JCST Papers

Only for academic and non-commercial use

Thanks for reading!

[Survey](#)

[Computer Architecture and Systems](#)

[Artificial Intelligence and Pattern Recognition](#)

[Computer Graphics and Multimedia](#)

[Data Management and Data Mining](#)

[Software Systems](#)

[Computer Networks and Distributed Computing](#)

[Theory and Algorithms](#)

[Emerging Areas](#)

Multi-Task Visual Semantic Embedding Network for Image-Text Retrieval

Xue-Yang Qin[1] (秦雪洋), *Member, CCF*, Li-Shuang Li[1, *] (李丽双), Jing-Yao Tang[1] (唐婧尧), Fei Hao[2] (郝　飞) Mei-Ling Ge[3] (盖枚岭), and Guang-Yao Pang[4] (庞光垚)

[1] *School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China*

[2] *School of Computer Science, Shaanxi Normal University, Xi'an 710119, China*

[3] *School of Computer Engineering, Weifang University, Weifang 261061, China*

[4] *Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software, Wuzhou University, Wuzhou 543002 China*

E-mail: qinxueyang@snnu.edu.cn; lils@dlut.edu.cn; tangjingyao@mail.dlut.edu.cn; fhao@snnu.edu.cn; gemeiling@wfu.edu.cn pangguangyao@gxuwz.edu.cn

Abstract   Image-text retrieval aims to capture the semantic correspondence between images and texts, which serves as a foundation and crucial component in multi-modal recommendations, search systems, and online shopping. Existing mainstream methods primarily focus on modeling the association of image-text pairs while neglecting the advantageous impact of multi-task learning on image-text retrieval. To this end, a multi-task visual semantic embedding network (MVSEN) is proposed for image-text retrieval. Specifically, we design two auxiliary tasks, including text-text matching and multi-label classification, for semantic constraints to improve the generalization and robustness of visual semantic embedding from a training perspective. Besides, we present an intra- and inter-modality interaction scheme to learn discriminative visual and textual feature representations by facilitating information flow within and between modalities. Subsequently, we utilize multi-layer graph convolutional networks in a cascading manner to infer the correlation of image-text pairs. Experimental results show that MVSEN outperforms state-of-the-art methods on two publicly available datasets, Flickr30K and MSCO-CO, with *rSum* improvements of 8.2% and 3.0%, respectively.

Keywords   image-text retrieval, cross-modal retrieval, multi-task learning, graph convolutional network

1   Introduction

Vision and language are fundamental information patterns that help people understand the world. The interplay between what we perceive visually and how we communicate through language forms the cornerstone of our experience. Images and texts, as the most direct reflections of vision and language, serve as potent conduits through which we exchange ideas, express emotions, and weave narratives. Exploring the relationship between them has become a hotspot in the field of multimodality and spawned some specific applications, such as image-text retrieval[1, 2], multi-modal recommendation[3, 4], and visual commonsense reasoning[5, 6]. This paper focuses on image-text retrieval that aims to bridge the semantic gap between images and textual descriptions. Despite significant efforts in recent years, there remains a challenge in measuring the relevance of images and texts due to the heterogeneity and distributional differences in the data of these two modalities.

To cope with the challenge, early approaches for image-text retrieval project the entire visual and textual information into a shared subspace, where the correlations between images and texts are easily measured. For instance, Wang *et al.*[7] presented a two-

---

branch embedding network to model images and texts, respectively, while employing a similarity network with element-wise product operation followed by a fully-connected layer to compute the correlations of image-text pairs. Similar methods, such as [8–13], also design different networks to acquire global semantic features of images and texts to mine visual-linguistic associations. Although such coarse-grained methods are impressive, they overlook the details of image-text alignment, resulting in poor performance. Intuitively, when observing an image, people tend to pay more attention to salient regions and less attention to non-salient regions. Considering this, some work begins to explore fine-grained alignments between image regions and text words to discover the connection between these two modalities. A common practice is to use pre-trained object detection tools to identify objects in an image and extract region-specific visual features using convolutional neural networks, while adopting recurrent neural networks to obtain word-level features. After that, the similarity scores between images and texts are inferred using paired region-word similarity matrices. SCAN[14] is a typical representative of such methods, which introduces a cross-modal attention network to explore the fine-grained relationship between images and texts. Subsequently, many attention-based methods[15–21] are proposed to realize fine-grained alignment. Compared with coarse-grained methods, fine-grained methods show great potential for image-text retrieval.

However, the fine-grained methods are more about improving performance from a model design perspective and ignore the view of training optimization. Therefore, we introduce two auxiliary tasks to enhance cross-modal retrieval performance from the model optimization perspective. Theoretically, the optimization of single-task learning moves toward loss reduction during training. If a model is trapped in a local minimum, it is difficult to be optimized further, which is not conducive to finding the global optimum. Distinctively, multi-task learning jointly trains multiple differentiated tasks. Due to the differences between tasks, the optimization directions between various tasks may be different. When the target task falls into a local optimum, it may jump out of the local optimum under the action of the other tasks, which provides a possibility for finding the global optimum. Furthermore, some sub-networks or parameters are shared in multi-task learning, which helps to learn general feature representations and improve the mod-

el's robustness and generalization ability.

Based on the above analyses and discussions, a multi-task visual semantic embedding network (MVSEN) is proposed to explore cross-modal image-text retrieval from both the training optimization and model design perspectives. Firstly, we design two semantically constrained auxiliary tasks, including text-text matching and multi-label classification, to train jointly with the target task image-text retrieval, aiming to improve model performance from the training optimization perspective. Secondly, we present an intra- and inter-modality interaction mechanism from the model design perspective to obtain discriminative transformed visual and textual features. Subsequently, a similarity vector function is utilized to acquire similarity matrix vectors that will be used to infer the correlations between images and texts through graph convolutional networks with residual connection followed by two fully-connected layers. Finally, we employ a weighted approach to combine the optimization objectives of these three tasks as the optimization loss of the proposed MVSEN. In brief, our contributions can be summarized as threefold.

● We present a multi-task visual semantic embedding network (MVSEN) that leverages multi-task collaborative training alongside intra- and inter-modality interaction strategies to enhance the discrimination of visual and textual features, thereby improving the performance of cross-modal image-text retrieval.

● We introduce two auxiliary tasks into image-text retrieval, which is beneficial for improving the robustness and generalization of visual-semantic embedding. To the best of our knowledge, we are the first to introduce the task of text-text matching as a semantic constraint into cross-modal retrieval.

● We conduct extensive experiments on two benchmarks, Flickr30K and MSCOCO, and the experimental results show that the proposed MVSEN achieves advanced performance compared with the state-of-the-art methods, with improvements of 8.2% and 3.0% on evaluation metric *rSum*, respectively.

2 Related Work

2.1 Image-Text Retrieval

According to the way of image-text modeling, existing methods can be roughly divided into two categories: global-based methods[8, 10–12, 22] and local-based methods[15, 18–20, 23, 24]. The global-based methods aim to project heterogeneous multimodal data into a joint

embedding space, generating corresponding global feature representations for each modality and then measuring their correlation using a distance metric function. Liu *et al.*[8] utilized ResNet152 and RNN to encode visual and textual information, and learned consistent visual-semantic embedding through deep mapping and reconstructed mapping. Sarafianos *et al.*[10] developed an adversarial network to learn discriminative feature representations in these two modalities jointly. Chen *et al.*[22] proposed an effective generalized pooling strategy to learn optimized visual-semantic features. Given that these methods solely focus on the global semantic information of image-text pairs, they may not effectively handle complex scenes. Compared with global-based methods, local-based methods pay more attention to the details of image-text alignment. Lee *et al.*[14] presented a stacked attention framework to identify the potential alignment between visual regions and textual words. Considering that different fragment-level features contribute differently to inferring the correspondence between vision and language, Liu *et al.*[15] proposed a bidirectional focal attention approach to focus on relevant regions and words and eliminate irrelevant ones. Wu *et al.*[20] adopted the method of reassigning region-word attention weights to alleviate the impact of unimportant fragment-level features on model performance. Additionally, some methods consider global- and local-based strategies to explore the correlation of image-text pairs from different perspectives. For example, Diao *et al.*[25] exploited graph convolutional networks to perform similarity inference on fused global-based alignment and local-based alignment. Wang *et al.*[26] designed a global-local alignment strategy that considers both global semantic and local segment information. Analogously, similar approaches such as [27–29] also optimize the whole network in a joint manner.

2.2 Multi-Task Learning

Different from single-task learning, multi-task learning learns a shared feature representation by jointly modeling multiple associated single tasks, where each task can act as a semantic constraint for other tasks, which is beneficial to improve the performance of models. In recent years, multi-task learning has been applied to various computer vision and natural language processing tasks, such as dense prediction[30], emotion recognition[31], biomedical relation ex-

traction[32], and so on. Vandenhende *et al.*[33] proposed a multi-scale task interaction framework to determine the information interaction between different tasks in multi-task learning through distillation units. Xu *et al.*[30] presented a shared encoder-decoder strategy to model multi-task learning and capture task-specific features via cross-task attention mechanisms. Similarly, Moscato *et al.*[32] designed shared encoder layers, including a lexicon encoder and a transformer encoder, and task-specific layers to realize biomedical relation extraction. Foggia *et al.*[31] developed a convolution-based shared encoding layer and a task-specific layer while employing an independent classification layer to make predictions for different tasks. Similarly, there are also some work[9, 34–37] that adopts multi-task learning in the field of cross-modal image-text retrieval, but the difference is that cross-modal retrieval is the target task, while the other tasks are auxiliary tasks. For example, Luo *et al.*[34] presented correlation recognition and context reconstruction tasks combined with two regularization terms to improve the performance of cross-modal image-text retrieval. Xu *et al.*[36] integrated image and text multi-classification tasks into cross-modal retrieval to constrain the global semantic features. Li *et al.*[9, 37] exploited the image captioning task to model the ground-truth caption and the generated caption association through the log-likelihood function. Analogously, Yuan *et al.*[35] adopted the joint training of image captioning and image-text retrieval to improve the performance of cross-modal retrieval.

2.3 Differences with Existing Methods

The proposed MVSEN utilizes the idea of multi-task learning which has been widely used to enhance the performance of image-text retrieval. However, it differs from previous methods as follows. First, compared with existing multi-task learning approaches, such as [9, 34–37], we introduce a new auxiliary task, text-text matching, into cross-modal retrieval. Theoretically, if an image strongly correlates with a specific text, other texts highly associated with that text are likely to correlate with the image. Considering this, we implement an auxiliary text-text matching task to constrain the global semantic information to improve the performance of cross-modal image-text retrieval. To the best of our knowledge, we are the first to introduce the text-text matching task into image-text retrieval. Second, although CASC[36] also em-

ploys a multi-label classification task, it treats images and texts as separate multi-label classifications. In contrast, our approach performs multi-label classification on the fused visual and textual information. Furthermore, we consider the interaction within each modality to preserve shared information between images and texts before the multi-label classification, while CASC ignores this. Based on the above analyses, it can be observed that MVSEN is different from existing cross-modal retrieval approaches.

3 Methodology

In this section, we will present the proposed Multi-Task Visual Semantic Embedding Network (MVSEN) in detail. As illustrated in Fig.1, the overall framework of MVSEN mainly consists of two components: feature representation for encoding shared visual and textual features, and multi-task learning for jointly learning and optimizing various tasks. Specifically, we introduce the feature representation module in Subsection 3.1, explain the multi-task learning in Subsection 3.2, and introduce the optimization objective of MVSEN in Subsection 3.3.

3.1 Feature Representation

Visual Feature Representation. Given an image I, we employ $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_k)$ to represent the features of image I, where $\boldsymbol{v}_i \in \mathbb{R}^{1 \times d}$ denotes the i-th

regional feature and k indicates the number of salient regions. Concretely, a well-known object detection tool Faster-RCNN pre-trained by Anderson *et al.*[38] on Visual Genome, is used to detect visual regions. Then, we select top-k ($k = 36$) salient regions with the highest confidence scores and utilize the pre-trained ResNet[39] to extract the feature of each region. This process can be formalized as

$$\boldsymbol{v} = f_{\text{rn}}(f_{\text{fr}}(I, \boldsymbol{\theta}_{\text{fr}}), \boldsymbol{\theta}_{\text{rn}}), \tag{1}$$

where $f_{\text{rn}}(\cdot)$ and $f_{\text{fr}}(\cdot)$ denote the Faster-RCNN and ResNet functions, respectively, and $\boldsymbol{\theta}_{\text{rn}}$ and $\boldsymbol{\theta}_{\text{fr}}$ are the corresponding learnable parameters. In the experiment, Faster-RCNN and ResNet are only used to extract visual features and do not participate in the joint training of the entire network.

Textual Feature Representation. Given a text T containing n words, following previous approaches[18–21], we utilize a bidirectional gated recurrent unit (Bi-GRU) to model the sequential information of the text. Bi-GRU consists of a forward GRU to capture semantic features in the forward direction and a backward GRU to capture semantic features in the reverse direction. Therefore, Bi-GRU helps capture the contextual relationships between words, facilitating the flow of information in both forward and backward directions, and thus enriching the feature representation of text to enhance the performance of cross-modal image-text retrieval. For the given text $T = \{w_1, w_2, \ldots, w_j, \ldots, w_n\}$, the encoding
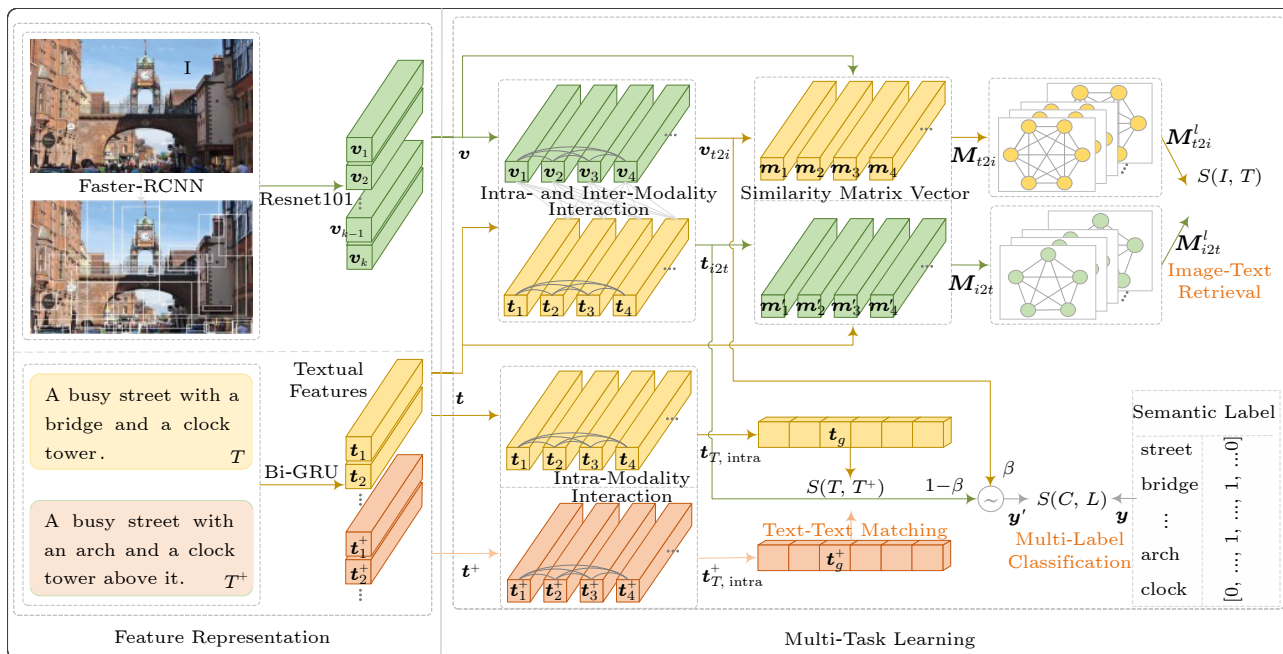

Fig.1. Overall framework of the proposed MVSEN.

using Bi-GRU can be expressed as

$$\overrightarrow{\boldsymbol{h}}_j = f_{\overrightarrow{\mathrm{g}}}(\overrightarrow{\boldsymbol{h}}_{j+1}, w_j, \boldsymbol{\theta}_{\overrightarrow{\mathrm{g}}}),$$

$$\overleftarrow{\boldsymbol{h}}_j = f_{\overleftarrow{\mathrm{g}}}(\overleftarrow{\boldsymbol{h}}_{j+1}, w_j, \boldsymbol{\theta}_{\overleftarrow{\mathrm{g}}}), \quad (2)$$

where $\overrightarrow{\boldsymbol{h}}_j$ and $\overleftarrow{\boldsymbol{h}}_j$ indicate the j-th hidden state representations from the forward $f_{\overrightarrow{\mathrm{g}}}(\cdot)$ and backward $f_{\overleftarrow{\mathrm{g}}}(\cdot)$, respectively. $\boldsymbol{\theta}_{\overrightarrow{\mathrm{g}}}$ and $\boldsymbol{\theta}_{\overleftarrow{\mathrm{g}}}$ are the learnable parameters. To better exploit the encoding information of Bi-GRU, we adopt the average of $\overrightarrow{\boldsymbol{h}}_j$ and $\overleftarrow{\boldsymbol{h}}_j$ as the feature representation $\boldsymbol{t}_j$ of the word w_j, i.e., $\boldsymbol{t}_j = (\overrightarrow{\boldsymbol{h}}_j + \overleftarrow{\boldsymbol{h}}_j)/2 \in \mathbb{R}^{1 \times d}$. Subsequently, the features of text T can be written as $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_j, \ldots, \boldsymbol{t}_n)$. Similarly, the features of the positive text sample T^+ of text T can be denoted as $\boldsymbol{t}^+ = (\boldsymbol{t}_1^+, \boldsymbol{t}_2^+, \ldots, \boldsymbol{t}_j^+, \ldots, \boldsymbol{t}_m^+)$. Furthermore, the method to obtain T^+ is as follows: since each image in the datasets we use is associated with five matching texts, we consider these five texts to be positive examples of each other. Specifically, we first assign numbers from 1 to 5 to these texts, respectively. Then the positive text example T^+ for the i-th (where $i \in \{1, 2, 3, 4\}$) text is set as the $(i+1)$-th text, and the positive text example T^+ for the last text is the text numbered 1.

3.2 Multi-Task Learning

In this subsection, we will illustrate the proposed multi-task learning module, including image-text retrieval, text-text matching, and multi-label classification tasks. We will first discuss the intra- and inter-modality interaction module since it is adopted in image-text retrieval and multi-label classification.

3.2.1 Intra-Modality Interaction

Intuitively, when describing an image, people will focus on some salient regions according to their preferences and enrich the image with words that have no real meaning. As shown in Fig.1, some contents such as "buildings" and "people" in the image cannot find the corresponding semantics in the sentence "A busy street with a bridge and a clock tower". Also, the words "a" and "with" in the sentence have no real meaning. Obviously, these will affect the performance of image text retrieval. Therefore, we adopt the self-attention mechanism to strengthen meaningful words and weaken meaningless words. At the same time, we integrate self-learned "external information" into the self-attention mechanism to alleviate the inconsisten-

cy between visual and textual information.

Specifically, taking textual feature $\boldsymbol{t}$ as an example, we first map the textual feature $\boldsymbol{t}$ into the query matrix $\boldsymbol{Q} = \boldsymbol{t}\boldsymbol{W}_{\mathrm{q}} \in \mathbb{R}^{n \times d_{\mathrm{v}}}$, key matrix $\boldsymbol{K} = \boldsymbol{t}\boldsymbol{W}_{\mathrm{k}} \in \mathbb{R}^{n \times d_{\mathrm{v}}}$, and value matrix $\boldsymbol{V} = \boldsymbol{t}\boldsymbol{W}_{\mathrm{v}} \in \mathbb{R}^{n \times d_{\mathrm{v}}}$, where $\boldsymbol{W}_{\mathrm{q}}$, $\boldsymbol{W}_{\mathrm{k}}$, and $\boldsymbol{W}_{\mathrm{v}}$ are learnable parameters. Subsequently, we employ Gaussian distribution to initialize two matrices $\boldsymbol{K}_{\mathrm{m}} \in \mathbb{R}^{l \times d_{\mathrm{v}}}$ and $\boldsymbol{V}_{\mathrm{m}} \in \mathbb{R}^{l \times d_{\mathrm{v}}}$ as self-learned "external information" to extend $\boldsymbol{K}$ and $\boldsymbol{V}$. A weighted output of text feature $\boldsymbol{t}$ can be formalized as

$$\boldsymbol{h}_i = f_{\mathrm{sm}}\left(\frac{\boldsymbol{Q}f_{\mathrm{cc}}(\boldsymbol{K}, \boldsymbol{K}_{\mathrm{m}})^{\mathrm{T}}}{\sqrt{d}}\right) f_{\mathrm{cc}}(\boldsymbol{V}, \boldsymbol{V}_{\mathrm{m}}), \quad (3)$$

where $f_{\mathrm{sm}}(\cdot)$ denotes the "softmax" function. $f_{\mathrm{cc}}(\cdot)$ means a function that performs a concatenation operation on the first dimension, such as $f_{\mathrm{cc}}(\boldsymbol{K}, \boldsymbol{K}_{\mathrm{m}}) \in \mathbb{R}^{(n+l) \times d_{\mathrm{v}}}$. Finally, we adopt H ($d = H \times d_{\mathrm{v}}$) substructures in (3) to ensure the diversity and robustness of the encoded information, i.e., $\boldsymbol{t}^{'} = f_{\mathrm{cc}}(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_i, \ldots, \boldsymbol{h}_H) \in \mathbb{R}^{n \times d}$. For simplicity, we utilize the function $f_{\mathrm{intra}}^T(\cdot)$ to represent the intra-modality interaction within the text T, and the entire above process can be rewritten as

$$\boldsymbol{t}^{'} = f_{\mathrm{intra}}^T(\boldsymbol{t}, \boldsymbol{\theta}_{\mathrm{intra}}^T), \quad (4)$$

where $\boldsymbol{\theta}_{\mathrm{intra}}^T$ is a learnable parameter. Analogously, we can obtain visual feature $\boldsymbol{v}^{'} = f_{\mathrm{intra}}^I(\boldsymbol{v}, \boldsymbol{\theta}_{\mathrm{intra}}^I) \in \mathbb{R}^{k \times d}$ after performing the intra-modality interaction.

3.2.2 Inter-Modality Interaction

Unlike the intra-modality interaction, the inter-modality interaction primarily focuses on the information correlation between different modalities, which is crucial for enhancing the performance of cross-modal retrieval. As we can see from Fig.1, some visual regions in the image can be matched with corresponding textual fragments and vice versa. To quantify this, we adopt the dot product followed by the "softmax" function to calculate the correlation scores between different visual regions and textual words. The process can be formalized as

$$\boldsymbol{a}_{t2i} = f_{\mathrm{sm}}(\alpha f_{\mathrm{lr}}(\boldsymbol{v}^{'}(\boldsymbol{t}^{'})^{\mathrm{T}})), \quad (5)$$

where $\boldsymbol{a}_{t2i} \in \mathbb{R}^{k \times n}$ denotes the correlation score matrix between visual regions and textual words. α is a control factor used to regulate the correlation scores. $f_{\mathrm{lr}}(\cdot)$ indicates the "LeakyRelu" activation function.

After that, we can consider this question: can textual information be converted into visual information?

Actually, the j-th column of the score matrix $\boldsymbol{a}_{t2i}$ can be interpreted as the importance of word w_j to each visual region. Therefore, an approximate approach is to obtain the transformed features through matrix multiplication. Formally, the visual feature $\boldsymbol{v}_{t2i}$ transformed from textual feature $\boldsymbol{t}'$ can be denoted as $\boldsymbol{v}_{t2i} = \boldsymbol{a}_{t2i}\boldsymbol{t}' \in \mathbb{R}^{k \times d}$. For convenience, we employ the function $\boldsymbol{f}_{\mathrm{inter}}^{t2i}$ to represent the inter-modality interaction from text to image direction, and the above process can be formalized as

$$\boldsymbol{v}_{t2i} = f_{\mathrm{inter}}^{t2i}(\boldsymbol{v}', \boldsymbol{t}', \alpha). \tag{6}$$

Similarly, we can acquire the textual feature $\boldsymbol{t}_{i2t}$ transformed from visual feature $\boldsymbol{v}'$, i.e., $\boldsymbol{t}_{i2t} = f_{\mathrm{inter}}^{i2t}(\boldsymbol{t}', \boldsymbol{v}', \alpha) \in \mathbb{R}^{n \times d}$. Theoretically, the transformed features retain some common information in both modalities, preventing the model from focusing on non-important information and helping improve the model's performance.

3.2.3 Image-Text Retrieval

As discussed earlier, the transformed features are beneficial to maintaining common information between these two modalities. Considering this, we utilize the transformed and original features to explore the correlations between images and texts. Specifically, taking the features $\boldsymbol{v}_{t2i}$ and $\boldsymbol{v}$ as an example, we first calculate the similarity matrix vector $\boldsymbol{M}_{t2i} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_k) \in \mathbb{R}^{k \times d_c}$ between $\boldsymbol{v}_{t2i}$ and $\boldsymbol{v}$ by

$$\boldsymbol{M}_{t2i} = \frac{|\boldsymbol{v}_{t2i} - \boldsymbol{v}|^2 \boldsymbol{w}}{|| |\boldsymbol{v}_{t2i} - \boldsymbol{v}|^2 \boldsymbol{w}||_2},$$

where $|\cdot|^2$ denotes element-wise square. $\boldsymbol{w} \in \mathbb{R}^{d \times d_c}$ is a learnable parameter matrix, and $||\cdot||_2$ is L_2-norm.

Subsequently, we treat $\boldsymbol{M}_{t2i}$ as a fully connected graph G_{t2i} containing k nodes, where each row in $\boldsymbol{M}_{t2i}$ contributes to a node. To enhance the representation of these nodes, we utilize stacked graph convolutional networks (GCNs) followed by residual connections to establish the association between them. We implement the update of node features of the l-th ($l \geqslant 0$) layer by

$$\boldsymbol{M}_{t2i}^{l+1} = f_{\mathrm{gcn}}(\boldsymbol{M}_{t2i}^l, \boldsymbol{w}_{t2i}^l, \boldsymbol{\theta}_{t2i}^l) + \boldsymbol{M}_{t2i}^l,$$

$$\boldsymbol{w}_{t2i}^l = (\boldsymbol{M}_{t2i}^l \cdot \boldsymbol{w}_1^l + \boldsymbol{b}_1^l)(\boldsymbol{M}_{t2i}^l \cdot \boldsymbol{w}_2^l + \boldsymbol{b}_2^l)^{\mathrm{T}},$$

$$\boldsymbol{M}_{t2i}^0 = \boldsymbol{M}_{t2i},$$

where $\boldsymbol{M}_{t2i}^l \in \mathbb{R}^{k \times d_c}$ denotes the node features of the l-th layer, and $\boldsymbol{w}_{t2i}^l \in \mathbb{R}^{k \times k}$ is the corresponding edge weight matrix. $\boldsymbol{w}_1^l \in \mathbb{R}^{d_c \times d_c}$, $\boldsymbol{w}_2^l \in \mathbb{R}^{d_c \times d_c}$, $\boldsymbol{b}_1^l \in \mathbb{R}^{d_c \times 1}$, and $\boldsymbol{b}_2^l \in \mathbb{R}^{d_c \times 1}$ are learnable parameters. Then, two cascaded fully connected layers are employed to infer semantic relevance score $\boldsymbol{S}_{t2i}$ of image-text pairs.

$$\boldsymbol{S}_{t2i} = f_{\mathrm{m}}((f_{\mathrm{th}}(\boldsymbol{M}_{t2i}^l \cdot \boldsymbol{w}_1 + \boldsymbol{b}_1)) \cdot \boldsymbol{w}_2 + \boldsymbol{b}_2),$$

where $f_{\mathrm{m}}(\boldsymbol{x})$ means averaging all elements in the matrix $\boldsymbol{x}$. $f_{\mathrm{th}}(\cdot)$ indicates the "tanh" activation function. $\boldsymbol{w}_1 \in \mathbb{R}^{d_c \times d_v}$, $\boldsymbol{w}_2 \in \mathbb{R}^{d_v \times 1}$, $\boldsymbol{b}_1 \in \mathbb{R}^{d_v \times 1}$, and $\boldsymbol{b}_2 \in \mathbb{R}^{d_v \times 1}$ are learnable parameters. Likewise, we can get $\boldsymbol{S}_{i2t}$.

To optimize image-text retrieval, following previous approaches[14, 24, 25], we employ the bidirectional triplet loss as the optimization objective, that is

$$\mathcal{L}_r = \sum_{(I,T)} [\lambda_1 - S(I, T) + S(I, T^-)]_+ + $$

$$[\lambda_1 - S(I, T) + S(I^-, T)]_+, \tag{7}$$

where λ_1 is a margin. I and T are the matched image-text pairs, and T^- and I^- are corresponding negatives, respectively. Since the selection processes for both T^- and I^- are the same, for convenience, we take the selection of T^- as an example. Specifically, for each batch of training data B, we consider all texts except T as candidate negative examples of image I. The top $C/B + 1$ candidates most relevant to the image I are chosen as negative example T^-, where C indicates the number of $\boldsymbol{S} + \lambda_1 - f_{\mathrm{diag}}(\boldsymbol{S}) > 0$. Here, $\boldsymbol{S} \in \{\boldsymbol{S}_{t2i}, \boldsymbol{S}_{i2t}\}$, and $f_{\mathrm{diag}}(\boldsymbol{X})$ means that for any $x_{ij} \in \boldsymbol{X}$, the element x_{ij} is replaced with the diagonal element x_{ii}, followed by adding $\lambda_1 + 0.01$ to x_{ii}.

3.2.4 Text-Text Matching

Different from image-text retrieval, we view the text-text matching task as an auxiliary task to perform semantic constraint from a global perspective. As shown in Fig.1, the texts T and T^+ are the ground-truth captions of the image I. Theoretically, if the text T and image I exhibit a high degree of semantic similarity, and text T^+ and image I also show a solid semantic correspondence, then it can be inferred that there exists a high semantic correlation between text T and text T^+. Based on the above analysis, we can improve the performance of image-text retrieval by constraining the semantic consistency of these two texts to alleviate the semantic bias

that may occur during the model's training process.

Specifically, we first utilize the intra-modality interaction module to capture textual semantics information. It is worth noting that the intra-modality interaction module here differs from (4). In (4), the matrices $\boldsymbol{K}$ and $\boldsymbol{V}$ are extended, but this operation is not performed here, constituting the sole distinction between them. To avoid duplication and distinguish them, we employ the function $g_{\text{intra}}^{T}(\cdot)$ to represent the intra-modal interaction here. Taking text T as an example, the feature $\boldsymbol{t}_{T,\,\text{intra}}$ of text T after intra-modality interaction can be written as

$$\boldsymbol{t}_{T,\,\text{intra}} = g_{\text{intra}}^{T}(\boldsymbol{t}, \boldsymbol{\theta}_{T,\,\text{intra}}),$$

where $\boldsymbol{t}_{T,\,\text{intra}} = (\boldsymbol{t}_{1,\,\text{intra}}, \boldsymbol{t}_{2,\,\text{intra}}, \ldots, \boldsymbol{t}_{n,\,\text{intra}})$ and $\boldsymbol{\theta}_{T,\,\text{intra}}$ is a learnable parameter. Then, we obtain the global semantic feature $\boldsymbol{t}_{\text{g}}$ of text T through a mean operation, i.e.,

$$\boldsymbol{t}_{\text{g}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{t}_{j,\,\text{intra}}.$$

Analogously, the global semantic feature of text T^{+} can be denoted as $\boldsymbol{t}_{\text{g}}^{+}$. Subsequently, we exploit cosine similarity to measure the correlation score between $\boldsymbol{t}_{\text{g}}$ and $\boldsymbol{t}_{\text{g}}^{+}$, that is, $S(T, T^{+}) = \cos(\boldsymbol{t}_{\text{g}}, \boldsymbol{t}_{\text{g}}^{+})$. Similar to image-text retrieval, we also employ bidirectional triplet loss as the optimization objective of text-text matching, i.e.,

$$\mathcal{L}_m = \sum_{(T,\,T^+)} [\lambda_2 - S(T, T^+) + S(T, \widehat{T}^+)]_+ +$$
$$[\lambda_2 - S(T, T^+) + S(\widehat{T}, T^+)]_+, \qquad (8)$$

where λ_2 is a margin. $\widehat{T}$ and $\widehat{T}^+$ are text negatives of T and T^+, respectively, obtained in the same manner as described in (7) for negatives.

3.2.5 Multi-Label Classification

Similar to the text-text matching task, the multi-label classification is also considered an auxiliary task for implementing global semantic constraints. Intuitively, each image or text will display some semantic information. For example, the image and texts in Fig.1 contain significant semantic features such as "street" and "clock tower". By constraining the shared semantic information between images and texts, it can be ensured that the matched images and

texts always maintain semantic consistency during the model training process. Since existing image-text retrieval datasets such as Flickr30K and MSCOCO do not contain semantic label information, a straightforward approach is to build semantic labels directly from these datasets. Given that nouns can embody the semantic information of images better than other types of words, we adopt nouns to construct semantic labels for image-text pairs.

Concretely, we first employ NLTK[①] to identify nouns in the text and then form the semantic label vector $\boldsymbol{L} \in \mathbb{R}^{1 \times N}$ by selecting the top-N nouns with the highest frequencies. Then, a "0-1" encoded semantic label $\boldsymbol{y} \in \mathbb{R}^{1 \times N}$ is assigned to each matched image-text pair (I, T), where "1" indicates that the semantic information at the same position in the label dictionary $\boldsymbol{L}$ is present in the image-text pair. Subsequently, we adopt the transformed features $\boldsymbol{v}_{t2i} = \{\boldsymbol{t}_{1,\,v}', \boldsymbol{t}_{2,\,v}', \ldots, \boldsymbol{t}_{k,\,v}'\}$ and $\boldsymbol{t}_{i2t} = \{\boldsymbol{v}_{1,\,t}', \boldsymbol{v}_{2,\,t}', \ldots, \boldsymbol{v}_{n,\,t}'\}$ to perform the multi-classification task. Formally, a mean function and two fully connected layers are utilized to predict the semantic labels $\boldsymbol{y}' \in \mathbb{R}^{1 \times N}$. The process can be formalized as

$$\boldsymbol{y}' = (((1 - \beta) \times \frac{1}{k} \sum_{r=1}^{k} \boldsymbol{t}_{r,\,v}' + \beta \times \frac{1}{n} \sum_{s=1}^{n} \boldsymbol{v}_{s,\,t}') \boldsymbol{w}_{11} +$$
$$\boldsymbol{b}_{11}) \boldsymbol{w}_{22} + \boldsymbol{b}_{22}, \qquad (9)$$

where β is a balancing factor. $\boldsymbol{w}_{11} \in \mathbb{R}^{d \times N}$, $\boldsymbol{w}_{22} \in \mathbb{R}^{N \times N}$, $\boldsymbol{b}_{11} \in \mathbb{R}^{d \times 1}$, and $\boldsymbol{b}_{22} \in \mathbb{R}^{N \times 1}$ are learnable parameters. Then, for the matched image-text pair (I, T), the predicted label $\boldsymbol{y}' = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots, \boldsymbol{y}_u', \ldots, \boldsymbol{y}_N')$ and ground-truth label $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_u, \ldots, \boldsymbol{y}_N)$ can be viewed as N binary classification problems, and the optimization objective can be written as

$$\mathcal{L}_c = -\sum_{u=1}^{N} \boldsymbol{y}_u \log \sigma(\boldsymbol{y}_u') + (1 - \boldsymbol{y}_u) \log \sigma(1 - \boldsymbol{y}_u'), \quad (10)$$

where σ indicates the "Sigmoid" activation function.

3.3 Optimization Objective

As discussed in Section 1, multi-task learning is beneficial to enhance the generalization and robustness of the model. Therefore, we adopt a joint approach to optimize these three tasks, consisting of im-

---

age-text retrieval, text-text matching, and multi-label classification. Finally, the optimization objective is defined as

$$\mathcal{L} = \mathcal{L}_r + \delta(\gamma\mathcal{L}_m + (1 - \gamma)\mathcal{L}_c), \qquad (11)$$

where δ is a balancing factor used to balance the weights between the target and auxiliary tasks. γ is employed to balance the weight between these two auxiliary tasks.

Additionally, it should be noted that all three tasks participate in the model training process, but the auxiliary tasks are not involved during the testing stage. Algorithm 1 is the entire training process, where θ is the parameter that the proposed MVSEN needs to train and optimize.

Algorithm 1. Model Training Process of Proposed MVSEN

Input: matched image-text pairs (I, T); positive sample T^+ of T; parameters: α, λ_1, B, λ_2, β, δ, γ; optimized network parameters: θ;
Output: θ
1 **for** $epoch = 1, 2, \ldots, E$ **do**
2   **for** each batch size B **do**
3   Initial feature representations: v, t, t^+ via (1) and (2);
4   Perform intra- and inter-modality feature interaction via (3) and (6);
5   Compute loss $\mathcal{L}_r$ for image-text retrieval via (7);
6   Obtain loss $\mathcal{L}_m$ for text-text matching via (8);
7   Calculate loss $\mathcal{L}_c$ for multi-label classification via (10);
8   **end**
9   Get losses $\mathcal{L}_r$, $\mathcal{L}_m$ and $\mathcal{L}_c$;
10   Compute $\mathcal{L}$ via (11);
11    $\theta \leftarrow$ Backward $(\mathcal{L})$
12 **end**

---

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the proposed approach MVSEN and all baselines on two publicly available datasets, Flickr30K[15] and MSCOCO[16]. Flickr30K consists of 31 000 images in total and each image is associated with five matched textual descriptions. Following the split protocol in [14], we adopt 29 000 images for training, 1 000 images for validation, and 1 000 images for testing. MSCOCO includes 123 287 images, and each image is manually annotated with five sentences, where 113 287 images are employed for training, 5 000 images for validation, and 5 000 images for testing. It should be noted that the performance of

MSCOCO is evaluated by averaging 5-fold of 1k and all 5k test images.

Evaluation Metrics. Following previous methods[23, 24], we utilize recall at K ($R@K, K = 1, 5, 10$) as an evaluation metric to assess the performance of cross-modal retrieval. Also, a comprehensive evaluation metric $rSum$, indicating the sum of $R@K$ in cross-modal retrieval, is also adopted to evaluate the model's performance.

Implementation Details. The proposed model MVSEN is implemented using Python 3.7.0 and PyTorch 1.7.0 frameworks and trained on an NVIDIA GeForce RTX 3090 GPU with the Adam optimizer. We set the training epoch to 30 with a learning rate of 0.000 2 on Flickr30K and MSCOCO, where the learning rate is reduced by 10% after 10 epochs. The batch size is set to 64. The control factor α in (5) is 10, i.e., $\alpha = 10$. The margins λ_1 in (7) and λ_2 in (8) are set to 0.2 and 0.1, respectively. The balancing factor β in (9) is set to 0.99 and δ in (11) is set to 0.01. Additionally, some detailed settings can be found in our code②.

4.2 Quantitative Results and Analysis

Results on Flickr30K. Table 1 reports the quantitative results on the Flickr30K dataset③. It can be observed that the proposed MVSEN exceeds all baseline models in the evaluation metric $rSum$, with a gain of 8.2%–26.2%. Compared with the state-of-the-art method NAAF[19], the proposed MVSEN obtains performance improvements of 0.4% and 2.3% on $R@1$ for text retrieval and image retrieval, respectively. Also, compared with baseline models, such as SGRAF[25], MEMBER[26], CGMN[24], and so on,

Table 1. Quantitative Results on Flickr30K

Method	Text Retrieval			Image Retrieval			$rSum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
SGRAF$_{2021}$[25]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
MEMBER$_{2021}$[29]	77.5	94.7	97.3	59.5	84.8	91.0	504.8
CGMN$^*_{2022}$[24]	77.9	93.8	96.8	59.9	85.1	90.6	504.1
UARDA$_{2022}$[18]	77.8	95.0	97.6	57.8	82.9	89.2	500.3
NAAF$^*_{2022}$[19]	81.3	95.6	98.1	60.8	84.8	90.7	511.3
GLFN$_{2023}$[1]	75.1	93.8	97.2	54.5	82.8	89.9	493.3
RAAN$_{2023}$[21]	77.1	93.6	97.3	56.0	82.4	89.1	495.5
VSRN++$_{2023}$[37]	79.2	94.6	97.5	60.6	85.6	91.4	508.9
MVSEN (ours)	**81.7**	**95.6**	**98.2**	**63.1**	**88.0**	**92.9**	**519.5**

MVSEN continues to demonstrate notable performance enhancements in terms of $R@1$, exhibiting improvements ranging from 2.5% to 6.6% for text retrieval and from 2.5% to 8.6% for image retrieval, which confirms the efficacy of MVSEN in the realm of image-text retrieval.

Results on MSCOCO. Tables 2 and 3 show the experimental performance on MSCOCO (1k) and MSCOCO (5k), respectively. From Table 2, we can observe that the proposed MVSEN obtains the best results $R@1 = 80.5\%$ for text retrieval and

Table 2.   Quantitative Results on MSCOCO (1k)

Method	Text Retrieval			Image Retrieval			$rSum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
SGRAF[25]$_{2021}$	79.6	96.2	98.5	63.2	90.7	96.1	524.3
MEMBER[29]$_{2021}$	78.5	**96.8**	98.5	63.7	90.7	95.6	523.8
CGMN*[24]$_{2022}$	76.8	95.4	98.3	63.8	90.7	95.7	520.7
UARDA[18]$_{2022}$	78.6	96.5	**98.9**	63.9	90.7	96.2	524.8
NAAF*[19]$_{2022}$	79.7	96.4	98.6	63.0	89.5	95.2	522.4
GLFN[1]$_{2023}$	78.4	96.0	98.5	62.6	89.6	95.4	520.5
RAAN[21]$_{2023}$	76.8	96.4	98.3	61.8	89.5	95.8	518.6
VSRN++[37]$_{2023}$	77.9	96.0	98.5	64.1	91.0	96.1	523.6
MVSEN (ours)	**80.5**	96.5	98.7	**64.6**	**91.1**	**96.4**	**527.8**

Table 3.   Quantitative Results on MSCOCO (5k)

Method	Text Retrieval			Image Retrieval			$rSum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
MEMBER[29]$_{2021}$	54.5	82.3	90.1	40.9	71.0	81.8	420.6
CGMN*[24]$_{2022}$	53.4	81.3	89.6	41.2	71.9	82.4	419.8
UARDA[18]$_{2022}$	56.2	83.8	91.3	40.6	69.5	80.9	422.3
NAAF*[19]$_{2022}$	58.1	**85.5**	**92.0**	42.1	70.7	80.8	429.2
VSRN++[37]$_{2023}$	54.7	82.9	90.9	42.0	72.2	82.7	425.4
MVSEN (ours)	**58.7**	84.0	91.7	**42.5**	72.0	82.7	**431.6**

$R@1 = 64.6\%$ for image retrieval compared with all baseline models. At the same time, MVSEN also achieves the best performance in terms of evaluation metric $rSum$, with $rSum = 527.8\%$. However, in text retrieval, the performance of MVSEN is lower than that of MEMBER[29] in terms of the evaluation metric $R@5$ and lower than that of UARDA[18] in terms of the evaluation metric $R@10$. The possible reasons are that MEMBER employs the global memory bank to improve visual features through relevant textual features, and UARDA utilizes optimized correlation thresholds to learn discriminative visual features, resulting in the model's performance being better than that of MVSEN. In addition, when the test data increases, it can be seen from Table 3 that MVSEN still has advantages in the evaluation metrics $R@1$ and $rSum$ compared with baseline models, which indirectly shows that the proposed MVSEN can handle cross-modal image-text retrieval under different data scales well. Furthermore, we can observe that in text retrieval, the performance of MVSEN is worse than that of NAAF[19] in $R@5$ and $R@10$. One possible reason is that NAAF considers the impact of misaligned segments on model performance, whereas MVSEN does not consider this.

4.3   Effect of Different Network Modules

To assess the influence of different modules on the model's performance, we conduct ablation studies on the Flickr30K and MSCOCO datasets, as shown in Table 4, where "✓" indicates that the corresponding module is adopted. R, M, and C represent image-text retrieval, text-text matching, and multi-classifi-

Table 4.   Ablation Studies on Flcikr30K and MSCOCO (1k)

Model	Setting					Flickr30K Dataset							MSCOCO Dataset (1k)						
						Text Retrieval			Image Retrieval			$rSum$	Text Retrieval			Image Retrieval			$rSum$
	R	M	C	$t2i$	$i2t$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
#1	✓			✓		76.9	93.5	96.9	60.6	85.7	91.6	505.2	78.0	95.8	98.2	63.0	90.5	96.0	521.5
#2	✓				✓	76.7	93.0	96.9	58.6	84.5	90.6	500.3	76.1	95.7	98.6	61.3	90.1	95.8	517.6
#3	✓			✓	✓	79.1	94.5	97.9	61.7	87.4	92.4	513.0	79.1	96.2	98.5	64.4	91.2	96.4	525.8
#4	✓	✓		✓		77.1	94.2	98.0	60.4	85.2	91.5	506.4	77.8	95.5	98.3	63.0	90.5	96.0	521.1
#5	✓	✓			✓	75.5	94.2	97.6	58.6	85.1	90.7	501.7	77.5	96.0	98.5	62.3	90.5	95.8	520.6
#6	✓	✓		✓	✓	80.2	95.5	97.9	62.2	86.9	92.6	515.3	79.7	96.3	98.6	64.3	**91.5**	**96.5**	526.9
#7	✓		✓	✓	77.2	93.2	97.1	61.4	85.8	91.8	506.5	77.8	95.3	98.4	62.1	90.7	96.1	520.4
#8	✓		✓		✓	75.4	93.5	96.9	58.9	85.2	90.8	500.7	77.1	95.7	98.5	61.4	90.2	95.9	518.8
#9	✓		✓	✓	✓	79.7	93.9	97.8	**63.6**	87.4	92.7	515.1	80.0	96.3	98.5	64.4	91.3	**96.5**	527.0
#10	✓	✓	✓	✓	79.5	94.6	97.7	61.2	86.6	91.7	511.3	77.5	96.0	98.3	62.7	90.3	95.8	520.6
#11	✓	✓	✓		✓	77.8	93.3	97.2	59.2	85.0	90.5	503.0	77.7	95.8	98.5	61.6	90.3	95.9	519.8
#12	✓	✓	✓	✓	✓	**81.7**	**95.6**	**98.2**	63.1	**88.0**	**92.9**	**519.5**	**80.5**	**96.5**	**98.7**	**64.6**	91.1	96.4	**527.8**

cation tasks, respectively. *t2i* and *i2t* denote that we employ $\boldsymbol{S}_{t2i}$ and $\boldsymbol{S}_{i2t}$ to measure the semantic correlation of image-text pairs, respectively. From Table 4, we can see that the model #12 achieves the best performance in the evaluation metric *rSum* when all auxiliary tasks are employed, with improvements of $rSum = 6.5\%$ on Flickr30K and $rSum = 2.0\%$ on MSCOCO compared with the baseline model #3. Additionally, when only one auxiliary task M or C is used, the model's performance is still enhanced, with a boost of 2.3% (#6 vs #3) and 2.1% (#9 vs #3) in terms of *rSum* on dataset Flickr30K. Similarly, we can observe similar properties on dataset MSCOCO, which indicates that both text-text matching and multi-label classification tasks contribute to improving the performance of cross-modal retrieval. Moreover, their combined use yields even better results. Furthermore, it can be seen that the multi-label classification task contributes more to Flickr30K than to MSCOCO. One possible reason is that Flickr30K describes people's daily lives, and the images share many similarities. The multi-label classification task effectively constrains sentences with high similarity. In contrast, MSCOCO encompasses diverse categories, and the images exhibit more considerable variations. In this case, the role of the multi-label classification task is relatively diminished, resulting in slight performance improvement on dataset MSCOCO.

4.4 Effect of Different Parameters

In this subsection, we select two representative hyper-parameters, l and γ, to explore the impact of different parameter settings on the model's performance, where l indicates the number of GCN layers, ranging from 1 to 5, and is critical for the overall performance of image-text retrieval. The hyper-parameter γ is employed to balance the loss weights between the target and auxiliary tasks, facilitating the training of multi-task learning. The value of γ is set from 0.1 to 0.9 with a step of 0.2.

The influence of parameter l on Flickr30K and MSCOCO is depicted in Fig.2. From these results, it can be observed that increasing the parameter l from 1 to 3 can enhance the performance of image-text retrieval on evaluation metrics $R@1$ and $R@5$. The possible reason is that increasing the parameter l explores deep feature representation, which is beneficial to improve the model's performance. However, the performance decreases when the parameter l is larger than 3, such as $l = 4$ or $l = 5$. One possible explanation is that the "over-smoothing" problem occurs when the number of GCN layers increases to a certain level, resulting in performance degradation.

Figs.3 and 4 show the impact of parameter γ on Flickr30K and MSCOCO, respectively, where $\gamma(o)$, $o \in \{t2i, i2t\}$, indicates that we employ $\boldsymbol{S}_o$ to measure the similarity of image-text pairs. Clearly, when the value of γ increases from 0.1 to 0.9 (the step size is 0.2), the experimental results also change, which indirectly shows that when the weight of the auxiliary task changes, the optimization direction of the model will also be different, resulting in performance difference. When $\gamma(t2i)=0.5$ and $\gamma(i2t)=0.1$, MVSEN obtains the best $R@1$ 81.7% for text retrieval and the best $rSum$ 519.5% for cross-modal retrieval on
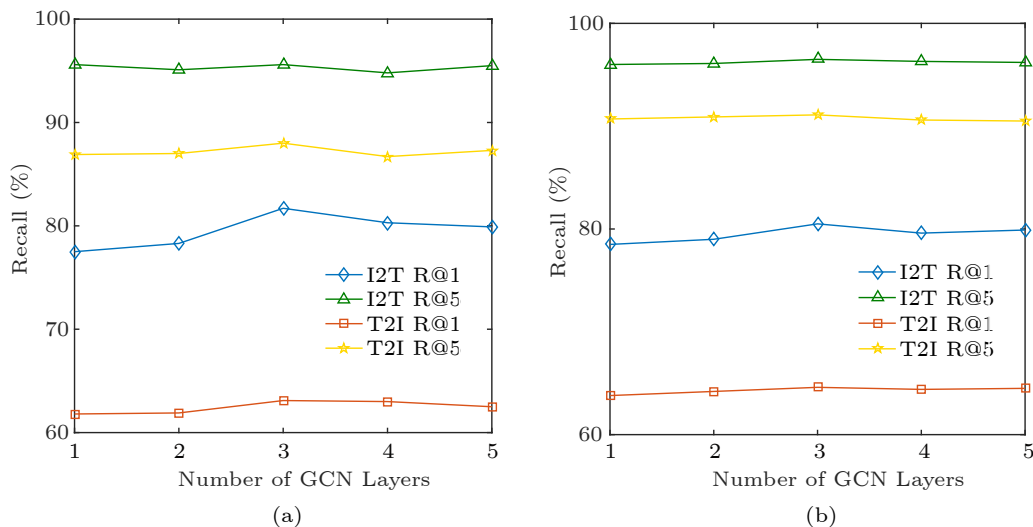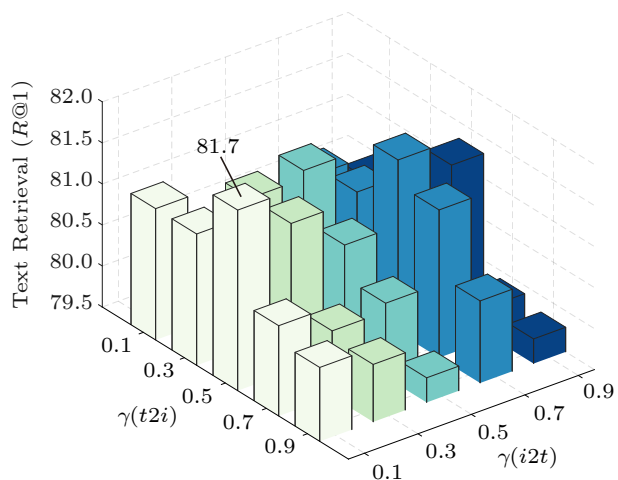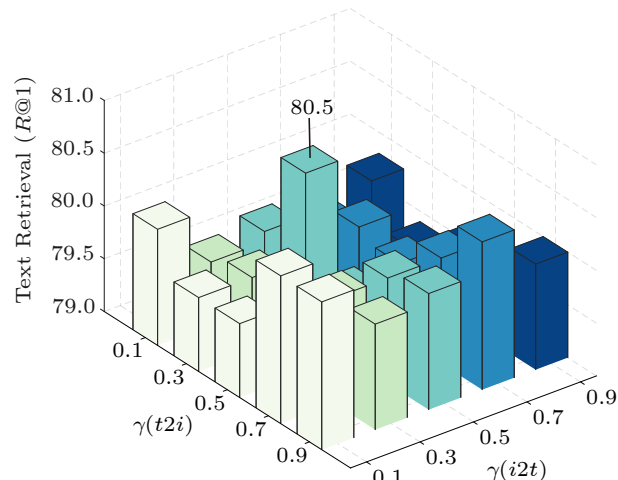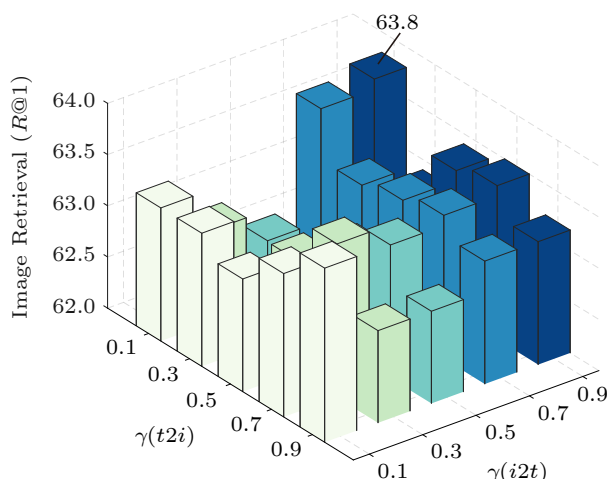

Fig.2. Effect of the number of GCN layers (l) on datasets (a) Flickr30k and (b) MSCOCO, where I2T and T2I indicate text retrieval and image retrieval, respectively.
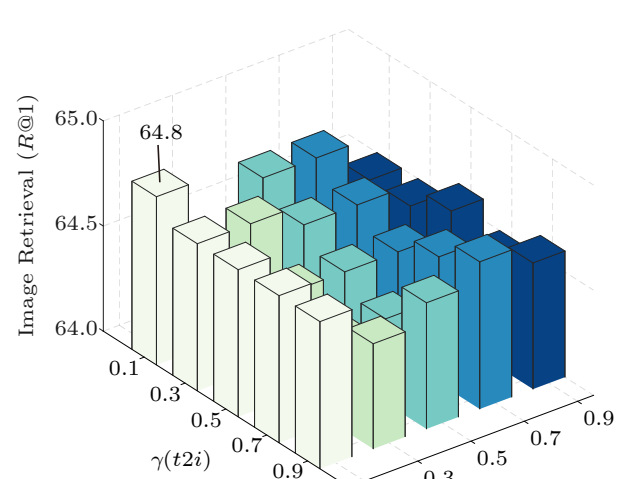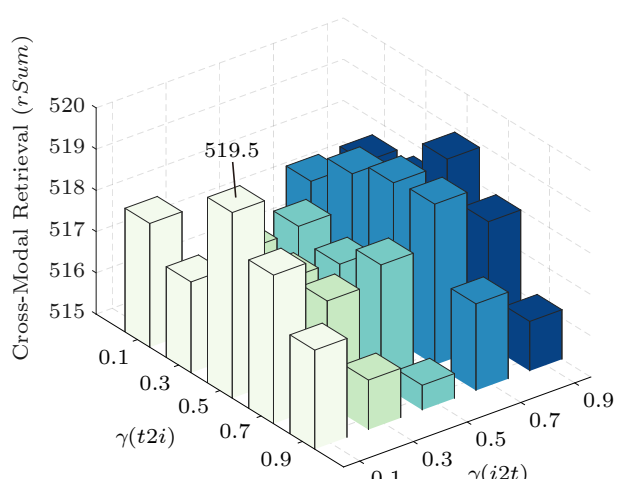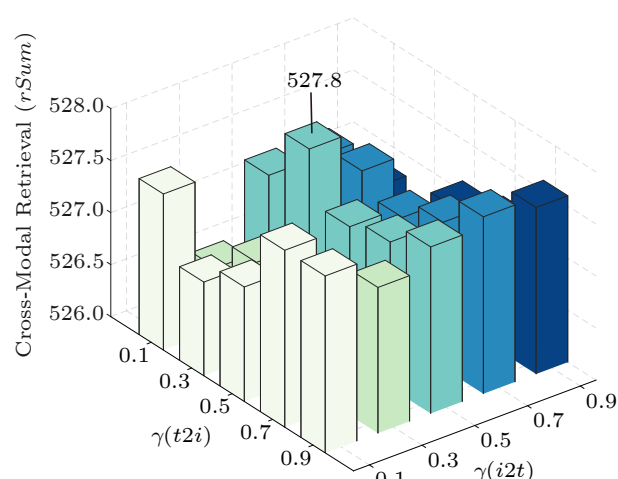
Fig.3. Effect of parameter γ on Flickr30k, and the best result is marked in the figure. (a) Text retrieval ($R@1$) on Flickr30k. (b) Image retrieval ($R@1$) on Flickr30k. (c) Cross-modal retrieval ($rSum$) on Flickr30k.

Fig.4. Effect of parameter γ on MSCOCO, and the best result is marked in the figure. (a) Text retrieval ($R@1$) on MSCOCO. (b) Image retrieval ($R@1$) on MSCOCO. (c) Cross-modal retrieval ($rSum$) on Flickr30k.

Flickr30K. Analogously, it can be seen from Fig.4 that MVSEN achieves the best text retrieval, $\gamma(t2i)=0.3$ and $\gamma(i2t)=0.5$ and the best image retrieval $\gamma(t2i)=0.1$ and $\gamma(i2t)=0.1$, by taking different values of γ.

By analyzing the impact of parameters l and γ on experimental results, we can discern that different parameter configurations substantially influence the model's performance. An appropriate parameter is helpful to improve the model's efficiency. Furthermore, it can be seen that when the model achieves the best performance, the same parameter, such as the parameter γ, may have different values on different datasets due to differences in data distribution.

4.5 Analysis of Retrieval Time

To validate that the proposed method achieves a good balance between performance and efficiency, we compare the retrieval time with three advanced approaches, SGRAF[25], CGMN[24], and NAAF[19] based on the publicly available codes they provide. For a fair comparison, we obtain the bidirectional retrieval time by averaging the retrieval time of 5 000 image-text pairs, and all experiments are performed on an Intel® Core™ i9-10920X CPU@3.50 GHz, 64 GB memory and an NVIDIA GeForce RTX 3090 GPU with 24 GB memory.

Fig.5 shows that when performing a bidirectional retrieval, the proposed approach MVSEN is significantly lower than SGRAF and NAAF but higher than CGMN in the retrieval time. Furthermore, we can observe that the retrieval time of NAAF is the longest among these four methods. The reason is that NAAF employs the correlations between all visual regions and textual words to assess the semantic simil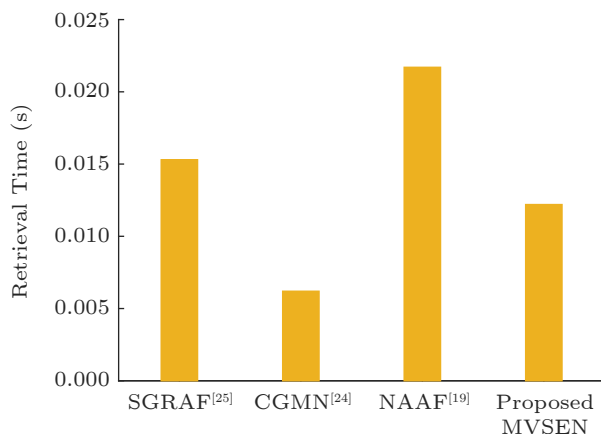arity of the entire image and text, resulting in a significant time overhead. In contrast, CGMN only utilizes cosine to measure the similarity of image-text pairs, and thus the cost is minimal. Although the retrieval time of the proposed MVSEN is higher than that of CGMN, it is still within the same order of magnitude. Moreover, the performance of MVSEN on the evaluation metric $rSum$ exceeds CGMN by a large margin, with 15.3% on Flickr30K and 7.1% on MSCOCO (1k). Therefore, the proposed approach performs better in balancing performance and retrieval efficiency.

4.6 Visualization and Analysis

To assess the performance of the proposed MVSEN further, we visualize the retrieval results and compare MVSEN with the advanced method NAAF[19]. For text retrieval, we offer the ground-truth caption (GT), and the top-ranked text retrieved by MVSEN and NAAF. Also, we report the top-2 results for image retrieval, where the correct result is marked with a green box. For each text query, we employ MVSEN to obtain the first two retrieval results and NAAF to obtain the last two retrieval results.

Figs.6 and 7 are the results of text retrieval on Flickr30K and MSCOCO datasets, respectively. As can be seen from these retrieval results, MVSEN performs better than NAAF. Taking the third example on Flcikr30K as an illustration, although NAAF iden-
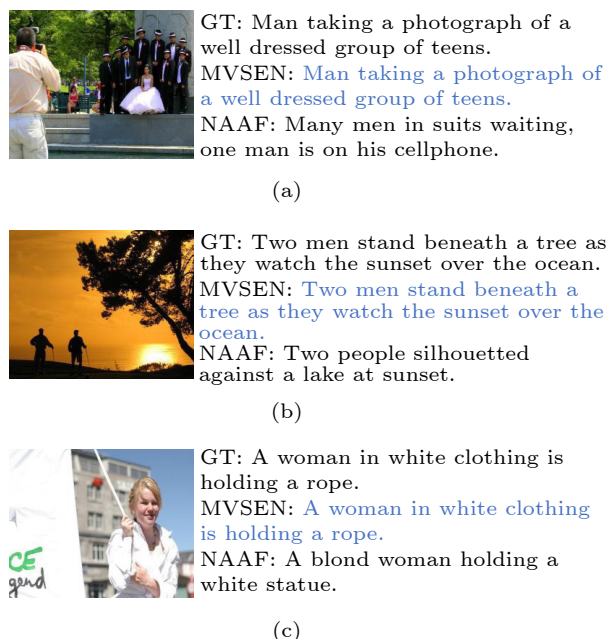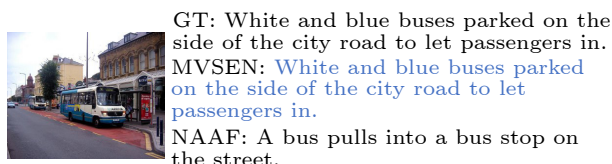

GT: Man taking a photograph of a well dressed group of teens.
MVSEN: Man taking a photograph of a well dressed group of teens.
NAAF: Many men in suits waiting, one man is on his cellphone.

(a)

GT: Two men stand beneath a tree as they watch the sunset over the ocean.
MVSEN: Two men stand beneath a tree as they watch the sunset over the ocean.
NAAF: Two people silhouetted against a lake at sunset.

(b)

GT: A woman in white clothing is holding a rope.
MVSEN: A woman in white clothing is holding a rope.
NAAF: A blond woman holding a white statue.

(c)

Fig.6. Qualitative results of text retrieval on Flickr30K.

Fig.5. Comparison of bidirectional retrieval time.

GT: White and blue buses parked on the side of the city road to let passengers in.
MVSEN: White and blue buses parked on the side of the city road to let passengers in.
NAAF: A bus pulls into a bus stop on the street.

(a)
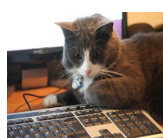
GT: A grey cat peers at a computer keyboard.
MVSEN: A grey cat peers at a computer keyboard.
NAAF: Cat sitting right next to keyboard on laptop.

(b)

GT: A photo taken in a car looking at a dog in the back seat.
MVSEN: A photo taken in a car looking at a dog in the back seat.
NAAF: A blissful dog laying against a windscreen of a car.

(c)

Fig.7. Qualitative results of text retrieval on Flickr30K.

tifies some vital elements such as "blond woman", "holding", and "white", it mistakes "rope" for "statue", leading to a wrong judgment. In contrast, MVSEN can detect these small gaps, which also verifies the effectiveness of multi-task learning in text retrieval.

Figs.8 and 9 show qualitative results of image retrieval on the Flickr30K and MSCOCO datasets, respectively. From these retrieval results, we can observe that for any given query text, all the images retrieved by NAAF and the proposed MVSEN exhibit high similarity. The difference is that MVSEN can distinguish them well and retrieve the ground-truth in the top-ranked result. However, NAAF is confused by

(a)

(b)

(c)

Fig.8. Qualitative results of image retrieval on Flickr30K. (a) Query 1: a woman gives a small child a piggyback ride. (b) Query 2: three people sit at an outdoor cafe. (c) Query 3: the girls dance to the street musicians.

(a)

(b)

(c)

Fig.9. Qualitative results of image retrieval on MSCOCO. (a) Query 1: young boy with T-ball and bat pointing at ball. (b) Query 2: a group of people play video games with controllers. (c) Query 3: a fridge and a sink in a home kitchen.

these similar results, leading to "incorrect" choices. Therefore, it can be inferred from these results that the proposed MVSEN can make sound judgments when facing high-similarity images.

5 Conclusions

In this paper, we proposed a multi-task visual semantic embedding network (MVSEN) that leverages the collaboration of different tasks to explore the semantic relevance of image-text pairs, aiming to improve the performance of cross-modal retrieval. Experimental results on two publicly available benchmarks, Flickr30K and MSCOCO, show that MVSEN performs better than state-of-the-art approaches, with rSum improvements of 8.2% and 3.0%, respectively. Additionally, through ablation experiments, it can be observed that both text-text matching and multi-label classification tasks contribute to improving the performance of cross-modal retrieval. Furthermore, qualitative results confirm the effectiveness of the proposed method. In fact, the two auxiliary tasks designed for semantic constraints are more suitable for handling highly similar images. The effects are less pronounced for images with significant differences, which is a limitation of the proposed multi-task scheme. In future work, we will explore the feasibility and efficacy of multi-task learning in other cross-modal tasks.

Conflict of Interest   The authors declare that they have no conflict of interest.

References

[1] Zhao G S, Zhang C F, Shang H, Wang Y X, Zhu L, Qian X M. Generative label fused network for image-text matching. *Knowledge-Based Systems*, 2023, 263: 110280. DOI: 10.1016/j.knosys.2023.110280.

[2] Qin X Y, Li L S, Hao F, Pang G Y, Wang Z H. Cross-modal information balance-aware reasoning network for image-text retrieval. *Engineering Applications of Artificial Intelligence*, 2023, 120: 105923. DOI: 10.1016/j.engappai.2023.105923.

[3] Liu K, Xue F, Guo D, Sun P J, Qian S S, Hong R C. Multimodal graph contrastive learning for multimedia-based recommendation. *IEEE Trans. Multimedia*, 2023, 25: 9343–9355. DOI: 10.1109/TMM.2023.3251108.

[4] Wu Y X, Liao L Z, Zhang G Y, Lei W Q, Zhao G S, Qian X M, Chua T S. State graph reasoning for multimodal conversational recommendation. *IEEE Trans. Multimedia*, 2023, 25: 3113–3124. DOI: 10.1109/TMM.2022.3155900.

[5] Wen Z, Peng Y X. Multi-level knowledge injecting for visual commonsense reasoning. *IEEE Trans. Circuits and Systems for Video Technology*, 2021, 31(3): 1042–1054. DOI: 10.1109/TCSVT.2020.2991866.

[6] Li Z Y, Guo Y Y, Wang K J, Wei Y W, Nie L Q, Kankanhalli M. Joint answering and explanation for visual commonsense reasoning. *IEEE Trans. Image Processing*, 2023, 32: 3836–3846. DOI: 10.1109/TIP.2023.3286259.

[7] Wang L W, Li Y, Huang J, Lazebnik S. Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2019, 41(2): 394–407. DOI: 10.1109/TPAMI.2018.2797921.

[8] Liu Y, Guo Y M, Liu L, Bakker E M, Lew M S. CycleMatch: A cycle-consistent embedding network for image-text matching. *Pattern Recognition*, 2019, 93: 365–379. DOI: 10.1016/j.patcog.2019.05.008.

[9] Li K P, Zhang Y L, Li K, Li Y Y, Fu Y. Visual semantic reasoning for image-text matching. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27–Nov. 2, 2019, pp.4654–4662. DOI: 10.1109/iccv.2019.00475.

[10] Sarafianos N, Xu X, Kakadiaris I A. Adversarial representation learning for text-to-image matching. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27–Nov. 2, 2019, pp.5814–5824. DOI: 10.1109/iccv.2019.00591.

[11] Peng Y X, Qi J W. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimedia Computing, Communications, and Applications*, 2019, 15(1): Article No. 22. DOI: 10.1145/3284750.

[12] Chi J Z, Peng Y X. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Trans. Circuits and Systems for Video Technology*, 2020, 30(4): 1173–1187. DOI: 10.1109/TCSVT.2019.2900171.

[13] Xie Y C, Zeng X H, Wang T H, Xu L M, Wang D J. Multiple deep neural networks with multiple labels for cross-modal hashing retrieval. *Engineering Applications of Artificial Intelligence*, 2022, 114: 105090. DOI: 10.1016/j.engappai.2022.105090.

[14] Lee K H, Chen X, Hua G, Hu H D, He X D. Stacked cross attention for image-text matching. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018, pp.201–216. DOI: 10.1007/978-3-030-01225-0_13.

[15] Liu C X, Mao Z D, Liu A A, Zhang T Z, Wang B, Zhang Y D. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proc. the 27th ACM International Conference on Multimedia*, Oct. 2019, pp.3–11. DOI: 10.1145/3343031.3350869.

[16] Wei X, Zhang T Z, Li Y, Zhang Y D, Wu F. Multi-modality cross attention network for image and sentence matching. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.10941–10950. DOI: 10.1109/cvpr42600.2020.01095.

[17] He Y, Liu X, Cheung Y M, Peng S J, Yi J H, Fan W T. Cross-graph attention enhanced multi-modal correlation learning for fine-grained image-text retrieval. In *Proc. the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2021, pp.1865–1869. DOI: 10.1145/3404835.3463031.

[18] Zhang K, Mao Z D, Liu A A, Zhang Y D. Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Trans. Multimedia*, 2023, 25: 1320–1332. DOI: 10.1109/TMM.2022.3141603.

[19] Zhang K, Mao Z D, Wang Q, Zhang Y D. Negative-aware attention framework for image-text matching. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.15661–15670. DOI: 10.1109/cvpr52688.2022.01521.

[20] Wu J, Wu C L, Lu J, Wang L Q, Cui X R. Region reinforcement network with topic constraint for image-text matching. *IEEE Trans. Circuits and Systems for Video Technology*, 2022, 32(1): 388–397. DOI: 10.1109/TCSVT.2021.3060713.

[21] Wang Y, Su Y T, Li W H, Sun Z Y, Wei Z Q, Nie J, Li X Y, Liu A A. Rare-aware attention network for image–text matching. *Information Processing & Management*, 2023, 60(3): 103280. DOI: 10.1016/j.ipm.2023.103280.

[22] Chen J C, Hu H X, Wu H, Jiang Y N, Wang C H. Learning the best pooling strategy for visual semantic embedding. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.15789–15798. DOI: 10.1109/cvpr46437.2021.01553.

[23] Liu C X, Mao Z D, Zhang T Z, Xie H T, Wang B, Zhang Y D. Graph structured network for image-text matching. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.10921–10930. DOI: 10.1109/cvpr42600.2020.01093.

[24] Cheng Y H, Zhu X G, Qian J C, Wen F, Liu P L. Cross-modal graph matching network for image-text retrieval. *ACM Trans. Multimedia Computing, Communications, and Applications*, 2022, 18(4): 95. DOI: 10.1145/3499027.

[25] Diao H W, Zhang Y, Ma L, Lu H C. Similarity reasoning and filtration for image-text matching. In *Proc. the 35th AAAI Conference on Artificial Intelligence*, Feb. 2021, pp.1218–1226. DOI: 10.1609/aaai.v35i2.16209.

[26] Wang X H, Zhu L C, Yang Y. T2VLAD: Global-local sequence alignment for text-video retrieval. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.5079–5088. DOI: 10.1109/cvpr46437.2021.00504.

[27] Ji Z, Chen K X, Wang H R. Step-wise hierarchical alignment network for image-text matching. In *Proc. the 30th International Joint Conference on Artificial Intelligence*, Aug. 2021, pp.765–771. DOI: 10.24963/ijcai.2021/106.

[28] Li W H, Yang S, Wang Y, Song D, Li X Y. Multi-level similarity learning for image-text retrieval. *Information Processing & Management*, 2021, 58(1): 102432. DOI: 10.1016/j.ipm.2020.102432.

[29] Li J T, Liu L, Niu L, Zhang L Q. Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval. *IEEE Trans. Image Processing*, 2021, 30: 9193–9207. DOI: 10.1109/TIP.2021.3123553.

[30] Xu Y Y, Li X T, Yuan H B, Yang Y B, Zhang L F. Multi-task learning with multi-query transformer for dense prediction. *IEEE Trans. Circuits and Systems for Video Technology*, 2024, 34(2): 1228–1240. DOI: 10.1109/tcsvt.2023.3292995.

[31] Foggia P, Greco A, Saggese A, Vento M. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. *Engineering Applications of Artificial Intelligence*, 2023, 118: 105651. DOI: 10.1016/j.engappai.2022.105651.

[32] Moscato V, Napolano G, Postiglione M, Sperlì G. Multi-task learning for few-shot biomedical relation extraction. *Artificial Intelligence Review*, 2023, 56(11): 13743–13763. DOI: 10.1007/s10462-023-10484-6.

[33] Vandenhende S, Georgoulis S, Van Gool L. MTI-Net: Multi-scale task interaction networks for multi-task learning. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.527–543. DOI: 10.1007/978-3-030-58548-8_31.

[34] Luo J Y, Shen Y, Ao X, Zhao Z, Yang M. Cross-modal image-text retrieval with multitask learning. In *Proc. the 28th ACM International Conference on Information and Knowledge Management*, Nov. 2019, pp.2309–2312. DOI: 10.1145/3357384.3358104.

[35] Yuan H, Huang Y, Zhang D B, Chen Z R, Cheng W L, Wang L. VSR++: Improving visual semantic reasoning for fine-grained image-text matching. In *Proc. the 25th International Conference on Pattern Recognition*, Jan. 2021, pp.3728–3735. DOI: 10.1109/icpr48806.2021.9413223.

[36] Xu X, Wang T, Yang Y, Zuo L, Shen F M, Shen H T. Cross-modal attention with semantic consistence for image–text matching. *IEEE Trans. Neural Networks and Learning Systems*, 2020, 31(12): 5412–5425. DOI: 10.1109/TNNLS.2020.2967597.

[37] Li K P, Zhang Y L, Li K, Li Y Y, Fu Y. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023, 45(1): 641–656. DOI: 10.1109/TPAMI.2022.3148470.

[38] Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.6077–6086. DOI: 10.1109/cvpr.2018.00636.

[39] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.770–778. DOI: 10.1109/cvpr.2016.90.

Xue-Yang Qin received his Ph.D. degree in computer software and theory from Dalian University of Technology, Dalian, in 2024, and received his M.S. degree in computer systems organization from Shaanxi Normal University, Xi'an, in 2019. His main research interests include cross-modal retrieval, information extraction, and multimodal data processing.

Li-Shuang Li received her Ph.D. degree in knowledge management from Dalian University of Technology, Dalian, in 2013. She is currently a professor with the School of Computer Science and Technology, Dalian University of Technology, Dalian. Her current research interests include data mining, natural language processing, and information extraction.

Jing-Yao Tang received her M.Sc. degree in computer software and theory from South China Normal University, Guangzhou, in 2021. She is working toward her Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, Dalian. Her main research interests include natural language processing, information extraction, causal inference, and low-resource learning.

Fei Hao received his Ph.D. degree in computer science and engineering from Soonchunhyang University, Asan, in 2016. Since 2016, he has been with Shaanxi Normal University, Xi'an, where he is an associate professor. His research interests include social computing, soft computing, big data analytics, pervasive computing, and data mining.

Mei-Ling Ge received her Master's degree in computer applied technology from the College of Computer Science, Shaanxi Normal University, Xi'an, in 2021. She is currently a teaching assistant with the School of Computer Engineering, Weifang University, Weifang. Her current research interests include recommendation, multimedia content analysis and retrieval, machine learning, and data mining.

Guang-Yao Pang received his Ph.D. degree in computer software and theory from Shaanxi Normal University, Xi'an, in 2021, and his M.S. degree in software engineering from University of Electronic Science and Technology of China, Chengdu, in 2013. He is currently an associate professor with the Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software, Wuzhou University, Wuzhou. His main research interests include deep learning, recommendation system, and multimodal data processing.