Pose-Guided Multi-Granularity
Attention 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 person images 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 granularities of semantic relevance, which is usually ignored in previous methods. To exploit the multilevel corresponding visual contents, we propose a pose-guided multi-granularity attention network (PMA). Firstly, we propose a coarse alignment network (CA) to select the related image regions to the global description by a similarity-based attention. To further capture the phrase-related visual body part, a fine-grained alignment network (FA) is proposed, 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 the top-1 metric.

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 (Zhou et al. 2018; Zheng, Gong, and Xiang 2011; Sarfraz et al. 2018), attribute-based query (Su et al. 2016; Vaquero et al. 2009) and text-based query (Li et al. 2017b; 2017a; Zheng et al. 2017b; Chen, Xu, and Luo 2018). Image-based person search needs at least one image of the queried person which in many cases is very difficult to obtain. Attribute-based person search has limited capability of
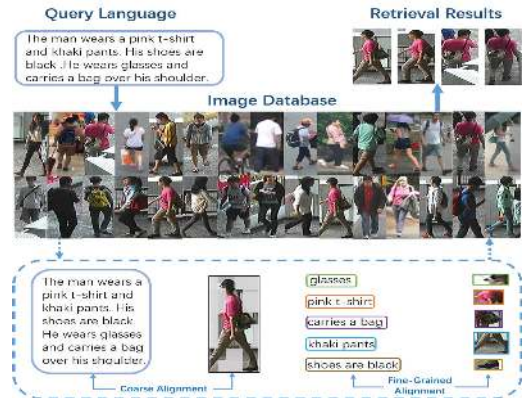

Figure 1: The illustration of text-based person search. Given a textual description of a person, the model aims to retrieve the corresponding person images from the given image database. There are two granularity alignments in our retrieval procedure: coarse alignment by textual description and fine-grained alignment by noun phrases.

describing persons' appearance. Since textual descriptions are more accessible and can describe persons of interest with more details in a more natural way, we study the task of text-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 multi-granularity semantic relevances between image and text. Previous methods (Zheng et al. 2017b; Zhang and Lu 2018) generally utilize Convolutional Neural Networks (CNNs) (Krizhevsky, Sutskever, and Hinton 2012) to obtain a global representation of the input image, which often cannot effectively extract visual contents corresponding to the person in 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, it is still not suitable to match the global features directly 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. Therefore, selecting description-related image regions is necessary for better matching. Previous methods (Chen, Xu, and Luo 2018; Chen et al. 2018) mainly focus on utilizing attention mechanisms for single-level alignment, which largely neglect the multi-granularity semantic relevances. In contrast, we propose to exploit the multi-granularity corresponding visual contents by the given textual description. First, we coarsely select the related visual regions by the whole text. In addition to sentence-level correlations, phrase-level relations are also significant for fine-grained 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 select the phrase-level visual contents, we learn the latent semantic alignment between noun phrase and visual human part. Considering that human pose is closely related to human body parts, we utilize pose information to guide the fine-grained alignment.

In summary, we propose a pose-guided multi-granularity attention network (PMA) for text-based person search as shown in Fig. 2. First, we estimate human pose from the input image, where global visual representations emphasize the human-related features by concatenating the pose confidence maps to the original input image. To further capture the description-related image regions, a coarse alignment network (CA) is performed between textual description and image regions, where a similarity-based attention is utilized. To exploit the phrase-related visual contents, a fine-grained alignment network (FA) is further proposed, where the pose information is utilized to guide the attention of noun phrases and image regions. Both ranking loss and identification loss are employed for better training. Our proposed method is evaluated on a challenging dataset CUHK-PEDES (Li et al. 2017b), which is currently the only available dataset for text-based person search. Experimental results show that our PMA outperforms the state-of-the-art methods on this dataset.

The main contributions of this paper can be summarized as follows: (1) Multi-granularity corresponding visual contents are extracted by the coarse alignment network and fine-grained alignment network. (2) 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. (3) The proposed PMA achieves the best results on the challenging dataset CUHK-PEDES. Extensive ablation studies verify the effectiveness of each component in the PMA.

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. (Li et al. 2017b) propose the task of text-based person search and further employ a recurrent neural network with gated neural attention (GNA-RNN) for this task. To utilize identity-level annotations, Li et al. (Li et al. 2017a) propose an identity-aware two-stage framework. PWM+ATH (Chen, Xu, and Luo 2018) utilizes 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, Dual Path (Zheng et al. 2017b) employs CNN for textual feature learning. CMPM+CMPC (Zhang and Lu 2018) utilizes 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 multi-granularity attention network to learn multilevel cross-modal relevances.

Human Pose for Person Search. With the development of pose estimation (Insafutdinov et al. 2016; Cao et al. 2017a; Carreira et al. 2015; Chu et al. 2016; 2017), many approaches in image-based person search extract human poses to improve the visual representation. To solve the problem of human pose variations, Pose-transfer (Liu et al. 2018) employs a pose transferrable person search framework through pose-transferred sample augmentations. PIE (Zheng et al. 2017a) utilizes human pose to normalize the person image. The normalized image and original image are both used to match the person. SSDAL (Su et al. 2016) also utilizes pose to normalize the person image, while it leverages human body parts and the global image to learn a robust feature representation. PSE (Sarfraz et al. 2018) directly concatenates 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. GNA-RNN (Li et al. 2017b) computes an attention map based on text representation to focus on visual units. IATV (Li et al. 2017a) employs a co-attention method, which extracts word-image features via spatial attention and aligns sentence structures via a latent semantic attention. Zhao et al. (Zhao et al. 2017) 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 fully exploit the hard and soft attentions in learning the semantic relevance between image and text.

Pose-Guided Multi-Granularity Attention Network

In this section, we explain the proposed pose-guided multi-granularity attention network in detail. First, we introduce the procedures of visual and textual representations extraction. Then, we describe the two alignment networks includ-
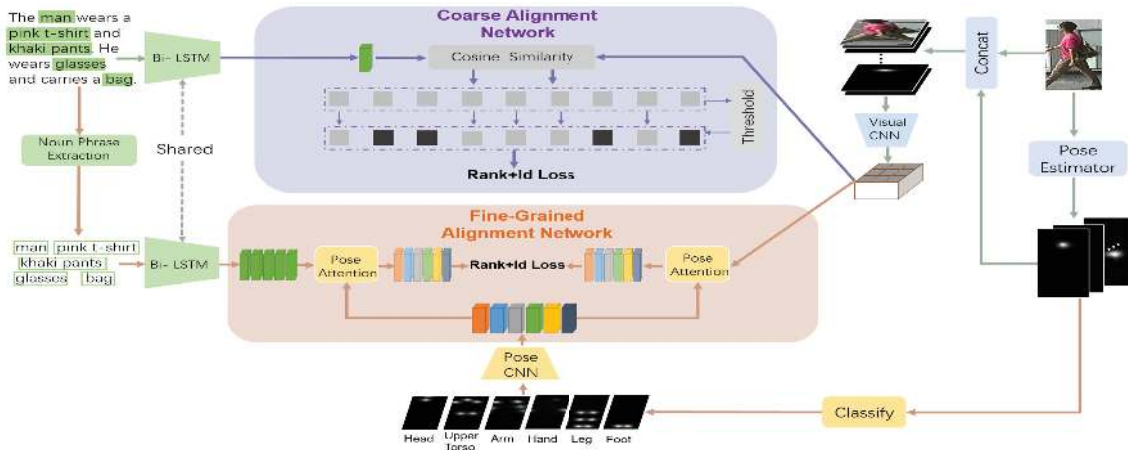
Figure 2: The architecture of our proposed pose-guided multi-granularity attention network (PMA) for text-based person search. It contains two matching networks (best viewed in colors): the coarse alignment network aims to select the description-related image regions by similarity-based attention and fine-grained alignment network aims to select phrase-related visual human parts by pose-guided attention. To better train PMA, both ranking loss and identification loss are employed in our method.

ing coarse alignment network and fine-grained alignment network. Finally, we give the details of learning the proposed model.

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 (Cao et al. 2017b) due to its high accuracy and realtime performance. To obtain more accurate human poses, we retrain PAF on a larger AI challenge dataset (Wu et al. 2017) which annotates 14 keypoints for each person.

However, in the experiments, the retrained PAF still cannot obtain accurate human poses in the challenging CUHK-PEDES dataset due to the occlusion and lighting change. 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 visual representation. We extract visual representation using both VGG-16 (Simonyan and Zisserman 2014) and ResNet-50 (He et al. 2016) on the augmented 17-channel input and compare their performance in the experiments. Taking VGG-16 for example, we first resize the input image to 384×128 and obtain the feature map $\phi'(I) \in \mathbf{R}^{12 \times 4 \times 512}$ before the last pooling layer of VGG-16. Then we parti-

tion the $\phi'(I)$ into 6 horizontal stripes inspired by (Sun et al. 2018) and average each stripe along the first dimension. The $\phi'(I)$ is transformed into $\phi(I) \in \mathbf{R}^{6 \times 4 \times 512}$, where $6 \times 4 \times 512$ means there are 24 regions and each region is represented by a 512-dimensional vector.

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 fine-grained alignment network.

Textual Representation Learning

Textual Description. Given a textual description T, we represent its j-th word as an one-hot vector $t_j \in \mathbf{R}^K$ 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:

$$x_j = W_t t_j, \tag{1}$$

where $W_t \in \mathbf{R}^{300 \times K}$ is the embedding matrix. To model the dependencies between adjacent words, we adopt a bi-directional long short-term memory network (bi-LSTM) (Hochreiter and Schmidhuber 1997) to handle the embedding vectors $X = (x_1, x_2, ..., x_r)$ which correspond to the r words in the text description:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t, \overrightarrow{h_{t-1}}), \tag{2}$$

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

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 $\overleftarrow{h_1}$:

$$e^t = concat(\overrightarrow{h_r}, \overleftarrow{h_1}) \tag{4}$$

Noun Phrase. For the given textual description, we utilize the NLTK (Loper and Bird 2002) to extract the noun
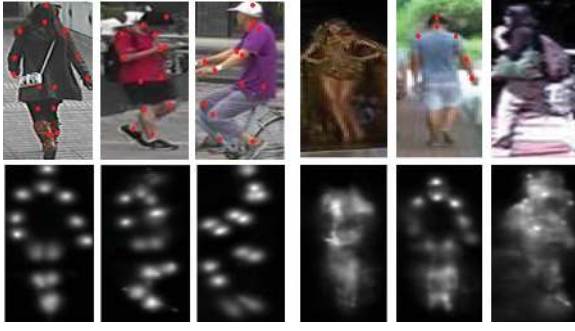
Figure 3: Examples of human pose estimation from the re-trained PAF. The first row shows the detected joint keypoints and the second row shows the superimposed results from 14 confidence maps.

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 1-4. Therefore, we can obtain the representations of all noun phrases $e^n = (e_1^n, e_2^n, ..., e_m^n)$. Note 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.

Coarse Alignment Network

Due to the fact that only partial image regions are related to textual description, we propose a coarse alignment network to select the most related image regions by attention mechanism.

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

$$\widetilde{\phi^{ij}(I)} = W_{\phi^i} \phi^{ij}(I), \tag{5}$$

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

where $W_{\phi^i} \in \mathbf{R}^{b \times 512}$ and $W_{e^t} \in \mathbf{R}^{b \times 2d}$ 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, we define the local cosine similarity between each image region representation $\widetilde{\phi^{ij}(I)}$ and textual representation $\tilde{e}^t$:

$$s_{ij} = \cos(\widetilde{\phi^{ij}(I)}, \tilde{e}^t). \tag{7}$$

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, we set a threshold τ and select the local similarities if their weights are higher than τ. The final similarity score S^{ca} is defined as follows:

$$q_{ij} = \frac{\exp(s_{ij})}{\sum_{i=1}^{6} \sum_{j=1}^{4} \exp(s_{ij})}, \tag{8}$$

$$S^{ca} = \sum_{q_{ij} \geq \tau} s_{ij}, \tag{9}$$

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.

Fine-Grained Alignment Network

In fact, a noun phrase in the textual description usually describes specific part of an image. To extract the phrase-level related visual contents, we propose a fine-grained alignment network to learn the latent semantic alignment between noun phrase and image regions. Considering that human pose is closely related to human body parts, we propose to use pose information to guide the matching. 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 $p \in \mathbf{R}^b$ 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 head part p_1, we first transform all noun phrases $(e_1^n, e_2^n, ..., e_m^n)$ to the same feature space as p_1 and then compute the cosine similarity between p_1 and all transformed noun phrases $(\tilde{e}_1^n, \tilde{e}_2^n, ..., \tilde{e}_m^n)$. The attended head-related textual representation is defined as follows:

$$e_1^{pn} = \sum_{i=1}^{m} \alpha_{1,i} \tilde{e}_i^n, \tag{10}$$

$$\alpha_{1,i} = \frac{\exp(\cos(p_1, \tilde{e}_i^n))}{\sum_{i=1}^{m} \exp(\cos(p_1, \tilde{e}_i^n))}, \tag{11}$$

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:

$$s_i = \cos(e_i^{pn}, \phi_i^p), i = 1, 2, .., 6 \tag{12}$$

$$S^{fa} = \sum_{i=1}^{6} s_i, \tag{13}$$

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

Learning PMA

The ranking loss L_r is a common objective function for the retrieval task. In this paper, we employ the triplet ranking loss as proposed in (Faghri et al. 2018) to train our PMA. 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 coarse alignment network, we obtain a ranking loss L_r^{ca}.

In addition, we adopt the identification loss and classify persons into different groups by their identities, which ensures the identity-level matching. The image and text identification losses $L_{id_i}^{ca}$ and $L_{id_t}^{ca}$ in coarse alignment network are defined as follows:

$$\widetilde{\phi(I)_{avg}} = avgpool(\widetilde{\phi(I)}), \tag{14}$$

$$L_{id_i}^{ca} = -y_{id}\log(softmax(W_{id}^{ca}\widetilde{\phi(I)_{avg}})), \tag{15}$$

$$L_{id_t}^{ca} = -y_{id}\log(softmax(W_{id}^{ca}\tilde{e^t})), \tag{16}$$

where $W_{id}^{ca} \in \mathbf{R}^{11003 \times b}$ is a shared transformation matrix to classify the 11003 different persons in the training set, y_{id} is the ground truth identity. We share W_{id}^{ca} between image and text to constrain their representations in the same feature space.

Then the total loss in CA is defined as:

$$L^{ca} = L_r^{ca} + \lambda_1 L_{id_i}^{ca} + \lambda_2 L_{id_t}^{ca}, \tag{17}$$

where λ_1 and λ_2 are empirically set to 1 in our experiments. Similarly, we can obtain the total loss L^{fa} in FA .

To make sure that different pose parts can attend to different parts of image and text in fine-grained alignment network, 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:

$$L_p = \frac{1}{6}\sum_{i=1}^{6} -y_{p_i}\log(softmax(W_p p_i)), \tag{18}$$

where $W_p \in \mathbf{R}^{6 \times b}$ is a transformation matrix and y_{p_i} is the ground truth label of p_i.

Finally, the total loss is defined as:

$$L = L^{ca} + \lambda_3 L^{fa} + \lambda_4 L_p, \tag{19}$$

where λ_3 and λ_4 are also empirically set to 1 in our experiments.

During testing, we rank the similarity score $S = S^{ca} + \lambda_3 S^{fa}$ to retrieve the person images based on the text query.

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 several attention maps.

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 best performance is bold.

Method	Visual	Top-1	Top-5	Top-10
CNN-RNN(2016)	VGG-16	8.07	-	32.47
Neural Talk(2015)	VGG-16	13.66	-	41.72
GNA-RNN(2017b)	VGG-16	19.05	-	53.64
IATV(2017a)	VGG-16	25.94	-	60.48
PWM-ATH(2018)	VGG-16	27.14	49.45	61.02
Dual Path(2017b)	VGG-16	32.15	54.42	64.30
PMA(ours)	**VGG-16**	**47.02**	**68.54**	**78.06**
Dual Path(2017b)	Res-50	44.40	66.26	75.07
GLA(2018)	Res-50	43.58	66.93	76.26
PMA(ours)	**Res-50**	**53.81**	**73.54**	**81.23**

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 (Li et al. 2017b). 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 each textual description contains more than 23 words. The dataset contains 9408 different words. In addition, the method we proposed can be utilized in other human related cross-modal tasks.

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 query 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 convolutional 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, the obtained vocabulary has 4984 words.

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 $1e^{-3}$. 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 (Kingma and Ba 2014) optimizer. The batch size and margin are 128 and 0.2, respectively.

Table 2: Effects of different vocabulary sizes and LSTMs on CUHK-PEDES. ResNet-50 is utilized as the visual CNN.

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), coarse alignment network (CA), and fine-grained alignment network (FA).

Con-pose	CA	FA	Top-1	Top-5	Top-10
×	×	×	48.67	70.53	79.22
√	×	×	50.06	71.27	79.84
√	√	×	52.20	72.41	80.48
√	√	√	**53.81**	**73.54**	**81.23**

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 PMA achieves the best performances under top-1, top-5 and top-10 metrics. Specifically, compared with the best competitor Dual Path (Zheng et al. 2017b), our PMA significantly outperforms it under top-1 metric by about 15% with the VGG-16 feature and 9% with the ResNet-50 feature, respectively. The improved performances over the best competitor indicate that our PMA is very effective for this task. Compared with the methods (GNA-RNN (Li et al. 2017b), IATV (Li et al. 2017a), PWM-ATH (Chen, Xu, and Luo 2018) and GLA (Chen et al. 2018)) which employ the attention mechanism to extract visual representations or textual representations, our PMA also achieves better performances under three evaluation metrics, which proves the superiorities of our similarity-based attention and pose-guided attention in selecting multi-granularity image-text relations so as to learn diverse and discriminative representations. Although GLA (Chen et al. 2018) also learns the cross-modal representations by global and local associations, the improved performance (10%) over it suggests that our pose-guided multi-granularity attention network can learn more discriminative and robust multi-granularity representations.

Ablation Experiments. To investigate the several components in PMA, 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 utilizing the global visual and textual features to compute the similarity in experiments. In addition, there is no pose information. As shown in Table 2, 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 ef-

Table 4: Effects of selecting different numbers of regions and utilizing different attention mechanisms in CA. Adaptive hard attention is our proposed similarity-based hard attention model.

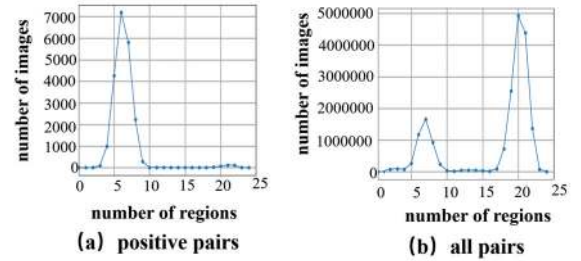| Method	Number of Regions	Top-1	Top-5	Top-10
PMA(CA-Hard)	5	51.24	72.05	79.87
PMA(CA-Hard)	10	53.56	73.16	80.97
PMA(CA-Hard)	20	53.21	72.98	80.78
PMA(CA-Soft)	24	52.35	72.66	80.24
PMA(CA-Hard)	Adaptive	**53.81**	**73.54**	**81.23**

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.

fective to encode textual description.

Table 3 illustrates the effectiveness of concatenated pose confidence maps before visual CNN (Con-pose), coarse alignment network (CA), and fine-grained alignment network (FA). The improved performances compared with the strong baseline demonstrate that Con-pose, CA and FA 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+CA over Con-pose by 2.2% in terms of top-1 metric indicates that CA can help our model select coarse description-related image regions and thus benefit the performances. In addition, the Con-pose+CA+FA outperforms the Con-pose+CA by 1.6% in terms of top-1 metric, which proves the effectiveness of pose-guided fine-grained aligned matching. The overall performances further indicate that exploiting multi-granularity cross-modal relations are effective in image-text matching by learning sufficient and diverse discriminative representations.

Analysis of Coarse Alignment Network. 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 pairs better. For positive pairs, our PMA 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 atten-
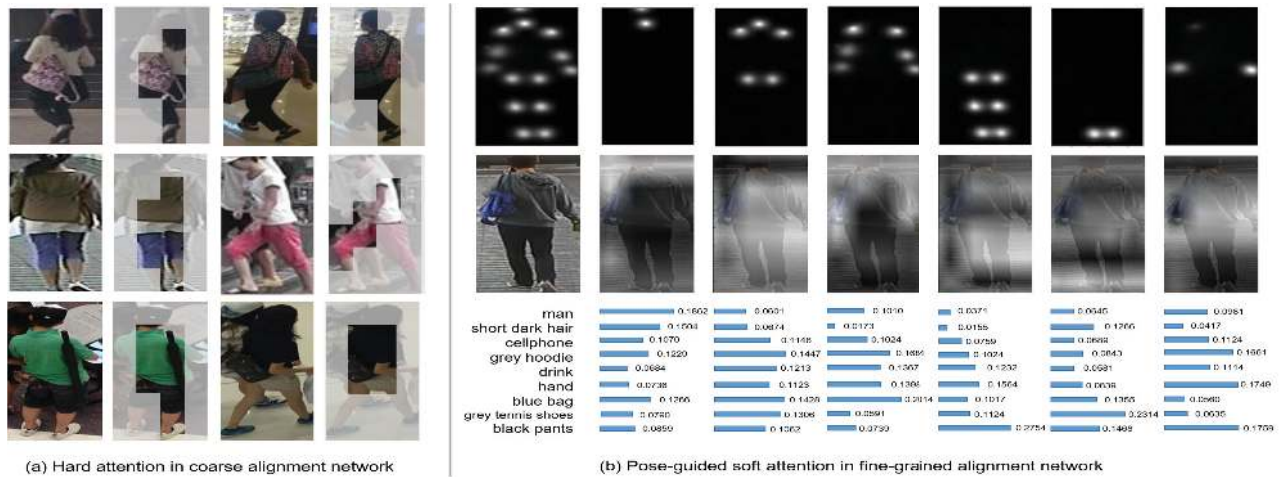
Figure 5: Visualization of the similarity-based hard attention regions and pose-guided soft attention map on serval examples. (a) shows the selected regions in coarse alignment network. (b) shows the attention map in fine-grained alignment network. The lighter regions and higher values indicate the attended areas.

Table 5: Effects of pose information and noun phrases used in fine-grained alignment network.

Method	Top-1	Top-5	Top-10
PMA (FA w/o pose)	52.54	72.69	80.73
PMA (FA w/o noun phrase)	52.71	72.87	80.69
PMA (FA)	**53.81**	**73.54**	**81.23**

tion network is unable to select the regions with high similarity score and thus tends to select all regions in image.

We also exploit how the number of selected regions affects the performance of our PMA. In experiments, we select the top m regions according to their local similarities. Table 4 shows the experimental results. We can see that the performance increases when increasing the number of selected 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, we can see that the soft attention performs worse than most of the hard attention, which demonstrates that our hard attention in CA is more effective due to filtering out unrelated visual features.

Analysis of Fine-Grained Alignment Network. In experiments, we utilize 6 body parts rather than 14 score maps of body joints to guide the attention due to the fact that 14 score maps of body joints are just the coordinates of the joints. Therefore, it is difficult to model the relationships with image regions, as well as with the text. To investigate the effectiveness of pose information and noun phrases used in fine-grained alignment network, we perform ablation analysis by dropping one of them. When dropping the pose information, we utilize noun phrases to guide the attention of image regions by their similarities. In addition, dropping the noun phrases means we utilize every word in sentence. As shown in Table 5, the performance decreases when drop-

ping one of the key components in FA, which indicates that pose information is effective in selecting aligned image-text contents and noun phrase has more complete semantic information than word. Moreover, we perform the FA with hard attention just like CA but do not obtain meaningful results.

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 attention weights of the similarity-based hard attention model and pose-guided soft 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 soft attention model can attend to the corresponding image regions and noun phrases to pose parts, which illustrates our model can learn accurate aligned fine-grained matching by the guidance of pose information. Specifically, for the foot part, the model mainly attends to the foot region in image and grey tennis shoes in noun phrases. This fine-grained matching benefits the inference of image-text similarity.

Conclusion

We propose a novel pose-guided multi-granularity attention network for text-based person search in this paper. Both coarse alignment network and fine-grained alignment network are proposed to learn multi-granularity cross-modal relevances. The coarse alignment network selects the description-related image regions by a similarity-based attention, while the fine-grained alignment network employs pose information to guide the attention of phrase-related visual contents. Extensive experiments with ablation analysis on a challenging dataset show that our approach outperforms the state-of-the-art methods by a large margin.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*.

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

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

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

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

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

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

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

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

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