Asymmetric Cross-Scale Alignment for Text-Based Person Search

Zhong Ji, Senior Member, IEEE, Junhua Hu, Deyin Liu, Lin Yuanbo Wu, Senior Member, IEEE, and Ye Zhao

Abstract—Text-based person search (TBPS) is of significant importance in intelligent surveillance, which aims to retrieve pedestrian images with high semantic relevance to a given text description. This retrieval task is characterized by both modal heterogeneity and fine-grained matching. To implement this task, one needs to extract multi-scale features from both image and text domains, and then perform the cross-modal alignment. However, most existing approaches only consider the alignment confined at their individual scales, e.g., an image-sentence or a region-phrase scale. Such a strategy adopts the presumable alignment in feature extraction, while overlooking the cross-scale alignment, e.g., image-phrase. In this paper, we present a transformer-based model to extract multi-scale representations, and perform Asymmetric Cross-Scale Alignment (ACSA) to precisely align the two modalities. Specifically, ACSA consists of a global-level alignment module and an asymmetric cross-attention module, where the former aligns an image and texts on a global scale, and the latter applies the cross-attention mechanism to dynamically align the cross-modal entities in region/image-phrase scales. Extensive experiments on two benchmark datasets CUHK-PEDES and RSTPReid demonstrate the effectiveness of our approach.

Index Terms—Text-based person search, Transformer, Cross-modal matching.

I. INTRODUCTION

TEXT-BASED Person Search (TBPS) aims to retrieve the shots of a target person with high semantic relevance to a given text description. It has attracted increasing attention due to the wide applications in modern cities wherein a large number of monitoring devices are deployed. Compared to image query based approaches [1], [2], [3], TBPS only requires a verbal description to query a target person. This setting is more practical in certain situations where the image query may not be accessible. Moreover, natural language can describe the target person more faithfully than alternative representations, such as attributes [4], [5], [6]. Therefore, TBPS extends high practical value in real-world applications, such as multiple people tracking and person re-identification [7], [8], [9].

However, the task of TBPS has two open challenges. First, as a cross-modal retrieval task, TBPS inherently encounters the modal heterogeneity problem, that is, the data distributions of different modalities are inconsistent [10], [11], and such modal heterogeneity makes it difficult to directly measure the correlation between visual and text representations. Second, TPBS is essentially a fine-grained visual search task, which requires the model to be effective in matching pedestrian images with high variations. In this case, a global-scale sentence may not be reliable cues to retrieve all correct images of the same identity. As shown in Fig. 1, given a text query, the retrieved pedestrian images only matching the global sentence could be a different identity, while fine-grained matching with appearance details such as “long hair” is more distinct to facilitate the matching. To address the modal heterogeneity problem, some studies have investigated the alignment between visual and text modalities. Early approaches only consider the global-level alignment [12], [13], [14], which aligns the global visual with overall textual information, see relation I in Fig. 2(a). Recall Fig. 1 that different pedestrians may look very similar in their overall appearance, and it is difficult to correctly distinguish them only

Manuscript received 16 June 2022; revised 22 October 2022; accepted 23 November 2022. Date of publication 5 December 2022; date of current version 8 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62176178, U19A2073, and 62002096, and in part by the Natural Science Foundation of Tianjin under Grant 19JCJC16000. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Palaiahnakote Shivakumara. (Corresponding author: Lin Yuanbo Wu.)

Zhong Ji and Junhua Hu are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: jizhong@tju.edu.cn; hujunhua@tju.edu.cn).
Deyin Liu is with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230039, China (e-mail: iedyzzu@outlook.com).
Lin Yuanbo Wu is with the Department of Computer Science, Swansea University, SA1 8EN, U.K. (e-mail: jolin.lwu@gmail.com).
Ye Zhao is with the Key Laboratory of Knowledge Engineering with Big Data, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: zhaoye@hfut.edu.cn).
Codes are available at https://github.com/mul-hjh/ACSA.
Digital Object Identifier 10.1109/TMM.2022.3225754

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
We develop a transformer-based framework to extract representations, that is, global/local visual representations, global textual complete texts correspond to the whole image again. Thus, it relations: overall description corresponds to a whole image, text description. In fact, this process essentially includes three "black pants". These information constitute the final complete such as "short black hair," "black, white and brown plaid shirt," "tall," "thin," "man". Then, he tends to describe the details, a witness is likely to outline a person's body and gender, such as "woman," "dress" and "slim" are different scales. For example, a word may correspond to a patch or the entire image, such as "woman," "dress" and "slim" are the overall description with regard to a person. Nevertheless, the words "glasses," "bag," and "shoes" can only describe some regions of an image. Thereby, they proposed the Non-local Alignment over Full-Scale representations (NAFS) [22], which considers full-scale adaptive alignment. In other words, four alignment relations I, II, III and IV are fully employed in NAFS [22] (see Fig. 2(a)).

However, such full-scale representation learning is based on the hypothesis that the cross-modal instances are symmetric in their information level, that is, each instance, e.g., the sentence can be associated with an entire image or a region, and vice versa. This may not hold true in TBPS. For instance, one may use a phrase to describe a whole image as outline but unlikely to use a long sentence to describe a region. As shown in Fig. 2(b), a witness is likely to outline a person's body and gender, such as "tall," "thin," "man". Then, he tends to describe the details, such as "short black hair," "black, white and brown plaid shirt," "black pants". These information constitute the final complete text description. In fact, this process essentially includes three relations: overall description corresponds to a whole image, detailed texts correspond to a few regions of an image, and the complete texts correspond to the whole image again. Thus, it requires to extract multi-scale visual and textual representations, that is, global/local visual representations, global textual representations and phrase representations. Global visual features can be easily extracted using a pre-trained off-the-shelf model [12], while extracting local visual features is inexplicable, due to the presence of occlusion, background clutter in natural pedestrian images. To extract local features, some methods [16], [22] applied object detection or additional branch networks to detect salient regions and then extract features. However, these methods yield high computational cost because of the external networks. Other methods [17], [23] directly sliced the global visual representations horizontally into non-overlapping slices as local visual representations. This fashion is simple but may inadvertently divide the same part into different slices.

In this paper, we propose an Asymmetric Cross-Scale Alignment (ACSA) approach for TBPS. Specifically, we employ the Swin Transformer [24] to extract the global visual representations, and then divide the global representation into four regions as local visual representations, namely head, upper body, lower body and feet. This partition strategy does not involve extra computational cost but can better preserve the salient body parts for fine-grained matching. In the text domain, we employ BERT [25] to extract the global textual representations and local phrase representations. Both Swin Transformer [24] and BERT [25] are based on the self-attention mechanism, which can fully leverage the information within each modality.

We further propose an asymmetric cross-scale alignment module (ACSA) that performs three alignments, i.e., relations I, II, III in Fig. 2. The proposed ACSA module consists of a global-level alignment and an asymmetric cross attention. The former is to perform the global image-text alignment, i.e., relation I, and the latter is to adaptively align the image/regions with phrases, i.e., relation II and relation III. Compared to the multi-granularity alignment based methods [15], [16], [17], [18], [19], the proposed ACSA performs cross-scale image-phrase alignment, and the phrase is adaptively aligned with an entire image or regions. In other words, a phrase may correspond to either a region or the whole image. This alignment can be automatically learned through the cross-attention mechanism in the network, rather than restricting the alignment at a certain scale, e.g., a global or local scale. Compared to a recent adaptive full-scale alignment method [22], the proposed ACSA does not perform the region-text alignment, i.e., relation IV in Fig. 2. This alignment is empirically demonstrated to be unnecessary for the task of TBPS. In other words, a full-scale alignment can cause the over-matching, while not contribute to the matching performance. Please refer to our experimental results in Table V.

The main contributions of this paper are summarized below.

- We propose an Asymmetric Cross-Scale Alignment approach, which exploits three effective alignments, namely image-text, region-phrase, and image-phrase alignments, to improve the performance of TBPS.
- We develop a transformer-based framework to extract multi-scale feature representations, including the global and local representations for both image and text domains. With such multi-scale features, the cross-modal matching
is performed with the proposed asymmetric cross-attention mechanism.

- Our approach achieves state-of-the-art performance on two public datasets. Extensive ablation studies and visualization demonstrate the effectiveness of our approach.

II. RELATED WORKS

A. Text-Based Person Search

Since Li et al. [12] first established the task of TBPS and released a large-scale dataset called CUHK-PEDES. Since then, TBPS has become a popular topic in intelligent surveillance. It can search relevant person images using natural language as the query instead of using images or attributes [2], [3], [5], [26], [27], and thus has shown high practical value in real-world applications [28]. Most existing studies employ the following steps: (1) Applying CNNs or RNNs to extract the respective visual and textual features from images and texts; (2) Projecting these cross-modal features into a common latent feature space, followed by alignment; (3) Calculating the similarity of the image-text pair. Roughly, existing methods, based on their different focuses, can be categorized into three streams: feature representation approaches, cross-modal alignment approaches and approaches focusing on loss functions of similarities.

The first group focuses on the feature representations. For the visual modality, CNNs are the most popular backbones, such as VGG-Net [12], [13], MobileNet [14], [29], and ResNet [30], [31]. As for the textual modality, early studies usually employed RNNs [12], [14], while some methods employed CNNs [15], [32]. For example, Zhang et al. [14] tokenized the sentence and split it into words, and then sequentially processed them with a bi-directional LSTM, which is a variant of RNNs. Zheng et al. [32] proposed a dual-path CNN to learn the image and text representations. Since TBPS is a fine-grained search task, local discriminative features are imperative. Therefore, attention mechanism can be leveraged, which seeks to boost local information or mitigate the noisy interference residing in global features. Li et al. [12] proposed a GNA-RNN model with the gated neural attention mechanism, taking into account that different words have different importance. Ji et al. [15] applied attention mechanism to learn discriminative representations, and offered an accurate guidance to a common space. Different from the existing works, we employ Swin Transformer [24] and BERT [25] to extract visual and textual representations, respectively. They both are based on self-attention mechanism.

The second group investigates the alignment of visual and textual modalities. For example, Zhang et al. [14] learned the global representations of images and texts, and then aligned the images with sentences, without involving local alignment. Afterwards, multi-scale alignment received great attention, in which the local-level alignment is employed as an important supplement to the global-level alignment. Jing et al. [16] utilized pose information to guide visual feature extraction, thereby learning latent semantic alignment between visual part and textual noun phrase. Niu et al. [17] proposed a multi-granularity image-text alignment model. Particularly, they first extracted the features of image parts and noun phrases as local representations, and then performed multi-granularity alignment, i.e., global-global alignment, global-local alignment, and local-local alignment. Zheng et al. [18] proposed a hierarchical Gumbel attention network, which adaptively selected the semantically relevant image regions and words/phrases for precise alignment. Their matching strategy includes three granularities, i.e., word-level, phrase-level, and sentence-level. Recently, cross-scale alignment was developed to indicate that the alignments between different scales are also beneficial. Gao et al. [22] proposed non-local alignment over full-scale representations. They designed a novel staircase CNN network and a locality-constrained BERT model to extract multi-scale visual and textual representations, and applied a contextual non-local attention mechanism to align the learned representations across different scales adaptively. Different from them, as discussed above, we extract multi-scale representations for performing asymmetric cross-scale alignment. We also propose a partition strategy to obtain local visual representations without computational cost.

The third group aims to develop different loss functions for similarities. For instance, Zheng et al. [32] proposed the instance loss that explicitly considers the intra-modal data distribution, and each image/text group is viewed as a class. Zhang et al. [14] proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for learning a discriminative image-text embedding. Both CMPM and CMPC losses are effective in global-level alignment. Besides, in this paper we design an asymmetric cross-scale alignment loss based on KL divergence for cross-scale alignment.

B. Transformer

Transformer was first proposed in [33] for addressing machine translation tasks. Instead of using recurrent formulation, it only employs the self-attention mechanism, and thus Transformer has a better parallel ability and yet alleviates the problem of long-distance dependence of texts. Based on Transformer, Devlin et al. [25] proposed a pre-training BERT model, which has good generalization ability and achieve promising progress on multiple NLP tasks.

With the resurgence a series of Transformer models in NLP [34], [35], [36], their applications in computer vision have also attracted increasing attention. Dosovitskiy et al. [37] proposed a seminal Vision Transformer model (ViT), which interprets an image as a sequence of patches and processes them with a standard Transformer encoder as that in NLP. However, there are two drawbacks in ViT. First, too many patches in high-resolution images will cause high computational complexity. Second, the fixed split patch mode is difficult to adapt to the problem of variable scale in computer vision. To address the above challenges, Liu et al. [24] proposed Swin Transformer, a hierarchical Transformer whose representation is calculated with shift windows. This way of grouping patches significantly reduces the computational complexity. In addition, by gradually merging patches, its view of field is gradually increased. This renders it more suitable for computer vision tasks.

Recently, Transformer has achieved state-of-the-art performance on multiple computer vision tasks [38], [39], [40].
Chen et al. [41] applied it to low-level computer vision task, and achieved state-of-the-art performance on several tasks like super-resolution, denoising, and de-raining. He et al. applied ViT to the Re-ID task [42], in which they employed a sliding window to generate overlapping patches as the input of ViT to maintain the local neighbor structure information of the patch. Liang et al. [43] employed Swin Transformer in image restoration task, where multiple Residual Swin Transformer Blocks (RSTB) is developed to extract deep features, and each RSTB is composed of multiple Swin Transformer layers and residual connection. The shift window mechanism in RSTB can perform long-distance dependency modeling.

III. OUR APPROACH

Text-Based Person Search (TBPS) can be regarded as a fine-grained cross-modal retrieval task which simultaneously deals with multi-granularity alignment and modal heterogeneity. The major challenge is to extract finer features from texts and images at multiple levels, and appropriately align these instances across modalities. To this end, we design a framework on the basis of Swin Transformer [24] and BERT [25] to extract multi-scale representations from images and texts, and perform the Asymmetric Cross-Scale Alignment (ACSA), i.e., global-level image-text, local-level region-phrase, and cross-scale image-phrase alignments. As shown in Fig. 3, we use the output of Swin Transformer [24] as the global image representation, and then divide this representation into head, upper body, lower body, and foot regions, which form the local image representations. BERT [25] is employed as an encoder to extract the multi-scale representations of a given text, i.e., a global text representation and a set of noun phrase representations. In the following, we first detail the feature extraction from images and texts (Section III-A and III-B). Then, we present the proposed Asymmetric Cross-Scale Alignment (Section III-C).

A. Visual Representations

We extract the representations from images in both global and local levels. The global representation integrates all the information of a person. The local features provide fine-grained details with respect to body parts and patterns, etc. We argue that both global and local features should be leveraged to faithfully describe the identities in images.

Global Representations: Given an image \( I \), we aim to encode it into a vector \( I_g \in \mathbb{R}^{1\times d} \), where \( d \) is the dimensionality of the feature vector. We adopt Swin Transformer [24] as the backbone feature extractor due to its capability of extracting hierarchical features at both global and local levels. Specifically, we first resize the image \( I \) to 224x224, then we divide it into patches to fit the Swin Transformer [24]. We apply the pre-trained Swin Transformer model [24], and fine-tune it on our training data. Since Swin Transformer [24] gradually increases its field of view as the network deepens, we use the feature output of Stage 4 (after the global pooling) as the global visual representation, i.e., \( I_g \).

Local Representations: We divide the pedestrian image into several regions according to its characteristics, i.e., the head (e.g., cap, hairstyle, glasses), the upper body (e.g., jacket, backpack), the lower body (e.g., pants, handbag), and the foot regions...
(e.g., shoes), we extract visual features from these four regions to form the local representations. Specifically, we divide the global visual representation into six parts horizontally, then employ the first and the second parts as the head representation, the second and third parts as the upper body representation, the fourth and fifth parts as the lower body representation, and the sixth part as the foot region representation. The head and upper body regions are partially overlapping, because some components may cross the two regions, such as long hair, scarf, etc. We additionally apply a fully connected layer to adjust all representations into the same dimension. Finally, we concatenate the four region embeddings to be the local image embeddings, denoted as \(I_r = [I_{\text{head}}, I_{\text{upper}}, I_{\text{lower}}, I_{\text{foot}}] \in \mathbb{R}^{k \times d}\), where \(k = 4\) and \(d\) is the dimensionality of the embedding.

**B. Textual Representations**

We employ the Bidirectional Encoder Representation from Transformers (BERT) [25] to extract textual features. The self-attention mechanism in BERT [25] can fully make use of the contextual relationship between words, allowing each word to establish a connection with any other word.

**Global Representations:** To fit the input requirement of BERT [25], we split the texts into words and tokenize each word to be a token, then we insert [CLS] and [SEP] tokens at the beginning and end, respectively. We set the maximum number of tokens to \(L\), if there are less than \(L\), we fill them with zeros, and if there are more than \(L\) tokens, we take the first \(L\) tokens. Then we input the processed texts into the pre-trained BERT [25], we take the [CLS] as the global text representation, denoted as \(T_g \in \mathbb{R}^{1 \times d}\).

**Local Representations:** Li et al. [12] suggest that nouns have more discriminant information. Thus, we extract noun phrases from texts as local text representations. Specifically, we use the TextBlob tool [44] to extract \(M\) noun phrases from every text, and encode them into feature vectors in a similar way to the above textual encoding. The final local text representations are represented as \(T_p = \{t_1, t_2, \ldots, t_M\} \in \mathbb{R}^{M \times d}\).

**C. Asymmetric Cross-Scale Alignment**

Our observation is that the region-text alignment does not hold true in TBPS, while other forms of alignment in terms of image-text, region-phrase and image-phrase are beneficial to the task. Therefore, we propose an Asymmetric Cross-Scale Alignment (ACSA) approach. The proposed ACSA consists of two major components: a global-level image-text alignment, and region-phrase/image-phrase alignments based on an asymmetric cross-scale attention module.

**Global Alignment:** We employ the Cross-Modal Projection Matching (CMPM) loss [14] and the Cross-Modal Projection Classification (CMPC) loss [14] for the global-level matching, which are demonstrated to be effective in learning cross-modal visual and textual representations. The CMPM loss associates the representations of different modalities by integrating the cross-modal projections into the KL divergence. Moreover, the CMPC loss applies identity-level annotations for cross-modal projection classification, so as to increase the differences of features for inter-class samples and enhance the compactness of features for intra-class samples. More details about the two losses can be found in [14].

**Asymmetric Cross-Attention Module:** Inspired by Lee et al. [45], we propose an Asymmetric Cross-Attention Module (ACAM) to perform the region-phrase alignment and the image-phrase alignment. As shown in Fig. 3, ACAM takes two inputs: a set of multi-scale visual embeddings \(I_m = [v_0, v_1, \ldots, v_k] \in \mathbb{R}^{(k+1) \times d}\), which are obtained by concatenating the global image embedding \(I_g\) and the region embeddings \(I_r\), and a set of noun phrase embeddings \(T_p = [t_1, t_2, \ldots, t_M] \in \mathbb{R}^{M \times d}\). The output of ACAM is a similarity score, which measures the similarity of an image-text pair by using \(I_m\) and \(T_p\).

Intuitively, ACAM attempts to align image-phrase and region-phrase instances in the sense that a noun phrase may correspond to the whole image or the image partially. However, we do not perform the region-text alignment because it is unlikely that all textual descriptions are trying to describe a specific part of an image. Therefore, the alignment relationship between texts and images is asymmetric. More specifically, in ACAM, we use visual entities and noun phrases as contexts for each other, while paying different attention to them to obtain the weighted visual and textual representations, and then calculate the similarity for the image-text pair. In the following, we detail the two directional cross-attention based alignment nested in ACAM.

**Image-Text Cross-Attention:** For a specific visual entity \(i\) (the entity may be a region or an image), we calculate the similarity between its embedding \(v_i\) and every noun phrase embedding \(t_j\) as follows:

\[
s_{ij} = \frac{v_i^T t_j}{\|v_i\| \|t_j\|} \quad i \in [0, k], \; j \in [1, M]. \tag{1}\]

Here, \(s_{ij}\) represents the similarity between the \(i\)-th visual entity and the \(j\)-th phrase. We normalize \(s_{ij}\) as

\[
\tilde{s}_{ij} = \frac{[s_{ij}]_+}{\sqrt{\sum_{k=0}^K [s_{ij}]_+^2}}, \tag{2}\]

where \([s_{ij}]_+ = \max(s_{ij}, 0)\). We further employ the attention mechanism to obtain the weighted text representations with respect to the visual entity \(i\):

\[
T_{v_i} = \sum_{j=1}^M \alpha_{ij} t_j, \tag{3}\]

where \(\alpha_{ij}\) is the attention weight that can be calculated as

\[
\alpha_{ij} = \frac{\exp(\lambda_1 \tilde{s}_{ij})}{\sum_{j=1}^M \exp(\lambda_1 \tilde{s}_{ij})}. \tag{4}\]

In (4), the parameter \(\lambda_1\) is the inverted temperature of the soft-max function. Then, the similarity between the visual representation \(v_i\) and the corresponding weighted text representation \(T_{v_i}\) is computed as:

\[
S(v_i, T_{v_i}) = \frac{v_i^T T_{v_i}}{\|v_i\| \|T_{v_i}\|}. \tag{5}\]
TABLE I
COMPARISON WITH STATE-OF-THE-ART APPROACHES ON CUHK-PEDES DATASET

| Method                      | Scale                | Top-1 | Top-5 | Top-10 | Total   |
|-----------------------------|----------------------|-------|-------|--------|---------|
| GNA-RNN (CVPR’17) [12]     | 19.05                | 53.64 |       |        |         |
| PWM-ATH (WACV’18) [13]     | 27.14                | 49.45 | 61.02 | 137.61 |         |
| Dual-Path (TOMCCAP’20) [32] | 44.40                | 66.26 | 75.07 | 185.73 |         |
| CMPM+CMPC (ECCV’18) [14]   | 49.37                | 79.27 |       |        |         |
| TIMM (ICCV’19) [30]        | 54.41                | 77.56 | 84.78 | 216.75 |         |
| PMA (AAAI’18) [16]         |                     |       | 81.23 | 208.58 |         |
| MIA (TIP’19) [17]          |                     |       | 82.90 | 211.00 |         |
| VDAA (ECCV’20) [20]        |                     | 83.52 | 215.33|        |         |
| HGAN (MM’20) [18]          |                     | 79.49 | 86.62 | 225.11 |         |
| T-MRS (TCSVT’21) [48]      |                     | 78.25 | 84.93 | 220.85 |         |
| MGEL (IJCAI’21) [49]       |                     | 80.01 | 86.74 | 227.02 |         |
| SSAN (arXiv’21) [50]       |                     | 80.15 | 87.63 | 228.25 |         |
| AXM-Net (arXiv’21) [23]    |                     | 79.41 | 85.75 | 227.06 |         |
| DSSL (MM’21) [46]          |                     | 80.41 | 87.56 | 227.95 |         |
| DSSL (MM’21) [46] + RR     |                     | 81.11 | 88.01 | 232.45 |         |
| TIPCR (Neurocomputing’22) [51] |                 | 82.82 | 89.01 | 235.46 |         |

RR stands for re-ranking algorithm. Best results are in boldface.

TABLE II
COMPARISON TO THE STATE-OF-THE-ART METHOD ON RSTPR-REID DATASET

| Method                      | Top-1 | Top-5 | Top-10 | Total   |
|-----------------------------|-------|-------|--------|---------|
| DSSL (ACM MM’21) [46]       | 32.43 | 55.08 | 63.19  | 150.70  |
| ACSA (Ours)                 | 48.40 | 71.85 | 81.45  | 201.70  |

Best results are in boldface.

By averaging all $S(v_i, T_{vi})$, we obtain the final similarity for an image-text pair:

$$S(I, T) = \frac{\sum_{i=0}^{k} S(v_i, T_{vi})}{k + 1}. \quad (6)$$

Fig. 4 shows the intuition of the above image-text cross-attention mechanism. If the visual entity $i$ is not described in the text, the similarity value between $v_i$ and $T_{vi}$ will be small. This is because when we calculate $T_{vi}$, the visual vector $v_i$ and each phrase embedding $t_j$ is assumed to have a uniform similarity distribution, which results in an unweighted attention to a specific phrase (For example, $\alpha_{1} = \alpha_{2} = \ldots = \alpha_{M} = 1/M$). If a phrase $t_j$ is important to describe a subject, paying no attention to $t_j$ will cause a small similarity between $v_i$ and $T_{vi}$ in (5). We particularly reflect the importance of each phrase associated with the image, and formulate this rationale into the embedding based similarity function. Therefore, we measure the importance of the visual entity $i$ with respect to the texts by calculating the similarity between $v_i$ and $T_{vi}$.

**Text-Image Cross-Attention:** Similarly, for the noun phrase $j$, we calculate the similarity between its embedding $t_j$ and all visual entities as follows:

$$s'_{ij} = \frac{v_i^T t_j}{\|v_i\| \|t_j\|}, \quad i \in [0, k], j \in [1, M]. \quad (7)$$

Here, $s'_{ij}$ represents the similarity between the $i$-th visual entity and the $j$-th phrase. We normalize it as

$$\tilde{s}'_{ij} = \max (s'_{ij}, 0).$$

With these similarity values, we employ the attention mechanism to obtain a weighted visual representation related to the phrase $i$:

$$V_{ti} = \sum_{j=1}^{M} \alpha'_{ij} t_j, \quad (9)$$

where $\alpha'_{ij} = \exp(s'_{ij}) / \sum_{j=1}^{M} \exp(s'_{ij})$ and $\lambda'$ is the inversed temperature of the SoftMax function. Then we calculate the similarity between the phrase embedding $t_j$ and the corresponding weighted visual representation $V_{ti}$ as follows:

$$S(t_j, V_{ti}) = \frac{t_j^T V_{ti}}{\|t_j\| \|V_{ti}\|}. \quad (10)$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE III
IMPACT OF DIFFERENT BACKBONES

| Image Model | Text Model   | Top-1     | Top-5     | Top-10    | Total   |
|-------------|--------------|-----------|-----------|-----------|---------|
| ResNet50 [52] Bi-LSTM [53] | 43.65 67.00 76.37 187.02 |
| ResNet50 [52] BERT [25] | 52.11 73.63 82.35 208.09 |
| ViT [37] Bi-LSTM [53] | 46.34 67.99 77.80 190.15 |
| ViT [37] BERT [25] | 59.61 79.19 85.56 224.36 |
| Swin Transformer [24] Bi-LSTM [53] | 45.89 69.60 78.66 194.15 |
| Swin Transformer [24] BERT [25] | 60.77 80.02 86.63 227.42 |

Best results are in boldface.

TABLE IV
STUDY ON DIFFERENT LOCAL IMAGE FEATURE EXTRACTION METHODS

| Strategy | Top-1 | Top-5 | Total   |
|----------|-------|-------|---------|
| Six Slices [17] | 62.39 60.64 87.05 230.08 |
| Partition | 63.56 81.40 87.70 232.66 |

Best results are in boldface.

TABLE V
IMPACT OF DIFFERENT ALIGNMENT RELATIONS

| Relation | Image Region | Text Phrase | Top-1 | Top-5 | Top-10 | Total   |
|----------|--------------|-------------|-------|-------|--------|---------|
| I        | x            | x           | 60.77 80.02 86.63 227.42 |
| I, II     | ✓             | ✓           | 62.13 80.91 87.04 230.08 |
| I, II, III, IV | ✓   | ✓        | 62.95 81.06 87.31 231.32 |
| I, II, III | ✓             | ✓           | 63.56 81.40 87.70 232.66 |

Best results are in boldface.

By averaging all $S(t_j, V_{t_j})$, we obtain the similarity of a text-image pair:

$$S(I, T) = \frac{1}{M} \sum_{j=1}^{M} S(t_j, V_{t_j}).$$ (11)

Asymmetric Cross-Scale Alignment Loss: We apply KL divergence to associate the representations across different modalities for cross-scale matching. Given a mini-batch with $N$ image-text pairs, for each image $x_a$ the image-text pair is constructed as $\{(x_a, z_b), y_{ab}\}^{N}_{a,b=1}$, where $y_{ab} = 1$ means that $(x_a, z_b)$ is a positive pair, while $y_{ab} = 0$ indicates $(x_a, z_b)$ is a negative pair. The probability of matching $x_a$ to $z_b$ is defined as

$$p_{a,b} = \frac{S(x_a, z_b)}{\sum_{b=1}^{N} S(x_a, z_b)},$$ (12)

where $S(x_a, z_b)$ is the similarity between $x_a$ and $z_b$.

Since each identity can be associated with multiple images and multiple texts, this may incur a multi-matching $z$ for $x_a$ in a mini-batch, so we normalize the true matching probability as

$$q_{a,b} = \frac{y_{ab}}{\sum_{m=1}^{N} y_{am}}.$$ (13)

Then we apply KL divergence to measure the distance between the actual distribution $p_{a,b}$ and the true distribution $q_{a,b}$:

$$\mathcal{L}_a = \frac{1}{N} \sum_{b=1}^{N} p_{a,b} \log \frac{p_{a,b}}{q_{a,b} + \varepsilon},$$ (14)

where $\varepsilon$ is a non-negligible value to avoid numerical problems. We compute the matching loss from texts to images in a mini-batch as follows:

$$\mathcal{L}_{i2t} = \frac{1}{N} \sum_{a=1}^{N} \mathcal{L}_a.$$ (15)

Similarly, we compute the matching loss from texts to images in the mini-batch as follows:

$$\mathcal{L}_{i2t} = \frac{1}{N} \sum_{b=1}^{N} \mathcal{L}_b.$$ (16)

The matching loss is calculated as $\mathcal{L}_{ACSA} = \mathcal{L}_{i2t} + \mathcal{L}_{i2t}$.

**Objective Function:** The final objective function is formulated as

$$\mathcal{L}_{total} = \mathcal{L}_{cmpp} + \mu \mathcal{L}_{cmpc} + \gamma \mathcal{L}_{ACSA},$$ (17)

where $\mu$ and $\gamma$ are hyperparameters to control the importance of different loss functions. We conduct a study on hyperparameters, and we set $\mu = 4$ and $\gamma = 0.1$ as default.

IV. EXPERIMENTS

A. Datasets and Evaluation Protocol

**CUHK-PEDES:** CUHK-PEDES dataset [12] is the first large public dataset for TBPS, which contains 40,206 images of 13,003 pedestrians, and each image has two text descriptions. The average length of text descriptions is 23.5 words. Following [12], we divided the dataset as follows. The training set has 11,003 pedestrians, 34,054 images and 68,126 captions. The validation set has 1,000 pedestrians, 3,078 images and 6,158 captions. The test set has 1,000 pedestrians, 3,074 images and 6,156 captions.

**RSTPReid:** RSTPReid dataset is a new dataset constructed by Zhu et al. [46] based on the MSMT17 dataset. This dataset is more challenging than CUHK-PEDES dataset. RSTPReid contains 20,505 images of 4,101 pedestrians. Each pedestrian has five images with each image corresponding to two text descriptions. Following DSSL [46], we divided the dataset into the training set (3,701 images), validation set (200 images) and test set (200 images).

**Evaluation Protocol:** We adopted the widely used top-k accuracy as the retrieval criterion [12], which ranks all gallery images according to their similarity with respect to the text query. If the correct pedestrian image is found in the first $k$ ranked images, the retrieval is considered to be successful.

B. Implementation Details

All images are resized to 224x224. We employed the pre-trained Swin Transformer Tiny [24] as the visual backbone to extract features from images. The BERT [25] pre-trained on the CUHK-PEDES dataset is used as the text backbone. The maximum number of tokens is set to 100. The embedding dimension is set to 768. The inverse temperature of softmax is set to 20.0. We employed the AdamW optimizer [47] for 30 epochs with the Cosine decay learning rate scheduler and 5 epochs of linear warm-up. The batch size is set to 16. The initial learning rate is 0.0001, and the minimum learning rate is 0.000005. The number of noun phrases $M$ is set to 10.
TABLE VI
EVALUATION ON DIFFERENT LANGUAGE MODELS ON CUHK-PEDES DATASET

| Method      | Visual Model | Language Model | Top-1 | Top-5 | Top-10 | Total  |
|-------------|--------------|----------------|-------|-------|--------|--------|
| TIPCB [51]  | ResNet50 [52]| BERT [25]      | 63.22 | 82.79 | 89.92  | 234.93 |
| ACSA (Ours) | Swin Transformer [24]| BERT [25]      | 63.56 | 81.40 | 87.70  | 232.66 |

Best results are in boldface.

C. Comparison With State-of-The-Arts (SOTAs)

We evaluated the proposed method by comparing to SOTA methods on both CUHK-PEDES and RSTPReid datasets. Experimental results are shown in Table I and II, respectively. The state-of-the-art methods can be classified into three categories: 1) global-scale approaches (GNA-RNN [12], PWM+ATH [13], DCMP [14], Dual-Path [32], and TIMAM [30]); 2) multi-scale approaches (PMA [16], MIA [17], ViTAA [20], HGAN [18], T-MRS [48], MGEL [49], SSAN [50], AXM-Net [23], DSSL [46], and TIPCB [51]) using both global and local scales; and 3) a full-scale approach (NAFS [22]), which uses global/local scales, and a cross scale.

From Table I, we make the following observations. First, the multi-scale approaches generally outperform the global-scale approaches. For instance, TIPCB [51] achieves 63.63 at top-1 vs. TIMAM [30] with 54.41 at top-1. This proves the necessity of employing finer scale alignment. The full-scale approach, i.e., NAFS [22] also shows competitive results, suggesting that the alignment between different scales is beneficial to the task of TBPS. Comparing to these methods, our approach achieves similar performance to TIPCB [51] in all evaluation indicators without re-ranking (e.g., ACSA→63.56 vs TIPCB [51]→63.63 on top-1). It is noteworthy that TIPCB [51] uses multiple branches on top of different layers of the neural network to extract multi-scale visual features. This leads to high computational cost. In contrast, our approach achieves competitive performance using a simple network architecture.

To further improve the retrieval performance, we empirically employed a re-ranking algorithm [22] only in the testing phase. Obtained results show that re-ranking can noticeably improve the top-1 accuracy of our method by 5.11%. As indicated by [22], the re-ranking process requires both text-image and image-image retrieval. As such, the model not only needs to align the data of different modality, but also needs to fully leverage the information in each modality, especially in the image modality. In this experiment, we employed the self-attention mechanism to fully discover the informative priors in the image modality, and applied the cross-attention mechanism to effectively reduce the modal gap.

Experimental results on RSTPReid dataset are reported in Table II. RSTPReid is a newly introduced dataset for TBPS. As such, only DSSL [46] is validated on this dataset. Table II shows that comparing to DSSL [46] our algorithm achieves performance gain by 15.97%, 16.77% and 18.26% in terms of top-1, top-5 and top-10 respectively.

D. Ablation Studies

To thoroughly study the effectiveness of each module in our method, we conducted a series of experiments on CUHK-PEDES dataset, including different backbones for visual/textual representations, image partitioning and the effect of the proposed ACSA.

1) Impact of Different Backbones: In this experiment, we studied the impact of applying different backbones for the visual and textual domains. Specifically, we considered three image backbones, i.e., ResNet50 [52], ViT [37] and Swin Transformer [24], and two text backbones, i.e., Bi-LSTM [53] and BERT [25]. After extracting image and textual features using...
respective backbones, we simply performed a multi-scale alignment, and the cross-modal matching was performed by using two losses: CMPM [14] and CMPC [14].

Experimental results are reported in Table III. When Bi-LSTM [53] is applied as the text backbone, we could observe that Swin Transformer [24] brings 2.24% performance gain on top-1 accuracy. When replacing Bi-LSTM [53] with BERT [25], the top-1 accuracy gain increases to 8.66%. Similar results can be seen on top-5 and top-10. These results demonstrate the effectiveness of employing the Swin Transformer [24] as the image backbone. Similarly, BERT [25] outperforms Bi-LSTM [53] as the text backbone. This demonstrates that the self-attention based network is effective in discovering the intra-modal information for feature representations. When we combine Swin Transformer [24] and BERT [25], a notable improvement of 17.12% is achieved on top-1 accuracy versus the combination of ResNet50 [52] and Bi-LSTM [53]. This empirically confirms the adoption of transformer-based networks in both domains.

2) Different Image Partitioning Strategies: We employed a partitioning strategy to obtain the embedding of head, upper body, lower body and foot regions as local visual embeddings. Unlike the simple slicing strategy of directly dividing the global image embedding into six slices [17], our partitioning strategy brings no extra computational cost and ensures the consistency of local regions in higher layers of the network. We compared our partitioning strategy with the simple slicing method, and report the results in Table IV. The results show that our partitioning strategy achieves better performance and effectively avoids splitting the same region into different slices.

3) Impact of Asymmetric Cross-Scale Alignment: To investigate the effectiveness of the proposed asymmetric cross-scale alignment, we implemented different scales of feature embedding in the cross-attention module, as shown in Table V. It should be noted that these experiments are based on global-level alignment, that is, the alignment in the cross-attention module is a supplement to global-level alignment.

Multi-Scale Alignment: We only employed local embeddings for visual and text in the cross-attention module, i.e., region embeddings and noun phrase embeddings. It can be regarded as a multi-scale alignment approach that includes global-level and local-level alignments, i.e., the relations I and II in Fig. 2. Table V shows that it improves the top-1 accuracy by 1.36% in comparing to the global-level alignment. This proves the effectiveness of using finer-scale alignment.

Adaptive Full-Scale Alignment: Based on the above multi-scale alignment, we further considered the cross-scale alignment, i.e., region-text and image-phrase alignment. That is, it includes relations I~IV. Specifically, we concatenated the region embeddings and global image embedding as the final visual embeddings. We also concatenated noun phrase embeddings and global text embedding as the final textual embeddings. Compared with the multi-scale alignment, e.g., relations I~IV, our method achieves better results by applying the cross-scale alignment.

Asymmetric Cross-Scale Alignment: Intuitively, we argue that region-text alignment in adaptive full-scale alignment is unnecessary. Thus, in the cross-attention module, we concatenated region embeddings and global image embedding as the final visual embeddings but only utilized the noun phrase embeddings as the final textual embeddings. That is, we considered the relations I~III. It can be seen from Table V that this combination shows inferior performance, which indicates that region-text alignment is unnecessary. In fact, the entire texts rarely correspond to a specific region of a person image. Therefore, employing this alignment offers no benefits to the task of TBPS.

4) Evaluation on Different Language Models: In this experiment, we evaluated the proposed method using different language models, i.e., BERT [25] and XLNet [34]. We considered the variations of our method and TIPCB [51] by using the two language models. Specifically, we replaced the BERT [25] in both the proposed method and TIPCB [51] by using a recent language model XLNet [34]. Experimental results are reported in Table VI. We have the following observations. Both our approach and TIPCB [51] show better results than the variants of using XLNet [34]. This affirms the effectiveness of using BERT [25] as the textual backbone. One possible reason is BERT [25] adopts the random sampling, which produces more robust textual representations.
E. Visualization

In Fig. 5, we show the saliency information in pedestrian images which can be learnt through the attention maps. In each group, we arrayed three images: the leftmost is the raw pedestrian images, the middle is the corresponding heat map based on the attention, and the right is the attentive saliency. We can see that the right image clearly shows the area where the saliency is located. In specific, our model can effectively ignore the background while focusing on its saliency on pedestrian images. In fact, the saliency information is mainly concentrated in four regions, i.e., head, upper body, lower body and the feet region, which verifies the rationale of our proposed partition strategy. Although there are some exceptions that some images may not meet this criterion, our partition strategy suits to human cognitive perception. Finally, it is noted that our model reveals a good correspondence between noun phrases and the saliency in the image. This demonstrates that our model is effective in aligning the cross-modal data. In Fig. 6, we further show retrieval results with pedestrian images captured under various conditions, such as illuminations and nighttime. The results show that most of the correct images appear in the first few ranked positions of the retrieval list. This proves that our model is robust against nuisance factors, and can be applied into various real-world conditions.

V. CONCLUSION

In this paper, we propose a transformer-based model for text-based person search by employing the Swin Transformer [24] and BERT [25] to extract multi-scale features from images and texts. To allow for fine-grained visual-text matching, we propose an Asymmetric Cross-Scale Alignment for adaptive cross-modal match, which consists of a global visual-text alignment, and an asymmetric cross-attention module for region/image-phrase alignments. Extensive experiments on CUHK-PEDES and RST-PRReid datasets demonstrated the effectiveness and superiority of our approach.

REFERENCES

[1] X. Gong et al., “LAG-Net: Multi-granularity network for person re-identification via local attention system,” IEEE Trans. Multimedia, vol. 24, pp. 217–229, 2022.
[2] L. Wu, R. Hong, Y. Wang, and M. Wang, “Cross-entropy adversarial view adaptation for person re-identification,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 7, pp. 2081–2092, Jul. 2020.
[3] L. Wu, Y. Wang, X. Li, and J. Gao, “What-and-where to match: Deep spatially multiplicative integration networks for person re-identification,” Pattern Recognit., vol. 76, pp. 727–738, 2018.
[4] Y. Shi et al., “Person retrieval in surveillance videos via deep attribute mining and reasoning,” IEEE Trans. Multimedia, vol. 23, pp. 4376–4387, 2021.
[5] Z. Ji, Z. Hu, E. He, J. Han, and Y. Xiong, “Pedestrian attribute recognition based on multiple time steps attention,” Pattern Recognit. Lett., vol. 138, pp. 170–176, 2020.
[6] H. Fan, H. Hu, S. Liu, W. Lu, and S. Pu, “Correlation graph convolutional network for pedestrian attribute recognition,” IEEE Trans. Multimedia, vol. 24, pp. 49–60, 2022.
[7] L. Wu et al., “Pseudo-pair based self-similarity learning for unsupervised person re-identification,” IEEE Trans. Image Process., vol. 31, pp. 4803–4816, 2022.
[8] L. Wu, D. Liu, X. Guo, R. Hong, and R. Zhang, “Multi-scale spatial representation learning via recursive polynomial networks,” in Proc. Int. Joint Conf. Artif. Intell., 2022, pp. 1465–1473.
[9] D. Liu et al., “Generative metric learning for adversarially robust open-world person re-identification,” ACM Trans. Multimedia Comput. Commun. Appl., 2022, doi: 10.1145/3522714.
[10] D. Liu, L. Wu, F. Zheng, L. Liu, and M. Wang, “Verbal-person nets: Pose-guided multi-granularity language-to-person generation,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 09, 2022, doi: 10.1109/TNNLS.2022.3151631.
[11] L. Chen et al., “Cross-modal retrieval with heterogenous graph embedding,” in Proc. 38th ACM Int. Conf. Multimedia, 2022, pp. 3291–3300.
[12] S. Li et al., “Person search with natural language description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5187–5196.
[13] T. Chen, C. Xu, and J. Luo, “Improving text-based person search by spatial matching and adaptive threshold,” in Proc. IEEE Winter Conf. Comput. Vis. Comput. Graph., 2018, pp. 1879–1887.
[14] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 686–701.
[15] Z. Ji and S. Li, “Multimodal alignment and attention-based person search via natural language description,” IEEE Internet Things J., vol. 7, no. 11, pp. 11147–11156, Nov. 2020.
[16] Y. Jing et al., “Pose-guided multi-granularity attention network for text-based person search,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 11189–11196.
[17] K. Niu, Y. Huang, W. Ouyang, and L. Wang, “Improving description-based person re-identification by multi-granularity image-text alignments,” IEEE Trans. Image Process., vol. 29, pp. 5542–5556, 2020.
[18] K. Zheng, W. Liu, J. Liu, Z. Zha, and T. Mei, “Hierarchical gumbel attention network for text-based person search,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 3441–3449.
[19] X. Wei, C. Zhang, L. Liu, C. Shen, and J. Wu, “Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification,” in Proc. Asilomar Conf. Signal, Sys. and Comp., 2018, vol. 11362, pp. 575–591.
[20] Z. Wang, Z. Fang, J. Wang, and Y. Yang, “VITAA: Visual-textual attributes alignment in person search by natural language,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 402–420.
[21] S. Aggarwal, R. V. Babu, and A. Chakraborty, “Text-based person search via attribute-aided matching,” in Proc. IEEE Winter Conf. Comput. Vis. Comput. Graph., 2020, pp. 2606–2614.
[22] C. Gao et al., “Contextual non-local alignment over full-scale representation for text-based person search,” 2021, arXiv:2101.03036.
[23] A. Farooq, M. Awais, J. Kittler, and S. X. P. F. “AXM-Net: Cross-modal context sharing attention network for person Re-ID,” CoRR, vol. abs/2101.08238, 2021.
[24] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 9992–10002.
[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT 2019, pp. 4171–4186.
[26] L. Wu et al., “Deep co-attention based comparators for relative representation learning in person re-identification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 2, pp. 722–735, Feb. 2021.
[27] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, “Few-shot deep adversarial learning for video-based person re-identification,” IEEE Trans. Image Process., vol. 29, pp. 1233–1245, 2020.
[28] Y. Wang, “Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 17, pp. 1–22, 2020.
[29] Y. Wang et al., “Language person search with mutually connected classification loss,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 2057–2061.
[30] N. Saraffianos, X. Xu, and I. A. Kakadiaris, “Adversarial representation learning for text-to-image matching,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 581–590.
[31] D. Chen et al., “Improving deep visual representation for person re-identification by global and local image-language association,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 56–73.
[32] Z. Zheng et al., “Dual-path convolutional image-text embeddings with instance loss,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 16, no. 2, pp. 1–23, 2020.
[33] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.
[34] Z. Yang et al., “Generalized autoregressive pretraining for language understanding,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 5754–5764.

[35] Y. Liu et al., “A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692.

[36] T. B. Brown et al., “Language models are few-shot learners,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 1–25.

[37] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021, pp. 1–21.

[38] N. Carion et al., “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213–229.

[39] S. Zheng et al., “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6881–6890.

[40] J. Yang et al., “Focal self-attention for local-global interactions in vision transformers,” 2021, arXiv:2107.00641.

[41] H. Chen et al., “Pre-trained image processing transformer,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 12299–12310.

[42] S. He et al., “TransReID: Transformer-based object re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 14993–15002.

[43] J. Liang et al., “SwinIR: Image restoration using swin transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2021, pp. 1833–1844.

[44] S. Bird and E. Loper, “NLTK: The natural language toolkit,” in Proc. Assoc. Comput. Linguistics., 2004, pp. 1–4.

[45] K. H. Lee, C. Xi, H. Gang, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 212–228.

[46] A. Zha et al., “DSSL: Deep surroundings-person separation learning for text-based person retrieval,” in Proc. 29th ACM Int. Conf. Multimedia, 2021, pp. 209–217.

[47] D. P. Kingma and J. Ba, “ADAM: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015, pp. 1–15.

[48] H. Li, J. Xiao, M. Sun, E. G. Lim, and Y. Zhao, “Transformer-based language-person search with multiple region slicing,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1624–1633, Mar. 2022.

[49] C. Wang, Z. Luo, Y. Lin, and S. Li, “Text-based person search via multi-granularity embedding learning,” in Proc. Int. Joint Conf. Artif. Intell., 2021, pp. 1068–1074.

[50] Z. Ding, C. Ding, Z. Shao, and D. Tao, “Semantically self-aligned network for text-to-image part-aware person re-identification,” 2021, arXiv:2107.12666.

[51] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, “TIPCB: A simple but effective part-based convolutional baseline for text-based person search,” Neurocomputing, vol. 494, pp. 171–181, 2022.

[52] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[53] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2013, pp. 6645–6649.

Zhong Ji (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Tianjin University, Tianjin, China, in 2008. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. He has authored more than 100 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, PR, CVPR, ICCV, ECCV, NeurIPS, AAAI, and IJCAI. His research interests include zero/few-shot learning and cross-modal analysis.

Deyin Liu received the B.E. and Ph.D. degrees from Zhengzhou University, Zhengzhou, China, in 2010 and 2021, respectively. He is currently a Lecturer with the School of Artificial Intelligence, Anhui University, Hefei, China. His research interests include optimization in computer vision, unsupervised learning, and sparse representation learning.

Lin Yuanbo Wu (Senior Member, IEEE) received the Ph.D. degree from the University of New South Wales, Sydney, NSW, Australia, in 2014. She is currently a Senior Lecturer (Associate Professor) with the Department of Computer Science, Swansea University, Swansea, U.K. She was previously with The University of Western Australia, Perth, WA, Australia, Hefei University of Technology, Hefei, China, The University of Queensland, Brisbane, QLD, Australia, and The University of Adelaide, Adelaide, SA, Australia. Her research outcome are expounded with more than 60 academic papers (including two book chapters) in premier journals and proceedings. She was the Area Chair with ACM Multimedia 2022.

Ye Zhao received the M.S. degree in communication and information system from Harbin Engineering University, Harbin, China, in 2005, and the Ph.D. degree in signal and information processing from the Hefei University of Technology, Hefei, China, in 2014. She is currently an Associate Professor with the School of Computer and Information, Hefei University of Technology. From 2016 to 2017, she was a Visiting Scholar with the Computer Science Department, University of Central Florida, Orlando, FL, USA. Her research interests include multimedia analysis and pattern recognition.