Knowledge-Based Systems

SUM: Serialized Updating and Matching for Text-based Person Retrieval

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	person retrieval; text-based person re-identification; cross-modal retrieval
Corresponding Author:	Aichun Zhu CHINA
First Author:	Zijie Wang
Order of Authors:	Zijie Wang
	Aichun Zhu
	Jingyi Xue
	Daihong Jiang
	Chao Liu
	Yifeng Li
	Hichem Snoussi
Abstract:	The central problem of text-based person retrieval is how to properly bridge the gap between heterogeneous cross-modal data. Most existing methods consider and align multi-modal semantics equally. Many of them either utilizing attention mechanism or directly mapping cross-modal information into a common space in a one-off manner, which can be inconsistent with the fact that humans usually follow a step-by-step process to properly recognize and match two objects. Intuitively, the large heterogeneity gap between multi-modal data can be better bridged by gradually analyzing the complex cross-modal relationships. In this paper, we propose a Serialized Updating and Matching (SUM) method for text-based person retrieval to bridge the heterogeneity gap between cross-modal data in a progressive manner. The core component of SUM is the proposed Memory Gating Modules (MGM), which can be stacked to gradually update and match features extracted from visual/textual modalities. To fully excavate the correlations lie within multi-granular cross-modal data, two variants are designed to care for both global and fine-grain local information, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) with which the updating rate of information at each step is dynamically determined after observing the feature in opposite modality. Moreover, SUM can be flexibly utilized as an add-on to any multi-granular text-based person retrieval methods to further improve the performance. We evaluate our proposed method on CUHK-PEDES, which is currently the only available dataset for text-based person re-identification. Experimental results present that the proposed SUM outperforms existing methods and achieves the state-of-the-art performance.

SUM: Serialized Updating and Matching for Text-based Person Retrieval

Zijie Wang[a], Aichun Zhu[a,*], Jingyi Xue[a], Daihong Jiang[b], Chao Liu[c], Yifeng Li[a], Hichem Snoussi[d]

[a]*School of Computer Science and Technology, Nanjing Tech University, Nanjing, China*
[b]*School of Information Engineering, Xuzhou University of Technology, Xuzhou, China*
[c]*School of Intelligent Science and Control Engineering, Jinling Institute of Technology, Nanjing, China*
[d]*ICD - LM2S, Université de Technologie de Troyes, France*

Abstract

The central problem of text-based person retrieval is how to properly bridge the gap between heterogeneous cross-modal data. Most existing methods consider and align multi-modal semantics equally. Many of them either utilizing attention mechanism or directly mapping cross-modal information into a common space in a one-off manner, which can be inconsistent with the fact that humans usually follow a step-by-step process to properly recognize and match two objects. Intuitively, the large heterogeneity gap between multi-modal data can be better bridged by gradually analyzing the complex cross-modal relationships. In this paper, we propose a Serialized Updating and Matching (SUM) method for text-based person retrieval to bridge the heterogeneity gap between cross-modal data in a progressive manner. The core component of SUM is the proposed Memory Gating Modules (MGM), which can be stacked to gradually update and match features extracted from visual/textual modalities. To fully excavate the correlations lie within multi-granular cross-modal data, two variants are designed to care for both global and fine-grain local information, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) with which the updating rate of information at each

*Corresponding author
Email address: aichun.zhu@njtech.edu.cn (Aichun Zhu)

step is dynamically determined after observing the feature in opposite modality. Moreover, SUM can be flexibly utilized as an add-on to any multi-granular text-based person retrieval methods to further improve the performance. We evaluate our proposed method on CUHK-PEDES, which is currently the only available dataset for text-based person re-identification. Experimental results present that the proposed SUM outperforms existing methods and achieves the state-of-the-art performance.

Key words: person retrieval, text-based person re-identification, cross-modal retrieval

2010 MSC: 00-01, 99-00

---

1. Introduction

Person retrieval aims at searching for the images of a target pedestrian in a large-scale image gallery according to a given query. Currently, researches of person retrieval mainly interest in image-based person retrieval [1–3], which is also known as person re-identification. However, it may suffer from the lack of query images of the targeted person in some real-world scenarios. Instead, queries in type of textual description are much easier to access in practical application, and hence text-based person retrieval [4–10] has drawn remarkable attention in recent years.

The central problem of the text-based person retrieval task is how to properly bridge the gap between heterogeneous cross-modal data, which requests for an effective feature extraction and matching paradigm with more detailed cross-modal interaction. Besides, text-based person retrieval has its own particularities compared with the general cross-modal retrieval task. The task of text-based person retrieval is proposed to handle multi-modal data, namely, gallery images of pedestrians and the corresponding textual descriptions, and thus it can be regarded as a subtask of cross-modal retrieval [11–23]. However, each image cared by the general cross-modal retrieval task contains various categories of objects, while images for text-based person retrieval only involve one

certain pedestrian. In addition, the textual description queries for text-based person retrieval offer much more detailed cues about the corresponding pedestrian rather than roughly mention the objects in an image or even just give abstract understanding of the image. The above mentioned particularities of text-based person retrieval make many previous methods proposed on general cross-modal retrieval benchmarks (e.g. Flickr30K [24] and MSCOCO [25]) generalize poorly on it, and also indicates that the cross-modal information should thoroughly interact with each other in a fine-grained manner to achieve better performance.

Many of the previous works [6–10] contrive to learn a latent common space to bridge the modality gap and extract modality-invariant feature vectors to enable better cross-modal information. Nevertheless, directly mapping cross-modal information into a common space in a one-off and unconstrained manner may not properly take full advantage of the high-dimensional data to catch discriminative clues. Instead, it can be preferable to enable information obtained from multi-modal data to interact with each other step by step and get refined gradually.

Intuitively, in order to properly recognize and match two variant objects, humans usually follow a step-by-step process as well. To be specific, at first we often coarsely recognize the objects separately in a low semantic level. And then by observing and comparing between them back and forth, we can progressively care for higher-level semantics such as fine-grained information and cross-modal relationships. After adequate comparison, whether the two objects match with each other is better determined than deciding at first glance. This is consistent with the nature of text-based person retrieval and indicates that the large heterogeneity gap between cross-modal data ought to be bridged by gradually analyzing the complex cross-modal relationships.

To this end, in this paper, we propose a novel Serialized Updating and Matching (SUM) method for text-based person retrieval to bridge the heterogeneity gap between cross-modal data in a progressive manner. The core component of SUM is the proposed Memory Gating Modules (MGM), which can

be stacked to gradually update and match the features extracted from visual and textual modalities. To fully excavate the correlations lie within the multi-granular cross-modal data, two variants of MGM are designed to care for both global and fine-grain local information, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM). With the employed GMGM and FMGM, visual and textual features interact with each other in serial and the updating rate of information at each step is dynamically determined after observing the feature in the opposite modality. Moreover, SUM can be flexibly utilized as an add-on to any multi-granular text-based person retrieval methods to further improve the performance. We evaluate our proposed method on the CUHK-PEDES dataset [4], which is a challenging dataset currently only available for text-based person retrieval. Experimental results present that the proposed SUM outperforms the existing methods and achieves the state-of-the-art performance.

The main contributions of this paper can be summarized as threefold:

- A Serialized Updating and Matching (SUM) method for text-based person retrieval is proposed to bridge the heterogeneity gap between cross-modal data in a progressive manner.
- The Memory Gating Modules (MGM) play the key role in SUM, which can be stacked to gradually update and match cross-modal features. To fully excavate the correlations lie within the multi-granular data, two variants of MGM, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM), are designed to care for both global and fine-grain local information.
- A comprehensive study is carried out to evaluate the proposed SUM method. Experimental results demonstrate that SUM significantly outperforms existing methods and achieves the state-of-the-art performance.

2. Related Works

2.1. Person Re-identification

Person re-identification has drawn increasing attention in both academical and industrial fields [4, 26–37]. This technology addresses the problem of matching pedestrian images across disjoint cameras. The key challenges lie in the large intra-class and small inter-class variation caused by different views, poses, illuminations, and occlusions. Existing methods can be grouped into handed-crafted descriptors, metric learning methods and deep learning methods. With the development of deep learning [38–50], deep learning methods are in general playing a major role in current state-of-the-art works. Yi et al. [1] firstly proposed deep learning methods to match people with the same identification. Xia et al. [3] proposed the Second-order Non-local Attention (SONA) Module to learn local/non-local information and relationships in a more end-to-end way. In order to strengthen the representation capability of the deep neural network, Hou et al. [2] proposed the Interaction-and-Aggregation (IA) Block, which consists of a Spatial Interaction-and-Aggregation (SIA) Module and a Channel Interaction-and-Aggregation (CIA) Module and can be inserted into deep CNNs at any depth. To bridge the gap between theoretical research and practical application, Zhang et al. [37] propose a large and real-scenario person re-identification dataset for night scenario named KnightReid. Image denoising networks combined with common used person re-identification networks can be adapted to this kind of problem. Yuan et al. [36] propose a Gabor convolution module for deep neural networks based on Gabor function, which has a good texture representation ability and is effective when it is embedded in the low layers of a network. Taking advantage of the hinge function, they also design a new regularizer loss function to make the proposed Gabor Convolution module meaningful.

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. [4] and they take an LSTM to handle the input image and text. An efficient patch-word matching model [51] is proposed to capture the local similarity between image and text. Jing et al. [7] utilize pose information as soft attention to localize the discriminative regions. Niu et al. [6] propose a Multi-granularity Image-text Alignments (MIA) model exploit the combination of multiple granularities. Nikolaos et al. [8] 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. [9] 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. [10] 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. [52] which learns an attribute-driven space along with a class-information driven space by introducing extra attribute annotation and prediction. Zheng et al. [53] 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.

3. Methodology

3.1. Problem Formulation

The goal of the proposed framework is to measure the similarity between cross-modal data, namely, a given textual description and a gallery person image. Formally, let $D = \{i_i, t_i\}_{i=1}^{N}$ denote a training set consists of N image-text pairs. Each pair contains a pedestrian image captured by one certain surveillance camera and its corresponding textual description. The IDs of pedestrian in X are $Y = \{y_i\}_{i=1}^{Q}$. Given a textual description, the aim is to identify images of the most relevant pedestrian from a large scale person image gallery.

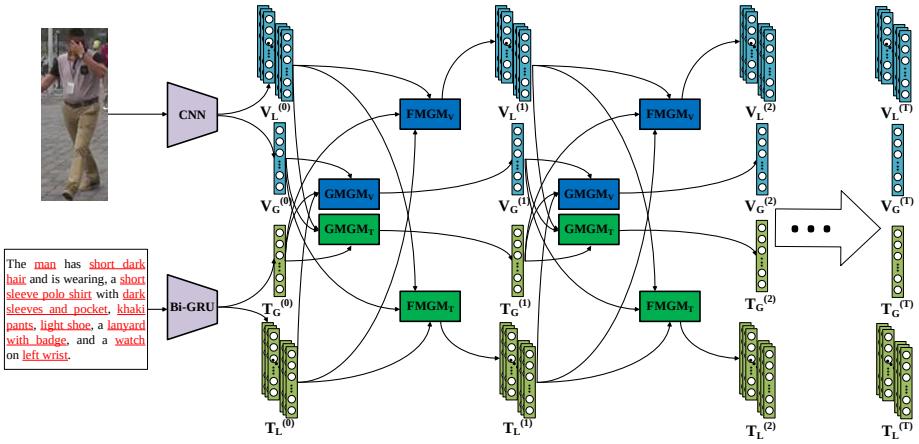


Figure 1: Illustration of the proposed Serialized Updating and Matching (SUM) method, which is proposed to bridge the heterogeneity gap between cross-modal data in a progressive manner. The Memory Gating Modules (MGM) play the key role in SUM, which can be stacked to gradually update and match cross-modal features. To fully excavate the correlations lie within the multi-granular data, two variants of MGM, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM), are designed to care for both global and fine-grain local information.

3.2. Feature Extraction

As mentioned in Section 1, our proposed SUM approach can be employed as a flexible add-on to any text-based person retrieval method which cares for multi-granular cross-modal data. In this paper, we extract multi-granular features from both the visual and textual modalities following a general paradigm that is commonly utilized in some of the existing methods [6, 9].

3.2.1. Visual Feature Extraction

A ResNet-50 [54] backbone pretrained on ImageNet is utilized to extract global/local visual features from a given image I. To obtain the global feature $V_G^{(0)} \in \mathbb{R}^p$, 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 $V_L^{(0)} \in \mathbb{R}^{k \times p}$.

3.2.2. Textual Feature Extraction

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 $2p$-dim feature vectors. The $2p$-dim vector got from the whole sentence is passed through a GN followed by an FC to form the global textual feature $T_G^{(0)} \in \mathbb{R}^p$. 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 $T_L^{(0)} \in \mathbb{R}^{n \times p}$.

3.3. Model Architecture

As shown in Fig. 1, the central part of the Serialized Updating and Matching (SUM) method is the proposed Memory Gating Modules (MGM). To fully excavate the correlations lie within the multi-granular data, we carried out two variants including Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) to consider both global and fine-grained local clues. To update either the global or fine-grained local features extracted from one certain modality, the message carried by multi-granular features obtained from the opposite modality is utilized. In the following part of this section, we first introduce the proposed Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) in detail in Sec. 3.3.1 and 3.3.2, respectively. And then the mechanism of the Serialized Updating and Matching (SUM) method is described in Sec. 3.3.3.

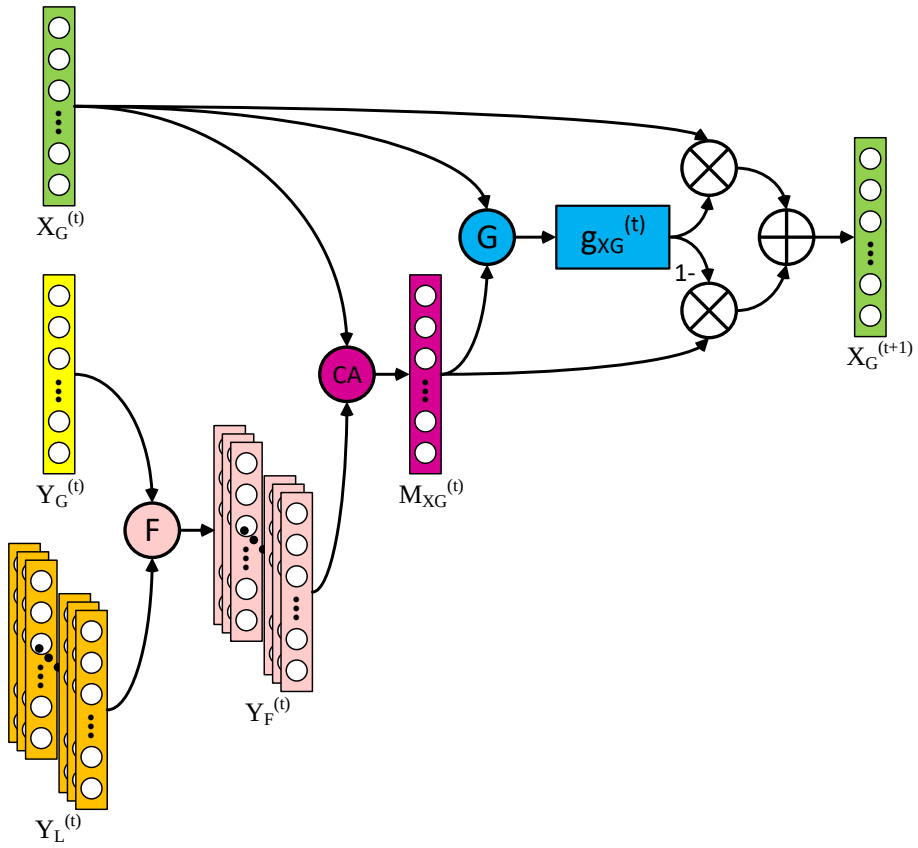


Figure 2: Illustration of the Global Memory Gating Module (GMGM), which is proposed to update and match both the visual and textual global features.

3.3.1. Global Memory Gating Module (GMGM)

The Global Memory Gating Module (GMGM) is proposed to update and match both the visual and textual global features (shown in Fig. 2), which is formulated as:

$$X_G^{(t)} = GMGM(X_G^{(t-1)}, Y_G^{(t-1)}, Y_L^{(t-1)}), \tag{1}$$

where X_G denotes the global feature to be updated which is obtained from one certain modality, while Y_G and Y_L denote the global and local features extracted from the opposite modality utilized as updating message. (X, Y) can be (V, T) or (T, V). $t \in \{1, 2, \cdots, T\}$ and T is the total time step number, namely, the

total number of stacked GMGM blocks.

To be specific, Y_G and Y_L are first fused to form a set of unified features Y_F, which contains both global and fine-grained local information:

$$Y_{Fi} = \frac{Y_G + Y_{Li}}{2},\ i \in \{1, 2, \cdots, num\}, \tag{2}$$

where num can be k or n for visual or textual data, respectively. The feature fusion paradigm can be implemented as several other methods (e.g. concatenation and addition), which will be further discussed in Sec. 4.2.3. Then a cross-modal attention ($\mathcal{CA}$) mechanism [6, 9] is employed to generate a updating message. The global updating message M_{XG} is calculated following

$$\alpha^i_{XG} = \frac{exp(cos(Y_{Fi}, X_G))}{\sum_{m=1}^{num} exp(cos(Y_{Fm}, X_G))}. \tag{3}$$

$$M_{XG} = \mathcal{CA}(Y_F, X_G) = \sum_{\alpha^i_{XG} > \frac{1}{num}} \alpha^i_{XG} Y_{Fi}, \tag{4}$$

α^i_{XG} represents the cross-modal relation between the i-th fused feature Y_{Fi} and global feature X_G.

At the t-th time step, a updating gate is calculated with the global feature X_G and the global updating message M_{XG}:

$$g^{(t)}_G = Gating(X^{(t-1)}_G, M^{(t-1)}_{XG}) = \sigma(\mathcal{F}_{gg}(X^{(t-1)}_G \oplus M^{(t-1)}_{XG})), \tag{5}$$

where $g^{(t)}_G$ is the global gating value at the t-th time step. $\mathcal{F}_{gg}(\cdot)$ denotes a linear transformation function and $\sigma(\cdot)$ stands for the sigmoid function. $\oplus$ is the feature fusion operation and can be implemented as several variants, which will be further discussed in Sec. 4.2.4.

With the obtained gating value, the global feature X_G is updated following

$$X^{(t)}_G = g^{(t)}_G X^{(t-1)}_G + (1 - g^{(t)}_G) M^{(t-1)}_{XG}. \tag{6}$$

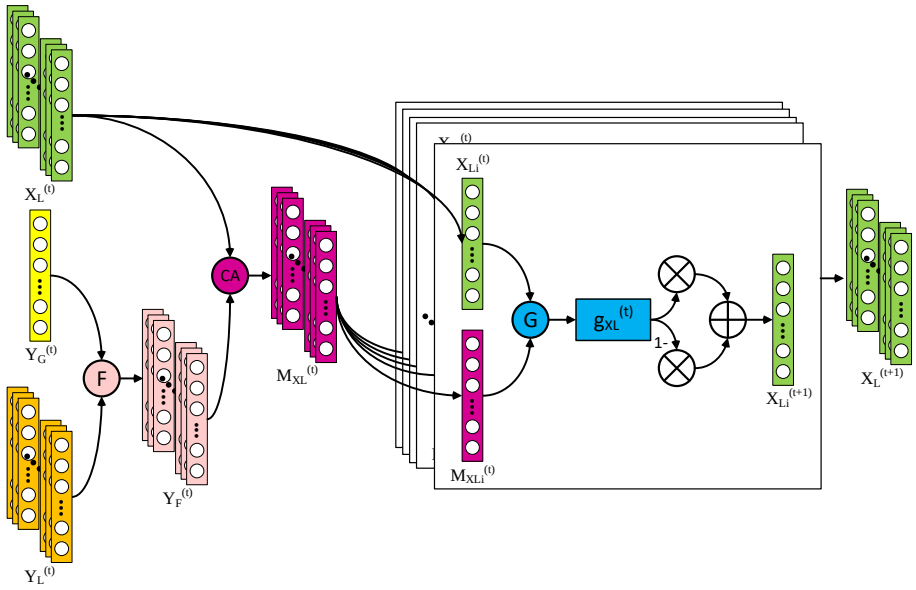


Figure 3: Illustration of the Fine-grained Memory Gating Module (FMGM), which is proposed to update and match both the visual and textual local features.

3.3.2. Fine-grained Memory Gating Module (FMGM)

The structure of the Fine-grained Memory Gating Module (FMGM) is illustrated in Fig. 3, which is formulated as

$$X_L^{(t)} = FMGM(X_L^{(t-1)}, Y_G^{(t-1)}, Y_L^{(t-1)}), \tag{7}$$

where X_L denotes the local features to be updated which is obtained from one certain modality, while Y_G and Y_L denote the global and local features extracted from the opposite modality which are utilized as updating message. (X, Y) can be (V, T) or (T, V). $t \in \{1, 2, \cdots, T\}$ and T is the total time step number, namely, the total number of stacked FMGM blocks.

Similar to GMGM, Y_G and Y_L are first fused to form a set of unified features Y_F which contains both global and fine-grained local information:

$$Y_{Fi} = \frac{Y_G + Y_{Li}}{2}, \ i \in \{1, 2, \cdots, num\}, \tag{8}$$

where num can be k or n for visual or textual data, respectively.

Then the cross-modal attention ($\mathcal{CA}$) mechanism is applied on each local feature vector in X_L to generate a set of local updating messages M_{XL}:

$$\alpha_{XL}^{ij} = \frac{exp(cos(Y_{Fi}, X_{Gj})))}{\sum_{m=1}^{num} exp(cos(Y_{Fm}, X_{Gj}))}. \tag{9}$$

$$M_{XLj} = \mathcal{CA}(Y_F, X_L^j) = \sum_{\alpha_{XL}^{ij} > \frac{1}{num}} \alpha_{XL}^{ij} Y_{Fi}, \tag{10}$$

α_{XL}^{ij} represents the cross-modal relation between the i-th fused feature Y_{Fi} and the j-th local feature X_{Lj}.

Then in num paralleled branches, num local updating gates $\{g_{L1}, g_{L2}, \cdots, g_{L(num)}\}$ are calculated according to the corresponding pairs of local feature and updating message $\{(X_{L1}, M_{XL1}), (X_{L2}, M_{XL2}), \cdots, (X_{L(num)}, M_{XL(num)}),\}$:

$$g_{Li}^{(t)} = Gating(X_{Li}^{(t-1)}, M_{XLi}^{(t-1)}) = \sigma(\mathcal{F}_{lg}(X_{Li}^{(t-1)} \oplus M_{XLi}^{(t-1)})), \tag{11}$$

where $g_{Li}^{(t)}$ is the i-th local updating gate at the t-th time step, $i \in \{1, 2, \cdots, num\}$. $\mathcal{F}_{lg}(\cdot)$ denotes a linear transformation function and $\sigma(\cdot)$ stands for the sigmoid function.

After that, each local feature is updated as

$$X_{Li}^{(t)} = g_{Li}^{(t)} X_{Li}^{(t-1)} + (1 - g_{Li}^{(t)}) M_{XLi}^{(t-1)}. \tag{12}$$

3.3.3. Serialized Updating and Matching (SUM)

By means of the proposed Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM), the extracted features can be updated in a serialized manner:

$$V_G^{(t)} = GMGM_V(V_G^{(t-1)}, T_G^{(t-1)}, T_L^{(t-1)}), \tag{13}$$

$$T_G^{(t)} = GMGM_T(T_G^{(t-1)}, V_G^{(t-1)}, V_L^{(t-1)}), \tag{14}$$

$$V_L^{(t)} = FMGM_V(V_L^{(t-1)}, T_G^{(t-1)}, T_L^{(t-1)}), \tag{15}$$

$$T_L^{(t)} = FMGM_T(T_L^{(t-1)}, V_G^{(t-1)}, V_L^{(t-1)}), \tag{16}$$

where $t \in \{1, 2, \cdots, T\}$. At each time step t, cross-modal similarities are calculated within three different combinations:

$$S_{GG}^{(t)} = Sim(V_G^{(t)}, T_G^{(t)}), \tag{17}$$

$$S_{GL}^{(t)} = \sum_{i=1}^{n} Sim(V_G^{(t)}, T_{Li}^{(t)}), \tag{18}$$

$$S_{LG}^{(t)} = \sum_{i=1}^{k} Sim(V_{Li}^{(t)}, T_G^{(t)}), \tag{19}$$

where $Sim(\cdot, \cdot)$ denotes the Cosine similarity between two feature vectors. The overall cross-modal similarity at the t-th time step is

$$S^{(t)} = S_{GG}^{(t)} + \lambda(S_{GL}^{(t)} + S_{LG}^{(t)}). \tag{20}$$

The final similarity for cross-modal matching is

$$S = \sum_{t=0}^{T} S^{(t)}. \tag{21}$$

3.4. Loss Function And Training Strategy

The complete training process includes 3 stages.

3.4.1. Stage-1

We first fix the parameters of the ResNet-50 backbone and train the left feature extraction parts with the identification (ID) loss

$$L_{id}(X) = -log(softmax(W_{id} \times GN(X)) \tag{22}$$

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 fully-connected (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 $V_G^{(0)}$ and $T_G^{(0)}$ are utilized here:

$$L_{ID}^{(0)} = L_{id}(V_G^{(0)}) + L_{id}(T_G^{(0)}). \tag{23}$$

And the entire loss in Stage-1 is

$$L_{Stage1} = L_{ID}^{(0)}. \tag{24}$$

3.4.2. Stage-2

In this stage, all the parameters of the feature extraction model are fine-tuned together including ones in the visual backbone. The ID loss $L_{ID}^{(0)}$ is still employed along with a triplet ranking loss $L_{TR}^{(0)}$.

The triplet ranking loss is commonly adopted in either person re-identification or text-based person retrieval tasks, which aims to constrain the matched pairs to be closer than the mismatched pairs in a mini-batch with a margin α. Following [55], we employ the sum of all pairs within each mini-batch when computing the hinge-based triplet ranking loss instead of utilizing the furthest positive and closest negative sampled pairs:

$$\begin{aligned} L_{ranking}(V,T) = \sum_{\widehat{T}} max\{\alpha - cos(V,T) + cos(V,\widehat{T}), 0\} \\ + \sum_{\widehat{V}} max\{\alpha - cos(V,T) + cos(\widehat{V},T), 0\}, \end{aligned} \tag{25}$$

where V can be V_G or V_L, while T can be T_G or T_L, respectively. (V,T) denotes the matched visual-textual pairs while $(V,\widehat{T})$ or $(\widehat{V},T)$ denotes the mismatched pairs and α is a margin. At time step 0, the general triplet ranking loss $L_{TR}^{(0)}$ on raw features without serialized updating is calculated following:

$$L_{TR}^{(0)} = L_{ranking}(V_G^{(0)}, T_G^{(0)}) + L_{ranking}(V_L^{(0)}, T_G^{(0)}) + L_{ranking}(V_G^{(0)}, T_L^{(0)}). \tag{26}$$

The complete loss function in Stage-2 is

$$L_{Stage2} = L_{ID}^{(0)} + L_{TR}^{(0)}. \tag{27}$$

Intuitively, the identification loss mainly focuses on the ID category of a given person, which functions more like a loose constraint thereby failing to provide adequate accuracy for the fine-grained matching task. As the triplet ranking loss regards the description sentences annotated for a certain image as negative for any other images even with the same person ID, it is much stricter. Thus, the ID loss in Stage-1 can eliminate obvious mismatched pairs and as well provide an initialization for Stage-2. Then in Stage-2 the triplet ranking losses are employed to catch more fine-grained information and in this stage the ID losses are still reserved to function as a regularization for the model.

3.4.3. Stage-3

Now that the feature extraction model is well pretrained, our proposed Serialized Updating and Matching (SUM) method is employed on top of it to further improve the retrieval performance. At each time step, the ID loss and the triplet ranking loss are calculated and summed up:

$$L_{ID} = \sum_{t=0}^{T} L_{ID}^{(t)} = \sum_{t=0}^{T} (L_{id}(V_G^{(t)}) + L_{id}(T_G^{(t)})), \tag{28}$$

$$L_{TR} = \sum_{t=0}^{T} L_{TR}^{(t)} = \sum_{t=0}^{T} (L_{ranking}(V_G^{(t)}, T_G^{(t)}) + L_{ranking}(V_L^{(t)}, T_G^{(t)}) + L_{ranking}(V_G^{(t)}, T_L^{(t)})). \tag{29}$$

Therefore, the complete loss for this stage is

$$L_{Stage3} = L_{ID} + L_{TR}. \tag{30}$$

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset and Metrics

The CUHK-PEDES dataset is currently the only dataset for the Text-Based Person Retrieval task. We follow the same data split approach as [4]. In detail,

the training set contains 34054 images, 11003 persons and 68126 textual descriptions. There are 3078 images, 1000 persons and 6158 textual descriptions in the validation set while 3074 images, 1000 persons and 6156 textual descriptions in the testing set. Almost every image has two descriptions, and each sentence is generally no shorter than 23 words.

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. We report the top-1, top-5, and top-10 accuracies for all experiments.

4.1.2. Implementation Details

In our experiments, we set the dimensionality $p = 1024$. The word number W is 4984 after dropping the words that appears less than twice and the dimensionality E of embedded word vectors is set to 300. We choose the pre-trained ResNet-50 [54] as the visual CNN backbone. We obtain noun phrases of each sentence with the Natural Language ToolKit (NLTK) by syntactic analysis, word segmentation and part-of-speech tagging. The total number of noun phrases obtained from each sentence is kept flexible. In training, we initialize the weights of the ResNet-50 backbone pre-trained on the ImageNet classification task. An Adam optimizer [56] is adopted to train the model with a batch size of 32. The margin α of ranking losses is set to 0.2 and λ is set to 0.5. In training stage-1, we start the iteration with a learning rate of 1×10^{-3} for 10 epochs with all weights in the ResNet-50 backbone fixed. In stage-2, we first initialize the learning rate to 2×10^{-4}. During the early 15 epochs, we just let the Adam optimizer to find its own way down. After that, the initial learning rate for later epochs is defined as:

$$lr = 2 \times 10^{-4} \times (\frac{1}{10})^{epoch//10}, \tag{31}$$

where lr means the learning rate and $\cdot//\cdot$ denotes a division operation only takes the integer part. We totally train the stage-2 for 30 epochs. Then in stage-3,

the learning rate is also initialized as 2×10^{-4} and is decayed by 1/10 every 10 epochs. With an Adam optimizer, the model is trained for 20 epochs in this stage.

4.2. Ablation Analysis

To further investigate the effectiveness and contribution of each proposed component in SUM, a series of ablation studies are performed on the CUHK-PEDES dataset, which is currently the only available dataset for text-based person retrieval. The top-1, top-5 and top-10 accuracies (%) are reported and the best result in each table is presented in bold.

4.2.1. Total Number T of Time Steps

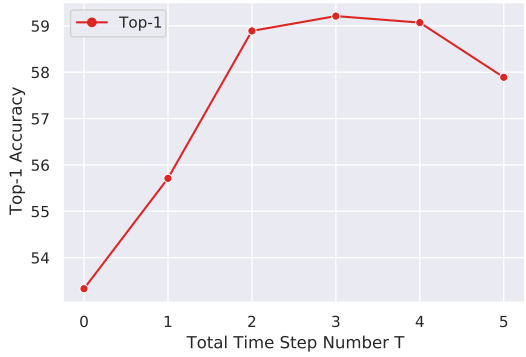
All of the ablation experiments are conducted with T increasing from 1 to 5 to search for the optimal number T of total time steps. As reported in Tab. 1, all combinations of MGMs gives a consistent and reasonable phenomenon that initially the performance of SUM keeps improving with the increase of T, and then after reaching a peak ($T = 3$), the retrieval accuracy begins to turn worse as T continues to go larger. In order to see the trend more clearly, the top-1, top-5 and top-10 accuracies achieved by the full SUM method on the CUHK-PEDES dataset are illustrated in Fig. 4. From $T = 0$ to $T = 1$, and especially from $T = 1$ to $T = 2$, the performance experiences significant improvement. This observation properly demonstrates the effectiveness of updating and matching cross-modal information in a serialized manner. Then from $T = 2$ to $T = 4$, after a relatively small improvement, performance of the model began to decline slightly. And there is a somewhat large drop in the retrieval accuracy when T achieves 5, which is reasonable as the multi-modal information can be over-smoothed after too much cross-modal interaction.

Table 1: Ablation analysis of the Memory Gating Modules (MGM) and total time step number T on CUHK-PEDES.

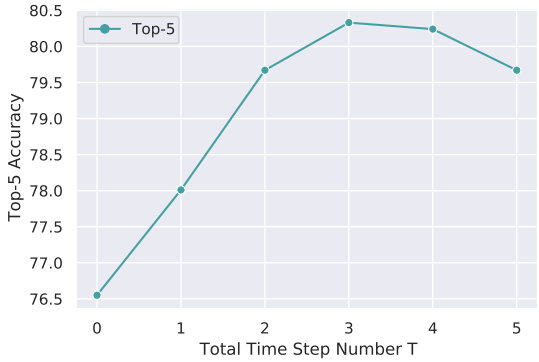
$GMGM_V$	$GMGM_T$	$FMGM_V$	$FMGM_T$	T	Top-1	Top-5	Top-10
×	×	×	×	0	53.33	76.55	85.91
✓	×	×	×	1	54.99	77.81	86.21
×	✓	×	×	1	55.03	77.81	86.21
×	×	✓	×	1	54.19	77.77	86.12
×	×	×	✓	1	54.00	76.89	85.91
✓	✓	×	×	1	55.48	77.94	86.25
×	×	✓	✓	1	55.41	77.66	86.03
✓	✓	✓	×	1	55.67	77.97	**86.27**
✓	✓	×	✓	1	55.55	77.91	86.19
✓	✓	✓	✓	1	**55.71**	**78.01**	86.14
✓	×	×	×	2	58.00	79.27	86.64
×	✓	×	×	2	57.89	79.19	86.58
×	×	✓	×	2	57.81	79.04	86.32
×	×	×	✓	2	57.12	78.81	86.01
✓	✓	×	×	2	58.13	79.33	86.67
×	×	✓	✓	2	58.36	79.12	86.27
✓	✓	✓	×	2	58.83	79.63	86.72
✓	✓	×	✓	2	58.78	79.43	86.88
✓	✓	✓	✓	2	**58.89**	**79.67**	**87.00**
✓	×	×	×	3	58.56	79.64	86.95
×	✓	×	×	3	58.49	79.71	86.88
×	×	✓	×	3	58.32	79.65	86.88
×	×	×	✓	3	57.67	79.04	86.48
✓	✓	×	×	3	58.61	79.80	86.91
×	×	✓	✓	3	58.80	79.67	86.91

Table 1: Ablation analysis of the Memory Gating Modules (MGM) and total time step number T on CUHK-PEDES.

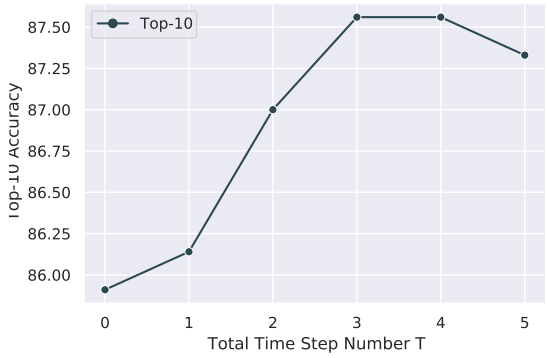
$GMGM_V$	$GMGM_T$	$FMGM_V$	$FMGM_T$	T	Top-1	Top-5	Top-10
✓	✓	✓	×	3	59.17	80.14	87.41
✓	✓	×	✓	3	59.11	80.20	87.36
✓	✓	✓	✓	3	**59.21**	**80.33**	**87.56**
✓	×	×	×	4	58.41	79.55	86.70
×	✓	×	×	4	58.36	79.43	86.81
×	×	✓	×	4	57.88	79.54	86.79
×	×	×	✓	4	57.28	79.14	86.67
✓	✓	×	×	4	58.44	79.49	86.88
×	×	✓	✓	4	58.94	79.52	87.01
✓	✓	✓	×	4	59.11	80.16	87.33
✓	✓	×	✓	4	59.04	79.83	87.26
✓	✓	✓	✓	4	**59.07**	**80.24**	**87.56**
✓	×	×	×	5	56.83	79.09	86.55
×	✓	×	×	5	56.83	79.09	86.55
×	×	✓	×	5	56.64	79.14	86.55
×	×	×	✓	5	55.91	78.73	86.32
✓	✓	×	×	5	56.83	79.09	86.55
×	×	✓	✓	5	57.52	78.81	86.67
✓	✓	✓	×	5	57.81	79.32	87.20
✓	✓	×	✓	5	57.77	79.11	87.19
✓	✓	✓	✓	5	**57.89**	**79.67**	**87.33**



(a) Top-1 Accuracies



(b) Top-5 Accuracies



(c) Top-10 Accuracies

Figure 4: Illustration of ablation analysis on the total time step number T. Top-1, Top-5 and Top-10 accuracies are illustrated, respectively.

4.2.2. Combinations of the Memory Gating Modules

The complete SUM method employs both a Global Memory Gating Module (GMGM) and a Fine-grained Memory Gating Module (FMGM) for each modality, termed $(GMGM_V, FMGM_V)$ and $(GMGM_T, FMGM_T)$, respectively. To in depth explore the effects of these modules, we conduct ablation analysis with various combinations of them on the CUHK-PEDES dataset and the results are reported in Tab 1. It can be observed that any combination of MGMs achieves better performance than only adopting a single one, which proves the significance of updating various features. And compared with the combination of either the two GMGMs or the two FMGMs, the performance given by combining three or all of the four MGMs which includes both the global and fine-grained MGMs is exactly better. This observation further indicates that it is of necessity to excavate the inherent correlations lie within the multi-granular information. Therefore, the full SUM with four MGMs outperforms any other variants for all different values of T.

4.2.3. Choice of the Feature Fusion Method for Unified Feature Generation

Before calculating the updating messages, the input global and local features Y_G and Y_L are fused to generate the unified features Y_F, which contains both global and fine-grained local information. As mentioned in Sec. 3.3.1, the feature fusion method can be implemented as several variants, which includes feature concatenation, feature addition and feature averaging and are described below. Ablation analysis are conducted to study the effectiveness of them on CUHK-PEDES and the experimental results are recorded in Tab. 2.

Feature Concatenation. The input global feature vector Y_G is concatenated with each local feature vectors Y_{Li}:

$$Y_{Fi} = [Y_G, Y_{Li}], i \in \{1, 2, \cdots, num\}, \tag{32}$$

where num can be k or n for visual or textual data, respectively.

Table 2: Ablation analysis of the fusion methods for generating unified features on CUHK-PEDES.

Method	T	Top-1	Top-5	Top-10
Feature Concatenation	1	55.61	77.98	86.08
Feature Addition	1	55.68	**78.08**	86.11
Feature Averaging	1	**55.71**	78.01	**86.14**
Feature Concatenation	2	58.84	79.69	86.96
Feature Addition	2	58.86	**79.82**	86.96
Feature Averaging	2	**58.89**	79.67	**87.00**
Feature Concatenation	3	59.19	80.28	87.61
Feature Addition	3	59.17	80.26	**87.64**
Feature Averaging	3	**59.21**	**80.33**	87.56
Feature Concatenation	4	**59.11**	80.22	**87.59**
Feature Addition	4	59.04	80.13	87.44
Feature Averaging	4	59.07	**80.24**	87.56
Feature Concatenation	5	57.84	79.64	87.33
Feature Addition	5	**57.92**	79.67	**87.35**
Feature Averaging	5	57.89	**79.67**	87.33

Table 3: Ablation analysis of the fusion methods $\oplus$ in the Memory Gating Modules (MGM) on CUHK-PEDES.

Method	T	Top-1	Top-5	Top-10
Feature Concatenation	1	55.68	77.64	85.76
Feature Addition	1	55.62	77.87	86.12
Feature Averaging	1	**55.71**	**78.01**	**86.14**
Feature Concatenation	2	58.77	79.49	86.61
Feature Addition	2	**58.94**	79.61	86.96
Feature Averaging	2	58.89	**79.67**	**87.00**
Feature Concatenation	3	59.21	80.08	87.22
Feature Addition	3	59.11	80.27	87.54
Feature Averaging	3	**59.21**	**80.33**	**87.56**
Feature Concatenation	4	58.98	80.01	87.17
Feature Addition	4	59.02	**80.33**	**87.58**
Feature Averaging	4	**59.07**	80.24	87.56
Feature Concatenation	5	57.54	79.38	87.19
Feature Addition	5	57.76	79.57	87.30
Feature Averaging	5	**57.89**	**79.67**	**87.33**

Feature Addition. Y_{Fi} is given by directly adding Y_G with each Y_{Li}:

$$Y_{Fi} = Y_G + Y_{Li}. \tag{33}$$

Feature Averaging. Y_G and each Y_{Li} are fused into Y_{Fi} by means of averaging the values of them at corresponding positions as Eq. (2) and Eq. (8) show:

$$Y_{Fi} = \frac{Y_G + Y_{Li}}{2}. \tag{34}$$

As recorded in Tab. 2, with the change of T, the three employed feature fusion approaches achieve similar performance.

4.2.4. Choice of the Feature Fusion Method in the Memory Gating Modules

When calculating the updating gates in MGMs, the input feature is first fused with the updating message. Commonly there are several feature fusion paradigms to choose. In this paper we carried out ablation experiments to compare them with each other, including feature concatenation, feature addition and feature averaging, which are briefly introduced below. The results on the CUHK-PEDES dataset are reported in Tab. 3.

Feature Concatenation. The input feature and its corresponding updating message are concatenated with each other to obtain the fused feature:

$$X \oplus M_X = [X, M_X], \tag{35}$$

where (X, M_X) can be (X_G, M_{XG}) or (X_{Li}, M_{XLi}), which denotes the global feature and updating message or the i-th local feature and updating message, respectively.

Feature Addition. The input feature vector is directly added with the corresponding updating message vector:

$$X \oplus M_X = X + M_X. \tag{36}$$

Feature Averaging. The fused feature is given by calculating the average values of the input feature vector and the updating the updating message vector at corresponding positions:

$$X \oplus M_X = \frac{X + M_X}{2}. \tag{37}$$

As shown in Tab. 3, the three employed feature fusion paradigms achieve comparative performance with different values of T. The performance of feature concatenation is relatively poor while the feature averaging paradigm slightly outperform the other two paradigms.

Table 4: Comparison with several variations of the Memory Gating Module (MGM) on CUHK-PEDES.

Method	T	Top-1	Top-5	Top-10
$DirectAddition$	1	53.39	76.68	85.88
$Averaging$	1	53.54	76.65	85.95
$DirectConcatenation$	1	53.37	76.72	85.93
$FurtherFuse + ScalarGate$	1	54.52	77.46	86.03
$ScalarGate$	1	54.98	77.63	86.09
$FurtherFuse + VectorGate$	1	55.68	77.97	86.06
$SUM(ours)$	1	**55.71**	**78.01**	**86.14**
$DirectAddition$	2	56.63	78.38	86.44
$Averaging$	2	56.67	78.29	86.41
$DirectConcatenation$	2	57.29	78.40	86.52
$FurtherFuse + ScalarGate$	2	57.73	79.15	86.96
$ScalarGate$	2	57.81	79.28	86.99
$FurtherFuse + VectorGate$	2	58.84	79.53	**87.08**
$SUM(ours)$	2	**58.89**	**79.67**	87.00
$DirectAddition$	3	56.94	79.06	87.01
$Averaging$	3	57.20	79.03	87.01
$DirectConcatenation$	3	57.35	79.00	87.11
$FurtherFuse + ScalarGate$	3	58.04	79.89	87.38
$ScalarGate$	3	58.37	80.26	87.41
$FurtherFuse + VectorGate$	3	58.72	80.23	87.43
$SUM(ours)$	3	**59.21**	**80.33**	**87.56**
$DirectAddition$	4	56.81	78.90	87.12
$Averaging$	4	57.04	78.96	87.13
$DirectConcatenation$	4	57.17	78.94	87.23
$FurtherFuse + ScalarGate$	4	57.82	79.71	87.34

Table 4: Comparison with several variations of the Memory Gating Module (MGM) on CUHK-PEDES.

Method	T	Top-1	Top-5	Top-10
$ScalarGate$	4	58.13	79.76	87.41
$FurtherFuse + VectorGate$	4	58.61	80.15	87.35
$SUM(ours)$	4	**59.07**	**80.24**	**87.56**
$DirectAddition$	5	55.64	78.12	86.74
$Averaging$	5	55.88	78.37	86.83
$DirectConcatenation$	5	55.43	78.34	86.95
$FurtherFuse + ScalarGate$	5	56.64	79.18	87.11
$ScalarGate$	5	56.82	79.31	86.15
$FurtherFuse + VectorGate$	5	57.49	79.46	87.17
$SUM(ours)$	5	**57.89**	**79.67**	**87.33**

4.2.5. Comparison with Several Variations of Memory Gating Module (MGM)

The information updating paradigm, namely, the Memory Gating Module (MGM) is one of the core components for SUM. We enumerate several paradigms for updating information, which are described below. Extensive experiments are carried out by employing them as substitution of MGM to investigate the effectiveness of our proposed method in depth.

DirectAddition. Global and local feature vectors are directly added with the corresponding global and local updating messages to obtain the updated feature vectors:

$$X_G^{(t)} = X_G^{(t-1)} + M_{XG}^{(t-1)}, \tag{38}$$

$$X_{Li}^{(t)} = X_{Li}^{(t-1)} + M_{XLi}^{(t-1)}. \tag{39}$$

Averaging. Feature vectors are updated by calculated the average value of the input feature vectors and the updating messages:

$$X_G^{(t)} = \frac{X_G^{(t-1)} + M_{XG}^{(t-1)}}{2}, \tag{40}$$

$$X_{Li}^{(t)} = \frac{X_{Li}^{(t-1)} + M_{XLi}^{(t-1)}}{2}. \tag{41}$$

DirectConcatenation. The updated features are generated by directly concatenate the input feature vectors with the updating messages:

$$X_G^{(t)} = [X_G^{(t-1)}, M_{XG}^{(t-1)}], \tag{42}$$

$$X_{Li}^{(t)} = [X_{Li}^{(t-1)}, M_{XLi}^{(t-1)}]. \tag{43}$$

ScalarGate. Instead of the real-valued vector gate calculated with the input feature and the updating message, here a learned real-valued scalar gate is used. The feature information is updating following

$$X_G^{(t)} = s_G^{(t)} X_G^{(t-1)} + (1 - s_G^{(t)}) M_{XG}^{(t-1)}, \tag{44}$$

$$X_{Li}^{(t)} = s_{Li}^{(t)} X_{Li}^{(t-1)} + (1 - s_{Li}^{(t)}) M_{XLi}^{(t-1)}, \tag{45}$$

where s_G and s_L are the global and local learned real-valued scalar gates, respectively.

FurtherFuse+ScalarGate. Before updating, the input feature vectors are first further fused with the corresponding updating messages:

$$f_G^{(t)} = FurtherFuse(X_G^{(t-1)}, M_{XG}^{(t-1)}) = \sigma(\mathcal{F}_{gff}(X_G^{(t-1)} \oplus M_{XG}^{(t-1)})), \tag{46}$$

$$f_{Li}^{(t)} = FurtherFuse(X_{Li}^{(t-1)}, M_{XLi}^{(t-1)}) = \sigma(\mathcal{F}_{lff}(X_{Li}^{(t-1)} \oplus M_{XLi}^{(t-1)})), \tag{47}$$

where $f_G^{(t)}$ is the global gating value at the t-th time step while $f_{Li}^{(t)}$ is the i-th local fused feature. $\mathcal{F}_{gff}(\cdot)$ and $\mathcal{F}_{lff}(\cdot)$ denote linear transformation functions for global and local features, respectively, and $\sigma(\cdot)$ stands for the sigmoid function. Here, the feature fusion operation $\oplus$ is implemented as element-wise averaging. Then the fused features are utilized for information updating under the control of the learned real-valued scalar gates:

$$X_G^{(t)} = s_G^{(t)} f_G^{(t-1)} + (1 - s_G^{(t)}) M_{XG}^{(t-1)}, \tag{48}$$

$$X_{Li}^{(t)} = s_{Li}^{(t)} f_{Li}^{(t-1)} + (1 - s_{Li}^{(t)}) M_{XLi}^{(t-1)}, \tag{49}$$

FurtherFuse+VectorGate. Here the further fused features are updated the control of the real-valued vector gates which is the same as ones used in MGM:

$$X_G^{(t)} = g_G^{(t)} f_G^{(t-1)} + (1 - g_G^{(t)}) M_{XG}^{(t-1)}, \tag{50}$$

$$X_{Li}^{(t)} = g_{Li}^{(t)} f_{Li}^{(t-1)} + (1 - g_{Li}^{(t)}) M_{XLi}^{(t-1)}. \tag{51}$$

As can be observed from Tab. 4, after comparing with all other kinds of variants, our proposed Memory Gating Modules (MGM) shows superiority with all values of the total time step number T. Although the retrieval accuracy is also improved after progressively updating and matching the features, it is obvious and understandable that the three relatively rough methods including *DirectAddition*, *Averaging* and *DirectConcatenation* are not so competitive in performance. The only difference between the $FurtherFuse + VectorGate$ variant and our proposed SUM method is that the input features are further fused with the corresponding updating messages before information updating. The top-1 retrieval accuracies drop by 0.03%, 0.05%, 0.49%, 0.46% and 0.40% for $T = 1, 2, \cdots, 5$, respectively. The impact becomes more pronounced as the total time step number T go larger. Besides, comparing with methods using a real-valued scalar as the the gate (*ScalarGate* and $FurtherFuse$+*ScalarGate*), approaches with real-valued vector gates ($FurtherFuse + VectorGate$ and our

proposed full SUM) achieve obvious higher retrieval accuracies, which proves the effectiveness of our proposed method.

Figure 5: Examples of the top-5 text-based person retrieval results by SUM with $T = 0, 1, 2, \cdots, 5$. $T = 0$ means that there is no feature updating employed after the multi-modal features are extracted. Images of the target pedestrian are marked by green rectangles.

4.3. Analysis of The Retrieval Results

Some of the examples of the top-5 text-based person retrieval results by our proposed SUM method (with $T = 0, 1, 2, \cdots, 5$) are displayed in Fig. 5. $T = 0$ means that there is no feature updating employed after the multi-modal features are extracted. Images of the targeted pedestrian are marked by green rectangles.

As can be seen from Fig. 5, when the total time step number T is changing from 0 to 3, it tends to contain more and more images of the targeted person in the candidate list and the ranks of the targeted images are going higher

and higher. And from $T = 3$ to $T = 5$, the change in the list of candidates is slightly in the opposite direction. Therefore, with T varying from 0 to 5, the variation tendency for the retrieved candidate list of most queries shows obvious consistency with the tendency discussed in Sec. 4.2.1 that initially the performance of SUM keeps improving with the increase of T, and then after reaching a peak ($T = 3$), the retrieval accuracy begins to turn worse as T continues to go larger.

Besides, images of the target person in the candidate list given by the SUM models trained with varied total time step number T may sometimes appear in different orders. This can be a reasonable phenomenon due to the non-convex optimization nature of the training of deep learning models. Though given in varied orders, images of the target person are always retrieved properly as long as the key information is caught by the model.

After observing in more detail, for a well trained SUM model with different values of T, the majority of the mismatched persons in the candidate list (including in some of the top-1 mismatched cases) show high similarity with the targeted person in appearance and are to some extent conform to the description given by the query sentence. To be more specific, let's take the query sentence 'The woman is wearing a bright pink shirt and wearing black shorts. She has a black backpack and white shoes.' for example. The discrepancy between some mismatched images and images of the targeted person is quite trivial. For the third image in the candidate list for $T = 3$, there are mainly 3 trivial differences. First, the person in this image is a little 'girl' rather than a 'woman' mentioned in the query. Nevertheless, this is too a tiny difference in semantics to be attended by the model. Second, the person in this image wears a pair of light pink shoes instead of white, which is also hard to be caught as the part of shoes in the images is rather small and there is not much difference between these two colors as well. Third, the girl in this image has no 'black backpack', but the dark shade near her back can be somewhat confusing for the model. Aside from this image, some of the other images of mismatched persons also show quite trivial discrepancies. What's more, in some of the cases, images of the targeted person

can be hard to retrieve for certain reasons. Still taking the same query item for example, the second image in the candidate list for $T = 3$ is also a proper match for the query sentence. However, due to the occlusion of the targeted person caused by other pedestrians, the local part for the mentioned 'bright pink shirt' is rather non-obvious, which increases the difficulty to recognize this image properly. According to the above discussion, it is reasonable for this query item that the third image in the candidate list for $T = 3$ ranks high for any value of the total time step number T. This image always ranks top-3 and even is the top-1 candidate for $T = 0$ as the multi-modal features are matched relatively roughly in this stage. And when T is 0 or 1, the above mentioned hard targeted image is even not in the top-5 lists, which indicates that more cross-modal interaction is needed to better suppress the highly similar mismatched image samples and catch all images of the targeted person. With more thorough serialized feature updating and matching, it can be observed that the hard targeted image gets to rank higher than the highly similar mismatched images, which proves the effectiveness of our proposed method.

4.4. Comparison with Other State-of-the-art Methods

Table 5 shows the comparison of SUM against 14 previous state-of-the-art methods including CNN-RNN [57], Neural Talk [58], GNA-RNN [4], IATV [5], PWM-ATH [51], Dual Path [59], GLA [60], MIA [6], A-GANet [10], GALM [7], TIMAM [8], IMG-Net [9], CMAAM [52] and HGAN [53] in terms of top-1, top-5 and top-10 accuracies in the text-based person retrieval task. Our proposed DSSL achieves 59.21%, 80.33% and 87.56% of top-1, top-5 and top-10 accuracies, respectively. It can be observed that SUM outperforms existing methods, which proves the effectiveness of our proposed method. By updating and matching information in a progressive manner, SUM surpasses approaches which directly map the multi-modal data into a common space in a one-off and unconstrained manner, which indicates that the step-by-step process proposed to gradually analyze the complicated cross-modal relationships can be a better choice to properly bridge the large heterogeneity gap between multi-modal data.

Table 5: Comparison with other state-of-the-art methods on CUHK-PEDES.

Method	Top-1	Top-5	Top-10
CNN-RNN [57]	8.07	-	32.47
Neural Talk [58]	13.66	-	41.72
GNA-RNN [4]	19.05	-	53.64
IATV [5]	25.94	-	60.48
PWM-ATH [51]	27.14	49.45	61.02
Dual Path [59]	44.40	66.26	75.07
GLA [60]	43.58	66.93	76.26
MIA [6]	53.10	75.00	82.90
A-GANet [10]	53.14	74.03	81.95
GALM [7]	54.12	75.45	82.97
TIMAM [8]	54.51	77.56	84.78
IMG-Net [9]	56.48	76.89	85.01
CMAAM [52]	56.68	77.18	84.86
HGAN [53]	59.00	79.49	86.62
SUM (ours)	**59.21**	**80.33**	**87.56**

5. Conclusion

The central problem of the text-based person retrieval task is how to properly bridge the gap between heterogeneous cross-modal data. Most of the existing methods consider and align the multi-modal semantics equally, and many of them either utilizing attention mechanism or directly mapping cross-modal information into a common space in a one-off manner, which can be inconsistent with the fact that humans usually follow a step-by-step process to properly recognize and match two variant objects. Intuitively, the large heterogeneity gap between multi-modal data can be better bridged by gradually analyzing the complex cross-modal relationships. In this paper, we propose a Serialized Updating and Matching (SUM) method for text-based person retrieval to bridge the heterogeneity gap between cross-modal data in a progressive manner. The core component of SUM is the proposed Memory Gating Modules (MGM), which can be stacked to gradually update and match the features extracted from visual and textual modalities. To fully excavate the correlations lie within the multi-granular cross-modal data, two variants of MGM are designed to care for both global and fine-grain local information, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) with which the visual and textual features interact with each other in serial and the updating rate of information at each step is dynamically determined after observing the feature in the opposite modality. Moreover, SUM can be flexibly utilized as an add-on to any multi-granular text-based person retrieval methods to further improve the performance. We evaluate our proposed method on the CUHK-PEDES [4] dataset, which is currently the only available dataset for text-based person re-identification. Experimental results present that the proposed SUM outperforms existing methods and achieves the state-of-the-art performance.

6. Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (Grant No. 62101245), China Postdoctoral Science Foundation (Grant

No.2019M661999) and Natural Science Research of Jiangsu Higher Education Institutions of China (19KJB520009).

References

[1] D. Yi, Z. Lei, S. Liao, S. Z. Li, Deep metric learning for person re-identification, in: 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 34–39.

[2] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-aggregation network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9317–9326.

[3] B. N. Xia, Y. Gong, Y. Zhang, C. Poellabauer, Second-order non-local attention networks for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3760–3769.

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

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

[6] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, IEEE Transactions on Image Processing 29 (2020) 5542–5556.

[7] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided multi-granularity attention network for text-based person search, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11189–11196.

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

[9] Z. Wang, A. Zhu, Z. Zheng, J. Jin, Z. Xue, G. Hua, Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification, Journal of Electronic Imaging 29 (4) (2020) 043028.

[10] J. Liu, Z.-J. Zha, R. Hong, M. Wang, Y. Zhang, Deep adversarial graph attention convolution network for text-based person search, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 665–673.

[11] F. Yan, K. Mikolajczyk, Deep correlation for matching images and text, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3441–3450.

[12] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[13] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 299–307.

[14] Y. Liu, Y. Guo, E. M. Bakker, M. S. Lew, Learning a recurrent residual fusion network for multimodal matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4107–4116.

[15] C. Sun, X. Song, F. Feng, W. X. Zhao, H. Zhang, L. Nie, Supervised hierarchical cross-modal hashing, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 725–734.

[16] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.

[17] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross-modal retrieval, Knowledge-Based Systems 180 (2019) 38–50.

[18] H. Qiang, Y. Wan, Z. Liu, L. Xiang, X. Meng, Discriminative deep asymmetric supervised hashing for cross-modal retrieval, Knowledge-Based Systems 204 (2020) 106188.

[19] X. Dong, H. Zhang, X. Dong, X. Lu, Iterative graph attention memory network for cross-modal retrieval, Knowledge-Based Systems 226 (2021) 107138.

[20] Z. Yang, L. Yang, O. I. Raymond, L. Zhu, W. Huang, Z. Liao, J. Long, Nsdh: A nonlinear supervised discrete hashing framework for large-scale cross-modal retrieval, Knowledge-Based Systems 217 (2021) 106818.

[21] X. Shen, H. Zhang, L. Li, Z. Zhang, D. Chen, L. Liu, Clustering-driven deep adversarial hashing for scalable unsupervised cross-modal retrieval, Neurocomputing 459 (2021) 152–164.

[22] J. Dong, Z. Long, X. Mao, C. Lin, Y. He, S. Ji, Multi-level alignment network for domain adaptive cross-modal retrieval, Neurocomputing 440 (2021) 207–219.

[23] X. Wang, X. Zou, E. M. Bakker, S. Wu, Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval, Neurocomputing 400 (2020) 255–271.

[24] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[26] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. 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), 2018, pp. 480–496.

[27] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: Exemplar memory for domain adaptive person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 598–607.

[28] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1179–1188.

[29] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 393–402.

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

[31] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.

[32] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in:

Proceedings of the iEEE conference on computer vision and pattern recognition, 2016, pp. 1335–1344.

[33] Y.-J. Cho, K.-J. Yoon, Pamm: Pose-aware multi-shot matching for improving person re-identification, IEEE Transactions on Image Processing 27 (8) (2018) 3739–3752.

[34] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, Q. Tian, Deep representation learning with part loss for person re-identification, IEEE Transactions on Image Processing 28 (6) (2019) 2860–2871.

[35] J. Dai, P. Zhang, D. Wang, H. Lu, H. Wang, Video person re-identification by temporal residual learning, IEEE Transactions on Image Processing 28 (3) (2018) 1366–1377.

[36] Y. Yuan, J. Zhang, Q. Wang, Deep gabor convolution network for person re-identification, Neurocomputing 378 (2020) 387–398.

[37] J. Zhang, Y. Yuan, Q. Wang, Night person re-identification and a benchmark, IEEE Access 7 (2019) 95496–95504.

[38] A. Zhu, Z. Zheng, Y. Huang, T. Wang, J. Jin, F. Hu, G. Hua, H. Snoussi, Cacrowdgan: Cascaded attentional generative adversarial network for crowd counting, IEEE Transactions on Intelligent Transportation Systems.

[39] A. Zhu, Q. Wu, R. Cui, T. Wang, W. Hang, G. Hua, H. Snoussi, Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional lstm-cnn, Neurocomputing 414 (2020) 90–100.

[40] Y. Chen, Y. Zhang, H. Shu, J. Yang, L. Luo, J.-L. Coatrieux, Q. Feng, Structure-adaptive fuzzy estimation for random-valued impulse noise suppression, IEEE Transactions on Circuits and Systems for Video Technology 28 (2) (2016) 414–427.

[41] X. Yin, Q. Zhao, J. Liu, W. Yang, J. Yang, G. Quan, Y. Chen, H. Shu, L. Luo, J.-L. Coatrieux, Domain progressive 3d residual convolution net-

work to improve low-dose ct imaging, IEEE transactions on medical imaging 38 (12) (2019) 2903–2913.

[42] H. Wang, D. Sahoo, C. Liu, K. Shu, P. Achananuparp, E.-p. Lim, C. S. Hoi, Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism, IEEE Transactions on Multimedia.

[43] J. Daihong, D. Lei, P. Jin, et al., Facial expression recognition based on attention mechanism, Scientific Programming 2021.

[44] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, A. G. Hauptmann, Adaptive semi-supervised feature selection for cross-modal retrieval, IEEE Transactions on Multimedia 21 (5) (2018) 1276–1288.

[45] Y. Wang, H. Yang, X. Bai, X. Qian, L. Ma, J. Lu, B. Li, X. Fan, Pfan++: Bi-directional image-text retrieval with position focused attention network, IEEE Transactions on Multimedia.

[46] L. Gu, J. Liu, X. Liu, J. Sun, Deep loss driven multi-scale hashing based on pyramid connected network, IEEE Transactions on Multimedia 23 (2020) 939–954.

[47] S. Qian, D. Xue, Q. Fang, C. Xu, Adaptive label-aware graph convolutional networks for cross-modal retrieval, IEEE Transactions on Multimedia.

[48] M. Zhao, J. Liu, Z. Zhang, J. Fan, A scalable sub-graph regularization for efficient content based image retrieval with long-term relevance feedback enhancement, Knowledge-based systems 212 (2021) 106505.

[49] Y. Fang, B. Li, X. Li, Y. Ren, Unsupervised cross-modal similarity via latent structure discrete hashing factorization, Knowledge-Based Systems 218 (2021) 106857.

[50] F. Li, T. Wang, L. Zhu, Z. Zhang, X. Wang, Task-adaptive asymmetric deep cross-modal hashing, Knowledge-Based Systems 219 (2021) 106851.

[51] T. Chen, C. Xu, J. Luo, Improving text-based person search by spatial matching and adaptive threshold, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1879–1887.

[52] S. Aggarwal, V. B. Radhakrishnan, A. Chakraborty, Text-based person search via attribute-aided matching, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2617–2625.

[53] K. Zheng, W. Liu, J. Liu, Z.-J. Zha, T. Mei, Hierarchical gumbel attention network for text-based person search, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3441–3449.

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[55] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives, in: Proceedings of the British Machine Vision Conference (BMVC), 2018.

[56] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, 2015.

[57] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 49–58.

[58] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[59] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16 (2) (2020) 1–23.

[60] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, X. Wang, Improving deep visual representation for person re-identification by global and local image-language association, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 54–70.

Cover Letter

Dear Editors:

We would like to submit the enclosed manuscript entitled “SUM: Serialized Updating and Matching for Text-based Person Retrieval”, which we wish to be considered for publication in “Knowledge-Based Systems”. No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

In this work, we address the problem in the text-based person retrieval. propose a Serialized Updating and Matching (SUM) method for text-based person retrieval to bridge the heterogeneity gap between cross-modal data in a progressive manner. The core component of SUM is the proposed Memory Gating Modules (MGM), which can be stacked to gradually update and match features extracted from visual/textual modalities. To fully excavate the correlations lie within multi-granular cross-modal data, two variants are designed to care for both global and fine-grain local information, namely, Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) with which the updating rate of information at each step is dynamically determined after observing the feature in opposite modality. Moreover, SUM can be flexibly utilized as an add-on to any multi-granular text-based person retrieval methods to further improve the performance. We evaluate our proposed method on CUHK-PEDES, which is currently the only available dataset for text-based person re-identification. Experimental results present that the proposed SUM outperforms existing methods and achieves the state-of-the-art performance. This work could be considered as a task in the computer vision and multimedia, and I hope this paper is suitable for “Knowledge-Based Systems”.

We deeply appreciate your consideration of our manuscript, and we look forward to receiving comments from the reviewers. If you have any queries, please don’t hesitate to contact me at the address below.

Thank you and best regards.

Yours sincerely,

Aichun ZHU

E-mail: aichun.zhu@njtech.edu.cn

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

CRediT Author Statement

Zijie Wang: Software, Writing- Original draft preparation.
Aichun Zhu: Conceptualization, Methodology, Writing- Reviewing and Editing.
Jingyi Xue: Software, Validation.
Daihong Jiang: Resources, Formal analysis.
Chao Liu: Data curation, Visualization.
Yifeng Li: Supervision, Resources.
Hichem Snoussi: Supervision.