Pose-Guided Joint Global and Attentive Local Matching Network for Text-Based Person Search

Ya Jing[1,3], Chenyang Si[1,3], Junbo Wang[1,3], Wei Wang[1,3,], Liang Wang[1,2,3], Tieniu Tan[1,2,3]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
{ya.jing, chenyang.si, junbo.wang}@cripac.ia.ac.cn, {wangwei, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Text-based person search aims to retrieve the corresponding persons in an image database by virtue of a describing sentence about the person, which poses great potential for various applications such as video surveillance. Extracting visual contents corresponding to the human description is the key to this cross-modal matching problem. Moreover, correlated images and descriptions involve different levels of semantic relevance. To exploit the multilevel relevances between human description and corresponding visual contents, we propose a pose-guided joint global and attentive local matching network (GALM), which includes global, uni-local and bi-local matching. The global matching network aims to learn global cross-modal representations. To further capture the meaningful local relations, we propose an uni-local matching network to compute the local similarities between image regions and textual description and then utilize a similarity-based hard attention to select the description-related image regions. In addition to sentence-level matching, the fine-grained phrase-level matching is captured by the bi-local matching network, which employs pose information to learn latent semantic alignment between visual body part and textual noun phrase. To verify the effectiveness of our model, we perform extensive experiments on the CUHK Person Description Dataset (CUHK-PEDES) which is currently the only available dataset for text-based person search. Experimental results show that our approach outperforms the state-of-the-art methods by 15 % in terms of top-1 metric.

1. Introduction

Person search has gained great attention in recent years due to its wide applications in video surveillance, e.g., missing persons searching and suspects tracking. With the explosive increase in the number of videos, manual person search in large-scale videos is unrealistic, so we need to design automatic methods to perform this task more efficiently. Existing methods of person search are mainly classified into three categories according to the query type, e.g., image-based query [32, 30, 20], attribute-based query [22, 24] and text-based query [16, 15, 31, 6]. Image-based person search needs at least one image of the queried person which in many cases is very difficult to obtain. Since textual descriptions are more accessible, text-based person search can solve person image missing problem. Compared to attribute-based person search, text-based methods describe persons of interest with more details in a more natural way. Considering the advantages above, we study the task of text-

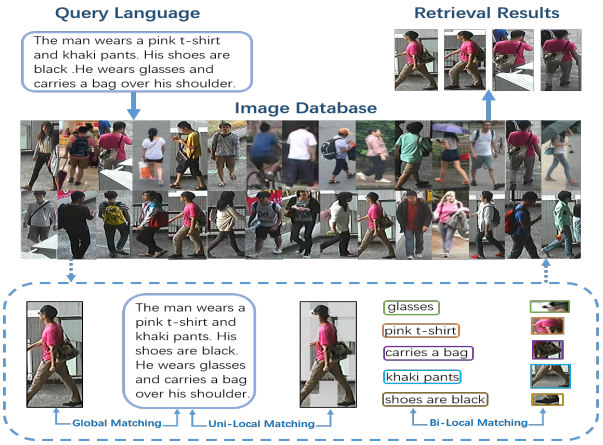

Figure 1. The illustration of text-based person search. Given a textual description of a person, the model aims to retrieve the corresponding persons from the given image database. There are three level matchings in our retrieval procedure: global matching, uni-local matching and bi-local matching.

based person search in this paper.

Text-based person search aims to retrieve the corresponding person images to a textual description from a large-scale image database, which is illustrated in Fig. 1. The main challenge of this task is to effectively extract corresponding visual contents to the human description under multilevel semantic relevances between image and text. Previous methods [31, 27] generally utilize Convolutional Neural Networks (CNNs) [14] to obtain a global representation of the input image which often cannot effectively extract visual contents corresponding to the person in the image. Considering that human pose is closely related to human body parts, we exploit pose information for effective visual feature extraction. To our knowledge, we are probably the first to employ human pose to handle the task of text-based person search.

Nonetheless, matching only the global features is coarse due to the fact that only partial image regions are corresponding to the given textual description. The learned global visual representation suffers from much irrelevant information brought by useless visual regions, e.g., meaningless background visual regions. Therefore, we expect to select description-related image regions for fine matching. There are some prior approaches [6, 5] which utilize local features for accurate matching. Different from them, we propose to compute the local similarities between image regions and the text. To further select the most related visual regions, we perform a hard attention over the local similarities.

In addition to sentence-level correlations, phrase-level relations are also significant for fine image-text matching. In fact, the noun phrase in textual description usually is related to the specific visual human part but not the whole image as shown in Fig. 1. To learn the latent semantic alignment between them and make our matching procedure more accurate and interpretable, we aim to learn the correspondence between noun phrase and visual human part. Considering that human pose is closely related to human body parts, we propose to utilize pose information to guide the aligned part matching.

In this paper, we propose a pose-guided joint global and attentive local matching network (GALM) for text-based person search, which includes global, uni-local and bi-local matching as shown in Fig. 2. First, we estimate human pose from the input image and perform global matching between global visual and textual representations, where global visual representation emphasizes the human-related features by concatenating the pose confidence maps to the original input image. To further capture the meaningful local relations, the uni-local matching is performed between textual description and image regions, where a similarity-based hard attention is utilized to select the most description-related image regions. In addition to sentence-level relations, phrase-level relations are also exploited by a bi-local matching network which uses the pose information to guide the attention of noun phrase and image regions and performs an aligned part matching. Both ranking loss and identification loss are employed for better training. Our proposed method is evaluated on a challenging dataset CUHK-PEDES [16], which is currently the only available dataset for text-based person search. Experimental results show that our GALM model outperforms the state-of-the-art methods on this dataset.

The main contributions of this paper can be summarized as fourfold. (1) A pose-guided joint global and attentive local matching network is proposed to learn multilevel cross-modal relevances. (2) A similarity-based hard attention is proposed in uni-local matching to effectively select the most related image regions to the textual description. (3) A novel pose-guided part aligned matching network is proposed to exploit the latent semantic alignment between visual body part and textual noun phrase, which is probably the first used in text-based person search. (4) The proposed GALM achieves the best results on the challenging dataset CUHK-PEDES. The extensive ablation studies verify the effectiveness of each component in the GALM.

2. Related Work

In this section, we briefly introduce the related work about prior studies on text-based person search, human pose for person search, as well as attention for person search.

Text-Based Person Search. Li et al. [16] propose the task of text-based person search and further propose a recurrent neural network with gated neural attention (GNA-RNN) for this task. To utilize identity-level annotations, Li et al. [15] propose an identity-aware two-stage framework. Chen et al. [6] propose a word-image patch matching model in order to capture the local similarity. Different from the above three methods which are all the CNN-RNN architectures, Zheng et al. [31] propose to employ CNN for textual feature learning. Zhang and Lu [27] propose a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss to learn discriminative image-text representations. In this paper, we propose a pose-guided joint global and attentive local matching network to learn multilevel cross-modal relevances.

Human Pose for Person Search. With the development of pose estimation [1, 12, 2, 4, 7, 8], many approaches in image-based person search extract human poses to improve the visual representation. To solve the problem of human pose variations, Liu et al. [17] propose a pose transferrable person search framework through utilizing pose-transferred sample augmentations. Zheng et al. [29] utilize pose to normalize the person image. The normalized image and original image are both used to match the person. Su et al. [22] also utilize pose to normalize the person image,

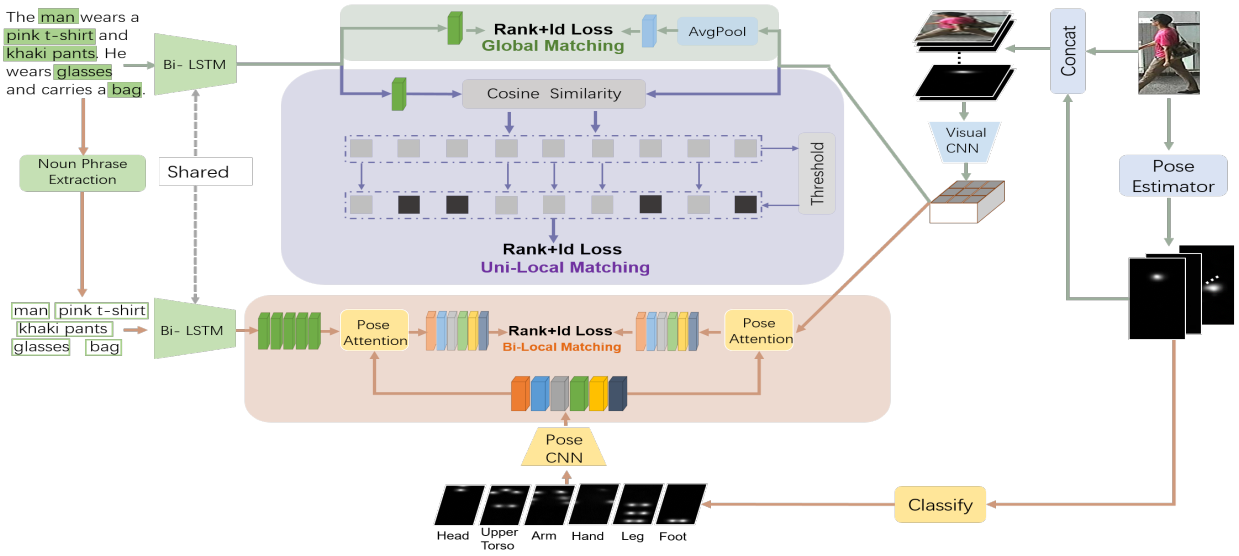

Figure 2. The architecture of our proposed pose-guided joint global and attentive local matching network (GALM) for text-based person search. It contains three level matchings (best viewed in colors): the global matching between the textual description and global visual feature, the uni-local matching between the textual description and local visual feature to select the description-related image regions by hard attention, the pose-guided bi-local matching between noun phrase and local visual feature to learn the latent semantic alignment. To better train GALM, both ranking loss and identification loss are employed in our method.

while they leverage human body parts and the global image to learn a robust feature representation. Sarfraz et al. [20] directly concatenate pose information to the input image to learn visual representation. In this paper, we use pose information for effective visual feature extraction and aligned part matching between noun phrase and image regions.

Attention for Person Search. Attention mechanism aims to select key parts of an input, which is generally divided into soft attention and hard attention. Soft attention computes a weight map and selects the input according to the weight map, while hard attention just preserves one or a few parts of the input and ignores the others. Recently, attention is widely used in person search, which selects either visual contents or textual information. Li et al. [16] compute an attention map based on text representation to focus on visual units. Li et al. [15] propose a co-attention method, which extracts word-image features via spatial attention and aligns sentence structures via a latent semantic attention. Zhao et al. [28] employ human body parts to weight the image feature map. As we know, hard attention is rarely exploited in person search. In this paper, we propose a similarity-based hard attention in uni-local matching and a pose-guided soft attention in bi-local matching.

3. Pose-Guided Joint Global and Attentive Local Matching Network

In this section, we explain the proposed pose-guided joint global and attentive local matching network in detail. First, we introduce the procedures of visual and textual representations extraction. Then, we describe the three-level matching including global matching, uni-local matching and pose-guided bi-local matching. Finally, we give the details of learning the proposed model.

3.1. Visual Representation Extraction

Considering that human pose is closely related to human body parts, we exploit pose information to learn human-related visual feature and guide the aligned part matching. In this work, we estimate human pose from the input image using the PAF approach proposed in [3] due to its high accuracy and realtime performance. To obtain more accurate human poses, we retrain PAF on a larger AI challenge dataset [26] which annotates 14 keypoints for each person.

However, in the experiments, the retrained PAF still cannot obtain accurate human poses in the challenging person search dataset due to the occlusion and lighting change, e.g., CUHK-PEDES. The upper right three images in Fig. 3 show the cases of partial and complete missing keypoints. Looking in detail at the procedure of pose estimation, we find that the 14 confidence maps prior to keypoints generation can convey more information about the person in the image when the estimated joint keypoints are incorrect or missing. The lower right images in Fig. 3 show the superimposed results of the 14 confidence maps which can still provide cues for human body and its parts.

The pose confidence maps play the two-fold role in our model. On one hand, the 14 confidence maps are concatenated with the 3-channel input image to augment vi-

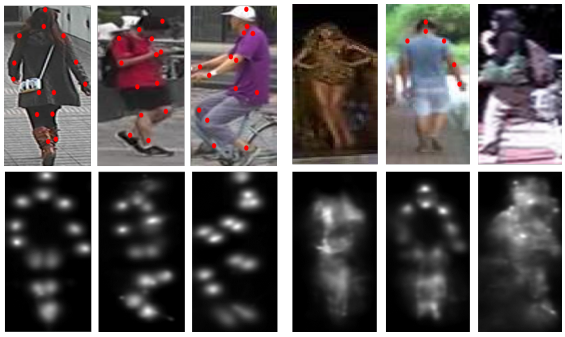

Figure 3. Examples of human pose estimation from the retrained PAF. The first row shows the detected joint keypoints and the second row shows the superimposed results from 14 confidence maps.

sual representation. We extract visual representation using both VGG-16 [21] and ResNet-50 [10] on the augmented 17-channel input and compare their performance in the experiments. Take VGG-16 for example. We first resize the input image to 384×128 inspired by [23] and obtain the feature map $\phi^{'}(I) \in \mathbb{R}^{12\times4\times512}$ before the last pooling layer of VGG-16. Then we partition the $\phi^{'}(I)$ into 6 horizontal stripes and average each stripe along the first dimension. The $\phi^{'}(I)$ is transformed into $\phi(I) \in \mathbb{R}^{6\times4\times512}$, where $6 \times 4 \times 512$ means there are 24 regions and each region is represented by a 512-dimensional vector. The global visual representation $\psi(I) \in \mathbb{R}^{6\times512}$ is defined as follows:

$$\psi(I) = avgpool(\phi(I)), \tag{1}$$

where avgpool means average pooling along the second dimension.

On the other hand, the 14 confidence maps are used to learn latent semantic alignment between noun phrase and image regions, which is explained in Section 3.5.

3.2. Textual Representation Learning

Textual Description. Given a textual description T, we represent its j-th word as an one-hot vector $t_j \in \mathbb{R}^K$ according to the index of this word in the vocabulary, where K is the vocabulary size. Then we embed the one-hot vector into a 300-dimensional embedding vector:

$$x_j = W_t t_j, \tag{2}$$

where $W_t \in \mathbb{R}^{300\times K}$ is the embedding matrix. To model the dependencies between adjacent words, we adopt a bi-directional long short-term memory network (bi-LSTM) [11] to handle the embedding vectors $X = (x_1, x_2, ..., x_r)$ which correspond to the r words of the text description:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t, \overrightarrow{h_{t-1}}), \tag{3}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(x_t, \overleftarrow{h_{t+1}}), \tag{4}$$

where $\overrightarrow{LSTM}$ and $\overleftarrow{LSTM}$ represent the forward and backward LSTMs, respectively.

The global textual representation e^t is defined as the concatenation of the last hidden states $\overrightarrow{h_r}$ and $\overleftarrow{h_1}$:

$$e^t = concat(\overrightarrow{h_r}, \overleftarrow{h_1}) \tag{5}$$

Noun Phrase. For the given textual description, we utilize the NLTK [18] to extract the noun phrase N. Similar to textual description, for the j-th noun phrase n_j in $N = (n_1, n_2, ..., n_m)$, we use the words in n_j to represent the noun phrase according to Equations 2-5. Therefore, we can obtain the representations of all noun phrases $e^n = (e^n_1, e^n_2, ..., e^n_m)$. It should be noted that we adopt the same bi-LSTM when encoding the textual description and noun phrase. Moreover, the number of noun phrase m varies in different textual descriptions.

3.3. Global Image-Text Matching

Given an image-text pair, the most direct way to judge whether they are of the same identity is to measure their global similarity. Therefore, we calculate the global correlation between visual representation $\psi(I)$ and textual representation e^t.

First, we transform visual representation $\psi(I)$ and textual representation e^t to the same feature space as follows:

$$\tilde{e^t} = W_{e^t} e^t, \tag{6}$$

$$\widetilde{\psi(I)} = W_\psi \psi(I), \tag{7}$$

where $W_{e^t} \in \mathbb{R}^{b\times2d}$ and $W_\psi \in \mathbb{R}^{b\times3072}$ are two transformation matrices, and b is the dimension of the feature space. Here d is the hidden dimension of the bi-LSTM in textual representation learning.

Then, the global similarity is computed as follows:

$$S^g = \cos(\widetilde{\psi(I)}, \tilde{e^t}). \tag{8}$$

3.4. Uni-Local Image-Text Matching

Due to the fact that only partial image regions are related to textual description, the global matching described above is not enough for fine matching. Therefore, we propose a hard attention to select the most related image regions to the textual description by the local similarity between image regions and text.

First, we transform local visual representation $\phi(I)$ to the same feature space as $\tilde{e^t}$. For different horizontal stripes in $\phi(I) = \{\phi^{ij}(I) \in \mathbb{R}^{512} \mid i = 1, 2, ..., 6, j = 1, 2, 3, 4\}$, we employ different transformation matrices as follows:

$$\widetilde{\phi^{ij}(I)} = W_{\phi^i} \phi^{ij}(I), \tag{9}$$

where transformation matrix $W_{\phi^i} \in \mathbb{R}^{b\times512}$.

Then, we define the local cosine similarity between each image region representation $\widetilde{\phi^{ij}(I)}$ and textual representation $\tilde{e^t}$:

$$s_{ij} = \cos(\widetilde{\phi^{ij}(I)}, \tilde{e^t}). \tag{10}$$

Accordingly, a set of local similarities can be obtained.

Rather than summing all the local similarities as the final score, we propose a hard attention to select the description-related image regions and ignore the irrelevant ones. Specifically, a threshold τ is set and local similarities can be selected if their weights are higher than this threshold. The final similarity score S^{ul} is defined as follows:

$$q_{ij} = \frac{\exp(s_{ij})}{\sum_{i=1}^{6}\sum_{j=1}^{4}\exp(s_{ij})}, \tag{11}$$

$$S^{ul} = \sum_{q_{ij} \geq \tau} s_{ij}, \tag{12}$$

where q_{ij} is the weight of local similarity and τ is set to $\frac{1}{6\times 4}$. Practically, we can select a fixed number of similarity scores according to their values, but the complex person images can easily fail this idea. In the ablation experiments, we compare the proposed hard attention with several other selection strategies. The experimental results demonstrate the effectiveness of our model.

3.5. Bi-Local Image-Text Matching

A noun phrase in the textual description usually describes specific part of an image, we expect to learn the latent semantic alignment between them and make image-text matching more accurate and interpretable. Considering that human pose is closely related to human body parts, we propose to use pose information to guide the aligned matching between noun phrase and image regions. As described above, for an image, we extract 14 confidence maps which are related to 14 keypoints. As shown in Fig. 2, 14 keypoints are classified into 6 body parts: head, upper torso, arm, hand, leg, foot. It should be noted that keypoints in different parts have overlap.

For each body part, we first add the corresponding confidence maps of keypoints. Next the confidence maps are embedded into a b-dimensional vector $p \in \mathbb{R}^b$ by a pose CNN, which has 4 convolutional layers and a fully connected layer. Then two pose-guided attention models are employed to concentrate on the noun phrases and image regions, respectively. The formulation is illustrated in Fig. 2. Taking textual attention for example, each body part $(p_1, p_2, ..., p_6) \in P$ attends to noun phrases with respect to the cosine similarities. Specifically, for the hand part p_1, we compute the cosine similarity between p_1 and all noun phrases $(e_1^n, e_2^n, ..., e_m^n)$ and employ an attention mechanism on these noun phrases. The attended part-related textual representation is defined as follows:

$$e_1^{pn} = \sum_{i=1}^{m} \alpha_{1,i} e_i^n, \tag{13}$$

$$\alpha_{1,i} = \frac{\exp(\cos(p_1, e_i^n))}{\sum_{i=1}^{m}\exp(\cos(p_1, e_i^n))}, \tag{14}$$

where $\alpha_{1,i}$ indicates the weight value of attention. Similarly, we can obtain the attended part-related visual representations $(\phi_1^p, \phi_2^p, ..., \phi_6^p)$. Then the similarity between attended image-text representations is defined as follows:

$$s_i = \cos(e_i^{pn}, \phi_i^p), i = 1, 2, .., 6 \tag{15}$$

$$S^{bl} = \sum_{i=1}^{6} s_i, \tag{16}$$

where s_i measures the local similarity between corresponding cross-modal parts.

3.6. Learning GALM

The ranking loss is a common objective function for the retrieval task. In this paper, we employ the triplet ranking loss as proposed in [9] to train our GALM, which is defined as follows:

$$\begin{aligned} L_r(I,T) = &max(\alpha - S(I_p, T_p) + S(I_p, T_{\widehat{n}}), 0) \\ &+ max(\alpha - S(I_p, T_p) + S(I_{\widehat{n}}, T_p), 0), \end{aligned} \tag{17}$$

where (I_p, T_p) is a positive pair, $T_{\widehat{n}}$ is the hardest negative text in a mini-batch given an image I_p, $I_{\widehat{n}}$ is the hardest negative image in a mini-batch given a text T_p and α is a margin.

This ranking loss function ensures that the positive pair is closer than the hardest negative pair in a mini-batch with a margin α. For our global matching score, we obtain a global ranking loss L_r^g.

In addition, we adopt the identification loss and classify persons into different groups by their identifications, which ensures the identification-level matching. The global image and text identification losses $L_{id_i}^g$ and $L_{id_t}^g$ are defined as follows:

$$L_{id_i}^g = -\log(softmax(W_{id}^g \widetilde{\psi(I)})), \tag{18}$$

$$L_{id_t}^g = -\log(softmax(W_{id}^g \tilde{e^t})), \tag{19}$$

where $W_{id}^g \in \mathbb{R}^{11003\times b}$ is a shared transformation matrix, and there are 11003 different persons in the training set. We share W_{id}^g between image and text to constrain their representations in the same feature space.

Then the total global loss is defined as:

$$L^g = \lambda_1 L_r^g + \lambda_2 L_{id_i}^g + \lambda_3 L_{id_t}^g, \tag{20}$$

where $\lambda_1, \lambda_2, \lambda_3$ are all set to 1 in our experiments. Similarly, we can obtain the total uni-local loss L^{ul} and bi-local loss L^{bl}.

To make sure that different pose parts can attend to different parts of image and text in bi-local matching, we add an additional part loss L_p to classify 6 pose parts into 6 kinds. The part loss L_p is defined as follows:

$$L_p = \frac{1}{6}\sum_{i=1}^{6} -\log(softmax(W_p p_i)), \tag{21}$$

where $W_p \in \mathbb{R}^{6\times b}$ is a transformation matrix.

Finally, the total loss is defined as:

$$L = \lambda_4 L^g + \lambda_5 L^{ul} + \lambda_6 L^{bl} + \lambda_7 L_p, \tag{22}$$

where $\lambda_4, \lambda_5, \lambda_6, \lambda_7$ are all set to 1 in our experiments.

During testing, we rank the similarity score $S = S^g + S^{ul} + S^{bl}$ to retrieve the person images based on the text query.

4. Experiments

In this section, we first introduce the experimental setup including dataset, evaluation metrics, and implementation details. Then, we analyze the quantitative results of our method and a set of baseline variants. Finally, we visualize some attention maps and retrieval results given text queries.

4.1. Experimental Setup

Dataset and Metrics. The CUHK-PEDES is currently the only dataset for text-based person search. We follow the same data split as [16]. The training set has 34054 images, 11003 persons and 68126 textual descriptions. The validation set has 3078 images, 1000 persons and 6158 textual descriptions. The test set has 3074 images, 1000 persons and 6156 textual descriptions. On average, each image contains 2 different textual descriptions and the textual descriptions contain more than 23 words. The dataset contains 9408 different words.

We adopt top-1, top-5 and top-10 accuracies to evaluate the performance. Given a textual description, we rank all test images by their similarities with the queried text. If top-k images contain any corresponding person, the search is successful.

Implementation Details. In our experiments, we set both the hidden dimension of bi-LSTM and dimension b of the feature space as 1024. For pose CNN, the kernel size of each convolutional layer is 3×3 and the numbers of the convolution channels are 64, 128, 256 and 256, respectively. The fully connected layer has 1024 nodes. In addition, the pose CNN is randomly initialized. After dropping the words that occur less than twice, we get a vocabulary with a size of 4984.

Table 1. Comparison with the state-of-the-art methods using the same visual CNN as us on CUHK-PEDES. Top-1, top-5 and top-10 accuracies (%) are reported. The top section employs the VGG-16 as the visual CNN and the lower section employs the ResNet-50. The best performance is **bold**. "-" represents that the result is not provided.

Method	Visual	Top-1	Top-5	Top-10
CNN-RNN[19]	VGG-16	8.07	-	32.47
Neural Talk[25]	VGG-16	13.66	-	41.72
GNA-RNN[16]	VGG-16	19.05	-	53.64
IATV[15]	VGG-16	25.94	-	60.48
PWM-ATH[6]	VGG-16	27.14	49.45	61.02
Dual Path[31]	VGG-16	32.15	54.42	64.30
GALM(ours)	**VGG-16**	**47.82**	**69.83**	**78.31**
Dual Path[31]	Res-50	44.40	66.26	75.07
GLA[5]	Res-50	43.58	66.93	76.26
GALM(ours)	**Res-50**	**54.12**	**75.45**	**82.97**

We initialize the weights of visual CNN with VGG-16 or ResNet-50 pre-trained on the ImageNet classification task. In order to match the dimension of the augmented first layer, we directly copy the averaged weight along the channel dimension to initialize the first layer. To better train our model, we divide the model training into two steps. First, we fix the parameters of pre-trained visual CNN and only train the other model parameters with a learning rate of $2e^{-3}$. Second, we release the parts of the visual CNN and train the entire model with a learning rate of $2e^{-4}$. We stop training when the loss converges. The model is optimized with the Adam [13] optimizer. The batch size and margin are 128 and 0.2, respectively.

4.2. Quantitative Results

Comparison with the State-of-the-art Methods. Table 1 shows the comparison results with the state-of-the-art methods which use the same visual CNN (VGG-16 or ResNet-50) as we do. Overall, it can be seen that our GALM achieves the best performances under top-1, top-5 and top-10 metrics. Specifically, compared with the best competitor Dual Path [31], our GALM significantly outperforms it under top-1 metric by about 15% with the VGG-16 feature and 10% with the ResNet-50 feature, respectively. The improved performances over the best competitor indicate that our GALM is very effective for this task. Compared with the methods (GNA-RNN [16], IATV [15], PWM-ATH [6] and GLA [5]) which employ the attention mechanism to extract visual representations or textual representations, our GALM also achieves better performances under three evaluation metrics, which proves the superiorities of our similarity-based hard attention and pose-guided part attention in selecting multilevel image-text relations so as to learn diverse and discriminative representations. Although GLA [5] also learns the cross-modal representations by global and local associations, the improved performance

Table 2. Effects of different vocabulary sizes and LSTM on CUHK-PEDES. ResNet-50 is utilized as the visual CNN. Top-1, top-5 and top-10 accuracies (%) are reported.

Method	Top-1	Top-5	Top-10
Base-9408	46.51	68.45	77.32
Base-LSTM	47.02	69.70	78.33
Base(bi-LSTM)	**48.67**	**70.53**	**79.22**

Table 3. Ablation analysis. We investigate the effectiveness of the concatenated pose before visual CNN (Con-pose), global image-text matching (Global), uni-local image-text matching (Uni-Local) and bi-local image-text matching (Bi-Local).

Con-pose	Global	Uni-Local	Bi-Local	Top-1	Top-5	Top-10
×	✓	×	×	48.67	70.53	79.22
✓	✓	×	×	50.06	71.27	79.84
✓	×	✓	×	52.20	72.41	80.48
✓	×	×	✓	41.56	64.84	75.31
✓	✓	✓	×	53.15	73.38	80.98
✓	✓	×	✓	51.97	72.06	80.17
✓	×	✓	✓	52.94	72.92	80.56
✓	✓	✓	✓	**54.12**	**75.45**	**82.97**

(10%) over this method suggests that our pose-guided joint global and attentive local matching network can learn more discriminative and robust multilevel representations by hard attention and pose-guided part matching.

Ablation Experiments. To investigate the several components in GALM, we perform a set of ablation studies. The ResNet-50 is employed as the visual CNN in experiments.

We first investigate the importance of vocabulary size by utilizing all the words in the dataset (9408). The baseline model indicates only the global matching is used in experiments and there is no pose information. As Table 2 shows, utilizing all the words in the dataset has a negative effect on the accuracy, which demonstrates that low-frequency words could make noise to the model. Then we investigate the effectiveness of bi-LSTM. Compared with unidirectional LSTM, the increased performances illustrate that bi-LSTM is more effective to encode textual description.

Table 3 illustrates the effectiveness of the concatenated pose confidence maps before visual CNN (Con-pose), global image-text matching (Global), uni-local image-text matching (Uni-Local) and bi-local image-text matching (Bi-Local). The improved performances demonstrate that these three level matchings can exploit different levels of visual-linguistic relations and Con-pose, Global, Uni-Local and Bi-Local are all effective for text-based person search. Specifically, concatenating the pose confidence maps with original input image indeed improves the matching performance, which demonstrates that pose information is effective in learning discriminative human-related representation. The performance improvement of Con-pose+Glocal+Uni-Local over Con-pose+Glocal by 3.1%

Table 4. Effects of selecting different numbers of regions and utilizing different attention mechanisms. Adaptive hard attention is our similarity-based hard attention model used in uni-local.

Method	Number of Regions	Top-1	Top-5	Top-10
Hard	5	51.47	72.50	80.16
Hard	10	53.86	73.87	81.14
Hard	20	53.29	73.24	80.96
Soft	24	52.34	72.31	80.24
Hard	Adaptive	**54.12**	**75.45**	**82.97**

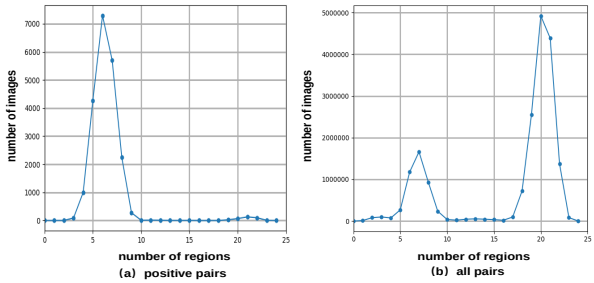

Figure 4. The distribution diagram of the selected number of regions by our similarity-based hard attention network. (a) is the distribution diagram of positive pairs. (b) is the distribution diagram of positive and negative pairs.

in terms of top-1 metric indicates that Uni-Local can help our model select description-related image regions and thus benefit the performances. In addition, the Con-pose+Glocal+Bi-Local outperforms the Con-pose+Glocal by 1.9% in terms of top-1 metric, which proves the effectiveness of aligned part matching in Bi-Local. It is worth noting that although Bi-Local alone does not achieve particularly high performance due to the fact that some estimated poses are not accurate in CUHK-PEDES dataset, it can improve the performances of Global and Uni-Local by adding it. This further indicates that exploiting different levels cross-modal relations are effective in image-text matching by learning sufficient and diverse discriminative representations.

Analysis of Uni-Local Matching. In our experiments, we select a variational number of regions by our hard attention. As illustrated in Fig. 4, our model mainly selects 6 regions with positive image-text pairs and 20 regions with negative image-text pairs, which demonstrates that our model can match the positive image-text pairs better. For positive pairs, our GALM can select significant description-related regions. But for negative pairs, due to the lack of regions corresponding to the textual description, the similarity-based hard attention network is unable to select the regions with high similarity and thus tends to select all regions in image.

We also explore how the number of selected regions af-

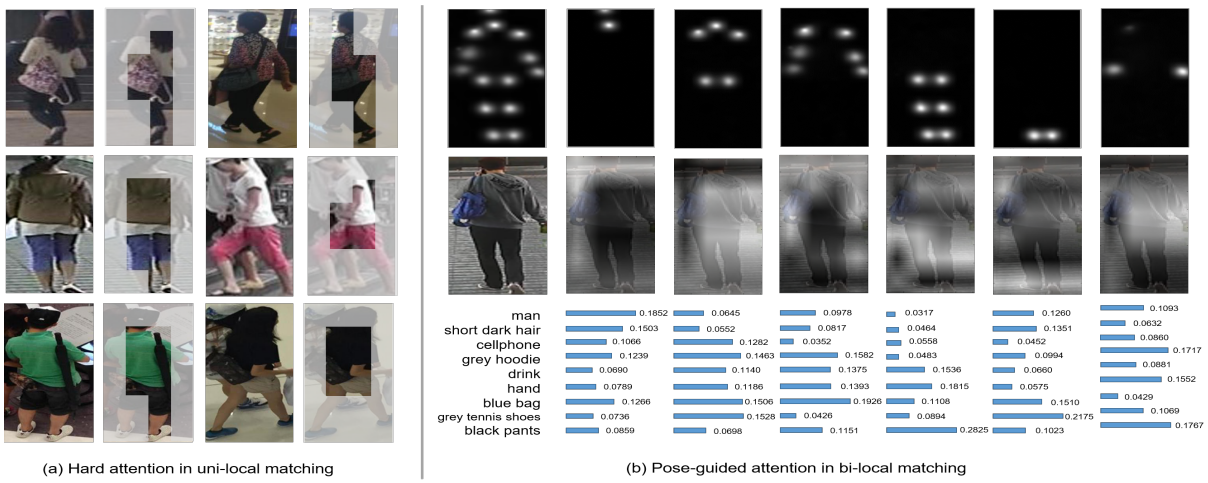

Figure 5. Visualization of the similarity-based hard attention regions and pose-guided part attention map on serval examples. (a) shows the hard attention regions in uni-local matching network. (b) shows the pose-guided part attention map in bi-local matching network. The lighter regions and higher values indicate the attended areas.

Figure 6. Examples of top-6 retrieved images based on the text query by our proposed GALM. Retrieval results are sorted by their similarity scores with text. Green boxes indicate the corresponding images and unmatched images are marked with red boxes. The first two rows are successful cases in which corresponding images are within the top-6 images and the last row is a failure case.

fects the model performance. In experiments, we select the top m regions according to their local similarities. Table 4 shows the experimental results of our similarity-based hard attention under selecting different numbers of regions. We can see that the performance increases when increasing the number of regions, and saturates soon. It denotes that only some description-related regions are useful for matching.

In addition, we compare our hard attention with the soft attention model which re-weights the local similarities according to their values. From the results shown in Table 4, we can see that the soft attention performs worse than most of the hard attention, which demonstrates that our similarity-based hard attention is more effective due to filtering out unrelated visual features.

4.3. Qualitative Results

To verify whether the proposed model can selectively attend to the corresponding regions and make our matching procedure more interpretable, we visualize the focus areas of the similarity-based hard attention model and pose-guided part attention model, respectively. Fig. 5 shows the results, where lighter regions and higher values indicate the attended areas. We can see that the similarity-based hard attention indeed selects the image regions about the described human and filters out the unrelated regions. The pose-guided part attention model can attend to the corresponding image regions and noun phrases to pose parts, which illustrates our model can learn accurate aligned part matching by the guidance of pose information. Specifically, for the hand part, the model mainly attends to the hand region in image and short dark hair in noun phrases. This part level matching benefits the inference of image-text similarity.

Fig. 6 shows the qualitative results of the person search based on the text query by proposed GALM. For each text, we show the top-6 retrieved images ranked by the similarity scores. The first two rows show successful cases and the last row shows a failure case. For successful cases, the corresponding images are within the top-6 images. Although some images are non-corresponding to the text, they also fit parts of the descriptions. For example, in the first case, almost all persons wear "green clothes", which demonstrates the effectiveness of our GALM in matching text and image. However, the case in the third row fails to capture the keywords "white and black shoes".

5. Conclusion

In this paper, we have proposed a novel pose-guided joint global and attentive local matching network for text-based

person search. The global, uni-local and bi-local matching are all utilized to learn multilevel cross-modal relevances. The uni-local matching network selects the description-related image regions by a similarity-based hard attention, while the bi-local matching network employs pose information to guide the aligned part matching. Extensive experiments with ablation analysis on a challenging dataset have shown that our approach outperforms the state-of-the-art methods by a large margin.

References

[1] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] D. Chen, H. Li, X. Liu, Y. Shen, Z. Yuan, and X. Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *European Conference on Computer Vision*, 2018.

[6] T. Chen, C. Xu, and J. Luo. Improving text-based person search by spatial matching and adaptive threshold. In *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[7] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 2017.

[9] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[12] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, 2016.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Conference and Workshop on Neural Information Processing Systems*, 2012.

[15] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *IEEE International Conference on Computer Vision*, 2017.

[16] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] E. Loper and S. Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[19] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[20] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[22] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, 2016.

[23] 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*, 2018.

[24] D. A. Vaquero, R. S. Feris, T. Duan, and L. Brown. Attribute-based people search in surveillance environments. In *IEEE Winter Conference on Applications of Computer Vision*, 2009.

[25] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[26] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, and Y. Fu. Ai challenger : A large-scale dataset for going deeper in image understanding. In *arXiv preprint arXiv:1711.06475*, 2017.

[27] Y. Zhang and H. Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision*, 2018.

[28] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[29] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. In *arXiv preprint arXiv:1701.07732*, 2017.

[30] W. S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[31] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen. Dual-path convolutional image-text embedding. In *arXiv preprint arXiv:1711.05535*, 2017.

[32] Q. Zhou, H. Fan, S. Zheng, H. Su, X. Li, S. Wu, and H. Ling. Graph correspondence transfer for person re-identification. In *The Association for the Advance of Artificial Intelligence*, 2018.