Transformer based Language-Person Search with Multiple Region Slicing

Hui Li, *Student Member, IEEE,* Jimin Xiao, *Member, IEEE,* Mingjie Sun, Eng Gee Lim, *Senior Member, IEEE,* and Yao Zhao, *Senior Member, IEEE*

***Abstract*—Language-person search is an essential technique for applications like criminal searching, where it is more feasible for a witness to provide language descriptions of a suspect than providing a photo. Most existing works treat the language-person pair as a black-box, neither considering the inner structure in a person picture, nor the correlations between image regions and referring words. In this work, we propose a transformer-based language-person search framework with matching conducted between words and image regions, where a person picture is vertically separated into multiple regions using two different ways, including the overlapped slicing and the key-point-based slicing. The co-attention between linguistic referring words and visual features are evaluated via transformer blocks. Besides the obtained outstanding searching performance, the proposed method enables to provide interpretability by visualizing the co-attention between image parts in the person picture and the corresponding referring words. Without bells and whistles, we achieve the state-of-the-art performance on the CUHK-PEDES dataset with Rank-1 score of 57.67% and the PA100K dataset with mAP of 22.88%, with simple yet elegant design. Code is available on https://github.com/detectiveli/T-MRS.**

***Index Terms*—Transformer, Language-Person Search.**

I. Introduction

TRADITIONAL person re-identification (Re-ID) task aims to identify the target person among a gallery of person pictures, according to only one portrait image. The Re-ID task exhibits tremendous potential in a variety of applications, such as terrorist tracking and missing person identification. However, traditional image-based Re-ID is unsuitable for some emerging real-world Re-ID applications. For instance, in the task of criminal searching, it is arduous for a witness to provide a clear digital picture of the criminal, but to some extents, the witness can describe the criminal with language descriptions.

Thus, the language-person search task, where language descriptions are provided as natural references, is drawing increasing attention recently. However, due to the semantic gap between the linguistic domain and the visual domain, aligning and comparing feature representations for languages and images is a challenging task.

H. Li, J. Xiao, M. Sun and E. G. Lim are with the School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: hui.li02@xjtlu.edu.cn; jimin.xiao@xjtlu.edu.cn; mingjie.sun@liverpool.ac.uk; enggee.lim@xjtlu.edu.cn). (Corresponding author: Jimin Xiao)

Y. Zhao is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: yzhao@bjtu.edu.cn).

The work was supported by National Natural Science Foundation of China under 61972323, and Key Program Special Fund in XJTLU under KSF-T-02, KSF-P-02.

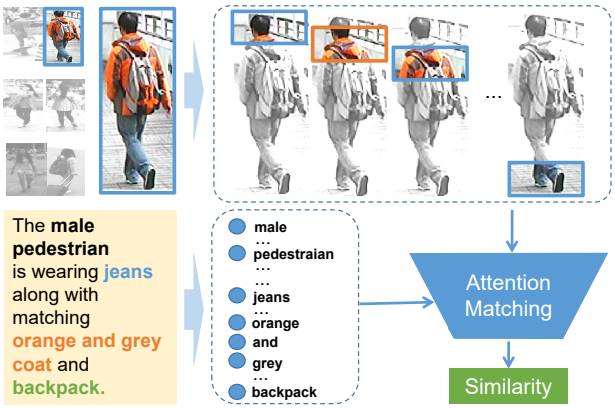

Fig. 1. The main workflow of our proposed method. In the language-person search task, a language description is taken as the natural reference to search for the target person from the image gallery. Different words in a sentence with different colors depict different vertical parts in a person image.

Many previous works attempt to tackle this task through three steps [1], [2], [3]. The first step is to generate the visual feature of the person image using a visual feature extractor and the linguistic feature of the person descriptions using a linguistic feature extractor. The second step is to align the features from two domains by mapping them to a common latent feature space. The final step is to calculate the similarity of the language-person pair according to the aforementioned features. However, most existing works match the sentence-image pair as a black-box, without considering the inner structure in a person picture that all human pictures contain the head, arms, legs and so on.

In this paper, we propose a transformer-based [4] language-person search framework with matching conducted between individual words and image regions, as shown in Fig. 1. Using the proposed framework, the original person image needs to be sliced into several regions. Existing person re-identification methods achieve great success by slicing a person image into equal-sized regions, such as [5], [6], [7]. However, using such a Vanilla Slicing (VS) method, one entirety of the human body might be sliced into multiple regions, which is not suitable for the language-based person search problem. Thus, we propose two new slicing methods which consider the structure of the human body, including the Overlapped Slicing (OS) method and the Key-point-based Slicing (KS) method. OS method maintains an overlap between neighbouring regions, while the

KS method slices the image regions based on the key-points of a human body [8]. It is worth noticing, OS and KS are complementary with each other, and can be jointly applied to further enhance the performance.

After the image slicing process, each image piece is extracted to obtain a visual feature. Accordingly, the corresponding language sentence is separated word by word to obtain the word features. Therefore, the correlation between each image piece and each word can be setup via the attention mechanism within the transformer blocks. For example, in Fig. 1, the descriptions of the cloth color and style matches the upper-body part of a person image.

Different from existing methods that work blindly, the proposed method can visualize the correlations between different parts in the person picture and the corresponding referring words, using a simple but powerful transformer model as the backbone. These visualization results provide interpretability for the final score. We believe such interpretability is quite important for applications like criminal searching, where searching results can be visualized with explanations. Meanwhile, the pre-training procedure based on the large image-text matching dataset using BERT can better align the visual-linguistic clues and benefit the language-person search task.

In summary, the contributions of the proposed method are listed as follows:

- To the best of our knowledge, it is the first work to introduce transformer blocks into language-person search, where the attentions between image parts and referring words are evaluated. Contrast to existing works that view the language-person pair as a black-box, interpretability is provided by visualizing the cross-attention between the human parts and the referring words.
- Different with general image-text matching task [9][10], where object detection methods, e.g., Faster RCNN [11], are used to generate object proposals and region features, we propose two new image region slicing methods to maintain the entirety of human body structure, including the overlapped slicing method and the key-point-based slicing method, which fully considers the intrinsic property of the language-person search task.
- We achieve the state-of-the-art language-person search performance on the CUHK-PEDES [1] dataset (Rank-1 of 57.67%) and the PA100K [12] dataset (mAP of 22.88%) using simple yet elegant design, without complex bells and whistles.

II. Related Work

A. Natural Language Embedding

The representation of a natural language sequence is a basic research area in natural language processing (NLP) community for a long time. The language embedding extractors evolve from Word2Vec [13], EIMO [14] to BERT [15]. Word2Vec [13] is the first one to provide a trainable word-to-vector mapping, which introduces the correlation distance into the embedded vector. To solve the multi-meaning problem in different contexts, EIMO [14] adopts LSTM [16] to represent a word by considering all previous words in the sentence. Thanks to the success of transformer [4], BERT [15] surprisingly boosts the performance on multiple NLP tasks by designing a parallel multi-head attention structure to fully utilize the correlations among words in the entire sentence.

B. Vision-Language Tasks

Vision-language tasks are attracting increasing attention from both computer vision (CV) [17], [18] and NLP communities [19], [20]. Such tasks involve a variety of real-world problems, including visual question answering (VQA) , image-text matching (ITM) [21], [22], referring expression grounding (REG) [23], [24] and so on. Many methods have been proposed to handle these tasks simultaneously. ViLBERT [9] modifies the standard encoder transformer block to a co-attention transformer layer, which is the first work to jointly consider the attention between texts and images. VL-BERT [10] proposes an end-to-end framework to train the parameters of the image feature encoder (F-RCNN [25]) and the attention parts in BERT. Video-BERT [26] applies BERT on video-level tasks, empowering BERT on both the video captioning and the video sequence generation. There are also some researches focus specific on image-text matching problem, like [27] match the semantic attention between regional objects with global context.

The image-text matching task focuses on the relationship between the objects in image regions and referring words, where the objects (e.g., cat, person, car) are significantly different from each other. However, the language-based person search task focuses on the correlation between the human parts and the words, where the persons' structure and body parts are much more similar to each other. Such difference makes the person pictures more difficult to be identified than natural images.

C. Language-Person Search

Many existing works perform well on traditional image-based person Re-ID [28], [29], [30], [31], [32], [33], [34], [35], [36]. Some works use the probe picture for the target person search [6], [7], [37]; some works align the original image and focus more on human body [38], which shares the similar principle in the face recognition task [39], [40], [41], [42], [43], [44], [45]. They separate a probe picture into equal-sized regions [6], [7] or based on the human key-point information [37], and such attempts prove to be effective for the image-based person Re-ID.

Language descriptions are used for the target person search in the language-person search task. Li et al. [1] establish the first benchmark in this task, utilizing LSTM to extract embedding from language descriptions. Sarafianos et al. [2] adopt the pre-trained BERT model to extract the initial language embedding before the matching process. Zhang et al. [3] propose a cross-modal projection method to minimize the KL divergence between linguistic and visual features. Niu et al. [5] slice the human image into equal-sized regions, and propose a multi-granularity image-text alignment method considering global-global, global-local and local-local level matching. Jing et al. [46] introduce the alignment network between words

and human-key-point confidence maps to reinforce the latent semantic alignment between noun phrases and image regions.

Different from existing methods, our proposed method attempts to utilize the co-attention between image regions and referring words based on the attention mechanism of transformer.

III. Methodology

A. Overview

To address the language-person search task, a human-part based end-to-end matching model is proposed in this work. As shown in Fig.2, it consists of three main steps.

1) Person images are slices into multiple regions using the proposed overlapped slicing method or key-point-based slicing method.
2) The visual features for sliced regions of a human picture are extracted, and the linguistic features of referring words in the language sentence are also extracted.
3) To unify the linguistic domain and visual domain into a latent feature space, the embedded features in a common latent space are generated.
4) Both visual and linguistic embedded features are fed into transformer blocks [15] in a parallel way, enabling the analysis of the correlations between human parts and referring words by visualization. The similarity score between the human picture and the referring sentence is calculated via an attention-based framework.

B. Feature Embedding Process

1) Part-based Visual Feature Generation: To extract the part-based visual feature from the original person image i, we need to vertically slice it into M_v bounding boxes $\{b_m\}_{m=1}^{M_v}$, where each b_m is defined as a 4-dim vector denoting the coordinates of top-left and bottom-right corners of the corresponding bounding box. For instance, in the **Vanilla Slicing (VS)** case, it is as follows:

$$b_m = [0, \frac{m-1}{M_v} i_H, i_W, \frac{m}{M_v} i_H], \tag{1}$$

where i_H and i_W are the height and width of a person image, respectively.

Overlapped Slicing (OS). To achieve a more precise word-to-region matching, we propose an overlapped image slicing method, which evolves from the previous vanilla method. We define the number of image regions with overlaps as $M_o = 2(M_v - 1) + 1$, where the two neighbouring regions have an overlap ratio of 0.5. Thus, the 4-dim bounding boxes $\{b_m\}_{m=1}^{M_o}$ for visual feature extraction is defined as:

$$b_m = [0, \frac{m-1}{2M_v} i_H, i_W, \frac{m-1}{2M_v} i_H + \frac{i_H}{M_v}]. \tag{2}$$

Key-point-based Slicing (KS). Instead of slicing the image into equal-sized parts, we also employ the key-point detector Openpose [8] to get human key points as reference for the image slicing process. The number of bounding boxes M_k is set to 6, including the blank space above the head, the head, upper-body, lower-body, feet and the blank space under feet. They are obtained based on five key points: the nose, the neck, the hip, the ankles and the toes from Openpose.

By using these bounding boxes based on key-points, each human structure part can be localized and processed separately inside a certain bounding box. For example, the clothes description can match the upper body, but the previous method has a large chance to cut the upper body into several pieces.

Visual Feature Extraction. The person image feature is denoted as $X^v = \{x_m^v\}_{m=0}^{M}$ ($M \in \{M_v, M_o, M_k\}$), which is composed of two components, including the region-level features and the global feature. For the region-level features $\{x_m^v\}_{m=1}^{M}$ ($M \in \{M_v, M_o, M_k\}$), each x_m^v consists of the geometry feature and the appearance feature from the m^{th} region. Specifically, the geometry feature of b_m is a 2048-dim vector encoded from the 4-dim b_m using Relation Network [47], and the appearance feature is generated from the image patch within b_m via the ROI pooling of faster RCNN [11]. In this way, the calculation of each region-level feature x_m^v can be defined as $\phi_v(i, b_m|\theta_v)$, where i is the current image and b_m is the corresponding bounding box, and θ_v is the learnable weight of ϕ_v.

The generation of global visual feature x_0^v is similar to the region-level feature. The only difference is that instead of adopting the bounding box of a region (b_m) inside the image, the bounding box b_0 ($b_0 = [0, 0, i_W, i_H]$) that covers the whole image is adopted to extract the geometry feature and appearance feature. The global visual feature x_0^v is defined as $x_0^v = \phi_v(i, b_0|\theta_v)$.

2) Linguistic Token Feature Generation: For a linguistic sentence $s = \{u_n\}_{n=1}^{N}$, say "a pedestrian with dark hair is wearing red and white shoe", u_n denotes the n^{th} token in s, and N is the length of the sentence. The feature of whole sentence s is denoted as $X^s = \{x_n^s\}_{n=1}^{N}$, with x_n^s corresponding to u_n, utilizing a linguistic feature extractor $\phi_s(u_n|\theta_s)$. Specifically, ϕ_s adopts a fixed word-to-vector mapping following WordPiece [48] with 30,522 vocabularies.

3) Unified Visual and Linguistic Embedding: To unify the linguistic domain and visual domain into a latent feature space, a cross-domain mapping process is adopted before the similarity matching. The mapping process consists of two parts: introducing the special linguistic feature into the visual feature and introducing the special visual feature into the linguistic feature.

To introduce special linguistic feature into the visual feature, the linguistic feature of a special token, "IMG", is added to the visual feature x^v for each region and the global image:

$$\hat{x}^v = w_v \cdot x^v + w_s \cdot x_{[IMG]}^s, \tag{3}$$

where w_v, as well as w_s, is the learnable weight, $x_{[IMG]}^s$ is the linguistic feature for the special token "IMG", and $\hat{x}^v$ is the mixed visual feature including certain linguistic information.

Similarly, to introduce special visual feature into the linguistic feature, the visual feature x_0^v for the entire image is added to the linguistic feature x^s for each token:

$$\hat{x}^s = w_v \cdot x_0^v + w_s \cdot x^s, \tag{4}$$

where $\hat{x}^s$ is the mixed linguistic feature for each token.

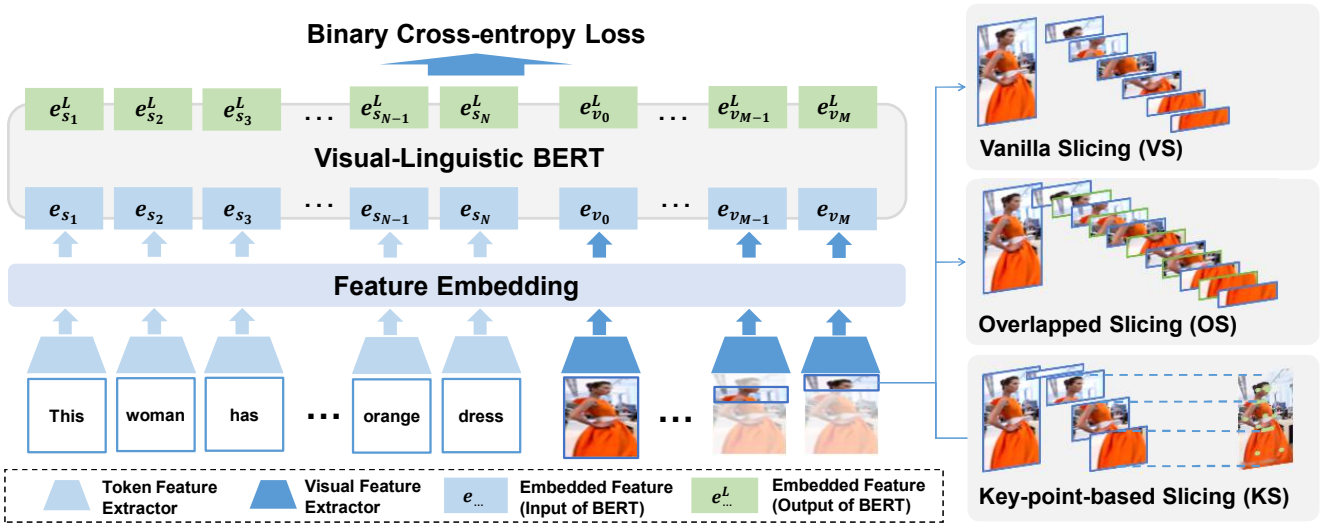

Fig. 2. Overview of the proposed language-person search framework. The main workflow of our framework is to generate the features of image-regions and words, unify the visual and linguistic features, and then feed the embedded features into a visual-linguistic BERT structure to predict the matching result. Specifically designed for our region-word matching method, we propose two new image slicing methods, including overlapped slicing (OS) method and the key-point-based slicing (KS) method, along with the original vanilla slicing (VS) method.

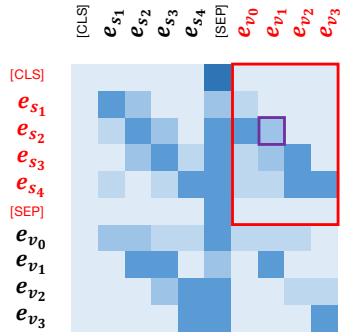

Fig. 3. Illustration of the attention matrix between embedded features. The color depth of each individual grid represents the correlation between each embedded feature pair (e.g., the grid in the small purple box).

Finally, Eq.(3) is used to evaluate all the region-level and global visual features, and all $\hat{x}^v$ compose the new visual feature set $\hat{X}^v$ for the entire image. Eq.(4) is used to evaluate all the linguistic features, and all $\hat{x}^s$ compose the new linguistic feature set $\hat{X}^s$ for the whole referring sentence. $[\hat{X}^s, \hat{X}^v]$ is combined with sequence position information as the input of the transformer blocks of BERT.

C. Attention-based Visualization Mechanism

The attention structure in BERT is utilized to compose the proposed visualization mechanism.

Original BERT consists of N_l layers, and each layer contains N_t transformer blocks, with N_h heads implemented inside each transformer block. Let $E^l = \{e^l_{s_1}, ..., e^l_{s_N}, e^l_{v_0}, e^l_{v_1}, ..., e^l_{v_M}\}$ be the input of l^{th} layer, with $\{e^l_{s_1}, ..., e^l_{s_N}\}$ being the embedded linguistic features and $\{e^l_{v_0}, e^l_{v_1}, ..., e^l_{v_M}\}$ being the embedded visual features. The number of transformer blocks, N_t, is the same as the number of elements in E^l, which is $N + M + 1$ in our case.

Specifically, for the first layer, E^1 is the initial embedded features, which is obtained by combining $[\hat{X}^s, \hat{X}^v]$ with sequence position information [1,2,..., $N + M + 1$], similar as in VL-BERT [10].

In BERT, multiple heads are used for attention evaluation. For layer l, head h, the attention matrix for the embedded features E^l is calculated as:

$$A^{l,h}_{E^l} = softmax[\frac{(W^{l,h}_Q \cdot E^l)(W^{l,h}_K \cdot E^l)^T}{\sqrt{d_k}}], \quad (5)$$

where $W^{l,h}_Q$ and $W^{l,h}_K$ are the learnable weight sets of query-generation model and key-generation model, respectively. d_k in is the default scale value from [10].

The attention matrix $A^{l,h}_{E^l}$ calculated according to Eq. (5) is with size $\mathbb{R}^{(N+M+1)\times(N+M+1)}$. Rather than adopting the entire $A^{l,h}_{E^l}$, a particular subset $\hat{A}^{l,h}_{E^l}$ (with size $\mathbb{R}^{(N)\times(M+1)}$) is used to visualize the correlation between each visual part and each linguistic token, as follows:

$$\hat{A}^{l,h}_{E^l} = \begin{bmatrix} A^{l,h}_{E^l}[1, \text{N+1}] & ... & A^{l,h}_{E^l}[1, \text{N+M+1}] \\ ... & ... & ... \\ A^{l,h}_{E^l}[\text{N}, \text{N+1}] & ... & A^{l,h}_{E^l}[\text{N}, \text{N+M+1}] \end{bmatrix}. \quad (6)$$

For better illustration, as showed in Fig.3, the grids in the red box correspond to $\hat{A}$ calculated according to Eq.(6).

As for the selection of the layer index l and head index h, motivated by [49], which provides sentence-sentence pair attention visualization on the NLP task, we choose the same layer index and the same head index as in [49] for our visual-sentence pair attention visualization.

D. Loss Function

In the training process, positive and negative pairs are organized to adapt to the language-person search task. A

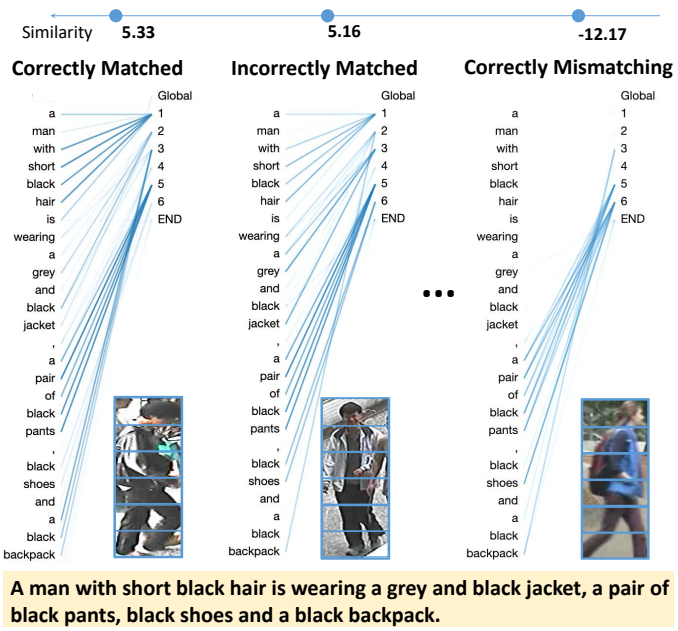

Fig. 4. Visualization of the co-attention for correct matching, wrong matching and correct mismatched images, when providing one language sentence. To make it clear, we choose the inference results from vanilla slicing method. The color depth represents the correlation value between human parts and referring words. Best viewed in color.

person appears in several images, and is described by multiple sentences. For a language-person pair $[i, s]$, if the image i and sentence s refer to the same person, $[i, s]$ is a positive language-person pair; otherwise, $[i, s]$ is a negative pair.

Let i^c ($i^c \in I^c$) be one image from the image group I^c with ID c. Followed the rule of the Re-ID, the positive sentences for the image i^c are $\{s^c_{k,j}\}^{k_c,j_k}_{k=1,j=1}$, where k_c is the number of images in class c , and j_k is the number of sentences for the k_{th} image.

The negative pair is defined as that the image and sentence are with different IDs. For a positive image sentence pair $[i^c, s^c_{k,j}]$, the corresponding negative pairs consist of two parts, the right image with negative sentence $[i^c, s^{\bar{c}}_{k',j'}]$ ($\bar{c} \neq c$), where k' and j' can be a random selection here; and the right sentence with negative image $[i^{\bar{c}}, s^c_{k,j}]$ ($\bar{c} \neq c$).

During the training process, one positive pair $[i^c, s^c_{k,j}]$, λ_{Neg} negative pairs $[i^c, s^{\bar{c}}_{k',j'}]$ ($\bar{c} \neq c$) , and λ_{Neg} negative pairs $[i^{\bar{c}}, s^c_{k,j}]$ ($\bar{c} \neq c$) are selected for each batch, where λ_{Neg} is a hyper-parameter to control the ratio of negative pairs.

Finally, the binary cross-entropy loss is adopted for each batch, as follows:

$$L = -\sum_{r=1}^{2\lambda_{Neg}+1} p_r(i,s)log(q_r(i,s)) + (1-p_r(i,s))log(1-q_r(i,s)), \quad (7)$$

where $q_r(i, s)$ refers to the prediction of r^{th} pair, and $p_r(i, s)$ refers to the ground truth label. $p_r(i, s) = 1$ when the pair is positive, and $p_r(i, s) = 0$ when the pair is negative.

TABLE I
LANGUAGE BASED PERSON RE-ID DATASET CUHK-PEDES AND ATTRIBUTE BASED PERSON RE-ID DATASET PA100K.

Parameters	CUHK-PEDES	PA100K
Image number	40206	100000
Train/Val/Test	34054/3078/3074	80000/10000/10000
Number of IDs	11003/1000/1000	2020/954/849
Description	Sentences	Attributes
Number of description	2	26

E. Application of Correlation

One promising application of the proposed framework is criminal investigation, which benefits a lot from the correlations for human part and referring word pairs. The main reason is that, given the language descriptions of the criminal from a witness, existing language-person search methods can only predict the matching score as a black-box, making the further investigation formidable when two suspects get similar matching scores. Whereas, the proposed model enables to access the correlation of human part and referring word pairs and predict the most distinct parts from both the person image and the linguistic sentence, which helps the witness recall the details of these crucial parts, and facilitates the identification of true criminal.

Specifically, as shown in Fig.4, when a witness provides a language description of the criminal with many common attributes, e.g., "A man with short black hair is wearing a grey and black jacket, a pair of black pants, black shoes and a black backpack.", it is arduous for conventional methods to identify the true criminal among all suspects, as the correct matching suspect and wrong matching suspect get quite similar matching scores. Nevertheless, with the correlations between human part and referring word pairs visualized, it can be observed that, "black backpack" turns crucial for the matching of the correct matching suspect. In this way, the police can ask the witness to recall the details about the backpack, which might be key to solve the case.

It is interesting to note that the proposed model can provide the correct correlations for human part and referring word pairs, even for a mismatched language-person pair. As shown in the final image of Fig.4, the current person is not the target person described by the referring sentence. However, the current person happens to wear similar pants, shoes and backpacks. Thus, the proposed model also predicts high correlations for these human part and referring word pairs, which demonstrates the robustness of the model.

IV. EXPERIMENTS

A. Experimental Settings

The proposed method is evaluated on two datasets for the language-person search task, including CUHK-PEDES [1] and PA100K [12]. The detailed information is listed in Table I. **CUHK-PEDES** is the first dataset for the language-person search task, containing 40,206 images, with 34,054/3,078/3,074 images for train/val/test, respectively. Images of CUHK-PEDES come from several widely-used image-based Re-ID datasets, such as Market1501 [60], DukeMTMC-reID [61] and CUHK01 [62]. Each image in CUHK-PEDES is

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE CUHK-PEDES DATASET. THE BOLD NUMBER REPRESENTS THE BEST SCORE, AND UNDERLINED NUMBER REPRESENTS THE SECOND BEST FROM PREVIOUS WORK.

Method	Visual Feature	CUHK-PEDES			
		rank-1	rank-5	rank-10	total
deeper LSTM [50] ICCV'15	global	17.19	-	57.82	-
GNA-RNN [1] CVPR'17	global	19.05	-	53.64	-
IATV [51] ICCV'17	global	25.94	-	60.48	-
PWM-ATH [52] WACV'18	global	27.14	49.45	61.02	137.61
GLA [53] ECCV'18	global	43.58	66.93	76.26	186.77
Dual Path [54] TOMM'17	global	44.40	66.26	75.07	185.73
CMPM-CMPC [3] ECCV'18	global	49.37	-	79.27	-
MIA [5] TIP'20	global+parts	53.10	75.00	82.90	211.00
A-GANet [55] ACMMM'19	proposals	53.14	74.03	81.95	209.12
PMA† [46] AAAI'20	global+key-points	53.81	73.54	81.23	208.58
TIMAM [2] ICCV'19	global	54.51	<u>77.56</u>	<u>84.78</u>	<u>216.85</u>
ViTAA† [56] ECCV'20	global+attribute	<u>55.97</u>	75.84	83.52	215.33
GARN [57] TIP'21	global+parts	52.75	74.36	81.85	208.96
Our method	global+OS	56.83	77.76	84.76	219.35
Our method†	global+OS+KS	**57.67**	**78.25**	**84.93**	**220.85**

† indicates the extra information is requested.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE PA100K DATASET. "ATT" DENOTES THE NUMBER OF ATTRIBUTES. THE BOLD NUMBER REPRESENTS THE BEST SCORE, AND THE UNDERLINED NUMBER REPRESENTS THE SECOND-BEST ONE.

Method	PA100K				
	att	rank-1	rank-5	rank-10	mAP
SCAN [58]	15	2.9	8.2	12.5	1.9
GNA-RNN [1]	15	20.3	30.8	38.2	9.3
CMCE [51]	15	25.8	34.9	45.4	13.1
AIHM [59]	15	**31.3**	<u>45.1</u>	51.0	<u>17.0</u>
CMCE [51]	26	18.14	42.52	52.65	10.88
CMPM-CMPC [3]	26	21.30	42.87	<u>53.24</u>	12.69
Our method	26	<u>26.73</u>	**48.88**	**60.30**	**22.88**

described by two referring sentences, with various vocabularies, phrases, and sentence structures. **PA100K** is an attribute-based Re-ID dataset, consisting of 100,000 images, where the ratio for the training, validation and test is 8:1:1. The main difference between PA100K and CUHK-PEDES is that the referring sentences in PA100K are not intact but a list of attributes, say whether the person has a hat or not. These attributes are viewed as a special kind of sentence to evaluate the proposed method, making the language-based Re-ID more difficult, as most attributes are not sufficiently discriminative to distinguish different persons.

The evaluation metric for the language-person search task remains the same as the traditional image-based Re-ID task, consisting of the top-k accuracy, and the mean Average Precision (mAP) score. Given one referring sentence, all test images are ranked according to the similarity score. For a successful search, all images of the corresponding person are ranked within top-k. The mAP score provides a wider view based on both precision and recall.

B. Implementation Details

The proposed method is implemented based on the pre-trained model of VL-BERT [10], which is trained on a large image-caption dataset Conceptual Captions [63] based on BERT [15] structure. We implement our method with BERT-small using 12 layers and 12 heads, and BERT-large using 24 layers and 16 heads, followed by fully connected layers to predict the matching scores. If it is not explicitly specified, the ensemble model from BERT-large and BERT-small is used. The visual feature extractor is ResNet-101 [64], with a 2048-D feature output. The linguistic feature extractor is WordPiece [48], whose glossary size is 30,522. The target dimension of weights w_v and w_s in Eq.(3) and Eq.(4) is set to 768.

To fine-tune the pre-trained model on the CUHK-PEDES and PA100K datasets, the first two stages of the ResNet-101 network are set frozen, with the remaining layers of ResNet-101 and all layers of BERT trainable. The optimizer is AdamW [65]. The total training epoch is set to 20. The learning rate is set to 1e-7, with a linear warm-up process on the first 20,000 steps. The dropout rate is 0.1 for all layers in BERT, and 0.5 for the fully-connected layers to predict matching scores.

The i_H and i_W for all images in the training set are 256 and 168, respectively. The hyper-parameter λ_{Neg} for negative pair control is set to 7, and the part number M_v is set to 6 for the VS method. For the key-point-based slicing method, we choose the 25-human-body-points model of Openpose [8] as reference for image cutting, which is pre-trained on COCO [66] plus foot-points.

The training process is conducted on a 24G TITAN RTX GPU, with a batch size of 8 for BERT-small and 5 for BERT-large. Apex [67] from Amp (Automatic Mixed Precision) is adopted for precision improvement and spatial assemble. Based on the aforementioned settings, where all the 24G GPU memory is used, the training process takes around 120 hours and 200 hours for BERT-small and BERT-large on the vanilla slicing method, respectively. All the obtained results have around (-0.2, 0.2) fluctuation with different seeds and random dropout. Thus, we used the averaged results over 5 trials.

C. Experimental Results

1) Comparison with State-of-the-Arts: In this section, the proposed method is compared with other state-of-the-art meth-

TABLE IV
ABLATION STUDY ON DIFFERENT REGION SLICING SETTINGS, CONDUCTED ON THE VALIDATION SET OF CUHK-PEDES. THE RESULTS ARE OBTAINED USING BERT-SMALL.

Settings	rank-1	rank-5	rank-10	total
global	44.18	68.27	77.59	190.04
global+VL-BERT	45.78	70.42	79.08	195.28
global+VS	53.03	74.45	82.58	210.06
global+KS	53.39	74.61	82.57	210.57
global+OS	54.21	75.24	82.73	212.18
global+VS+KS	55.46	75.84	83.56	214.86
global+OS+KS	**55.67**	**76.74**	**83.97**	**216.38**

TABLE V
ABLATION STUDY ON DIFFERENT BERT STRUCTURES, CONDUCTED ON THE VALIDATION SET OF CUHK-PEDES. THE REGION FEATURES ARE OBTAINED USING FASTER RCNN.

Settings	rank-1	rank-5	rank-10	total
ViL-BERT	36.19	60.74	70.81	167.74
VL-BERT	45.78	70.42	79.08	195.28

ods according to rank-1, rank-5, rank-10 and mAP on the CUHK-PEDES and PA100K datasets.

Table II reports the comparison results on the CUHK-PEDES [1] dataset. The proposed method achieves a new state-of-the-art rank-1 score of 57.67%, showing the effectiveness of the proposed method with only one simple matching loss. The results of "global+OS" are obtained using overlapped slicing method, where the results of "global+OS+KS" are obtained by the model-ensemble from overlapped slicing method and key-point-based slicing method. Compared with TIMAM [2], the best existing method without extra information, our "global+OS" with overlapped slicing method boots the performance by 2.32%. It is worth noticing that even compared with the latest method ViTAA [56], which uses extra annotations on visual and linguistic grouping, our method "global+OS+KS" improves the rank-1 score by 1.70%.

The proposed method is also evaluated on the PA100K [3] dataset, as reported in Table III. The compared benchmark methods are categorized into two groups. The first group are trained on the selected 15 attributes of PA100K dataset. Their performances are reported in AIHM [59], including SCAN [58], GNA-RNN [1], CMCE [51] and AIHM [59]. The second group consists of methods trained on the entire official 26 attributes of the PA100K dataset, including our repetition of CMPM-CMPC [3], CMCE [51], and our method. The difference between these two settings is that, the official 26 attributes might provide more information during training, but might also introduce more noise in some circumstances. Note that the proposed method is trained with the same settings as other methods in the second group. As shown in Table III, compared with methods in the second group, the proposed method achieves new state-of-the-art rank-1 score (26.73%), rank-5 score (48.88%), rank-10 score (60.30%), and mAP score (22.88%). In terms of comparison with methods in the first group, the proposed method also improves the mAP score significantly, demonstrating its robustness.

2) Ablation Study on Components: The ablation study results on different image slicing methods are reported in Table IV. "global" is obtained when the person image is not sliced. "global+VL-BERT" is the baseline method which generates proposals and region features based on BUA[68] the same as VL-BERT. "global+VS", "global+OS" and "global+KS" are obtained using different region slicing methods combined with the global box. As can be observed, if the entire image is adopted without any separated parts, "global" only achieves a rank-1 score of 44.18%. "global+VL-BERT" slightly improves the performance when image proposal features are introduced, which indicates that the proposals generated from the pre-trained Faster R-CNN are not perfectly suitable for

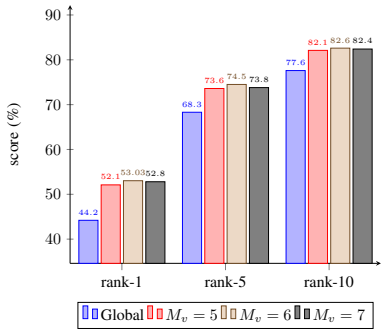

Fig. 5. Analysis on how the hyper-parameter M_v (number of the part in VS method) affects the performance, conducted on the validation set of CUHK-PEDES. The results are obtained using BERT-small.

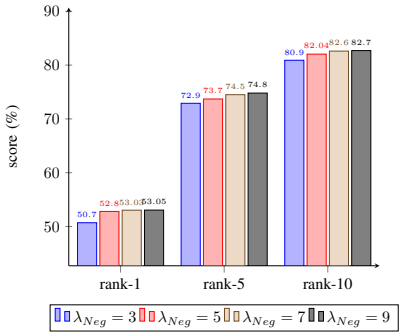

Fig. 6. Analysis on how the hyper-parameter λ_{Neg} affects the performance, conducted on the validation set of CUHK-PEDES. The results are obtained using BERT-small.

Fig. 7. Case study: qualitative comparison of our method with A-GANet [55] on top-10 matched images. The yellow boxes are correct matching from A-GANet, and blue boxes are the correct matching from our method. The different colors in the sentence represent different human parts. Best viewed in color.

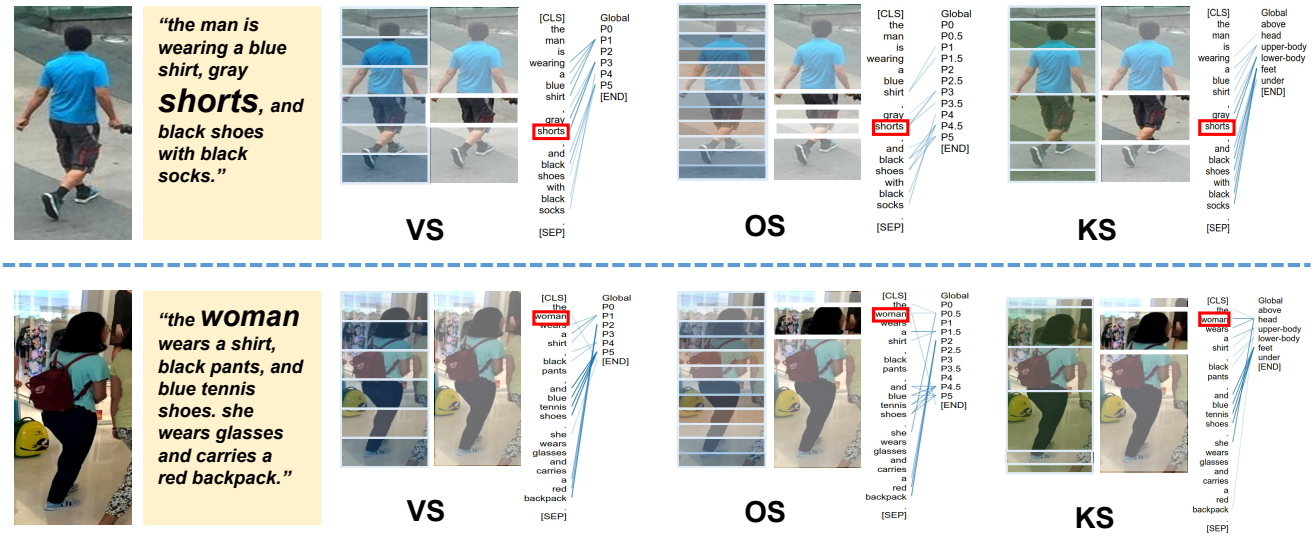

Fig. 8. Visualization of the co-attention for different slicing methods VS, OS, and KS, where one specific word from the sentence is provided ("**shorts**" for the upper case, "**woman**" for the lower case). For each method, the related parts and the correlation attention map to the referring word are highlighted. The color depth in attention map represents the correlation between human parts and the referring words.

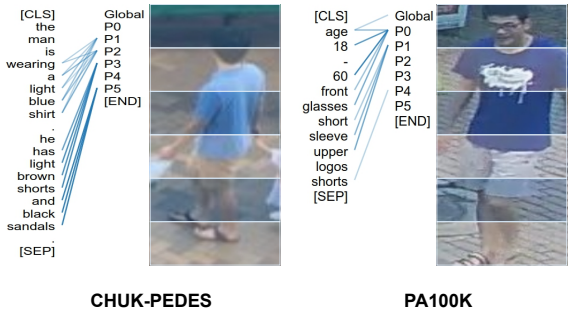

Fig. 9. Visualization of the co-attention for different datasets CUHK-PEDES [1] and PA100K [12] on similar images with vanilla slicing method. The color depth in attention map represents the correlation between human parts and the referring word. Best viewed in color.

the language-person search task. However, a 7.25% rank-1 improvement is obtained when separated parts (VS) are introduced, demonstrating that the vanilla slicing method better represents the human body structure than the aforementioned proposals to match the description sentence. Employing the proposed overlapped slicing method, our "global+OS" method outperforms "global+VS" by 1.18%, which indicates the effectiveness of taking overlapped regions in our transformer-based framework, where more precise attention connections between image parts and the words are obtained.

The results of "global+VS+KS" are obtained by model ensemble from "global+VS" and "global+KS". Similarly, "global+OS+KS" are obtained by model ensemble from "global+OS" and "global+KS". They achieve the rank-1 score of 55.46% and 55.67%, respectively. The results are higher than each individual model. It shows that overlapped slicing method and key-point-based slicing method are complementary, and can be jointly used. It also proves that, though "global+OS" and "global+KS" have limited performance improvements compared to the classical vanilla slicing method, the fused methods of different slicing methods are complementary to each other and boost the final performance for the language-person search task.

We also conduct an ablation study of the performances on different BERT structures (i.e., VL-BERT and ViL-BERT), evaluated on the CUHK-PEDES dataset. As reported in Table V, "ViL-BERT" achieves the rank-1 score of 36.19%. When comparing "VL-BERT" with "ViL-BERT", all ranking scores are dramatically increased, and the rank-1 performance improves to 45.78%. It proves that "VL-BERT" has better feature representation in this task, since "VL-BERT" optimizes the Faster R-CNN module simultaneously with the BERT during the training process, while "ViL-BERT" uses a fix a pre-trained Faster R-CNN module.

3) Analysis on Hyper-parameter: The first analysis is about how the number of image parts M_v from the vanilla slicing (VS) method affects the performance. As can be observed from Fig. 5, without any separated parts, our method achieves rank-1 of 44.2% in our framework. However, the performance is improved when separated parts are introduced, and the proposed setting with $M_v = 6$ performs the best among all settings, demonstrating the the necessity of the separated parts for the person images. When $M_v = 6$, it reaches a good balance, because too few slicing parts cannot cover different human body structures, while too many slicing parts may split one body part into multiple pieces. A similar result of the number of parts is reported from PCB [6], for traditional person re-identification task.

The second analysis is on the effect of λ_{Neg}, which controls the ratio between negative pairs and positive pairs. As shown in Fig. 6, the accuracy of Rank-1 score increases from 50.7%, 52.8%, 53.03% to 53.05%, when λ_{Neg} rises from 3, 5, 7 to 9 respectively. By increasing the number of negative pairs, each training batch contains more situations and promote the variety

of negative pairs. But the effectiveness tends to be saturated when $\lambda_{Neg} = 7$, because the random selection process is able to cover most situations of possible negative pair selections.

4) Qualitative Analysis: Some qualitative comparison results on the CUHK-PEDES [1] dataset are reported in Fig. 7. The top-10 results predicted by our proposed method and A-GANet [55] are visualized for two referring sentences. In terms of the first case, though both methods' predicted results (rank-1, rank-2, rank-3) are correct, our proposed method enables to identify and focus on the most crucial attribute "stripe". For the second case, our proposed method is able to predict correct results from rank-1 to rank-4, but A-GANet fails in Rank-3 and Rank-4, which shows the outstanding searching ability of our method.

Meanwhile, in Fig. 8, we provide two more sets of visualization results to illustrate different slicing mechanisms. In terms of the first case, we focus on the word "shorts" in the referring sentence. All three methods connect the "shorts" with the image part reasonably. It is worth to notice that with the OS method, two image parts with overlaps are activated, which enhances the connection between word "shorts" and the image region. For the second case, we focus on the word "woman". From the visualization results, the VS method fails to identify the related image region, because the hair of the woman is separated into two parts, making it less recognizable. However, for the KS method, the word "woman" has strong attention on the head part of the woman, which indicates that a precise slicing method is necessary for the language-person search task.

In Fig. 9, we provide the visualization results to compare PA100K with CUHK-PEDES. As can be observed, the transformer block can match the human parts with corresponding attributes in PA100K dataset, though it lacks continuous language information (e.g., is, has) and detailed attributes (e.g., light, brown).

V. Conclusion

In this paper, we have proposed an image region based language-person search framework, which separates the person picture into several regions to match the linguistic referring words with visual-language co-attention. Our framework is able to learn robust correlations between image parts and referring words, which helps for the language-person search task. Furthermore, unlike conventional methods that conduct the language-person search as a black-box, the proposed framework enables interpretability by visualizing the co-attention between image parts and referring words. Using the proposed language-person search framework, we also achieved state-of-the-art performance on two public datasets.

References

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *The European Conference on Computer Vision (ECCV)*, 2018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving description-based person re-identification by multi-granularity image-text alignments," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 5542–5556, 2020.

[6] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *The European Conference on Computer Vision (ECCV)*, 2018.

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 43, no. 1, pp. 172–186, 2019.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[10] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations (ICLR)*, 2020.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[12] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[14] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] F. Zhao, J. Zhao, S. Yan, and J. Feng, "Dynamic conditional networks for few-shot learning," in *The European Conference on Computer Vision (ECCV)*, 2018.

[18] J. Zhao, J. Li, F. Zhao, S. Yan, and J. Feng, "Marginalized CNN: Learning deep invariant representations," in *The British Machine Vision Conference (BMVC)*, 2017.

[19] W. Zhang, C. Ma, Q. Wu, and X. Yang, "Language-guided navigation via cross-modal grounding and alternate adversarial learning," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, pp. 1–1, 2020.

[20] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 29, no. 9, pp. 2822–2832, 2018.

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," *arXiv*, 2014.

[23] M. Sun, J. Xiao, and E. G. Lim, "Iterative shrinking for referring expression grounding using deep reinforcement learning," *arXiv*, 2021.

[24] M. Sun, J. Xiao, E. G. Lim, S. Liu, and J. Y. Goulermas, "Discriminative triad matching and reconstruction for weakly referring expression grounding," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, pp. 1–1, 2021.

[25] R. Girshick, "Fast R-CNN," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[26] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation

learning," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[27] K. Wen, X. Gu, and Q. Cheng, "Learning dual semantic relations with graph attention for image-text matching," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, pp. 1–1, 2020.

[28] Z. Jian, "Deep learning for human-centric image analysis," *Ph.D. dissertation*, 2018.

[29] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan, "Self-supervised neural aggregation networks for human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[30] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng, "Multiple-human parsing in the wild," *arXiv*, 2017.

[31] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*, 2018.

[32] J. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Fine-grained multi-human parsing," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 8, pp. 2185–2203, 2020.

[33] J. Li, J. Zhao, C. Lang, Y. Li, Y. Wei, G. Guo, T. Sim, S. Yan, and J. Feng, "Multi-human parsing with a graph-based generative adversarial model," *The Journal of the ACM*, vol. 37, no. 4, 2020.

[34] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "IAN: the individual aggregation network for person search," *Pattern Recognition*, vol. 87, pp. 332–340, 2019.

[35] H. Li, J. Xiao, M. Sun, E. G. Lim, and Y. Zhao, "Progressive sample mining and representation learning for one-shot person re-identification," *Pattern Recognition*, vol. 110, p. 107614, 2021.

[36] D. Zheng, J. Xiao, K. Huang, and Y. Zhao, "Segmentation mask guided end-to-end person search," *arXiv*, 2019.

[37] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 29, no. 10, pp. 3037–3045, 2018.

[39] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-Aided dual-agent GANs for unconstrained face recognition," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 41, no. 10, pp. 2380–2394, 2018.

[40] J. Zhao, L. Xiong, J. Karlekar, J. Li, F. Zhao, Z. Wang, S. Pranata, S. Shen, S. Yan, and J. Feng, "Dual-agent GANs for photorealistic and identity preserving profile face synthesis." in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[41] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, "Towards pose invariant face recognition in the wild," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, F. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition," in *The AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[43] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan, and J. Feng, "Multi-prototype networks for unconstrained set-based face recognition," in *The International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[44] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 2, pp. 460–478, 2020.

[45] J. Zhao, S. Yan, and J. Feng, "Towards age-invariant face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2020.

[46] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search." in *The AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[47] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[48] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv*, 2016.

[49] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2019.

[50] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual question answering," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[51] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

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

[54] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y. Shen, "Dual-path convolutional image-text embeddings with instance loss," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2017.

[55] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, 2019.

[56] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "ViTAA: Visual-textual attributes alignment in person search by natural language," in *The European Conference on Computer Vision (ECCV)*, 2020.

[57] Y. Jing, W. Wang, L. Wang, and T. Tan, "Learning aligned image-text representations using graph attentive relational network," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 1840–1852, 2021.

[58] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *The European Conference on Computer Vision (ECCV)*, 2018.

[59] Q. Dong, S. Gong, and X. Zhu, "Person search by text attribute query as zero-shot learning," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[60] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[61] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[62] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *The Asian Conference on Computer Vision (ACCV)*, 2012.

[63] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2018.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[65] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[66] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *The European Conference on Computer Vision (ECCV)*, 2014.

[67] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *International Conference on Learning Representations (ICLR)*, 2018.

[68] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Hui Li received the B.S. degree in the automation from the University of Chongqing, Chongqing, PR China, in 2015, and obtained the M.S. degree in the Electronic and Computer Engineering from the University of Birmingham, Birmingham, U.K., in 2017. After that, she was with the IDriverPlus Auto-driving company, Beijing, from 2017 to 2018. She is now a Ph.D. student in the Department of the Electrical and Electronic Engineering of the Xi'an Jiaotong-Liverpool University, Suzhou, PR China. Her current research interests include computer vision, person re-identification, and semi-supervised learning.

Jimin Xiao received the B.S. and M.E. degrees in telecommunication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electrical engineering and electronics from the University of Liverpool, Liverpool, U.K., in 2013. From 2013 to 2014, he was a Senior Researcher with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and an External Researcher with the Nokia Research Center, Tampere. Since 2014, he has been a Faculty Member with Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include image and video processing, computer vision, and deep learning.

Mingjie Sun received the B.S. degree in the telecommunication engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016, and the M.E. degree in the computer science from the Xi'dian University, Xian, China, in 2019. He is now a Ph.D. student in the Department of the Electrical and Electronic Engineering of the Xi'an Jiaotong-Liverpool University, Suzhou, PR China. His current research interest is video object segmentation and visual understanding.

Eng Gee Lim received the Ph.D. degree from the University of Northumbia, in 2002. He is currently a Professor with the Department of Electrical and Engineering, Xian Jiaotong-Liverpool University, Suzhou, China. His research interests include artificial intelligence, antennas, RF and radio propagation for wireless communications and systems.

Yao Zhao received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996, where he became an Associate Professor and a Professor in 1998 and 2001, respectively. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. In 2015, he visited the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland. From 2017 to 2018, he visited University of Southern California. He is currently the Director with the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, video analysis and understanding, and artificial intelligence. Dr. Zhao is a Fellow of the IET. He serves on the Editorial Boards of several international journals, including as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, a Senior Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor for Signal Processing: Image Communication. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010 and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013.