DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval

Aichun Zhu*
Nanjing Tech University
Nanjing, China
aichun.zhu@njtech.edu.cn

Zijie Wang
Nanjing Tech University
Nanjing, China
zijiewang9928@gmail.com

Yifeng Li
Nanjing Tech University
Nanjing, China
lyffz4637@163.com

Xili Wan
Nanjing Tech University
Nanjing, China
xiliwan@njtech.edu.cn

Jing Jin
Nanjing Tech University
Nanjing, China
janeking1015@163.com

Tian Wang
Beihang University
Beijing, China
wangtian@buaa.edu.cn

Fangqiang Hu
Nanjing Tech University
Nanjing, China
hufq@njtech.edu.cn

Gang Hua
China University of Mining and Technology
XuZhou, China
ghua@cumt.edu.cn

ABSTRACT

Many previous methods on text-based person retrieval tasks are devoted to learning a latent common space mapping, with the purpose of extracting modality-invariant features from both visual and textual modality. Nevertheless, due to the complexity of high-dimensional data, the unconstrained mapping paradigms are not able to properly catch discriminative clues about the corresponding person while drop the misaligned information. Intuitively, the information contained in visual data can be divided into person information (PI) and surroundings information (SI), which are mutually exclusive from each other. To this end, we propose a novel Deep Surroundings-person Separation Learning (DSSL) model in this paper to effectively extract and match person information, and hence achieve a superior retrieval accuracy. A surroundings-person separation and fusion mechanism plays the key role to realize an accurate and effective surroundings-person separation under a mutually exclusion constraint. In order to adequately utilize multi-modal and multi-granular information for a higher retrieval accuracy, five diverse alignment paradigms are adopted. Extensive experiments are carried out to evaluate the proposed DSSL on CUHK-PEDES, which is currently the only accessible dataset for text-base person retrieval task. DSSL achieves the state-of-the-art performance on CUHK-PEDES. To properly evaluate our proposed DSSL in the real scenarios, a Real Scenarios Text-based Person Reidentification (RSTPReid) dataset is constructed to benefit future research on text-based person retrieval, which will be publicly available.

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475369

CCS CONCEPTS

• **Information systems → Image search**; • **Computing methodologies → Object identification**.

KEYWORDS

person retrieval, text-based person re-identification, cross-modal retrieval, surroundings-person separation

ACM Reference Format:
Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475369

1 INTRODUCTION

Person retrieval is a basic task in the field of video surveillance, which aims to identify the corresponding pedestrian in a large-scale person image database with a given query. Current researches of person retrieval chiefly focus on image-based person retrieval [7, 27, 29] (aka. person re-identification), which may sometimes suffer from lacking query images of the target pedestrian in practical application. Considering that in most of the real-world scenes, textual description queries are much more accessible, text-based person retrieval [8, 11, 12, 14, 17, 18, 25] has drawn remarkable attention for its effectiveness and applicability.

As text-based person retrieval involves processing multi-modal data, it can be deemed as a specific subtask of cross-modal retrieval [9, 10, 15, 16, 21, 28]. Nevertheless, instead of containing various categories of objects in an image, each image cared by text-based person retrieval contains just one certain pedestrian. The textual description queries, meanwhile, offer much more details about the corresponding person rather than roughly mention the objects in an image. Owing to the particularity of text-based person retrieval, many previous methods proposed on general cross-modal retrieval benchmarks (e.g. Flickr30K [19] and MSCOCO [13]) generalize on

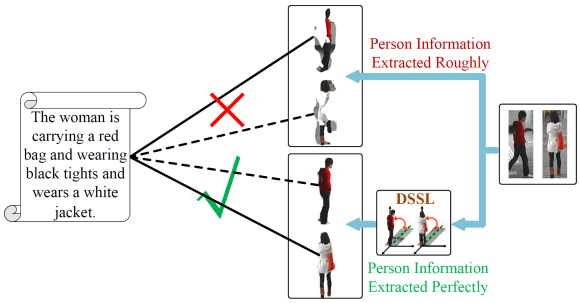

Figure 1: Due to the complexity of high-dimensional data, roughly extracting the person information without proper constraints may miss key clues while fail to drop redundant inferences, which further leads to a mismatched case. Within our proposed Deep Surroundings-person Separation Learning (DSSL) model, the person information is properly separated with the surroundings information, which hence gives a superior retrieval performance.

it poorly. In addition, CUHK-PEDES [12] is currently the only accessible dataset for text-base person retrieval. It is large in scale and contains images collected from various re-identification datasets under different scenes, view points and camera specifications. Nevertheless, images of each specific person are mostly caught by a same camera under similar conditions of time and space, which is not consistent with the real application scenarios. Therefore, we construct a Real Scenarios Text-based Person Reidentification (RSTPReid) dataset based on MSMT17 [26] to further train and evaluate the performance of our work, which also benefit future research. For each person, RSTPReid pools 5 images caught by 15 different cameras with complex both indoor and outdoor scene transformations and backgrounds in various periods of time, which makes RSTPReid much more challenging and more adaptable to real scenarios. Extensive experiments on RSTPReid and CUHK-PEDES can better validate the promising accuracy and efficiency of our work.

The major challenge of text-based person retrieval is to effectively extract and match features from both raw images and textual descriptions. Many previous methods [8, 14, 17, 18, 25] are devoted to learning a latent common space mapping, with the purpose of extracting modality-invariant features from both the visual and textual modalities. These proposed approaches are mainly based on the assumption that through a latent common space mapping, the intersection of information carried by the two modalities, namely, information of the targeting person can be retained into extracted modality-invariant common features. Nevertheless, due to the complexity of high-dimensional data, the unconstrained mapping paradigms are not able to properly catch discriminative clues about the corresponding person while drop the misaligned information (shown in Fig. 1).

Intuitively, information contained in visual data can be divided into person information (PI) and surroundings information (SI), which are **mutually exclusive** from each other. Meanwhile, the given textual description queries commonly describe the gender, appearance, clothing, carry-on items, possible movement, etc. of a certain pedestrian. In most of the real scenarios, the describer who offers a query nearly knows nothing about what kind of surroundings the target person is exactly in when captured by surveillance cameras, where the light conditions, viewpoints, etc. can be varied. Therefore, the given textual description basically contains only person information and there is no surroundings information included. On account of the structure of natural language sentences, noise signals (NS) like semantically irrelevant words and incorrect grammar are also inevitably included. Based on the above discussion, an efficient algorithm to accurately separate person and surroundings information in visual data and properly denoise features extracted from textual data is essential to enhance the retrieval performance.

To this end, we propose a novel Deep Surroundings-person Separation Learning (DSSL) model in this paper to effectively extract and match person information, and hence achieve a superior retrieval accuracy. DSSL takes raw images and textual descriptions as input and first extracts global and fine-grained local information from both modalities. As shown in Fig. 2, DSSL aims to properly separate surroundings and person information. To achieve this goal, a novel Surroundings-Person Separation Module (SPSM) is proposed to split the visual information as person and surroundings features (denoted as V_P and V_S) in a mutually exclusive manner. Then we adopt a Signal Denoising Module (SDM) to denoise and refine the extracted person feature (denoted as T_P) from the textual modality. As discussed above, ideally the person features V_P and T_P are purely about the target person without modality-specific interference. Hence the alignment between them (*AlignI*) can be regarded as matching the pedestrian cut out of the gallery image with the pedestrian in mind of the describer. In addition, through a proposed Surroundings-Person Fusion Module (SPFM), T_P is fused with V_S and reconstructed into the visual modality as V_R. Then an alignment between V_R and the visual feature before partitioned by SPSM (*AlignII*) is conducted, which can be viewed as placing the described person into the same surroundings as the gallery person and then matching it with the complete gallery image including the surroundings in the visual modality space. Besides, a Person Describing Module (PDM) is employed to reconstruct V_P into textual modality as T_R, which is then aligned with the non-refined textual feature (*AlignIII*). This proposed alignment can be regarded as describing the person in the gallery image with a text and then matching the text with the given query sentence in the textual modality space. Due to the mutually exclusion constraint in SPSM, V_P and V_S are orthogonal to each other, so the visual information is distributed between them without overlap. As shown in Fig. 2, during the training process, *AlignI* and *AlignIII* will form a constraint which forces V_P to contain more complete information about the person. Based on the mutually exclusion precondition, person information in V_S will accordingly be taken away into V_P. Meanwhile, *AlignII* is conducted by putting the described person into the surroundings of the gallery person, which requires the surroundings information to be properly included in V_S while peeled off in V_P. As a result, these three alignments work in complementary to guide the correct information exchange between V_P and V_S under a mutually exclusion constraint, and finally lead to an accurate and effective surroundings-person separation. To adequately exploit fine-grained clues, a cross-modal attention (CA) mechanism [18, 25] is utilized to further align a local feature

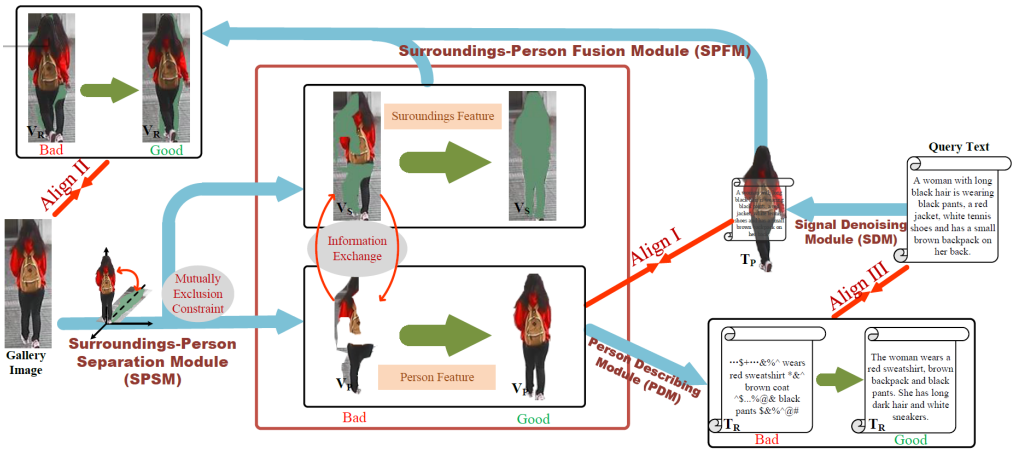

Figure 2: Illustration of the complementary relationship among *AlignI*, *AlignII* and *AlignIII* when training DSSL. During the training process, *AlignI* and *AlignIII* will form a constraint which forces more complete information about the person to be contained in the person feature V_P. Based on the mutually exclusion precondition, person information in the surroundings feature V_S will accordingly be taken away into V_P. Meanwhile, *AlignII* is conducted by putting the described person into the surroundings of the gallery person, which requires the surroundings information to be properly included in V_S while peeled off in V_P. As a result, these three alignments work in complementary to guide the correct information exchange under a mutually exclusion constraint, which finally leads to the change of extracted information from bad to good and an accurate and effective surroundings-person separation.

matrix extracted from one modality with the global feature in the other (*AlignIV* and *AlignV* shown in Fig. 3).

Our contributions can be summarized as five folds: (1) A novel Deep Surroundings-person Separation Learning (DSSL) model is proposed to properly extract and match person information. A proposed surroundings-person separation and fusion mechanism plays the key role to realize an accurate and effective surroundings-person separation under a mutually exclusion constraint. (2) Five diverse alignment paradigms are adopted to adequately utilize multi-modal and multi-granular information and hence improve the retrieval accuracy. (3) A Signal Denoising Module (SDM) is employed to denoise and refine the extracted person feature from the textual modality. (4) Extensive experiments are carried out to evaluate the proposed DSSL on CUHK-PEDES [12]. DSSL outperforms previous methods and achieves the state-of-the-art performance on CUHK-PEDES. (5) A Real Scenarios Text-based Person Reidentification (RSTPReid) dataset is constructed to benefit future research on text-based person retrieval, which will be publicly available.

2 RELATED WORKS

2.1 Person Re-identification

Person re-identification has drawn increasing attention in both academical and industrial fields, and deep learning methods generally plays a major role in current state-of-the-art works. Yi et al. [29] firstly proposed deep learning methods to match people with the same identification. Hou et al. [7] proposed an Interaction-and-Aggregation (IA) Block, which consists of Spatial Interaction-and-Aggregation (SIA) and Channel Interaction-and-Aggregation (CIA) Modules to strengthen the representation capability of the deep neural network. Xia et al. [27] proposed the Second-order Non-local Attention (SONA) Module to learn local/non-local information in a more end-to-end way.

2.2 Text-based Person Retrieval

Text-based person retrieval aims to search for the corresponding pedestrian image according to a given text query. This task is first put forward by Li et al. [12] and they take an LSTM to handle the input image and text. An efficient patch-word matching model [3] is proposed to capture the local similarity between image and text. Jing et al. [8] utilize pose information as soft attention to localize the discriminative regions. Niu et al. [18] propose a Multi-granularity Image-text Alignments (MIA) model exploit the combination of multiple granularities. Nikolaos et al. [17] propose a Text-Image Modality Adversarial Matching approach (TIMAM) to learn modality-invariant feature representation by means of adversarial and cross-modal matching objectives. Besides that, in order to better extract word embeddings, they employ the pre-trained publicly-available language model BERT. An IMG-Net model is proposed by Wang et al. [25] to incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model for extracting the multi-granular semantic information. Liu et al. [14] generate fine-grained structured representations from images and texts of pedestrians with an A-GANet model to exploit semantic scene graphs. A new approach CMAAM is introduced by Aggarwal et al. [1] which learns an attribute-driven space along with a class-information driven space by introducing extra attribute annotation and prediction. Zheng et al. [30] propose a Gumbel attention module to alleviate the matching redundancy problem and a hierarchical adaptive matching model is employed to learn subtle feature representations from three different granularities. Recently, the NAFS proposed by Gao et al. [5] is designed to extract full-scale

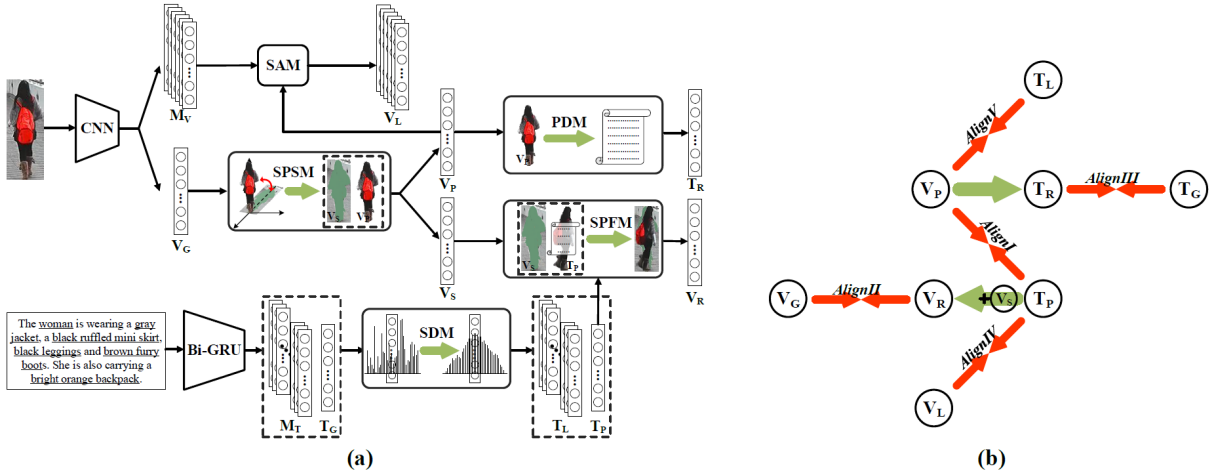

Figure 3: The overall framework of the proposed Deep Surroundings-person Separation Learning (DSSL) model. A surroundings-person separation and fusion mechanism plays the key role to realize an accurate and effective surroundings-person separation under a mutually exclusion constraint. (a) Illustration of the proposed feature extraction procedure in DSSL. (b) Illustration of the five diverse alignment paradigms adopted to adequately utilize multi-modal and multi-granular information and hence improve the retrieval accuracy. V_G/T_G, V_P/T_P, V_R/T_R and V_L/T_L denote the extracted visual/textual global, person, reconstructed and local features, respectively.

image and textual representations with a novel staircase CNN network and a local constrained BERT model. Besides, a multi-modal re-ranking algorithm by comparing the visual neighbors of the query to the gallery (RVN) is utilized to further improve the retrieval performance.

3 METHODOLOGY

In this section, we describe the proposed Deep Surroundings-person Separation Learning (DSSL) model in detail (shown in Fig. 3), which consists of a Surroundings-Person Separation Module (SPSM), a Surroundings-Person Fusion Module (SPFM), a Signal Denoising Module (SDM), a Person Describing Module (PDM) and a Salient Attention Module (SAM).

3.1 Feature Extraction And Refinement

3.1.1 Feature Extraction. We utilize a ResNet-50 [6] backbone pre-trained on ImageNet to extract global/local visual features from a given image I. To obtain the global feature $V_G \in \mathbb{R}^p$, the feature map before the last pooling layer of ResNet-50 is down-scaled to a vector $\in \mathbb{R}^{1\times1\times2048}$ with an average pooling layer and then passed through a group normalization (GN) layer followed by a fully-connected (FC) layer. In the local branch, the same feature map is first horizontally k-partitioned by pooling it to $k \times 1 \times 2048$, and then the local strips are separately passed through a GN and two FCs with a ReLU layer between them to form k p-dim vectors, which are finally concatenated to obtained the local visual feature matrix $M_V \in \mathbb{R}^{k\times p}$.

For textual feature extraction, we take a whole sentence and the n phrases extracted from it as textual materials, which are handled by a bi-directional GRU (bi-GRU). The last hidden states of the forward and backward GRUs are concatenated to give global/local p-dim feature vectors. The p-dim vector got from the whole sentence is passed through a GN followed by an FC to form the global textual feature $T_G \in \mathbb{R}^p$. With each certain input phrase, the corresponding output p-dim vector is processed consecutively by a GN and two FCs with a ReLU layer between them and then concatenated with each other to form the local textual feature matrix $M_T \in \mathbb{R}^{n\times p}$.

3.1.2 Textual Person Information Refinement. To further remove the noise signals in the textual data, so as to refine the extracted person information, the global feature vector T_G and local feature vectors in M_T are separately handled by a **Signal Denoising Module (SDM)**. With a fixed zeroing ratio r, a fixed number of elements in an input vector is set to zero [22]. And then the processed vector is reconstructed following an autoencoder manner to obtain the textual person feature vector T_P and the local textual person feature matrix T_L:

$$y = Dec(Enc(Z(x, r))), \tag{1}$$

where $Z(x, r)$ denotes the zero setting operation with ratio r, $x \in \{T_G\} \cup M_T$ and $y \in \{T_P\} \cup T_L$. With the zeroing and reconstruction mechanism, the input vectors are required to fully retain effective information while discarding redundant noise signals. The reconstruction loss of SDM is defined as:

$$L_{SDM} = L_{rank}(T_G, T_P) + \sum_{i=1}^{k} L_{rank}((M_T)_i, (T_L)_i), \tag{2}$$

where $(M_T)_i$ and $(T_L)_i$ denote the i-th vector in matrices M_T and T_L. Rather than being superficially look-alike, the denoised vector ought to be properly matched with the original vector because of the special nature of a retrieval task. Therefore, instead of utilizing the traditional Euclidean Distance to guide the reconstruction, a

triplet ranking loss is adopted:

$$L_{rank}(x_1, x_2) = \sum_{\widehat{x_2}} max\{\alpha - S(x_1, x_2) + S(x_1, \widehat{x_2}), 0\} + \sum_{\widehat{x_1}} max\{\alpha - S(x_1, x_2) + S(\widehat{x_1}, x_2), 0\}, \quad (3)$$

to more accurately constrain the matched pairs to be closer than the mismatched pairs with a margin α, where $(x_1, \widehat{x_2})$ or $(\widehat{x_1}, x_2)$ denotes a mismatched pair and $S(\cdot, \cdot)$ is the cosine similarity between two vectors. Instead of using the furthest positive and closest negative sampled pairs, we adopt the sum of all pairs within each mini-batch when computing the loss following [4].

3.2 Deep Surroundings-Person Separation Learning

As shown in Fig. 3 (b), five alignment paradigms are adopted to adequately utilize multi-modal and multi-granular information for a robust Deep Surroundings-Person Separation Learning process, thereby improving the retrieval accuracy.

3.2.1 Align I. To process the visual data, the person feature V_P and surroundings feature V_S are separated through a **Surroundings-Person Separation Module (SPSM)**, which is implemented as two paralleled multi-layer perceptrons (MLP) (with the feature dimension conversion as $p \rightarrow 2p \rightarrow p$) followed by a *tanh* layer:

$$V_P, V_S = SPSM(V_G). \quad (4)$$

The person features extracted from both modalities are first aligned. The alignment loss for *AlignI* is

$$L_{AlignI} = L_{rank}(V_P, T_P). \quad (5)$$

Besides, a **Mutually Exclusion Constraint (MEC)** is proposed to ensure that V_P and V_S are orthogonal to each other and the visual information is distributed between them without overlap. Let $P = \{V_P^i\}_{i=1}^B \in \mathbb{R}^{B \times p}$ and $S = \{V_S^i\}_{i=1}^B \in \mathbb{R}^{B \times p}$ denote matrices whose rows are person and surroundings features in a training batch, respectively, where B is the batch size, and then the mutually exclusion loss is

$$L_{MEC} = \|P^T S\|. \quad (6)$$

3.2.2 Align II. With a proposed **Surroundings-Person Fusion Module (SPFM)**, T_P is fused with V_S and reconstructed into the visual modality as V_R:

$$V_R = SPFM(T_P, V_S), \quad (7)$$

which is then aligned with V_G and the alignment loss for *AlignII* is

$$L_{AlignII} = L_{rank}(V_G, V_R). \quad (8)$$

SPFM first combine the two input vectors by addition or concatenation (compared in Section 4.2.1), and then the combined feature is processed by an MLP similar to *SPSM*.

3.2.3 Align III. A **Person Describing Module (PDM)**, which is implemented as an MLP with a *tanh* activation function is employed to reconstruct V_P into the textual modality as T_R and then aligned with T_G:

$$T_R = PDM(V_P). \quad (9)$$

The alignment loss for *AlignIII* is

$$L_{AlignIII} = L_{rank}(T_G, T_R). \quad (10)$$

3.2.4 Align IV. A **Salient Attention Module (SAM)** is first employed to highlight person information in the local visual feature matrix M_V:

$$(V_L)_i = Sigmoid(W_2(GN(ReLU(W_1(V_P)+b_1)))+b_2) \cdot (M_V)_i, \quad (11)$$

where GN denotes the group normalization layer while W_1, W_2 and b_1, b_2 denote the linear transformation. To adequately exploit fine-grained clues, a cross-modal attention (CA) mechanism [18, 25] is utilized to align V_L with the textual person feature T_P and form a p-dim vector:

$$CA(V_L, T_P) = \sum_{\alpha_V^i > \frac{1}{k}} \alpha_V^i (V_L)_i, \alpha_V^i = \frac{exp(cos((V_L)_i, T_P))}{\sum_{j=1}^k exp(cos((V_L)_j, T_P))}, \quad (12)$$

where α_V^i represents the relation between the i-th local visual part and textual person feature. And the alignment loss for *AlignIV* is

$$L_{AlignIV} = L_{rank}(CA(T_P, V_L), T_P). \quad (13)$$

3.2.5 Align V. Similar with *AlignIV*, the alignment loss for *AlignV* is

$$L_{AlignV} = L_{rank}(CA(V_P, T_L), V_P), \quad (14)$$

$$CA(T_L, V_P) = \sum_{\alpha_T^i > \frac{1}{n}} \alpha_T^i (T_L)_i, \alpha_T^i = \frac{exp(cos((T_L)_i, V_P))}{\sum_{j=1}^n exp(cos((T_L)_j, V_P))}. \quad (15)$$

3.3 Loss Function for Training

The complete training process includes 2 stages.

3.3.1 Stage-1. We first fix the parameters of the ResNet-50 backbone and train the left feature extraction part of DSSL with the identification (ID) loss

$$L_{id}(X) = -log(softmax(W_{id} \times GN(X)) \quad (16)$$

to cluster person images into groups according to their identification, where $W_{id} \in \mathbb{R}^{Q \times p}$ is a shared transformation matrix implemented as a FC layer without bias and Q is the number of different people in the training set. As global features can provide more complete information for clustering, only V_G and T_G are utilized here:

$$L_{ID1} = L_{id}(V_G) + L_{id}(T_G). \quad (17)$$

And the entire loss in Stage-1 is

$$L_{Stage1} = L_{ID1}. \quad (18)$$

3.3.2 Stage-2. In this stage, all the parameters of DSSL are fine-tuned together. Here the ID loss is also employed to ensure that the person features and reconstructed features can be correctly related to the corresponding person:

$$L_{ID2} = L_{ID1} + L_{id}(V_P) + L_{id}(T_P) + L_{id}(V_R) + L_{id}(T_R). \quad (19)$$

The five alignment losses are utilized to improve retrieval accuracy:

$$L_{Alignment} = L_{AlignI} + L_{AlignII} + L_{AlignIII} + L_{AlignIV} + L_{AlignV}. \quad (20)$$

Along with the mutually exclusion loss, the entire loss in Stage-2 is

$$L_{Stage2} = L_{ID2} + L_{Alignment} + L_{MEC}. \quad (21)$$

4 EXPERIMENTS

4.1 Experimental setup

4.1.1 Dataset and metrics. Our approach is evaluated on two challenging datasets: CUHK-PEDES [12] and our proposed Real Scenario Text-based Person Re-identification (RSTPReid) dataset.

CUHK-PEDES. Previously, CUHK-PEDES [12] is the only available dataset for text-based person retrieval task. Following the official data split approach, the training set contains 34054 images, 11003 persons and 68126 textual descriptions. The validation set contains 3078 images, 1000 persons and 6158 textual descriptions while the testing set has 3074 images, 1000 persons and 6156 descriptions. Every image generally has two descriptions, and each sentence is commonly no shorter than 23 words. After dropping words that appear less than twice, the word number is 4984.

RSTPReid. To properly handle real scenarios, we construct a new dataset called Real Scenario Text-based Person Re-identification (RSTPReid) based on MSMT17 [26]. RSTPReid contains 20505 images of 4,101 persons from 15 cameras. Each person has 5 corresponding images taken by different cameras and each image is annotated with 2 textual descriptions. For data division, 3701, 200 and 200 identities are utilized for training, validation and testing, respectively. Each sentence is no shorter than 23 words. After dropping words that appear less than twice, the word number is 2204. High-frequency words and examples of person images in RSTPReid are shown in Fig. 4.

The performance is evaluated by the top-k accuracy. Given a query description, all test images are ranked by their similarities with this sentence. If any image of the corresponding person is contained in the top-k images, we call this a successful search. The top-1, top-5, and top-10 accuracy for all experiments are reported.

4.1.2 Implementation details. The feature dimension p is set to 1024 and the number of local strips k is set to 6. The total number of phrases n obtained from each sentence is kept flexible with an upper bound 26, which are obtained with the Natural Language ToolKit (NLTK) by syntactic analysis, word segmentation and part-of-speech tagging. We adopt an Adam optimizer to train DSSL with a batch size of 32. The margin α of ranking losses is set to 0.2. In training stage-1, DSSL is trained with a learning rate of 1×10^{-3} for 10 epochs with the ResNet-50 backbone fixed. In stage-2, the learning rate is initialized as 2×10^{-4} to optimize all parameters including the visual backbone for extra 30 epochs. The learning rate is down-scaled by $\frac{1}{10}$ every 10 epochs. λ_1 and λ_2 in Sim_{TI} are both set to 0.5. In testing and real application, a cross-modal re-ranking scheme (RR) based on [24] is employed to further improve the retrieval accuracy in testing and real application.

4.2 Ablation Analysis

To further investigate the effectiveness and contribution of each proposed component in DSSL, we perform a series of ablation studies on the CUHK-PEDES dataset. The top-1, top-5 and top-10 accuracies (%) are reported and the best result in each table is presented in bold. As shown in Table 1, comparing with a baseline which is proposed following IMG-Net [25] without the Inner-Modal Self-Attention Module, DSSL achieves superior performance on both

Figure 4: High-frequency words and person images in our constructed RSTPReid dataset.

CUHK-PEDES [12] and our proposed RSTPReid with the aid of proper surroundings-person separation. Additionally, images of each person in RSTPReid are caught by different ones out of 15 independent cameras in both indoor and outdoor scenarios in various periods of time and thereby differ in illumination condition, weather, view angle, body position, etc., which makes RSTPReid obviously a much more challenging benchmark, on which the retrieval performance stumbles, hence leaving much space for further research.

4.2.1 Surroundings-person separation and fusion mechanism. As shown in Table 2, the retrieval result in the first row is given by a model without the surroundings-person separation and fusion mechanism (*SPSM* + *SPFM*) along with the mutually exclusion constraint (*MEC*). It directly mapping multi-modal data into a latent common space as many of the existing methods do. The top-1, top-5 and top-10 performances drop sharply by 4.46%, 2.79% and 2.37%, respectively, which demonstrates that our proposed method is more able to properly catch discriminative clues about the corresponding person while drop the misaligned information from complex high-dimensional multi-modal data than unconstrained mapping paradigms. By merely utilizing a *SPSM* + *SPFM* mechanism, without a mutually exclusion constraint, the top-1, top-5 and top-10 performances improve by 1.79%, 1.71% and 1.67%, respectively, which further proves the validity of *SPSM*+*SPFM*. However, the performance is still 2.67%, 1.08% and 0.70% respectively worse than the complete DSSL. This suggests that without the mutually exclusion constraint to ensure the orthogonality between person and surroundings features, information is not able to be well distributed between them. In Table 3, the addition and concatenation methods for combing the two input vectors before handled by the MLP in SPFM are compared. It turns out that the two methods give similar results, with the addition method slightly better.

Some examples of the top-5 text-based person retrieval results by DSSL are shown in Fig. 5. Images of the target pedestrian are marked by red rectangles. As can be seen in the figure, many of the pedestrians in mismatched person images also look quite similar to the target one, which is consistent with the distribution pattern of features discussed above. It seems necessary to find ways

Table 1: Ablation analysis of the five alignment paradigms in DSSL on CUHK-PEDES and RSTPReid.

-	-	-	-	-	-	CUHK-PEDES			RSTPReid		
Baseline	*AlignI*	*AlignII*	*AlignIII*	*AlignIV*	*AlignV*	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
✓	×	×	×	×	×	52.42	76.06	84.94	26.31	46.90	58.33
×	✓	×	×	×	×	56.18	79.56	86.45	28.74	50.88	61.69
×	×	✓	×	×	×	55.01	77.68	85.25	26.83	49.55	59.91
×	×	×	✓	×	×	54.56	78.49	85.64	27.01	50.02	60.67
×	×	×	×	✓	×	54.65	78.30	85.51	26.73	50.71	60.25
×	×	×	×	×	✓	51.01	75.47	83.01	25.73	48.99	59.82
×	✓	✓	×	×	×	57.31	79.56	86.42	29.23	51.55	61.77
×	✓	×	✓	×	×	56.73	79.21	86.65	28.87	51.81	62.43
×	×	✓	✓	×	×	57.08	79.11	86.06	29.51	51.89	62.22
×	✓	✓	✓	×	×	58.86	79.70	86.95	31.00	53.83	62.63
×	✓	×	×	✓	✓	58.19	79.41	86.52	30.81	53.67	62.71
×	✓	✓	✓	✓	✓	**59.98**	**80.41**	**87.56**	**32.43**	**55.08**	**63.19**

Table 2: Ablation analysis of the mutually exclusion constraint (MEC), surroundings-person separation and fusion (SPSM + SPFM), salient attention module (SAM) and signal denoising module (SDM) on CUHK-PEDES.

MEC	*SPSM + SPFM*	*SAM*	*SDM*	Top-1	Top-5	Top-10
×	×	✓	✓	55.52	77.62	85.19
×	✓	✓	✓	57.31	79.33	86.86
✓	✓	✓	✓	**59.98**	**80.41**	**87.56**
✓	✓	×	✓	57.95	79.89	87.20
✓	✓	✓	×	57.46	79.67	87.19

Table 3: Performance comparison of the feature combination method utilized in SPFM on CUHK-PEDES.

Method	Top-1	Top-5	Top-10
Addition	**59.98**	80.41	**87.56**
Concatenation	59.54	**80.45**	87.17

Table 4: Performance comparison of zeroing rate r in SDM on CUHK-PEDES.

r	Top-1	Top-5	Top-10
0	57.84	79.97	87.47
0.25	58.61	**81.05**	87.28
0.5	**59.98**	80.41	**87.56**
0.75	58.48	80.18	87.29
0.9	54.35	77.71	86.09

to dig deeper into the semantic information and draw similar clusters closer without mixing with each other, which remains for our future work.

Table 5: Performance comparison of Euclidean distance and ranking loss utilized in SDM on CUHK-PEDES.

Method	Top-1	Top-5	Top-10
Euclidean Distance	58.93	80.32	87.47
Ranking Loss	**59.98**	**80.41**	**87.56**

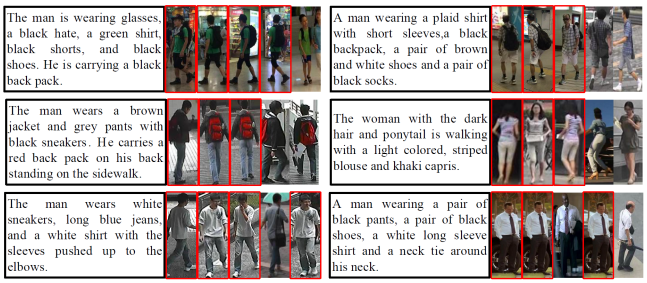

Figure 5: Examples of top-5 text-based person retrieval results by DSSL. Images of the target pedestrian are marked by red rectangles.

4.2.2 Alignment paradigms. To adequately utilize multi-modal and multi-granular information for a higher retrieval accuracy, five different alignment paradigms are adopted. Extensive ablation experiments are conducted on both CUHK-PEDES and RSTPReid to prove the effectiveness of them and the results are reported in Table 1. The results show that utilizing more than one single alignment brings performance gain, which indicates that the use of multi-modal and multi-granular features in DSSL can provide more comprehensive information, hence leading to a more accurate retrieval. By combining *AlignII* and *AlignIII* with *AlignI*, SPSM can more completely separate person and surroundings information with the aid of SPFM and PDM. Besides, comparing the third row from the bottom with the last row in Table 1, the top-1, top-5 and top-10 performance increase by 1.12%, 0.71%, 0.61% and 1.43, 1.25, 0.56

Table 6: Comparison with other state-of-the-art methods on CUHK-PEDES.

Method	Top-1	Top-5	Top-10
CNN-RNN [20]	8.07	-	32.47
Neural Talk [23]	13.66	-	41.72
GNA-RNN [12]	19.05	-	53.64
IATV [11]	25.94	-	60.48
PWM-ATH [3]	27.14	49.45	61.02
Dual Path [31]	44.40	66.26	75.07
GLA [2]	43.58	66.93	76.26
MIA [18]	53.10	75.00	82.90
A-GANet [14]	53.14	74.03	81.95
GALM [8]	54.12	75.45	82.97
TIMAM [17]	54.51	77.56	84.78
IMG-Net [25]	56.48	76.89	85.01
CMAAM [1]	56.68	77.18	84.86
HGAN [30]	59.00	79.49	86.6
NAFS [5]	59.94	79.86	86.70
DSSL (ours)	**59.98**	**80.41**	**87.56**
NAFS + RVN [5]	61.50	81.19	87.51
DSSL + RR (ours)	**62.33**	**82.11**	**88.01**

respectively on CUHK-PEDES and RSTPReid after the two fine-grained alignments *AlignIV* and *AlignV* are added, which reals the effect of utilizing multi-granular clues.

4.2.3 Signal denoising module (SDM). Comprehensive experimental analysis is as well carried out to study the proposed signal denoising module (SDM). As shown in Table 4, ablation experiments are conducted to search for the optimal zeroing rate r. It can be observed that initially the performance of DSSL follows a increasing tendency with the growth of r. After reaching a peak, the performance begins to turn worse as r continues to go larger. It is conceivable that by randomly dropping a certain number of elements in the input vector at random and then reconstructing it following a autoencoder manner, which is required to be well matched with the original feature under a ranking loss, redundant noise signals are inclined to be removed. Note that when r is set to 0, there is no zero setting process before the input vector is reconstructed. With the growth of r, SDM gradually finds an optimal zeroing rate that the noise signal are just properly dropped while person information is well retained, which gives a summit in performance. After bypassing the summit, an excess of amount of information will be discarded, and hence the retrieval performance will undoubtedly go down. As can be seen in Table 4, when r reaches 0.9, the accuracies of top-1, top-5 and top-10 all fall sharply.

Besides, we compare the performance of utilizing Euclidean distance and ranking loss in SDM (shown in Table 5). The top-1 accuracy of DSSL with ranking loss in SDM is 1.05% higher than the one with Euclidean distance, which indicates that rather than the commonly used Euclidean distance for reconstruction, ranking loss is better at dealing with the particularity of retrieval problems. We also train and evaluate DSSL without the whole SDM (shown in Table 2). The top-1, top-5 and top-10 performance drop by 2.52%, 0.74% and 0.37% respectively, which reveals the effect of SDM as well.

4.2.4 Salient attention module (SAM). As shown in Table 2, the top-1 accuracy drops by 2.03% without the salient attention module (SAM) which utilize the extracted person information to highlight and catch body part information in the visual local features. The results indicate the effectiveness of SAM.

4.3 Comparison With Other State-of-the-art Methods

Table 6 shows the comparison of DSSL against 15 previous state-of-the-art methods including CNN-RNN [20], Neural Talk [23], GNA-RNN [12], IATV [11], PWM-ATH [3], Dual Path [31], GLA [2], MIA [18], A-GANet [14], GALM [8], TIMAM [17], IMG-Net [25], CMAAM [1], HGAN [30] and NAFS [5] in terms of top-1, top-5 and top-10 accuracies in the text-based person retrieval task. Our proposed DSSL achieves 59.98%, 80.41% and 87.56% of top-1, top-5 and top-10 accuracies, respectively. It can be observed that DSSL outperforms existing methods, which proves the effectiveness of our proposed method. Both with a cross-modal re-ranking method, DSSL outperforms NAFS as well. With person and surroundings information separated properly, DSSL surpasses methods which directly map data into a common space. Moreover, compared to methods building similarities based on attention mechanism, DSSL achieves a significant performance improvement, which indicates that our proposed surroundings-person separation mechanism is more able to properly capture detailed person information.

5 CONCLUSION

In this paper, we propose a novel Deep Surroundings-person Separation Learning (DSSL) model to effectively extract and match person information, and hence achieve a superior retrieval accuracy. A surroundings-person separation and fusion mechanism plays the key role to realize an accurate and effective surroundings-person separation under a mutually exclusion constraint. In order to adequately utilize multi-modal and multi-granular information for a higher retrieval accuracy, Five diverse alignment paradigms are adopted. Extensive experiments are carried out to evaluate the proposed DSSL on CUHK-PEDES, which is currently the only accessible dataset for text-base person retrieval task. DSSL outperforms previous methods and achieves the state-of-the-art performance on CUHK-PEDES. To properly evaluate the proposed method in the real scenarios, a Real Scenarios Text-based Person Reidentification (RSTPReid) dataset is further constructed to benefit future research on text-based person retrieval.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61972016 and 61802176), China Postdoctoral Science Foundation (Grant No.2019M661999) and Natural Science Foundation of Jiangsu Higher Education Institutions of China (19KJB520009).

REFERENCES

[1] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. 2020. Text-based person search via attribute-aided matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2617–2625.

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. 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)*. 54–70.

[3] T. Chen, C. Xu, and J. Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1879–1887.

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[5] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. 2021. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search. *arXiv preprint arXiv:2101.03036* (2021).

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9317–9326.

[8] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search. *arXiv preprint arXiv:1809.08440* (2018).

[9] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[10] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[11] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 1890–1899.

[12] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[14] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep adversarial graph attention convolution network for text-based person search. In *Proceedings of the 27th ACM International Conference on Multimedia*. 665–673.

[15] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4107–4116.

[16] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 299–307.

[17] Ioannis A. Kakadiaris Nikolaos Sarafianos, Xiang Xu. 2019. Adversarial Representation Learning for Text-to-Image Matching, In ICCV. *ICCV*, 5813–5823.

[18] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing* 29 (2020), 5542–5556.

[19] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[20] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–58.

[21] Changchang Sun, Xuemeng Song, Fuli Feng, Wayne Xin Zhao, Hao Zhang, and Liqiang Nie. 2019. Supervised hierarchical cross-modal hashing. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 725–734.

[22] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[24] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. 2019. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*. 12–20.

[25] Zijie Wang, Aichun Zhu, Zhe Zheng, Jing Jin, Zhouxin Xue, and Gang Hua. 2020. IMG-Net: inner-cross-modal attentional multigranular network for description-based person re-identification. *Journal of Electronic Imaging* 29, 4 (2020), 043028.

[26] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[27] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. 2019. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3760–3769.

[28] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3441–3450.

[29] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 34–39.

[30] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. 2020. Hierarchical Gumbel Attention Network for Text-based Person Search. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3441–3449.

[31] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.