

# Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search

Chenyang Gao<sup>1\*</sup>, Guanyu Cai<sup>2</sup>, Xinyang Jiang<sup>2†</sup>, Feng Zheng<sup>1</sup>, Jun Zhang<sup>2</sup>  
Yifei Gong<sup>2</sup>, Pai Peng<sup>2</sup>, Xiaowei Guo<sup>2</sup>, Xing Sun<sup>2</sup>

<sup>1</sup> Southern University of Science and Technology, <sup>2</sup> Tencent Youtu Lab

## Abstract

Text-based person search aims at retrieving target person in an image gallery using a descriptive sentence of that person. It is very challenging since modality gap makes effectively extracting discriminative features more difficult. Moreover, the inter-class variance of both pedestrian images and descriptions is small. Hence, comprehensive information is needed to align visual and textual clues across all scales. Most existing methods merely consider the local alignment between images and texts within a single scale (e.g. only global scale or only partial scale) or simply construct alignment at each scale separately. To address this problem, we propose a method that is able to **adaptively align image and textual features across all scales**, called NAFS (i.e. Non-local Alignment over Full-Scale representations). Firstly, a novel staircase network structure is proposed to extract full-scale image features with better locality. Secondly, a BERT with locality-constrained attention is proposed to obtain representations of descriptions at different scales. Then, instead of separately aligning features at each scale, a novel **contextual non-local attention mechanism** is applied to simultaneously discover latent alignments across all scales. The experimental results show that our method outperforms the state-of-the-art methods by 5.53% in terms of top-1 and 5.35% in terms of top-5 on text-based person search dataset. The code is available at <https://github.com/TencentYoutuResearch/PersonReID-NAFS>

## 1. Introduction

Text-based person search aims at retrieving target person in an image gallery using a descriptive sentence of that person. Compared to classical person re-identification (Re-id), it does not need an image of the target person as query

\*This work was done when Chenyang Gao was an intern at Tencent Youtu Lab and this work was supported by 2020 Tencent Rhino-Bird Elite Training Program

†Correspondance Author: xinyangj@zju.edu.cn

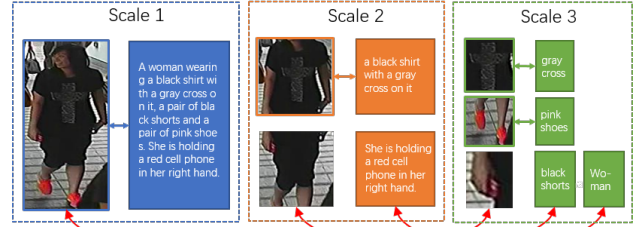


Figure 1. Illustration of image text alignment both within similar scales and across different scales.

which could be difficult to obtain. In addition, **text-based person search is more user-friendly since it can support open-form natural language queries**. Thus it has the potential to have much broader applications.

**Compared with general image text matching task where an image may contain several objects, text-based person search is a much more challenging task since the high-level semantics among different pedestrian images are very similar, causing small inter-class variance of both pedestrian images and textual descriptions.** Thus, in order to explore more distinctive and comprehensive information, text-based person search requires an algorithm to extract image and textual features from all scales. For example, both of the image and textual description in Figure 1 can be decomposed into representations at different scales. The sentence can be represented as short phrases, such as “black shorts” at scale 3, or longer sub-sentences at scale 2. Similarly, the image can also be partitioned into sub-regions with different sizes at scale 3 and scale 2. Since correct alignment between these image representations and textual representations are the basis of image text matching task, it is essential to represent the image and textual description at all scales. In this paper, we call it **full-scale representation**. However, the complex relevance at different scales makes it difficult to build a reasonable scheme of alignment. As shown in Figure 1, in most cases, the alignment occur at similar scales, such as the sub-sentence “a black shirt with a gray cross on it” and the image region in scale 2, and the

short phrase “gray cross” and the smaller image region at scale 3. But occasionally alignment could also occur across different scales. For instance, as shown with the red arrows in Figure 1, a single word “woman” in scale 3 aligns with the whole image in scale 1. These phenomena illustrate the importance to **jointly align image and description both within similar scales and across different scales**. Therefore, a reasonable text-based person search method generally contains two key components. One is to learn image and textual representations at all scales in a coarse-to-fine fashion, the other is to explore an appropriate alignment to automatically and adaptively match these representations of different scales.

Most of the existing works [25, 7, 19] are unable to fully satisfy the aforementioned two perspectives. On one hand, for multi-scale representations, most methods merely learn representations for images and textual descriptions at a certain scale. Several coarse-grained methods [13, 12, 2, 26, 30] focus on learning representations at the global scale, i.e. the whole image and sentence as shown in Figure 1 Scale 1. Fine-grained methods [25, 7, 19] model the images and textual descriptions at the finest scale, e.g. image regions and short phrases as shown in Figure 1 Scale 3. Although some fine-grained methods [7, 19] consider combining the finest scale with the global scale, they still **lack mid-scale information** causing some description segments (image regions) fail to correctly align with proper image regions (description segments).

On the other hand, for the cross-scale alignment, existing methods [26, 30, 25, 7, 19] try to employ pre-defined rules to align images and textual descriptions of different scales. Zhang et al. and Zheng et al. [26, 30] only consider the global matching of images and textual descriptions. Some other methods [25, 7] add alignments between short phrases and image regions as shown in Figure 1 Scale 3, but ignore alignments across different scales. Recently, Niu et al. [19] further add extra alignments between the whole image and short phrases, as well as small image stripes and the whole sentence. These methods show that utilizing multi-scale features can significantly improve performance. However, all of them pre-define several alignment rules among image representations and textual representations of different scales (e.g. global-global, local-local), and build alignment within these fixed scale pairs separately. Hence, it limits the alignment to a certain scope, causing the alignment between image representations and textual representations outside the scale pairs are completely ignored.

To address above problems, in this paper, we propose a novel text-based person search method that builds full-scale representations for both images and textual representations, and adaptively aligns them across all scales, called NAFS (Non-local Alignment over Full-Scale representations). First, we propose a staircase network with a novel

**stripe shuffling operation** that incorporates better locality to the learned full scale image features. Then a modified BERT language model by adding a **locality-constrained attention** is adopted to extract full-scale textual features. Next, instead of aligning features under several pre-defined scales (e.g., local-local, global-global), we develop a much more flexible alignment mechanism called **contextual non-local attention**, which is able to jointly take image representations and textual representations from all scales as input then adaptively build the alignment across all scales. Finally, a novel **re-ranking algorithm** based on the nearest visual neighbors is proposed to further improve the ranking quality.

The main contributions of this paper can be summarized as follows: (1) A novel staircase CNN network and a local constrained BERT model are specially developed to extract full-scale image and textual representations. (2) A contextual non-local attention mechanism is proposed to adaptively align the learned representations across all scales. (3) The proposed framework achieves state-of-the-art results on the challenging dataset CUHK-PEDES[13]. Extensive ablation studies clearly demonstrate the effectiveness of each component in our method.

## 2. Related Work

### 2.1. Person Re-identification (ReID).

Generally, most of the ReID methods [5, 18] use deep CNNs to extract a global discriminative representation for each person image, while some part-based models [22, 24, 28] try to exploit local information. For example, PCB [22] horizontally cuts the output feature map into six parts to learn six different local features. MGN [24] and Pyramid Network [28] propose a pyramid structured network to extract features in a coarse-to-fine manner. Moreover, some approaches propose to learn local features from local regions with semantic meanings like human part segmentation or pose [29, 17, 8, 21, 27]. However such methods highly rely on the accuracy of pose estimation and semantic parsing algorithms.

### 2.2. Text-Based Person Search.

Li et al. [13] first introduce the text-based person search task and propose a GNA-RNN model to learn an affinity score between the query description and the image in the gallery. Later, Li et al. [12] propose an identity-aware two-stage network to efficiently locate simple incorrect matchings and make the result insensitive to changes in sentence structure. In [2], a patch-wise word matching model is introduced to exploit the local matching information and obtain the proper affinity between the text and image. Zhang et al. [26] design cross-modal objective functions for learning discriminative image-text embeddings. Moreover, a new

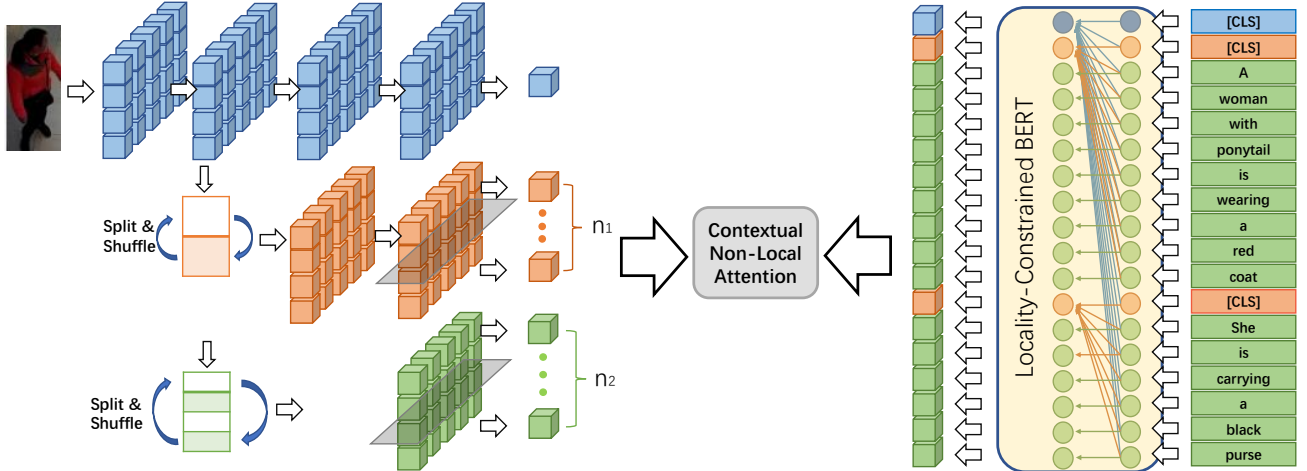


Figure 2. The overall framework consists of a stair-case network for visual representation extraction, a locality-constrained BERT for textual representation extraction and a contextual non-local attention module for joint alignment.

method CMAAM [1] treats the task as a multi-task training framework that significantly boosts the performance of global features by introducing extra attribute annotation and prediction. Recently, Niu et al. [19] propose to define three types of alignment, namely global-global, global-local and local-local, and learn separate alignment within these three scale pairs. In addition, Wang et al. [25] exploit an extra segmentation model to align person partial features and textual attribute features with a k-reciprocal sampling align loss. While, a pose-guided multi-granularity attention network is explored in [7], which contains a fine-grained alignment component and a coarse alignment component to exploit multi-granularity cross-modal relations.

### 2.3. Image Text Matching.

The goal of general image text matching is to learn a joint latent space where the embeddings of visual inputs and textual annotations can be compared directly. Besides global representations, some state-of-the-art methods including SCAN [10] and BFAN [16] also exploit alignment between images and textual fragments such as objects and words. Recently, methods like Unicoder and OSCAR [11, 14] propose to use BERT [3] or transformer [23] like network to model the text-image matching problem as a binary classification task, improving the retrieval performance at the cost of much longer inference time.

## 3. Our Method

In this section, we explain the proposed NAFS in detail. First, we introduce the procedures of extracting the visual and textual representations. Then we describe our contextual non-local attention mechanism. Finally, we introduce proposed re-ranking by visual neighbors to further improve the performance.

### 3.1. Extracting Visual Representation

**Staircase Backbone Structure.** Firstly, we elaborate on the implementation details of the proposed staircase network. As shown in Figure 2, it contains three branches, each of which is responsible for extracting visual features at different scales, from coarse to fine, namely the **global branch** (colored in blue), the **region branch** (colored in yellow) and the **patch branch** (colored in green). A general ResNet [4] network is used as the backbone. 1) The global branch is used to extract global and coarsest features. 2) The region branch extracts finer features from large sub-regions in the image. It takes the feature map at the second stage of global branch as its input, then fed into two consecutive res-blocks to extract features at region scales. The output feature map of the region branch is then horizontally partitioned into  $n_1$  stripes, each of which is further encoded as a local feature of a certain region. 3) The patch branch extracts the finest features from small patches in the image. It takes the feature map at the third stage of global branch as its input, which is then fed into one res-block to extract features at small patch scales. Then we horizontally partition the output feature map into  $n_2$  stripes to extract  $n_2$  feature vectors for local patches.

**Split and Shuffle Operation.** A challenge of stripe-based ReID models is that **due to the large perception field of CNN models, the stripe of feature maps in the deep layers may contain global information as well**. Thus, to guarantee a better locality for the multi-scale image features, we introduce a novel split&shuffle operation. It takes the intermediate feature map as input then equally partitions the feature map into several horizontal stripes denoted as a list  $F = \{f_1, f_2, \dots, f_n\}$ , where  $f_i$  is the  $i$ -th stripe starting from the top of the feature map. Then, such set of the partitioned stripes are randomly shuffled and re-concatenated

along the vertical axis to form a complete feature map as the output. Both feature maps at stage 2 and stage 3 will be first split and shuffled before feeding into the range and patch branches, respectively. By shuffling the partitioned stripes randomly, it enables to break the inter-relationship between consecutive stripes so that the model can focus on the information within each stripe. Since our contextual non-local attention does not rely on the order of feature map fragments, it is not necessary to re-organize the partitioned stripes to the original order.

The visual representation extraction module takes a pedestrian image as input, and then a list of image features of different scales can be obtained and notated as  $I = \{i_{p1}, i_{p2}, \dots, i_{pn}\}$  where  $i_{pi} \in \mathbb{R}^D$ .

### 3.2. Extracting Textual Representation

Given a textual description  $E$ , we add a **locality constraint** to BERT to extract different scale representations of  $E$ . In our method, a textual description will be represented in three scales, separately. 1) At the **sentence-level**, we add a special classification token ([CLS]) to the beginning of sentence  $E$ . The final hidden state corresponding to this token can be used as the sentence-level representation of the whole sentence  $E$  in a global view. 2) At the **middle-level**, we separate the sentence  $E$  by commas resulting in a list of shorter sub-sentences. For every sub-sentence in the list, the [CLS] token is also attached to the beginning of the sub-sentence, whose final hidden state is used as the representation of each sub-sentence as well. 3) At the finest **word-level**, the final hidden state of each word is directly used as the word-level representation.

For a common BERT-based model [3], the hidden variables of all tokens have the same global perception field. Each token can attend to any tokens in the entire input sentence. To provide locality to representations of sub-regions in a sentence (the [CLS] token for the sub-sentence), we propose a **locality-constrained attention** module to attend tokens within a certain range. Similar to the original BERT, given the query of a [CLS] token that corresponds to a sub-sentence, denoted as  $q_{CLS}$ , the locality-constrained attention is computed as follows:

$$\text{Attention}(q_{CLS}) = \sum_i \frac{e^{q_{CLS} k_i}}{\sum_i e^{q_{CLS} k_i} \mathbf{1}(i \in U)} v_i \mathbf{1}(i \in U), \quad (1)$$

where  $k_i$  and  $v_i$  denote keys and values corresponding to all tokens in a sentence, respectively.  $U$  is the set of tokens within the range of this sub-sentence, and  $\mathbf{1}(\cdot)$  is an indication function that returns 1 when  $i$ -th token is in  $U$ .

The textual representation extraction module takes a pedestrian description as input, and then a list of textual embeddings of different scales can be obtained and denoted as  $T = \{t_{p1}, t_{p2}, \dots, t_{pn}\}$  where  $t_{pi} \in \mathbb{R}^D$ .

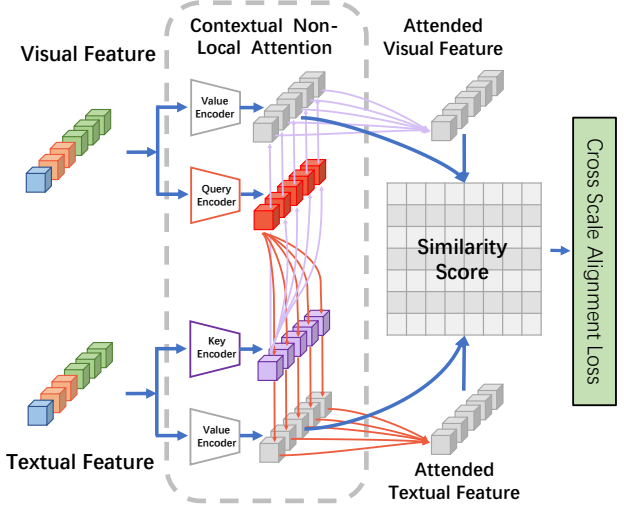


Figure 3. The illustration of proposed contextual non-local attention module.

### 3.3. Contextual Non-Local Attention Mechanism

As shown in Figure 3, the contextual non-local attention expects two inputs: a set of **visual features**  $I = \{i_{p1}, i_{p2}, \dots, i_{pn}\}$  and a set of **textual features**  $T = \{t_{p1}, t_{p2}, \dots, t_{pn}\}$ . The output of the attention module is a **similarity score** that measures the relevance of a image-text pair. In a nutshell, the contextual non-local attention enables cross-modal features to align with each other in a coarse-to-fine fashion according to their semantics, instead of merely using pre-defined and fixed rules (e.g., local-local, global-global).

Inspired by the spirit of self attention [23], we can explain our proposed attention mechanism as mapping a query and a set of key-value pairs to an output. For visual features, two learned linear projections are used to map  $I$  to visual queries  $I_Q = \{I_{q1}, I_{q2}, \dots, I_{qm}\}$  and visual values  $I_V = \{i_{v1}, i_{v2}, \dots, i_{vm}\}$ . Similarly, two linear projections are explored to map  $T$  to textual keys  $T_K = \{t_{k1}, t_{k2}, \dots, t_{kn}\}$  and textual values  $T_V = \{t_{v1}, t_{v2}, \dots, t_{vn}\}$ . Based on  $I_Q$ ,  $I_V$ ,  $T_K$  and  $T_V$ , we introduce our proposed attention mechanism in both Image-Text and Text-Image ways.

**Image-Text Contextual Non-Local Attention.** The proposed image-Text attention module includes two stages. First, each visual query attend to textual keys to get a corresponding attended textual value. Then, considering all visual values and their attended textual values, similarity between a image-text pair can be determined. In detail, to obtain attended textual values, we first compute the cosine similarity matrix of  $I_Q$  and  $T_K$  to obtain the weights on  $T_V$



as follows:

$$s_{a,b} = \left[ \frac{i_{qa}^T t_{kb}}{\|i_{qa}\| \|t_{kb}\|} \right]_+, \quad a \in m, b \in n, [x]_+ = \max(x, 0) \quad (2)$$

where  $s_{a,b}$  denotes the similarity between the  $a$ -th visual query and  $b$ -th textual key. Further, we normalize it as  $\hat{s}_{a,b} = \frac{s_{a,b}}{\sum_{a=1}^m s_{a,b}}$ . Moreover, to filter out irrelevant textual values, a **focal attention trick** which is similar to [15] is used, where  $\tilde{s}_{a,b} = [\sum_{c=1}^n \hat{s}_{a,b} - \hat{s}_{a,c}]_+ \hat{s}_{a,b}$ . Then, we compute the weighted textual values as:

$$r_{va} = \sum_{b=1}^n \alpha_{a,b} t_{vb}, \quad \alpha_{a,b} = \frac{\exp(\lambda_1 \tilde{s}_{a,b})}{\sum_{b=1}^n \exp(\lambda_1 \tilde{s}_{a,b})} \quad (3)$$

where  $\lambda_1$  is the inverse temperature of the softmax function.

In the second stage, we define the relevance between  $a$ -th visual value and its corresponding textual context using the cosine similarity between  $i_{va}$  and  $r_{va}$ :

$$R(i_{va}, r_{va}) = \frac{i_{va}^T r_{va}}{\|i_{va}\| \|r_{va}\|} \quad (4)$$

By averaging all  $R(i_{va}, r_{va})$ , we obtain the similarity of a image-text pair as

$$S(I, T) = \frac{\sum_{a=1}^m R(i_{va}, r_{va})}{m} \quad (5)$$

As illustrated in our proposed attention mechanism, each visual feature pays more attention to relevant textual features. The relevant textual features may come from a word, a short phrase or a whole sentence, merely depending on whether the visual and textual features share similar semantics. While, instead, previous methods [25, 7] tend to build the correspondence in a fixed way. We relax such constraints by enabling a semantic-based attention mechanism to build correspondence across different scales, which helps us align a image-text pair more adaptively and correctly.

**Text-Image Contextual Non-Local Attention.** Similar to the Image-Text contextual non-local attention, we regard textual keys as queries and visual queries as keys, respectively, and attend textual keys with respect to visual queries. Then, with textual values and attended visual values, we compute the similarity between a image-text pair. Specifically, the weight of  $b$ -th visual value with respect to  $a$ -th textual value is defined as  $s'_{a,b} = \left[ \frac{t_{ka}^T i_{qb}}{\|t_{ka}\| \|i_{qb}\|} \right]_+$ ,  $a \in n, b \in m$ . The normalized and focal attended weight is defined as  $\tilde{s}'_{a,b} = [\sum_{c=1}^m \hat{s}'_{a,b} - \hat{s}'_{a,c}]_+ \hat{s}'_{a,b}$ , where  $\hat{s}'_{a,b} = \frac{s'_{a,b}}{\sum_{a=1}^n s'_{a,b}}$ .

Then, we define the weighted visual value as  $r'_{va} = \sum_{b=1}^m \alpha'_{a,b} i_{vb}$ , where  $\alpha'_{a,b} = \frac{\exp(\lambda_2 \tilde{s}'_{a,b})}{\sum_{b=1}^m \exp(\lambda_2 \tilde{s}'_{a,b})}$ . Using the weighted visual value  $r'_{va}$  and textual value, we compute

the similarity of them as  $R(t_{va}, r'_{va}) = \frac{t_{va}^T r'_{va}}{\|t_{va}\| \|r'_{va}\|}$ . The final similarity of a image-text pair is obtained by averaging operation  $S'(T, I) = \frac{\sum_{a=1}^n R(t_{va}, r'_{va})}{n}$ .

**Alignment Objective.** We introduce an objective function named **Cross-Scale Alignment Loss** (CSAL) to optimize the proposed algorithm. Given a mini-batch of images  $\{I_i\}_{i=1}^B$ , captions  $\{T_j\}_{j=1}^B$  and all image-text pairs  $\{(I_i, T_j), y_{i,j}\}_{i=1, j=1}^{B \times B}$  where  $y_{i,j} = 1$  if  $(I_i, T_j)$  is a matched pair otherwise 0, we define the image-to-text similarity of  $(I_i, T_j)$  as  $S(I_i, T_j)$  and text-to-image similarity as  $S'(T_j, I_i)$ . To maximize similarities between the matched pairs and restrain correspondences of unmatched pairs, we define CSAL as:

$$\mathcal{L}_{CSAL} = \mathcal{L}_i + \mathcal{L}_t, \text{ where} \quad (6)$$

$$\mathcal{L}_i = \frac{1}{B} \sum_{i=1}^B \sum_{j=1}^B S(I_i, T_j) \log \frac{S(I_i, T_j)}{q_{i,j} + \epsilon}, \quad q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^B y_{i,k}}$$

$$\mathcal{L}_t = \frac{1}{B} \sum_{j=1}^B \sum_{i=1}^B S'(T_j, I_i) \log \frac{S'(T_j, I_i)}{q_{i,j} + \epsilon}, \quad q_{i,j} = \frac{y_{k,j}}{\sum_{k=1}^B y_{k,j}}$$

where  $\epsilon$  denotes a small number to avoid numerical problems.

Considering that the backbone is essential to features from multiple scales, we use Cross-Modal Projection Matching (CMPM)  $\mathcal{L}_{CMPM}$  and Cross-Modal Projection Classification  $\mathcal{L}_{CMPC}$  (CMPC) proposed by Zhang et al. [26] to stabilize the training procedure by adding CMPM and CMPC loss on features extracted from the global branch. Thus, the final objective function is:

$$\mathcal{L} = \lambda_2 \mathcal{L}_{CMPM} + \lambda_3 \mathcal{L}_{CMPC} + \lambda_4 \mathcal{L}_{CSAL} \quad (7)$$

### 3.4. Re-Ranking by Visual Neighbors

We propose a multi-modal re-ranking algorithm to further improve the performance by **comparing the visual neighbors of the query to the gallery** (RVN). Given a textual query  $T$ , the initial ranking list is obtained by sorting the images based on their similarities to the query obtained by Eq.5. Then, for each image  $I$  in the initial list, we obtain its  $l$ -nearest neighboring images based on the similarity of their visual representations, denoted as  $N_{i2i}(I, l)$ . Similarly, the nearest neighbors of the textual query can be obtained based on the similarity between its textual representations and the visual representation of images, denoted as  $N_{t2i}(T, l)$ . Here, to accelerate the computation, only the global feature is used for finding nearest neighbors. Then, we re-calculate the pair-wise similarity between the textual query and each image in the gallery by **comparing the  $l$ -nearest neighbors with Jaccard Distance**:

$$D_J(I, T) = 1 - \frac{N_{i2i}(I, l) \cap N_{t2i}(T, l)}{N_{i2i}(I, l) \cup N_{t2i}(T, l)} \quad (8)$$

Finally, the gallery is re-sorted based on **the averaged scores of the original similarity and the Jaccard Distance**.

## 4. Experiments

In this section, we evaluate our proposed NAFS by comparing the person search performance with state-of-the-art methods. Furthermore, we conduct ablation studies to demonstrate the effectiveness of each component. Finally, the attentions between images and textual descriptions are visualized to demonstrate NAFS’s ability to discover joint alignment across multiple scales.

### 4.1. Experimental Setup

**Dataset and Evaluation Protocol.** We evaluate our proposed model on the CUHK-PEDES dataset, which is currently the only benchmark for text-based person search. It contains 40206 images from 13003 unique person IDs in total. The training set has 34054 images, 11003 person IDs and 68126 textual descriptions. The validation set has 3078 images, 1000 person IDs and 6158 textual descriptions. The test set has 3074 images, 1000 person IDs and 6156 textual descriptions. On average, each image contains 2 different textual descriptions and each textual description is generally longer than 23 words. The vocabulary of the dataset contains 9408 different words. Following the standard evaluation setting, the performance is measured by top-k accuracy ( $K = 1, 5, 10$ ). Specially, given a person description, if top-k images contain any person corresponding to the given description, the search is successful. Top-k accuracy is the percentage of successful searches among all searches.

**Implementation Details.** For the visual representation extraction module, we use ResNet-50 as our backbone for fair comparisons with previous methods. The region branch splits the feature map into two stripes equally and the patch branch splits the feature map into three stripes equally. The number of output strides of the convolution layer at the last stage of the backbone is set to 1. The dimension  $D$  of the image features at different scales is 768. We use horizontally flipping (50% probability) as data augmenting. All images are normalized and resized to  $384 \times 128$  before sending into the network. For the textual representation extraction module, we use BERT-Base-Uncased model as our backbone. The dimension  $D$  of different scale textual features is set to 768 as well.

We initialize ResNet-50 with the weights pre-trained on the ImageNet classification task. And we initialize the weights of BERT-Base-Uncased model with the weights pre-trained using a combination of masked language modeling objective and next sentence prediction on a large corpus including the Toronto Book Corpus and Wikipedia. The model is optimized with the Adam [9] optimizer and the importance hyperparameters of each loss function  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  in 7 are 1, 1 and 0.1 respectively. The learning rate

Table 1. Comparison with state-of-the-art methods. Top-1, top-5 and top-10 accuracies (%) are reported. The best performance is bold. In the second column, “g” stands for global scale, “g+l” stands for global scale and local scale, “m” stands for full-scale representations from coarse to fine. “RVN” stands for our proposed re-ranking by visual neighbors.

Method	Scale	Top1	Top5	Top10
GNA-RNN [13]	g	19.05	-	53.64
CMCE [12]	g	25.94	-	60.48
PWM+ATH [2]	g	27.14	49.45	61.02
Dual Path [30]	g	44.40	66.26	75.07
CMPM+CMPC [26]	g	49.37	-	79.27
MIA [19]	g+l	53.10	75.00	82.90
PMA [7]	g+l	53.81	73.54	81.23
ViTAA [25]	g+l	55.97	75.84	83.52
NAFS(ours)	m	59.94	79.86	86.70
NAFS with RVN (ours)	m	<b>61.50</b>	<b>81.19</b>	<b>87.51</b>

Table 2. Comparison with the methods using feature representations at different scales. Top-1, top-5 and top-10 accuracies (%) are reported.

Feature	Top1	Top5	Top10
Global	55.47	77.29	84.36
Local+Global	56.90	77.92	84.81
Full Scale	59.94	79.86	86.70

for the visual and textual feature extraction branch is set to 0.00011 and for the rest of the network layers is set to 0.0011. The batch size is 64.

### 4.2. Comparison with State-of-The-Art Methods

Table 1 demonstrates our results compared with state-of-the-art methods on CUHK-PEDES. GNA-RNN, CMCE and PWM+ATH use VGG-16[20] as visual representation extraction backbone. While Dual Path, CMPM+CMPC, MIA, PMA, ViTAA and our NAFS use ResNet-50 as visual representation extraction backbone. Overall, our NAFS achieves the highest performance both with and without RVN. In Table 1, it can be observed that methods utilizing global and local information achieve better performance than those merely use global information. This verifies the effectiveness of adapting representations at finer scale. Compared with ViTAA, which is the state-of-the-art method using both global and local features, NAFS gains 5.53%, 5.35% and 3.99% performance improvement in terms of top1, top5 and top10 metrics respectively. This clearly verifies the effectiveness of introducing full-scale representation and contextual non-local attention mechanism.

### 4.3. Extensive Ablation Studies

**Full-Scale Representation and Joint Alignment.** We conduct ablation studies to compare the performance of using image and textual representations at different scales. The results of the following three methods are shown in Ta-

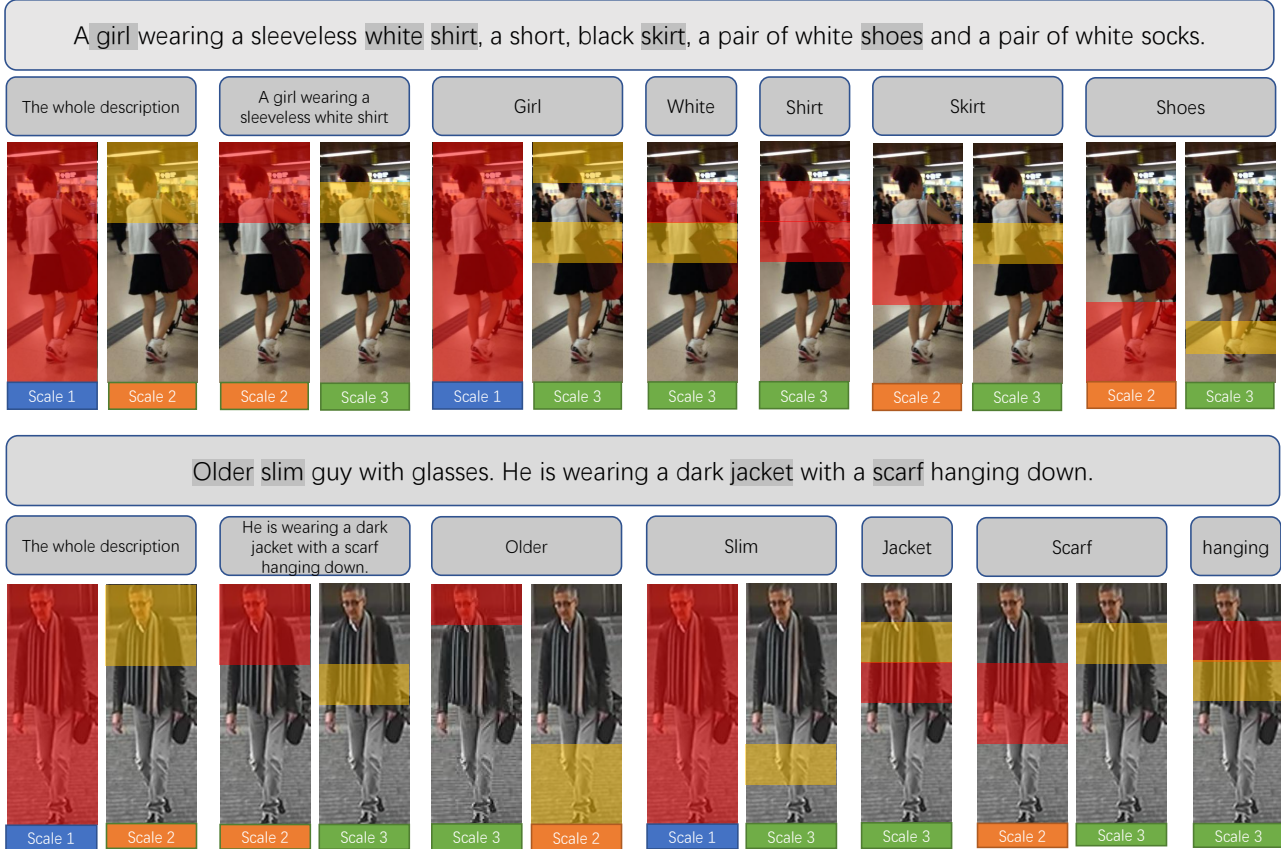


Figure 4. Visualization of the joint alignment between words and image regions across different scales. The image regions highlighted in red and yellow color are the ones have the highest and the second highest attention weights to the corresponding words.

Table 3. Performance Comparison of separate alignment and joint alignment. Top-1, top-5 and top-10 accuracies (%) are reported.

Method	Top1	Top5	Top10
Separate Alignment	57.98	78.22	85.43
Joint Alignment	59.94	79.86	86.70

ble 2.

- **Global Features.** This method only extracts global features for the entire images and descriptions. Since contextual non-local attention is not suitable for methods with only global representations, only CMPM and CMPC loss is applied to train the model.
- **Local + Global Features.** Besides the global features, this method further adds the image and text representations at the finest scale (scale 3). Apart from missing the mid-scale representations, the other components are exactly the same as our proposed method. We select this method to verify the effectiveness of adding finest image and textual representations to text-based person search tasks.
- **Full Scale Features.** This is the full implementation

of our proposed NAFS, with images and textual representations at three different scales, from coarse to fine. This comparison is to verify the effectiveness of adding representations of mid-level scale (scale 2).

Table 2 shows the performance of using representations under different scales. It is observed that the Top1 performance increases from 55.47 to 56.90 after adding local information. After adding mid-scale information, the top1 performance increases from 56.90 to 59.94. This implies different scale information is beneficial to the alignment procedure.

In order to verify the effectiveness of introducing joint alignment to the representations across different scales, we compare our joint alignment with methods using pre-define alignment. As shown in Table 3, separate alignment refers to separately aligning image and text within three scale pairs: the whole image to the whole sentence, large image regions (n1-stripe partition) to sub-sentences and small image regions (n2-stripe partition) to words. From table 3, we observe that our proposed joint alignment outperforms the separate alignment. This verifies that jointly aligning images and textual descriptions across different scales effec-

Table 4. Performance comparison of different components in our methods. Top-1, top-5 and top-10 accuracies (%) are reported.

BERT	Staircase Network	Contextual Non-local	Split&Shuffle	RVN	Top1	Top5	Top10
×	×	×	×	×	54.76	77.10	84.86
✓	×	×	×	×	55.47	77.29	84.36
✓	✓	×	×	×	57.59	78.22	85.71
✓	✓	✓	×	×	59.63	79.53	86.42
✓	✓	✓	✓	×	59.94	79.86	86.70
✓	✓	✓	✓	✓	61.50	81.19	87.51

tively boosts the performance of text-based person search.

**Model Components.** We divide our proposed methods into 5 different components and observe the performance improvement by adding each component, as shown in Table 4:

- **Baseline.** The first row of Table 4 is a baseline model without any NAFS components. A standard bi-LSTM[6] and ResNet-50 is used for feature extraction. CPM and CMPC loss is used for training.
- **BERT.** We replace the bi-LSTM in the baseline with our proposed locality-constrained BERT (denoted as ‘BERT’ in Table 4). The features from different scales are concatenated together to obtain one unified feature representation for image text matching. The locality-constrained BERT gains 0.71 performance improvement in terms of top1 accuracy.
- **Staircase Network.** The normal ResNet-50 backbone is replaced with the proposed staircase backbone structure that extracts representations at multiple scales. The features from different scales are concatenated together to obtain one unified feature vector for image text matching. The staircase network brings 2.12 performance improvement in terms of top1 accuracy.
- **Contextual Non-Local.** Instead of concatenating the multi-scale features, the joint alignment by the contextual non-local attention mechanism is applied, which gives 2.04 performance improvement in terms of top1 accuracy.
- **Split&shuffle.** The split and shuffle operation is added to the staircase backbone structure.
- **RVN.** Our proposed re-ranking method by visual neighbors is applied after the initial ranking, improving the top1 accuracy by 1.56.

#### 4.4. Visualization Analysis

To demonstrate NAFS’s ability to discover joint alignment across different scales, we visualize the alignment results between textual descriptions and image regions at different scales, which is shown in Figure 4. For better visualization of proposed contextual non-Local attention mechanism, we horizontally partition the output feature map into three stripes in region branch and six stripes in patch branch

respectively. The image regions highlighted with red and yellow colors have the highest and the second highest attention weights to the corresponding textual descriptions. In the case of two sub-regions with similar attention weights, both of them will be highlighted.

From Figure 4, we observe that NAFS is able to align textual descriptions with image regions across different scales, from coarse to fine. As shown in the top half of Figure 4, the whole description aligns with the whole image and the sub-sentence “A girl wearing a sleeveless white shirt” aligns with the image part at scale 2. Word “girl” aligns with the whole image at scale 1, because a person’s gender is determined by the clues from the entire image. Words “white” and “shirt” align with the image regions at scale 3 because a small part of the images contains the white shirt. Words like “skirt” and “shoes” align with both scale 3 and scale 2 image regions because the object skirt and shoes exists in both small and mid-level image regions. Similarly, in the bottom half of Figure 4, the whole description aligns with the whole image and the sub-sentence “He is wearing a dark jacket with a scarf hanging down” aligns with the image part at scale 2. Word “older” matches the top image stripe at scale 3, because we can tell this man’s age by his face, while the word “slim” matches the whole image because we need to example the full body of the person to tell if he is slim. The visualization results verify the effectiveness of proposed joint alignment and the necessity of full-scale representations.

## 5. Conclusion

We propose a novel text-based person search method that conducts joint alignment over full-scale representations, called NAFS. A novel staircase CNN network and a locality-constrained BERT model are proposed to extract multi-scale image and textual representations. A contextual non-local attention mechanism adaptively aligns the learned representations across different scales. Extensive ablation studies on the CUHK-PEDES dataset demonstrate that our approach outperforms state-of-the-art methods by a large margin.



## References

- [1] Surbhi Aggarwal, Venkatesh Babu RADHAKRISHNAN, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2617–2625, 2020. 3
- [2] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1879–1887. IEEE, 2018. 2, 6
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 4
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [5] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2
- [6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 8
- [7] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. In *AAAI*, pages 11189–11196, 2020. 2, 3, 5, 6
- [8] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018. 2
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [10] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 3
- [11] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 3
- [12] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1890–1899, 2017. 2, 6
- [13] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, 2017. 2, 6
- [14] Xiujuan Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020. 3
- [15] C. Liu, Zhendong Mao, Anan Liu, Tianzhu Zhang, Bo Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 5
- [16] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11, 2019. 3
- [17] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 2
- [18] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2
- [19] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020. 2, 3, 6
- [20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [21] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3960–3969, 2017. 2
- [22] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval

- with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 2
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4
- [24] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 2
- [25] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. *arXiv preprint arXiv:2005.07327*, 2020. 2, 3, 5, 6
- [26] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701, 2018. 2, 5, 6
- [27] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017. 2
- [28] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019. 2
- [29] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 2019. 2
- [30] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020. 2, 6