Text-based Person Search in Full Images via Semantic-Driven Proposal Generation

Shizhou Zhang, Duo Long, Yitao Gao, Liying Gao, Qian Zhang, Kai Niu, Yanning Zhang

***Abstract*—Finding target persons in full scene images with a query of text description has important practical applications in intelligent video surveillance. However, different from the real-world scenarios where the bounding boxes are not available, existing text-based person retrieval methods mainly focus on the cross modal matching between the query text descriptions and the gallery of cropped pedestrian images. To close the gap, we study the problem of text-based person search in full images by proposing a new end-to-end learning framework which jointly optimize the pedestrian detection, identification and visual-semantic feature embedding tasks. To take full advantage of the query text, the semantic features are leveraged to instruct the Region Proposal Network to pay more attention to the text-described proposals. Besides, a cross-scale visual-semantic embedding mechanism is utilized to improve the performance. To validate the proposed method, we collect and annotate two large-scale benchmark datasets based on the widely adopted image-based person search datasets CUHK-SYSU and PRW. Comprehensive experiments are conducted on the two datasets and compared with the baseline methods, our method achieves the state-of-the-art performance.**

***Index Terms*—text-based Person Search, semantic-driven RPN, cross scale alignment.**

I. Introduction

RECENTLY, image-based person re-identification [1]–[9] and person search [10]–[20] problems (Figure 1(a) and 1(b)), which aim at matching a specific person with a gallery of cropped pedestrian images or finding a target person in a gallery of full (whole scene) images, have been widely studied in the computer vision community as their great application values in cross camera tracking [21]–[32], criminal investigation, person activity, intention analysis, etc. In many real-world scenarios, such as finding criminals/suspects, a query image of the target person can not always be easily obtained, while text descriptions given by witnesses are available. In such scenarios, it is necessary to develop the techniques for finding a target person with a given query text description.

Although the text-based person retrieval task (Figure 1(c)), which aims to match a given text query with a gallery of cropped person images, has been explored in recent years [33]–[42]. However, there is still a step distance from the real-world scenarios as the bounding box annotations are unavailable and the query-described person needs to be searched in a gallery of full images. To close the gap, we study the text-based person search (Figure 1(d)) problem in this paper.

Fig. 1: Comparison of the four tasks. (a) Person ReID. Query: cropped person image. Gallery: cropped person images. (b) Person Search. Query: cropped person image. Gallery: full scene images. (c) Text-based person retrieval. Query: text description. Gallery: cropped person images. (d) Text-based person search. Query: text description. Gallery: full scene images.

Note that it is straightforward to breaking down the problem into two independent tasks: person detection and text-based person retrieval. As an off-the-shelf person detector would unavoidably introduce misdetections and misalignments, it could not be optimal when taking the detection results as inputs of the second stage retrieval model.

In this paper, we propose a new end-to-end learning framework which integrates person detection, identification and image-text cross-modal matching and jointly optimizes the three tasks together. As shown in Figure 4, The detection network follows the Faster-RCNN pipeline, *i.e.* a Region Proposal Network (RPN) for person candidate generation is built on top of a Base-Net which is shared with the iden-

Shizhou Zhang, Duo Long, Yitao Gao, Liying Gao, Qian Zhang, Kai Niu and Yanning Zhang are with National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710071, China.

tification network. Alongside the conventional RPN which aims to output proposals according to the objectness scores, to pay more attention on the text-described proposals and filter out text-irrelevant ones, we propose a novel Semantic-Driven Region Proposal Net (SDRPN) where the RPN features are dynamicly instructed by the semantic feature of the input text description. After obtaining the features of the proposals, the commonness and positions of the persons are supervised by the detection branch (Det-Net) with a Softmax classification loss and a regression loss, while the uniqueness of each person IDs are discriminated by enforcing the OIM loss [10] on top of the identification branch (ID-Net). Furthermore, the BERT language model [43] is utilized to extract the text features from sentence, sub-sentence and word-levels. And the visual features are extracted from global, regional and local scales via splitting and shffuling the proposal feature maps. The similarity scores of the proposal-text pairs can be computed with the help of cross attention mechnanism to achieve a cross-scale visual-semantic feature matching.

To validate the proposed method, we collect and annotate two large scale benchmarks for text-based person search based on the widely adopted person search datasets CUHK-SYSU [10] and PRW [11]. To eliminate ambiguity, we name the corresponding datasets as CUHK-SYSU-TBPS and PRW-TBPS, respectively. As the person datasets already have the annotations of person bounding-boxes and IDs, we merely need to annotate text descriptions for each person bounding-box. In total, we collect and annotate 54969 sentences for CUHK-SYSU-TBPS and PRW-TBPS datasets. The textual descriptions contain abundant noun phrases and various sentence structures. And we give a statistical analysis of the text description in both datasets. Extensive experiments are conducted on these two datasets and the results demonstrate the superiority of our proposed method. Compared with many baseline methods, the proposed method outperforms them by a large margin.

The main contribution of our paper is three-fold and can be summarized as:

- We make the first attempt to conduct text-based person search in full images, which has more practical application values than text-based person retrieval from cropped pedestrian images. To support this research direction, two benchmark datasets CUHK-SYSU-TBPS and PRW-TBPS with large scale full images and rich text annotations are collected and annotated.
- We propose a novel end-to-end learning framework where person detection, identification and image-text embedding tasks are jointly optimized together. And it is worth noting that a SDRPN module is devised aiming to care about text description-related person proposals. The proposed SDRPN can boost the final performance by 1.21% mAP, 1.86% Rank-1 on CUHK-SYSU-TBPS dataset, and 0.73% mAP, 1.11% Rank-1 on PRW-TBPS dataset.
- We conduct comprehensive experiments on the two datasets and compare our method with many baselines. The experimental results showed that our method outperforms baseline methods by a large margin and achieves state-of-the-art performances.

The rest of the paper is organized as follows: we briefly review related work of our paper in Section II. Section III gives a statistical analysis of the collected datasets. In Section IV we elaborate the proposed framework. The experimental results are reported and analyzed in Section V. And we conclude the paper in Section VI.

II. Related Work

In this section, we briefly review the related works from the following three aspects:

A. Person search

Person search is to localize and identify a target person in a gallery of full images other than cropped pedestrian images in ReID task. Some approaches [11], [44], [45] proposed to break down the problem into two separate tasks, *i.e*. pedestrian detection and person re-identification. Different from the two-stage methods, some works devoted their efforts to propose an end-to-end learning strategy [10], [16], [28] aiming to jointly optimize the detection and re-identification tasks. Xiao *et al*. [10] firstly introduced an end-to-end person search network and proposed the Online Instance Matching (OIM) loss function for fast convergence. Han *et al*. [28] proposed to refine the detection bounding boxes supervised by the re-identification training. Munjal *et al*. [16] took full advantage of both the query and gallery images to jointly optimize detection and re-id network. Additionally, Liu *et al*. [46] proposed Conv-LSTM based Neural Person Search Machines (NPSM) to perform the target person localization as an search area iterative shrinkage process. Chang *et al*. [47] tranformed the search problem into a conditional decision-making process and trained relational context-aware agents to learn the localization actions via reinforcement learning.

Different from the image based person search whose query is a cropped pedestrian image, in this work, we investigate the text-based person search problem which is much more challenging and able to meet the requirement of the scenarios where query image is not available in many situations.

B. Text-Based Person Retrieval

Considering that the image query is not always available in real-world scenes, Li *et al*. [48] firstly introduced the text-based person retrieval task and collected a benchmark named CUHK-PEDES. Early methods about text-based person retrieval concentrate on global feature alignment, like [37], [38], [49], which employed universal feature extraction networks to extract global feature representations for images and descriptions and made efforts to design more proper objective functions for this task. Such as in [37], a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss were proposed for computing the similarity of image-text pair data. Meanwhile, there are also several methods [50], [51] employing local feature alignment to provide complementary information for global feature alignment. For example, Li *et al*. [50] applied the spatial attention, which relates each word with corresponding image regions, to

refine the global alignment in the stage-1 training. Recently, many methods [35], [36], [52] have applied global and local features of images and text descriptions to realize multi-scale matching and achieved better performance. Niu *et al*. [35] proposed a Multi-granularity Image-text Alignment (MIA) module, including global-global, global-local and local-local alignment, and improved the accuracy of retrieval by multi-grained feature alignment between visual representations and textual representations. Although the multi-scale alignment provide supplement for global feature matching, the alignment for each scale is fixed. Gao *et al*. [33] realized the need to align visual and textual clues across all scales and proposed cross-scale alignment for text-based person search.

Text-based person retrieval has achieved great performance improvement in recent years, while the task setting still has a gap with the real-world scenarios. Therefore, in this paper, we study text based person search in full images.

C. Variants of Region Proposal Network

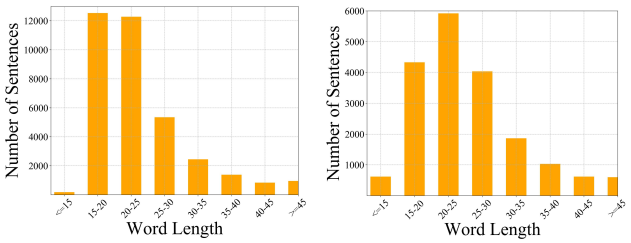
Region proposal network (RPN) is a significant component in series of detection networks, and there are many studies making efforts to improve it in order to generate more accurate or task-relevant proposals. In order to produce high-quality proposals and improve detection performance, Wang *et al*. [14] proposed Guided Anchoring Region Proposal Network, which learns to guide a sparse anchoring scheme and can be seamlessly integrated into proposal methods and detectors. Besides, [53] introduces Cascade RPN, which systematically address the limitation of the conventional RPN that heuristically defines the anchors and aligns the features to the anchors for improving the region-proposal quality and detection performance. To improve the generalization ability of neural networks for few-shot instance segmentation, Fan *et al*. [54] proposed attention guided RPN in order to generate class-aware proposals by making full use of the guidance effect from the support set.

Inspired by the above works, in this paper, we propose a Semantic-Driven Region Proposal Network for text-based person search, which employs semantic information from the query text description to generate semantically similar proposals.

III. Benchmarks for Text-based Person Search

Since there is no existing datasets for the task of text-based person search, we build new benchmarks for evaluating our method. Aiming at this task, the dataset need to include visual information of person bounding-box positions accompanied with person IDs and textual information of language descriptions. Therefore, based on the two widely adopted image-based person search datasets, CUHK-SYSU and PRW, which already contain person bounding-box and ID labels, we match each person box from train set and query set with text descriptions and propose our Text-based Person Search benchmark CUHK-SYSU-TBPS and PRW-TBPS.

For CUHK-SYSU-TBPS, there are 11,206 scene images and 15,080 person boxes with 5532 different IDs in train set, while 2,900 person boxes in query set. And we collect corresponding text descriptions from existing person retrieval dataset CUHK-PEDES (train_query&test_query), where each person box was labeled with two sentences. As for PRW-TBPS, there are 5,704 images and 14,897 boxes, with 483 different IDs in train set and 2,056 boxes in query set. And text descriptions of all person boxes were annotated, in which the boxes from the train set were labeled with one sentence, and the boxes from the query set were labeled twice independently. Here, we



(a) CUHK-SYSU-TBPS (b) PRW-TBPS

Fig. 2: The word length distributions on the benchmark datasets CUHK-SYSU-TBPS and PRW-TBPS.

Fig. 3: High-frequency words and labeled person image samples in the datasets.

labeled the training person box once due to the fact that the large amount of repetition of person box share with the same ID, and the average number of each ID occurrence is ten times individuals than that in CUHK-SYSU-TBPS. Therefore, we believe one sentence of each box in PRW-TBPS dataset is capable of providing enough samples for each identity for training.

The text descriptions of the datasets not only focus on person appearances, including clothes and body shape, but also pay attention to person actions, gestures and other details. To some extent, vocabulary and sentence length are vital indicators to evaluate the capacity of the dataset. In total, there are 1,318,445 words and 5,934 unique words in the datasets.

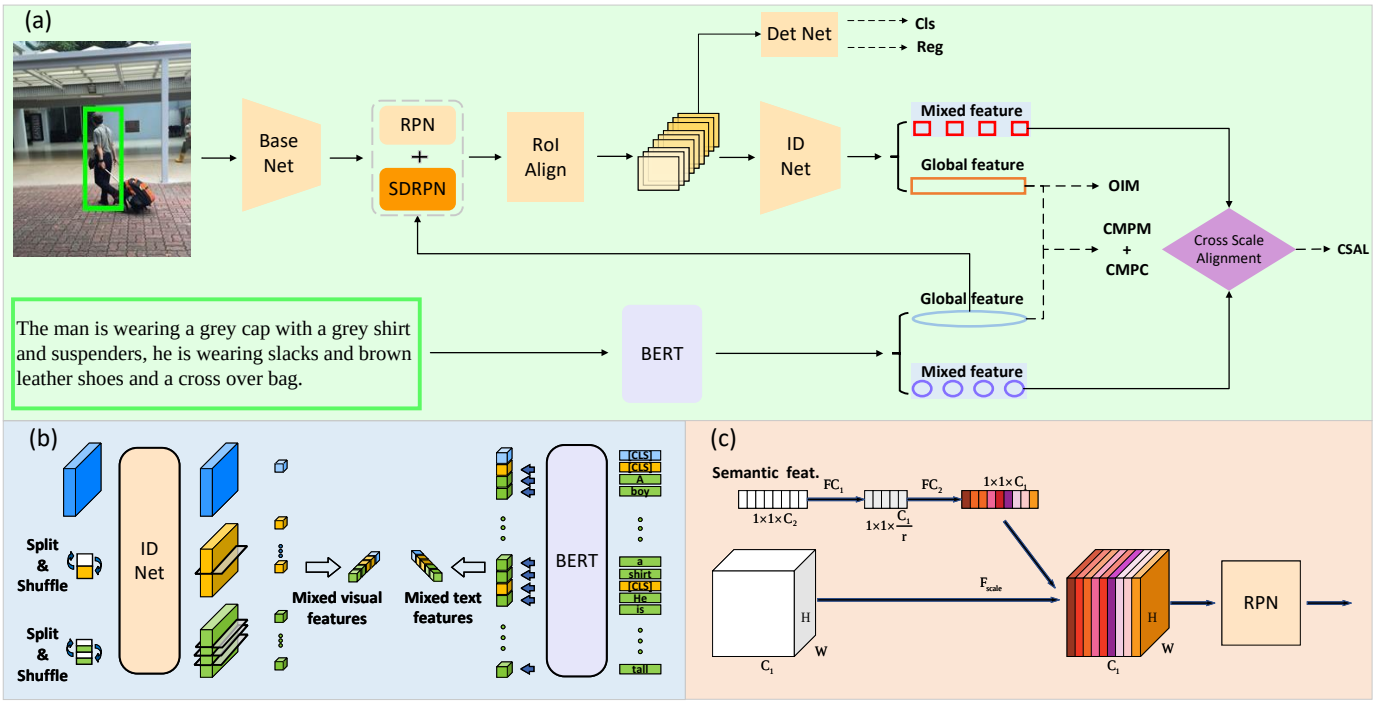


Fig. 4: (a) depicts The overall architecture of the proposed learning framework. (b) shows the process of mixed feature extraction. (c) exhibits the SDRPN.

As Figure 2 shows, most sentences have 15 to 45 words in length, and the average word lengths of the datasets are 23.9 and 24.96 words respectively, which is much longer compared with other image-caption datasets like MS-COCO [55] (5.18 words in average) and Visual Genome [56] (10.45 words in average). Figure 3 displays high-frequency words and example images with labeled bounding boxes and identities in the datasets.

IV. Method

In this section, we introduce the proposed end-to-end learning method as illustrated in Figure 4. Firstly, we briefly introduce the overview of the framework (Figure 4 (a)). Then two major components, namely Semantic-Driven RPN module and Proposal-Text Embedding module are elaborated. Finally, we give the total loss function of training the proposed method.

A. Overview

Our goal is to search the target person from full images with a query of text description. To be exact, we decompose the task to three sub-tasks, namely pedestrian detection, identification and image-text cross-modal feature embedding/matching. As shown in Figure 4 (a), the whole framework takes full images with text descriptions of labeled persons as input when training. During the inference stage, we obtain text representation as query to compare with the RoIs detected from the gallery images. The framework includes two paths, namely image-path and text-path.

For the image path, we follow the structure of one-stage person search method and add a multi-task head for localization, detection, and identification on top of the convolutional features of Faster-RCNN. We first exploit the ResNet-50 as backbone and split it into two parts as Base-Net (Conv1 to Conv4-3) and ID-Net (Conv4-4 to Conv5). The Base-Net takes the whole image as input and outputs 1024-channel feature maps. Then a novel Semantic-Driven Region Proposal Network (SDRPN) is built on top of the Base-Net feature maps to generate text-relevant region-of-interests (RoIs). After non-maximum suppression, we adopt the RoI-Align layer to obtain the pooled features for the resulted proposals. Then, those RoIs features are sampled and fed into two branches. For the identification branch, we sample top-k RoIs, based on objectness scores and Intersection over Union (IoU) threshold. After passing through the ID Net with global average pooling, an additional 1×1 convolution layer is used to project each RoI feature into a 768-dimensional feature vector. And we adopt the OIM loss [10] to supervise the feature learning process for each identity. For the detection branch, we select top 256 RoIs and pass them through a Det-Net. A softmax classifier and a linear regression are utilized to reject backgrounds and refine the person locations.

As for the text-path, the semantic feature of the query text is encoded by a BERT language model [43]. Then both the visual feature and semantic feature from the two paths are fed into proposal-text embedding module. Inspired by [33], we propose a RoI-level cross scale matching scheme which utilizes mixed multi-scale features extracted from person proposals and text descriptions for feature embedding with the help of the non-local attention mechanism. Besides, the CMPM and CMPC loss are enforced on top of the global features to supervise the proposal-text cross-modal feature embedding process.

B. Semantic-Driven RPN

Traditional RPN generates class-agnostic proposals for generic object detection. To take full advantage of the text query which contains full semantic information, we devise a Semantic-Driven RPN which leverages the semantic features from the text query to instruct the proposal generation process, aiming at paying more attention to those semanticly more similar candidates with the text description. Sepecifically, inspired by the SENet [57], the semantic features are utilized to re-weight the Base-Net feature maps. As illustrated in Figure 4 (c), SDRPN includes a channel-wise attention mechanism to guide a standard RPN, generating the proposal boxes from the re-weighted image features.

In more detail, we use the semantic feature extracted from a BERT language model and unsqueeze it to $1 \times 1 \times C_2$. The resulted feature is denoted as $\mathbf{z}$ and then we apply two fully connected layers FC_1 and FC_2 to squeeze and expand the feature $\mathbf{z}$, from C_2 to C_1/r then to C_1, to emphasize important signal correlations. Based upon the sigmoid activation $\boldsymbol{\sigma}$, the resulted excitation $\mathbf{s}$ is computed as follows.

$$\mathbf{s} = FC_2(FC_1(\mathbf{z})) = \boldsymbol{\sigma}(\mathbf{W}_2\delta(\mathbf{W_1 z})). \tag{1}$$

Then the excitation $\mathbf{s}$ is applied to the BaseNet feature maps $\mathbf{X}$ from a gallery image through channel-wise multiplication as follows:

$$\hat{\mathbf{X}} = F_{scale}(\mathbf{X}, \mathbf{s}) = \mathbf{s} \odot \mathbf{X} \tag{2}$$

Note that SDRPN extracts proposals featuring at a text-similarity score and RPN pursues the standard objectness score, Therefore, as shown in Figure 4 (a), we use SDRPN in parallel with RPN when generating proposals by summing up the scores of corresponding anchor boxes to obtain better performance.

C. Proposal-Text Embedding

The proposal-text feature embedding module aims to learn a common space for both the visual and text modality. To improve the performance, a cross-scale alignment scheme is borrowed in the embedding process.

Multi-Scale Visual Feature Extraction. In visual path, proposal features will be represented in three scales from coarse to fine, namely as global-scale, region-scale and local-scale. As illustrated in Figure 4 (b), we take the output of ID-Net as the global-scale representations of the proposals. Further, to better focus on local features and reduce the influence of large receptive field of CNN, we do the split and shuffle operation on the RoI-Aligned proposal features, which equally partitions the feature map into several horizontal stripes, and these set of the partitioned stripes are randomly shuffled and re-concatenated. The re-concatenated feature maps then are passed through the ID-Net. After that, the output feature map of the region-scale branch is horizontally partitioned into n stripes, each of which is further encoded as a region-scale feature corresponding to a certain region. Finally, a finer partition scheme is used to produce the local-scale features,

Multi-Scale Semantic Feature Extraction. As for the text path, we use the BERT language model to extract the semantic representation from three levels, namely sentence-level, sub sentence-level and word-level (Figure 4 (b)). We use the final hidden state of token [CLS], which is added at the beginning of the sentence, as the representation of the whole sentence. For the sub sentence-level, sentences are separated by commas resulting in several shorter sub-sentences. And we attach the [CLS] token to the beginning of each sub-sentence, whose final hidden state is treated as the representation of each sub-sentence. While as for the word-level, each final hidden state of word is considered as the word-level representation.

Proposal-Text cross scale alignment. After proposal and text feature extraction, we obtain a set of three-scale visual and semantic features. We concatenate them to get the mixed visual features $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_m\}$ and mixed textual features $\mathbf{T} = \{\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_n\}$, where m and n corresponds to the m_{th} and n_{th} part. To get the cross attended features, fully connected layers are used to map the mixed visual features $\mathbf{I}$ to visual queries, keys and values, denoted by $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ respectively. And the mixed semantic features $\mathbf{T}$ are mapped to semantic queries, keys and values, denoted by $\mathcal{Q}$, $\mathcal{K}$ and $\mathcal{V}$.

Firstly, the attended semantic feature $\mathbf{A}$ can be computed from the view of text-to-image attention mechanism as,

$$\mathbf{A} = norm(\mathbf{Q}\mathcal{K}^T) \odot \mathcal{V} \tag{3}$$

Then we can obtain the relevance between the visual value and its corresponding semantic context by calculating the cosine similarity between $\mathbf{V}$ and $\mathbf{A}$,

$$\mathbf{R} = \cos(\mathbf{V} \odot \mathbf{A}) \tag{4}$$

where $\mathbf{R}$ denotes the relevance scores. The similarity of text-to-image pair S is then computed by averaging all components of $\mathbf{R}$. Meanwhile, by alternating the semantic keys as queries and visual queries as keys respectively, and following the above procedure, the similarity of image-to-text pair denoted by S' can be computed.

Then, assuming that a mini-batch of B person boxes and captions are given, and all image-to-text pairs are constructed as $\{(\mathbf{I}_i, \mathbf{T}_j), y_{i,j}\}_{i=1,j=1}^{B}$. Note that $y_{i,j} = 1$ if $(\mathbf{I}_i, \mathbf{T}_j)$ is a matched pair, otherwise $y_{i,j} = 0$. To maximize similarities between the matched pairs and push away the unmatched pairs, KL divergence is enforced to diminish the modality discrepancy. Considering that the normalized similarities $\overline{\mathbf{S}}$ can be treated as the predicted matching probability and the normalized label vector $\overline{\mathbf{y}}$ can denote ground-truth label distribution.

$$\begin{aligned} L_I &= D_{KL}(\overline{\mathbf{S}}, \overline{\mathbf{y}}) \\ L_T &= D_{KL}(\overline{\mathbf{S}'}, \overline{\mathbf{y}'}) \end{aligned} \tag{5}$$

Finally, the Cross-Scale Alignment Loss (CSAL) is calculated by,

$$L_{CSAL} = L_I + L_T \tag{6}$$

Global matching. Besides the cross scale alignment with mixed features, we additionally use CMPM loss and CMPC loss to supervise the cross modal matching of the global-scale features. CMPM loss computes the matching probability of the proposal-text pair by calculating a scalar projection $\mathbf{p}_{i,j}$

TABLE I: Performance comparison of the baseline methods and the proposed method on CUHK-SYSU-TBPS and PRW-TBPS.

Method	CUHK-SYSU-TBPS				PRW-TBPS			
	mAP(%)	top-1(%)	top-5(%)	top-10(%)	mAP(%)	top-1(%)	top-5(%)	top-10(%)
OIM+BiLSTM	23.74	17.41	38.48	49.21	4.58	6.66	16.33	22.99
NAE+BiLSTM	23.48	16.62	38.45	49.66	5.20	7.54	17.21	24.11
BSL+BiLSTM	26.91	20.97	42.31	52.31	3.60	6.42	15.41	22.46
OIM+BERT	43.39	36.59	62.03	72.66	8.52	14.44	30.68	39.77
NAE+BERT	45.70	39.14	64.62	74.34	9.20	14.44	31.55	39.91
BSL+BERT	48.39	40.83	67.52	76.86	10.70	16.82	34.86	45.36
Ours	**50.36**	**49.34**	**74.48**	**82.14**	**11.93**	**21.63**	**42.54**	**52.99**

of global proposal feature $\mathbf{I}_i$ onto the normalized global text feature $\overline{\mathbf{T}}_j$,

$$p_{i,j} = \frac{\exp(\mathbf{I}_i^T \overline{\mathbf{T}}_j)}{\sum_{k=1}^{B} \exp(\mathbf{I}_k^T \overline{\mathbf{T}}_j)}, \tag{7}$$

KL divergence is enforced to measure the distance between the predicted and ground-truth label distributions.

$$L_{i2t} = D_{KL}(\mathbf{p} || \overline{\mathbf{y}}) \tag{8}$$

where $\overline{\mathbf{y}}$ is the same with Eqn. (5). Finally, CMPM loss is computed by summing up L_{i2t} and L_{i2t}.

$$L_{cmpm} = L_{i2t} + L_{t2i} \tag{9}$$

Meanwhile, CMPC is a variant of norm-softmax classification loss. In our setting, it is formulated as,

$$L_{ipt} = -\frac{1}{N} \sum_i \log(\frac{exp(\mathbf{W}_{y_i}^T \hat{\mathbf{I}}_i)}{\sum_j exp(\mathbf{W}_j^T \hat{\mathbf{I}}_i)}),\ \hat{\mathbf{I}}_i = \mathbf{I}_i^T \overline{\mathbf{T}}_i \cdot \overline{\mathbf{T}}_i \tag{10}$$

where $\hat{\mathbf{I}}_i$ represents the vector projection of image feature $\mathbf{I}_i$ onto normalized text feature $\hat{\mathbf{T}}_i$, y_i indicates the label of $\mathbf{I}_i$, and $\mathbf{W}$ denotes the weight matrix.

Similarly, CMPC loss is computed in two directions:

$$L_{cmpc} = L_{ipt} + L_{tpi} \tag{11}$$

D. Overall Loss Function

The whole framework is trained via an end-to-end strategy and pursue the joint optimization of all the loss functions for each task. More specifically, the sub-network for person detection is supervised with a classification loss (L_{cls}), a regression loss (L_{reg}), a RPN objectness loss (L_{rpn_cls}), and a RPN box regression loss (L_{rpn_reg}). While for the supervision of the identification network, the adopted loss function is the OIM loss (L_{oim}). To learn a common feature space for proposals and text descriptions, we adopt CMPM Loss (L_{cmpm}), CMPC Loss (L_{cmpc}), and cross-scale alignment loss (L_{csal}). Therefore, the overall loss function is formulated as:

$$\begin{aligned} L = \lambda_1 L_{rpn\_cls} + \lambda_2 L_{rpn\_reg} + \lambda_3 L_{cls} + \lambda_4 L_{reg} \\ + \lambda_5 L_{oim} + \lambda_6 L_{cmpm} + \lambda_7 L_{cmpc} + \lambda_8 L_{csal} \end{aligned} \tag{12}$$

where $\lambda_1 - \lambda_8$ are responsible for the relative loss importance.

V. Experiment

In this section, we report and analyze the experimental results on the collected datasets. Firstly, we describe the details of datasets and evaluation protocols as well as the implementation details. To verify the effectiveness of the proposed end-to-end approach, we investigate several two-stage solutions as baseline methods. In addition, we conduct ablation studies to analyze the influence of each component in our proposed method. Finally, both quantitative and qualitative results are exhibited.

A. Datasets and Protocol

The collected datasets are built upon the two existing image-based person search datasets CUHK-SYSU and PRW. On CUHK-SYSU-TBPS, The testing set includes 2,900 query descriptions and 6,978 gallery images. For each query, different gallery sizes are set to assess the scaling ability of different models. We use the gallery size of 100 by default. As for PRW-TBPS, the testing set contains 2,057 query persons and each of them are to be searched in a gallery with 6,112 images.

To measure the performance of text-based person search task, the widely adopted mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC top-K) are used as standard metrics. However, different from the conventional retrieval tasks, a candidate in the ranking list would only be considered correct when its IoU with the ground truth bounding box is greater than 0.5.

B. Implementation Details

We take ImageNet-pretrained ResNet-50 to initialize parameters of Base-Net and ID-Net. When obtaining the mixed visual features, the region branch splits the proposal feature map into two stripes equally and the local branch splits the feature map into three stripes equally. The dimension of the visual features at different scales is set to 768-d. For the text semantic feature extraction module, we use BERT-Base-Uncased model as the backbone, which is pretrained on a large corpus including Book Corpus and Wikipedia. The dimension of different scale textual features is also set to 768.

In RPN and SDRPN, we adopt the same anchors and adjust the anchor sizes to the objects in the dataset. We adopt scales $\{8, 16, 32\}$ and aspect ratios $\{1, 2\}$. Non maximum Suppression (NMS) with a threshold of 0.7 is used to filter out redundant boxes and 2000 bounding boxes are left from original 12,000 bounding boxes after NMS. Then, we select

TABLE II: Performance comparison of different components in our method.

BERT	Cross-scale Alignment	SDRPN	CUHK-SYSU-TBPS		PRW-TBPS	
			mAP(%)	top-1(%)	mAP(%)	top-1(%)
√	×	×	48.77	45.86	10.48	19.40
√	√	×	49.15	47.48	11.20	20.52
√	√	√	**50.36**	**49.34**	**11.93**	**21.63**

top 4 proposals for each person identity from 2000 RoIs based on objectiveness score, and meanwhile the IoU of the selected proposals have to be bigger than a threshold of 0.5. During inference, we keep 300 boxes from 6000 bounding boxes and sent them into detection branch. In SDRPN, the reduction ratio r is set to 16. The loss function of SDRPN and RPN are both cross-entropy loss. Note that in SDRPN, anchor boxes that overlap with the text-relevant persons are marked as positives. While in the standard RPN, all persons in the image are positive samples.

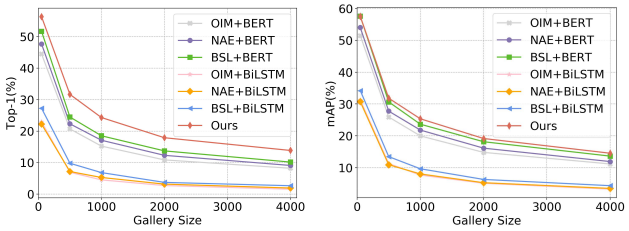
The batch size is set to 4, and we use horizontally flipping as data augmentation. The model contains three groups of parameters, namely detection, identification and projection parameters. The detection parameters are optimized with SGD optimizer with momentum of 0.9, and identification and projection parameters adopt Adam optimizer. The learning rate of three groups parameters are set to 0.0001, 0.001, 0.0001 respectively, and the model is trained for 12 epochs in total. The hyper-parameters of each loss function are set to 1, except the one for CSAL loss which is set to 0.1.

C. Compared Methods

Since there is no existing method specifically designed for text-based person search, we explore typical methods of related tasks and split the task into two parts, detection and text-image alignment, which are combined together as two-stage method to compared with our proposed one stage model.

Specifically, we take fully trained person search model to extract visual features of labeled person image. Also we use language model to extract textual features of language description. The distances between visual feature and text feature are measured under the supervision of CMPM and CMPC loss. During inference, the similarity between the query text and the detected person bounding boxes is calculated based on their embedded features.

The chosen person search method contains OIM [10], NAE [30], and BSL, which are all based on Faster-RCNN while the model architectures are different. In OIM, the box regression and region classification losses remain the same as in Faster-RCNN, with an additional identity classification loss as supervision. In contrast, NAE removes the original region classification branch and uses the embedding norm as the binary person/background classification confidence. BSL is the network used in our framework which is also evaluated as an image-based person search method. Different with OIM and NAE, BSL uses one convolution layer instead of identification net for detection branch, meanwhile the output feature of identification net is directly encoded as final feature vector for matching without further projection to reduce the feature dimension.



(a) Top-1 result (b) mAP result

Fig. 5: Results comparison with different gallery size of CUHK-SYSU-TBPS

As for language model, BiLSTM and BERT are both used as text feature extractors. Notebly, the number of hidden units of BiLSTM is set to 2048 when matching visual features extracted by BSL, otherwise the hidden units number is 256. While, for BERT, we use 1×1 convolution to adjust the shape of text features. All BiLSTM networks are trained for 150 epochs and BERT is trained with 50 epochs.

D. Quantitative and Qualitative Results

Comparison with baseline methods. Table I shows the results of our proposed framework and the compared two stage methods on both the datasets. On CUHK-SYSU-TBPS dataset, our method acheived 49.34% Rank-1 accuracy and 50.36% mAP, which is +8.51% Rank-1 and +1.97% mAP better than the superior compared method BSL+BERT. On the more challenging PRW-TBPS dataset, our method acheived 21.63% Rank-1 accuracy and 11.93% mAP, which is +4.81% Rank-1 and +1.23% mAP better than BSL+BERT method. As can be seen, our approach achieves state-of-the-art performances in terms of both mAP and CMC top-1 to 10 accuracies.

It can also be clearly seen from Table I that (1) When using BERT as the text feature extraction model, it brings significant improvement for our task compared with BiLSTM on both datasets. It indicates that BERT is more capable of encoding complex text descriptions into semantic feature vectors for joint alignment with visual features in a certain way. (2) BSL architecture is more suitable for the task compared with OIM and NAE, as about 1%-3% improvement can be obtained in terms of both CMC top-1 accuracy and mAP on both datasets. We infer that the usage of separate Det-Net and ID-Net for detection and identification, is better for person search model to obtain more accurate location of detection and more discriminative visual features. (3) The proposed end-to-end

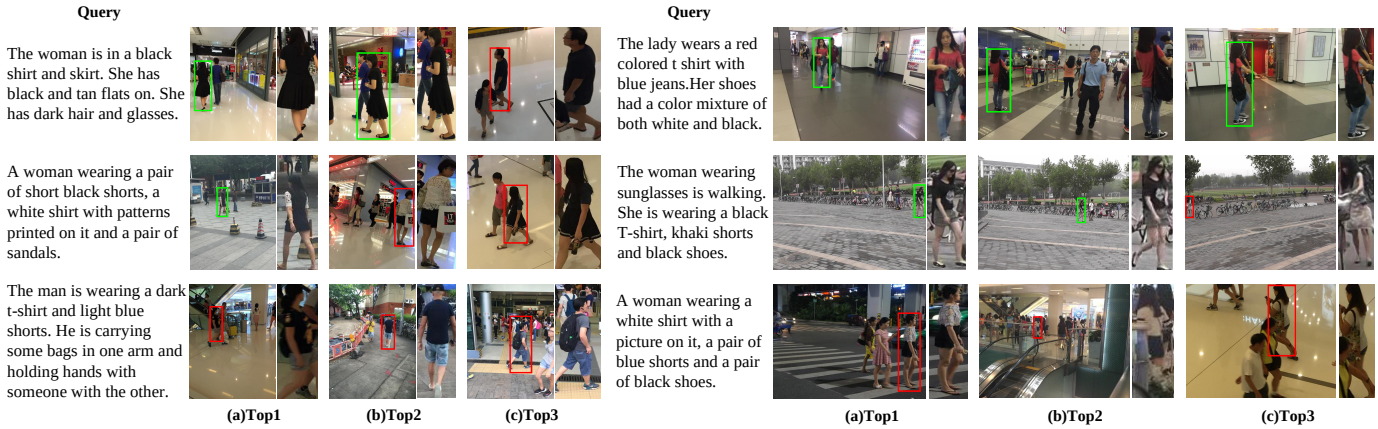


Fig. 6: Examples of text-based person search.

solution has clearly advantages as it can beat all the two-stage counterparts.

Results over varying gallery sizes on CUHK-SYSU-TBPS. As shown in Figure 5, when the gallery size of CUHK-SYSU-TBPS is adjusted from 50, 500, 1000, 2000 to 4000, all of the methods degraded the performance while our method exhibits the consistent advantages compared with others.

Component analysis. We analyze three major components of our method, namely BERT, Cross-scale Alignment and SDRPN, by observing the performance improvement when progressively adding each component. The results are reported in Table II. The first row of Table II is a baseline one stage model which adopts BERT to extract text features with a standard RPN. CMPC and CMPM loss are used for training the model. Note that even the baseline one stage model outperforms the best two stage model. Then, we introduce cross-scale alignment for extracted mixed features and add CSAL loss for joint text-image embedding, which brings +1.62% and +1.12% performance improvement in terms of CMC top-1 accuracy on the two datasets. Based upon that, SDRPN when combined together with standard RPN as aforementioned improves the CMC top-1 accuracy by additional +1.86% and +1.11% on the two datasets, respectively.

Qualitative results. Figure 6 illustrates some text-based person search results. The boxes with green lines represent correct search results, while the boxes with red lines denote failure results. The top 2 rows demonstrate successful cases where correct person boxes are within the top-3 retrieved full images. From these successful cases, we can observe that our method can spot the target person occoured with different angle and background in full scene images. Even though in some cases, like the second case of the middle row in Figure 6, the size of person box is relative small compared to the full scene images, it can also be correctly searched through a text description by our model. Meanwhile, in failure cases, some search results have some characteristics that partially fits the query description, such as the bottom-left case in Figure 6, the first two persons both wear black T-shirt and the third man carries a black backpack. And they all wear blue pants, which are very close to part of the query description.

VI. Conclusion

In this paper, we investigate the problem of text-based person search in full scene images to meet the real-world scenarios where both the query image and the bounding boxes are not available. Specifically, instead of a straightforward two-stage method, we proposed a new end-to-end learning framework which integrated the pedestrian detection, person identification and image-text cross-modal feature embedding tasks together and jointly optimize them to achieve better performance. To take full advantage of the query text description, we devise a Semantic-Driven Region Proposal Network where the proposal generation process is instructed to pay attention to those candidates which are more similar with the semantic features of the text description. Furthermore, a cross-scale visual-semantic feature matching mechanism is introduced to improve the final searching results. To validate the proposed approach, we collect and annotate two large scale text-based person search benchmark datasets named as CUHK-SYSU-TBPS and PRW-TBPS which are built on top of the widely adopted image-based person search datasets CUHK-SYSU and PRW, respectively. We conduct extensive experiments and the experimental results on the two datasets demonstrated that our proposed method achieved state-of-the-art performance compared with many classical baseline methods.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[2] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person re-identification in aerial imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281–291, 2021.

[3] Q. Zhou, H. Fan, S. Zheng, H. Su, X. Li, S. Wu, and H. Ling, "Graph correspondence transfer for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[4] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 420–429.

[5] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.

[6] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4099–4108.

[7] Martinel, Niki, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 858—-877.

[8] Z. Li, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1239–1248.

[9] B. Jiao, X. Tan, L. Yang, Y. Wang, and P. Wang, "Instance and pair-aware dynamic networks for re-identification," *arXiv preprint arXiv:2103.05395*, 2021.

[10] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 3415–3424.

[11] L. Zheng, H. Zhang, S. Sun, Y. Chandrakerand Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 1367–1376.

[12] Z. He and L. Zhang, "End-to-end detection and re-identification integrated net for person search," in *Asian Conference on Computer Vision*, 2018, pp. 349–364.

[13] X. Zhang, X. Wang, J. Bian, C. Shen, and M. You, "Diverse knowledge distillation for end-to-end person search," *arXiv preprint arXiv:2012.11187*, 2020.

[14] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2965–2974.

[15] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 553–569.

[16] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 811–820.

[17] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2158–2167.

[18] B. Munjal, F. Galasso, and S. Amin, "Knowledge distillation for end-to-end person search." in *BMVC*, 2019, p. 216.

[19] K. Islam, "Person search: New paradigm of person re-identification: A survey and outlook of recent works," *Image and Vision Computing*, vol. 101, p. 103970, 2020.

[20] Y. Zhong, X. Wang, and S. Zhang, "Robust partial matching for person search in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6827–6835.

[21] Y. Chen, W. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 3402–3408.

[22] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[23] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 649–656.

[24] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 460–469.

[25] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Tcts: A task-consistent two-stage framework for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 952–11 961.

[26] D. Zheng, J. Xiao, K. Huang, and Y. Zhao, "Segmentation mask guided end-to-end person search," *Signal Processing-image Communication*, vol. 86, p. 115876, 2020.

[27] S. Zhai, S. Liu, X. Wang, and J. Tang, "Fmt: fusing multi-task convolutional neural network for person search," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31 605–31 616, 2019.

[28] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, "Re-id driven localization refinement for person search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9814–9823.

[29] J. Dai, P. Zhang, H. Lu, and H. Wang, "Dynamic imposter based online instance matching for person search," *Pattern Recognition*, vol. 100, p. 107120, 2020.

[30] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 615–12 624.

[31] W. Dong, Z. Zhang, C. Song, and T. Tan, "Bi-directional interaction network for person search," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2839–2848.

[32] X. Chen, W. Liu, X. Liu, Y. Zhang, and T. Mei, "A cross-modality and progressive person search system," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4550–4552.

[33] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, and X. Sun, "Contextual non-local alignment over full-scale representation for text-based person search," *arXiv preprint arXiv:2101.03036*, 2021.

[34] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 189–11 196.

[35] 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*, vol. 29, pp. 5542–5556, 2020.

[36] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 54–70.

[37] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 686–701.

[38] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5814–5824.

[39] Q. Dong, X. Zhu, and S. Gong, "Person search by text attribute query as zero-shot learning," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019, pp. 3652–3661.

[40] J. Liu, Z. 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*, 2019, pp. 665–673.

[41] Z. Ji, S. Li, and Y. Pang, "Fusion-attention network for person search with free-form natural language," *Pattern Recognition Letters*, vol. 116, pp. 205–211, 2018.

[42] Y. Jing, W. Wang, L. Wang, and T. Tan, "Cross-modal cross-domain moment alignment network for person search," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 678–10 686.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[44] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 937–940.

[45] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 764–781.

[46] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *Proceedings of the IEEE international conference on computer vision*, 2017, p. 493–501.

[47] X. Chang, P. Huang, Y. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "Rcaa: Relational context-aware agents for person search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 84–100.

[48] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[49] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y. Shen, "Dual-path convolutional image-text embedding with instance loss," *arXiv preprint arXiv:1711.05535*, 2017.

[50] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1890–1899.

[51] 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)*. IEEE, 2018, pp. 1879–1887.

[52] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 744–11 752.
[53] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 1432–1442.
[54] Z. Fan, J. Yu, Z. Liang, J. Ou, C. Gao, G. Xia, and Y. Li, "Fgn: Fully guided network for few-shot instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9172–9181.
[55] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.
[56] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.
[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.