Expert Systems With Applications 216 (2023) 119415

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Complementarity is the king: Multi-modal and multi-grained hierarchical semantic enhancement network for cross-modal retrieval

Xinlei Pei [a,b], Zheng Liu [a,b,*], Shanshan Gao [a,b], Yijun Su [c]

[a] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, 250014, Shandong, China
[b] Shandong Provincial Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan, 250014, Shandong, China
[c] School of Information Engineering, Minzu University of China, Beijing, 100081, China

ARTICLE INFO

Keywords:
Cross-modal retrieval
Primary similarity
Auxiliary similarity
Semantic enhancement
Multi-spring balance loss

ABSTRACT

Cross-modal retrieval takes a query of one modality to retrieve relevant results from another modality, and its key issue lies in how to learn the cross-modal similarity. Note that the complete semantic information of a specific concept is widely scattered over the multi-modal and multi-grained data, and it cannot be thoroughly captured by most existing methods to learn the cross-modal similarity accurately. Therefore, we propose a Multi-modal and Multi-grained Hierarchical Semantic Enhancement network (M^2HSE), which contains two stages to obtain more complete semantic information by fusing the complementarity in multi-modal and multi-grained data. In stage 1, two classes of cross-modal similarity (primary similarity and auxiliary similarity) are calculated more comprehensively in two subnetworks. Especially, the primary similarities from two subnetworks are fused to perform the cross-modal retrieval, while the auxiliary similarity provides a valuable complement for the primary similarity. In stage 2, the multi-spring balance loss is proposed to optimize the cross-modal similarity more flexibly. Utilizing this loss, the most representative samples are selected to establish the multi-spring balance system, which adaptively optimizes the cross-modal similarities until reaching the equilibrium state. Extensive experiments conducted on public benchmark datasets clearly prove the effectiveness of our proposed method and show its competitive performance with the state-of-the-arts.

1. Introduction

In the era of digital multimedia, the amount of multi-modal data (*e.g.*, texts, images, and videos) is surging at an unprecedented rate. For example, when a hot social event happens, users may take pictures, record videos, and make comments. Therefore, effectively retrieving multi-modal data has been a significant issue. Under this situation, cross-modal retrieval (Peng, Huang et al., 2017) has got extensive concerns in industry and academia in recent years. When carrying out the task of cross-modal retrieval, users utilize the query term of a specific modality to search the most relevant results of another modality, *e.g.*, retrieving images that describe the same semantic concepts as the given text and vice versa.

The principal goal of cross-modal retrieval is how to learn more precise cross-modal similarity. Nevertheless, the challenge of "heterogeneity gap" (Baltrušaitis et al., 2018), *i.e.*, inconsistent representations and distributions of various modalities, has brought tremendous difficulty to directly estimate the cross-modal similarity. In order to eliminate such heterogeneity, the mainstream of cross-modal retrieval methods focus on learning a common embedding space for various modalities (Peng, Huang et al., 2017). Over the past decade, several pioneers have been devoted to exploring the paradigm of common embedding space learning, which ranges from the early traditional shallow methods (Peng et al., 2015; Rasiwasia et al., 2010; Wang et al., 2015) to the recent deep learning methods (Diao et al., 2021; Li et al., 2022; Zhang et al., 2022).

In general, traditional shallow methods aim to project different heterogeneous features to a common space through finding a linear mapping transformation for each modality. However, their main shortcoming is that they learn the common representations from shallow networks with the low-order linear projection functions, which cannot well model the high-level semantics of multi-modal data. Due to the significant advantages of the Deep Neural Network (DNN) in generating more effective embedding features and uncovering the nonlinear cross-modal correlations, the DNNs have been successfully used in common embedding space learning. Most DNN-based methods build several subnetworks for different modalities and then optimize them together. Although these methods have overcome the disadvantages of traditional approaches and gained noticeable improvement, they

* Corresponding author at: School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, 250014, Shandong, China.
E-mail addresses: peixinlei@mail.sdufe.edu.cn (X. Pei), liuzheng@sdufe.edu.cn (Z. Liu), gss_sdufe@sdufe.edu.cn (S. Gao), 22302055@muc.edu.cn (Y. Su).

https://doi.org/10.1016/j.eswa.2022.119415
Received 27 September 2022; Received in revised form 25 November 2022; Accepted 4 December 2022
Available online 6 December 2022

actually neglect the significance of the cross-modal similarity learning yet, which is the critical factor to promote the performance of cross-modal retrieval. Concretely, there are two main defects existing in the previous methods.

Firstly, *the calculation of cross-modal similarity lacks comprehensiveness*. On the one hand, most of them only consider two modalities (*e.g.*, texts and images) for cross-modal correlation learning, while the valuable clues in other modalities are ignored. On the other hand, some methods only exploit uni-grained features to calculate the similarity, however, features of different granularities play different roles in cross-modal similarity calculation, because they emphasize distinct levels of the same semantic concept (*i.e.*, "granularity gap"). For instance, coarse-grained features extracted from whole images mainly contain the global-level information, while fine-grained features extracted from image patches pay more attention to the local-level information. The performance of existing methods suffers from limited number of modalities and granularities, because they mine incomplete semantic information to calculate the cross-modal similarity, which is difficult to bridge the "heterogeneity gap" and the "granularity gap".

Secondly, *the optimization of cross-modal similarity lacks flexibility*. Existing cross-modal retrieval methods mainly use the hinge-based triplet ranking loss (Frome et al., 2013) as the objective function, which is trained with the random sampling and then generates lots of uninformative pairs. The redundant pairs cannot provide valuable information for training, and thus lead a slow convergence and poor performance. Furthermore, the triplet loss considers all selected samples equally, which is a rigid manner for cross-modal similarity optimizing. Hence, it is still a crucial problem for cross-modal retrieval to select and weight informative pairs.

It is recognized that some characteristics existing in one modality or granularity cannot be accurately expressed with other modalities or granularities. Thus, we argue that when describing a specific semantic concept, different modalities may contain unequal amounts and views of information, that is to say, there actually exist complementary relationships between them. Similarly, complementary relationships also exist in different granularities. As shown in Fig. 1, we provide several examples about the concept of "airplane" to illustrate the above mentioned two types of complementary relationships. Particularly, the image–text pair in Fig. 1(a) is collected from Wikipedia.[1] As we often say that "a picture is worth a thousand words", the image in Fig. 1(a) contains more details that do not exist in the corresponding text, such as "sky" and "cloud". On the contrary, the text in Fig. 1(a) expresses some high-level semantic information that cannot be embodied in its paired image, such as the designation of this airplane and the company it belongs to. Therefore, we can conclude that each modality has its unique semantic information that other modalities do not have, that is, there are explicit complementary relationships between different modalities. Furthermore, as shown in Fig. 1(b), the left part describes the semantic concept "airplane" as a whole, while the right part divides the "airplane" into several components, such as "Engine" and "Wing". Therefore, Fig. 1(b) explicitly indicates that the whole image and its patches describe the semantic information at the global-level and the local-level respectively, and they are actually complementary to each other.

As illustrated in Fig. 1, the complete information of a semantic concept contains lots of semantic pieces, which are widely scattered over in different modalities and different granularities. Therefore, the key problem behind the cross-modal similarity learning is how to collect and integrate as many semantic pieces as possible. Based on the above analysis, the key idea "complementarity is the king" is proposed to guide us to handle the task of cross-modal retrieval from a fresh perspective, that is, bridging the "heterogeneity gap" and the "granularity gap" via thoroughly exploring and exploiting the complementarity in multi-modal and multi-grained data. Specifically, for the sake of bridging these two gaps in a unified framework, a Multi-modal and Multi-grained Hierarchical Semantic Enhancement network (M^2HSE) is proposed to learn the cross-modal similarity more comprehensively and more accurately through two stages.

In stage 1, the concept of "auxiliary modality" is introduced in order to exploit the semantic knowledge in other modalities, and two subnetworks are constructed based on the coarse-grained and fine-grained data respectively. Thus, the cross-modal similarity can be calculated more comprehensively with the multi-modal and multi-grained data. In stage 2, to optimize the cross-modal similarity more flexibly, a novel loss function is designed based on the multi-spring balance system, and it integrates two modules, that is, "samples selecting" and "weights assigning", into a unified framework. In summary, the major contributions of our work are outlined as follows.

- **Multi-modal and Multi-grained Hierarchical Semantic Enhancement network** (M^2HSE) is proposed to simultaneously mine and fuse the complementary semantic information distributed in different modalities and different granularities, which significantly promotes the performance of cross-modal retrieval through semantic enhancement.
- **Two classes of cross-modal similarity** (*i.e.*, primary similarity and auxiliary similarity) are defined and calculated in M^2HSE. To capture more accurate cross-modal correlations, the valuable semantic knowledge in auxiliary similarity is transferred to primary similarity with two subnetworks. Furthermore, the final fused primary similarity is used to perform the cross-modal retrieval.
- **Multi-Spring Balance loss** (MSB) is proposed to precisely optimize the cross-modal similarity between the anchor and the sample through two steps. In step 1, the critical area is defined to select the most representative samples for each anchor. In step 2, the adaptive weights are learned automatically for different similarities when the multi-spring balance system reaches its equilibrium state.

The rest of this paper is organized as follows. In Section 2, we briefly review the related works. In Section 3, we elaborate the proposed M^2HSE method. Then, the detailed experiments are conducted in Section 4. Finally, the whole work is concluded in Section 5.

2. Related works

In this section, we first present a briefly overview of cross-modal retrieval, then we further discuss the applications of metric learning in cross-modal retrieval.

2.1. Cross-modal retrieval

In order to bridge the "heterogeneity gap", the mainstream of cross-modal retrieval methods concentrate on constructing a common embedding space to assess the cross-modal similarity. The similarities among different modalities can be directly measured by mapping their original features into a common embedding space. According to the characteristics of models, the existing methods can be divided into non-DNN-based methods and DNN-based methods.

(1) non-DNN-based methods: The key issue of non-DNN-based methods is to design linear projection functions, which can effectively project the input features of various modalities into a common space. Canonical Correlation Analysis (CCA) (Rasiwasia et al., 2010) is the most typical method for cross-modal retrieval, which learns linear projection matrices to maximize the pairwise correlations between heterogeneous data in the subspace. Afterwards, in order to learn more effective common spaces, several constraints are imposed on projection function learning. For example, the constraints of relevant and irrelevant cross-modal correlations (Zhai et al., 2012a), multi-modal graph regularization (Wang et al., 2015), unified patch graph

[1] https://en.wikipedia.org/wiki/Airplane

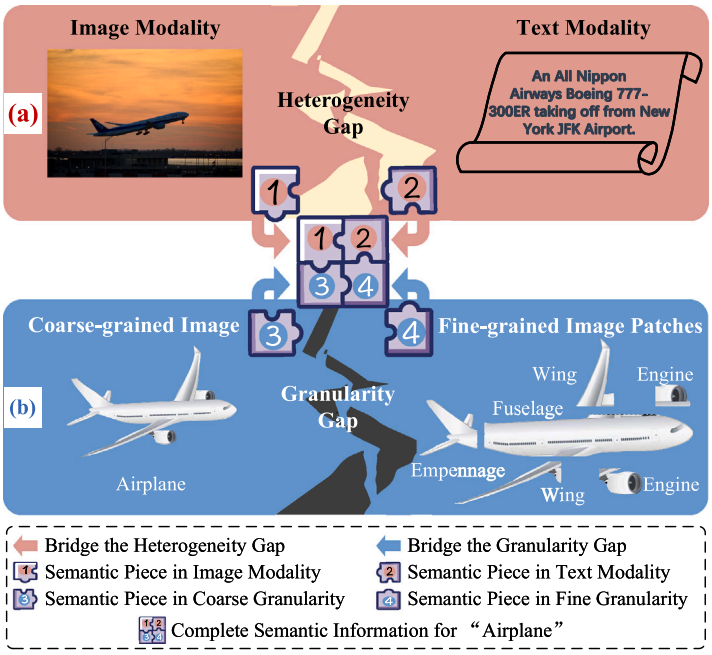


Fig. 1. Illustrating the complementary relationships existing in different modalities and different granularities about the semantic concept of "airplane". (a) Complementary relationship between different modalities. (b) Complementary relationship between different granularities. Like completing a jigsaw puzzle, the "heterogeneity gap" and the "granularity gap" can be effectively bridged through collecting and combining all semantic pieces that are scattered over multi-modal and multi-grained data.

regularization (Peng et al., 2015), have been confirmed impressive for cross-modal retrieval.

However, these non-DNN-based methods mainly use the coarse-grained features extracted from the whole image and full text to estimate the cross-modal similarity. That is, these methods only consider the global-level information of different modalities. It is worth mentioning that the fine-grained features of image patches and words can provide the complementary local-level information, which should be considered to explore the fine-grained alignments for learning more thorough cross-modal correlations.

(2) DNN-based methods: Different DNNs have been exploited in common space learning, such as Convolutional Neural Network (CNN) (Andrew et al., 2013) and Deep Belief Network (DBN) (Peng, Qi et al., 2017), for taking the advantage of the powerful ability on modeling highly nonlinear correlations. Recently, some methods based on DNNs have become the mainstream of cross-modal retrieval, including methods based on Attention Mechanism (AM) (Diao et al., 2021; Lee et al., 2018; Zhang et al., 2022), Graph Convolutional Network (GCN) (Cheng et al., 2022; Diao et al., 2021; Li et al., 2022), and Generative Adversarial Network (GAN) (Liu et al., 2022; Wang et al., 2017; Xu et al., 2020), etc.

As one of the latest advancements in deep learning, AM can be adopted to attend the discriminative fine-grained parts of different modalities. As a pioneering work, Stacked Cross Attention Network (SCAN) (Lee et al., 2018) exploited the cross-attention mechanism to estimate the cross-modal similarity, which captures the total latent alignments between visual and textual patches. Thereafter, several works (Chen et al., 2019; Peng et al., 2019; Zhang et al., 2022) extended SCAN to achieve better performance. However, these methods only make use of the inter-modality relationship for the cross-modal similarity learning.

Inspired by the utilization of Transformer (Vaswani et al., 2017) in machine translation, many recent works of cross-modal retrieval use it to mine the contextual information with self-attention mechanism. Wei et al. (2020) proposed to model the intra-modal relationship for visual and textual patches based on the self-attention module. Qu et al. (2021) proposed to aggregate the local-level information within texts based on the pre-trained BERT (Devlin et al., 2018).

Similarly, GCN (Kipf & Welling, 2016) also has been widely used for reasoning the global semantic knowledge in each modality, in which the fine-grained data serve as vertices in one or more graphs, and edges describe the correlations of them. Diao et al. (2021) constructed a similarity graph reasoning module to infer the image–text similarity with graph reasoning. Li et al. (2022) utilized GCN to establish communications between visual areas and obtained unique features with the semantic association knowledge. Cheng et al. (2022) further achieved the intra-relation and inter-relation reasoning for images and texts without affecting the search efficiency.

Besides, some cross-modal GAN methods have been proposed to exploit the adversarial learning strategy (Goodfellow et al., 2020) to either improve the effectiveness of common embedding learning or ensure the generation of synthetic features in each modality. Wang et al. (2017) proposed the adversarial cross-modal retrieval that firstly adopted the adversarial learning to learn a modality-specific common space. Gu et al. (2018) proposed a generative cross-modal feature learning framework, which incorporates the image–text generative module and the text–image generative module. More recently, Liu et al. (2022) proposed to simultaneously preserve the intrinsic structure between the original features and the projected features for images and texts with two discriminators.

These DNN-based methods calculate the cross-modal similarity with fine-grained features, which make full use of the local-level information. However, they only consider the cross-modal correlations between two modalities (*i.e.*, the primary modality in our work) that participate in cross-modal retrieval, while the valuable complementary semantic knowledge in other modalities (*i.e.*, the auxiliary modality in our work) are ignored.

In summary, existing methods cannot capture detailed correlations between different modalities and different granularities to calculate the cross-modal similarity. Notably, different from the aforementioned studies, we propose to mine and fuse the complementary semantic information distributed in multi-modal and multi-grained data to calculate more comprehensive similarity.

2.2. Metric learning for cross-modal retrieval

Metric learning aims to measure the similarity with a loss function, which ensures semantical relevant samples be closer, while pushes irrelevant ones away from each other. In previous literature, a variety of metric learning methods have been developed for various tasks.

Hadsell et al. (2006) proposed the contrastive loss to learn an invariant mapping for dimensionality reduction, where similar vectors are pulled together and dissimilar vectors are pushed apart. Schroff et al. (2015) proposed the triplet loss for face recognition with an online triplet mining scheme, which enables the positive similarity to be larger than the negative similarity with a pre-defined margin. Besides, as a type of quadruplet loss, the histogram loss (Ustinova & Lempitsky, 2016) can make the distributions of positive and negative similarities less overlapping without any additional hyper-parameters. Given N training samples, there are $O(N^2)$ pairs, $O(N^3)$ triplets, and $O(N^4)$ quadruplets, and it is impracticable to traverse all these training tuples during training. Therefore, representative samples selecting plays a key role in metric learning. Several strategies to select representative samples have been discussed by many scholars, such as smart mining (Harwood et al., 2017) and dynamic sampling (Ge, 2018).

However, the above metric learning approaches mainly concentrate on unimodal related tasks, which cannot precisely capture the relations between different modalities. Cross-modal retrieval is generally driven by metric learning, because it needs to measure the similarity among samples of different modalities. In recent years, more literature on cross-modal metric learning appears. Frome et al. (2013) used an unweighted triplet loss to ensure semantically similar examples closer to one other by projecting images and texts into a common embedding space. Faghri et al. (2017) proposed a hard-triplet loss via mining the hardest negatives within a mini-batch during training. Liong et al. (2016) proposed a dual perceptron network, which learns two groups of nonlinear transformations for multi-modal features. However, the above works cannot accurately distinguish samples according to their importance, and thus result in slow convergence and poor performance.

As revealed by recent works (Chen et al., 2020; Wei, Xu et al., 2021; Wei, Yang et al., 2021), a proper weighting strategy can further improve the performance. Chen et al. (2020) proposed a quintuple loss to adaptively penalize cross-modal similarity by sampling negatives offline from the whole training set. Wei, Yang et al. (2021) proposed the Universal Weighting Framework (UWF), in which the polynomial loss is used to choose informative negative pairs and then provides accurate weights for various pairs. Wei, Xu et al. (2021) designed the Meta Self-Paced Network (MSPN) that fully considers the potential interactions among different similarities, and automatically estimates the weights for them.

In summary, there are two critical problems in the process of cross-modal similarity optimizing: (1) How to select the most representative samples to accelerate convergence? (2) How to assign accurate weights for various samples corresponding to their importance to improve performance? In this paper, to better cope with the above two problems simultaneously, we present a novel multi-spring balance loss to optimize the cross-modal similarity more accurately using a unified framework, in which two modules are integrated. Specifically, the first module is samples selecting, and the second module is weights assigning.

3. Proposed method

In this section, we elaborate the proposed M^{2}HSE method. Firstly, the problem formulation and the overall framework are presented in Sections 3.1 and 3.2, respectively. Then, we introduce the multi-modal and multi-grained feature encoders in Section 3.3. Afterwards, two cross-modal similarity calculation modules are described detailedly in Section 3.4. Finally, the proposed MSB loss for cross-modal similarity optimization is explained in Section 3.5.

3.1. Problem formulation

Suppose that there are N labeled multi-modal documents with Q modalities denoted as $\mathcal{D} = \{\boldsymbol{s}_i^1, \boldsymbol{s}_i^2, \dots, \boldsymbol{s}_i^Q\}_{i=1}^N$, where each document $D_i = \{\boldsymbol{s}_i^1, \boldsymbol{s}_i^2, \dots, \boldsymbol{s}_i^Q\}$ contains Q samples that describe the same semantic concept, and each sample belongs to a unique modality. Formally, the generalized problem definition for cross-modal retrieval is as follows:

Definition 1 (*Cross-modal Retrieval*)**.** Suppose that a sample $\boldsymbol{s}_i^a$ from the ath modality is regarded as a query, the goal of the cross-modal retrieval is to search the most relevant results from the bth modality, where $1 \le a, b \le Q$ and $a \ne b$.

In this paper, the multi-modal documents are divided into two subsets. One is the **primary modality** set denoted as $\mathcal{PM} = \{\boldsymbol{s}_i^a, \boldsymbol{s}_i^b\}_{i=1}^N$, which provides candidates from two modalities to perform the cross-modal retrieval. The other is the **auxiliary modality** set denoted as $\mathcal{AM} = \{\boldsymbol{s}_i^1, \dots, \boldsymbol{s}_i^{a-1}, \boldsymbol{s}_i^{a+1}, \dots, \boldsymbol{s}_i^{b-1}, \boldsymbol{s}_i^{b+1}, \dots, \boldsymbol{s}_i^Q\}_{i=1}^N$, which contains samples from the remaining $Q-2$ modalities. Then, three related matrices are defined as follows:

Definition 2 (*Primary Similarity Matrix* $\boldsymbol{P}$)**.** The matrix consists of similarities between samples from different modalities in $\mathcal{PM}$. For example, P_{ij} represents the cross-modal similarity between $\boldsymbol{s}_i^a$ and $\boldsymbol{s}_j^b$.

Definition 3 (*Auxiliary Similarity Matrix* $\boldsymbol{A}$)**.** The matrix consists of similarities between samples from different modalities in $\mathcal{PM}$ and $\mathcal{AM}$. Note that, A_{ij} represents the cross-modal similarity between the ith sample of a modality in $\mathcal{PM}$ and the jth sample of a modality in $\mathcal{AM}$.

Definition 4 (*Cross-modal Affinity Matrix* $\boldsymbol{C}$)**.** The matrix contains the supervision information for cross-modal retrieval. If the ith sample and the jth sample belong to different modalities, and they represent the same semantic concept, $C_{ij} = +1$, otherwise, $C_{ij} = -1$.

In general, baseline methods mainly concentrate on the mutual retrieval between two modalities in $\mathcal{PM}$, and they only consider how to obtain the primary similarity matrix $\boldsymbol{P}$ to perform cross-modal retrieval. However, the complete information of a semantic concept distributes in all modalities, which means that the valuable semantic information in $\mathcal{AM}$ is ignored by existing methods.

Therefore, we innovatively propose to calculate the auxiliary similarity matrix $\boldsymbol{A}$, which is jointly optimized with the primary similarity matrix $\boldsymbol{P}$ to capture more cross-modal correlations. Actually, the goal of our work is to enhance the performance of cross-modal retrieval within $\mathcal{PM}$ by leveraging the semantic knowledge gleaned from $\mathcal{AM}$.

3.2. Framework of M^{2}HSE

We aim to mine and fuse the complementarity in multi-modal and multi-grained data through the proposed M^{2}HSE, which contains a global-level subnetwork and a local-level subnetwork. In each subnetwork, we can calculate one primary similarity matrix and $2(Q-2)$ auxiliary similarity matrices. Afterwards, these matrices are jointly optimized via our proposed MSB loss with the supervision of the cross-modal affinity matrix.

Specifically, the global-level objective function $\mathcal{J}^G$ and the local-level objective function $\mathcal{J}^L$ are defined as follows:

$$\begin{aligned} \mathcal{J}^G &= \mathcal{H}(\boldsymbol{P}^G, \boldsymbol{\Theta}^G) + \sum_{i=1}^{2(Q-2)} \alpha_i \mathcal{H}(\boldsymbol{A}_i^G, \boldsymbol{\Theta}^G) \\ \mathcal{J}^L &= \mathcal{H}(\boldsymbol{P}^L, \boldsymbol{\Theta}^L) + \sum_{i=1}^{2(Q-2)} \beta_i \mathcal{H}(\boldsymbol{A}_i^L, \boldsymbol{\Theta}^L) \end{aligned} \tag{1}$$

where $\mathcal{H}(\cdot)$ is the MSB loss, $\boldsymbol{P}^G$ is the global-level primary similarity matrix, $\{\boldsymbol{A}_i^G \mid i = 1, \dots, 2(Q-2)\}$ are the global-level auxiliary similarity

matrices, $\boldsymbol{P}^L$ is the local-level primary similarity matrix, $\{\boldsymbol{A}_i^L \mid i = 1,\dots,2(Q-2)\}$ are the local-level auxiliary similarity matrices. Besides, $\boldsymbol{\Theta}^G$ and $\boldsymbol{\Theta}^L$ denote the parameters of two subnetworks, and the hyper-parameters $\{\alpha_i \mid i = 1,\dots,2(Q-2)\}$ and $\{\beta_i \mid i = 1,\dots,2(Q-2)\}$ denote the multi-modal complementarity adjustment factors in two subnetworks.

In this paper, cross-modal image–text retrieval is adopted to prove the effectiveness of M^{2}HSE, where CNN features of images and Bi-GRU features of texts are determined as the primary modalities. Learned from Wang et al. (2015) and Zheng et al. (2017), the SIFT-BoVW is one kind of hand-crafted visual features which actually characterizes the different aspects of images. Inspired by this theory, the SIFT-BoVW features can be completely regarded as the auxiliary modality. Consequently, we utilize the above three modalities in our work, *i.e.*, $Q = 3$, then Eq. (1) can be accordingly simplified as follows:

$$\begin{aligned}\mathcal{J}^G &= \mathcal{H}(\boldsymbol{P}^G,\boldsymbol{\Theta}^G) + \alpha_1\mathcal{H}(\boldsymbol{A}_1^G,\boldsymbol{\Theta}^G) + \alpha_2\mathcal{H}(\boldsymbol{A}_2^G,\boldsymbol{\Theta}^G)\\ \mathcal{J}^L &= \mathcal{H}(\boldsymbol{P}^L,\boldsymbol{\Theta}^L) + \beta_1\mathcal{H}(\boldsymbol{A}_1^L,\boldsymbol{\Theta}^L) + \beta_2\mathcal{H}(\boldsymbol{A}_2^L,\boldsymbol{\Theta}^L)\end{aligned} \tag{2}$$

where $\boldsymbol{P}^G$, $\boldsymbol{A}_1^G$, $\boldsymbol{A}_2^G$ are calculated with three kinds of coarse-grained features by the global-level cross-modal similarity calculation module (GCS), which is discussed in Section 3.4.1, and $\boldsymbol{P}^L$, $\boldsymbol{A}_1^L$, $\boldsymbol{A}_2^L$ are calculated with three kinds of fine-grained features by the local-level cross-modal similarity calculation module (LCS), which is discussed in Section 3.4.2. By minimizing the objective function $\mathcal{J}^G$ and $\mathcal{J}^L$, optimal parameters of two subnetworks are estimated as follows:

$$\begin{aligned}\tilde{\boldsymbol{\Theta}}^G &= arg\min_{\boldsymbol{\Theta}^G}\mathcal{J}^G\\ \tilde{\boldsymbol{\Theta}}^L &= arg\min_{\boldsymbol{\Theta}^L}\mathcal{J}^L\end{aligned} \tag{3}$$

Then the optimal primary similarity matrices $\tilde{\boldsymbol{P}}^G$ and $\tilde{\boldsymbol{P}}^L$ can be obtained. In practice, we use the gradient descent method (Ruder, 2016) to accomplish the above optimization procedures of two subnetworks, respectively. The gradients with respect to parameters $\boldsymbol{\Theta}^G$ and $\boldsymbol{\Theta}^L$ are denoted as follows:

$$\begin{aligned}\frac{\partial\mathcal{J}^G}{\partial\boldsymbol{\Theta}^G} &= \frac{\partial\mathcal{H}(\boldsymbol{P}^G,\boldsymbol{\Theta}^G)}{\partial\boldsymbol{\Theta}^G} + \alpha_1\frac{\partial\mathcal{H}(\boldsymbol{A}_1^G,\boldsymbol{\Theta}^G)}{\partial\boldsymbol{\Theta}^G} + \alpha_2\frac{\partial\mathcal{H}(\boldsymbol{A}_2^G,\boldsymbol{\Theta}^G)}{\partial\boldsymbol{\Theta}^G}\\ \frac{\partial\mathcal{J}^L}{\partial\boldsymbol{\Theta}^L} &= \frac{\partial\mathcal{H}(\boldsymbol{P}^L,\boldsymbol{\Theta}^L)}{\partial\boldsymbol{\Theta}^L} + \beta_1\frac{\partial\mathcal{H}(\boldsymbol{A}_1^L,\boldsymbol{\Theta}^L)}{\partial\boldsymbol{\Theta}^L} + \beta_2\frac{\partial\mathcal{H}(\boldsymbol{A}_2^L,\boldsymbol{\Theta}^L)}{\partial\boldsymbol{\Theta}^L}\end{aligned} \tag{4}$$

For simplicity, we take the global-level subnetwork as an example to illustrate the optimization procedure of parameters $\boldsymbol{\Theta}^G$ in Algorithm 1.

Finally, in order to investigate the complementary relationship between multi-grained data, a linear weighted fusion strategy is utilized to generate the final primary similarity matrix:

$$\begin{aligned}&\tilde{\boldsymbol{P}} = \theta_1\tilde{\boldsymbol{P}}^G + \theta_2\tilde{\boldsymbol{P}}^L\\ &s.t.\ 0 \le \theta_1,\theta_2 \le 1\end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{P}}$ is finally used to perform the cross-modal retrieval, and the hyper-parameters θ_1,θ_2 denote the multi-grained complementarity adjustment factors.

The framework of our proposed M^{2}HSE method is illustrated in Fig. 2. There are ten image–text pairs belonging to three semantic concepts (*i.e.*, "airplane", "dog", "bus"), and the colors of them are represented by "green", "orange", "blue" respectively. For each cross-modal similarity matrix, the first four rows and columns represent the "airplane", the middle three rows and columns denote the "dog", and the last three rows and columns refer to the "bus". Then we rank all elements in each row according to their values, and only the top ranked instances that belong to the correct semantic concept are displayed with the above pre-defined colors. Like completing a jigsaw puzzle, three diagonal blocks can be clearly seen in matrix $\tilde{\boldsymbol{P}}$ after two fusion operations.

Note that $\tilde{\boldsymbol{P}}$ is the result of semantic enhancement, which not only comprehensively contains the complementary semantic information existing in multi-modal and multi-grained data, but also is optimized accurately via the proposed MSB loss.

Algorithm 1 The optimization procedure of the global-level subnetwork in M^{2}HSE.

Input: The training set $\Omega = \{(I_i,T_i)\}_{i=1}^{N^t}$, the validation set $\Phi = \{(I_i,T_i)\}_{i=1}^{N^v}$, the number of epochs E, the batch size B, the learning rate η and the hyper-parameters α_1, α_2, γ_1, γ_2.

Output: The optimized parameters $\tilde{\boldsymbol{\Theta}}^G$ of the global-level subnetwork.

```
Initialize parameters Θ^G of the global-level network and set mAP_max = 0;
for δ = 1, 2, ..., E do
    for ρ = 1, 2, ..., ⌈N^t/B⌉ do
        Randomly sample B image–text pair from Ω to construct a mini-batch with C;
        Compute {x_i^G, y_i^G, z_i^G}_{i=1}^B, then construct P^G, A_1^G, A_2^G;
        Compute the result of J^G in Eq. (2), and calculate the gradients ∇_{Θ^G} according to Eq. (4);
        Update the parameters Θ^G through: Θ^G ← Θ^G − η∇_{Θ^G};
    end for
    Validate the performance of the current model on Φ to obtain mAP_curr;
    if mAP_curr > mAP_max then
        mAP_max = mAP_curr;
        Θ̃^G = Θ^G;
    end if
end for
```

3.3. Feature encoders

Suppose that an image–text dataset $\{(I_i,T_j)\}_{i,j=1}^{N}$ consists of N pairs, where I_i and T_j are the ith image and the jth text respectively. We first extract three kinds of coarse-grained and fine-grained features. Then we encode them as feature vectors in a common embedding space.

3.3.1. Primary modality 1: CNN

Pre-trained convolutional neural networks (CNNs) with large scale datasets like ImageNet (Deng et al., 2009), have been highlighted to be effective in catching the discriminative information for images.

In our work, the coarse-grained feature vector of image I_i is denoted as $\boldsymbol{v}_i^G \in \mathbb{R}^{d_1}$, and a set of fine-grained feature vectors of I_i is represented as $\boldsymbol{V}_i^L = \{\boldsymbol{v}_{i,p}^L \mid p = 1,\dots,n, \boldsymbol{v}_{i,p}^L \in \mathbb{R}^{d_1}\}$. Note that the CNN feature space has d_1 dimensions and each image has n patches. In the global-level subnetwork, the coarse-grained feature vector of image I_i is projected into a d-dimensional embedding space via the fully-connect layer:

$$\boldsymbol{x}_i^G = \boldsymbol{W}^G * \boldsymbol{v}_i^G + \boldsymbol{b}^G \tag{6}$$

Similarly, in the local-level subnetwork, the fine-grained feature vector of the jth patch in image I_i is also projected into a d-dimensional embedding space as follows:

$$\boldsymbol{x}_{i,p}^L = \boldsymbol{W}^L * \boldsymbol{v}_{i,p}^L + \boldsymbol{b}^L \tag{7}$$

where $\boldsymbol{W}^G$, $\boldsymbol{W}^L$ and $\boldsymbol{b}^G$, $\boldsymbol{b}^L$ refer to the weight matrices and bias terms that are required to be optimized.

Thus, we obtain the coarse-grained CNN feature vector $\boldsymbol{x}_i^G \in \mathbb{R}^d$, and a set of fine-grained CNN feature vectors $\boldsymbol{X}_i^L = \{\boldsymbol{x}_{i,p}^L \mid p = 1,\dots,n, \boldsymbol{x}_{i,p}^L \in \mathbb{R}^d\}$ for image I_i.

3.3.2. Primary modality 2: Bi-GRU

Recently, a series of word embedding technologies, such as Word2vec (Mikolov et al., 2013), RNN (Hochreiter & Schmidhuber, 1997), have been introduced to model textual features, and perform amazingly in a wide range of tasks.

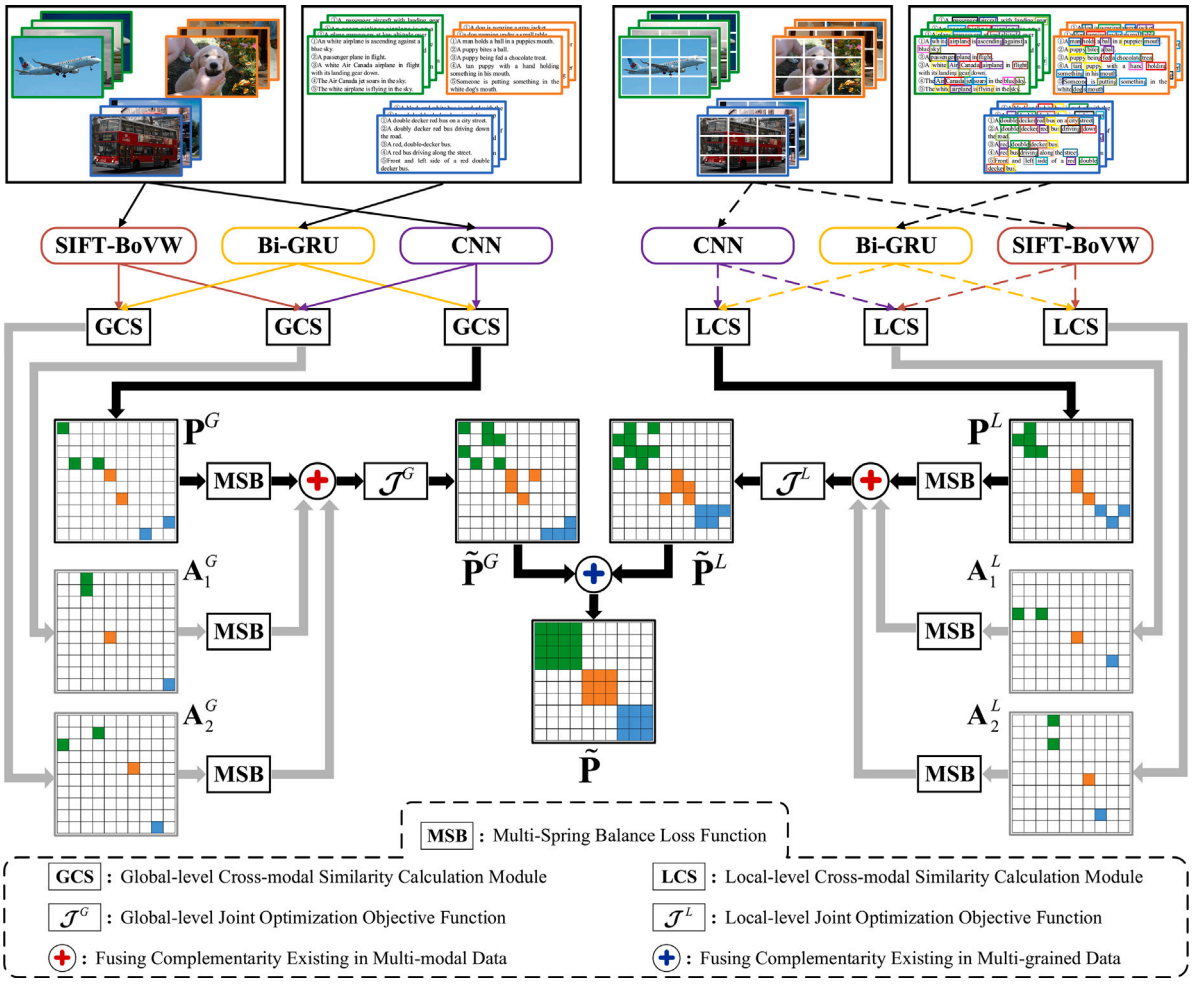


Fig. 2. The overall framework of our proposed M^{2}HSE method. It includes a global-level subnetwork (left half) and a local-level subnetwork (right half) to mine and fuse the complementarity in multi-modal and multi-grained data for semantic enhancement. GCS and LCS are designed to calculate the cross-modal similarity matrix with coarse-grained and fine-grained features, respectively. Besides, $\mathcal{J}^G$ and $\mathcal{J}^L$ are defined to jointly optimize the primary similarity and auxiliary similarity based on the proposed MSB loss.

Therefore, as a variant of RNNs, the Gated Recurrent Unit (GRU) (Cho et al., 2014) is adopted to fully mine the context information of words during learning the word embeddings. For text T_j, we suppose that it is composed of m words, and the qth word in T_j is represented as a d'-dimensional one-hot vector $\boldsymbol{e}_{j,q}$, $\forall q \in [1, m]$. Specifically, d' refers to the total number of words in the dictionary, and just the position that exactly corresponds to the word is set 1 in $\boldsymbol{e}_{j,q}$, while other positions are set 0. Next, each word is embedded into a continuous space by a mapping matrix $\boldsymbol{W}^e$: $\boldsymbol{w}_{j,q} = \boldsymbol{W}^e \boldsymbol{e}_{j,q}$, and $\boldsymbol{w}_{j,q} \in \mathbb{R}^{d_2}$ is the embedding vector for the qth word in text T_j with d_2 dimensions.

To directly compare image and text features, we also map the text features into a d-dimensional embedding space. The Bi-directional GRU (Bi-GRU) is adopted to model textual context in text T_j in two different directions, and $\overrightarrow{\boldsymbol{h}}_{j,q}$, $\overleftarrow{\boldsymbol{h}}_{j,q}$ are used to indicate the GRU's hidden states (forward and backward) as follows:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}}_{j,q} &= \overrightarrow{GRU}(\boldsymbol{w}_{j,q}, \overrightarrow{\boldsymbol{h}}_{j,q-1}) \\ \overleftarrow{\boldsymbol{h}}_{j,q} &= \overleftarrow{GRU}(\boldsymbol{w}_{j,q}, \overleftarrow{\boldsymbol{h}}_{j,q+1}) \end{aligned} \tag{8}$$

Afterwards, the feature vector $\boldsymbol{y}_{j,q}$ for the qth word in text T_j is computed with $\boldsymbol{y}_{j,q} = (\overrightarrow{\boldsymbol{h}}_{j,q} + \overleftarrow{\boldsymbol{h}}_{j,q})/2$.

Thus, in the global-level subnetwork, the coarse-grained Bi-GRU feature vector of text T_j is generated via averaging all word vectors, that is, $\boldsymbol{y}_j^G = \frac{1}{m}\sum_{q=1}^{m} \boldsymbol{y}_{j,q}^G$. In the local-level subnetwork, a set of fine-grained Bi-GRU feature vectors for T_j is represented as $\boldsymbol{Y}_j^L = \{\, \boldsymbol{y}_{j,q}^L \,|\, q = 1, \ldots, m, \boldsymbol{y}_{j,q}^L \in \mathbb{R}^d \}$.

3.3.3. Auxiliary modality: SIFT-BoVW

Scale Invariant Feature Transform (SIFT) (Lowe, 2004) is a kind of local feature, and it has been widely and successfully utilized in image processing. In our work, the SIFT descriptor is integrated with the Bag-of-Visual-Words (BoVW) (Csurka et al., 2004) model, and the process of extracting the SIFT-based BoVW (SIFT-BoVW) features is divided into the following two steps.

Step 1: Codebook construction. Firstly, a vocabulary is established through clustering visual features of images. Secondly, the codebook model is constructed by describing an image as a bag of visual words that are chosen from the vocabulary, where the occurrence frequency of visual words can represent visual features of images. Specifically, we suppose that a pool of 128-d SIFT descriptors $\{\boldsymbol{f}_k\}_{k=1}^{D}$ are extracted from the training set, where D is the total number of them. K-means is used to partition the D points into K clusters, and the centroid of each cluster is corresponding to a visual word.

Step 2: Feature extraction. Firstly, SIFT descriptors of image I_i and its n patches are extracted, which are denoted as $\{\boldsymbol{f}_k\}_{k=1}^{g}$ and $\{\{\boldsymbol{f}_k\}_{k=1}^{l_p} \,|\, p = 1, \ldots, n\}$ respectively, where $\{\boldsymbol{f}_k\}_{k=1}^{l_p}$ are extracted from the pth patch of I_i. Secondly, $\{\boldsymbol{f}_k\}_{k=1}^{g}$ are encoded according to their distances to the visual words in the codebook, and the SIFT descriptor is represented as its nearest visual word. Then, the occurrences of visual words in I_i are used to construct a histogram, which represents the d_3-dimensional ($d_3 = K$) coarse-grained feature vector $\boldsymbol{b}_i^G \in \mathbb{R}^{d_3}$. Similarly, for $\{\boldsymbol{f}_k\}_{k=1}^{l_p}$, the corresponding fine-grained feature vector $\boldsymbol{b}_{i,p}^L \in \mathbb{R}^{d_3}$

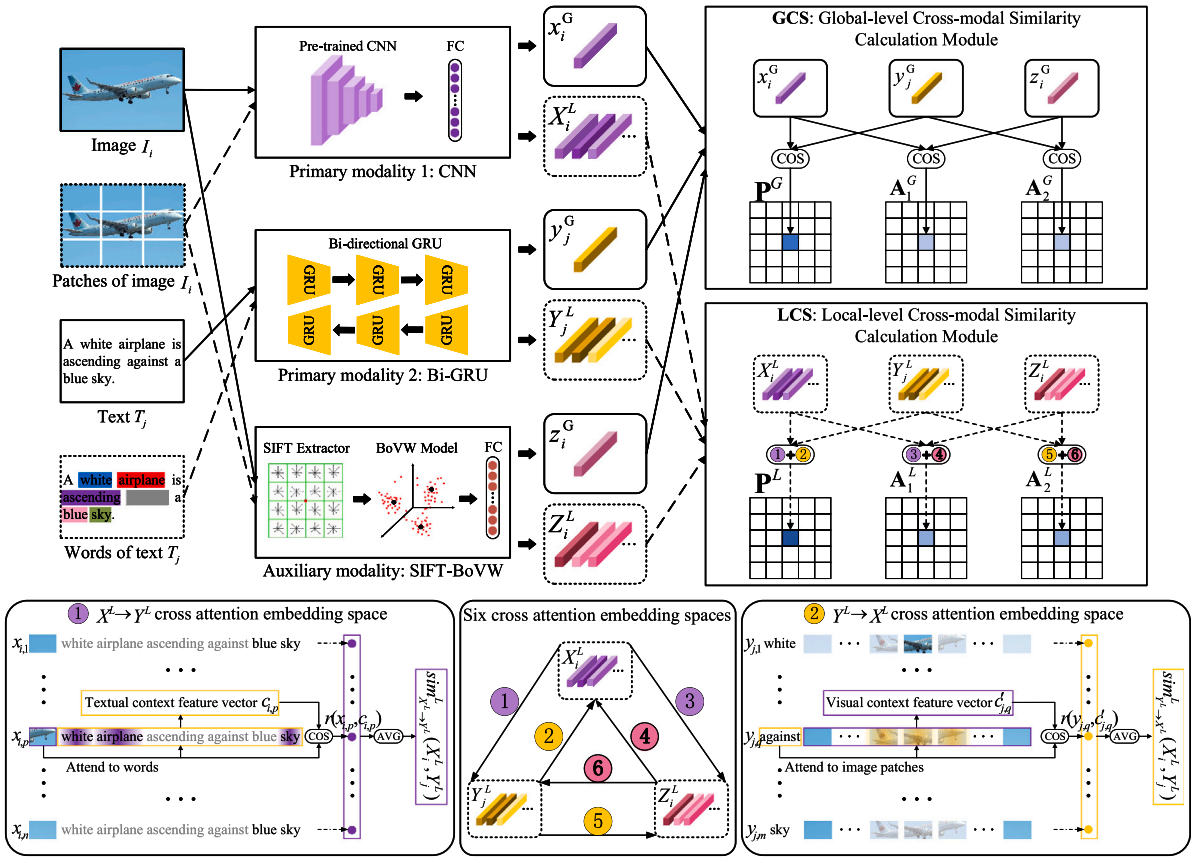


Fig. 3. The flowchart of cross-modal similarity calculation. GCS and LCS are based on three coarse-grained features (*i.e.*, $\boldsymbol{x}_i^G, \boldsymbol{y}_j^G, \boldsymbol{z}_i^G$) and three fine-grained features (*i.e.*, $\boldsymbol{X}_i^L, \boldsymbol{Y}_j^L, \boldsymbol{Z}_i^L$), respectively. In LCS, there are six cross attention embedding spaces in pairs: ① $\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L$ and ② $\boldsymbol{Y}^L \rightarrow \boldsymbol{X}^L$, ③ $\boldsymbol{X}^L \rightarrow \boldsymbol{Z}^L$ and ④ $\boldsymbol{Z}^L \rightarrow \boldsymbol{X}^L$, ⑤ $\boldsymbol{Y}^L \rightarrow \boldsymbol{Z}^L$ and ⑥ $\boldsymbol{Z}^L \rightarrow \boldsymbol{Y}^L$, which are constructed to capture more fine-grained correlations between multi-modal data. Details of ① $\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L$ and ② $\boldsymbol{Y}^L \rightarrow \boldsymbol{X}^L$ are shown to illustrate the calculation process of local-level primary similarity, and the remaining four cross attention embedding spaces are similar.

can also obtained. Thus, a set of fine-grained feature vectors of image I_i is denoted as $\boldsymbol{B}_i^L = \{\, \boldsymbol{b}_{i,p}^L \,|\, p = 1, \ldots, n, \boldsymbol{b}_{i,p}^L \in \mathbb{R}^{d_3} \}$.

Then, just like Eqs. (6) and (7), fully-connect layers are adapted to map these features into a d-dimensional embedding space. Afterwards, we obtain the coarse-grained SIFT-BoVW feature vector $\boldsymbol{z}_i^G \in \mathbb{R}^d$ and a set of fine-grained SIFT-BoVW feature vectors $\boldsymbol{Z}_i^L = \{\, \boldsymbol{z}_{i,p}^L \,|\, p = 1, \ldots, n, \boldsymbol{z}_{i,p}^L \in \mathbb{R}^d \}$ for image I_i.

3.4. Cross-modal similarity calculation

There are two modules, GCS and LCS, which are designed to comprehensively calculate the cross-modal similarity with multi-modal and multi-grained features. In each module, we calculate one primary similarity matrix and two auxiliary similarity matrices. Notably, to learn more precise local-level correlations between multi-modal data, the attention mechanism is adopted in LCS to fully aggregate the fine-grained matches between visual patches and words.

As shown in Fig. 3, given image I_i and text T_j, we illustrate how to calculate the cross-modal similarity at the global-level and the local-level, respectively. Moreover, to make a clear distinction between different granularities, the data streams of coarse-grained and fine-grained instances are represented as solid arrows and dotted arrows, respectively.

3.4.1. GCS: global-level cross-modal similarity calculation module

Three kinds of coarse-grained feature vectors (*i.e.*, $\boldsymbol{x}^G, \boldsymbol{y}^G$ and $\boldsymbol{z}^G$) are input to GCS. Specifically, as the element at the ith row and jth column of the global-level primary similarity matrix, $\boldsymbol{P}^G(i,j)$ is calculated as follows:

$$P^G(i,j) = sim^G(\boldsymbol{x}_i^G, \boldsymbol{y}_j^G) = \frac{\boldsymbol{x}_i^{G\top} \boldsymbol{y}_j^G}{\|\boldsymbol{x}_i^G\| \cdot \|\boldsymbol{y}_j^G\|} \tag{9}$$

where $\boldsymbol{x}_i^G$ and $\boldsymbol{y}_j^G$ represent the coarse-grained CNN feature vector of I_i, and the coarse-grained Bi-GRU feature vector of T_j, respectively.

Similarly, the elements at the ith row and jth column of the global-level auxiliary matrices $A_1^G(i,j)$ and $A_2^G(i,j)$ are defined as follows:

$$\begin{aligned} A_1^G(i,j) &= sim^G(\boldsymbol{x}_i^G, \boldsymbol{z}_j^G) = \frac{\boldsymbol{x}_i^{G\top} \boldsymbol{z}_j^G}{\|\boldsymbol{x}_i^G\| \cdot \|\boldsymbol{z}_j^G\|} \\ A_2^G(i,j) &= sim^G(\boldsymbol{y}_i^G, \boldsymbol{z}_j^G) = \frac{\boldsymbol{y}_i^{G\top} \boldsymbol{z}_j^G}{\|\boldsymbol{y}_i^G\| \cdot \|\boldsymbol{z}_j^G\|} \end{aligned} \tag{10}$$

where $\boldsymbol{z}_j^G$ represents the coarse-grained SIFT-BoVW feature vector of I_j.

3.4.2. LCS: local-level cross-modal similarity calculation module

Correspondingly, three kinds of fine-grained feature vectors (*i.e.*, $\boldsymbol{X}^L, \boldsymbol{Y}^L$ and $\boldsymbol{Z}^L$) are input to LCS. Inspired by SCAN (Lee et al., 2018), we construct some cross attention embedding spaces in pairs to measure the local-level cross-modal similarity. Instead of learning different spaces independently like SCAN, we jointly learn six cross attention embedding spaces to capture the latent alignments between any two kinds of fine-grained features.

Specifically, $P^L(i,j)$ represents the element at the ith row and the jth column of the local-level primary similarity matrix, which is computed with two cross attention embedding spaces (*i.e.*, ① $\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L$

and ② $\boldsymbol{Y}^L \rightarrow \boldsymbol{X}^L$) as follows:

$$\begin{aligned} P^L(i,j) &= sim^L(\boldsymbol{X}_i^L, \boldsymbol{Y}_j^L) \\ &= sim^L_{\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L}(\boldsymbol{X}_i^L, \boldsymbol{Y}_j^L) + sim^L_{\boldsymbol{Y}^L \rightarrow \boldsymbol{X}^L}(\boldsymbol{X}_i^L, \boldsymbol{Y}_j^L) \end{aligned} \tag{11}$$

where $\boldsymbol{X}_i^L$ is a set of fine-grained CNN feature vectors of I_i, and $\boldsymbol{Y}_j^L$ is a set of fine-grained Bi-GRU feature vectors of T_j.

Similarly, the elements at the ith row and the jth column of the local-level auxiliary matrices $\boldsymbol{A}_1^L$ and $\boldsymbol{A}_2^L$ are calculated as follows:

$$\begin{aligned} A_1^L(i,j) &= sim^L(\boldsymbol{X}_i^L, \boldsymbol{Z}_j^L) \\ &= sim^L_{\boldsymbol{X}^L \rightarrow \boldsymbol{Z}^L}(\boldsymbol{X}_i^L, \boldsymbol{Z}_j^L) + sim^L_{\boldsymbol{Z}^L \rightarrow \boldsymbol{X}^L}(\boldsymbol{X}_i^L, \boldsymbol{Z}_j^L) \\ A_2^L(i,j) &= sim^L(\boldsymbol{Y}_i^L, \boldsymbol{Z}_j^L) \\ &= sim^L_{\boldsymbol{Y}^L \rightarrow \boldsymbol{Z}^L}(\boldsymbol{Y}_i^L, \boldsymbol{Z}_j^L) + sim^L_{\boldsymbol{Z}^L \rightarrow \boldsymbol{Y}^L}(\boldsymbol{Y}_i^L, \boldsymbol{Z}_j^L) \end{aligned} \tag{12}$$

where $\boldsymbol{Z}_j^L$ is a set of fine-grained SIFT-BoVW feature vectors of the I_j.

There are six types of cross-modal similarity in Eqs. (11) and (12), and each type of similarity is corresponding to a specific cross attention embedding space. For simplicity, we take $\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L$ cross attention embedding space as an example to explain how to calculate the local-level cross-modal similarity $sim^L_{\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L}(\boldsymbol{X}_i^L, \boldsymbol{Y}_j^L)$ as follows.

Specifically, given image I_i with n patches and text T_j with m words, to uncover associations among all possible pairs, a cosine similarity matrix $\boldsymbol{U}$ is initially constructed with $\boldsymbol{X}_i^L$ and $\boldsymbol{Y}_j^L$. Particularly, $U_{pq} = \frac{\boldsymbol{x}_{i,p}^{\top} \boldsymbol{y}_{j,q}}{\|\boldsymbol{x}_{i,p}\| \cdot \|\boldsymbol{y}_{j,q}\|}, \forall p \in [1,n], \forall q \in [1,m]$ denotes the similarity between the pth patch and the qth word, where $\boldsymbol{x}_{i,p}$ is the feature vector of the pth patch in I_i, and $\boldsymbol{y}_{j,q}$ is the feature vector of the qth word in T_j. Then, $\boldsymbol{U}$ is normalized according to its column dimension $\bar{U}_{pq} = \frac{relu(U_{pq})}{\sqrt{\sum_{p=1}^{n} relu(U_{pq})^2}}$, where $relu(x) = \max(0, x)$.

Next, for the pth patch of I_i, the textual context feature vector $\boldsymbol{c}_{i,p} = \sum_{q=1}^{m} \alpha_{pq} \boldsymbol{y}_{j,q}$ is defined as a weighted integration with representations of words using the attention mechanism, where $\alpha_{pq} = \frac{exp(\lambda \bar{U}_{pq})}{\sum_{q=1}^{m} exp(\lambda \bar{U}_{pq})}$ is satisfied. Especially, λ is the temperature-inverse parameter in the softmax function to adjust the smoothness of the attention distribution. Afterwards, to assess the significance of each image patch given the text context, we compute the relevance score between $\boldsymbol{x}_{i,p}$ and $\boldsymbol{c}_{i,p}$ by cosine function $r(\boldsymbol{x}_{i,p}, \boldsymbol{c}_{i,p}) = \frac{\boldsymbol{x}_{i,p}^{\top} \boldsymbol{c}_{i,p}}{\|\boldsymbol{x}_{i,p}\| \cdot \|\boldsymbol{c}_{i,p}\|}$. Consequently, the similarity between $\boldsymbol{X}_i^L$ and $\boldsymbol{Y}_j^L$ is obtained by averaging all relevance scores:

$$sim^L_{\boldsymbol{X}^L \rightarrow \boldsymbol{Y}^L}(\boldsymbol{X}_i^L, \boldsymbol{Y}_j^L) = \frac{1}{n} \sum_{p=1}^{n} r(\boldsymbol{x}_{i,p}, \boldsymbol{c}_{i,p}) \tag{13}$$

In addition, the procedures of calculating other five types local-level cross-modal similarities are analogous to the above. Finally, we obtain three global-level cross-modal similarity matrices $\boldsymbol{P}^G$, $\boldsymbol{A}_1^G$, $\boldsymbol{A}_2^G$, and three local-level cross-modal similarity matrices $\boldsymbol{P}^L$, $\boldsymbol{A}_1^L$, $\boldsymbol{A}_2^L$ with GCS and LCS, respectively.

To capture more accurate cross-modal correlations, each primary similarity matrix is jointly optimized with two auxiliary similarity matrices via the MSB loss in the corresponding subnetwork. Afterwards, the final fused primary similarity matrix is used to perform cross-modal retrieval.

3.5. Cross-modal similarity optimization

A novel multi-spring balance loss is proposed to optimize the cross-modal similarity more accurately through two steps. In step 1, we propose to select representative samples for optimizing, which ensures not only the available information is fully utilized, but also the time efficiency is greatly improved. In step 2, we design an adaptive similarity weighting strategy based on the multi-spring balance system, which can accurately discriminate the selected representative samples according to their significance.

The framework of our proposed MSB loss is illustrated in Fig. 4. Firstly, four positive samples and five negative samples are selected from the critical area (see Definition 5), and the unselected samples no longer need to be optimized. Then, the selected positive and negative samples are used to construct two kinds of multi-spring balance systems, in which the elastic coefficient of each spring is estimated with the cross-modal similarity between the anchor and the selected sample. Finally, the selected positive and negative samples are drawn close to or pushed away from the anchor according to the adaptive weight values, which are obtained when the multi-spring balance systems reach the equilibrium states. The following are the details of the two steps mentioned above.

Step 1: Representative Samples Selecting. In cross-modal retrieval, given an image/text query as the *anchor*, text/image candidates are treated as *positive* samples when they belong to the same semantic concept with the anchor, otherwise, they are regarded as *negative* samples. We consider that both positive samples with higher similarity to anchors and negative samples with smaller similarity to anchors contain less information for optimization (the parameters of model have been appropriately fitted to them). For each anchor, we propose to select informative positive and negative samples to optimize during the training stage.

Formally, suppose that there is a cross-modal similarity matrix $\boldsymbol{M}$ with the corresponding cross-modal affinity matrix $\boldsymbol{C}$. If $C_{ij} = +1$, let M_{ij}^+ denote the similarity between anchor $\boldsymbol{s}_i$ and positive sample $\boldsymbol{s}_j^+$, otherwise, if $C_{ij} = -1$, the similarity between $\boldsymbol{s}_i$ and the negative sample $\boldsymbol{s}_j^-$ is represented as M_{ij}^-. As shown in Fig. 4, for a fixed anchor $\boldsymbol{s}_i$ within a mini-batch, *the hardest positive sample* is the furthest one from $\boldsymbol{s}_i$ compared to all positive samples, thus, the similarity between the hardest positive sample and $\boldsymbol{s}_i$ is lowest in all positive pairs, which is represented as H_i^+.

$$H_i^+ = \min_{C_{ik}=+1} M_{ik}^+ \tag{14}$$

Besides, *the hardest negative sample* is the nearest one from $\boldsymbol{s}_i$ compared to all negative samples, which has the highest similarity H_i^- with $\boldsymbol{s}_i$ in all negative pairs.

$$H_i^- = \max_{C_{ik}=-1} M_{ik}^- \tag{15}$$

Utilizing the hardest positive and negative samples, the definition of the *critical area* is given as follows.

Definition 5 (*Critical Area*)**.** For each anchor $\boldsymbol{s}_i$, the outer and inner boundaries of the critical area are determined by the hardest positive and negative samples, respectively. For the positive sample $\boldsymbol{s}_j^+$ in the critical area, the similarity M_{ij}^+ satisfies the condition: $M_{ij}^+ < H_i^-$, otherwise, for the negative sample $\boldsymbol{s}_j^-$ in the critical area, the similarity M_{ij}^- satisfies the condition: $M_{ij}^- > H_i^+$.

We propose to select samples from the critical area, which means that unselected samples outside the critical area will be filtered out. Afterwards, considering the critical area, we provide the redefinition of the cross-modal affinity matrix $\tilde{\boldsymbol{C}}$ as follows:

$$\tilde{C}_{ij} = \begin{cases} +1 & \text{if } \boldsymbol{s}_j \text{ is a positive sample in the critical area,} \\ -1 & \text{if } \boldsymbol{s}_j \text{ is a negative sample in the critical area,} \\ 0 & \text{if } \boldsymbol{s}_j \text{ is not in the critical area.} \end{cases} \tag{16}$$

where for the anchor $\boldsymbol{s}_i$, the affinity $\tilde{C}_{ij}$ between it and a sample $\boldsymbol{s}_j$ outside the critical area is changed into 0, whether $\boldsymbol{s}_j$ is positive or negative. Furthermore, in step 2, we discard the samples corresponding to all zero-valued elements in $\tilde{\boldsymbol{C}}$, and only select the representative samples in the critical area for optimization.

Step 2: Representative Samples Discriminating. As a visualization technology, the implementation principle of RadViz (Hoffman et al., 1997) is inspired by the multi-spring balance system, which follows the elastic force balance theorem of objects in physics. As shown in Fig. 5, there are n springs uniformly locating at a circle in a two-dimensional coordinate system. We suppose that the original

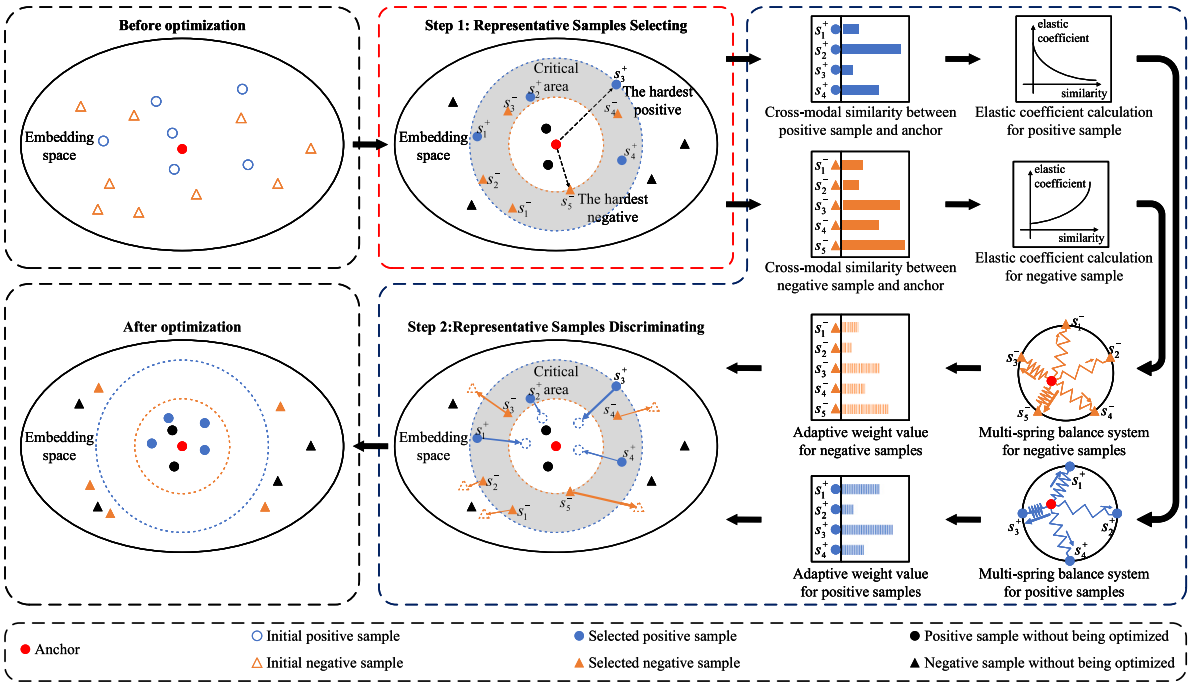


Fig. 4. Framework of the MSB loss for optimizing the cross-modal similarity. For each anchor, the representative samples are selected from the critical area in Step 1, then they are accurately discriminated based on two multi-spring balance systems in Step 2. For the selected samples, the forces of their corresponding springs and the penalty strengths for them are presented as different colorful arrows, where blue is for positive samples, while orange is for negative samples. Note that the length and width of the arrows approximately give the magnitude of the force and strength of the penalty.

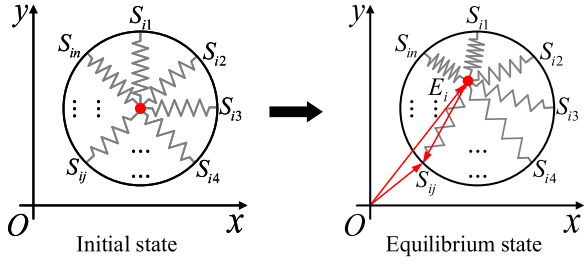


Fig. 5. Schematic diagram of the multi-spring balance system. The elastic force of all springs makes the point continuously move until it reaches the equilibrium state, then the deformation of each spring can be estimated according to the vector operation principle.

length of each spring is zero and the other ends of all springs are fixed to the same point, which is initially located at the center of the circle. Given the ith point, the position vector and the elastic coefficient of the jth spring are represented as $\boldsymbol{S}_{ij}$ and K_{ij}, respectively. Then, the elastic force of all springs makes the ith point continuously move until it reaches the equilibrium state, which means that the point is static now and the resultant force of all springs is zero. Let $\boldsymbol{E}_i$ denote the position vector of the ith point at the equilibrium state.

Hooke's law states that the elastic force $\boldsymbol{F}$ required to compress or stretch a spring via its deformation $\boldsymbol{X}$ scales according to the distance in a linear mode. That is to say, $\boldsymbol{F} = K\boldsymbol{X}$, where K is the elastic coefficient of the spring. Therefore, according to Hooke's Law, for the ith point, the elastic force of the jth spring is represented as $K_{ij}(\boldsymbol{S}_{ij} - \boldsymbol{E}_i)$, where the vector $\boldsymbol{S}_{ij} - \boldsymbol{E}_i$ represents the elastic deformation of the jth spring. Consequently, the resultant force of all springs on the ith point at the equilibrium state is zero:

$$\sum_{j=1}^{n} K_{ij}(\boldsymbol{S}_{ij} - \boldsymbol{E}_i) = 0 \tag{17}$$

Thus, the relationship between $\boldsymbol{E}_i$ and $\boldsymbol{S}_{ij}$ can be represented as follows:

$$\boldsymbol{E}_i = \sum_{j=1}^{n} w_{ij}\boldsymbol{S}_{ij} \tag{18}$$

$$w_{ij} = \frac{K_{ij}}{\sum_{j=1}^{n} K_{ij}} \tag{19}$$

where w_{ij} is the adaptive weight value. Especially, w_{ij} represents the relative significance of the jth spring compared with all other springs according to their elastic coefficients. Furthermore, for the ith point, w_{ij} also reveals the relative contribution of the jth spring in its balancing process.

For an anchor, the amount of information contained in its selected representative samples is unbalanced. Thus, we intuitively deem that the penalty strength for cross-modal similarities between anchor and them should be different. If a similarity score significantly deviates from the ideal, it should be penalized severely. Otherwise, if a similarity score is close to the ideal, it should be moderately optimized. With these insights, we aim to precisely discriminate the selected samples via an adaptive similarity weighting strategy based on the multi-spring balance system.

Concretely, the selected sample $\boldsymbol{s}_j$ of anchor $\boldsymbol{s}_i$ is corresponding to the jth spring, and its elastic coefficient is used to represent the importance score of sample $\boldsymbol{s}_j$ on anchor $\boldsymbol{s}_i$, which is defined as follows:

$$K_{ij} = G(\tilde{C}_{ij}, M_{ij}), \text{ if } \tilde{C}_{ij} \neq 0. \tag{20}$$

where $G(\cdot)$ is a function that generates the importance score for the selected sample according to the cross-modal similarity.

Considering that the importance score of a positive sample decreases when the cross-modal similarity between it and the anchor increases. On the contrary, the importance score of a negative sample increases as the cross-modal similarity between it and the anchor increases. In order to coincide with the above two opposite changing trends of importance scores, we choose the exponential function to describe

the relationship between the importance score and the cross-modal similarity. Therefore, function $G(\cdot)$ is defined as follows:

$$G(\tilde{C}_{ij}, M_{ij}) = exp(\tilde{C}_{ij}(\gamma_2 - \gamma_1 M_{ij})) \tag{21}$$

where γ_1 and γ_2 are hyper-parameters to let the MSB loss more flexible and adaptable. It can be observed from Eqs. (20) and (21) that the higher the cross-modal similarity between the negative sample and the anchor, the greater the elastic coefficient of the corresponding spring will be. However, the changing rule of the elastic coefficient for the positive sample is just the opposite of the negative one.

Therefore, due to different optimization directions for positive and negative samples, we build two multi-spring balance systems for each anchor. One is for the selected positive samples, and the other is for the selected negative samples. Specifically, w_{ij}^{+} refers to the weight value for the selected positive sample $\boldsymbol{s}_j^{+}$ of anchor $\boldsymbol{s}_i$, and w_{ij}^{-} is the weight value for the selected negative sample $\boldsymbol{s}_j^{-}$ of anchor $\boldsymbol{s}_i$. w_{ij}^{+} and w_{ij}^{-} are computed as follows:

$$w_{ij}^{+} = \frac{K_{ij}}{\sum_{\tilde{C}_{ik}=+1} K_{ik}} \tag{22}$$

$$w_{ij}^{-} = \frac{K_{ij}}{\sum_{\tilde{C}_{ik}=-1} K_{ik}} \tag{23}$$

In the following part, we take the optimization process of the similarity matrix $\boldsymbol{M}$ as an example to elaborate the proposed MSB loss. Notably, with the consideration that the reflexivity of the query and candidate, there are two directions in cross-modal retrieval, which is presented as follows:

$$\begin{aligned}\mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta}) = & \overbrace{\frac{1}{\gamma_1}\frac{1}{B}\sum_{i=1}^{B}\left\{\ln\left[\sum_{\tilde{C}_{ik}=+1} K_{ik}\right] + \ln\left[\sum_{\tilde{C}_{ik}=-1} K_{ik}\right]\right\}}^{\text{The } a\text{th modality queries the } b\text{th modality}} \\ & + \overbrace{\frac{1}{\gamma_1}\frac{1}{B}\sum_{i=1}^{B}\left\{\ln\left[\sum_{\tilde{C}_{ik}^{\mathsf{T}}=+1} K_{ik}^{*}\right] + \ln\left[\sum_{\tilde{C}_{ik}^{\mathsf{T}}=-1} K_{ik}^{*}\right]\right\}}^{\text{The } b\text{th modality queries the } a\text{th modality}}\end{aligned} \tag{24}$$

where B is the size of the mini-batch in training, $\boldsymbol{\Theta}$ is the parameters of the deep neural network, and $K_{ik}^{*} = G(\tilde{C}_{ik}^{\mathsf{T}}, M_{ik}^{\mathsf{T}})$, if $\tilde{C}_{ij}^{\mathsf{T}} \neq 0$. Specifically, $\boldsymbol{M}$ is corresponding to the cross-modal retrieval that the ath modality queries the bth modality. Particularly, M_{ij} represents the cross-modal similarity between sample $\boldsymbol{s}_i^a$ of the ath modality and sample $\boldsymbol{s}_j^b$ of the bth modality. Differently, $\boldsymbol{M}^{\mathsf{T}}$ represents the opposite direction of cross-modal retrieval. According to the chain rule of the composite function, we have:

$$\frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial \boldsymbol{M}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{\Theta}} = \sum_{i=1}^{B}\sum_{j=1}^{B} \frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial M_{ij}} \frac{\partial M_{ij}}{\partial \boldsymbol{\Theta}} \tag{25}$$

Remarkably, as indicated in Eq. (25), the gradient of the loss function is a weighted summation about the gradients of all similarities, and $\frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial M_{ij}}$ is used as weight. We believe that the influence of each M_{ij} in computing the gradient of a loss function is worthy of being distinguished, hence, we exploit the above two multi-spring balance systems to discriminate the selected positive and negative samples, respectively.

As the MSB loss encourages positive samples to be closer to the anchor, and negative samples to be farther away from the anchor, we can obtain that $\frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial M_{ij}} \leq 0$ when $\tilde{C}_{ij} = +1$, while $\frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial M_{ij}} \geq 0$ when $\tilde{C}_{ij} = -1$. Consequently, we have the following equation:

$$\frac{\partial \mathcal{H}(\boldsymbol{M}, \boldsymbol{\Theta})}{\partial M_{ij}} = \begin{cases} -w_{ij}^{+} & \text{if } \tilde{C}_{ij} = +1, \\ w_{ij}^{-} & \text{if } \tilde{C}_{ij} = -1. \end{cases} \tag{26}$$

According to the calculus theory, we can obtain the primitive integral with Eq. (26). Details of the derivation process of MSB loss (*i.e.*, $\mathcal{H}(\cdot)$) are presented in Appendix A.

The MSB loss can effectively mine the informative samples for discriminative optimization, which is a more flexible and efficient loss function for cross-modal retrieval. In our global-level and local-level subnetworks, the MSB loss is applied to optimize all primary and auxiliary similarity matrices.

4. Experiment

In this section, we conduct a series of experiments on several benchmark multi-modal datasets and compare the performance with 18 state-of-the-art methods to verify the effectiveness of our proposed M^{2}HSE method. Furthermore, parameter sensitivity and convergence analysis are presented, as well as the ablation studies to testify the contribution of each component in M^{2}HSE.

4.1. Datasets and evaluation metrics

Firstly, we introduce several widely used multi-modal datasets adopted in the experiments briefly, including Corel 5K (Duygulu et al., 2002), Pascal Sentence (Rashtchian et al., 2010) and NUS-WIDE (Chua et al., 2009). Each dataset is split to three subsets, namely training set, validation set and testing set.

Corel 5K dataset contains a total number of 5000 images collected by Corel company, which covers 50 semantic concepts. Each semantic concept contains 100 images, the number of tags in the dictionary is 260, and each image has 1-5 tags. Particularly, eight images without tags are removed by data cleaning, hence, we further divide this dataset into three parts: 4493 for training, 250 for validation, and 249 for testing.

Pascal Sentence dataset is made up of 1000 images, which is collected from the 2008 pascal development kit. Each image is tagged by Amazon Mechanical Turk through crowdsourcing to construct five sentences from various annotators, which produce one text. Specifically, Pascal Sentence is classified into 20 categories with three subsets, 800 for training, 100 for validation, and 100 for testing.

NUS-WIDE dataset consists of 269,648 images that are collected from Flickr. Furthermore, this dataset has a total number of 5018 unique tags and 81 semantic concepts. It is worth mentioning that some noisy tags do not have generalized meanings, which are useless for our task and need to be eliminated. Therefore, we choose 560 tags and a total of 25,084 image–text pairs from NUS-WIDE to establish the subdataset **NUS-WIDE-25K**, in which the training set, validation set and testing set contain 20,000, 2500 and 2584 image–text pairs, respectively. Furthermore, six semantic concepts (*i.e.*, "animal", "clouds", "person", "sky", "water" and "window") in NUS-WIDE-25K are reserved to establish a new subdataset **NUS-WIDE-5K**, which contains 4996 image–text pairs with three subsets: 4500 pairs for training, 250 pairs for validating, and 246 pairs for testing.

Then, we conduct two cross-modal retrieval tasks: (1) searching text by image (I→T), (2) searching image by text (T→I) on the above datasets. The mean Average Precision (mAP), Recall@K, and Precision–Recall curve (PR curve) are adopted to evaluate the experimental results of all methods on I→T and T→I.

mAP is used to calculate the mean value of all Average Precision (AP) of every query, which is defined as:

$$AP = \frac{1}{R}\sum_{n=1}^{N}\frac{R_n}{n} * \varphi(n) \tag{27}$$

where N is the number of samples, R is the number of true items in the retrieval results, and R_n is the number of true items at the top of n retrieval results. $\varphi(n)$ is set to 1 if the nth returned item is true, otherwise, $\varphi(n)$ is set to 0. Besides, the average mAP scores (*i.e.*, Avg.) of I→T and T→I are also reported to show the general performance of

methods. The higher value of mAP means better performance of the cross-modal retrieval.

Recall@K is the proportion of relevant items found in the top-K positions of the ranking list. The greater the value of Recall@K is, the better performance is.

$$Recall@K = \frac{1}{N}\sum_{i=1}^{N} r_K \tag{28}$$

The testing set has a total of N instances, and r_K is 1 if the ground-truth result is in the top-K returned results, else r_K is 0. We exhibit the recall rates at top 1 result (R@1), top 5 results (R@5) and top 10 results (R@10). To elaborate the overall performance of Recall@K, we also provide an additional criterion R@sum by summing R@1, R@5 and R@10 together at both I→T and T→I.

PR curve shows the changing trends of the retrieval precision under all recall values. The larger the area enclosed by the curve, the better the performance of the model.

4.2. Implementation details

The source codes of M^{2}HSE will be released at https://github.com/Boreas-pxl/M2HSE, and we briefly introduce some significant implementation details, such as data preprocess strategy, experimental parameter settings, and network training details.

Data preprocess strategy. For each image, the coarse-grained and fine-grained visual features are extracted from the whole image and its patches, respectively. Especially, to balance the computation cost and data capacity of each patch, we uniformly divide each image into 3*3 patches, that is, $n = 9$. Concretely, we exploit the AlexNet (Krizhevsky et al., 2012), pre-trained on ImageNet, to extract the CNN features for a fair comparison with most existing works. The dimension of features extracted from the seventh fully-connected (FC7) layer in AlexNet is 4096, *i.e.*, $d_1 = 4096$. Meanwhile, for acquiring the SIFT-BoVW features, we empirically discover it instructive to set the number of clusters to 500 in K-means, that is, $d_3 = 500$.

Experimental parameter settings. Both CNN and SIFT-BoVW features are mapped into the d-dimensional embedding space through fully-connected layers in two subnetworks, and $d = 1024$. Additionally, for each text, the dimensionality of the word embeddings is set as 300, *i.e.*, $d_2 = 300$. Then, to connect the domains of vision and language, we use a bi-directional GRU with only one layer, and the dimensionality of hidden state (*i.e.*, $\overrightarrow{\boldsymbol{h}}_j$ and $\overleftarrow{\boldsymbol{h}}_j$ in Eq. (8)) is also set as 1024. The inverse temperature parameter λ in softmax function is set as 9 following Lee et al. (2018). In addition, the sensitivity of hyper-parameters involved in M^{2}HSE is discussed detailedly in Section 4.5.

Network training details. The proposed M^{2}HSE approach is implemented by Pytorch (Paszke et al., 2017) using an NVIDIA GeForce RTX 2080 GPU. Both the global-level subnetwork and the local-level subnetwork are trained E epochs in a mini-batch by the Adam optimizer (Kingma & Ba, 2014), and the batch-size is B. For all models, we train with a learning rate of 0.0002 for the first $E/2$ epochs, and then decay the learning rate by 0.1 for the remainder epochs. Specifically, for Corel 5K and NUS-WIDE-25K, the batch size is set as 100 and we utilize 100 epochs; for Pascal Sentence, the batch size is 10 and 30 epochs are exploited; as for NUS-WIDE-5K, the batch size is 64 and the number of epochs is 20. We evaluate the effectiveness of each model on the validation set at every epoch, and obtain the best model according to mAP scores. Then, the best model is evaluated on the testing set to provide experimental results.

4.3. Compared methods

To verify the effectiveness of our proposed M^{2}HSE, we choose totally 18 state-of-the-art methods for a comprehensive comparison. There are 7 conventional non-DNN-based methods for cross-modal retrieval, and their introductions are presented as follows:

- **CCA** (Rasiwasia et al., 2010) is a classic dimensionality reduction technology, which learns heterogeneous spaces for multi-modal data, some joint information is reflected by the correlations across two or more spaces.
- **CMCP** (Zhai et al., 2012a) simultaneously handles the relevant and irrelevant correlations between different modalities, meanwhile propagates the correlations between any combinations of heterogeneous data.
- **HSNN** (Zhai et al., 2012b) obtains the specific similarity by calculating the probability of different modalities belonging to the same class. This probability can be estimated by exploiting the nearest neighbors of each item.
- **JGRHML** (Zhai et al., 2013a) is a heterogeneous metric learning method with a joint graph regularization. This algorithm mines the complementary information between different modalities.
- **JRL** (Zhai et al., 2013b) integrates sparse and semi-supervised regularization into a joint optimization framework for multi-modal data. It also explores the effect of labeled data and unlabeled data of various modalities.
- **JFSSL** (Wang et al., 2015) learns the projection matrix for each modality separately, and uses a graph regularization term to map data into a common space while maintaining the inter-modality and intra-modality relationships.
- **S^{2}UPG** (Peng et al., 2015) uses the joint graph to exploit semantic correlations between multi-modal data. Meanwhile, it takes fine-grained data to emphasize the important parts and makes cross-modal correlations more precise.

Furthermore, the remaining 11 DNN-based compared methods are briefly introduced as follows:

- **DCCA** (Andrew et al., 2013) learns sophisticated nonlinear transformations between two perspectives of data, leading in highly linearly connected representations.
- **CCL** (Peng, Qi et al., 2017) designs a hierarchical network to take advantage of the multi-level implication with joint optimization, which keeps both the inter-modality and intra-modality correlations.
- **SCAN** (Lee et al., 2018) regards image–text matching as the central of cross-modal retrieval task, and it maps words and image regions into a common embedding space to discover the full latent alignments between them.
- **GXN** (Gu et al., 2018) combines two generative models into the feature embedding for cross-modal retrieval, which learns the high-level abstract representation and the local grounded representation for images and texts.
- **VSESC** (Chen et al., 2019) constructs a visual-based space and a textual-based space, then incorporates a semantic consistency constraint to learn them simultaneously.
- **MAVA** (Peng et al., 2019) proposes a vision-language dual-attention mechanism, which discriminates the fine-grained data with various importance at the relation level and the local level.
- **SGRAF** (Diao et al., 2021) learns the vector-based similarity representations and infers relation-aware similarities. Then, it integrates the global and local similarities by the attention mechanism.
- **SCL** (Liu et al., 2022) utilizes the unsupervised contrastive learning to obtain more discriminative features for multi-modal data, and exploits the correlations among intra- and inter-modality items.
- **CGMN** (Cheng et al., 2022) explores the intra-relation in images and sentences with graph convolutional networks, and achieves inter-relation reasoning between regions and words without affecting the search efficiency.
- **NAAF** (Zhang et al., 2022) explicitly uses the positive influence of matched patches and the negative influence of mismatched patches to estimate the similarities between images and texts.
- **VSRN++** (Li et al., 2022) proposes an image–text embedding learning framework, in which the visual and textual semantic reasoning modules are implemented to obtain the global representations for images and texts.

Table 1
The mAP scores of cross-modal retrieval for M^{2}HSE and other compared methods on all datasets.

Methods	Corel 5K			Pascal Sentence			NUS-WIDE-25K			NUS-WIDE-5K		
	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.
CCA (Rasiwasia et al., 2010)	0.1006	0.1129	0.1068	0.0927	0.0963	0.0945	0.7222	0.6854	0.7038	0.2239	0.2537	0.2388
CMCP (Zhai et al., 2012a)	0.3681	0.3638	0.3660	0.4717	0.4376	0.4546	–	–	–	0.3951	0.4329	0.4140
HSNN (Zhai et al., 2012b)	0.3708	0.3673	0.3691	0.4136	0.3915	0.4025	–	–	–	0.4761	0.4818	0.4790
JGRHML (Zhai et al., 2013a)	0.3996	0.4097	0.4047	0.4733	0.4381	0.4557	0.7607	0.7346	0.7476	0.4824	0.4865	0.4845
JRL (Zhai et al., 2013b)	0.4081	0.4197	0.4139	0.5208	0.5067	0.5138	0.7420	0.6973	0.7197	0.4982	0.5330	0.5156
JFSSL (Wang et al., 2015)	0.4141	0.4139	0.4140	0.5073	0.4640	0.4856	0.7092	0.7148	0.7120	0.5224	0.5098	0.5161
S^{2}UPG (Peng et al., 2015)	0.4289	0.4249	0.4269	0.5521	0.5475	0.5498	0.7591	0.6820	0.7205	0.4943	0.4974	0.4959
DCCA (Andrew et al., 2013)	0.3107	0.3064	0.3086	0.4754	0.4719	0.4737	0.7089	0.7103	0.7096	0.4264	0.4330	0.4297
CCL (Peng, Qi et al., 2017)	0.4354	0.4413	0.4384	0.5679	0.5633	0.5656	–	–	–	0.4329	0.5036	0.4683
SCAN* (Lee et al., 2018)	0.4916	0.4886	0.4901	0.5662	0.5709	0.5686	0.7536	0.8014	0.7775	0.5315	0.5302	0.5309
GXN* (Gu et al., 2018)	0.5254	0.5186	0.5220	0.5981	0.5785	0.5883	0.8020	0.8019	0.8020	0.5654	0.5572	0.5613
VSESC* (Chen et al., 2019)	0.5001	0.4947	0.4974	0.5757	0.5721	0.5739	0.7768	0.7980	0.7874	0.5421	0.5567	0.5494
MAVA* (Peng et al., 2019)	0.5217	0.5134	0.5176	0.5723	0.5711	0.5717	0.8012	0.8092	0.8052	0.5977	0.5697	0.5837
SGRAF* (Diao et al., 2021)	0.5241	0.5136	0.5189	0.5876	0.5727	0.5802	0.8319	0.8376	0.8348	0.6099	0.6025	0.6063
SCL (Liu et al., 2022)	0.5404	0.5501	0.5453	0.6185	0.6219	0.6202	0.8255	0.8197	0.8226	0.6105	0.6188	0.6147
CGMN (Cheng et al., 2022)	0.5266	0.5231	0.5249	0.6218	0.6059	0.6139	0.8401	0.8264	0.8333	0.6415	0.6301	0.6358
NAAF (Zhang et al., 2022)	0.5493	0.5538	0.5516	0.6156	0.6286	0.6221	0.8468	0.8380	0.8424	0.6252	0.6300	0.6276
VSRN++ (Li et al., 2022)	0.5589	0.5546	0.5568	**0.6475**	0.6104	0.6290	0.8498	0.8454	0.8476	0.6357	0.6401	0.6379
M^{2}HSE (ours)	**0.5715**	**0.5804**	**0.5760**	0.6429	**0.6421**	**0.6425**	**0.8600**	**0.8664**	**0.8632**	**0.6550**	**0.6531**	**0.6541**

Table 2
The Recall@K scores on Corel 5K dataset.

Methods	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA (Rasiwasia et al., 2010)	0.158	0.307	0.427	0.251	0.439	0.581	2.163
CMCP (Zhai et al., 2012a)	0.385	0.596	0.694	0.445	0.482	0.707	3.309
HSNN (Zhai et al., 2012b)	0.430	0.587	0.712	0.425	0.531	0.718	3.403
JGRHML (Zhai et al., 2013a)	0.489	0.583	0.746	0.485	0.577	0.733	3.613
JRL (Zhai et al., 2013b)	0.465	0.592	0.681	0.496	0.636	0.758	3.628
JFSSL (Wang et al., 2015)	0.461	0.637	0.759	0.513	0.663	0.774	3.807
S^{2}UPG (Peng et al., 2015)	0.487	0.660	0.777	0.523	0.687	0.782	3.916
DCCA (Andrew et al., 2013)	0.343	0.461	0.644	0.383	0.530	0.704	3.065
CCL (Peng, Qi et al., 2017)	0.545	0.758	0.838	0.495	0.721	0.802	4.159
SCAN* (Lee et al., 2018)	0.571	0.689	0.790	0.605	0.762	0.830	4.247
GXN* (Gu et al., 2018)	0.664	0.792	0.913	0.702	0.867	0.917	4.855
VSESC* (Chen et al., 2019)	0.573	0.730	0.817	0.627	0.808	0.844	4.339
MAVA* (Peng et al., 2019)	0.601	0.750	0.820	0.666	0.818	0.882	4.537
SGRAF* (Diao et al., 2021)	0.643	0.784	0.808	0.701	0.876	0.918	4.730
SCL (Liu et al., 2022)	0.671	0.804	0.924	0.713	0.860	0.921	4.893
CGMN (Cheng et al., 2022)	0.655	0.806	0.870	0.736	0.874	0.912	4.853
NAAF (Zhang et al., 2022)	0.671	0.811	0.908	0.724	0.870	0.915	4.899
VSRN++ (Li et al., 2022)	0.668	0.808	0.917	0.718	0.881	0.923	4.915
M^{2}HSE (ours)	**0.697**	**0.830**	**0.940**	**0.738**	**0.894**	**0.958**	**5.057**

Notably, all compared methods are implemented with the source codes published by the authors, and the best results of mAP and Recall@K are highlighted in bold. As image–text pairs in NUS-WIDE-25K belong to multiple semantic concepts, CMCP, HSNN and CCL cannot be carried out on it, we use '–' to represent the unreported experimental results. Furthermore, for some DNN-based methods, the symbol '∗' is used to indicate the performance of their ensemble models.

4.4. Comparison results

The mAP scores of our M^{2}HSE approach with the compared methods on all datasets are presented in Table 1, which indicates that our method achieves the best overall performance on mAP scores. Specifically, compared with the previous best model VSRN++, the average mAP scores of M^{2}HSE on Corel 5K, Pascal Sentence, NUS-WIDE-25K, and NUS-WIDE-5K, are obviously improved by 1.92%, 1.35%, 1.56%, and 1.62%, respectively. Note that VSRN++ only outperforms our M^{2}HSE by 0.46% in I→T retrieval task on Pascal Sentence. However, our M^{2}HSE method achieves almost balanced performance in both I→T and T→I on all datasets, which means that M^{2}HSE can effectively bridge the "heterogeneity gap" and the "granularity gap", and then demonstrates its effectiveness on two different directions of cross-modal retrieval.

Recall@K is further utilized to verify the performance of cross-modal retrieval. From Tables 2 to 5, we observe that M^{2}HSE also achieves the best overall performance on all datasets. Specifically, our M^{2}HSE outperforms the previous best model VSRN++ by a margin of 14.2%, 12.5%, 12.0%, and 13.2% in terms of the overall performance R@sum on Corel 5K, Pascal Sentence, NUS-WIDE-25K, and NUS-WIDE-5K, respectively. Notably, VSRN++ only achieves the highest R@10 score in I→T retrieval task on Pascal Sentence, which outperforms our M^{2}HSE by a small margin of 0.9%. However, these compared methods cannot maintain the continuous effectiveness on all datasets.

Besides, the scores of mAP and Recall@K on NUS-WIDE-25K are noticeably higher than that on other datasets, because image–text pairs in NUS-WIDE-25K belong to multiple semantic concepts and thus there are more chances to hit the true semantic concepts. Moreover, scores of mAP and Recall@K on NUS-WIDE-5K are higher than that on Corel 5K, because (1) the quality of tagged words in NUS-WIDE-5K is significantly better than that in Corel 5K, and (2) the number of semantic concepts in NUS-WIDE-5K is remarkably less than that in Corel 5K. Notably, the text modality of Pascal Sentence is a set of sentences, while that of other datasets is a set of tags. Experimental results of mAP and Recall@K

Table 3
The Recall@K scores on Pascal Sentence dataset.

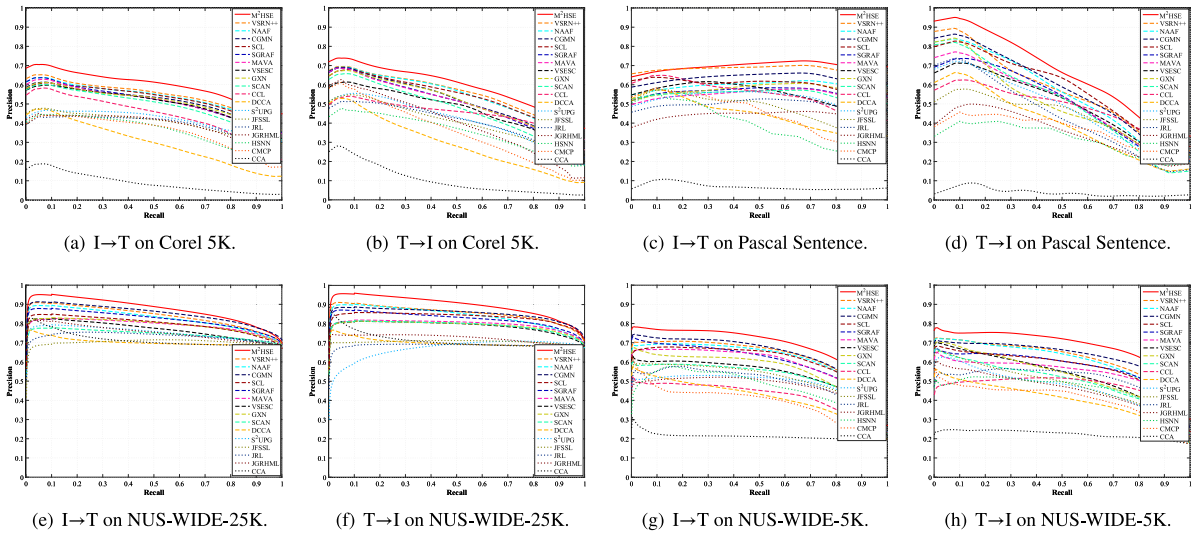
Methods	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA (Rasiwasia et al., 2010)	0.062	0.244	0.365	0.072	0.193	0.338	1.274
CMCP (Zhai et al., 2012a)	0.421	0.724	0.855	0.401	0.778	0.876	4.055
HSNN (Zhai et al., 2012b)	0.483	0.598	0.764	0.331	0.498	0.801	3.475
JGRHML (Zhai et al., 2013a)	0.381	0.601	0.881	0.419	0.875	0.942	4.099
JRL (Zhai et al., 2013b)	0.541	0.671	0.904	0.603	0.894	0.966	4.579
JFSSL (Wang et al., 2015)	0.425	0.764	0.896	0.552	0.912	0.907	4.456
S^2UPG (Peng et al., 2015)	0.521	0.735	0.872	0.697	0.918	0.970	4.713
DCCA (Andrew et al., 2013)	0.407	0.601	0.822	0.584	0.902	0.915	4.231
CCL (Peng, Qi et al., 2017)	0.613	0.862	0.919	0.599	0.854	0.933	4.780
SCAN* (Lee et al., 2018)	0.588	0.878	0.900	0.761	0.828	0.882	4.837
GXN* (Gu et al., 2018)	0.596	0.875	0.960	0.821	0.884	0.968	5.104
VSESC* (Chen et al., 2019)	0.510	0.789	0.921	0.823	0.911	0.964	4.918
MAVA* (Peng et al., 2019)	0.561	0.824	0.892	0.685	0.932	0.961	4.855
SGRAF* (Diao et al., 2021)	0.631	0.852	0.895	0.757	0.901	0.981	5.017
SCL (Liu et al., 2022)	0.615	0.870	0.928	0.771	0.926	0.980	5.090
CGMN (Cheng et al., 2022)	0.611	0.876	0.951	0.813	0.888	0.974	5.113
NAAF (Zhang et al., 2022)	0.613	0.869	0.932	0.784	0.928	0.981	5.107
VSRN++ (Li et al., 2022)	0.622	0.873	**0.966**	0.793	0.917	0.969	5.140
M^2HSE (ours)	**0.664**	**0.900**	0.957	**0.826**	**0.942**	**0.993**	**5.265**

Table 4
The Recall@K scores on NUS-WIDE-25K dataset.

Methods	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA (Rasiwasia et al., 2010)	0.205	0.824	0.931	0.142	0.899	0.942	3.943
CMCP (Zhai et al., 2012a)	–	–	–	–	–	–	–
HSNN (Zhai et al., 2012b)	–	–	–	–	–	–	–
JGRHML (Zhai et al., 2013a)	0.447	0.816	0.959	0.409	0.840	0.957	4.428
JRL (Zhai et al., 2013b)	0.497	0.814	0.936	0.114	0.829	0.946	4.136
JFSSL (Wang et al., 2015)	0.329	0.815	0.898	0.297	0.816	0.908	4.063
S^2UPG (Peng et al., 2015)	0.443	0.868	0.908	0.335	0.819	0.883	4.256
DCCA (Andrew et al., 2013)	0.257	0.897	0.894	0.159	0.891	0.904	4.002
CCL (Peng, Qi et al., 2017)	–	–	–	–	–	–	–
SCAN* (Lee et al., 2018)	0.849	0.914	0.945	0.784	0.877	0.908	5.277
GXN* (Gu et al., 2018)	0.811	0.953	0.968	0.818	0.862	0.893	5.305
VSESC* (Chen et al., 2019)	0.852	0.902	0.933	0.793	0.894	0.916	5.290
MAVA* (Peng et al., 2019)	0.809	0.944	0.970	0.825	0.865	0.900	5.313
SGRAF* (Diao et al., 2021)	0.873	0.956	0.969	0.876	0.928	0.943	5.545
SCL (Liu et al., 2022)	0.884	0.951	0.963	0.869	0.904	0.958	5.529
CGMN (Cheng et al., 2022)	0.893	0.954	0.970	0.873	0.914	0.950	5.554
NAAF (Zhang et al., 2022)	0.906	0.958	0.965	0.917	0.959	0.976	5.681
VSRN++ (Li et al., 2022)	0.905	0.956	0.971	0.909	0.950	0.974	5.675
M^2HSE (ours)	**0.948**	**0.976**	**0.980**	**0.933**	**0.972**	**0.986**	**5.795**

Table 5
The Recall@K scores on NUS-WIDE-5K dataset.

Methods	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA (Rasiwasia et al., 2010)	0.246	0.627	0.809	0.242	0.335	0.379	2.638
CMCP (Zhai et al., 2012a)	0.489	0.702	0.827	0.228	0.608	0.897	3.751
HSNN (Zhai et al., 2012b)	0.492	0.657	0.746	0.399	0.784	0.921	3.999
JGRHML (Zhai et al., 2013a)	0.429	0.696	0.809	0.579	0.725	0.901	4.139
JRL (Zhai et al., 2013b)	0.575	0.814	0.872	0.503	0.846	0.914	4.524
JFSSL (Wang et al., 2015)	0.564	0.732	0.896	0.552	0.812	0.927	4.483
S^2UPG (Peng et al., 2015)	0.476	0.681	0.752	0.612	0.774	0.909	4.204
DCCA (Andrew et al., 2013)	0.520	0.751	0.921	0.434	0.643	0.910	4.179
CCL (Peng, Qi et al., 2017)	0.492	0.631	0.853	0.308	0.643	0.875	3.802
SCAN* (Lee et al., 2018)	0.589	0.716	0.852	0.673	0.867	0.919	4.616
GXN* (Gu et al., 2018)	0.612	0.785	0.876	0.645	0.849	0.928	4.695
VSESC* (Chen et al., 2019)	0.603	0.789	0.837	0.657	0.841	0.885	4.612
MAVA* (Peng et al., 2019)	0.625	0.788	0.908	0.629	0.856	0.935	4.741
SGRAF* (Diao et al., 2021)	0.782	0.889	0.923	0.794	0.869	0.915	5.172
SCL (Liu et al., 2022)	0.793	0.876	0.919	0.804	0.880	0.906	5.178
CGMN (Cheng et al., 2022)	0.794	0.885	0.921	0.826	0.871	0.907	5.204
NAAF (Zhang et al., 2022)	0.788	0.873	0.919	0.810	0.862	0.929	5.181
VSRN++ (Li et al., 2022)	0.801	0.883	0.924	0.817	0.869	0.911	5.205
M^2HSE (ours)	**0.815**	**0.907**	**0.940**	**0.834**	**0.899**	**0.942**	**5.337**



(a) I→T on Corel 5K. (b) T→I on Corel 5K. (c) I→T on Pascal Sentence. (d) T→I on Pascal Sentence.

(e) I→T on NUS-WIDE-25K. (f) T→I on NUS-WIDE-25K. (g) I→T on NUS-WIDE-5K. (h) T→I on NUS-WIDE-5K.

Fig. 6. The PR curves of cross-modal retrieval for M^{2}HSE and other compared methods on all datasets.

Table 6
Comparison of the training time with recent state-of-the-art methods on all datasets, and the symbol 's' represents for second.

Methods	Corel 5K	Pascal Sentence	NUS-WIDE-25K	NUS-WIDE-5K
DCCA (Andrew et al., 2013)	8764 s	2165 s	14442 s	3296 s
CCL (Peng, Qi et al., 2017)	9957 s	3683 s	–	6261 s
SCAN (Lee et al., 2018)	12633 s	3487 s	49750 s	11900 s
GXN (Gu et al., 2018)	28298 s	5314 s	55419 s	17244 s
VSESC (Chen et al., 2019)	23970 s	4716 s	52186 s	14315 s
MAVA (Peng et al., 2019)	20013 s	4341 s	52092 s	12093 s
SGRAF (Diao et al., 2021)	27979 s	5007 s	57435 s	16757 s
SCL (Liu et al., 2022)	25334 s	3851 s	50178 s	12570 s
CGMN (Cheng et al., 2022)	32217 s	6019 s	59048 s	17331 s
NAAF (Zhang et al., 2022)	26839 s	4596 s	58584 s	15553 s
VSRN++ (Li et al., 2022)	34917 s	6558 s	60109 s	18365 s
M^{2}HSE (ours)	29722 s	5818 s	57212 s	14671 s

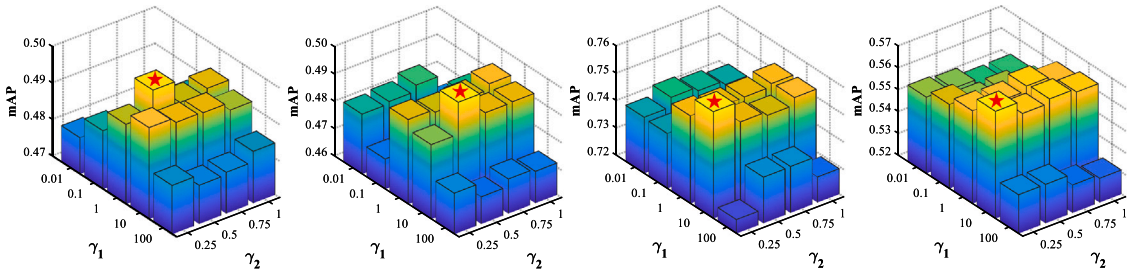
also demonstrate that M^{2}HSE can obtain better performance whether using sentences or tags in cross-modal retrieval. We also find that DNN-based methods generally outperform other traditional non-DNN-based methods.

Next, the PR curves of both I→T and T→I cross-modal retrieval tasks on all datasets are shown in Fig. 6, from which we can observe that M^{2}HSE achieves the best overall performance since the PR curves generated by M^{2}HSE cover more areas than other methods. It should be noted that VSRN++ achieves similar performance on PR curve to M^{2}HSE in Fig. 6(c), while M^{2}HSE outperforms VSRN++ in all other cases. Just like the experimental results of mAP and Recall@K, the PR curves of DNN-based methods generally outperform traditional non-DNN-based methods.

To better evaluate the proposed method, we make comparative experiments on the training time of DNN-based methods in Table 6. It is worth mentioning that source codes of all methods run on the same machine with only one GPU. What we observe from Table 6 can be summarized by the following aspects. Firstly, DCCA, CCL, and SCAN take the least training time, however, their performance of cross-modal retrieval is not competitive when compared to other DNN based methods. Secondly, although the training time of GXN, VSESC, MAVA, SGRAF, SCL, CGMN, and NAAF is roughly the same as M^{2}HSE, M^{2}HSE performs significantly better than them on the task of cross-modal retrieval. Thirdly, VSRN++ is second only to M^{2}HSE on cross-modal retrieval, but it requires the longest training time.

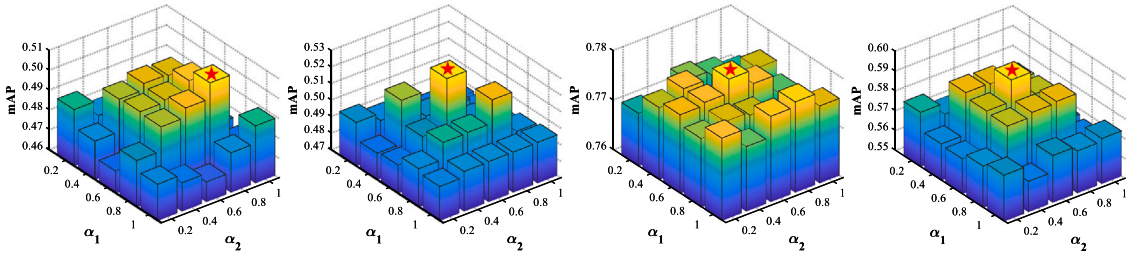
By carefully analyzing the above experimental results, we gain the following observations:

- We compare seven traditional non-DNN-based methods firstly. Specifically, S^{2}UPG is superior to other methods in most cases because S^{2}UPG exploits fine-grained features in the cross-modal similarity learning, while other methods only use coarse-grained features and omit complementary fine-grained clues. In addition, as CMCP, HSNN, JGRHML, JRL and JFSSL make full use of semantic concepts to enlarge interval distances among different semantic concepts, they are clearly better than CCA.
- Benefiting from the powerful capability of deep neural networks in uncovering the nonlinear cross-modal correlations, most DNN-based methods outperform non-DNN-based methods. For instance, DCCA achieves significant performance improvement compared with CCA, due to DCCA maximizes the association between the output layers of two distinct subnetworks with coarse-grained data. Besides, benefiting from fusing multi-grained features with a hierarchical network, CCL outperforms DCCA. Moreover, SCL proposes a self-supervised correlation learning framework based on the contrastive learning, which designs a weight-sharing scheme and minimizes the modality-invariant loss in the common space. SCL significantly outperforms DCCA and CCL, and even achieves comparable performance to attention-based methods.
- Attention mechanism based models are significantly superior to DCCA and CCL, because they can effectively estimate the cross-modal similarity by achieving the latent matches between image patches and words. Concretely, SCAN and VSESC attend to visual regions and words with each other as the corresponding context to compute cross-modal similarity. As VSESC incorporates a constraint of the semantic consistency in the objective function, it outperforms SCAN.



(a) Corel 5K (b) Pascal Sentence (c) NUS-WIDE-25K (d) NUS-WIDE-5K

Fig. 7. mAP scores variation with respect to γ_1 and γ_2 on all datasets.



(a) Corel 5K (b) Pascal Sentence (c) NUS-WIDE-25K (d) NUS-WIDE-5K

Fig. 8. mAP scores variation with respect to α_1 and α_2 when we fix γ_1, γ_2 on all datasets.

However, both of them only exploit fine-grained relations. MAVA measures the cross-modal similarity from the instance-level, region-level, and relation-level, so it achieves better performance than SCAN and VSESC. Besides, through suppressing the irrelevant interactions at global and local level, SGRAF significantly outperforms MAVA. Note that NAAF outperforms the above methods, because it simultaneously explores the positive influence of matched patches and the negative influence of mismatched patches to calculate the similarity.

- As a typical cross-modal GAN approach, GXN achieves fairly good performance because it generates effective feature representations with two generative modules, which are immensely helpful for matching the right image–text pairs. Additionally, CGMN uses graphs to represent images and sentences, and then explores the potential relations between image regions or words by GCN, which outperforms GXN in most cases. VSRN++ establishes connections between image patches to obtain features with the semantic association knowledge, and then it performs global semantic reasoning to select the distinctive information for cross-modal similarity learning. Although it is second to M^{2}HSE, it needs more training time than our method.
- M^{2}HSE achieves the best overall performance on all datasets. The reasons lie in that M^{2}HSE can mine and fuse the complementarity in multi-modal and multi-grained data to bridge the "heterogeneity gap" and the "granularity gap". Concretely, M^{2}HSE accurately describes the complex and nonlinear cross-modal relationships, which is a distinct advantage compared with non-DNN-based methods. As M^{2}HSE exploits both coarse-grained and fine-grained relations, it also easily outperforms SCAN, VSESC, GXN, SCL, CGMN, and NAAF. Although MAVA and SGRAF fully utilize both coarse-grained and fine-grained data, M^{2}HSE still performs better than them due to the fact that: (1) M^{2}HSE discovers lots of latent cross-modal correlations in the auxiliary modality to calculate the cross-modal similarity more comprehensively. (2) M^{2}HSE adopts the proposed multi-spring balance loss to optimize the cross-modal similarity more accurately through selecting important samples and further assigning suitable weights in a unified framework.

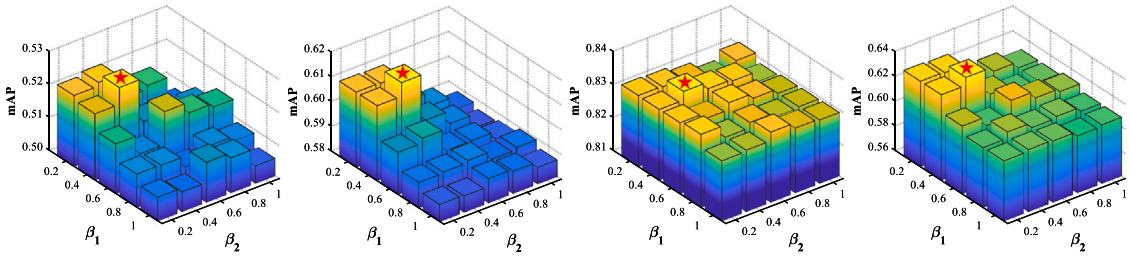
4.5. Parameter sensitivity and convergence analysis

There are several hyper-parameters involved in M^{2}HSE, *i.e.*, (γ_1, γ_2) in MSB loss, (α_1, α_2) in $\mathcal{J}^G$, (β_1, β_2) in $\mathcal{J}^L$, and (θ_1, θ_2) for fusing the global-level and local-level primary similarity matrices. They are selected on the validation set through the grid-search method with pre-defined range of values, and the optimal combination of each set of parameters is marked with a red five-pointed star in the figure. All experiments related to parameter sensitivity analysis are conducted using the averaging mAP scores on both I→T and T→I.

Firstly, we set the value of γ_1 and γ_2 in the range of $\{0.01, 0.1, 1, 10, 100\}$ and $\{0.25, 0.5, 0.75, 1\}$, respectively, and conduct parameter analysis with the M^{2}HSE-GP (refers to Table 7) on all datasets in Fig. 7. Generally speaking, for all datasets, better performance can be achieved when γ_1 is equal to 1 or 10, and the proposed method is not much sensitive to γ_2. Especially, the best parameter settings of (γ_1, γ_2) for Corel 5K, Pascal Sentence, NUS-WIDE-25K and NUS-WIDE-5K are set to (1, 0.5), (10, 0.5), (10, 0.25), and (10, 0.25), respectively.

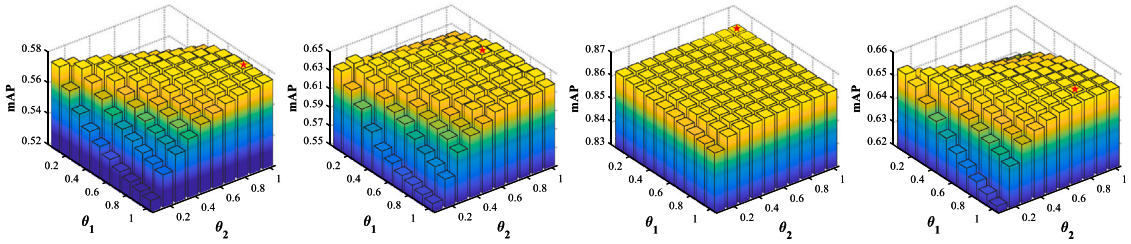
Secondly, we fix γ_1 and γ_2 as their optimal values for each dataset, then report the mAP scores with α_1 and α_2 varying in the global-level subnetwork (shown in Fig. 8), and with β_1 and β_2 varying in the local-level subnetwork (shown in Fig. 9). Particularly, α_1, α_2, β_1, β_2 are all ranged in $\{0.2, 0.4, 0.6, 0.8, 1\}$. In Fig. 8, we can see that (1) if the values of α_1 and α_2 are too small, the auxiliary modality cannot provide enough complementary semantic information, which exerts a negative influence on the performance of cross-modal retrieval, (2) if the values of α_1 and α_2 are too large, the primary modality cannot play a critical role in cross-modal retrieval, and thus obtain lower mAP scores. Hence, the best values of (α_1, α_2) for Corel 5K, Pascal Sentence, NUS-WIDE-25K and NUS-WIDE-5K are set to (0.8, 0.8), (0.6, 0.6), (0.6, 0.6), and (0.6, 0.6), respectively. Similarly, experimental results in Fig. 9 demonstrate that too small or too large values of β_1 and β_2 may decrease mAP scores as well. Furthermore, the best parameters setting of (β_1, β_2) should be (0.4, 0.4) for all datasets. Note that M^{2}HSE is not much sensitive to α_1, α_2, β_1, β_2 on NUS-WIDE-25K, due to the fact that each image–text pair of this dataset belongs to multiple semantic concepts and thus there are more chances to hit the ground-truth.

Thirdly, we tune θ_1 and θ_2 in the range of $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, and experimental results are shown in Fig. 10. Specifically, the best parameter settings of (θ_1, θ_2) for Corel 5K, Pascal Sentence, NUS-WIDE-25K and NUS-WIDE-5K are set to (0.8, 1), (0.5, 0.9), (0.1, 1), and (0.8, 0.8), respectively. It can be observed that for Corel 5K, Pascal Sentence and NUS-WIDE-5K, mAP scores drop if θ_1



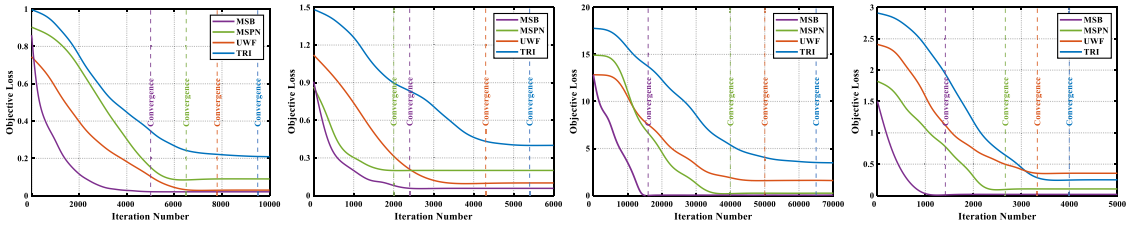
(a) Corel 5K (b) Pascal Sentence (c) NUS-WIDE-25K (d) NUS-WIDE-5K

Fig. 9. mAP scores variation with respect to β_1 and β_2 when we fix γ_1, γ_2 on all datasets.



(a) Corel 5K (b) Pascal Sentence (c) NUS-WIDE-25K (d) NUS-WIDE-5K

Fig. 10. mAP scores variation with respect to θ_1 and θ_2 when we fix γ_1, γ_2, α_1, α_2, β_1, β_2 on all datasets.



(a) Corel 5K (b) Pascal Sentence (c) NUS-WIDE-25K (d) NUS-WIDE-5K

Fig. 11. Convergence curves of the objective function $\mathcal{J}^G$ with four algorithms on all datasets.

and θ_2 are quite different, which means that coarse and fine granularity have almost the same amount of semantic information and they can well complement each other. Besides, M^{2}HSE is less sensitive to θ_1 and θ_2 on NUS-WIDE-25K.

Finally, as shown in Fig. 11, we conduct the convergence analysis with the global-level subnetwork on all datasets. As the size of the training set of these datasets is different, the number of their iterations are not fixed. For instance, the iteration number for NUS-WIDE-25K is significantly higher than that for other datasets, because its scale is much larger than theirs. Particularly, the convergence pattern of the local-level subnetwork is quite similar to that of the global-level subnetwork. For simplicity, we do not report the convergence curves of $\mathcal{J}^L$. In order to fully demonstrate the effectiveness of our proposed MSB loss, we compare it with Triplet loss (TRI Frome et al., 2013), Universal Weighting Framework (UWF Wei, Yang et al., 2021) and Meta Self-Paced Network (MSPN Wei, Xu et al., 2021). Notably, TRI, UWF, and MSPN are used to substitute for MSB in $\mathcal{J}^G$.

Specifically, the training samples are selected randomly and considered equally in TRI, which leads a slow convergence in the training stage. Compared with TRI, we can observe that the convergence speed of $\mathcal{J}^G$ is greatly improved on each dataset by adopting our proposed MSB loss, which can accelerate convergence and improve performance through effectively mining informative samples for discriminative optimization. Additionally, UWF and MSPN also notice the problem of samples selecting and weights assigning in cross-modal scenario, hence, UWF and MSPN converge faster than TRI.

To further prove the effectiveness and advantages of our strategy, the MSB loss is compared with UWF and MSPN. Concretely, $\mathcal{J}^G$ converges faster with MSB instead of UWF on all datasets. The convergence speed is further improved with MSPN, and even faster on Pascal Sentence than using MSB. In addition to compare the convergence speed of the above algorithms, we also set up a series of ablation models with them to show their performance on cross-modal retrieval. Note that the detailed experimental results and analysis are discussed in Section 4.6.2.

Table 7
Experimental configurations of different models in ablation study 1.

Models	Subnetwork		Modality	
	GLO	LOC	PRI	AUX
M^{2}HSE-GP	√		√	
M^{2}HSE-GPA	√		√	√
M^{2}HSE-LP		√	√	
M^{2}HSE-LPA		√	√	√
M^{2}HSE-GLP	√	√	√	
M^{2}HSE	√	√	√	√

4.6. Ablation studies

In this section, to investigate the contribution of each key component in M^{2}HSE, a series of ablation experiments are carried out under different configurations.

4.6.1. Ablation study 1

As shown in Table 7, we select four components, GLO, LOC, PRI and AUX, to construct some ablation models in ablation study 1.

Table 8
Experimental results of ablation study 1 on all datasets with mAP scores.

Model	Corel 5K			Pascal Sentence			NUS-WIDE-25K			NUS-WIDE-5K		
	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.
M^2HSE-GP	0.4859	0.5048	0.4953	0.5008	0.4967	0.4988	0.7501	0.7617	0.7559	0.5635	0.5700	0.5668
M^2HSE-GPA	**0.4983**	**0.5180**	**0.5082**	**0.5342**	**0.5216**	**0.5279**	**0.7650**	**0.7933**	**0.7792**	**0.5887**	**0.5952**	**0.5919**
M^2HSE-LP	0.5091	0.5167	0.5109	0.5953	0.5788	0.5871	0.8163	0.8082	0.8123	0.6013	0.5901	0.5957
M^2HSE-LPA	**0.5268**	**0.5235**	**0.5252**	**0.6210**	**0.6065**	**0.6138**	**0.8388**	**0.8274**	**0.8331**	**0.6416**	**0.6289**	**0.6353**
M^2HSE-GLP	0.5403	0.5497	0.5450	0.6186	0.6217	0.6201	0.8247	0.8193	0.8220	0.6108	0.6197	0.6153
M^2HSE	**0.5715**	**0.5804**	**0.5760**	**0.6429**	**0.6421**	**0.6425**	**0.8600**	**0.8664**	**0.8632**	**0.6550**	**0.6531**	**0.6541**

Table 9
Experimental configurations of different models in ablation study 2.

Model	Subnetwork		Loss function				
	GLO	LOC	MSB		TRI	UWF	MSPN
			Step1	Step2			
M^2HSE-GPA-TRI	√				√		
M^2HSE-GPA-UWF	√					√	
M^2HSE-GPA-MSPN	√						√
M^2HSE-GPA-MSB(Step1)	√		√				
M^2HSE-GPA-MSB(Step2)	√			√			
M^2HSE-GPA	√		√	√			
M^2HSE-LPA-TRI		√			√		
M^2HSE-LPA-UWF		√				√	
M^2HSE-LPA-MSPN		√					√
M^2HSE-LPA-MSB(Step1)		√	√				
M^2HSE-GPA-MSB(Step2)		√		√			
M^2HSE-LPA		√	√	√			
M^2HSE-TRI	√	√			√		
M^2HSE-UWF	√	√				√	
M^2HSE-MSPN	√	√					√
M^2HSE-MSB(Step1)	√	√	√				
M^2HSE-MSB(Step2)	√	√		√			
M^2HSE	√	√	√	√			

Specifically, GLO and LOC represent the global-level and local-level subnetwork, respectively, PRI and AUX denote the primary and auxiliary modality, respectively. There are six combinations of the above four components, for example, M^2HSE-GPA represents the model with only the global-level subnetwork and using both primary and auxiliary modalities. The experimental results are shown in Table 8, in which all ablation models are implemented with the proposed MSB loss. Some important observations are listed as follows.

- Compared with M^2HSE-GP and M^2HSE-LP, M^2HSE-GPA and M^2HSE-LPA perform better on all datasets, which proves that the auxiliary modality plays an important role in cross-modal similarity learning in each subnetwork. To further verify the contribution of the auxiliary modality, we compare the performance of two ensemble models. Compared with M^2HSE-GLP on four datasets, M^2HSE gains the improvements of 3.10%, 2.24%, 4.12%, and 3.88% with "Avg.", respectively.
- It is apparent that M^2HSE-GLP outperforms both M^2HSE-GP and M^2HSE-LP, M^2HSE outperforms M^2HSE-GPA and M^2HSE-LPA, which proves that the performance of cross-modal retrieval can be significantly improved by integrating two subnetworks, which are constructed with the coarse-grained and the fine-grained data, respectively. Besides, M^2HSE-LP outperforms M^2HSE-GP, M^2HSE-LPA outperforms M^2HSE-GPA, due to the fact that the local-level subnetwork captures more fine-grained details and more valuable semantic information with the attention mechanism.
- Integrated with all components, M^2HSE always achieves the greatest results on all datasets. In summary, it can be concluded that: (1) each component in M^2HSE plays a very positive role for semantic enhancement in cross-modal similarity learning, (2) the auxiliary modality contains valuable complementary semantic information that does not exist in the primary modality, (3) data with different granularities emphasizes distinct and complementary views in cross-modal correlation learning, (4) M^2HSE achieves the best performance by effectively mining and fusing the complementarity in multi-modal and multi-grained data to bridge the "heterogeneity gap" and the "granularity gap".

4.6.2. Ablation study 2

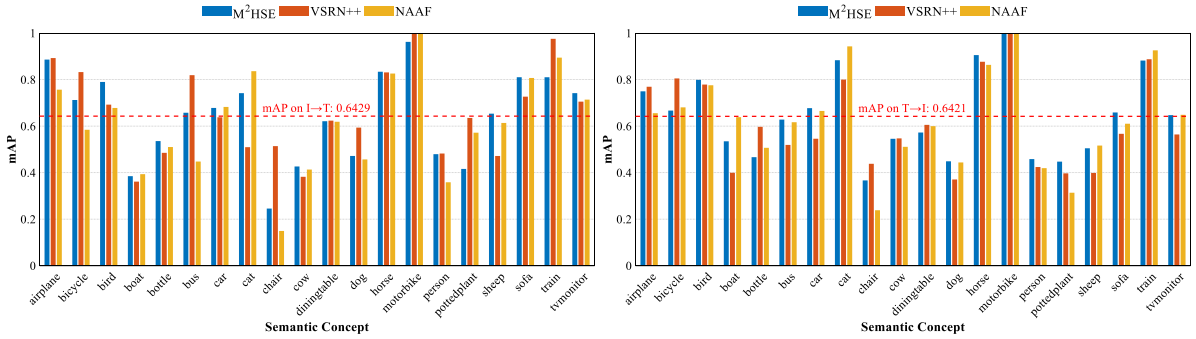
As illustrated in Table 9, we provide some ablation models to reveal the effectiveness of each step of MSB loss with GLO and LOC components in ablation study 2. Concretely, MSB(Step1) denotes that representative samples are selected through step 1, while they share the same weights, that is, all elastic coefficients are set to one in step 2. On the contrary, MSB(Step2) represents that all samples are discriminated via step 2, however, the process of representative samples selecting is omitted.

Moreover, TRI, UWF and MSPN are used to substitute for MSB in the ablation models to compare their effects on cross-modal retrieval performance. Notably, UWF and MSPN also consider the issue of selecting the most representative samples and assigning appropriate weights to optimize the cross-modal similarity. The experimental results are shown in Table 10, and more detailed analysis are presented as follows.

- The performance of MSB(Step1) and MSB(Step2) are very close to each other, whether in two subnetworks or the ensemble models, which shows that these two steps play almost the same role in cross-modal similarity optimization. Besides, MSB(Step1) and MSB(Step2) significantly outperform than triplet loss, because there are two main defects in triplet loss: (1) training samples are randomly selected, thus, some good enough samples may interfere with the optimization of network parameters, resulting in the degradation of performance. (2) positive and negative samples are separated with a pre-defined margin, which cannot be accurately optimized according to their importance. But the above two defects are well handled

Table 10
Experimental results of ablation study 2 on all datasets with mAP scores.

Model	Corel 5K			Pascal Sentence			NUS-WIDE-25K			NUS-WIDE-5K		
	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.
M^{2}HSE-GPA-TRI	0.4556	0.4601	0.4579	0.4763	0.4685	0.4724	0.7127	0.7382	0.7255	0.5406	0.5429	0.5418
M^{2}HSE-GPA-UWF	0.4864	0.4903	0.4884	0.5242	0.5235	0.5239	0.7320	0.7533	0.7427	0.5615	0.5623	0.5619
M^{2}HSE-GPA-MSPN	**0.5003**	0.5025	0.5014	**0.5411**	**0.5375**	**0.5393**	0.7501	0.7623	0.7562	0.5600	0.5644	0.5622
M^{2}HSE-GPA-MSB(Step1)	0.4709	0.4732	0.4721	0.5015	0.4917	0.4966	0.7349	0.7358	0.7354	0.5605	0.5634	0.5620
M^{2}HSE-GPA-MSB(Step2)	0.4747	0.4844	0.4796	0.4924	0.4935	0.4930	0.7402	0.7411	0.7407	0.5555	0.5629	0.5592
M^{2}HSE-GPA	0.4983	**0.5180**	**0.5082**	0.5342	0.5216	0.5279	**0.7650**	**0.7933**	**0.7792**	**0.5887**	**0.5952**	**0.5919**
M^{2}HSE-LPA-TRI	0.4644	0.4614	0.4629	0.5669	0.5446	0.5558	0.7821	0.7679	0.7750	0.5872	0.5596	0.5734
M^{2}HSE-LPA-UWF	0.5192	0.5187	0.5190	0.5883	0.5794	0.5839	0.7986	0.8035	0.8011	0.6094	0.6028	0.6061
M^{2}HSE-LPA-MSPN	0.5111	0.5203	0.5157	0.6169	**0.6146**	**0.6158**	0.8208	0.8193	0.8201	0.6106	0.6010	0.6058
M^{2}HSE-LPA-MSB(Step1)	0.4900	0.4874	0.4887	0.5598	0.5636	0.5617	0.8112	0.7947	0.8030	0.6213	0.5995	0.6104
M^{2}HSE-LPA-MSB(Step2)	0.4823	0.4851	0.4837	0.5714	0.5603	0.5659	0.8117	0.8058	0.8088	0.6126	0.6048	0.6087
M^{2}HSE-LPA	**0.5268**	**0.5235**	**0.5252**	**0.6210**	0.6065	0.6138	**0.8388**	**0.8274**	**0.8331**	**0.6416**	**0.6289**	**0.6353**
M^{2}HSE-TRI	0.4910	0.5020	0.4965	0.5738	0.5523	0.5631	0.7957	0.7742	0.7850	0.5911	0.5600	0.5756
M^{2}HSE-UWF	0.5643	0.5661	0.5652	0.6345	0.6278	0.6312	0.8364	0.8357	0.8361	0.6416	0.6329	0.6373
M^{2}HSE-MSPN	0.5667	0.5784	0.5726	**0.6442**	**0.6501**	**0.6472**	0.8472	0.8480	0.8476	0.6436	0.6439	0.6438
M^{2}HSE-MSB(Step1)	0.5328	0.5415	0.5372	0.5972	0.6014	0.5993	0.8332	0.8245	0.8289	0.6266	0.6146	0.6256
M^{2}HSE-MSB(Step2)	0.5243	0.5421	0.5332	0.6057	0.6012	0.6032	0.8217	0.8202	0.8210	0.6355	0.6292	0.6324
M^{2}HSE	**0.5715**	**0.5804**	**0.5760**	0.6429	0.6421	0.6425	**0.8600**	**0.8664**	**0.8632**	**0.6550**	**0.6531**	**0.6541**



(a) The mAP scores of I→T on each semantic concept.

(b) The mAP scores of T→I on each semantic concept.

Fig. 12. The mAP scores of I→T and T→I on each semantic concept in Pascal Sentence dataset by M^{2}HSE as well as comparison methods VSRN++ and NAAF. Note that the red horizontal dotted lines indicate the mAP scores obtained by our M^{2}HSE method.

by step 1 and step 2 of our proposed MSB loss. We can observe that M^{2}HSE-MSB(Step1) and M^{2}HSE-MSB(Step2) achieve an average improvement of 4.27% and 4.24% on all datasets than M^{2}HSE-TRI, respectively.

- Integrated with step 1 and step 2 together, MSB loss achieves better performance than only using either of them alone, whether in two subnetworks or the ensemble models. For example, compared with M^{2}HSE-MSB(Step1) on four datasets, M^{2}HSE gets a promotion of 3.88%, 4.32%, 3.43%, and 2.85% with "Avg.", respectively. For MSB(Step1), the selected samples do not be further discriminated, therefore, important samples have not received adequate attention, which exerts a negative influence on the cross-modal similarity optimization. For MSB(Step2), it is not the best strategy to discriminate all training samples in a mini-batch to optimize the cross-modal similarity, because some samples have already been close to the optimal states, and they may bring negative impacts on weight assigning.
- Compared with two related works, we observe that all ablation models integrated with MSB loss significantly outperform UWF, mainly because polynomial functions used in UWF require more hyper-parameters than MSB loss. It is easy to understand that the more hyper-parameters that a metric learning framework contains, the more difficult it is to achieve an optimal solution. In general, MSPN is only inferior to MSB loss, and it even performs the best on Pascal Sentence. However, MSPN relies on a fully connected neural network to fit the weight function, which is complex and weakly interpretable. On the contrary, MSB loss makes better use of potential interactions among the samples based on the multi-spring balance system to automatically learn weights and further adaptively optimize the cross-modal similarity. Hence, MSB loss has low computational complexity and strong interpretability.
- Besides, the local-level subnetwork always outperforms the global-level subnetwork no matter what types of loss functions are used. However, the former cannot substitute for the latter in the task of cross-modal retrieval, due to the fact that they are complementary to each other. In general, by integrating these two subnetworks, M^{2}HSE achieves a better performance than any single one subnetwork.

4.7. Qualitative results and analysis

Firstly, taking Pascal Sentence dataset as an example, the mAP scores of I→T and T→I on each semantic concept by our M^{2}HSE and two most recent DNN-based solutions VSRN++, NAAF are reported in Fig. 12. On the one hand, these three methods have different performance on I→T and T→I. Specifically, M^{2}HSE and VSRN++ outperform NAAF on most semantic concepts. Note that the performance of M^{2}HSE is not the best on all semantic concepts. On the other hand, the mAP scores of M^{2}HSE on all semantic concept vary widely. Moreover, VSRN++ and NAAF also exhibit similar patterns. That is to say, the intrinsic properties of the dataset have an impact on the performance of cross-modal retrieval. For instance, all these three methods achieve higher mAP scores on "motorbike", "horse", "airplane", "bird", "sofa", "bicycle", etc, while they achieve lower mAP scores on "chair", "boat", "cow", "dog", "person", etc.

Then, as described in Fig. 13, in order to explore the disparity of mAP scores on different semantic concepts, we provide several typical examples of cross-modal retrieval by M^{2}HSE, VSRN++ and NAAF with

(a) Examples of I→T retrieval results on Pascal Sentence dataset.

(b) Examples of T→I retrieval results on Pascal Sentence dataset.

Fig. 13. Examples of the I→T and T→I retrieval results on Pascal Sentence dataset by M^{2}HSE as well as compared methods VSRN++ and NAAF. The ground-truth semantic concept of each query is presented for instruction. Besides, the true matches are marked in green rectangles with check marks, while the incorrect retrieval results are indicated by red rectangles and cross marks.

the Pascal Sentence dataset, and display the top five results of I→T and T→I corresponding to a specific query. Notably, for each cross-modal retrieval task, we select four queries that belong to different semantic concepts. Particularly, higher mAP scores are achieved on the first two queries, while lower mAP scores are acquired on the latter two.

Specifically, on I→T, VSRN++ shows better overall performance, and the results returned by the first two queries are all correct. However, all methods make more mistakes under the queries of "boat" and "chair". The reason is that the images about "boat" and "chair" contain many objects, which may be occluded by other salient objects. Therefore, the returned wrong results are mostly influenced by these category-irrelevant and semantic-relevant objects, such as "cat" and "pottedplant" in the fourth image. Besides, on T→I, the images searched by M^{2}HSE under four queries contain less wrong results than VSRN++ and NAAF, while it also makes more mistakes in the latter two queries (*i.e.*, "dog" and "person"). It can be observed that the query texts about "dog" and "person" contain rich semantic information, which leads M^{2}HSE to make mistakes in the task of semantic category recognition. In conclusion, M^{2}HSE has some limitations when dealing with the images and texts containing complex scenes, but it still exhibits better overall performance than VSRN++ and NAAF.

5. Conclusion

In this paper, we propose a multi-modal and multi-grained hierarchical semantic enhancement network to handle the task of cross-modal retrieval. To bridge the "heterogeneity gap" and the "granularity gap", M^{2}HSE follows the key idea "complementarity is the king", which well conforms to the intrinsic rules of the sophisticated and disordered distributions of semantic information. Specifically, M^{2}HSE aims to gather and fuse all semantic pieces scattered over various modalities and various granularities through two stages. The initial primary and auxiliary cross-modal similarity are calculated with coarse-grained and fine-grained data in the first stage, and all types of cross-modal similarities are optimized by the MSB loss in the second stage.

Experimental results demonstrate the superiority of our proposed M^{2}HSE compared with 18 state-of-the-art methods on several widely-used cross-modal datasets, and ablation studies further verify the effectiveness of each component in M^{2}HSE. Concretely, compared with the previous best model VSRN++, the average mAP scores of M^{2}HSE on Corel 5K, Pascal Sentence, NUS-WIDE-25K, and NUS-WIDE-5K, are obviously improved by 1.92%, 1.35%, 1.56%, and 1.62%, respectively. Besides, our M^{2}HSE outperforms VSRN++ by a margin of 14.2%, 12.5%, 12.0%, and 13.2% in terms of the overall performance R@sum on these datasets, respectively.

Note that we also find some limitations of our method through experiments, for instance, as M^{2}HSE does not perform well with complex multi-modal scenarios, and the training time of M^{2}HSE is not very competitive. Therefore, the future works mainly lie in four aspects. Firstly, we will try to extend two levels of granularity to multiple levels, and further deeply mine the complementarity among them. Secondly, as a kind of important semantic information, the positional relations among fine-grained patches of each modality will be considered in cross-modal similarity learning. Thirdly, we will attempt to accelerate the training process of M^{2}HSE through improving its architecture. Fourthly, we will verify the scalability of M^{2}HSE, that is, other types of modality will be used to perform the cross-modal retrieval, for example, video queries text.

CRediT authorship contribution statement

Xinlei Pei: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Zheng Liu:** Investigation, Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition. **Shanshan Gao:** Conceptualization, Writing – review & editing. **Yijun Su:** Data curation, Visualization.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by Humanities and Social Sciences Project of Education Ministry (20YJA870013), Natural Science Foundation of Shandong Province (ZR2019MF016, ZR2020MF037), NSFC-Zhejiang Joint Fund of the Integration of Informatization and Industrialization (U1909210), Introduction and Education Plan of Young Creative Talents in Colleges and Universities of Shandong Province, Scientific Research Studio in Colleges and Universities of Ji'nan City (2021GXRC092), Research Project of Undergraduate Teaching Reform in Shandong Province (Z2020025), Key Research and Development Project of Shandong Province (2019GSF109112), Innovation Team of Youth Innovation Science and Technology Plan in Colleges and Universities of Shandong Province (2020KJN007).

Appendix A

Here we present the details of the derivation process of the proposed MSB loss. The gradient of $\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})$ with respect to each cross-modal similarity score $\boldsymbol{M}_{ij}$ is defined in Eq. (26), which is designed based on the multi-spring balance system for positive samples or negative samples. Eq. (26) can be transformed to the following equation according to Eqs. (20)–(23):

$$\frac{\partial\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})}{\partial M_{ij}}=\begin{cases}-\dfrac{K_{ij}}{\sum_{\tilde{C}_{ik}=+1}K_{ik}}=-\dfrac{G(\tilde{C}_{ij},M_{ij})}{\sum_{\tilde{C}_{ik}=+1}G(\tilde{C}_{ik},M_{ik})}\\ =-\dfrac{exp(\tilde{C}_{ij}(\gamma_2-\gamma_1M_{ij}))}{\sum_{\tilde{C}_{ik}=+1}exp(\tilde{C}_{ik}(\gamma_2-\gamma_1M_{ik}))} & \text{if } \tilde{C}_{ij}=+1,\\ \dfrac{K_{ij}}{\sum_{\tilde{C}_{ik}=-1}K_{ik}}=\dfrac{G(\tilde{C}_{ij},M_{ij})}{\sum_{\tilde{C}_{ik}=-1}G(\tilde{C}_{ik},M_{ik})}\\ =\dfrac{exp(\tilde{C}_{ij}(\gamma_2-\gamma_1M_{ij}))}{\sum_{\tilde{C}_{ik}=-1}exp(\tilde{C}_{ik}(\gamma_2-\gamma_1M_{ik}))} & \text{if } \tilde{C}_{ij}=-1.\end{cases}\tag{29}$$

Thus, we have:

$$\frac{\partial\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})}{\partial M_{ij}}=\begin{cases}-\dfrac{exp(\gamma_2-\gamma_1M_{ij})}{\sum_{\tilde{C}_{ik}=+1}exp(\gamma_2-\gamma_1M_{ik})} & \text{if } \tilde{C}_{ij}=+1,\\ \dfrac{exp(-\gamma_2+\gamma_1M_{ij})}{\sum_{\tilde{C}_{ik}=-1}exp(-\gamma_2+\gamma_1M_{ik})} & \text{if } \tilde{C}_{ij}=-1.\end{cases}\tag{30}$$

Then, we obtain the primitive integral (*i.e.* $\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})$) of Eq. (30) according to the calculus theory:

$$\begin{aligned}\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})&=\int\frac{\partial\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})}{\partial M_{ij}}dM_{ij}\\&=\int\Bigg(-\frac{exp(\gamma_2-\gamma_1M_{ij})}{\sum_{\tilde{C}_{ik}=+1}exp(\gamma_2-\gamma_1M_{ik})}\\&\quad+\frac{exp(-\gamma_2+\gamma_1M_{ij})}{\sum_{\tilde{C}_{ik}=-1}exp(-\gamma_2+\gamma_1M_{ik})}\Bigg)dM_{ij}\end{aligned}\tag{31}$$

Note that $\boldsymbol{M}$ is a matrix of size $B*B$, where B is the mini-batch size in training, so we can infer that:

$$\begin{aligned}\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})&=\sum_{i=1}^{B}\Bigg\{\frac{1}{\gamma_1}\ln\Bigg[\sum_{\tilde{C}_{ik}=+1}exp(\gamma_2-\gamma_1M_{ik})\Bigg]\\&\quad+\frac{1}{\gamma_1}\ln\Bigg[\sum_{\tilde{C}_{ik}=-1}exp(-\gamma_2+\gamma_1M_{ik})\Bigg]\Bigg\}\\&=\frac{1}{\gamma_1}\sum_{i=1}^{B}\Bigg\{\ln\Bigg[\sum_{\tilde{C}_{ik}=+1}exp(\gamma_2-\gamma_1M_{ik})\Bigg]\\&\quad+\ln\Bigg[\sum_{\tilde{C}_{ik}=-1}exp(-\gamma_2+\gamma_1M_{ik})\Bigg]\Bigg\}\end{aligned}\tag{32}$$

According to Eqs. (20) and (21), the exponential terms in Eq. (32) refer to the elastic coefficients in the multi-spring balance systems, then Eq. (32) can be rewritten as follows:

$$\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})=\frac{1}{\gamma_1}\sum_{i=1}^{B}\Bigg\{\ln\Bigg[\sum_{\tilde{C}_{ik}=+1}K_{ik}\Bigg]+\ln\Bigg[\sum_{\tilde{C}_{ik}=-1}K_{ik}\Bigg]\Bigg\}\tag{33}$$

Finally, a fraction term is added for normalization:

$$\mathcal{H}(\boldsymbol{M},\boldsymbol{\Theta})=\frac{1}{\gamma_1}\frac{1}{B}\sum_{i=1}^{B}\Bigg\{\ln\Bigg[\sum_{\tilde{C}_{ik}=+1}K_{ik}\Bigg]+\ln\Bigg[\sum_{\tilde{C}_{ik}=-1}K_{ik}\Bigg]\Bigg\}\tag{34}$$

References

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 1247–1255). https://dl.acm.org/doi/10.5555/3042817.3043076.

Baltrušaitis, T., Ahuja, C., & Morency, L. -P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*, 423–443. http://dx.doi.org/10.1109/tpami.2018.2798607.

Chen, T., Deng, J., & Luo, J. (2020). Adaptive offline quintuplet loss for image-text matching. In *European conference on computer vision* (pp. 549–565). Springer, http://dx.doi.org/10.1007/978-3-030-58601-0_33.

Chen, H., Ding, G., Lin, Z., Zhao, S., & Han, J. (2019). Cross-modal image-text retrieval with semantic consistency. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1749–1757). http://dx.doi.org/10.1145/3343031.3351055.

Cheng, Y., Zhu, X., Qian, J., Wen, F., & Liu, P. (2022). Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *18*, 1–23. http://dx.doi.org/10.1145/3499027.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. http://dx.doi.org/10.3115/v1/d14-1179, arXiv preprint.

Chua, T. -S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). Nus-wide: A real-world web image database from national university of Singapore. In *Proceedings of the ACM international conference on image and video retrieval* (pp. 1–9). http://dx.doi.org/10.1145/1646396.1646452.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision: Vol. 1*, (pp. 1–2). Prague.

Deng, J., Dong, W., Socher, R., Li, L. -J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE, http://dx.doi.org/10.1109/cvpr.2009.5206848.

Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Diao, H., Zhang, Y., Ma, L., & Lu, H. (2021). Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence: Vol. 35*, (pp. 1218–1226). http://dx.doi.org/10.1609/aaai.v35i2.16209.

Duygulu, P., Barnard, K., de Freitas, J. F. G., & Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision* (pp. 97–112). Springer, http://dx.doi.org/10.1007/3-540-47979-1_7.

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, *26*.

Ge, W. (2018). Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision* (pp. 269–285). http://dx.doi.org/10.1007/978-3-030-01231-1_17.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*, 139–144.

Gu, J., Cai, J., Joty, S. R., Niu, L., & Wang, G. (2018). 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* (pp. 7181–7189). http://dx.doi.org/10.1109/cvpr.2018.00750.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2*, In *2006 IEEE computer society conference on computer vision and pattern recognition* (pp. 1735–1742). IEEE, http://dx.doi.org/10.1109/cvpr.2006.100.

Harwood, B., Kumar BG, V., Carneiro, G., Reid, I., & Drummond, T. (2017). Smart mining for deep metric learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2821–2829). http://dx.doi.org/10.1109/iccv.2017.307.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hoffman, P., Grinstein, G., Marx, K., Grosse, I., & Stanley, E. (1997). DNA visual and analytic data mining. In *Proceedings. Visualization'97 (Cat. No. 97CB36155)* (pp. 437–441). IEEE, http://dx.doi.org/10.1109/visual.1997.663916.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, http://dx.doi.org/10.1145/3065386.

Lee, K. -H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision* (pp. 201–216). http://dx.doi.org/10.1007/978-3-030-01225-0_13.

Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2022). Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/tpami.2022.3148470.

Liong, V. E., Lu, J., Tan, Y. -P., & Zhou, J. (2016). Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, *19*, 1234–1244. http://dx.doi.org/10.1109/tmm.2016.2646180.

Liu, Y., Wu, J., Qu, L., Gan, T., Yin, J., & Nie, L. (2022). Self-supervised correlation learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, http://dx.doi.org/10.1109/tmm.2022.3152086.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110. http://dx.doi.org/10.1023/b:visi.0000029664.99615.94.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch.

Peng, Y., Huang, X., & Zhao, Y. (2017). An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*, 2372–2385. http://dx.doi.org/10.1109/tcsvt.2017.2705068.

Peng, Y., Qi, J., Huang, X., & Yuan, Y. (2017). CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, *20*, 405–420. http://dx.doi.org/10.1109/tmm.2017.2742704.

Peng, Y., Qi, J., & Zhuo, Y. (2019). MAVA: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism. *IEEE Transactions on Image Processing*, *29*, 2728–2741. http://dx.doi.org/10.1109/tip.2019.2952085.

Peng, Y., Zhai, X., Zhao, Y., & Huang, X. (2015). Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*, 583–596. http://dx.doi.org/10.1109/tcsvt.2015.2400779.

Qu, L., Liu, M., Wu, J., Gao, Z., & Nie, L. (2021). Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1104–1113). http://dx.doi.org/10.1145/3404835.3462829.

Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). 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* (pp. 139–147).

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R. G., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 251–260).

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823). http://dx.doi.org/10.1109/cvpr.2015.7298682.

Ustinova, E., & Lempitsky, V. (2016). Learning deep embeddings with histogram loss. *Advances in Neural Information Processing Systems*, *29*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, K., He, R., Wang, L., Wang, W., & Tan, T. (2015). Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*, 2010–2023. http://dx.doi.org/10.1109/tpami.2015.2505311.

Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017). Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 154–162).

Wei, J., Xu, X., Wang, Z., & Wang, G. (2021). Meta self-paced learning for cross-modal matching. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3835–3843). http://dx.doi.org/10.1145/3474085.3475451.

Wei, J., Yang, Y., Xu, X., Zhu, X., & Shen, H. T. (2021). Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, http://dx.doi.org/10.1109/tpami.2021.3088863.

Wei, X., Zhang, T., Li, Y., Zhang, Y., & Wu, F. (2020). Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10941–10950). http://dx.doi.org/10.1109/cvpr42600.2020.01095.

Xu, X., Lin, K., Yang, Y., Hanjalic, A., & Shen, H. T. (2020). Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, http://dx.doi.org/10.1109/tpami.2020.3045530.

Zhai, X., Peng, Y., & Xiao, J. (2012a). Cross-modality correlation propagation for cross-media retrieval. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 2337–2340). IEEE, http://dx.doi.org/10.1109/icassp.2012.6288383.

Zhai, X., Peng, Y., & Xiao, J. (2012b). Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval. In *International conference on multimedia modeling* (pp. 312–322). Springer, http://dx.doi.org/10.1007/978-3-642-27355-1_30.

Zhai, X., Peng, Y., & Xiao, J. (2013a). Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *Twenty-seventh AAAI conference on artificial intelligence*. http://dx.doi.org/10.1609/aaai.v27i1.8464.

Zhai, X., Peng, Y., & Xiao, J. (2013b). Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*, 965–978. http://dx.doi.org/10.1109/tcsvt.2013.2276704.

Zhang, K., Mao, Z., Wang, Q., & Zhang, Y. (2022). Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15661–15670). http://dx.doi.org/10.1109/cvpr52688.2022.01521.

Zheng, L., Yang, Y., & Tian, Q. (2017). SIFT meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 1224–1244. http://dx.doi.org/10.1109/tpami.2017.2709749.