Knowledge-Based Systems 248 (2022) 108891

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

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], Fangqiang Hu [a]

[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

ARTICLE INFO

Article history:
Received 31 August 2021
Received in revised form 20 April 2022
Accepted 22 April 2022
Available online 29 April 2022

MSC:
00-01
99-00

Keywords:
Person retrieval
Text-based person re-identification
Cross-modal retrieval

ABSTRACT

The central problem of text-based person retrieval is how to properly bridge the gap between heterogeneous cross-modal data. Many of the previous works contrive to learn a latent common space to bridge the modality gap and extract modality-invariant feature vectors. Within these methods, the common space mapping and cross-modal information matching operations are conducted in a one-off manner, which aims to extract sufficient discriminative clues from the high-dimensional multi-modal data *at first glance*, but it is 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 step-by-step 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 two text-based person retrieval datasets CUHK-PEDES and RSTPReid along with two general cross-modal retrieval datasets Flickr8K and Flickr30K to see its generalization ability. Experimental results present that the proposed SUM outperforms existing methods and achieves the state-of-the-art performance.

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 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 [11–18]. However, text-based person retrieval has its own particularities compared with the general cross-modal retrieval task. To be specific, 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 [19] and MSCOCO [20]) 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.

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

* Corresponding author.
E-mail address: aichun.zhu@njtech.edu.cn (A. Zhu).

https://doi.org/10.1016/j.knosys.2022.108891
0950-7051/

and matching paradigm with more detailed cross-modal interaction. 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. The major limitation of these methods comes from the one-off manner for the common space mapping and cross-modal information matching operations. In other words, these paradigms aim to extract discriminative clues from the high-dimensional multi-modal data *at first glance*, which may fail to properly catch sufficient helpful details and give a sub-optimal retrieval performance. Intuitively, in order to properly recognize and match two variant objects, humans usually inversely follow a *step-by-step process*. 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 comparisons, whether the two objects match with each other can be better determined than decided at first glance. This habitual pattern of human beings is just consistent with the nature of text-based person retrieval, which indicates that the large heterogeneity gap between cross-modal data ought to be bridged by gradually analyzing the complex cross-modal relationships. Based on the above discussion, it seems preferable to enable information interaction within the proposed method progressively, and hence the discriminative features can get captured and refined gradually.

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 step-by-step manner. The core component of SUM is the proposed Memory Gating Modules (MGM), which can be stacked to gradually to update and match features extracted from the visual and textual modalities. To thoroughly excavate the correlations lying within the multi-granular cross-modal data, two variants of MGM are designed to care for both global and fine-grain local information, namely, a Global Memory Gating Module (GMGM) and a Fine-grained Memory Gating Module (FMGM). With the employed GMGM and FMGM, visual and textual features are enabled to interact with each other in serial and the updating rate of information at each step is dynamically determined after observing 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 CUHK-PEDES [4] and RSTPReid [21] datasets. In addition, we also conduct experimental analysis on the Flickr8K [22] and Flickr30K [19] datasets to further explore the generalization ability of SUM to the general cross-modal retrieval task. 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 fourfold:

- 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 step-by-step manner.
- The Memory Gating Modules (MGM) play a key role in SUM, which can be stacked to gradually update and match features extracted from the visual and textual modalities. To fully excavate the correlations lie within the multi-granular data, two variants of MGM, namely, a Global Memory Gating Module (GMGM) and a Fine-grained Memory Gating Module (FMGM), are designed to care for both global and fine-grain local information.
- To our knowledge, we are the first to integrate the step-by-step updating and matching mechanism into the task of text-based person retrieval.
- A comprehensive study is carried out to evaluate the proposed SUM method on CUHK-PEDES and RSTPReid. Experimental analysis are also conducted on the Flickr8K and Flickr30K datasets to further explore the generalization ability of SUM to the general cross-modal retrieval task. 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 [4,23–34]. 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 person re-ID approached either dedicate to design discriminative appearance representations or focus on learning a robust distance metric in the feature space. With the development of deeplearning [35–40], [?], deep learning methods are in general playing a major role in current appearance representation extraction works. The success of deep learning in image classification [41] spreads to re-ID in 2014, when Yi et al. [1] firstly proposed deep learning methods which employ a siamese neural network [42] to determine if a pair of input images belong to the same ID. The reason for choosing the siamese model is probably that the number of training samples for each identity is limited(usually two). 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. [34] 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. [33] 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. For metric learning, Bak et al. [43] proposed a one-shot learning algorithm for person re-ID where re-ID metric is decomposed into independent texture and color components. Hermans et al. [44] design a new variant of triplet loss called batch hard loss which makes the re-ID network easier to converge during training compared to the traditional triplet loss and improves the performance. Taking advantage of the hinge function, they also design a new regularizer loss function to make the proposed Gabor Convolution module meaningful. Hao et al. [45] propose a Modality Confusion Learning Network (MCLNet), of which the basic idea is to confuse two modalities, ensuring that the optimization is explicitly concentrated on the modality-irrelevant perspective. In recent years, methods for unsupervised person re-identification have gradually emerged. Unsupervised person re-identification means that the target data set is unlabeled but the auxiliary source data set is not necessarily unlabeled [46,47]. Existing unsupervised person ReID works can be concluded into three categories. The first

category utilizes hand-craft features [48]. But the features made by hand cannot be robust and discriminative. To solve this problem, second category [49] adopts clustering to estimate pseudo labels to train the CNN. However, these methods require good trained model. Recently, the third category is proposed, which improves unsupervised person ReID by using transfer learning. Some works [50,51] utilize transfer learning and minimize the attribute-level discrepancy by using extra attribute annotations.

2.2. RGB-infrared person retrieval

Compared with conventional RGB re-ID, the RGB-infrared re-ID is a relatively new problem and it needs to notice not only intra-class variations but also the modality discrepancy issue caused by different wavelength ranges of visible and infrared cameras [52]. Wu et al. [53] contribute a new multiple modality re-ID dataset named SYSU-MM01, including RGB and IR images, and proposes a deep zero-padding method to train a one-stream network towards automatically evolving domain-specific nodes in the network for cross-modality matching. Ye et al. [54] proposed a two-stage framework that included a TwO-stream CNN NEtwork (TONE) to learn multi-modality shareable feature representations and a Hierarchical Cross-modal Metric Learning (HCML) method. A novel and end-to-end Alignment Generative Adversarial Network (AlignGAN) for the RGB-IR re-ID task is generated by Wang et al. [55]. The proposed model consists of a pixel generator, a feature generator and a joint discriminator, which can exploit pixel alignment and feature alignment jointly.

2.3. Text-based person retrieval

Text-based person retrieval aims to search for the corresponding pedestrian image from a large-scaled person image database according to a given text query. Ye et al. [56] in the year of 2015 first come up with the task of Specific Person Retrieval via Incomplete Text Description, which aims to retrieve person images according to user-provided attributes. This task can be deemed as a prototype of the text-based person retrieval task. A specific attribute completion is proposed to enrich the original text query and generate a more expressive attribute vector. Formally, the task of text-based person retrieval is first put forward by Li et al. [4]. They collect the CUHK-PEDES dataset with detailed textual description annotations and take an LSTM to handle the input image and text. Most of the existing works adopt the cross-modality attention mechanism to attend to all the image regions of images and the corresponding words in the textual description. The core idea of these approaches is to obtain weighted alignments between image and text to alleviate the irrelevant matching. For instance, Li et al. proposed an identity-aware two-stage framework for the textual–visual matching. Identity-aware representation is learned in Stage 1, while in Stage 2 salient image regions and latent semantic concepts are matched for the following textual–visual affinity estimation. Following this work, Chen et al. [57] propose an efficient patch-word matching model in order to capture the local similarity between image and text. To exploit the multilevel corresponding visual contents, Jing et al. [58] propose a pose-guided multi-granularity attention network (PMA), which utilize pose information as soft attention to localize the discriminative regions. Due to the domain gap between data the extra model pre-trained on and data to be processed in the text-based person re-identification task, however, external cues generated by directly applying the pre-trained model without fine-tuning may suffer from great deviations. Unfortunately, as there is no annotation of body part in the dataset of text-based person re-identification, to fine-tune or re-train the proposed extra model seems impossible. Besides, introducing additional models is computationally costly as well. Besides, the joint embedding based methods directly compute the matching score for image-text pair in a shared latent space. For example, Zheng et al. [59] present a new system which can discriminatively embed the image and text to a shared visual-textual space. They also propose the instance loss, which explicitly considers the intra-modal data distribution. 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. These approaches are relatively computational efficient at th test stage, but they ignore the part representations which play a key role in text-based person retrieval. Besides TIMAM, there are also some other attempts at adversarial learning to reduce modality-gap. Liu et al. [10] design an A-GANet model to exploit semantic scene graphs, which generates fine-grained structured representations for multi-modal data. Recently, more and more works focus on extract fine-grained representations. Niu et al. [6] adopt a Multi-granularity Image-text Alignments (MIA) model exploit the combination of multiple granularities. 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. CMAAM is introduced by Aggarwal et al. [60] which learns an attribute-driven space along with a class-information driven space by introducing extra attribute annotation and prediction. Zheng et al. [61] 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. Wang et al. [62] propose a multi-granularity embedding learning (MGEL) method which generates multi-granularity embeddings of partial person bodies in a coarse-to-fine manner by re visiting the person image at different spatial scales. Zhu et al. [21] proposed a Deep Surroundings-person Separation Learning (DSSL) model to effectively extract and match person information. Besides, they construct a Real Scenarios Text-based Person Re-identification (RSTPReid) dataset to benefit future research on text-based person retrieval. A Semantically Self-Aligned Network (SSAN) is proposed by Ding et al. [63] to efficiently extracts semantically aligned visual and textual part features. Zhao et al. [64] introduce the weakly supervised person retrieval task and proposed a Cross-Modal Mutual Training (CMMT) framework. Besides, Shree et al. [65] consider the problem of searching people in an unconstrained environment with natural language descriptions and proposed an iterative question–answering (QA) strategy, which enables robots to request additional information from the users about the appearance of the target person.

It can be observed that most of the existing approaches extract and match the cross-modal information in a one-off manner, which may not properly take full advantage of the high-dimensional data to catch discriminative clues. Instead, our proposed Serialized Updating and Matching (SUM) method bridges the heterogeneity gap between cross-modal data in a progressive manner, which enables information obtained from multi-modal data to interact with each other step by step and get refined gradually.

2.4. Text-based person search in full images

Recently, Zhang et al. [66] put forward a new task which aims to search for the target person in full images via natural language descriptions. It is referred to as the Text-based Person Search in Full Images Task in order to distinguish from the conventional text-based person retrieval task. A novel end-to-end

learning framework is proposed to handle this new task, in which person detection, identification and image-text embedding tasks are jointly optimized together.

2.5. Video-related text-based person retrieval

During the recent years, some video-related text-based person retrieval tasks are put forward as well. Yamaguchi et al. manages to solve [67] the problem of spatio-temporal person retrieval from videos using a natural language query. By proposing a model including spatio-temporal human detection and multi-modal retrieval, a tube, namely, a sequence of bounding boxes which encloses the person described by the query can be output. Fan et al. [68] come up with the task of person tube retrieval via language description. Different from the videos-based spatio-temporal person retrieval task, which includes person detection and tracking, person tube retrieval is consistent with re-ID and only cares for the retrieval part. In addition to appearance, person tube also contains information about action and scene [69], so they propose a Multi-Scale Structure Preservation (MSSP) approach to solve this problem.

2.6. Visual Dialog

In addition, for some of the other vision–language tasks like Visual Dialog [70–72], the nature of these tasks requires the methods to answer a series of temporally ordered questions according to an image and a dialog history in a step-by-step manner. For instance, to effectively parse the temporal context and attend to the region of interest in an image given a question, Fan et al. [72] propose a recurrent attention network based on LSTM and reinforcement learning to observe an image several times to collect information for answering questions. Instead of answering a series of questions, text-based person retrieval can be deemed as answering one question that whether a pair of visual and textual data are matched with other or not. The model is supposed to excavate the multi-modal data for discriminative clues to answer this yes or no question. By means of the proposed Global Memory Gating Module (GMGM) and Fine-grained Memory Gating Module (FMGM) in our SUM method, the correlations lie within the multi-granular cross-modal data can be gradually checked from coarse to fine in a step-by-step manner, and hence a superior text-based person retrieval performance can be obtained.

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.

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 [73] 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 Sections 3.3.1 and 3.3.2, respectively. And then the mechanism of the Serialized Updating and Matching (SUM) method is described in Section 3.3.3.

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). $t \in \{1, 2, \ldots, T\}$ 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. The feature fusion paradigm can be implemented as several variants (e.g. averaging, concatenation and addition), which will be further discussed in Section 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

$$\alpha_{XG}^{i} = \frac{exp(cos(Y_{Fi}, X_G))}{\sum_{m=1}^{num} exp(cos(Y_{Fm}, X_G))}. \tag{2}$$

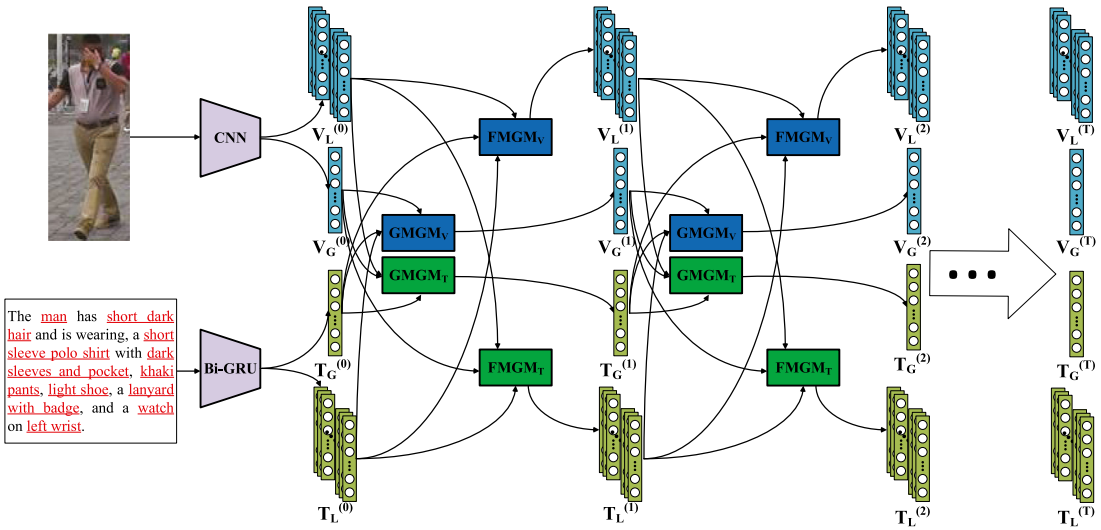


Fig. 1. Illustration of the proposed Serialized Updating and Matching (SUM) method, which is proposed to bridge the heterogeneity gap between cross-modal data in a progressive manner. The Memory Gating Modules (MGM) play the key role in SUM, which can be stacked to gradually update and match cross-modal 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.

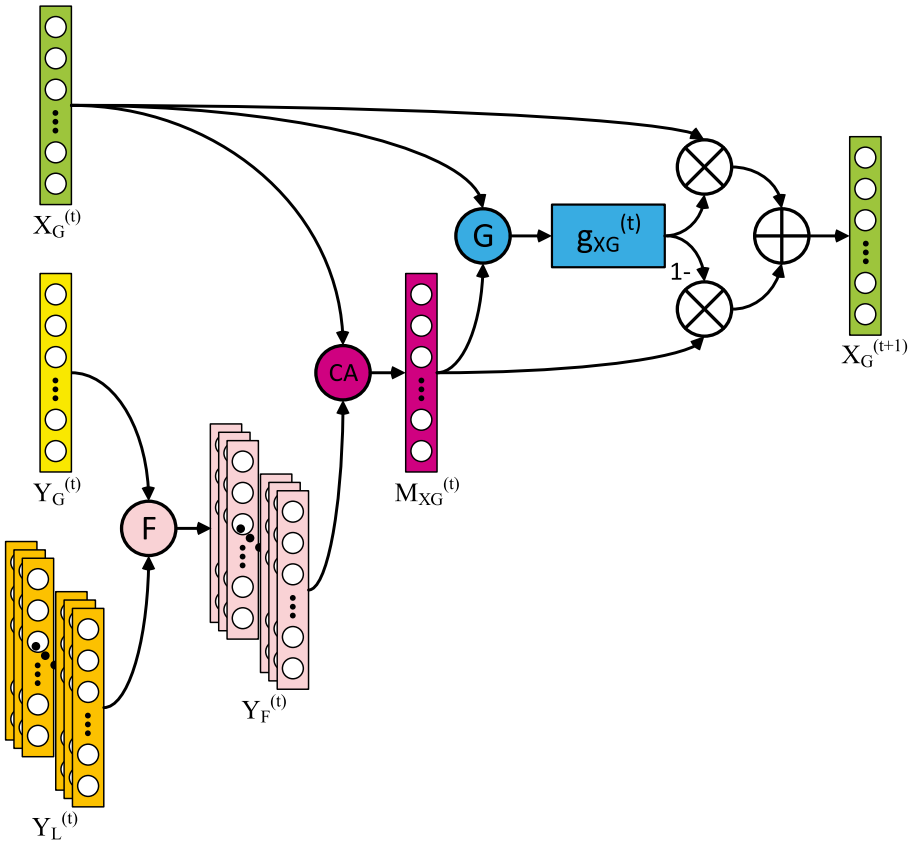


Fig. 2. Illustration of the Global Memory Gating Module (GMGM), which is proposed to update and match both the visual and textual global features.

$$M_{XG} = \mathcal{CA}(Y_F, X_G) = \sum_{\alpha^i_{XG} > \frac{1}{num}} \alpha^i_{XG} Y_{Fi}, \tag{3}$$

where *num* can be k or n for visual or textual data, respectively. α^i_{XG} represents the cross-modal relation between the ith fused feature Y_{Fi} and global feature X_G.

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

$$g^{(t)}_G = Gating(X^{(t-1)}_G, M^{(t-1)}_{XG}) = \sigma(\mathcal{F}_{gg}(X^{(t-1)}_G \oplus M^{(t-1)}_{XG})), \tag{4}$$

where $g^{(t)}_G$ is the global gating value at the tth 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

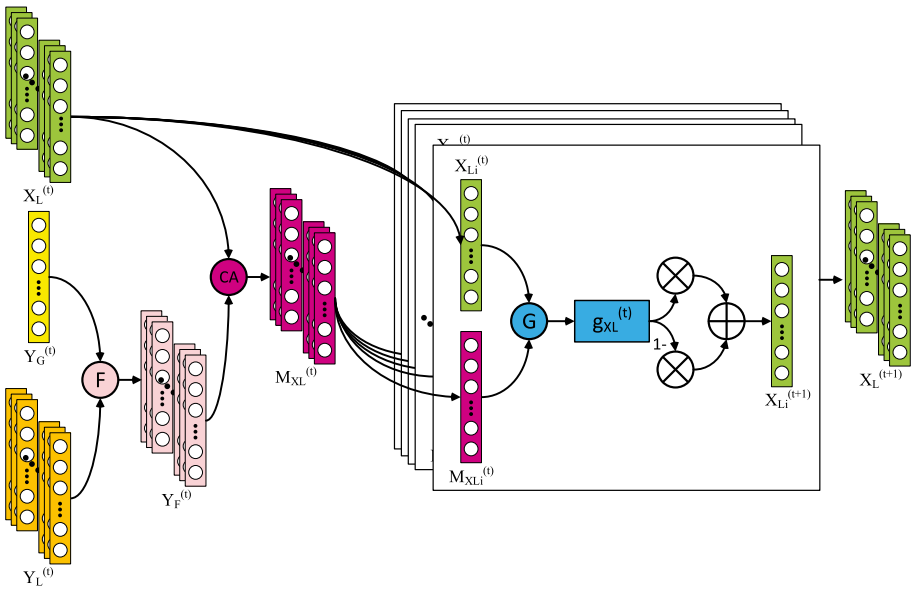


Fig. 3. Illustration of the Fine-grained Memory Gating Module (FMGM), which is proposed to update and match both the visual and textual local features.

implemented as several variants, which will be further discussed in Section 4.2.4.

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

$$X_G^{(t)} = g_G^{(t)} X_G^{(t-1)} + (1 - g_G^{(t)}) M_{XG}^{(t-1)}. \tag{5}$$

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

$$X_L^{(t)} = FMGM(X_L^{(t-1)}, Y_G^{(t-1)}, Y_L^{(t-1)}), \tag{6}$$

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). $t \in \{1, 2, \ldots, T\}$ 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. And 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}:

$$\alpha_{XL}^{ij} = \frac{exp(cos(Y_{Fi}, X_{Gj}))}{\sum_{m=1}^{num} exp(cos(Y_{Fm}, X_{Gj}))}. \tag{7}$$

$$M_{XLj} = \mathcal{CA}(Y_F, X_L^j) = \sum_{\alpha_{XL}^{ij} > \frac{1}{num}} \alpha_{XL}^{ij} Y_{Fi}, \tag{8}$$

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

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

$$g_{Li}^{(t)} = Gating(X_{Li}^{(t-1)}, M_{XLi}^{(t-1)}) = \sigma(\mathcal{F}_{lg}(X_{Li}^{(t-1)} \oplus M_{XLi}^{(t-1)})), \tag{9}$$

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

After that, each local feature is updated as

$$X_{Li}^{(t)} = g_{Li}^{(t)} X_{Li}^{(t-1)} + (1 - g_{Li}^{(t)}) M_{XLi}^{(t-1)}. \tag{10}$$

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:

$$\begin{aligned} V_G^{(t)} &= GMGM_V(V_G^{(t-1)}, T_G^{(t-1)}, T_L^{(t-1)}), \ T_G^{(t)} \\ &= GMGM_T(T_G^{(t-1)}, V_G^{(t-1)}, V_L^{(t-1)}), \end{aligned} \tag{11}$$

$$\begin{aligned} V_L^{(t)} &= FMGM_V(V_L^{(t-1)}, T_G^{(t-1)}, T_L^{(t-1)}), \ T_L^{(t)} \\ &= FMGM_T(T_L^{(t-1)}, V_G^{(t-1)}, V_L^{(t-1)}), \end{aligned} \tag{12}$$

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

$$\begin{aligned} S_{GG}^{(t)} &= Sim(V_G^{(t)}, T_G^{(t)}), \ S_{GL}^{(t)} = \sum_{i=1}^{n} Sim(V_G^{(t)}, T_{Li}^{(t)}), \ S_{LG}^{(t)} \\ &= \sum_{i=1}^{k} Sim(V_{Li}^{(t)}, T_G^{(t)}), \end{aligned} \tag{13}$$

where $Sim(\cdot, \cdot)$ denotes the Cosine similarity between two feature vectors. The overall cross-modal similarity at the tth time step is $S^{(t)} = S_{GG}^{(t)} + \lambda(S_{GL}^{(t)} + S_{LG}^{(t)})$ and hence the final similarity for cross-modal matching is $S = \sum_{t=0}^{T} S^{(t)}$.

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

$$L_{id}(X) = -log(softmax(W_{id} \times GN(X))) \tag{14}$$

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:

$$L_{ID}^{(0)} = L_{id}(V_G^{(0)}) + L_{id}(T_G^{(0)}). \tag{15}$$

And the entire loss in Stage-1 is

$$L_{Stage1} = L_{ID}^{(0)}. \tag{16}$$

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 [74], 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:

$$\begin{aligned} L_{ranking}(V,T) = &\sum_{\widehat{T}} max\{\alpha - cos(V,T) + cos(V,\widehat{T}), 0\} \\ &+ \sum_{\widehat{V}} max\{\alpha - cos(V,T) + cos(\widehat{V},T), 0\}, \end{aligned} \tag{17}$$

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:

$$L_{TR}^{(0)} = L_{ranking}(V_G^{(0)}, T_G^{(0)}) + L_{ranking}(V_L^{(0)}, T_G^{(0)}) + L_{ranking}(V_G^{(0)}, T_L^{(0)}). \tag{18}$$

The complete loss function in Stage-2 is

$$L_{Stage2} = L_{ID}^{(0)} + L_{TR}^{(0)}. \tag{19}$$

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:

$$L_{ID} = \sum_{t=0}^{T} L_{ID}^{(t)} = \sum_{t=0}^{T} (L_{id}(V_G^{(t)}) + L_{id}(T_G^{(t)})), \tag{20}$$

$$\begin{aligned} L_{TR} = &\sum_{t=0}^{T} L_{TR}^{(t)} = \sum_{t=0}^{T} (L_{ranking}(V_G^{(t)}, T_G^{(t)}) + L_{ranking}(V_L^{(t)}, T_G^{(t)}) \\ &+ L_{ranking}(V_G^{(t)}, T_L^{(t)})). \end{aligned} \tag{21}$$

Therefore, the complete loss for this stage is

$$L_{Stage3} = L_{ID} + L_{TR}. \tag{22}$$

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

Our approach is evaluated on two challenging Text-based Person Retrieval datasets : CUHK-PEDES [4] and RSTPReid [21]. (1) **CUHK-PEDES**: Following the official data split approach [4], the training set of CUHK-PEDES contains 34 054 images, 11 003 persons and 68 126 textual descriptions. The validation set contains 3078 images, 1000 persons and 6158 textual descriptions while the test set has 3074 images, 1000 persons and 6156 descriptions. (2) **RSTPReid**: The RSTPReid dataset [21] contains 20 505 images of 4101 persons. 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 test, respectively.

4.1.2. Evaluation metrics

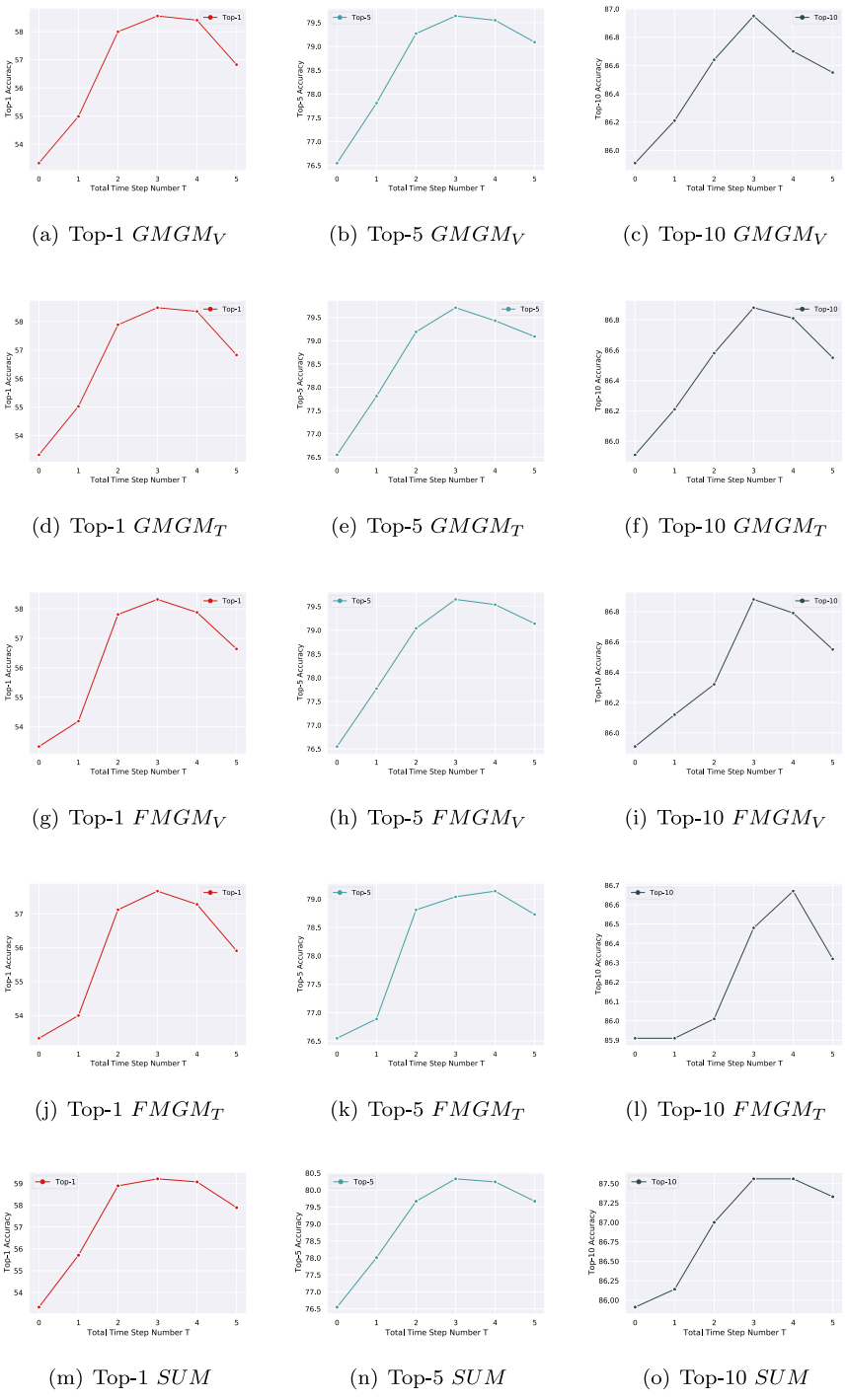
The performance is evaluated by the top-k accuracy and mean average precision (mAP) [?]. Given a query description, all test images are ranked by their similarities with this sentence. If any image of the corresponding person is contained in the top-k images, we call this a successful search. We report the top-1, top-5, and top-10 accuracies for all experiments.

4.1.3. Implementation details

In our experiments, we set the dimensionality $p = 1024$. The word number W is 4593 and 2128 after dropping the words that appears less than twice and the dimensionality E of embedded word vectors is set to 500. We choose the pre-trained ResNet-50 [73] as the visual CNN backbone. The input images are resized to $384 \times 128 \times 3$ and the random horizontal flipping strategy is employed for data augmentation. 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 [75]. An Adam optimizer [76] 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 $lr = 2 \times 10^{-4} \times (\frac{1}{10})^{epoch//10}$, 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 carried out. The top-1, top-5 and top-10 accuracies (%) are reported and the best result in each table is presented in bold. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.



(a) Top-1 $GMGM_V$ (b) Top-5 $GMGM_V$ (c) Top-10 $GMGM_V$

(d) Top-1 $GMGM_T$ (e) Top-5 $GMGM_T$ (f) Top-10 $GMGM_T$

(g) Top-1 $FMGM_V$ (h) Top-5 $FMGM_V$ (i) Top-10 $FMGM_V$

(j) Top-1 $FMGM_T$ (k) Top-5 $FMGM_T$ (l) Top-10 $FMGM_T$

(m) Top-1 SUM (n) Top-5 SUM (o) Top-10 SUM

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

4.2.1. Total number T of time steps

All of the ablation experiments are conducted with T increasing from 1 to 5 to check for the optimal number T of total time steps on the CUHK-PEDES and RSTPReid datasets. The experimental results are reported in Tables 1 and 9 for CUHK-PEDES and RSTPReid, respectively. Besides, 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. It can be observed from the results that initially the performance of SUM keeps improving with the increase of T, and then after reaching a peak ($T = 3$ for CUHK-PEDES while $T = 4$ for RSTPReid), the retrieval accuracy begins to drop slightly as T continues to go larger. To be specific, 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 $T = 5$, after a relatively small improvement, performance of the model began to decline slightly. This phenomenon is reasonable

Table 1
Ablation analysis of the Memory Gating Modules (MGM) and total time step number T on CUHK-PEDES. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results for each T are marked as bold type. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.

$GMGM_V$	$GMGM_T$	$FMGM_V$	$FMGM_T$	T	Top-1	Top-5	Top-10
×	×	×	×	0	53.30 ± 0.02	76.55 ± 0.04	85.93 ± 0.04
✓	×	×	×	1	54.99 ± 0.02	77.82 ± 0.01	86.20 ± 0.02
×	✓	×	×	1	55.06 ± 0.02	77.81 ± 0.02	86.23 ± 0.04
×	×	✓	×	1	54.17 ± 0.04	77.77 ± 0.03	86.12 ± 0.02
×	×	×	✓	1	54.02 ± 0.03	76.89 ± 0.02	85.87 ± 0.04
✓	✓	×	×	1	55.49 ± 0.02	77.93 ± 0.02	86.23 ± 0.02
×	×	✓	✓	1	55.46 ± 0.04	77.67 ± 0.02	86.03 ± 0.02
✓	✓	✓	×	1	55.69 ± 0.02	77.98 ± 0.02	**86.21 ± 0.05**
✓	✓	×	✓	1	55.57 ± 0.02	77.94 ± 0.03	86.19 ± 0.02
✓	✓	✓	✓	1	**55.72 ± 0.02**	**78.01 ± 0.03**	86.14 ± 0.02
✓	×	×	×	2	58.01 ± 0.02	79.27 ± 0.02	86.64 ± 0.02
×	✓	×	×	2	57.88 ± 0.02	79.19 ± 0.03	86.60 ± 0.03
×	×	✓	×	2	57.80 ± 0.01	79.04 ± 0.01	86.34 ± 0.04
×	×	×	✓	2	57.11 ± 0.04	78.81 ± 0.02	86.04 ± 0.04
✓	✓	×	×	2	58.13 ± 0.02	79.33 ± 0.01	86.67 ± 0.02
×	×	✓	✓	2	58.35 ± 0.01	79.12 ± 0.02	86.26 ± 0.04
✓	✓	✓	×	2	58.85 ± 0.04	79.63 ± 0.02	86.72 ± 0.02
✓	✓	×	✓	2	58.78 ± 0.02	79.43 ± 0.02	86.91 ± 0.03
✓	✓	✓	✓	2	**58.91 ± 0.02**	**79.66 ± 0.02**	**87.00 ± 0.03**
✓	×	×	×	3	58.53 ± 0.03	79.64 ± 0.02	86.95 ± 0.02
×	✓	×	×	3	58.49 ± 0.02	79.71 ± 0.02	86.89 ± 0.02
×	×	✓	×	3	58.32 ± 0.01	79.65 ± 0.02	86.88 ± 0.01
×	×	×	✓	3	57.66 ± 0.02	79.06 ± 0.04	86.45 ± 0.05
✓	✓	×	×	3	58.62 ± 0.03	79.80 ± 0.02	86.90 ± 0.02
×	×	✓	✓	3	58.82 ± 0.02	79.67 ± 0.02	86.89 ± 0.04
✓	✓	✓	×	3	59.17 ± 0.02	80.14 ± 0.02	87.41 ± 0.02
✓	✓	×	✓	3	59.14 ± 0.04	80.20 ± 0.02	87.36 ± 0.02
✓	✓	✓	✓	3	**59.22 ± 0.02**	**80.35 ± 0.02**	**87.60 ± 0.03**
✓	×	×	×	4	58.41 ± 0.02	79.55 ± 0.03	86.70 ± 0.02
×	✓	×	×	4	58.36 ± 0.02	79.43 ± 0.02	86.81 ± 0.02
×	×	✓	×	4	57.88 ± 0.02	79.55 ± 0.01	86.81 ± 0.04
×	×	×	✓	4	57.28 ± 0.02	79.15 ± 0.02	86.67 ± 0.02
✓	✓	×	×	4	58.44 ± 0.02	79.49 ± 0.02	86.88 ± 0.02
×	×	✓	✓	4	58.94 ± 0.01	79.52 ± 0.02	87.01 ± 0.02
✓	✓	✓	×	4	59.09 ± 0.04	80.16 ± 0.02	87.30 ± 0.03
✓	✓	×	✓	4	59.05 ± 0.04	79.81 ± 0.04	87.26 ± 0.02
✓	✓	✓	✓	4	**59.07 ± 0.02**	**80.26 ± 0.03**	**87.56 ± 0.02**
✓	×	×	×	5	56.85 ± 0.02	79.09 ± 0.02	86.55 ± 0.03
×	✓	×	×	5	56.81 ± 0.03	79.11 ± 0.02	86.52 ± 0.04
×	×	✓	×	5	56.64 ± 0.02	79.14 ± 0.02	86.54 ± 0.03
×	×	×	✓	5	55.95 ± 0.04	78.73 ± 0.02	86.32 ± 0.02
✓	✓	×	×	5	56.91 ± 0.03	79.10 ± 0.02	86.59 ± 0.02
×	×	✓	✓	5	57.52 ± 0.03	78.78 ± 0.04	86.67 ± 0.02
✓	✓	✓	×	5	57.81 ± 0.02	79.32 ± 0.02	87.24 ± 0.05
✓	✓	×	✓	5	57.77 ± 0.02	79.13 ± 0.02	87.19 ± 0.02
✓	✓	✓	✓	5	**57.92 ± 0.04**	**79.67 ± 0.02**	**87.31 ± 0.02**

as the multi-modal information can be over-smoothed after too much cross-modal interaction.

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 Table 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 Section 3.3.1, the feature fusion method can be implemented as several variants, which includes feature concatenation, feature addition and feature averaging. Ablation analysis are conducted to study the effectiveness of them on CUHK-PEDES and the experimental results are recorded in Table 2. As recorded in Table 2, with the change of T, the three employed feature fusion approaches achieve similar performance.

Table 2
Ablation analysis of the fusion methods for generating unified features on CUHK-PEDES. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results for each T are marked as bold type. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.

Method	T	Top-1	Top-5	Top-10
Feature Concatenation	1	55.61 ± 0.01	77.97 ± 0.02	86.08 ± 0.02
Feature Addition	1	55.68 ± 0.02	**78.08 ± 0.02**	86.13 ± 0.02
Feature Averaging	1	**55.72 ± 0.02**	78.01 ± 0.03	**86.14 ± 0.02**
Feature Concatenation	2	58.84 ± 0.02	79.69 ± 0.01	86.96 ± 0.02
Feature Addition	2	58.86 ± 0.02	**79.80 ± 0.04**	86.97 ± 0.03
Feature Averaging	2	**58.91 ± 0.02**	79.66 ± 0.02	**87.00 ± 0.03**
Feature Concatenation	3	59.19 ± 0.03	80.27 ± 0.03	87.61 ± 0.02
Feature Addition	3	59.16 ± 0.04	80.26 ± 0.02	**87.65 ± 0.02**
Feature Averaging	3	**59.22 ± 0.02**	**80.35 ± 0.02**	87.60 ± 0.03
Feature Concatenation	4	**59.10 ± 0.01**	80.23 ± 0.02	**87.58 ± 0.03**
Feature Addition	4	59.08 ± 0.04	80.15 ± 0.03	87.44 ± 0.02
Feature Averaging	4	59.07 ± 0.02	**80.26 ± 0.03**	87.56 ± 0.02
Feature Concatenation	5	57.86 ± 0.03	79.65 ± 0.02	87.33 ± 0.02
Feature Addition	5	57.91 ± 0.02	79.65 ± 0.04	**87.35 ± 0.03**
Feature Averaging	5	**57.92 ± 0.04**	**79.67 ± 0.02**	87.31 ± 0.02

Table 3
Ablation analysis of the fusion methods $\oplus$ in the Memory Gating Modules (MGM) on CUHK-PEDES. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results for each T are marked as bold type. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.

Method	T	Top-1	Top-5	Top-10
Feature Concatenation	1	55.67 ± 0.02	77.65 ± 0.01	85.77 ± 0.02
Feature Addition	1	55.62 ± 0.02	77.86 ± 0.04	86.14 ± 0.02
Feature Averaging	1	**55.72 ± 0.02**	**78.01 ± 0.03**	**86.14 ± 0.02**
Feature Concatenation	2	58.75 ± 0.02	79.51 ± 0.04	86.64 ± 0.04
Feature Addition	2	58.90 ± 0.03	79.61 ± 0.02	**87.02 ± 0.02**
Feature Averaging	2	**58.91 ± 0.02**	**79.66 ± 0.02**	87.00 ± 0.03
Feature Concatenation	3	59.20 ± 0.03	80.08 ± 0.02	87.22 ± 0.02
Feature Addition	3	59.14 ± 0.04	80.27 ± 0.02	87.57 ± 0.03
Feature Averaging	3	**59.22 ± 0.02**	**80.35 ± 0.02**	**87.60 ± 0.03**
Feature Concatenation	4	58.97 ± 0.01	80.01 ± 0.02	87.16 ± 0.02
Feature Addition	4	59.04 ± 0.02	**80.34 ± 0.02**	**87.59 ± 0.03**
Feature Averaging	4	**59.07 ± 0.02**	80.26 ± 0.03	87.56 ± 0.02
Feature Concatenation	5	57.58 ± 0.04	79.38 ± 0.02	87.19 ± 0.02
Feature Addition	5	57.76 ± 0.02	79.59 ± 0.01	**87.34 ± 0.02**
Feature Averaging	5	**57.92 ± 0.04**	**79.67 ± 0.02**	87.31 ± 0.02

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 can be formulated as:

$$X \oplus M_X = \begin{cases} [X, M_X], & \textit{Feature Concatenation}, \\ X + M_X, & \textit{Feature Addition}, \\ \frac{X+M_X}{2}, & \textit{Feature Averaging}, \end{cases} \tag{23}$$

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 ith local feature and updating message, respectively.

The results on the CUHK-PEDES dataset are reported in Table 3. As shown in the results, 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. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results for each T are marked as bold type. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.

Method	T	Top-1	Top-5	Top-10
DirectAddition	1	53.39 ± 0.02	76.68 ± 0.01	85.89 ± 0.02
Averaging	1	53.55 ± 0.02	76.65 ± 0.03	85.95 ± 0.02
DirectConcatenation	1	53.37 ± 0.02	76.71 ± 0.02	85.95 ± 0.03
FurtherFuse + ScalarGate	1	54.51 ± 0.01	77.49 ± 0.04	86.03 ± 0.02
ScalarGate	1	54.98 ± 0.03	77.63 ± 0.02	86.09 ± 0.02
FurtherFuse + VectorGate	1	55.68 ± 0.03	77.97 ± 0.03	86.06 ± 0.02
SUM(ours)	1	**55.72 ± 0.02**	**78.01 ± 0.03**	**86.14 ± 0.02**
DirectAddition	2	56.62 ± 0.02	78.38 ± 0.01	86.44 ± 0.01
Averaging	2	56.67 ± 0.02	78.27 ± 0.04	86.41 ± 0.03
DirectConcatenation	2	57.28 ± 0.03	78.41 ± 0.01	86.52 ± 0.02
FurtherFuse + ScalarGate	2	57.72 ± 0.02	79.15 ± 0.03	86.94 ± 0.04
ScalarGate	2	57.80 ± 0.02	79.28 ± 0.02	86.97 ± 0.03
FurtherFuse + VectorGate	2	58.87 ± 0.02	79.56 ± 0.04	**87.06 ± 0.04**
SUM(ours)	2	**58.91 ± 0.02**	**79.66 ± 0.02**	87.00 ± 0.03
DirectAddition	3	56.93 ± 0.03	79.04 ± 0.02	87.01 ± 0.02
Averaging	3	57.20 ± 0.02	79.03 ± 0.04	87.02 ± 0.02
DirectConcatenation	3	57.35 ± 0.02	79.02 ± 0.04	87.08 ± 0.03
FurtherFuse + ScalarGate	3	58.06 ± 0.04	79.89 ± 0.01	87.38 ± 0.01
ScalarGate	3	58.37 ± 0.02	80.23 ± 0.03	87.40 ± 0.02
FurtherFuse + VectorGate	3	58.72 ± 0.02	80.26 ± 0.05	87.43 ± 0.02
SUM(ours)	3	**59.22 ± 0.02**	**80.35 ± 0.02**	**87.60 ± 0.03**
DirectAddition	4	56.81 ± 0.02	78.90 ± 0.02	87.12 ± 0.02
Averaging	4	57.04 ± 0.03	78.94 ± 0.02	87.13 ± 0.02
DirectConcatenation	4	57.17 ± 0.02	78.94 ± 0.02	87.26 ± 0.04
FurtherFuse + ScalarGate	4	57.80 ± 0.03	79.70 ± 0.01	87.33 ± 0.02
ScalarGate	4	58.11 ± 0.02	79.76 ± 0.01	87.42 ± 0.04
FurtherFuse + VectorGate	4	58.58 ± 0.04	80.18 ± 0.03	87.33 ± 0.03
SUM(ours)	4	**59.07 ± 0.02**	**80.26 ± 0.03**	**87.56 ± 0.02**
DirectAddition	5	55.64 ± 0.02	78.12 ± 0.03	86.74 ± 0.02
Averaging	5	55.88 ± 0.02	78.39 ± 0.02	86.83 ± 0.02
DirectConcatenation	5	55.41 ± 0.01	78.34 ± 0.02	86.95 ± 0.02
FurtherFuse + ScalarGate	5	56.66 ± 0.04	79.18 ± 0.02	87.11 ± 0.02
ScalarGate	5	56.82 ± 0.02	79.34 ± 0.04	86.15 ± 0.02
FurtherFuse + VectorGate	5	57.49 ± 0.02	79.46 ± 0.02	87.19 ± 0.03
SUM(ours)	5	**57.92 ± 0.04**	**79.67 ± 0.02**	**87.31 ± 0.02**

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 can be formulated as:

$$X_\dagger^{(t)} = \begin{cases} X_\dagger^{(t-1)} + M_{X\dagger}^{(t-1)}, & \textit{DirectAddition}, \\ \frac{X_\dagger^{(t-1)} + M_{X\dagger}^{(t-1)}}{2}, & \textit{Averaging}, \\ [X_\dagger^{(t-1)}, M_{X\dagger}^{(t-1)}], & \textit{DirectConcatenation}, \\ s_\dagger^{(t)} X_\dagger^{(t-1)} + (1 - s_\dagger^{(t)}) M_{X\dagger}^{(t-1)}, & \textit{ScalarGate}, \\ s_\dagger^{(t)} f_\dagger^{(t-1)} + (1 - s_\dagger^{(t)}) M_{X\dagger}^{(t-1)}, & \textit{FurtherFuse + ScalarGate}, \\ g_\dagger^{(t)} f_\dagger^{(t-1)} + (1 - g_\dagger^{(t)}) M_{X\dagger}^{(t-1)}, & \textit{FurtherFuse + VectorGate}, \end{cases} \tag{24}$$

where $\dagger$ can be G or L_i which denotes the global or ith local feature. $s_\dagger$ denotes the global or local learned real-valued scalar gates, respectively. Before the feature updating step of *FurtherFuse+ScalarGate*, the input feature vectors are first further fused with the corresponding updating messages:

$$f_\dagger^{(t)} = \textit{FurtherFuse}(X_\dagger^{(t-1)}, M_{X\dagger}^{(t-1)}) = \sigma(\mathcal{F}_{ff}(X_\dagger^{(t-1)} \oplus M_{X\dagger}^{(t-1)})), \tag{25}$$

Table 5
Ablation analysis of the feature dimension p on CUHK-PEDES. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results are marked as bold type. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.

p	Top-1	Top-5	Top-10
256	58.42 ± 0.03	79.89 ± 0.02	87.19 ± 0.03
512	59.04 ± 0.02	80.19 ± 0.03	87.41 ± 0.02
1024	**59.22 ± 0.02**	80.35 ± 0.02	**87.60 ± 0.03**
2048	59.17 ± 0.02	**80.41 ± 0.02**	87.53 ± 0.02

Table 6
Comparison between cosine similarity and BLAB-SM on CUHK-PEDES. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results are marked as bold type. Note that we repeat every experiment for 10 times and the mean and std of the results are reported.

Method	Backbone	Top-1	Top-5	Top-10
Cosine similarity	VGG-16	**55.32 ± 0.02**	**76.29 ± 0.04**	**84.98 ± 0.03**
BLAB-SM	VGG-16	55.18 ± 0.04	76.23 ± 0.03	84.91 ± 0.03
Cosine similarity	ResNet-50	**59.22 ± 0.02**	**80.35 ± 0.02**	87.60 ± 0.03
BLAB-SM	ResNet-50	59.08 ± 0.03	80.32 ± 0.02	**87.62 ± 0.02**

where $f_{\dagger}^{(t)}$ is the global or the ith local gating value at the tth time step. $\mathcal{F}_{ff}(\cdot)$ is the linear transformation function, and $\sigma(\cdot)$ stands for the sigmoid function. Here, the feature fusion operation $\oplus$ is implemented as element-wise averaging.

Extensive experiments are carried out by employing them as substitution of MGM to investigate the effectiveness of our proposed method in depth. As can be observed from Table 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 $T = 1, 2, \ldots, 5$, 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 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.

4.2.6. Choice of the feature dimension p

Experiments are conducted to see the impact of the choice of the feature dimension p and the experimental results are reported in Table 5. It can be observed that by setting p as 1024 or 2048, SUM achieves better performance than 256 or 512, with $p = 1024$ slightly better.

4.2.7. Choice of feature distance calculation

For feature distance calculation, we further carry out experimental analysis on CUHK-PEDES with pretrained VGG-16 and ResNet-50 visual backbones to compare the performance between the conventional employed cosine similarity and BLAB-SM proposed by Abdalla and Amer [77]. The experimental results are reported in Table 6. As can be observed from the table, the two methods achieve similar performance on the task of text-based person retrieval, with cosine similarity slightly better.

Table 7
Statistics of the 4 utilized datasets. '#' stands for the number of the respective item.

Dataset	#IDs	#Images	#Captions	Category
CUHK-PEDES [4]	13003	40206	80440	Only pedestrian
RSTPReid [21]	4101	20505	41010	Only pedestrian
Flickr8k [22]	8091	8091	40455	Varied categories
Flickr30k [19]	31783	31783	158915	Varied categories

Table 8
Comparison with other state-of-the-art methods on CUHK-PEDES. Top-1, Top-5 and Top-10 accuracies along with the mean average precision (mAP) are reported. The highest the results are marked as bold type.

Method	Backbone	Top-1	Top-5	Top-10	mAP
CNN-RNN [78]	VGG-16	8.07	–	32.47	–
Neural Talk [79]	VGG-16	13.66	–	41.72	–
GNA-RNN [4]	VGG-16	19.05	–	53.64	–
IATV [5]	VGG-16	25.94	–	60.48	–
PWM-ATH [80]	VGG-16	27.14	49.45	61.02	–
Dual Path [59]	VGG-16	32.15	54.42	64.30	–
ITMeetsAL [81]	VGG-16	44.43	68.26	77.50	–
GALM [7]	VGG-16	47.82	69.83	78.31	–
MIA [6]	VGG-16	48.00	70.70	79.30	–
IMG-Net [9]	VGG-16	54.32	75.93	84.21	–
SUM (ours)	**VGG-16**	**55.32 ± 0.02**	**76.29 ± 0.02**	**84.98 ± 0.03**	**35.37 ± 0.02**
Dual Path [59]	ResNet-50	44.40	66.26	75.07	–
GLA [80]	ResNet-50	43.58	66.93	76.26	–
CMPC + CMPM [82]	ResNet-50	49.37	71.69	79.27	31.37
ITMeetsAL [81]	ResNet-50	50.63	73.33	81.34	–
MIA [6]	ResNet-50	53.10	75.00	82.90	–
A-GANet [10]	ResNet-50	53.14	74.03	81.95	–
GALM [7]	ResNet-50	54.12	75.45	82.97	–
TIMAM [8]	ResNet-101	54.51	77.56	84.78	35.13
IMG-Net [9]	ResNet-50	56.48	76.89	85.01	–
CMAAM [60]	ResNet-50	56.68	77.18	84.86	–
HGAN [61]	ResNet-50	59.00	79.49	86.62	37.80
SUM (ours)	**ResNet-50**	**59.22 ± 0.02**	**80.35 ± 0.02**	**87.60 ± 0.03**	**37.91 ± 0.01**

4.3. Comparison with other state-of-the-art methods

Table 8 shows the comparison of SUM against previous methods in terms of top-1, top-5 and top-10 accuracies in the text-based person retrieval task on CUHK-PEDES. In order to fairly compare with the previous approaches, we train and evaluate our proposed SUM method with both VGG-16 and ResNet-50 as the visual backbone. With VGG-16, we compare SUM with 10 existing approaches including CNN-RNN [78], Neural Talk [79], GNA-RNN [4], IATV [5], PWM-ATH [57], Dual Path [59], ITMeetsAL [81], GALM [7], MIA [6] and IMG-Net [9], while with ResNet-50, SUM is compared with 11 previous works including Dual Path [59], GLA [80], CMPC + CMPM [82], ITMeetsAL [81], MIA [6], A-GANet [10], GALM [7], TIMAM [8], IMG-Net [9], CMAAM [60] and HGAN [61]. All the results in the table are sorted according to the top-1 accuracy. As can be seen from Table 8, by means of the step-by-step paradigm for feature updating and matching, SUM respectively achieves 55.32%, 76.29%, 84.98% and 59.22%, 80.35%, 87.60% of top-1/5/10 accuracies with VGG-16 and ResNet-50 visual backbones, which outperforms the previous methods. GALM and MIA are typical approaches which extracts and aligns multi-granular features from cross-modal data in a one-off manner. Both with a VGG-16 backbone, our proposed SUM method surpasses GALM by 7.50%, 6.46% and 6.67% under the top-1/5/10 accuracy metrics, while outperforms MIA by 7.32%, 5.59% and 5.68t%, respectively. Moreover, simply with a VGG-16 backbone, SUM is also capable of surpassing GALM and MIA with a ResNet-50 backbone and even outperforming TIMAM which utilizing a ResNet-101 visual backbone in terms of top-1 and

Table 9
Comparison with other state-of-the-art methods on RSTPReid. Top-1, Top-5 and Top-10 accuracies are reported. The highest the results are marked as bold type.

Method	Top-1	Top-5	Top-10
DSSL [21]	39.05	62.60	73.95
SUM ($T = 0$, ours)	36.68 ± 0.02	62.88 ± 0.01	74.62 ± 0.02
SUM ($T = 1$, ours)	37.98 ± 0.03	64.15 ± 0.02	75.22 ± 0.02
SUM ($T = 2$, ours)	40.67 ± 0.02	66.84 ± 0.01	75.84 ± 0.02
SUM ($T = 3$, ours)	41.33 ± 0.02	**67.53 ± 0.03**	76.44 ± 0.02
SUM ($T = 4$, ours)	**41.38 ± 0.01**	67.48 ± 0.03	**76.48 ± 0.02**
SUM ($T = 5$, ours)	40.45 ± 0.02	67.11 ± 0.02	76.02 ± 0.03

Table 10
Comparison with other state-of-the-art methods on Flickr8k (text-to-image). Top-1, Top-5 and Top-10 accuracies are reported. The highest the results are marked as bold type.

Method	Top-1	Top-5	Top-10
Method	Text-to-image		
	Rank-1	Rank-5	Rank-10
Word2VisualVec [83]	33.4	63.1	75.3
ITMeetsAL [81]	40.1	67.8	79.2
Joint learning [84]	40.6	67.8	78.6
SUM (ours)	**45.4 ± 0.02**	**73.3 ± 0.03**	**83.2 ± 0.01**

Table 11
Comparison with other state-of-the-art methods on Flickr30k (text-to-image). Top-1, Top-5 and Top-10 accuracies are reported. The highest the results are marked as bold type.

Method	Top-1	Top-5	Top-10
Method	Text-to-image		
	Rank-1	Rank-5	Rank-10
RRF-Net [85]	35.4	68.3	79.9
CMPM-CMPC [86]	37.3	65.7	75.5
Dual Path [59]	39.1	69.2	80.9
DANs [87]	39.4	69.2	79.1
NAR [88]	39.4	68.8	79.9
A-GANet [10]	39.5	69.9	80.9
VSE++ [74]	39.6	70.1	79.5
SCO [89]	41.1	70.5	80.1
GXN [90]	41.5	–	80.1
TIMAM [8]	42.6	71.6	81.9
ITMeetsAL [81]	43.5	71.8	80.2
SUM (ours)	**46.9 ± 0.01**	**74.8 ± 0.02**	**83.4 ± 0.02**

top-10 accuracies. By adopting ResNet-50 as the visual backbone, SUM is able to tackle the task of text-based person retrieval better and achieve the state-of-the-art performance. All these observation demonstrates that compared with the one-off manner which matches cross-modal data at first glance, progressively updating and matching the cross-modal data in a step-by-step manner is more consistent with the nature of text-based person retrieval and is able to give a superior retrieval performance.

4.4. Generalization ability on general cross-modal retrieval task

As discussed in the Introduction, text-based person retrieval can be deemed as a specific sub-task of the general cross-modal retrieval task. Therefore, we conduct experimental analysis on two of the general cross-modal retrieval benchmarks Flickr8K [22] and Flickr30k [19] to evaluate the generalization ability of our proposed methods. Both Flickr8K and Flickr30K contain various categories of objects instead of only including pedestrians as CUHK-PEDES and RSTPReid. Following the standard data splitting method, there are 6091, 1000 and 1000 images in the training, validation and test sets, respectively, for the Flickr8K dataset. And for the Flickr30K dataset, there are 29,783, 1000 and 1000 images in the training, validation and test sets, respectively. For each images, there exist 5 natural language captions. The statistics of the 4 utilized datasets are displayed in Table 7. The input images are resized to $224 \times 224 \times 3$. We compare our proposed SUM method with 11 previous methods including RRF-Net [85], CMPM-CMPC [86], Dual Path [59], DANs [87], NAR [88], A-GANet [10], VSE++ [74], SCO [89], GXN [90], TIMAM [8] and ITMeetsAL [81] on Flickr30K and 3 previous methods including Word2VisualVec [83], ITMeetsAL [81] and Joint learning [84] on Flickr8K. The experimental results are reported in Tables 10 and 11 for Flickr8K and Flickr30K, respectively. And as the existing methods listed in Table 11 all utilize a pre-trained ResNet-152 as the visual backbone, we also use it in SUM for a fair comparison. It can be observed that our proposed SUM method outperforms all other methods at top-1/5/10 accuracies, which indicates the effectiveness and generalization ability of SUM.

4.5. Analysis of the retrieval results

Some of the examples of the top-5 text-based person retrieval results by our proposed SUM method (with $T = 0, 1, 2, \ldots, 5$) 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 Section 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 us 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 is more, in some of the cases, images of the targeted person can be hard to retrieve for certain reasons. Still

Fig. 5. Examples of the top-5 text-based person retrieval results by SUM with $T = 0, 1, 2, \ldots, 5$. $T = 0$ 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.

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. In addition, it can be observed that many of the mismatched person images contained in the top list are quite similar with the matched images in terms of color information, which indicates that the color information are somehow that more preferable for text-based person retrieval models than other types of discriminative clues like textual information, structural information, etc.. Therefore, enabling the proposed method to care for all kinds of clues beyond color information in an effective and balanced manner can be a key to further boost the research on the task of text-based person retrieval, which remains our future work.

5. Conclusion

The central problem of text-based person retrieval is how to properly bridge the gap between heterogeneous cross-modal data. Many of the previous works contrive to learn a latent common space to bridge the modality gap and extract modality-invariant feature vectors. Within these methods, the common space mapping and cross-modal information matching operations are conducted in a one-off manner, which aims to extract sufficient discriminative clues from the high-dimensional multi-modal data at first glance, which is just 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 step-by-step 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 two text-based person retrieval datasets CUHK-PEDES and RSTPReid along with two general cross-modal retrieval datasets Flickr8K and Flickr30K to see its generalization ability. Experimental results present that the proposed SUM outperforms existing methods and achieves the state-of-the-art performance. In addition, as it can be observed that many of the mismatched person images contained in the retrieved top list are quite similar with the matched images in terms of color information, which indicates that the color information are somehow that more preferable for text-based person retrieval models than other types of discriminative clues like textual information, structural information, etc.. Therefore, enabling the proposed method to care for all kinds of clues beyond color information in an effective and balanced manner can be a key to further boost the research on the task of text-based person retrieval, which remains our future work.

Broader impacts. Our work focuses on Text-based Person Retrieval by bridging the heterogeneity gap between cross-modal data in a step-by-step manner. The positive impact is that our work provides a novel view to update and align cross-modal data progressively, instead of determine whether a person image matches a query sentence at first glance. Besides, our work can be flexibly utilized as an add-on to any multi-granular text-based person retrieval methods to further improve the performance. Moreover, while we do not foresee our framework causing any direct negative societal impact, it may be utilized to build malicious applications for person search. It may be indirectly used to identify personal information and hence raising privacy concerns. Therefore, we strongly urge readers to limit the usage of the proposed method and ensure that to be strictly ethical and legal.

CRediT authorship contribution statement

Zijie Wang: Software, Writing – original draft. **Aichun Zhu:** Conceptualization, Methodology, Writing – review & editing. **Jingyi Xue:** Software, Validation. **Daihong Jiang:** Resources, Formal analysis. **Chao Liu:** Data curation, Visualization. **Yifeng Li:** Supervision, Resources.

Declaration of competing interest

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.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 62101245), China Postdoctoral Science Foundation (Grant No. 2019M661999), Natural Science Research of Jiangsu Higher Education Institutions of China (19KJB520009) and Future Network Scientific Research Fund Project (Grant No. FNSRFP-2021-YB-21).

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 Trans. Image Process. 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, 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, J. Electron. 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] 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. pp. 725–734.
[14] 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.
[15] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross-modal retrieval, Knowl.-Based Syst. 180 (2019) 38–50.
[16] H. Qiang, Y. Wan, Z. Liu, L. Xiang, X. Meng, Discriminative deep asymmetric supervised hashing for cross-modal retrieval, Knowl.-Based Syst. 204 (2020) 106188.
[17] X. Dong, H. Zhang, X. Dong, X. Lu, Iterative graph attention memory network for cross-modal retrieval, Knowl.-Based Syst. 226 (2021) 107138.
[18] 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, Knowl.-Based Syst. 217 (2021) 106818.
[19] 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.
[20] 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.
[21] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, G. Hua, Dssl: Deep surroundings-person separation learning for text-based person retrieval, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. pp. 209–217.
[22] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting image annotations using amazon's mechanical turk, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp. 139–147.
[23] 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.
[24] 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.
[25] 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.
[26] 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.

[27] 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.
[28] 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.
[29] 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.
[30] Y.-J. Cho, K.-J. Yoon, Pamm: Pose-aware multi-shot matching for improving person re-identification, IEEE Trans. Image Process. 27 (8) (2018) 3739–3752.
[31] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, Q. Tian, Deep representation learning with part loss for person re-identification, IEEE Trans. Image Process. 28 (6) (2019) 2860–2871.
[32] J. Dai, P. Zhang, D. Wang, H. Lu, H. Wang, Video person re-identification by temporal residual learning, IEEE Trans. Image Process. 28 (3) (2018) 1366–1377.
[33] Y. Yuan, J. Zhang, Q. Wang, Deep gabor convolution network for person re-identification, Neurocomputing 378 (2020) 387–398.
[34] J. Zhang, Y. Yuan, Q. Wang, Night person re-identification and a benchmark, IEEE Access 7 (2019) 95496–95504.
[35] 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 Trans. Intell. Transp. Syst..
[36] 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.
[37] 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, Knowl.-Based Syst. 212 (2021) 106505.
[38] Y. Fang, B. Li, X. Li, Y. Ren, Unsupervised cross-modal similarity via latent structure discrete hashing factorization, Knowl.-Based Syst. 218 (2021) 106857.
[39] F. Li, T. Wang, L. Zhu, Z. Zhang, X. Wang, Task-adaptive asymmetric deep cross-modal hashing, Knowl.-Based Syst. 219 (2021) 106851.
[40] Q. Zhang, Z. Lei, Z. Zhang, S.Z. Li, Context-aware attention network for image-text retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3536–3545.
[41] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.
[42] J. Bromley, J.W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, R. Shah, Signature verification using a siamese time delay neural network, Int. J. Pattern Recognit. Artif. Intell. 7 (04) (1993) 669–688.
[43] S. Bak, P. Carr, One-shot metric learning for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2990–2999.
[44] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737.
[45] X. Hao, S. Zhao, M. Ye, J. Shen, Cross-modality person re-identification via modality confusion and center aggregation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16403–16412.
[46] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: Clustering and fine-tuning, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 14 (4) (2018) 1–18.
[47] Y. Ding, H. Fan, M. Xu, Y. Yang, Adaptive exploration for unsupervised person re-identification, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 16 (1) (2020) 1–19.
[48] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2197–2206.
[49] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, T.S. Huang, Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6112–6121.
[50] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. pp. 2275–2284.
[51] S. Lin, H. Li, C.-T. Li, A.C. Kot, Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification, arXiv preprint arXiv:1807.01440.
[52] G.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, Z.-G. Hou, Cross-modality paired-images generation for rgb-infrared person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 12144–12151.
[53] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5380–5389.
[54] M. Ye, X. Lan, J. Li, P. Yuen, Hierarchical discriminative learning for visible thermal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[55] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[56] M. Ye, C. Liang, Z. Wang, Q. Leng, J. Chen, J. Liu, Specific person retrieval via incomplete text description, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 547–550.
[57] 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.
[58] 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.
[59] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 16 (2) (2020) 1–23.
[60] 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.
[61] 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.
[62] C. Wang, Z. Luo, Y. Lin, S. Li, Text-based person search via multi-granularity embedding learning, in: IJCAI, 2021.
[63] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, arXiv preprint arXiv:2107.12666.
[64] S. Zhao, C. Gao, Y. Shao, W.-S. Zheng, N. Sang, Weakly supervised text-based person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11395–11404.
[65] V. Shree, W.-L. Chao, M. Campbell, Interactive natural language-based person search, IEEE Robot. Autom. Lett. 5 (2) (2020) 1851–1858.
[66] S. Zhang, D. Long, Y. Gao, L. Gao, Q. Zhang, K. Niu, Y. Zhang, Text-based person search in full images via semantic-driven proposal generation, arXiv preprint arXiv:2109.12965.
[67] M. Yamaguchi, K. Saito, Y. Ushiku, T. Harada, Spatio-temporal person retrieval via natural language queries, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1453–1462.
[68] H. Fan, Y. Yang, Person tube retrieval via language description, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10754–10761.
[69] L. Zhu, H. Fan, Y. Luo, M. Xu, Y. Yang, Temporal cross-layer correlation mining for action recognition, IEEE Trans. Multimed..
[70] D. Kong, F. Wu, Visual dialog with multi-turn attentional memory network, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 611–621.
[71] M. Gu, Z. Zhao, W. Jin, D. Cai, F. Wu, Video dialog via multi-grained convolutional self-attention context multi-modal networks, IEEE Trans. Circuits Syst. Video Technol. 30 (12) (2019) 4453–4466.
[72] H. Fan, L. Zhu, Y. Yang, F. Wu, Recurrent attention network with reinforced generator for visual dialog, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 16 (3) (2020) 1–16.
[73] 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.
[74] 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.
[75] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[76] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, 2015.
[77] H.I. Abdalla, A.A. Amer, Boolean logic algebra driven similarity measure for text based applications, PeerJ Comput. Sci. 7 (2021) e641.

[78] 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.
[79] 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.
[80] 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.
[81] W. Chen, Y. Liu, E.M. Bakker, M.S. Lew, Integrating information theory and adversarial learning for cross-modal retrieval, Pattern Recognit. 117 (2021) 107983.
[82] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 686–701.
[83] J. Dong, X. Li, C.G. Snoek, Predicting visual features from text for image and video caption retrieval, IEEE Trans. Multimed. 20 (12) (2018) 3377–3388.
[84] S. Wang, D. Guo, X. Xu, L. Zhuo, M. Wang, Cross-modality retrieval by joint correlation learning, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 15 (2s) (2019) 1–16.
[85] 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.
[86] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 686–701.
[87] 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.
[88] C. Liu, Z. Mao, W. Zang, B. Wang, A neighbor-aware approach for image-text matching, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3970–3974.
[89] Y. Huang, Q. Wu, C. Song, L. Wang, Learning semantic concepts and order for image and sentence matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6163–6171.
[90] J. Gu, J. Cai, S.R. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189.