Displays 69 (2021) 102039

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Description-based person search with multi-grained matching networks

Ji Zhu [a,b], Hua Yang [a,*], Jia Wang [a], Wenjun Zhang [a]

[a] The Institute of Image Communication and Network Engineering, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Visbody Inc, Shenzhen 518000, China

ARTICLE INFO

Keywords:
Description-based person search
Visual-textual matching
Cross-modal matching
Attention mechanism
Multi-grained matching networks

ABSTRACT

Description-based person search aims to retrieve a person in the image database based on a description about that person. It is a challenging task since the visual image and the textual description belong to different modalities. To fully capture the relevance between person images and textual descriptions, we propose a multi-grained framework with three branches for visual-textual matching. Specifically, in the global-grained branch, we extract global contexts from the entire images and descriptions. In the fine-grained branch, we adopt visual human parsing and linguistic parsing to split images and descriptions into semantic components related to different body parts. We design two attention mechanisms including segmentation-based and linguistics-based attention to align visual and textual semantic components for fine-grained matching. To further exploit the spatial relations between fine-grained semantic components, we construct a body graph in the coarse-grained branch and exploit graph convolutional neural networks to aggregate fine-grained components into coarse-grained representations. The visual and textual representations learned by three branches are complementary to each other which enhance the visual-textual matching performance. Experimental results on the CUHK-PEDES dataset show that our approach performs favorably against state-of-the-art description-based person search methods.

1. Introduction

Description-based person search aims to automatically retrieve a person in images based on a corresponding textual description about that person. It has wide applications in video surveillance such as tracking suspects and searching missing persons since it is unrealistic to manually find the target in such a large-scale video database. Different from the traditional person re-identification (re-ID) task [1–6] which requires an existing image of the queried person, description-based person search only uses a natural description of the target person and directly applies the cross-modal matching between a sentence and candidate person images. Thus, it can deal with more cases where the image of a queried person is hard to obtain.

Despite its advantage in flexibility and convenience, description-based person search also brings new challenges. First, due to the large flexibility in natural language, an image can be described in many ways with different words and sentences. The language model is required to be able to extract consistent semantics from variable textual descriptions for the same image. Second, the textual description and visual appearance belong to different modalities. Thus, to measure the similarity between them, they need to be further mapped to a shared visual-textual feature space. Existing methods like [7,8] extract the visual and textual features independently in a holistic mode without considering the interaction between fine-grained visual regions and textual concepts. Some methods [9–11] adopt attention mechanisms to compute the affinities between visual regions and textual words. However, they simply split the visual feature map into grid regions and separate the description word by word, rather than explicitly locate the meaningful semantic parts in images and descriptions. Recent methods [12,13] employ stripe partition or pose estimation to locate semantic body parts in a person image and learn the alignment with noun phrases in a description. However, both stripes and body key points are not accurate enough to locate the semantic visual parts and will inevitably introduce some background noise.

As shown in Fig. 1, since the textual description for a queried person is mainly composed of noun phrases related to human body parts, it is intuitive that the key for description-based person search is to find the correspondences between body parts in an image and semantic concepts in a description. Motivated by this, we propose a multi-grained matching framework as shown in Fig. 2. First, like many existing works, we have a

* Corresponding author.
E-mail address: hyang@sjtu.edu.cn (H. Yang).

https://doi.org/10.1016/j.displa.2021.102039
Received 11 April 2021; Received in revised form 18 May 2021; Accepted 31 May 2021
Available online 15 June 2021
0141-9382/© 2021 Elsevier B.V. All rights reserved.

global-grained branch to extract the global contexts from images and descriptions. Second, in the fine-grained branch, we employ the human parsing model to accurately segment each person image into body parts and apply linguistic parsing to decompose each sentence into noun phrases. We design two attention mechanisms including segmentation-based attention and linguistics-based attention to align visual body parts and noun phrases. Some previous works [12–14] also exploit attention mechanisms for visual-textual alignment but the problem is that they do not exclude the descriptive information from interacted visual and textual features when generating the attention. However, the extra descriptive information (e.g., color information) encoded in the interacted features might interfere with the alignment of corresponding visual and textual components. For example, the visual body part with a white shirt in the image and the noun phrase of "a blue shirt" in the description should be aligned for comparison but the inconsistent color information might hamper the alignment. Different from these works, our attention models exploit the visual segmentation mask of each body part and the last noun word of each noun phrase which exclude the descriptive information for more reliable visual-textual alignment. Third, inspired by recent computer vision works [15–18] which model domain knowledge as a graph to encode the spatial and semantic connections between visual objects, we build a body graph in the coarse-grained branch based on the external knowledge of human body structure. We employ the Graph Convolutional Networks (GCNs) [19] to aggregate fine-grained semantic components into coarse-grained representations which capture the spatial relations between fine-grained components.

The contributions of this work are summarized as follows:

- We propose a multi-grained visual-textual matching framework for description-based person search. The visual and textual representations with different granularities are complementary to each other which enhance the image-text matching performance.
- In the fine-grained visual-textual matching branch, we split images and descriptions into fine-grained components related to different body parts and design two attention mechanisms to align visual regions with textual concepts. Different from previous works, our attention modules are designed to avoid being interfered by extra descriptive information for more reliable visual-textual alignment.
- In the coarse-grained visual-textual matching branch, we further inject the external knowledge of human body structure by building a body graph and employ GCNs to aggregate fine-grained visual body parts and textual concepts into coarse-grained representations. The coarse-grained representations capture the spatial relations between fine-grained components and eliminate the background interference.
- Experimental results on the CUHK-PEDES dataset show that our approach performs favorably against state-of-the-art description-based person search methods. We further carry out ablation studies to demonstrate the effectiveness of each contribution in this work.

2. Related works

The traditional person re-ID task aims to retrieve a person based on a query image. It has attracted great attention in both computer vision research and industry communities. Existing person re-ID approaches either dedicate to design discriminative appearance representations [20–27] or focus on learning a robust distance metric in the feature space [28–33]. For appearance representation extraction, most recent works employ deep learning methods. For example, Ahmed et al. [22] design a convolutional network which produces the cross-input neighborhood difference map on mid-level features to represent local relevances between an input pair of person images. A patch summary layer is proposed to further generate a holistic representation of the

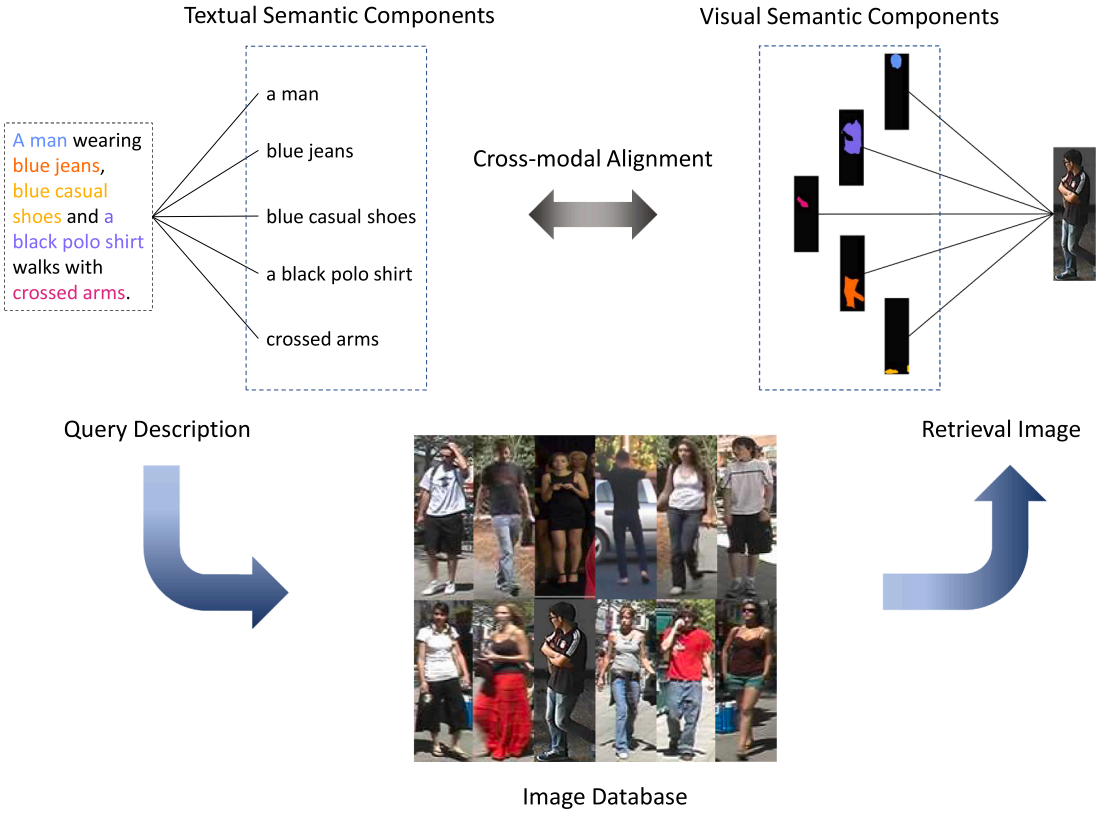


Fig. 1. Description-based person search aims to retrieve a person from the image database based on a textual description about that person. The textual description is mainly composed of noun phrases related to different human body parts. One key for description-based person search is to find the cross-modal alignment between visual and textual semantic components for meaningful fine-grained comparison (best viewed in color).

neighborhood difference map. Sun et al. [25] design a model named Part-based Convolutional Baseline (PCB) to generate part-level features and adopt an adaptive part pooling method to refine the within-part consistency. Kalayeh et al. [26] employ human parsing to provide extra local semantic information for person re-ID. Zhang et al. [27] further exploit the DensePose [34] to estimate the dense semantics of person images and address the misalignment problem caused by viewpoint variations. For metric learning, Bak et al. [30] propose a one-shot learning algorithm for person re-ID where the re-ID metric is decomposed into independent texture and color components. Hermans et al. [31] design a new variant of triplet loss called batch hard loss which makes the re-ID network easier to converge during training compared to the traditional triplet loss and improves the performance.

Different from person re-ID based on the query image, description-based person search aims to retrieve a person in images based on a textual description of that person. It brings new challenges due to the modality heterogeneity. To learn joint embeddings for visual-textual matching, Ying et al. [7] employ a CNN-RNN architecture and propose a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss. Zheng et al. [8] design a dual path CNN model and propose a large-number classification loss named instance loss. These two methods embed the image and description into holistic representations in a shared visual-textual space without considering the local similarities between fine-grained visual regions and textual concepts. Methods like the GNA-RNN [9], IATV [10], PWM-ATH [11], and GLA [35] design attention mechanisms to capture the local relations between the image and description. However, these methods do not explicitly locate and align meaningful semantic body parts in an image or description. To exploit relevances between visual and textual semantic components, Niu et al. [12] split a person image into six horizontal stripes and decompose a description into noun phrases. Jing et al. [13] employ the pose information to guide the alignment between visual body parts and textual noun phrases. Wang et al. [14] employ the human parsing model to accurately locate body parts in an image and parse each description into noun phrases. However, they [12–14] align visual and textual components based on features with descriptive information which might interfere with the alignment.

In this work, we explicitly align fine-grained visual and textual semantic components through attention mechanisms. Different from attention models designed in previous works, to generate the visual-textual attention, we proposed to exploit the segmentation masks and the last noun words in noun phrases which exclude descriptive information so that the visual and textual semantic components corresponding to the same body part but with inconsistent descriptive information (e.g., with different colors) can still be aligned for comparison without interference. We further employ GCNs [19] to aggregate fine-grained visual and textual representations of different body parts into coarse-grained representations based on the human body structure constraints. The multi-grained representations are complementary to each other which enhance the visual-textual matching performance.

3. Proposed Method

We propose a multi-grained image-text matching framework with three branches as shown in Fig. 2. In the global-grained branch, we extract the visual and textual representations from the entire image and description respectively to capture the global semantic contexts. In the fine-grained branch, we extract the visual and textual representations related to different body parts by performing human parsing on each

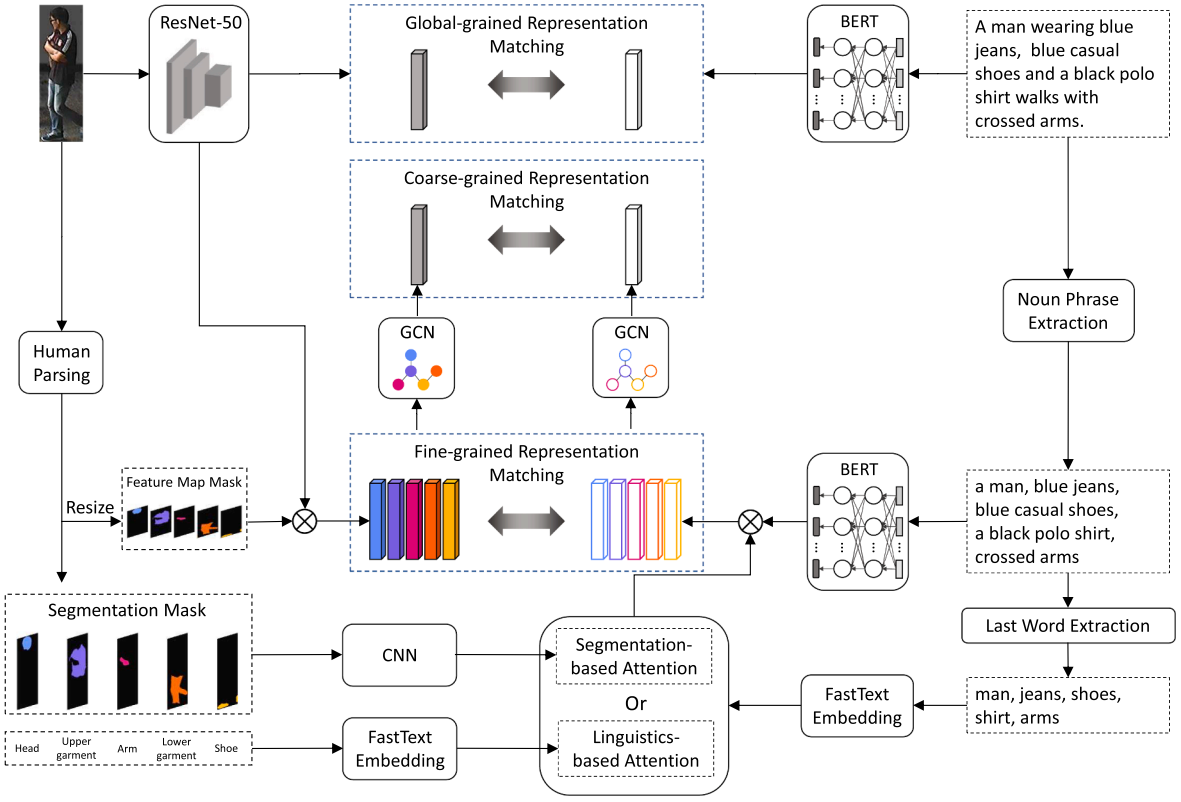


Fig. 2. Architecture of the proposed multi-grained visual-textual matching networks. Given a person image and a description, we first use the ResNet-50 and BERT to extract holistic visual and textual features respectively. The extracted features are mapped into a shared embedding space for global-grained visual-textual matching. Then we split the input image and description into semantic components related to different body parts and exploit attention mechanisms to align visual and textual components for fine-grained visual-textual matching. Furthermore, we build a body graph to model the spatial relations between fine-grained semantic components and exploit GCNs to aggregate them into coarse-grained representations for coarse-grained visual-textual matching. The multi-grained representations are complementary to each other and we use the sum of their similarity scores as the matching score for the input image-description pair (best viewed in color).

person image and linguistic parsing on each textual description. To align fine-grained visual and textual components, we design two attention mechanisms which utilize the interaction of body parts and noun phrases to model the visual-textual relevance. In the coarse-grained branch, we introduce the external knowledge of human body structures to build a body graph and employ the GCNs [19] to aggregate fine-grained representations into coarse-grained representations. In the following sections, we explain the details for computing representations of each branch respectively and describe the learning method.

3.1. Global-grained Representation

In the global-grained branch, we extract representations to capture the global semantic information of visual images and textual descriptions. Specifically, given a person image, we adopt the ResNet-50 [36] pre-trained on the ImageNet [37] as the base network and extract the intermediate visual feature from the last global pooling layer. The global visual representation $\mathbf{x}^g \in \mathbb{R}^{512}$ is then obtained by mapping the intermediate visual feature into a shared visual-textual space through a fully connected layer.

Given a textual description, we extract word embeddings using the BERT [38] model which is able to capture contextual relations between words in a textual description. We obtain the intermediate textual feature through a max-pooling layer on all the word embeddings. And the global textual representation $\mathbf{z}^g \in \mathbb{R}^{512}$ is then generated by projecting the intermediate textual feature into a shared visual-textual space through a fully connected layer.

3.2. Fine-grained Representation

We propose the fine-grained branch to exploit the relations between the visual body parts in an image and textual concepts in a description. To extract fine-grained visual representations from different body parts, we adopt the Graphonomy [18] method for human parsing on a person image. The universal human parsing model learned by the Graphonomy has a good generalization ability and achieves good parsing results on the CUHK-PEDES dataset. Fig. 3 shows a human parsing example generated by the Graphonomy model. For each input person image, the Graphonomy model outputs a confidence map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where H denotes the image height, W is the image width, and C is equal to the number of body part categories. Each channel of the confidence map corresponds to a body part and each value in that channel indicates the confidence of a spatial location belonging to the corresponding body part. The original 0–1 binary segmentation mask $\mathbf{M}^o \in \mathbb{R}^{C\times H\times W}$ is then obtained by performing the *argmax* operation along the channel dimension on the confidence map $\mathbf{F}$. To extract the visual feature for each body part, we resize each channel of the confidence map $\mathbf{F}$ to 7×7 (i.e., the same spatial size as the last convolutional feature map of the ResNet-50 used in the global-grained branch) and apply the *argmax* operation along the channel dimension to generate the original feature map mask $\mathbf{M}_f^o \in \mathbb{R}^{C\times 7\times 7}$ for each body part. The original segmentation mask $\mathbf{M}^o$ and feature map mask $\mathbf{M}_f^o$ contains C channels corresponding to very detailed body part labels such as hair and face. We remove the channel corresponding to the background and group the rest of body part channels into 5 categories: head, upper garment, arm, lower garment, and shoe as shown in Fig. 3. By summing up the mask channels grouped into the same body part category, we get the merged segmentation mask $\mathbf{M} \in \mathbb{R}^{5\times H\times W}$ and merged feature map mask $\mathbf{M}_f \in \mathbb{R}^{5\times 7\times 7}$. We further reshape the feature map mask to $\mathbf{M}_f \in \mathbb{R}^{5\times 49}$ where each row of $\mathbf{M}_f$ is a flattened mask vector corresponding to a body part category. Similarly, we reshape the last convolutional feature map of the ResNet-50 to $\mathbf{X} \in \mathbb{R}^{2048\times 49}$ and generate the masked visual feature $\widetilde{\mathbf{X}} \in \mathbb{R}^{5\times 2048}$ as:

$$\widetilde{\mathbf{X}} = \mathbf{M}_f \mathbf{X}^\top. \tag{1}$$

The masked visual feature $\widetilde{\mathbf{X}}$ is further mapped to a shared visual-textual space to get the fine-grained visual representations $\mathbf{X}^f \in \mathbb{R}^{5\times 512}$ as:

$$\mathbf{X}^f = \widetilde{\mathbf{X}}\mathbf{W}_x^f, \tag{2}$$

where $\mathbf{W}_x^f \in \mathbb{R}^{2048\times 512}$ is a trainable weight matrix for feature mapping. The i-th row of $\mathbf{X}^f$ is the fine-grained visual representation for the i-th body part.

To extract fine-grained textual representations, we decompose a textual description into N noun phrases using the NLTK [39]. The noun phrase features $\mathbf{P} = [\mathbf{p}_1; \cdots; \mathbf{p}_N] \in \mathbb{R}^{N\times 512}$ are then extracted using the BERT [38] model. To generate fine-grained textual representations aligned to the 5 body part categories, we employ the attention mechanism to aggregate noun phrase features $\mathbf{P}$ based on their relevances with each body part. We design and compare two kinds of attention mechanisms as explained below.

Segmentation-Based Attention. Considering that the position and shape of a body part segmentation mask are important cues to decide which body part it belongs to, we design an attention module which exploits the interplay of the segmentation masks and the last noun words in noun phrases to model the relevances between body parts and noun phrases. Note that we exploit features of the segmentation mask and the last noun word but not directly use features of the visual body part and the whole noun phrase because the extra descriptive information (e.g., color information) encoded in the body part or noun phrase features might interfere with the alignment between them. For example, the visual body part with a white shirt in an image and the noun phrase of "a blue shirt" in a description should be aligned for fine-grained comparison despite that they are in different colors. Specifically, we first exploit a CNN with 4 convolutional layers and a fully connected layer to transform each i-th channel of the segmentation mask $\mathbf{M}$ into a 256-dimensional mask feature vector $\widehat{\mathbf{m}}_i$. Then we map both the mask feature vectors $\widehat{\mathbf{M}} = [\widehat{\mathbf{m}}_1; \cdots; \widehat{\mathbf{m}}_5] \in \mathbb{R}^{5\times 256}$ in an image and the last word embeddings $\mathbf{E} = [\mathbf{e}_1; \cdots; \mathbf{e}_N] \in \mathbb{R}^{N\times 300}$ of N noun phrases in a description into a shared latent space and calculate the dot-product affinity matrix $\mathbf{A}^{seg} \in \mathbb{R}^{5\times N}$ between them as:

$$\mathbf{A}^{seg} = \left(\widehat{\mathbf{M}}\mathbf{W}_m\right)\left(\mathbf{E}\mathbf{W}_e\right)^\top, \tag{3}$$

where $\mathbf{W}_m \in \mathbb{R}^{256\times 256}$, $\mathbf{W}_e \in \mathbb{R}^{300\times 256}$ are trainable matrices which map the mask feature vectors and the last word embeddings into a 256-dimensional feature space. To compute the attention on each noun phrase with respect to each body part category, we normalize the affinity matrix $\mathbf{A}^{seg}$ along the noun phase dimension (i.e., column dimension) to obtain the attention matrix $\widetilde{\mathbf{A}}^{seg} \in \mathbb{R}^{5\times N}$ as:

$$\widetilde{\mathbf{A}}^{seg} = \mathrm{softmax}\left(\frac{\mathbf{A}^{seg}}{\sqrt{256}}\right), \tag{4}$$

where the scale factor $1/\sqrt{256}$ is introduced to avoid extremely small

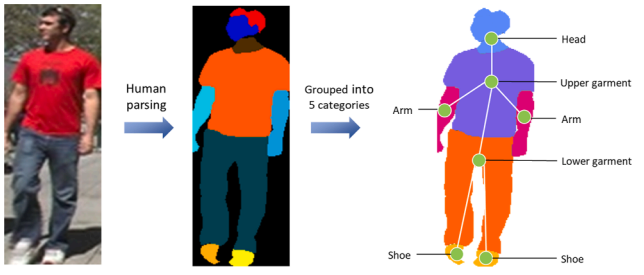


Fig. 3. An example of the human parsing result. We further group the original body part segmentations into 5 categories and construct a graph to capture spatial relations between body parts (best viewed in color).

gradients as suggested in [40].

Linguistics-Based Attention. Besides exploiting the visual segmentation information, we also explore the linguistic knowledge to model the relevances between noun phrases and body part categories. For each body part category, we construct a set of related words (e.g., words like "hair", "glasses", "face" are in the set of the head category) and map each word into a 300-dimensional word embedding using the FastText model [41]. Here we use the FastText but not the BERT language model because we want to extract embeddings of each single word without the need to capture contextual information. To obtain the affinity matrix $\mathbf{A}^{lin} \in \mathbb{R}^{5\times N}$ between 5 body part categories and N noun phrases in a textual description, for each body part category i and noun phrase j, we regard the last noun word in the phrase j as the keyword and calculate its cosine similarity with each word belonging to the body part category i. The word in the set of body part category i which has the highest cosine similarity to the keyword in noun phrase j is selected as the matched word and the similarity score between them is regarded as the affinity value A_{ij}^{lin} between the body part category i and noun phrase j. Then we normalize the affinity matrix $\mathbf{A}^{lin}$ along the noun phase dimension (i.e., column dimension) to obtain the attention matrix $\widetilde{\mathbf{A}}^{lin} \in \mathbb{R}^{5\times N}$ as:

$$\widetilde{\mathbf{A}}^{lin} = \mathrm{softmax}\left(\mathbf{A}^{lin}\right). \tag{5}$$

Note that we do not introduce the scale factor like in (4) because the cosine similarity has performed a normalization on the dot product of two feature vectors which counteracts the effect of the feature dimension.

We then aggregate the N noun phrase features $\mathbf{P} = [\mathbf{p}_1; \cdots; \mathbf{p}_N] \in \mathbb{R}^{N\times 512}$ using the attention matrix $\widetilde{\mathbf{A}}^{seg}$ or $\widetilde{\mathbf{A}}^{lin}$ to get the noun phrase features $\widetilde{\mathbf{P}} \in \mathbb{R}^{5\times 512}$ aligned to 5 body part categories as:

$$\widetilde{\mathbf{P}} = \widetilde{\mathbf{A}}^{seg/lin}\mathbf{P}, \tag{6}$$

where the i-th row of $\widetilde{\mathbf{P}}$ denotes the noun phrase feature attended by the i-th body part. The attended noun phrase features $\widetilde{\mathbf{P}}$ are finally projected to the fine-grained textual representations $\mathbf{Z}^f \in \mathbb{R}^{5\times 512}$ in a shared visual-textual space using a learnable weight matrix $\mathbf{W}_z^f \in \mathbb{R}^{512\times 512}$ as:

$$\mathbf{Z}^f = \widetilde{\mathbf{P}}\mathbf{W}_z^f. \tag{7}$$

3.3. Coarse-grained Representation

To further capture the spatial relations between fine-grained semantic components, we introduce the external knowledge of human body structure constraints and exploit the GCNs [19] to aggregate fine-grained representations into coarse-grained representations. Specifically, considering the fine-grained visual representations $\mathbf{X}^f \in \mathbb{R}^{5\times 512}$ or textual representations $\mathbf{Z}^f \in \mathbb{R}^{5\times 512}$ as graph nodes, we build an undirected graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$ to model the relationship between nodes based on the connections between body parts as shown in Fig. 3. For example, the head node is connected to the upper garment node but is disconnected to the shoe node. We then apply the graph convolution to enhance the fine-grained representations as:

$$\mathbf{Y}_x = \sigma\left(\mathbf{Q}\mathbf{X}^f\mathbf{W}_x^q\right), \quad \mathbf{Y}_z = \sigma\left(\mathbf{Q}\mathbf{Z}^f\mathbf{W}_z^q\right), \tag{8}$$

where σ is a non-linear activation function. $\mathbf{W}_x^q, \mathbf{W}_z^q \in \mathbb{R}^{512\times 512}$ are learnable weight matrices and the superscript q indicates the weight matrix of GCNs. The adjacency matrix $\mathbf{Q} \in \mathbb{R}^{5\times 5}$ is a 0–1 binary matrix defined based on the graph edge connections. We perform such graph convolution 2 times and finally apply a fully connected layer to generate the 512-dimensional coarse-grained representation $\mathbf{x}^c$, $\mathbf{z}^c \in \mathbb{R}^{512}$, respectively.

3.4. Multi-grained representations learning

To learn multi-grained visual and textual representations, we adopt the cross-modal projection matching (CMPM) loss and cross-modal projection classification (CMPC) loss [7], which incorporate the cross-modal feature projection operation and are demonstrated to be effective for learning joint visual and textual embeddings.

First, we briefly review the formulations of CMPM and CMPC loss [7]. Given a mini-batch with n pairs of visual images and textual descriptions, we denote the visual representation of each image as $\mathbf{x}_i \in \mathbb{R}^d$ and the textual representation of each description as $\mathbf{z}_j \in \mathbb{R}^d$, where d denotes the dimension of representations. For each $\mathbf{x}_i$, we construct the cross-modal pairs as $\{(\mathbf{x}_i, \mathbf{z}_j), y_{ij}\}_{j=1}^n$, where $y_{ij} = 1$ indicates that $(\mathbf{x}_i, \mathbf{z}_j)$ is a matched pair while $y_{ij} = 0$ means that they are unmatched. The probability of $\mathbf{z}_j$ being a match of $\mathbf{x}_i$ is defined as:

$$p_{i,j} = \frac{\exp\left(\mathbf{x}_i^\top \overline{\mathbf{z}}_j\right)}{\sum_{k=1}^{n}\exp\left(\mathbf{x}_i^\top \overline{\mathbf{z}}_k\right)} \quad s.t.\ \overline{\mathbf{z}}_j = \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}, \tag{9}$$

where $\overline{\mathbf{z}}_j$ denotes the normalized textual representation. Since there might be more than one matched descriptions for $\mathbf{x}_i$ in a mini-batch, the ground-truth matching probability of $(\mathbf{x}_i, \mathbf{z}_j)$ is normalized as:

$$q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{n} y_{i,k}}. \tag{10}$$

Then the matching loss $\mathscr{L}_{i2t}$ from image to text in a mini-batch (i.e., given a query image and retrieve its matched textual description in the mini-batch) is formulated to minimize the KL divergence from distribution $\mathbf{q}_i$ to $\mathbf{p}_i$ as:

$$\begin{aligned}\mathscr{L}_{i2t} &= \frac{1}{n}\sum_{i=1}^{n} KL\left(\mathbf{p}_i \| \mathbf{q}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{i,j}\log\frac{p_{i,j}}{q_{i,j}+\epsilon},\end{aligned} \tag{11}$$

where ϵ is a small number to avoid numercial problems. Similarly, the matching loss $\mathscr{L}_{t2i}$ from text to image in a mini-batch (i.e., given a query description and retrieve its matched image in the mini-batch) can be defined by exchanging $\mathbf{x}$ and $\mathbf{z}$ in (9)–(11). Finally, the CMPM loss is computed by:

$$\mathscr{L}_{cmpm} = \mathscr{L}_{i2t} + \mathscr{L}_{t2i}. \tag{12}$$

Instead of traditional classification loss (e.g., softmax loss) which aims to categorize the visual and textual representations independently, the CMPC loss integrates the cross-modal projection into the norm-softmax loss and attempts to classify the vector projection of representations from one modality onto another. The loss to classify the projection of visual representations onto the corresponding textual representations is defined as:

$$\mathscr{L}_{ipt} = \frac{1}{n}\sum_{i=1}^{n} -\log\left(\frac{\exp\left(\mathbf{W}_{y_i}^\top \widehat{\mathbf{x}}_i\right)}{\sum_{j=1}^{n}\exp\left(\mathbf{W}_j^\top \widehat{\mathbf{x}}_i\right)}\right) s.t.\ \|\mathbf{W}_j\| = 1,\ \widehat{\mathbf{x}}_i = \mathbf{x}_i^\top \overline{\mathbf{z}}_i \cdot \overline{\mathbf{z}}_i, \tag{13}$$

where y_i denotes the class label of $\mathbf{x}_i$. $\mathbf{W}_{y_i}$ and $\mathbf{W}_j$ indicate the y_i-th and j-th column of weight matrix $\mathbf{W}$ for the last fully connected layer. The vector $\widehat{\mathbf{x}}_i$ denotes the projection of the visual representation $\mathbf{x}_i$ onto the normalized textual representation $\overline{\mathbf{z}}_i$. Similarly, the loss $\mathscr{L}_{tpi}$ to classify the projection of textual representations onto the visual representations is formulated by exchanging $\mathbf{x}$ and $\mathbf{z}$ in (13). The final CMPC loss is

computed by:

$$\mathscr{L}_{cmpc} = \mathscr{L}_{ipt} + \mathscr{L}_{tpi}. \tag{14}$$

More details of the CMPM and CMPC loss can be found in [7].

For the global-grained representation learning, we adopt both the CMPM and CMPC losses. Each training mini-batch are constructed with n pairs of global-grained visual and textual representations $\{\mathbf{x}_i^g\}_{i=1}^n$, $\{\mathbf{z}_j^g\}_{j=1}^n$. Thus, the CMPM loss $\mathscr{L}_{cmpm}^g$ and CMPC loss $\mathscr{L}_{cmpc}^g$ for global-grained representation learning can be formulated by exchanging $\mathbf{x}_i, \mathbf{z}_j$ with $\mathbf{x}_i^g, \mathbf{z}_j^g$ in (9)–(14). And the total loss $\mathscr{L}^g$ for global-grained representation learning is defined as:

$$\mathscr{L}^g = \mathscr{L}_{cmpm}^g + \mathscr{L}_{cmpc}^g. \tag{15}$$

Similarly, for the coarse-grained representation learning, we also adopt both the CMPM and CMPC losses fomulated by exchanging $\mathbf{x}_i, \mathbf{z}_j$ with coarse-grained representation samples $\mathbf{x}_i^c, \mathbf{z}_j^c$ in (9)–(14). The total loss $\mathscr{L}^c$ for coarse-grained representation learning is then defined as:

$$\mathscr{L}^c = \mathscr{L}_{cmpm}^c + \mathscr{L}_{cmpc}^c. \tag{16}$$

For the fine-grained representation learning, we adopt only the CMPM loss. We reshape each $\mathbf{X}_i^f, \mathbf{Z}_j^f \in \mathbb{R}^{5\times512}$ into vector $\mathbf{x}_i^f, \mathbf{z}_j^f \in \mathbb{R}^{2560}$ and obtain the CMPM loss by exchanging $\mathbf{x}_i, \mathbf{z}_j$ with $\mathbf{x}_i^f, \mathbf{z}_j^f$ in (9)–(12). The loss for fine-grained representation learning is then defined as:

$$\mathscr{L}^f = \mathscr{L}_{cmpm}^f. \tag{17}$$

During training, we adopt a three-step strategy. First, we only train the global-grained branch with loss $\mathscr{L}^g$ for 20 epochs. Second, we fix parameters of the ResNet-50 and jointly train the fine-grained and coarse-grained branches with loss $\mathscr{L}^f + \mathscr{L}^c$ for 20 epochs. Finally, we jointly train the three branches with loss $\mathscr{L}^g + \mathscr{L}^f + \mathscr{L}^c$ until the loss converges.

During testing, the similarity score between a visual image sample i and a textual description sample j is defined as:

$$S = \cos\left(\mathbf{x}_i^g, \mathbf{z}_j^g\right) + \cos\left(\mathbf{x}_i^f, \mathbf{z}_j^f\right) + \cos\left(\mathbf{x}_i^c, \mathbf{z}_j^c\right), \tag{18}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity between two vectors. We then rank the similarity scores to retrieve the matched person image based on the query description.

4. Experimental Results

4.1. Dataset and Implementation Details

We evaluate the proposed method on the CUHK-PEDES [9] dataset. It is currently the *only* large-scale dataset for person search based on textual descriptions. It is split into a training set with 34,054 images containing 11,003 identities, a validation set with 3,078 images containing 1,000 identities, and a test set with 3,074 images containing 1,000 identities. Each image has at least two descriptions. We use the Recall@K (K = 1, 5, 10) to evaluate the retrieval performance, which denotes the percentage of query descriptions where at least one ground-truth person image is retrieved among the top-K results.

The proposed method is implemented on an NVIDIA Geforce GTX 1080 GPU. For human parsing, we use the universal trained model of Graphonomy [18]. We remove the background and group the rest of body part segmentations into 5 categories as shown in Fig. 3. To stabilize the segmentation results, we perform the inference by averaging predictions of multi-scale inputs with the scale from 1.75 to 2.25 in increments of 0.25. We multiply 0.5 to the background channel of the confidence map before applying the *argmax* operation to generate the segmentation mask. We find this operation helps to generate better segmentation results on person images with low resolution in the CUHK-PEDES dataset. For description-based person search, we resize all input person images to 224 × 224 and use the crawl-300d-2 M FastText model [41] for word embeddings extraction. Each mini-batch contains 16 image-description pairs during training. The CNN to extract the mask feature vectors $\widehat{\mathbf{M}}$ contains 4 convolutional layers with 3 × 3 filters and a 256-dimensional fully connected layer. Each convolutional layer is followed by a max-pooling layer for downsampling. The numbers of filters for the 4 convolutional layers are 64, 64, 128, and 256 respectively. The weights of the pre-trained BERT model for text embedding extraction is fixed during training.

4.2. Comparison with State-of-the-art Methods

Table 1 shows our performance on the CUHK-PEDES dataset against state-of-the-art methods including GNA-RNN [9], IATV [10], PWM-ATH [11], GLA [35], Dual Path [8], CMPM + CMPC [7], MIA [12], PMA [13], and ViTAA [14]. Methods like the CMPM + CMPC [7], and Dual Path [8] only extract global representations from the whole image and description without considering the relevances between local visual and textual components. GNA-RNN [9], IATV [10], PWM-ATH [11], and GLA [35] design attention mechanisms to model relations between visual and textual features without explicitly locating and aligning corresponding visual and textual semantic components. MIA [12], PMA [13], and ViTAA [14] split images and descriptions into meaningful components and perform the cross-modal alignment between them. However, they align visual and textual components based on features with descriptive information which might interfere with the alignment. In our work, we adopt human parsing and linguistics parsing to decompose images and descriptions into fine-grained components related to different body parts. Unlike previous works, we exclude the descriptive information when generating the visual-textual attention which provides more reliable alignments between visual body parts and textual noun phrases. Moreover, to model the spatial relations between fine-grained components, we further build a body graph based on the external knowledge of human body structures and exploit GCNs [19] to aggregate fine-grained components into coarse-grained representations for visual-textual matching. As shown in Table 1, our approach performs favorably against the state-of-the-art methods, which demonstrate the effectiveness of the proposed multi-grained matching networks. In addition, our approach using the linguistics-based attention model (ours-linguistics) performs better than that adopting the segmentation-based attention model (ours-segment), which reflects that the linguistics-based attention mechanism provides a more reliable visual-textual alignment compared with the segmentation-based attention mechanism.

4.3. Ablation Studies

To demonstrate the contribution of each branch in the proposed multi-grained visual-textual matching framework, we evaluate the

Table 1
Comparison with state-of-the-art description-based person search methods on the CUHK-PEDES dataset.

Method	R@1	R@5	R@10
GNA-RNN [9]	19.05	-	53.64
IATV [10]	25.94	-	60.48
PWM-ATH [11]	27.14	49.45	61.02
GLA [35]	43.58	66.93	76.26
Dual Path [8]	44.40	66.26	75.07
CMPM + CMPC [7]	49.37	-	79.27
MIA [12]	53.10	75.00	82.90
PMA [13]	53.81	73.54	81.23
ViTAA [14]	55.97	75.84	83.52
Ours-segment	56.76	76.98	84.23
Ours-linguistics	**57.81**	**78.36**	**85.53**

matching performances of methods with three different configurations including global-grained only (G), global-grained + fine-grained (G + F), and global-grained + fine-grained + coarse-grained (G + F+C). As shown in Table 2, both fine-grained and coarse-grained branches make improvements to the matching performance when employing either the segmentation-based or linguistics-based attention mechanism. Compared with methods with only the global branch (G), adding the fine-grained branch (G + F) improves the R@1 by 3.93% and 5.32% when adopting the segmentation-based and linguistics-based attention mechanisms respectively. This is because the fine-grained branch performs the alignment between local visual and textual components, and exploits their local similarities for more accurate cross-modal matching. Compared with G + F, integrating the coarse-grained branch (G + F+C) further improves the R@1 by 2.03% and 1.69% when adopting the segmentation-based and linguistics-based attention mechanisms respectively. This is because the coarse-grained branch injects the external knowledge of human body structure into representations and eliminates the background interference, which enhances the discriminative power of visual and textual representations.

In the fine-grained branch, we design two kinds of attention mechanisms to aggregate noun phrase features based on their relevances with each body part for visual-textual alignment. Different from previous works, our attention modules are designed to exclude the descriptive information from interacted visual and textual features when generating the attention. The reason is that when applying the alignment between noun phrases and body parts, the descriptive information are redundant and even might interfere with the alignment. To show the effect of our proposed attention mechanisms, we evaluate the performances of methods using attention modules with descriptive information (WD) and without descriptive information (W/OD) for comparison. Each method is described as follows:

WD in the segmentation-based attention module: We use an attention module which exploits the interplay of fine-grained visual representations $\mathbf{X}^f$ and noun phrase features $\mathbf{P}$ to model the relevances between body parts and noun phrases. Both $\mathbf{X}^f$ and $\mathbf{P}$ encode extra descriptive information (e.g., color information).

W/OD in the segmentation-based attention module: We use an attention module which exploits the interplay of visual segmentation mask features $\widehat{\mathbf{M}}$ and the last noun word embeddings $\mathbf{E}$ in noun phrases to model the relevances between body parts and noun phrases. Both $\widehat{\mathbf{M}}$ and $\mathbf{E}$ contain the necessary information to decide corresponding body part categories while excluding redundant descriptive information (e.g., color information).

WD in the linguistics-based attention module: We use an attention module which exploits the interplay of noun phrases features $\mathbf{P}$ and word sets related to different body part categories to capture the relevances between noun phrases and body parts. The noun phrases contain adjectives and thus $\mathbf{P}$ encode extra descriptive information (e.g, color information).

W/OD in the linguistics-based attention module: We use an attention module which exploits the interplay of the last noun word embeddings $\mathbf{E}$ of noun phrases and word sets related to different body part categories to capture the relevances between noun phrases and body parts. The last noun word embeddings $\mathbf{E}$ contain the necessary information to decide corresponding body part categories while excluding redundant descriptive information (e.g., color information).

As shown in Table 3, compared to involving descriptive information (WD) in the interacted visual and textual features, our strategy of excluding descriptive information (W/OD) improves the R@1 by 2.13% and 1.59% when adopting the segmentation-based and linguistics-based attention mechanisms respectively. These experimental results demonstrate the effectiveness of our proposed attention mechanisms.

Table 2
Performances of methods with different configurations on the CUHK-PEDES dataset.

Attention Mechanism	Method	R@1	R@5	R@10
Segmentation-based	G	50.80	72.24	80.93
	G + F	54.73	75.68	82.86
	G + F+C	**56.76**	**76.98**	**84.23**
Linguistics-based	G	50.80	72.24	80.93
	G + F	56.12	76.32	83.95
	G + F+C	**57.81**	**78.36**	**85.53**

4.4. Attention Visualization

To further demonstrate the effectiveness of our proposed attention mechanisms, we visualize the attention score of each noun phrase with respect to each body part category in Fig. 4. We can see that for each body part category, both the segmentation-based and the linguistics-based attention modules assign the highest attention score to the most related noun phrase. These results reflect that the proposed attention mechanisms make reasonable alignments between fine-grained visual and textual components. Compared with the segmentation-based attention modules, the linguistics-based attention model assigns higher weights to related noun phrases, which indicates that the linguistics-based similarity is more reliable than the cross-modal segmentation-based similarity for aligning noun phrases to corresponding body parts.

4.5. Running time analysis

In Table VIII, we present the average running time to extract visual and textual representations from each single image-text pair with three different configurations including global-grained only (G), global-grained + fine-grained (G + F), and global-grained + fine-grained + coarse-grained (G + F+C). We conduct experiments on a platform with an 4-core Intel Xeon W-2123 CPU @3.6 GHz and an NVIDIA GeForce GTX 1080 Ti GPU. Adding the fine-grained branch increases the running time by 0.164 s and 0.159 s when adopting the segmentation-based and linguistics-based attention mechanisms respectively. The increased running time is mainly attributed to the human parsing network. Adding the coarse-grained branch has only a small effect which further increases the running time by about 0.004s. (Table 4).

5. Conclusion

In this paper, we propose a multi-grained matching framework with global-grained, fine-grained, and coarse-grained branches to address the description-based person search task. Besides extracting holistic visual and textual representations in the global-grained branch, we propose to adopt human parsing and linguistic parsing to accurately locate semantic components in both visual images and textual descriptions. Unlike previous works, we design two attention mechanisms which exclude redundant descriptive information from interacted features to perform a more reliable alignment between visual and textual semantic components. Furthermore, we inject the extern knowledge of human body structure by building a body graph and exploit GCNs to aggregate fine-grained semantic components into coarse-grained representations. Experimental results on the public description-based person search dataset demonstrate the effectiveness of our proposed multi-grained

Table 3
Comparison of attention mechanisms for visual-textual alignment with descriptive information (WD) and without descriptive information (W/OD).

Attention Mechanism	Method	R@1	R@5	R@10
Segmentation-based	WD	54.63	75.02	82.36
	W/OD	**56.76**	**76.98**	**84.23**
Linguistics-based	WD	56.22	76.64	83.87
	W/OD	**57.81**	**78.36**	**85.53**

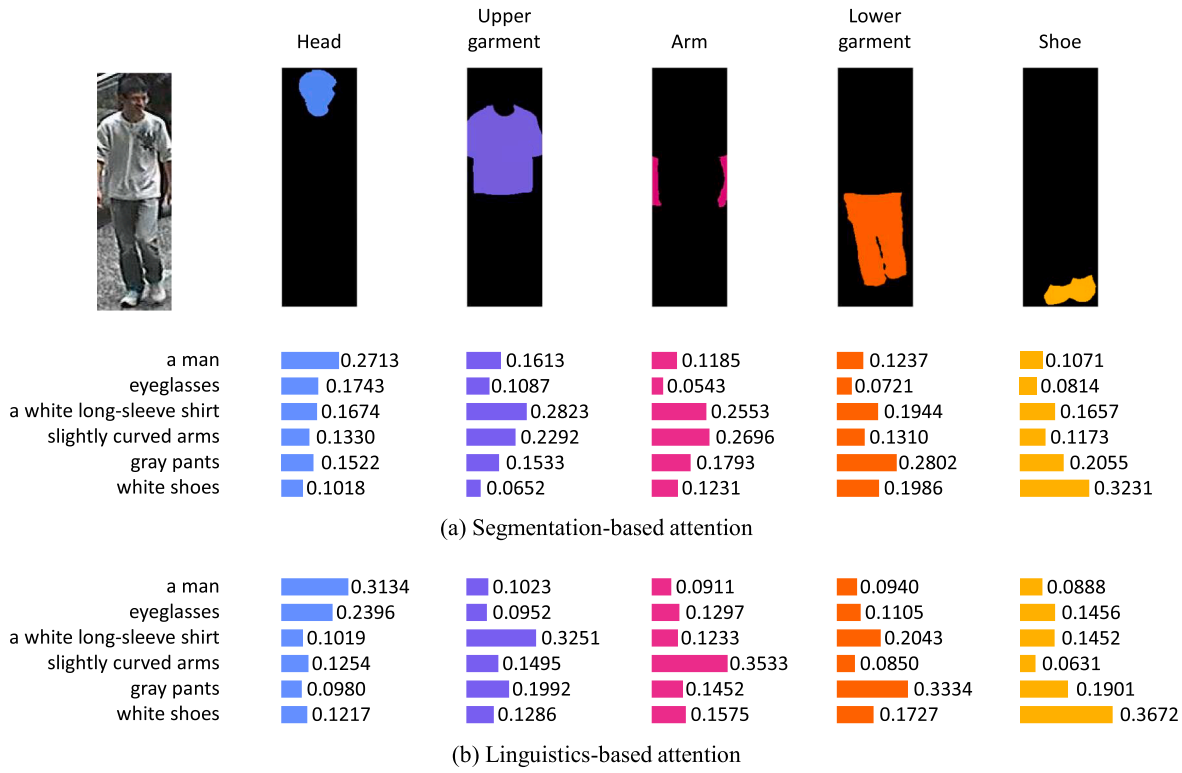


Fig. 4. Visualization of (a) segmentation-based and (b) linguistics-based attention values (best viewed in color).

Table 4
Running time of representation extraction with different model configurations.

Attention Mechanism	Method	Running Time
Segmentation-based	G	0.046 s
	G + F	0.210 s
	G + F+C	0.214 s
Linguistics-based	G	0.046 s
	G + F	0.205 s
	G + F+C	0.209 s

visual-textual matching approach against state-of-the-art methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

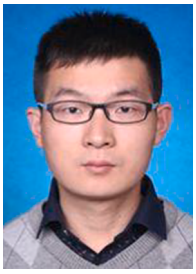
Acknowledgment

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant Nos. 61771303, 61771305), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 19DZ1209303, 20DZ1200203, 2021SHZDZX0102), and SJTU-Yitu/Thinkforce Joint Laboratory for Visual Computing and Application.

References

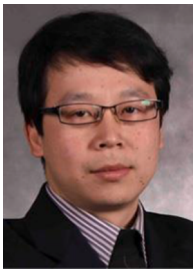
[1] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in, in: Proceedings of the Asia Conference on Computer Vision, 2012.
[2] W. Li, X. Wang, Locally aligned feature transforms across views, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[3] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[4] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1116–1124.
[5] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3774–3782.
[6] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3376–3385.
[7] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 707–723.
[8] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multimedia Comput. Commun. Appl. 16 (2020).
[9] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5187–5196.
[10] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1908–1917.
[11] T. Chen, C. Xu, J. Luo, Improving text-based person search by spatial matching and adaptive threshold, in, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1879–1887.
[12] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, IEEE Trans. Image Process. 29 (2020) 5542–5556.
[13] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided multi-granularity attention network for text-based person search, in: Association for the Advancement of Artificial Intelligence (AAAI), 2020, pp. 11189–11196.
[14] Z. Wang, Z. Fang, J. Wang, Y. Yang, Vitaa: Visual-textual attributes alignment in person search by natural language, in, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 402–420.
[15] X. Chen, L.J. Li, F.F. Li, A. Gupta, Iterative visual reasoning beyond convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7239–7248.
[16] W. Norcliffe-Brown, E. Vafeias, S. Parisot, Learning conditioned graph structures for interpretable visual question answering (2018) 8344–8353.
[17] C. Jiang, H. Xu, X. Liang, L. Lin, Hybrid knowledge routed modules for large-scale object detection (2018) 1559–1570.
[18] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, L. Lin, Graphonomy: Universal human parsing via graph transfer learning, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7442–7451.
[19] T. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in, in: Proceedings of the International Conference on Learning Representations, 2017.

[20] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in, in: Proceedings of the International Conference on Pattern Recognition, 2014, pp. 34–39.
[21] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[22] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[23] L. Chen, H. Yang, J. Zhu, Q. Zhou, S. Wu, Z. Gao, Deep spatial-temporal fusion network for video-based person reidentification, in: in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2017.
[24] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.
[25] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: in: Proceedings of the European Conference on Computer Vision, 2018, pp. 480–496.
[26] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1062–1071.
[27] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Densely semantically aligned person re-identification, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 667–676.
[28] S. Paisitkriangkrai, C. Shen, A. Van Den Hengel, Learning to rank in person re-identification with metric ensembles, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.
[29] S. Bak, P. Carr, Person re-identification using deformable patch metric learning, in, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–9.
[30] S. Bak, P. Carr, One-shot metric learning for person re-identification, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2990–2999.
[31] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, 2017 arXiv preprint arXiv:1703.07737.
[32] Z. Liu, D. Wang, H. Lu, Stepwise metric promotion for unsupervised video person re-identification, in, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2429–2438.
[33] H.X. Yu, A. Wu, W.S. Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, in, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 994–1002.
[34] I.K. Riza Alp Güler, Natalia Neverova, Densepose: Dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.
[35] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, X. Wang, Improving deep visual representation for person re-identification by global and local image-language association, in: in: Proceedings of the European Conference on Computer Vision, 2018, pp. 56–73.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[37] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F.F. Li, Imagenet: A large-scale hierarchical image database, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[38] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805 (2018).
[39] B. Steven, K. Ewan, L. Edward, Natural Language Processing with Python, O'Reilly Media Inc, 2009.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, u. Kaiser, I. Polosukhin, Attention is all you need (2017) 6000–6010.
[41] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for, Computational Linguistics 5 (2017) 135–146.



Ji Zhu received the B.E. degree from the School of Electronic Engineering, Xidian University, China, and is currently pursuing the Ph.D. degree at the Department of Electronic Engineering, Shanghai Jiao Tong University, China. He also works as a research scientist at Visbody. His research interests include computer vision, deep learning, and computer graphics.

Hua Yang received the Ph.D. degree in communication and information from Shanghai Jiaotong University, in 2004, and both the B.S. and M.S. degrees in communication and information from Haerbin Engineering University, China in 1998 and 2001, respectively. She is currently an associate professor in the Department of Electronic Engineering, Shanghai Jiaotong University, China. She received the first prize in Shanghai technical invention in 2017 and champion of wider person search as an advisor in ECCV2018. Her current research interests include computer vision, machine learning, and smart video surveillance applications.



Jia Wang received the B.Sc. degree in electronic engineering, the M.S. degree in pattern recognition and intelligence control, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 1997, 1999, and 2002, respectively. He is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, and also a member of the Shanghai Key Laboratory of Digital Media Processing and Transmission. His research interests include multiuser information theory and mathematics in artificial intelligence.

Wenjun Zhang received his B.S., M.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984, 1987 and 1989, respectively. After three years' working as an engineer at Philips in Nuremberg, Germany, he went back to his Alma Mater in 1993 and became a full professor of Electronic Engineering in 1995. He is the Chief Scientist of the Chinese Digital TV Engineering Research Centre (NERC-DTV) and the director of Cooperative MediaNet Innovation Center (CMIC). His main research interests include video coding and wireless transmission, multimedia semantic analysis, and broadcast/broadband network convergence.