# A Simple and Robust Correlation Filtering method for text-based person search

Wei Suo[1,4], Mengyang Sun[2,4], Kai Niu[1,4], Yiqi Gao[1,4], Peng Wang[1,4†], Yanning Zhang[1,4†], and Qi Wu[3]

[1] School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, China.
[2] School of Cybersecurity, Northwestern Polytechnical University, China.
[3] University of Adelaide, Australia
[4] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, China.
{suowei1994, sunmenmian, gyqiz}@mail.nwpu.edu.cn, {kai.niu, ynzhang, peng.wang}@nwpu.edu.cn, qi.wu01@adelaide.edu.au

**Abstract.** Text-based person search aims to associate pedestrian images with natural language descriptions. In this task, extracting differentiated representations and aligning them among identities and descriptions is an essential yet challenging problem. Most of the previous methods depend on additional language parsers or vision techniques to select the relevant regions or words from noise inputs. But there exists heavy computation cost and inevitable error accumulation. Meanwhile, simply using horizontal segmentation images to obtain local-level features would harm the reliability of models as well. In this paper, we present a novel end-to-end Simple and Robust Correlation Filtering (SRCF) method which can effectively extract key clues and adaptively align the discriminative features. Different from previous works, our framework focuses on computing the similarity between templates and inputs. In particular, we design two different types of filtering modules (*i.e.,* denoising filters and dictionary filters) to extract crucial features and establish multimodal mappings. Extensive experiments have shown that our method improves the robustness of the model and achieves better performance on the two text-based person search datasets. Source code is available at https://github.com/Suo-Wei/SRCF.

**Keywords:** Text-based Person Search, Correlation Filtering, Vision and Language

## 1 Introduction

Text-based person search [17,16,6] aims to retrieve the corresponding pedestrian in an image database by given language descriptions, which provides various potential applications such as missing person searching and suspects tracking. As

† Corresponding author


**Fig. 1.** (a) The illustration of the text-based person search. The left texts represent the language-based queries, and the right image indicates the corresponding pedestrian. Models are required to extract differentiated clues from noisy queries and images. Then aligning the body parts and corresponding descriptions in a common space. (b) The monitoring gallery we expect has clean backgrounds and the body parts of pedestrians are evenly distributed in each stripe to help the model achieve local alignments. (c) The actual gallery is with a cluttered background, and it is difficult to ensure that the body parts are fixed due to the changes in the viewpoints and pedestrian postures.

shown in Fig.1(a), to retrieve the pedestrian "the woman with the white dotted backpack . . . a pair of brown shoes.", models must have the ability to collect the differentiated clues (*e.g.* "black shirt, jean shorts") and relevant regions from noisy inputs. Then aligning these features in a common space (*e.g.* "white dotted backpack" and corresponding regions). Hence, the main challenges of text-based person search are how to accurately localize the discriminative regions (or key words) and how to align these multi-modal representations.

To refine crucial visual and textual contents from the background noise, previous methods usually depend on additional tools (*e.g.* the NLTK [1,23,39] and image parsing [14,30]) or attention mechanism [37,5,22,8] to localize key words and regions of interest (ROIs). However, on the one hand, these frameworks are capped by the tools with an inevitable error accumulation (if the discriminative noun phrases or significant body parts can not be captured, the retrieval capability of models would be weakened). Besides, these auxiliary operations could bring a huge amount of computational expenses, it would destroy the real-time requirement in video surveillance. On the other hand, the attention mechanism would ignore some relatively important regions due to the sum of attention weights being 1 among all image or text features.

For the second issue, since the parts of the human body are evenly arranged in the images as shown in Fig. 1 (b), existing methods mostly use sliced images

as supervision to guide models to achieve multi-modal alignment [8,5,22,4,7]. However, this strategy is vulnerable and sensitive to the changes of viewing conditions. For example, as shown in Fig. 1 (c), "head" does not always occur in the first strip. This phenomenon would significantly hamper the robustness of existing methods (more details in section 4.3) and make them difficult to apply in practice. For the language domain, simply using fully-connected layers [5] or adding tokens [8] to model the complex and changeable language is also difficult. As we see in Fig. 1(a), there exist enormous differences in linguistic descriptions, even facing the same person.

In order to solve the above challenges, we propose a novel Simple and Robust Correlation Filtering (SRCF) framework that through building a group of general semantic-templates (*i.e.,* filters) to effectively extract the key clues and adaptively align the multi-modal features without any auxiliary tools. The critical intuition behind SRCF is that no matter how the language or image changes, distinguishing components are unchanging. As far as we know, our paper is the first work that utilizes the idea of filtering to solve text-based image retrieval. Compared with the previous attention-based methods [37,5,22] which aim to learn a group of sparse wights by derived from inputs, our framework focuses on computing the similarity between templates and inputs. Specially, we design two different types of filtering modules (*i.e.,* denoising filters and dictionary filters) to achieve foreground separation and multi-modal alignment. Based on this conclusion that the similarity between the foreground is greater than the similarity with the background [34], we design two lightweight but effective denoising filters to help model separate pedestrian regions as well as meaningful words. Based on this insights that all pedestrians have the same body parts (such as heads, upper body, legs) and corresponding descriptions, we propose a novel matching search method with dictionary filters. It can be dynamically updated by moving-averaged strategy in the forward propagation. Moreover, the response maps from each dictionary filter would be used to local the specific semantic features with global search. As for flexible and diverse language, similar dictionary filters are also adopted to find out the words that correspond to body parts. To summarize, the main contributions of our paper are as follows:

1) We propose a novel simple and robust end-to-end method SRCF which can effectively extract key clues and adaptively align local features without any helps from external tools. 2) We design two lightweight denoising filters to refine the regions of interest (or meaningful words) from noisy inputs. Meanwhile, we build the dictionary filters to align body parts and corresponding texts by matching search. 3) The proposed SRCF not only achieves new state-of-the-art performance on CUHK-PEDES [17] and ICFG-PEDES [5] benchmark datasets, but also has better robustness and reliability compared with previous methods.

## 2 Related Work

**Text-based person search.** Text-based person search is the task of finding the best matched pedestrian with a ranking strategy based on a given expression.

This task needs to separate the distinguished regions from cluttered inputs and align these features in a common space. Previous methods usually depend on extra tools to extract useful information [23,30,1,22,38,14]. Specially, [22] focuses on multi-granularity alignment, they use extra alignment between the global image and noun phrases, as well as horizontal image stripes and the whole sentence to build the cross-granularity mapping. To localize the ROIs in images, [30] builds an additional attribute segmentation strategy to guide the alignment. [14] introduce a multi-granularities attention structure to align vision-and-language local information with human pose estimation.

Due to the above approaches being sensitive to the reliability of the external tools, several text-based person search methods are proposed to avoid the preprocessing and reduce additional computation such as [26,8,9,5]. [26] attempts to learn modality-invariant representations in a shared space by adversarial learning. [8] utilizes joint alignments over multi-scaled representations with a novel structure and a locality-constrained BERT. [5] uses an attention mechanism to select words in sentences and achieve joint alignments over full-scale representations. [33] constructs a representation learning approach, which depends on color-reasoning sub-tasks to align the cross-modal representations.

Different from the above methods, SRCF can effectively extract the key regions and discriminative words with an end-to-end training. Moreover, based on dictionary filters and expanded search space, our method also improves the robustness and reliability of the model.

**Correlation Filtering.** The correlation filtering is a popular technology that is used in many different fields. [2] introduces the correlation filter into the Fourier domain and achieves fast and accurate tracking. Classification network [15,27] can be translated into a correlation filtering task, where the output of global pooling is a filter kernel and weight matrix is search space. [21] propose vector correlation filter (VCF), which can adapt remarkable within-class variations while being discriminative against background. [18] reformulates the referring expression comprehension as a correlation filtering process. The text is converted as a filter and performs correlation filtering on the image.

Inspired by previous works, we introduce two different types of filters to achieve the text-based person search. For global-level alignment, the learnable filters are used to help model separate the discriminative foreground and the parameters are learnable by normal Back Propagation (BP) [11]. While the dictionary filters would be updated only in the forward propagation and it would be utilized to alleviate the difficulties of alignments due to the changes of viewpoints and the order of the words.

## 3 Our Approach

In this section, we introduce our SRCF framework. Our aim is to find the best matched pedestrian by the queries expression without any pre-processing. The SRCF is composed of a "global-level alignment" module and a "local-level alignment" module, as shown in Fig. 2.

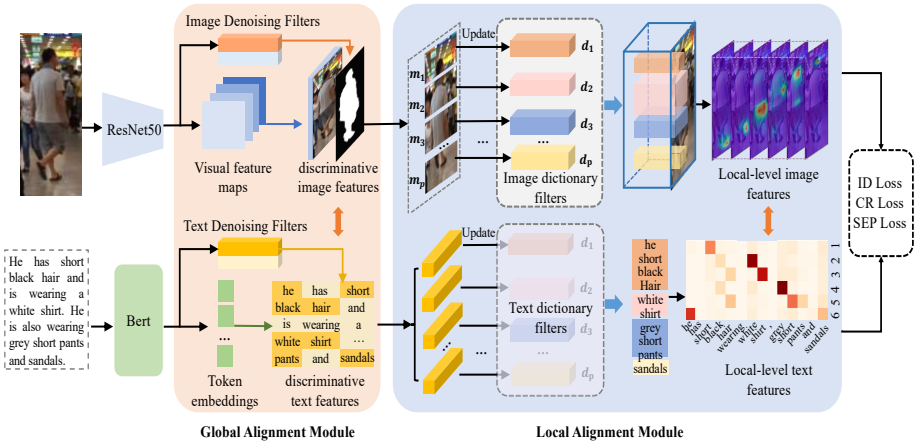


**Fig. 2.** The architecture for our framework. It consists of three main modules: (a) Encoder module: CNN and BERT models are used for image and language features extraction, respectively. (b) Global Alignment module: this module utilizes denoising filters to separate the discriminative regions and words. (c) Local Alignment module: this module utilizes dictionary filters to model local alignments by correlation matching in the global scope.

Different from previous methods, our model focus on extracting the key clues and adaptively aligning image-text inputs. Specially, representations of queries and images are first extracted by the language and visual encoders respectively. Image features would be further divided into "foreground" and "background" parts based on denoising filters. Then we feed the foreground features to the local-level alignment module, where several dictionary filters are built to learn the correlation between the body parts of different persons. Similar calculations also are executed in the language domain. Next, we describe the components of our model in details.

### 3.1 Encoder

**Visual encoder.** For each given pedestrian image $\boldsymbol{I} \in R^{W \times H \times 3}$, where $W \times H \times 3$ denotes the size of the image, we follow [8,5] that adopt the ResNet-50 [10] model as the backbone to extract visual features. Specially, we first resize the given image $\boldsymbol{I}$ to the size of $3 \times 384 \times 128$. Then it is fed into the encoder network to obtain the feature maps $\boldsymbol{G} = \{\boldsymbol{g}_i\}_{i=1}^{w \times h}, \boldsymbol{g}_i \in R^d$, where the feature maps spatial resolution is $w \times h$, and each $\boldsymbol{g}_i$ represents a grid feature for the output feature map $\boldsymbol{G}$.

**Language encoder.** Following [8,33,37], we use the uncased BERT[29] as our language encoder. For a given query $Q = \{q_t\}_{t=1}^{T}$, where $q_t$ represents the $t$-th word in this sentence, each word in this description is first mapped to the corresponding word embedding. Then, each $q_t$ and its index $t$ ($q_t$'s absolute position

in the sentence) is fed into language encoder. Before entering subsequent modules, we remove special tokens such as [CLS], [SEP] and [PAD] due to semantic fuzziness. Finally, we obtain text feature $\boldsymbol{E} = \{\boldsymbol{e}_t\}_{t=1}^{T}, \boldsymbol{e}_t \in R^d$, note that here we add a $1 \times 1$ convolution layer with batch normalization and RELU to map them all to the same dimension as images $d$.

### 3.2 Global-level alignment module

Inspired by [34,39], from a global-level point of view, each pedestrian image can be simply divided into two parts ("foreground" and "background"). We expand this idea to language domain and achieve tool-free separation. For simplicity, the "differentiated words" and the "undifferentiated words" in the sentences are also called "foreground" and "background", respectively. Intuitively, foregrounds are beneficial to retrieval, instead, the background noise would be harmful to the models. To filter out these noises, we introduce the lightweight denoising filters which contain foreground filters and background filters to learn the global-level correlation of foregrounds and background respectively.

Taking images for an example, the input of global-level alignment module is image features $\boldsymbol{G} = \{\boldsymbol{g}_i\}_{i=1}^{w \times h}$. As shown in Fig. 2, we set image denoising filters as two learnable filters that are named "foreground filter" $\boldsymbol{v}_f \in R^d$ and "background filter" $\boldsymbol{v}_b \in R^d$, respectively. We first use $\boldsymbol{v}_f$ and $\boldsymbol{v}_b$ to compute the similarity with the visual feature map $\boldsymbol{G}$, which can be formulated as:

$$\begin{aligned} s_i^f &= \frac{\boldsymbol{v}_f^{\mathrm{T}} \boldsymbol{g}_i}{\|\boldsymbol{v}_f\| \, \|\boldsymbol{g}_i\|}, \\ s_i^b &= \frac{\boldsymbol{v}_b^{\mathrm{T}} \boldsymbol{g}_i}{\|\boldsymbol{v}_b\| \, \|\boldsymbol{g}_i\|}, \end{aligned} \tag{1}$$

where $\|\cdot\|$ represents the $L_2$-normalization, the $s_i^f$ and $s_i^b$ are response maps which are utilized to estimate if the $g_i$ should be accepted. We would divide images into two classes rely on $s_i^f$ and $s_i^b$. However, because the foreground annotations are unavailable and simply using the strategy of heuristic tuning (such as setting threshold) would introduce additional hyper-parameters as well [37]. Therefore, we add a mutual-exclusion-loss [35] to ensure the $\boldsymbol{s}^f$ and $\boldsymbol{s}^b$ are orthogonal to each other (more details in section 3.4). Then, the foreground $\boldsymbol{G}^f = \{\boldsymbol{g}_i^f\}_{i=1}^{w \times h}$ can be obtained by

$$a_i^f, a_i^b = \mathrm{softmax}([s_i^f; s_i^b]), \tag{2}$$

$$\boldsymbol{g}_i^f = a_i^f \boldsymbol{g}_i, \tag{3}$$

where [;] indicates to concatenate these two response maps $s_i^f$ and $s_i^b$, we compute the softmax over the response maps to obtain foreground and background response scores $\boldsymbol{a}_i^f$ and $\boldsymbol{a}_i^b$, and then all $a_i^f$ and $a_i^b$ are jointed spatially to get $a^f$ and $a^b$ respectively. Moreover, we also perform a similar calculation in the domain of language and obtain the differentiated words $\boldsymbol{E}^f = \{\boldsymbol{e}_t^f\}_{t=1}^{T}$. To align

images and language, we utilize Global MAX Pooling (GMP) on $\boldsymbol{G}^f$ and $\boldsymbol{E}^f$ to obtain the global-level image feature $\boldsymbol{g}_g \in R^d$ and text feature $\boldsymbol{e}_g \in R^d$. The similarity matrix between global-level features of one image-text pair is denoted as follow:

$$\boldsymbol{S}_g = \frac{\boldsymbol{g}_g^{\mathrm{T}} \boldsymbol{e}_g}{\|\boldsymbol{g}_g\| \, \|\boldsymbol{e}_g\|} \tag{4}$$

### 3.3 Local-level alignment module

As discussed in the previous section, most of previous methods adopt horizontal segmentation images and multi-branch fully-connected layers [22,23,5,8] to offer local-level features. But they are vulnerable and sensitive to various viewpoints and the order of the words. To achieve the local-level alignments, we build the dictionary filters which can learn the correlation of body parts between different pedestrians and improve the robustness of model.

**Dictionary filters.** As shown in Fig. 2, we introduce the dictionary filters and a strategy of momentum update to learn the body parts correlation between different persons. Taking images for example, following [22,23,5], the output of global-level alignment module $\boldsymbol{G}^f$ is first segmented into $P$ horizontal stripes, which are denoted as $\{\boldsymbol{G}_1^f, \boldsymbol{G}_2^f, \cdots, \boldsymbol{G}_P^f\}$, $\boldsymbol{G}_p^f \in R^{\frac{h}{P} \times w \times d}$. For each strip $\boldsymbol{G}_p^f$, the GMP is used to obtain body parts feature vector $\boldsymbol{m}_p \in R^d$. Then, we define visual dictionary filters as $\boldsymbol{D}_g \in R^{P \times d}$, which is randomly initialized and further updated by a moving average operation [13] in one mini-batch. It can be formulated as:

$$\hat{\boldsymbol{d}}_p = \alpha * \boldsymbol{d}_p + (1 - \alpha) * \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}_{pi}, \tag{5}$$

$$\boldsymbol{m}_p = \mathrm{GMP}(\boldsymbol{G}_p^f), \tag{6}$$

where $\hat{\boldsymbol{d}}_p$ indicates that the $p$-th filter in the $\boldsymbol{D}_g$ is updated, the $\alpha$ and $n$ are momentum coefficient and batch size.

**Expanded search space.** Different from previous methods that simply use striped images to extract local-level features, we expand the search space to global scope. Specially, each filter $\hat{\boldsymbol{d}}_p$ in the dictionary would compute one response map with $\boldsymbol{G}^f$, then the softmax is utilized to obtain response scores. The local-level image features $\boldsymbol{g}_{lp}$ are obtained by summarizing $\boldsymbol{g}_i^f$ based on corresponding scores. The above computations can be formulated as:

$$s_i^p = \frac{\hat{\boldsymbol{d}}_p^{\,\mathrm{T}} \boldsymbol{g}_i^f}{\left\|\hat{\boldsymbol{d}}_p\right\| \left\|\boldsymbol{g}_i^f\right\|}, \tag{7}$$

$$a_1^p, a_2^p, \cdots, a_{w \times h}^p = \mathrm{softmax}(s_1^p, s_2^p, \cdots, s_{w \times h}^p), \tag{8}$$

$$\boldsymbol{g}_{lp} = \sum_{i=1}^{w \times h} (a_i^p \cdot \boldsymbol{g}_i^f). \tag{9}$$

Note that our dictionary filters are updated only in the forward propagation and use "straight-through" trick[24] to avoid the gradient stop. As for language, we use multi-branch fully connected layers to obtain $P$ language features followed by [8,5]. Then the similar dictionary filters and updating mechanism are implemented to learn the correlation for body parts' descriptions. Finally, we can obtain the similarity matrices $\boldsymbol{S}_l = \{\boldsymbol{S}_{lp}\}_{p=1}^{P}$. To align the local-level images and language, the $\boldsymbol{S}_{lp}$ is denoted as follow:

$$\boldsymbol{S}_{lp} = \frac{\boldsymbol{g}_{lp}^{\mathrm{T}} \boldsymbol{e}_l}{\|\boldsymbol{g}_{lp}\| \, \|\boldsymbol{e}_l\|}, \tag{10}$$

where $\boldsymbol{e}_l$ denotes local-level language features. During testing, we would directly obtain the local-level features based on updated dictionary filters and the strategy of horizontal partitioning is discarded. Following the previous work[5], we also use a transformer-style non-local module to establish the connection between local-level features. The outputs are used to obtain non-local similarity matrices $\boldsymbol{S}_n$.

### 3.4 Optimization

Compound Ranking (CR) loss $L_{cr}$ is utilized to optimize the $\boldsymbol{S}_g$, $\boldsymbol{S}_l$ and $\boldsymbol{S}_n$ respectively. In particular, $L_{cr}$ applies a constraint that the intra-class similarity score must be larger than the inter-class similarity by a margin. Meanwhile, $L_{cr}$ also exploits more diversely textual descriptions as complementary sentences for each image. More information on $L_{cr}$ can be found in the [5].

Besides, following [5,39,8,14], we also add ID loss $L_{id}$ to achieve identity-level matching. More importantly, it is also used as labels for our foreground and differentiated words in this paper. For $\boldsymbol{g}_g$ and $\boldsymbol{e}_g$, the identification loss is defined as follows:

$$\begin{aligned} L_{id} = & - (\boldsymbol{y}_{id} \log(\mathrm{softmax}(\boldsymbol{W}_{id} \boldsymbol{g}_g)) \\ & + \boldsymbol{y}_{id} \log(\mathrm{softmax}(\boldsymbol{W}_{id} \boldsymbol{e}_g))), \end{aligned} \tag{11}$$

where $\boldsymbol{W}_{id}$ is a shared transformation matrix to classify the different persons and $\boldsymbol{y}_{id}$ is the ground true identity. In addition, a mutual-exclusion-loss is used to separate the response maps:

$$L_{sep} = \left\| \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \odot (\boldsymbol{1} - \boldsymbol{K}) \right\|_F^2, \tag{12}$$

where matrix $\boldsymbol{A}$ is the $\boldsymbol{a}^f$ and $\boldsymbol{a}^b$ in Eq. 2 by concatenated in the last dimension. $\boldsymbol{1}$ is the matrix of ones and $\boldsymbol{K}$ is an identity matrix. The total loss in SRCF is defined as:

$$Loss = L_{cr} + \lambda_1 L_{id} + \lambda_2 L_{sep}, \tag{13}$$

following [5] and [35], where $\lambda_1$ and $\lambda_2$ is set to 1.

At inference time, the similarity score between a text-image pair is the sum of $\boldsymbol{S}_g$, $\boldsymbol{S}_l$ and $\boldsymbol{S}_n$. Note that our model completely avoided the traditional horizontal partitioning during testing and local-level features are straightforward obtained by dictionary filters.

# 4 Experiments

## 4.1 Experimental Setting

**Datasets.** We evaluate the proposed SRCF on the CUHK-PEDES [17] dataset and ICFG-PEDES [5] dataset. The CUHK-PEDES is the first large-scale dataset for text-to-image person search. It contains totally 80,412 textual descriptions for 13,003 different persons in 40,206 pedestrian images. This dataset is split into 34,054 images for 11,003 identities with 68,108 descriptions in the training set, 3,078 images for 1,000 identities in validation set and the test set contains 3,074 images for 1,000 persons.

Recently, a new dataset ICFG-PEDES [5] is conducted, this dataset has more identities and textual descriptions. It contains 54,522 pedestrian images from MSMT17 [31] of 4,102 different persons. The ICFG-PEDES is divided into 34,674 image-text pairs of 3,102 identities in the training set, and the test set contains 19,848 image-text pairs for 1,000 persons.

**Implementation Details.** Following previous methods [8,33], we also use ResNet-50 pretrained on ImageNet [25] as our visual backbone and adopt the BETR-Base-Uncase [29] for textual encoder. We follow [5] to resize an input image to $384 \times 128$ and use random horizontal flipping for data augmentation [30]. The size of the feature map is $24 \times 8 \times 2048$. The text length and the number of body part dictionary is set to 64 and 6, respectively. We follow [13] to set the momentum coefficient $\alpha$ to 0.99.

During training, we use Adam as our optimizer and the batch size is set to 32. The initial learning rate of our overall model is $5e-5$ and decreased by 0.1 per 10 epochs after 20 epochs. Following [8], the BERT encoder is frozen and we fine-tune the visual backbone with the learning rate to 1/10 of the whole model. Our dictionary filters are only updated to 25 epochs. We train our model on one 2080Ti GPU for 60 epochs.

**Evaluation.** Following the standard evaluation setting, we adopt top-K accuracy (K=1, 5, 10) as our evaluation metric. Specially, given a person description, the search is considered as correct if top-K images contain at least one corresponding person to the given description.

## 4.2 Quantitative Results

We compare our proposed SRCF with the state-of-the-art methods on the CUHK-PEDES dataset and ICFG-PEDES dataset, as shown in Table 1. For a fair comparison, all methods use the ResNet-50 as the backbone to extract visual representations. Depending on which query embedding is adopted, we divide these approaches into two types: LSTM-based methods and BERT-based methods. In the "Language or Image Parsing" column, we list whether the approach uses additional language parsing tools or image pre-processing methods.

In Table 1, it can be observed that although these networks based on pre-processing methods [22,14,30,1,39,23,37] also achieve good performance, they have to depend on additional language parsing [20,19,12] or vision models [3,28]

**Table 1.** Comparisons with the state-of-the-art methods on the CUHK-PEDES and ICFG-PEDES datasets. We report Top-1, Top-5 and Top-10 accuracies.

| Methods | Language embedding | Language or Image Parsing | CUHK-PEDES | | | ICFG-PEDES | | |
|---|---|---|---|---|---|---|---|---|
| | | | Top-1 | Top-5 | Top-10 | Top-1 | Top-5 | Top-10 |
| Dual Path [38] | Word2vec | - | 44.40 | 66.26 | 75.07 | 38.99 | 59.44 | 68.41 |
| CMPM/C [36] | LSTM | - | 49.37 | - | 79.27 | 43.51 | 65.44 | 74.26 |
| MIA [22] | LSTM | NLTK [19] | 53.10 | 75.00 | 82.90 | 46.49 | 67.14 | 75.18 |
| CMPM/C+TC&IC [33] | LSTM | - | 53.33 | - | 83.20 | - | - | - |
| PMA [14] | LSTM | NLTK [19]+Pose [3] | 53.81 | 73.54 | 81.23 | - | - | - |
| ViTAA [30] | LSTM | CoreNLP [20]+Seg [28] | 55.97 | 75.84 | 83.52 | 50.98 | 68.79 | 75.78 |
| CMAAM [1] | LSTM | NLTK [19] | 56.68 | 77.18 | 84.86 | - | - | - |
| DSSL [39] | LSTM | NLTK [19] | 59.98 | 80.41 | 87.56 | - | - | - |
| SSAN [5] | LSTM | - | 61.37 | 80.15 | 86.73 | 54.23 | 72.63 | 79.53 |
| TIMAM [26] | BERT | - | 54.51 | 77.56 | 84.78 | - | - | - |
| TDE [23] | BERT | SpaCy [12] | 55.25 | 77.46 | 84.56 | - | - | - |
| CMP_adv [33] | BERT | - | 55.05 | - | 85.09 | - | - | - |
| CMP_adv+TC&IC [33] | BERT | - | 57.00 | - | 85.62 | - | - | - |
| HGAN [37] | BERT | NLTK [19] | 59.00 | 79.49 | 86.62 | - | - | - |
| NAFS [8] | BERT | - | 59.94 | 79.86 | 86.70 | - | - | - |
| SRCF-LSTM (ours) | LSTM | - | **62.87** | **81.81** | **87.85** | **55.69** | **73.07** | **80.84** |
| SRCF-BERT (ours) | BERT | - | **64.04** | **82.99** | **88.81** | **57.18** | **75.01** | **81.49** |

to extract useful information. Instead, our proposed SRCF can adaptively extract key clues by end-to-end training with higher accuracy.

In addition, we compare the performance of our method with the methods without any tools [5,8,26,33,36,38]. We observe that our SRCF uses the denoising filters and the dictionary filters to achieve better performance than the previous methods. Specially, in CUHK-PEDS dataset, we observe that our method outperforms the start-of-the-art method [5] and [8] 1.5% and 4.1% respectively on Top-1 accuracies. In ICFG-PEDES dataset, the performance of SRCF over [5] 1.46% with LSTM and 2.95% with BERT.

### 4.3 Robustness of model

In this section, we further design experiments to verify the robustness and reliability of our SRCF by simulating real-world situations. As shown in the Fig. 3, the first row is the original gallery image examples from CUHK-PEDES dataset (short for "Raw"). In 2-5 rows of Fig. 3, we implement four different experimental settings on this dataset which include random horizontal translation, random vertical translation, random rotating and random cropping (they are short for "New").

We use the weights trained on the "Raw" dataset provided by the state-of-the-art methods [8,5] and test them on the "New" galleries. The Top-1 accuracy results between our proposed SRCF and the state-of-the-art methods [8,5] are shown in the three columns at the right side of Fig. 3. In order to compare the decline degree under these four different galleries, we use the red font to represent the largest decrease, the green font for the second, and the blue font for the lowest percentage decrease.

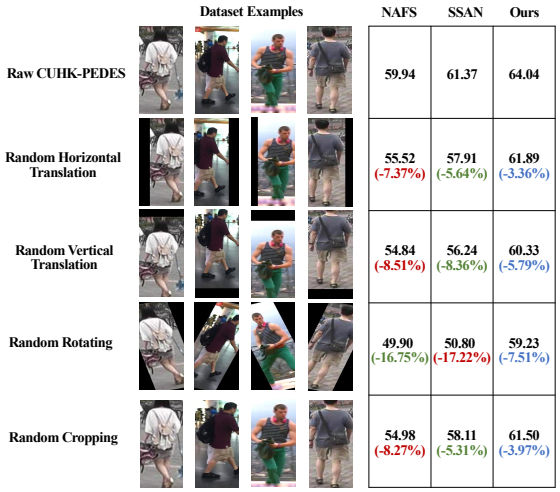


**Fig. 3.** Compare the performance between our method and the state-of-the-art methods under four different experimental settings, which include horizontal translation, vertical translation, rotating and cropping on the original CUHK-PEDES dataset. We use red font to represent the largest decrease, the green font for second, and the blue font for the lowest percentage decrease.

As we mentioned in Sec. 1, the state-of-the-arts methods [8,5] are vulnerable and sensitive to changeable conditions. For example, with the upper and lower offset of the pedestrian position in the third row of the Fig. 3, both NAFS [8] and SSAN [5] methods have declined more than 8%. Furthermore, when we rotate the person images randomly, the retrieval accuracy of NAFS and SSAN methods even decreases by more than 15%. It is worth noting that our SRCF has the lowest percentage decline in all cases. These quantitative results effectively demonstrate the robustness and stability of our model under changeable conditions.

### 4.4 Ablation Studies

In this section, we conduct several ablation studies on the CUHK-PEDES to demonstrate the effectiveness.

**Effectiveness of main modules.** First, to systematically evaluate the contributions of different model components, we design various ablation experiments. As shown in Table 2, "Global" and "Local" are global-global and local-local alignment, respectively. Specially, "Global" indicates Max-pooling is performed on the output of CNN and BERT. Then we use ranking loss and identification loss to align them. "Local" indicates images are horizontally partitioned into 6 parts and six textual fully connected layers are used to align corresponding the stripes of images. We observe that the performance can increase by multi-grained alignment and they would be used as our baselines. In Table 2, the "Image

**Table 2.** Ablation studies on the CUHK-PEDES dataset.

| | Global | Local | Image Denoising Filters | Text Denoising Filters | Image Dictionary Filters | Text Dictionary Filters | Top-1 | Top-5 | Top-10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ✓ | | | | | | 59.76 | 79.82 | 86.45 |
| 2 | ✓ | ✓ | | | | | 61.27 | 80.43 | 87.07 |
| 3 | ✓ | ✓ | ✓ | | | | 62.59 | 81.56 | 87.88 |
| 4 | ✓ | ✓ | | ✓ | | | 62.69 | 81.40 | 87.93 |
| 5 | ✓ | ✓ | ✓ | ✓ | | | 63.16 | 81.82 | 88.14 |
| 6 | ✓ | ✓ | ✓ | ✓ | ✓ | | 63.60 | 82.55 | 88.74 |
| 7 | ✓ | ✓ | ✓ | ✓ | | ✓ | 63.71 | 82.86 | 88.61 |
| 8 | ✓ | ✓ | | | ✓ | ✓ | 63.19 | 81.34 | 88.16 |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **64.04** | **82.99** | **88.81** |

**Table 3.** The effects of different model settings.

| | Method | Top-1 | Top-5 | Top-10 |
|---|---|---|---|---|
| 1 | SA + WA + Dictionary filtering | 61.89 | 80.77 | 87.24 |
| 2 | Denoising filtering + CNLA | 60.85 | 80.15 | 87.31 |
| 3 | Denoising filters moving-update | 62.69 | 81.35 | 87.64 |
| 4 | Dictionary filters BP-update | 62.15 | 81.61 | 87.67 |
| 5 | Share denoising filters | 61.99 | 81.34 | 87.70 |
| 6 | Share dictionary filters | 61.45 | 80.70 | 87.62 |
| 7 | Avg-pooling | 63.97 | 82.31 | 88.40 |
| 8 | Without mutual-exclusion-loss | 62.44 | 81.61 | 87.43 |
| 9 | Ours (Denoising filtering + Dictionary filtering) | **64.04** | **82.99** | **88.81** |

(Text) Denoising Filters" denotes denoising filters in image domain or language domain. The column of "Image (Text) Dictionary Filters" indicates image-based or language-based dictionary filters. As shown in 3-5 rows of Table 2, we follow the baseline and add the two learnable denoising filters to demonstrate the effectiveness of our method. The results show that denoising filters can effectively improve the performances and combining them in both domains would boost the baseline accuracy by 3.40% and 1.89% on Top-1. In addition, in 8 row of Table 2, we find employing "Dictionary Filters" in both domains can outperform the baseline by 3.43% and 1.92% on Top-1, respectively. Meanwhile, the language-based and image-based dictionary filters have similar contributions to the model (in 6-7 rows of Table 2). While, when we integrate the denoising filters and dictionary filters together, the performance improves to 64.04 on Top-1 over the baseline model by 4.28% and 2.77%.

**Alternative model settings.** In Table 3, we exploit some comparative experiments about the different model settings. We first compare the performance of our method with attention-based methods. In the first row of Table 3, we use popular Space Attention (SA) [32] and Word Attention (WA) [5] as global attention to replace our denoising filters. In row 2, we utilize Contextual Non-Local


**Fig. 4.** Visualizations of response maps of images and texts from denoising filters.

Attention (CNLA) as local-level attention module from SOTA method [8] to take the place of our dictionary filters. The results prove that our correlation-based approach is significant better than previous attention-based methods.

In the third row of Table 3, the moving-averaged mechanism is used to replace the back propagation for our "Denoising filters". The results show this strategy of moving-average is not suitable for localizing the discriminative foreground. The main reason is that comparing with stabilized body parts and corresponding descriptions, the changes of background are more drastic. If the filters are updated in the forward propagation, a lot of irrelevant noises would be mixed into denoising filters and it is harmful to the correlation learning. Instead, in row 4 of Table 3, when we utilize Back Propagation to learn the parameters in dictionary filters, the overall performance would also be degraded. This result shows that learning the consistency of body parts between different pedestrians can bring greater gains.

As shown in rows 5-6 of Table 3, we share the denoising filters and dictionary filters in both domains. The results show that the shared filters would influence the performance. It can also prove that the inter-modal variations are larger than intra-modal variations.

The row 7 of Table 3 shows the impact of avg-pooling on the performance. We observe our method is insensitive to different the strategy of pooling. In the row 8 of Table 3, we remove the mutual-exclusion-loss. We find that this loss is important for separating useful information.

### 4.5 Qualitative Results

To better explore the aligning processes learned by our method, we visualise sample results along with the response distribution from the "foreground filter". We take Fig. 4 Query (a) for example, the red regions in the image and text represent higher response scores for foreground filter, while the blue regions in the image and light color in the text indicate the lower scores. It can be observed that the background noises in the image (such as passers-by, tiles and plants) have lower response scores. On the other hand, our method is able to reserve gender, attributes ("long dark hair", "white tennis shoes"), and accessories ("pink

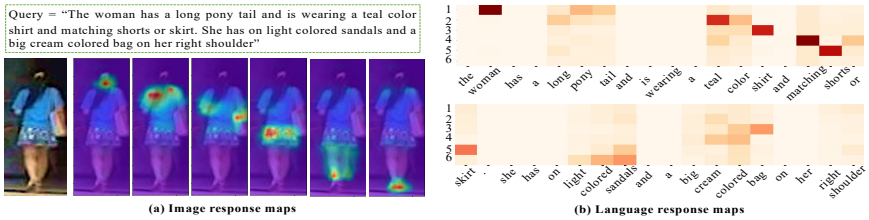


**Fig. 5.** Visualizations of response maps of image and text from dictionary filters.

bag", "black and grey backpack") from textual queries, which are key clues for subsequent local alignments of the model.

We further demonstrate response maps of our "dictionary filters" during the testing. Fig. 5 (a) shows the response maps of six different body parts which are correspond with "dictionary filters". Meanwhile, the response distribution of our language-based dictionary filters are shown in Fig. 5 (b). It can be observed that "woman" and "long pony tail" have relatively higher response scores in the first row of the query text. Accordingly, the crucial regions (*i.e.,* head) have higher response scores in the first response map of the image. The important words that each language-based dictionary filter pays attention to can be found in the visual dictionary feature maps. It suggests that our method can adaptively localize the differentiated regions or words without any additional information (such as semantic annotations or language processing toolkits).

## 5 Conclusion

In this paper, we propose a novel end-to-end Simple and Robust Correlation Filtering (SRCF) framework which can effectively extract the key clues and adaptively align the local features without any auxiliary tools. Meanwhile, we introduce two different types of filters to achieve the global-level and local-level alignments. They can help model refine pedestrian regions as well as meaningful words. The experimental results show that SRCF achieves better performance on both accuracy and robustness.

**Acknowledgement** This work was supported by National Key R&D Program of China (No.2020AAA0106900), the National Natural Science Foundation of China (No.62101451, No.U19B2037), Shaanxi Provincial Key R&D Program (No.2021KWZ-03), Natural Science Basic Research Program of Shaanxi (No.2021JCW-03), Open Projects Program of National Laboratory of Pattern Recognition (202100028), and Fundamental Research Funds for the Central Universities (D5000210733).

## References


1. Aggarwal, S., RADHAKRISHNAN, V.B., Chakraborty, A.: Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2617–2625 (2020)
2. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 2544–2550. IEEE (2010)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
4. Chen, Y., Zhang, G., Lu, Y., Wang, Z., Zheng, Y.: Tipcb: A simple but effective part-based convolutional baseline for text-based person search. Neurocomputing **494**, 171–181 (2022)
5. Ding, Z., Ding, C., Shao, Z.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)
6. Dong, Q., Gong, S., Zhu, X.: Person search by text attribute query as zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3652–3661 (2019)
7. Farooq, A., Awais, M., Kittler, J., Khalid, S.S.: Axm-net: Implicit cross-modal feature alignment for person re-identification (2022)
8. Gao, C., Cai, G., Jiang, X., Zheng, F., Zhang, J., Gong, Y., Peng, P., Guo, X., Sun, X.: Contextual non-local alignment over full-scale representation for text-based person search. arXiv preprint arXiv:2101.03036 (2021)
9. Han, X., He, S., Zhang, L., Xiang, T.: Text-based person search with limited data. arXiv preprint arXiv:2110.10807 (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: Neural networks for perception, pp. 65–93. Elsevier (1992)
12. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1373–1378 (2015)
13. Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J.: Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12976–12985 (2021)
14. Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T.: Pose-guided multi-granularity attention network for text-based person search. In: AAAI. vol. 34, pp. 11189–11196 (2020)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
16. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1890–1899 (2017)
17. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1970–1979 (2017)

18. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10880–10889 (2020)
19. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)
20. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
21. Naresh Boddeti, V., Kanade, T., Vijaya Kumar, B.V.K.: Correlation filters for object alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2291–2298 (2013)
22. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. TIP **29**, 5542–5556 (2020)
23. Niu, K., Huang, Y., Wang, L.: Textual dependency embedding for person search by language. In: ACM MM. pp. 4032–4040 (2020)
24. Oord, A.v.d., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. arXiv preprint arXiv:1711.00937 (2017)
25. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
26. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5814–5824 (2019)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
30. Wang, Z., Fang, Z., Wang, J., Yang, Y.: Vitaa: Visual-textual attributes alignment in person search by natural language. In: ECCV. pp. 402–420. Springer (2020)
31. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018)
32. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
33. Wu, Y., Yan, Z., Han, X.: Lapscore: Language-guided person search via color reasoning. In: ICCV. pp. 1624–1633 (2021)
34. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3166–3173 (2013)
35. Yang, Z., Chen, T., Wang, L., Luo, J.: Improving one-stage visual grounding by recursive sub-query construction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 387–404. Springer (2020)

36. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: ECCV. pp. 686–701 (2018)
37. Zheng, K., Liu, W., Liu, J., Zha, Z.J., Mei, T.: Hierarchical gumbel attention network for text-based person search. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3441–3449 (2020)
38. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16**(2), 1–23 (2020)
39. Zhu, A., Wang, Z., Li, Y.: Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 209–217 (2021)