed of the image and text from the same tuple. The other one adopts the loss in [63], where the positive image-text pairs are obtained by arbitrary image-text combinations from the same person. We modify $basel.+GDA$ by replacing $\mathcal{L}_{dis}$ with the two loss functions, and denote them by $basel.+rank^1$ and $basel.+rank^2$, respectively. The results in Table 2 show that both ranking losses can boost the baseline. Besides, $basel.+rank^2$ is better than $basel.+rank^1$ by incorporating more abundant positive samples for discrimination. The proposed $basel.+GDA$ further improves the mAP by 2.3% and 1.4% on Market-1501 and CUHK-SYSU, verifying the effectiveness of our relevance estimation strategy (Eqn. 4).

The importance of $\mathcal{L}_T$. To preserve separability of the visual feature, the associated linguistic feature $\theta^g(T)$ is supposed to be discriminative for different persons, thus $\mathcal{L}_T$ is employed along with $\mathcal{L}_{dis}$. We investigate the importance of $\mathcal{L}_T$ based upon $basel.+GDA$ and observe how the performance changes with λ_T in Table 3. Slightly worse results are observed when $\lambda_T = 0$, indicating $\mathcal{L}_T$ is indispensable. On the other hand, the optimal results are achieved when λ_T is around 0.1. One possible reason is that language description is sometimes more ambiguous to describe a specific person, making $\mathcal{L}_I$ and $\mathcal{L}_T$ not equally important. For example, "The man wears a blue shirt" can simultaneously describe different persons wearing a dark blue shirt and a light blue shirt.

4.3   The Effect of Local Reconstructive Association (LRA)

Comparison with non-reconstructive variants. We evaluate the effects of local reconstructive association by comparing the variants with and without using the local phrase feature $\theta^l(P)$. The performance gap between $basel.$ and $basel.+LRA$ proves the effectiveness of LRA for visual feature learning. Employing LRA brings 5.2% and 3.9% mAP gain over the two datasets, which is close

$basel.+GDA$	Market-1501				CUHK-SYSU			
λ_T	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
0	78.9	91.1	96.4	97.6	89.3	90.3	95.7	97.0
0.05	79.2	91.2	96.5	97.8	89.2	90.2	95.6	96.9
0.1	**80.0**	**91.5**	96.4	**98.0**	**90.2**	**91.0**	**96.3**	**97.5**
0.5	79.6	91.3	96.6	97.9	89.8	90.9	96.2	97.3
1	78.9	91.0	96.6	97.9	89.1	90.0	95.8	97.1

Table 3: Importance analysis of $\mathcal{L}_T$ in $basel.+GDA$. We fix $\lambda_{dis} = 1$ and adjust λ_T over 0, 0.05, 0.1, 0.5, 1. Top-1,-5,-10 accuracies (%) and mAP(%) are reported.

Methods	top-1	top-5	top-10	Methods	top-1	top-5	top-10
GNA-RNN [26]	19.05	- -	53.64	DPCE [63]	**44.40**	66.26	75.07
IATV [25]	25.94	- -	60.49	Ours	43.58	**66.93**	**76.26**

Table 4: Results on CUHK-PEDES.

to the gain of employing GDA. Besides, the fact that the proposed method is better than *basel.+GDA* also indicates the effectiveness of LRA.

Visualization of phrase-guided attention weights. We compute the attention weights for a specific phrase (Eqn. (4)), align the weights to the corresponding image, and obtain the heat map for the phrase. The heat maps are displayed in Fig. 5, showing that the attention weights can roughly capture the local regions described by the phrases.

4.4   Results on Text-to-image Retrieval

As a by-product, our method can also be utilized for text-to-image retrieval, which is fulfilled by ranking the cross-modal relevance (Eqn. (4)). We report the retrieval results on CUHK-PEDES following the standard protocol, where there are 3,074 test images with 6,156 captions, 3,078 validation images with 6,158 captions, and 34,054 training images with 68,126 captions. The quantitative and qualitative results are reported in Table 4 and Fig. 6, respectively. Although our method is not specifically designed for this task, it achieves competitive results to the current state-of-the-art methods.

4.5   Comparison with the State-of-the-Art Approaches

We compare our method with the current state-of-the-arts on the Market1501, CUHK03, and CUHK01 datasets. The results on Market-1501 are reported in Table 6 **left**. Our method outperforms all the other approaches regarding mAP and top-1 accuracy under both single-query and multi-query protocols. Note that the baseline of our method is quite competitive to the most of the previous methods, which is partly because of well initialized ResNet-50 backbone and proper data augmentation strategies. The proposed image-language association scheme can largely boost the well-performed baseline, making our method better than the recent state-of-the-arts [2, 6].

Methods	CUHK01		
	top-1	top-5	top-10
[31] XQDA(CVPR15)	63.2	83.9	90.0
[55] JSTL(CVPR16)	66.6	-	-
[58] DNS(CVPR16)	69.1	86.9	91.8
[12] Quad(CVPR17)	62.6	83.4	89.7
[14] CRAFT(PAMI17)	74.5	91.2	94.8
[59] Spindle(CVPR17)	79.9	94.4	97.1
[60] DLPAR (ICCV17)	75.0	93.5	95.5
basel.	77.0	93.2	95.3
proposed	**84.8**	**95.1**	**98.4**

Table 5: Results on CUHK01. Top-1,-5,-10 accuracies(%) are reported.

CUHK03 has two types of person bounding boxes: one is manually labelled, and the other is obtained by a pedestrian detector. We compare our methods and others on both types, and report the top-1 and top-5 accuracies in Table 6 **right**. It can be seen that our method
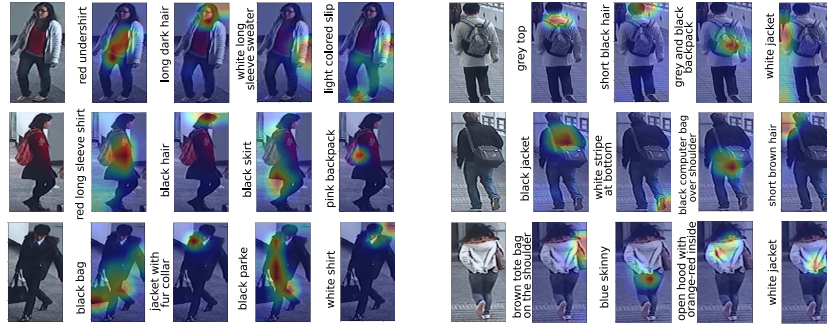
Fig. 5: Heat maps of the attention weights. The phrases are placed in on the left of the corresponding heat maps. Zoom in the figure for better view of the phrases.

A man in a blue shirt a pair of black shorts and a pair of glasses on his face.

A woman in a white shirt a pair of blue jean pants and a pair of gray shoes

He is wearing a light coloured hooded jacket and light brown pants, carrying something in one hand.

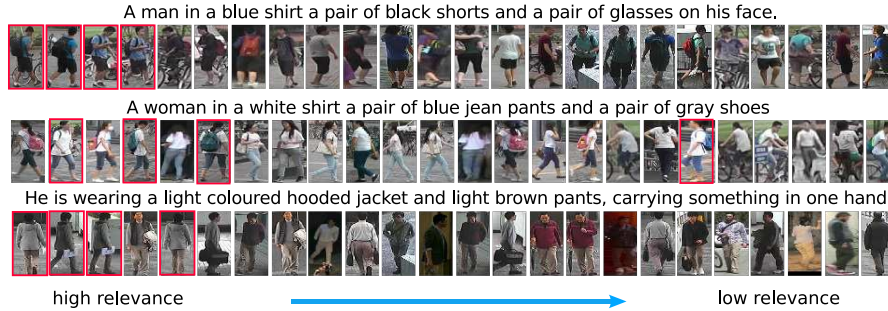high relevance   low relevance

Fig. 6: Examples of the text-to-image search. The most relevant 24 images are displayed. Red boxes indicate the ground truth.

has significant advantages over the top-1 accuracy, but is 0.2% less than D-person [6] on the top-5 accuracy for the labelled bounding boxes. As D-person only utilizes image data, it is promising to apply our language association scheme to D-person for better performance. Compared with Market-1501 and CUHK03, CUHK01 has fewer images for training as described in Sec. 5. As in Table 5, the proposed association schemes have 7.8% top-1 accuracy gain over the baseline on CUHK01. The results confirm the effectiveness of language description, and indicate the schemes may be more useful when the image data are not enough.

Among the compared approaches, Spindle [59] and PDC [47] utilize pose landmarks, CADL [32] employs the camera ID labels, and ACN [44] makes use of the attributes for training. We achieve better results than them on all the three datasets (Tables 5 and 6). The results indicate language description is also a kind of useful auxiliary information for person re-ID. With the proposed schemes, it can achieve the superior performance with the standard CNN architecture.

Methods	Market-1501				CUHK03			
	Single Query		Multi-Query		Labeled		Detected	
	mAP	top-1	mAP	top-1	top-1	top-5	top-1	top-5
[67] P2S(CVPR17)	44.3	70.7	55.7	85.8	-	-	-	-
[32] CADL(CVPR17)	47.1	73.8	55.6	80.9	-	-	-	-
[24] MSCAN(CVPR17)	57.5	80.3	66.7	86.8	74.2	94.3	68.0	91.0
[5] SSM(CVPR17)	68.8	82.2	76.2	88.2	76.6	94.6	72.7	92.4
[65] k-rank(CVPR17)	63.4	77.1	-	-	61.6	-	58.5	-
[44] ACN(CVPRW17)	62.6	83.6	-	-	-	-	62.6	89.7
[49] SVDNet(ICCV17)	62.1	82.3	-	-	-	-	81.8	95.2
[60] DLPAR(ICCV17)	63.4	81.0	-	-	85.4	97.6	81.6	97.3
[66] OLMANS(ICCV17)	60.7	-	66.8	-	61.7	88.4	62.7	87.6
[40] MuDeep(ICCV17)	-	-	-	-	76.9	76.3	75.6	94.4
[47] PDC(ICCV17)	63.4	84.1	-	-	88.7	98.6	78.3	94.8
[64] VI+LSRO(ICCV17)	66.1	84.0	76.1	88.4	-	-	84.6	97.6
[15] DPFL(ICCVW17)	73.1	88.9	80.7	92.3	86.7	-	82.0	-
[29] JLMT(IJCAI17)	65.5	85.1	74.5	89.7	83.2	98.0	80.6	96.9
[6] D-Person(Arxiv17)	79.6	92.3	94.5	85.1	91.5	**99.0**	89.4	**98.2**
[2] TGP(Arxiv18)	81.2	92.2	87.3	94.7	-	-	-	-
basel.	74.4	89.2	82.3	93.3	88.4	98.1	87.9	97.5
proposed	**81.8**	**93.3**	**87.9**	**95.3**	**92.5**	98.8	**90.9**	**98.2**

Table 6: Comparison with the state-of-the-art methods on the Market-1501 and CUHK03 datasets. The results on Market-1501 are under single-query and multi-query protocols. MAP (%) and top-1 accuracy (%) are reported. Meanwhile, the performances on CUHK03 are evaluated with labeled and detected bounding boxes. Top-1 and Top-5 accuracies(%) are reported.

5 Conclusions

We utilized language descriptions as additional training supervisions to improve the visual features for person re-identification. The global and local image-language association schemes have been proposed. The former learns better global visual features with the discriminative supervision of the overall language descriptions, while the latter enforces the semantic consistencies between local visual features and noun phrases by phrase reconstruction. Our ablation studies show that the proposed image-language association schemes can remarkably improve the learning of the visual feature and are more effective than the existing image-text joint embedding methods. The proposed method achieves state-of-the-art performance on three public person re-ID datasets.

Acknowledgements

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
2. Almazan, J., Gajic, B., Murray, N., Larlus, D.: Re-id done right: towards good practices for person re-identification. arXiv preprint arXiv:1801.05339 (2018)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. arXiv preprint arXiv:1707.07998 (2017)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015)
5. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: CVPR (2017)
6. Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., Xu, Y.: Deep-person: Learning discriminative deep features for person re-identification. CoRR **abs/1711.10658** (2017)
7. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: ECCV (2012)
8. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep crf for person re-identification. In: CVPR (2018)
9. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: CVPR (2016)
10. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: CVPR (2015)
11. Chen, D., Yuan, Z., Wang, J., Chen, B., Hua, G., Zheng, N.: Exemplar-guided similarity learning on polynomial kernel feature map for person re-identification. International Journal of Computer Vision **123**(3), 392–414 (2017)
12. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: A deep quadruplet network for person re-identification. In: CVPR (2017)
13. Chen, X., Zitnick, C.L.: Mind's eye: A recurrent visual representation for image caption generation. In: CVPR (June 2015)
14. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 392–408 (Feb 2018)
15. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: ICCVW (2017)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
17. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
18. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: NIPS. pp. 2121–2129 (2013)
19. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: CVPR (2016)
20. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4) (Apr 2017)
21. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: NIPS (2014)
22. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR. pp. 2288–2295 (2012)

23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS (2012)
24. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR (2017)
25. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: ICCV (2017)
26. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: CVPR (2017)
27. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: ACCV (2012)
28. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
29. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. IJCAI (2017)
30. Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., Zhou, M.: Visual question generation as dual task of visual question answering. In: CVPR (2018)
31. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
32. Lin, J., Ren, L., Lu, J., Feng, J., Zhou, J.: Consistent-aware deep learning for person re-identification in a camera network. In: CVPR (2017)
33. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. CoRR **abs/1703.07220** (2017)
34. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: ECCV (2016)
35. Liu, X., Li, H., Shao, J., Chen, D., Wang, X.: Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In: ECCV (2018)
36. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV (2017)
37. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: British Machive Vision Conference. pp. 11–pages (2012)
38. Mignon, A., Jurie, F.: Pcca: A new approach for distance learning from sparse pairwise constraints. In: CVPR. IEEE (2012)
39. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
40. Qian, X., Fu, Y., Jiang, Y.G., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. In: ICCV (2017)
41. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR (2016)
42. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
43. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
44. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: CVPRW (2017)
45. Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., Wang, X.: Deep group-shuffling random walk for person re-identification. In: CVPR (2018)
46. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: End-to-end deep kronecker-product matching for person re-identification. In: CVPR (2018)

47. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)
48. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: ECCV (2016)
49. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV (2017)
50. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: ECCV (2016)
51. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. IEEE (2015)
52. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: CVPR (2016)
53. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: ICCV (2007)
54. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: ICCV (2017)
55. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR (2016)
56. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. CoRR **abs/1604.01850** (2016)
57. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)
58. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR (2016)
59. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR (2017)
60. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)
61. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. CoRR **abs/1701.07732** (2017)
62. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
63. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Shen, Y.: Dual-path convolutional image-text embedding. CoRR **abs/1711.05535** (2017), `http://arxiv.org/abs/1711.05535`
64. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017)
65. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)
66. Zhou, J., Yu, P., Tang, W., Wu, Y.: Efficient online local metric adaptation via negative samples for person re-identification. In: ICCV (2017)
67. Zhou, S., Wang, J., Wang, J., Gong, Y., Zheng, N.: Point to set similarity based deep feature learning for person re-identification. In: CVPR (2017)