IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 March 2023, accepted 26 March 2023, date of publication 31 March 2023, date of current version 17 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263512

RESEARCH ARTICLE

FB-Net: Dual-Branch Foreground-Background Fusion Network With Multi-Scale Semantic Scanning for Image-Text Retrieval

JUNHAO XU[1], ZHENG LIU[1,2], XINLEI PEI[1,2], SHUHUAI WANG[1], AND SHANSHAN GAO[1,2,3]

[1]School of Computer Science and Technology, Shandong University of Finance and Economics, Ji'nan, Shandong 250014, China
[2]Shandong Provincial Key Laboratory of Digital Media Technology, Ji'nan, Shandong 250014, China
[3]Shandong China–U.S. Digital Media International Cooperation Research Center, Ji'nan, Shandong 250014, China

Corresponding author: Zheng Liu (Liuzheng@sdufe.edu.cn)

This work was supported in part by the Humanities and Social Sciences Project of Education Ministry under Grant 20YJA870013; in part by the Natural Science Foundation of Shandong Province under Grant ZR2019MF016 and Grant ZR2020MF037; in part by NSFC-Zhejiang Joint Fund of the Integration of Informatization and Industrialization under Grant U1909210; in part by the Scientific Research Studio in Colleges and Universities of Ji'nan City under Grant 202228105 and Grant 2021GXRC092; in part by the Introduction and Education Plan of Young Creative Talents in Colleges and Universities of Shandong Province, Research Project of Undergraduate Teaching Reform in Shandong Province, under Grant Z2020025; in part by the Key Research and Development Project of Shandong Province under Grant 2019GSF109112; and in part by the Innovation Team of Youth Innovation Science and Technology Plan in Colleges and Universities of Shandong Province under Grant 2020KJN007.

ABSTRACT As a fundamental branch in cross-modal retrieval, image-text retrieval is still a challenging problem largely due to the complementary and imbalanced relationship between different modalities. However, existing works have not effectively scanned and aligned the semantic units distributed in different granularities of images and texts. To address these issues, we propose a dual-branch foreground-background fusion network (FB-Net), which is implemented by fully exploring and fusing the complementarity in semantic units collected from the foreground and background areas of instances (e.g., images and texts). Firstly, to generate multi-granularity semantic units from images and texts, multi-scale semantic scanning is conducted on both foreground and background areas through multi-level overlapped sliding windows. Secondly, to align semantic units between images and texts, the stacked cross-attention mechanism is used to calculate the initial image-text similarity. Thirdly, to further adaptively optimize the image-text similarity, the dynamically self-adaptive weighted loss is designed. Finally, to perform the image-text retrieval, the similarities between multi-granularity foreground and background semantic units are fused to obtain the final image-text similarity. Experimental results show that our proposed FB-Net outperforms representative state-of-the-art methods for image-text retrieval, and ablation studies further verify the effectiveness of each component in FB-Net.

INDEX TERMS Image-text retrieval, foreground-background fusion, multi-scale semantic scanning, semantic unit, overlapped sliding window.

I. INTRODUCTION

In the digital multimedia era, a large amount of multimedia data is being generated every moment, and digital multimedia information has exploded in recent years. Cross-modal retrieval [1] aims to implement a retrieval task across different media types, such as retrieving images using sounds or retrieving videos using images. The key issue of information retrieval is to rank the relevant items by calculating the similarities between the queries and the relevant items. However, it is impossible to directly measure the cross-modal similarity due to the inconsistency of feature encoding in different modalities, that is, "heterogeneity gap [2]." Particularly, there are many popular cross-media datasets that are

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

used to test the performance of cross-modal retrieval [24], [32], [33], [41]. In addition, deep learning-based approaches tend to outperform traditional methods in most fields, such as face recognition. Especially, convolutional neural network (CNN) has shown its superiority in many applications. Therefore, deep learning technology provides a huge opportunity for cross-modal retrieval.

As a typical pattern of cross-modal retrieval, the researches on image-text retrieval have a positive effect to promote the development of cross-modal retrieval. Previous research [3]on image-text retrieval used more straightforward strategies to mine semantic information, for instance, mapping the original features of images and texts to one embedding space [4]. In this stage, the research of image-text retrieval mainly focuses on the global-level instances [5], while more detailed semantic information existing in the local-level patches is ignored. Therefore, to capture more complete semantic information, numerous studies about image-text retrieval have concentrated on the local-level fragments [15], [16], [17]. Specifically, to generate fine-grained fragments for images, the frequently-used partition strategies are uniform blocking or salient object detection [38]. Similarly, fine-grained fragments of texts are obtained by simply cutting the texts into several sentences or lots of words. However, the above-mentioned methods of fine-grained data partition may destroy the integrity of semantic units, which brings negative impacts on semantic understanding and further degrades the performance of image-text retrieval.

For the task of image-text retrieval, the semantic unit refers to a specific semantic concept with its attributes, such as the breed of a dog, the color of a car, the architectural style of a building, etc. Particularly, a semantic unit is corresponding to a local region of an image or a sequence of words in a text. It is actually difficult to accurately locate the semantic units in images and texts. Therefore, it is of great importance to correctly divide the fine-grained fragments with fully covering the semantic units. However, previous works cannot well solve the problem of semantic unit capturing, and the reasons lie in the following aspects: 1) Uniformly dividing an image may cut an object into multiple blocks, 2) The techniques of object detection can only find the salient objects in the image, while totally ignoring the semantic units in the background of the image, 3) The simple policy to partition a text into sentences or words may also fail to completely locate semantic units.

Although great progress in image-text retrieval has been made in recent years, it still faces many difficulties and challenges. In particular, the following key issues should be solved.

- **The semantic units existing in the foreground and background areas should be completely identified.** Inspired by related studies in computer vision, we classify the semantic units in images and texts into two types: 1) the foreground semantic unit, and 2) the background semantic unit. It is worth noting that there actually exists a complementary relationship between them. Unlike the tasks in computer vision, such as salient object detection, these two types of semantic units in images and texts are equally essential and complementary to each other for image-text retrieval. However, previous works on image-text retrieval have paid more attention to identifying the foreground semantic units, which can be extracted by image segmentation or salient object detection. Therefore, to enhance the performance of image-text retrieval, the complementary relationship between foreground and background semantic units should be carefully explored and exploited.
- **The semantic units distributed in different data granularities should be accurately captured.** Note that the semantic information existing in image blocks and word sequences at different levels of data granularities are complementary to each other. Furthermore, the complete semantic information of images and texts is widely distributed in different levels of data granularities. However, the goal of completely generating semantic units cannot be achieved by existing strategies about fine-grained fragments partitioning for images and texts, and the main reasons contain the following two folds. First, semantic units of images and texts do exist in different levels of data granularities, therefore, partitioning images and texts into single-grained local regions cannot identify all the semantic units. Second, previous works often use the uniform grid division scheme to partition fine-grained fragments, such as cutting an image into blocks of the same size or dividing a sentence into words. Such a simple strategy may cut a complete semantic unit into several pieces, and thus result in incomplete semantic units.

To better illustrate the complementary relationship between foreground and background semantic units in images and texts at different levels of granularities, we provide an image-text pair about the topic ‘‘Kayaking’’ in Figure 1, in which two things are discovered. First, the foreground semantic units and the background semantic units focus on different aspects of ‘‘Kayaking’’. Specifically, the former pays much attention to the attributes and components of the kayak (i.e., red kayak) and the actions of the two athletes (i.e., paddle a red kayak), while, the latter concentrates on the surrounding environment (i.e., on a clear lake). Second, semantic units do exist at different levels of granularity. For the text, semantic units distribute at the level of word sequences, such as ‘‘paddle a red kayak’’ and ‘‘a clear blue sky’’, or exist at the level of a single word, such as ‘‘shadows’’ and ‘‘backpacks’’. For the image, semantic units exist in different sizes of image blocks, due to the fact that a single receptive field cannot ‘‘observe’’ all semantic units in images. For instance, the semantics units distribute at the level of the whole image, or at the level of multi-granularity of foreground and background image blocks, such as ‘‘man wearing life jacket’’ and ‘‘paddle’’.

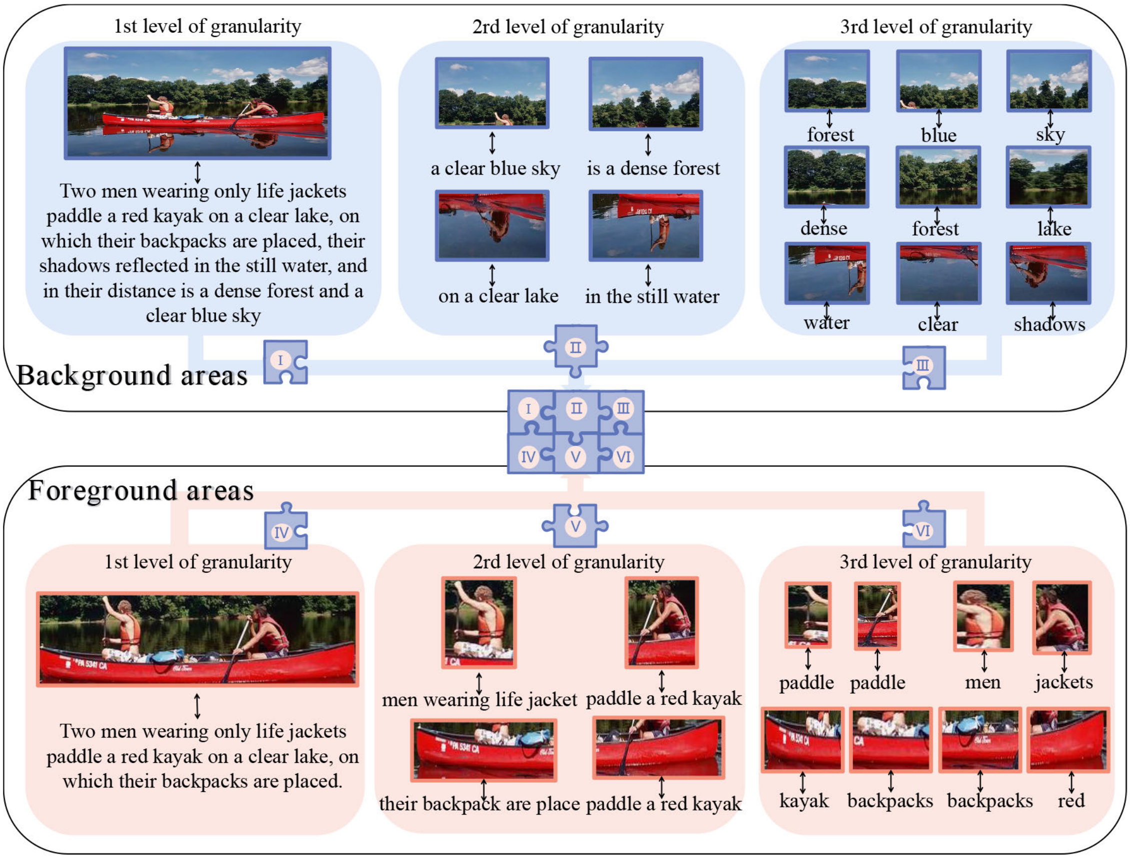


FIGURE 1. Examples of semantic unit distribution in images and texts. The complete semantic information (*i.e.*, the jigsaw with pieces I, II, III, IV, V, and VI) about the topic "Kayaking" is distributed in lots of semantic units, which are captured from both Background areas (*i.e.*, pieces I, II, and III) and Foreground areas (*i.e.*, pieces IV, V, and VI) of image and text at different levels of data granularities.

Consequently, to capture more complete semantic information and thus promote the performance of image-text retrieval, we would like to identify and collect as many semantic units as possible. Thus, we design the multi-scale semantic scanning strategy to capture semantic units with multi-level overlapped sliding windows. The above analysis inspires us to solve the image-text retrieval problem from a new perspective, that is, completely identifying and accurately partitioning the semantic units that are distributed in the foreground and background areas in images and texts at different levels of granularities. Therefore, in this work, we propose a two-branch foreground-background fusion network via multi-scale semantic scanning to achieve two goals: 1) Completely identifying the semantic units existing in the foreground and background areas, and 2) Accurately dividing the semantic units distributed in different data granularities. The main contributions of this work are summarized as follows:

- **A dual-branch foreground-background fusion network is proposed to solve the task of image-text retrieval**. The proposed network consists of: 1) Multi-granularity foreground semantic unit mining sub-network (F-Net), and 2) Multi-granularity background semantic unit mining sub-network (B-Net). These two sub-networks aim to thoroughly capture the multi-granularity semantic units in the foreground and background areas. Afterward, two types of image-text similarities are learned separately from F-Net and B-Net.
- **The multi-scale semantic scanning and alignment are proposed to accurately capture the semantic units in images and texts.** First, we scan the images and texts with multi-level overlapped sliding windows to obtain the semantic units. Notably, to ensure that the semantic units are captured as completely as possible, the adjacent sliding windows are overlapped with each other to maintain the continuity of local regions. Second, the stacked cross-attention is used to realize the alignment between the semantic units of images and texts, and the matched pairs of semantic units contain more complete semantic information than the unmatched ones.
- **Dynamically Self-Adaptive Weighted loss (DSAW for short) is proposed to optimize the image-text similarity through dynamically assigning proper weights to each selected sample**. Specifically, DSAW loss assigns higher weight scores to two types of samples: 1) positive

samples with lower similarity to the anchor, and 2) negative samples with higher similarity to the anchor. Particularly, DSAW loss estimates the importance of each selected sample by comparing it with other selected ones.

The remainder of this paper is organized as follows. We review related work in Section II. In Section III, we describe the proposed FB-Net. Then, Section IV discusses the extensive experiments and compares our method with the state-of-the-art of image-text retrieval. Finally, we conclude the paper in Section V.

II. RELATED WORK

In recent years, due to the rapid development of deep learning models in computer vision and natural language processing, image-text retrieval has also received increasing attention. Due to the variety of multi-modal data, it is crucial to align image and text features. In the following, existing studies of image-text retrieval are classified into two categories: 1) Global alignment and 2) Local alignment.

A. GLOBAL ALIGNMENT

In global alignment, many studies explore mapping images and sentences to a common embedding space. Thus, a global representation is generated for each image-text pair. Some typical traditional approaches, such as CCA [3] and its variants [4], [5], [6], project visual and textual features into a common embedding space to learn their mapping matrix before applying deep neural networks (DNNs) to image-text retrieval. With recent advances in deep learning for image and text modeling, deep learning approaches have become the mainstream for image-text retrieval. Andrew et al. [7] proposed Deep Typical Correlation Analysis (DCCA), which combines traditional CCA with deep networks to maximize the correlation of two sub-networks. Feng et al. [8] proposed a correspondence self-encoder model (Corr-AE) to reduce the correlation learning errors between different modalities. Then, Rehman et al. [43] comprehensively compute the performance of three modified versions of the Correspondence Auto Encoder (Corr-AE). Faghri et al. [9] proposed a two-stream global feature learning network and computed pairwise similarity by global features. After this, Zheng et al. [10] further improved the two-stream network architecture to align the full global features better. Recently, Jia et al. [11] used a global alignment-driven approach of pre-processing and fine-tuning, which also yields satisfactory results. Gu et al. [28] conducted a study to investigate richer representations by integrating the generative model into image-text embedding.

However, the above approaches only consider global-level semantic information in images and texts. A single level of data granularity cannot well bridge the ‘‘granularity gap’’. Therefore, image-text retrieval with only global alignment cannot obtain excellent performance, because fine-grained fragments are also useful for image-text similarity learning.

B. LOCAL ALIGNMENT

Local alignment refers to learning the correspondence between local regions in the image and words in the sentence. Fine-grained fragments can provide rich complementary information for coarse-grained instances. Recently, many works have fully explored the fine-grained alignment between local image regions and keywords. First, Peng et al. [12] proposed cross-modal correlation learning (CCL) to model inter-modal and intra-modal correlations from global and local information. After that, Karpathy et al. [13] proposed a deep neural network model to extract fine-grained features for image regions. Thus, the potential alignment is inferred between sentences and image regions. Unlike previous work embedding image regions and sentences, Karpathy et al. [14] further divided sentences into words to model the association of finer-grained texts with image regions. Rehman et al. [24] proposed a semantic-based cross-media retrieval method, which discussed a new similarity measurement in the embedded space to conduct cross-media retrieval.

Thereafter, Lee et al. [15] proposed the stacked cross-attention mechanism (SCAN), which uses an attention mechanism to align all the fine-grained fragments of the two modalities of images and texts, and the correspondence of the fine-grained fragments are deeply mined to infer the image-text similarity. Chen et al. [16] and Kim et al. [17] used the attention mechanism to explore the correspondence between local images and words, and further mined fine-grained fragments to infer image-text similarity. However, due to the complexity of semantic information mining, these approaches may not capture the best fine-grained correspondences very well. Liu et al. [18] proposed the Bidirectional Focused Attention Network (BFAN), which eliminates small weight-contributing regions and redistributes attention to the weight-bearing regions. Zhang et al. [19] proposed a context-aware attention network (CAAN) that uses inter-modal relationships between image regions and words to complement and enhance each other for image and sentence matching. After that, Zhuge et al. [20] proposed an alignment-guided masking strategy. Diao et al. [21] developed attention-filtering techniques for performing the local alignment. Wang et al. [22] proposed a Position-Focused Attention Network (PFAN) to improve the accuracy of image-text similarity.

Although the above works have achieved some achievements in image-text retrieval, more and more researchers are finding that it is difficult to further improve the performance only with fine-grained fragments.Â Therefore, Ma et al. [42] proposed a global and local semantics-preserving-based deep hashing method for cross-modal retrieval. In fact, different granularities of data are complementary to each other when representing specific semantic concepts.

Therefore, this paper focuses on obtaining the alignment relationships between multi-granularity foreground areas and multi-granularity background areas of images and texts, which may provide additional clues for image-text retrieval.

There is a significant complementary relationship between the foreground target and background areas of an image. Meanwhile, for the text, foreground words are used to describe the foreground semantics, such as salient objects, while background words are exploited to represent the background semantics, such as the atmosphere, environment and so on. Hence, these two types of words are also complementary to each other.

III. THE PROPOSED METHOD

In this section, we will elaborate on the proposed Dual-branch Foreground-background Fusion Network (FB-Net) to improve the performance of image-text retrieval. Particularly, FB-Net fully exploits the complementary relationships between the foreground and background semantic units of different granularities.

A. PROBLEM FORMULATION

As a typical representative of cross-modal retrieval, image-text retrieval includes two retrieval tasks: 1) image-to-text (i2t), and 2) text-to-image (t2i).

Definition 1 (Image-Text Retrieval): Let $\mathcal{D}$ be a multimedia database that contains n image samples $\{\boldsymbol{I}_i\}_{i=1}^{n}$ and m text samples $\left\{\boldsymbol{T}_j\right\}_{j=1}^{m}$, namely, $\mathcal{D} = \left\{\{\boldsymbol{I}_i\}_{i=1}^{n}, \left\{\boldsymbol{T}_j\right\}_{j=1}^{m}\right\}$. Let Q_I be an image query, the image-to-text retrieval is to return a set of texts R_T. Additionally, for a text query Q_T, the text-to-image retrieval returns a set of images R_I.

To achieve the Multi-scale Semantic Scanning, n level overlapped sliding windows are defined in our work. Particularly, we provide the definitions of the uth ($1 \leq u \leq n$) level sliding window for image and text respectively.

Definition 2 (The uth Level Overlapped Sliding Window for Image): Suppose that the width and height of the image are w and h respectively, and thus the width and height of the sliding window are $win\_w = \lceil w/x \rceil$ and $win\_h = \lceil h/x \rceil$ respectively, where $x = \frac{u+1}{2}$. Furthermore, the sliding step at the direction of the width and height is set to $stepsize\_w = w/2x$, and $stepsize\_h = h/2x$ respectively. Therefore, after the process of sliding, the total number of image blocks is $M = (2x-1)^2 = u^2$.

Definition 3 (The uth Level Overlapped Sliding Window for Text): For text T with K words, we suppose that the set of word features of T is represented as $T_q = \{\mathrm{w}_1, \ldots, \mathrm{w}_k, \ldots, \mathrm{w}_K\}$. st is the sliding step, $win = u$ is the window size, $t_{\mathrm{n}}^{\text{level } u}$ is the text feature in the nth window of the uth overlapped level sliding window, and the text features in the overlapped sliding window are merged: $\boldsymbol{t}_n^{\text{level } u} = \sum_{g=1+st\times(n-1)}^{st\times(n-1)+win} w_g, n \in [1, N]$, where N is the number of merged text areas, $N = \left\lfloor \frac{K-win}{st} \right\rfloor + 1$.

B. FRAMEWORK OF FB-Net

In this paper, we propose a dual-branch foreground-background fusion network based on the stacked cross-attention mechanism. The complete recognition of semantics in both foreground and background areas significantly improves the accuracy of image-text retrieval. Specifically, FB-Net contains two subnetworks: 1) F-Net and 2) B-Net.

Algorithm 1 The Optimization Procedure of B-Net in Proposed FB-Net

```
Input: 1) The multi-modal training set $\Omega = \{(I_i, T_i)\}_{i=1}^{N^t}$ and validation set $\Phi = \{(I_i, T_i)\}_{i=1}^{N^v}$,the number of epochs $E$, the batch size $B$ the learning rate $\eta$ and the hyper-parameters $\beta_1, \beta_2 \ldots \beta_u$.
Output: The optimized parameters $\tilde{\boldsymbol{\Theta}}^{\mathrm{B}}$ of the B-Net.
Initialize parameters $\boldsymbol{\Theta}^B$of the B-Net and set $MAP_{\max} = 0$
for $\delta = 1, 2, \ldots, E$ do
    for $\rho = 1, 2, \ldots, \lceil N^t/B \rceil$ do
        Randomly sample $B$ image-text pair from $\Omega$ to construct a mini-batch;
        Compute the similarity of sample pairs, then construct the B-Net matrix $S^{\text{background}}$, ..., $S^{\text{background } u}$ with SUA by the forward propagation;
        Calculate the gradients $\nabla_{\boldsymbol{\Theta}}{}^{B}$ by the backward propagation;
        Update the parameters $\boldsymbol{\Theta}^B$ through descending their stochastic gradients: $\boldsymbol{\Theta}^B \leftarrow \boldsymbol{\Theta}^B - \eta\nabla_{\boldsymbol{\Theta}^B}$;
    end for
    Linearly weighting of the B-Net matrix matrix: $\beta_1 S^{\text{background } 1} + \ldots + \beta_u S^{\text{background } u}$
    Validate the performance of the current cross-modal retrieval model on $\Phi$, then obtain $MAP_{\text{curr}}$;
    if $MAP_{\max} > MAP_{curr}$ then
        $MAP_{\max} = MAP_{curr}$
        $\tilde{\Theta}^B = \Theta^B$;
    end if
end for
```

Object detection techniques are used to extract foreground areas through F-Net. The foreground areas of images and texts are input to the multi-level overlapped sliding windows for feature extraction. By designing sliding windows of different scales, we construct n levels of multi-granularity subnetworks. To align semantic units between images and texts, the stacked cross-attention mechanism is used to calculate the initial image-text similarity. Finally, to perform the image-text retrieval, the similarities between multi-granularity foreground semantic units are fused and the final similarity matrix is represented as $S^{\text{foreground } 1}, \ldots, S^{\text{foreground } n}$. Particularly, $S^{\text{foreground } 1}$ represents the coarse-grained foreground subnetwork.

For the uth level of granularity of F-Net (F-Net(L_u)), the set of feature vectors for foreground image areas is as follows:

$$I_p^{\text{level } u} = \left\{ f_1^{\text{level } u}, \ldots, f_m^{\text{level } u}, \ldots, f_{\mathrm{M}}^{\text{level } u} \right\}, \forall \mathrm{m} \in [1, \mathrm{M}] \quad (1)$$

where $f_m^{\text{level } u}$ is the feature vector of the mth image area at the uth level of granularity for F-Net, M is the number of foreground areas of the image.

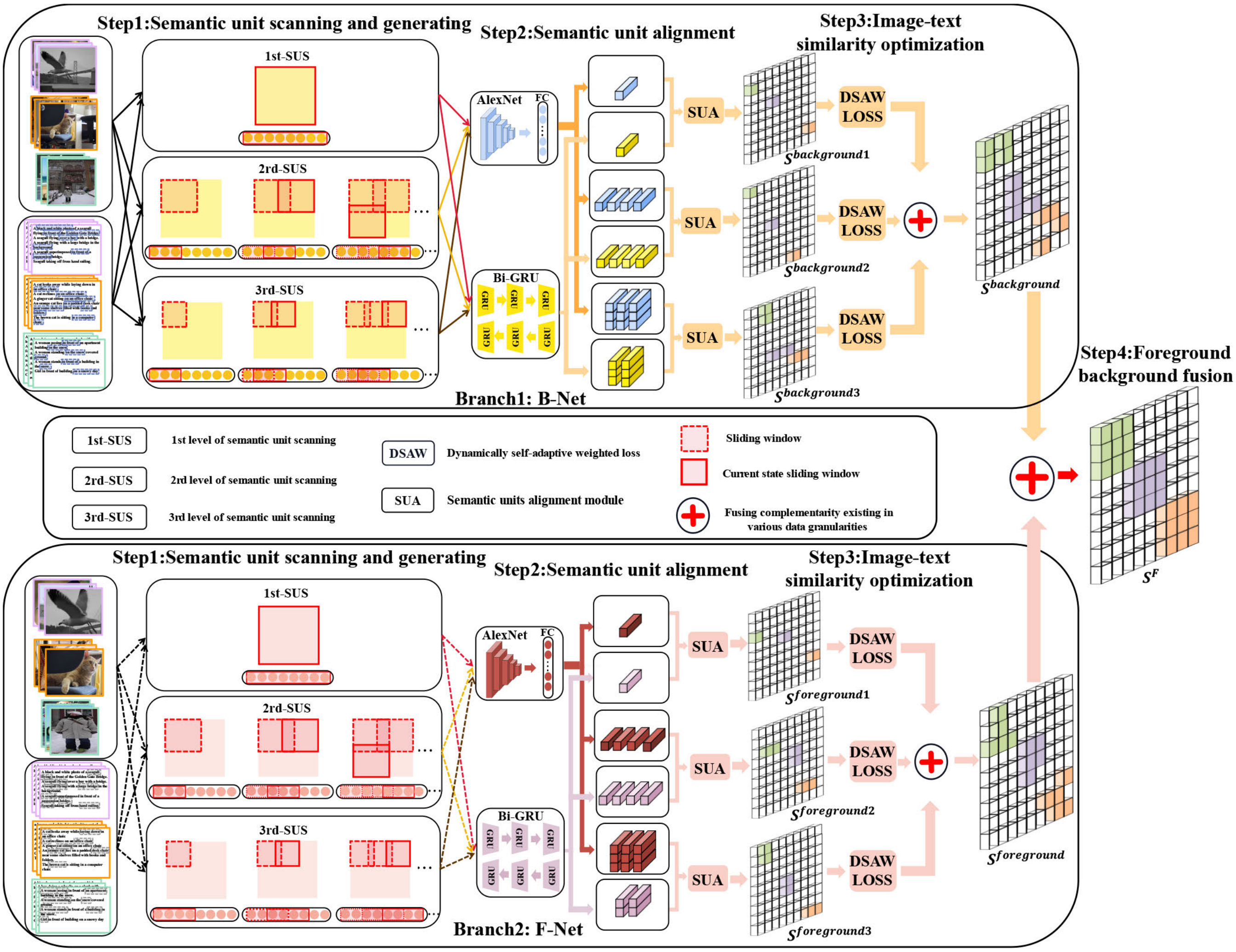

Step1:Semantic unit scanning and generating
Step2:Semantic unit alignment
Step3:Image-text similarity optimization
1st-SUS
2rd-SUS
3rd-SUS
AlexNet
FC
Bi-GRU
GRU
SUA
DSAW LOSS
$S^{background1}$
$S^{background2}$
$S^{background3}$
$S^{background}$
Branch1: B-Net
Step4:Foreground background fusion
1st-SUS 1st level of semantic unit scanning
2rd-SUS 2rd level of semantic unit scanning
3rd-SUS 3rd level of semantic unit scanning
DSAW Dynamically self-adaptive weighted loss
SUA Semantic units alignment module
Sliding window
Current state sliding window
Fusing complementarity existing in various data granularities
S^F
$S^{foreground1}$
$S^{foreground2}$
$S^{foreground3}$
$S^{foreground}$
Branch2: F-Net

FIGURE 2. Framework of our proposed FB-Net. It contains two branches: 1) B-Net and 2) F-Net. There are three steps to calculate the image-text similarity for each branch, that is, 1) Step1 (Semantic unit scanning and generating), 2) Step2 (Semantic unit alignment), 3) Step3 (Image-text similarity optimization). Afterwards, in Step4 (Foreground-background Fusion), the image-text similarities obtained from B-Net and F-Net are fused together to perform image-text retrieval.

Similarly, the set of feature vectors of foreground text areas for F-Net(L_u) is as follows:

$$T_e^{\text{level } u} = \left\{ \mathrm{e}_1^{\text{level } u}, \ldots, \mathrm{e}_{\mathrm{n}}^{\text{level } u}, \ldots, \mathrm{e}_{\mathrm{N}}^{\text{level } u} \right\}, \forall \mathrm{n} \in [1,\ \mathrm{N}] \quad (2)$$

where $\mathrm{e}_{\mathrm{n}}^{\text{level } u}$ is the feature vector of the nth text area at the uth level of granularity for F-Net, N is the number of foreground areas of the text.

The following formula is used to determine the similarity in the mth row and nth column of the matrix $S^{\text{foreground } u}$:

$$\begin{aligned} S^{\text{foreground } u}(m, n) &= S(m, n) = S^{i}(m, n) + S^{\mathrm{t}}(m, n) \\ &= \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \frac{\left(f_m^{\text{level } u}\right)^T \boldsymbol{r}_m^t}{\left\| f_m^{\text{level } u} \right\| \left\| \boldsymbol{r}_m^t \right\|} + \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} \frac{\left(\mathbf{e}_n^{\text{level } u}\right)^T \boldsymbol{r}_n^i}{\left\| \mathbf{e}_n^{\text{level } u} \right\| \left\| \boldsymbol{r}_n^i \right\|} \end{aligned} \quad (3)$$

The structure of B-Net is similar to that of F-Net, and the similarity scores are calculated by the attention weighting of features. In this paper, we propose the multi-level sliding window strategy for both image modality and text modality. We design sliding windows of different scales and construct u levels of multi-granularity sub-networks. The similarity matrix of each sub-network can be written as $S^{\text{background } 1}, \ldots, S^{\text{background } u}$, in which $S^{\text{background } 1}$ represents the coarse-grained background sub-network.

For the u level of granularity of B-Net (B-Net(L_u)), the set of feature vectors for background image areas is as follows:

$$\mathrm{I}_{\mathrm{r}}^{\text{level } u} = \left\{ i_1^{\text{level } u}, \ldots, i_x^{\text{level } u}, \ldots, i_{\mathrm{x}}^{\text{level } u} \right\}, \forall \mathrm{x} \in [1, \mathrm{X}] \quad (4)$$

where $i_x^{\text{level } u}$ is the feature vector of the xth image area at the uth level of granularity for B-Net, and X is the number of background areas of the image.

Furthermore, the set of feature vectors of background text areas for B-Net(L_u) is as follows:

$$T_q^{\text{level } u} = \left\{ t_1^{\text{level } u}, \ldots, t_{\mathrm{y}}^{\text{level } u}, \ldots, t_{\mathrm{Y}}^{\text{level } u} \right\}, \forall \mathrm{y} \in [1, \mathrm{Y}] \quad (5)$$

where $t_y^{\text{level } u}$ is the feature vector of the yth text area at the uth level of granularity for B-Net, and Y is the number of background areas of the text.

The following formula is used to determine the similarity in the xth row and yth column of the matrix $S^{\text{background } u}$:

$$\begin{aligned} S^{\text{background } u}&(x, y) \\ &= S(x, y) = S^{i}(x, y) + S^{\mathrm{t}}(x, y) \\ &= \frac{1}{\mathrm{X}} \sum_{x=1}^{\mathrm{X}} \frac{\left(\boldsymbol{i}_x^{\text{level } u}\right)^T \boldsymbol{r}_x^t}{\left\|\boldsymbol{i}_x^{\text{level } u}\right\| \left\|\boldsymbol{r}_x^t\right\|} + \frac{1}{\mathrm{Y}} \sum_{y=1}^{\mathrm{Y}} \frac{\left(\boldsymbol{t}_{\mathrm{y}}^{\text{level } u}\right)^T \boldsymbol{r}_y^i}{\left\|\boldsymbol{t}_{\mathrm{y}}^{\text{level } u}\right\| \left\|\boldsymbol{r}_y^i\right\|} \end{aligned} \quad (6)$$

Finally, we linearly fuse the similarity matrix of each sub-network to obtain the final image-text similarity matrix S^F as follows.

$$\begin{aligned} S^F = \lambda_1 &\left(\alpha_1 S^{\text{foreground } 1} + \ldots + \alpha_{\mathrm{u}} S^{\text{foreground } u}\right) \\ &+ \lambda_2 \left(\beta_1 S^{\text{background } 1} + \ldots + \beta_{\mathrm{u}} S^{\text{background } u}\right) \end{aligned} \quad (7)$$

where $\alpha_1, \ldots \alpha_{\mathrm{n}}, \beta_1, \ldots \beta_{\mathrm{n}}, \lambda_1$ and λ_2 are the balance parameters. Note that, the matrix S^F well reflects the complementary relationship between multi-granularity foreground and background semantic units. Thus, we use S^F for image-text retrieval.

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

C. SEMANTIC UNIT GENERATION WITH MULTI-SCALE SEMANTIC SCANNING

To completely and accurately capture the semantic units and improve the performance of image-text retrieval, in this section, we will illustrate how to implement the multi-level sliding window strategy to achieve multi-scale semantic scanning. Since the image is intrinsically a two-dimensional array and the text is a one-dimensional array, the multi-scale semantic scanning mechanisms for images and texts are quite different. Therefore, we discuss the multi-level sliding window strategy for images and texts respectively.

1) MULTI-LEVEL OVERLAPPED SLIDING WINDOW STRATEGY FOR IMAGES

To ensure that the semantic units can be completely captured, the adjacent sliding windows in images should be overlapped with each other to maintain the continuity of local regions. Thus, there are more chances for overlapped sliding windows to obtain the semantic units with more complete semantic information. Therefore, we set a certain overlap between adjacent sliding windows when sliding on images. The width and height of the sliding window are defined as follows:

$$\begin{aligned} win\_w &= \lceil w/x \rceil \\ win\_h &= \lceil h/x \rceil \end{aligned} \quad (8)$$

where $win\_w = \lceil w/x \rceil$, and $win\_h = \lceil h/x \rceil$ refer to the width and height of the sliding window respectively, and the width and height of the image are w, and h respectively. To make the adjacent sliding windows be overlapped to each other, the stepsize of the sliding window on the image is defined as follows:

$$\begin{aligned} stepsize\_w &= w/2x \\ stepsize\_h &= h/2x \end{aligned} \quad (9)$$

where $stepsize\_w$, and $stepsize\_h$ denote the stepsize at the direction of width and height respectively, and parameter x is used to control the overlapping degree between adjacent sliding windows. After scanning the whole image with overlapped sliding windows, the number of image blocks generated by the proposed multi-level overlapped sliding window strategy is $M = (2x - 1)^2$.

2) MULTI-LEVEL OVERLAPPED SLIDING WINDOW STRATEGY FOR TEXT

Similar to images, neighboring words in a text have a high correlation with each other. Moreover, the number of words in a semantic unit of a text is unable to determine in advance. Therefore, we also use a multi-level overlapped sliding window strategy to segment a text at different levels of granularities.

Different from images, the sliding window for texts uses the one-dimensional data scanning mode. For text T, the set of word features is represented as $T_q = \{\mathrm{w}_1, \ldots, \mathrm{w}_k, \ldots, \mathrm{w}_K\}$, where w_k denotes the kth word feature, and K is the number of words in the text. We use the Bi-directional GRU (Bi-GRU) with only one layer to extract the word features within the text, and the word features in T_q are calculated as follows:

$$\mathrm{w}_k = \frac{\overrightarrow{h_k} + \overleftarrow{h_k}}{2}, k \in [1, K] \quad (10)$$

where $\overrightarrow{h_k}$ and $\overleftarrow{h_k}$ denote the hidden states from the forward GRU and the backward GRU respectively.

Then, features of the words covered by the sliding window are merged to generate the feature of the whole text. For instance, the text feature in the nth sliding window at the uth level is calculated as follows:

$$\boldsymbol{t}_n^{\text{tevel } u} = \sum_{g=1+st\times(n-1)}^{st\times(n-1)+win} \mathbf{w}_g, n \in [1, N] \quad (11)$$

where st is the sliding step, and win is the window size. Particularly, the number of sliding windows N is calculated as follows:

$$N = \left\lfloor \frac{K - \text{ win}}{st} \right\rfloor + 1 \quad (12)$$

D. SEMANTIC UNIT ALIGNMENT WITH STACKED CROSS-ATTENTION

In Figure4, to align the candidate semantic units of images and texts, we use the stacked cross-attention with attention

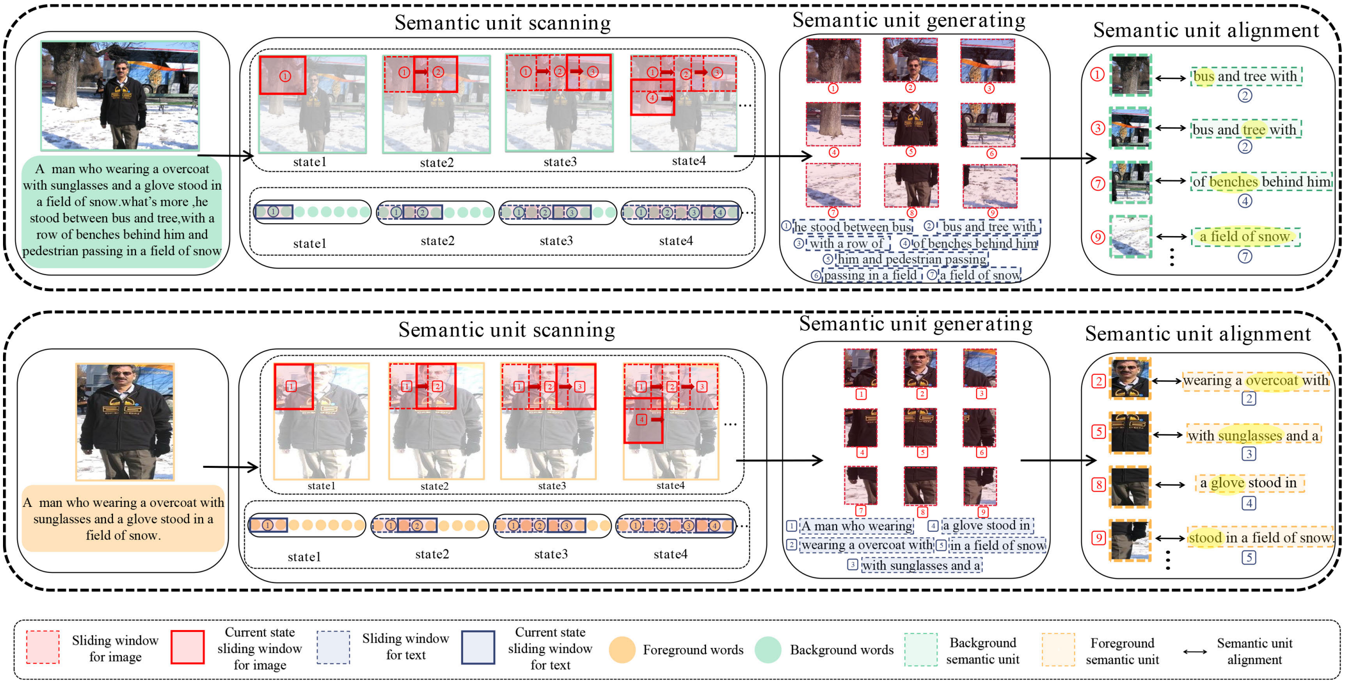

Semantic unit scanning
Semantic unit generating
Semantic unit alignment
A man who wearing a overcoat with sunglasses and a glove stood in a field of snow.what's more ,he stood between bus and tree,with a row of benches behind him and pedestrian passing in a field of snow
state1
state2
state3
state4
①he stood between bus ②bus and tree with
③with a row of ④of benches behind him
⑤him and pedestrian passing
⑥passing in a field ⑦a field of snow
bus and tree with
of benches behind him
a field of snow.
A man who wearing a overcoat with sunglasses and a glove stood in a field of snow.
①A man who wearing ④a glove stood in
②wearing a overcoat with ⑤in a field of snow
③with sunglasses and a
wearing a overcoat with
with sunglasses and a
a glove stood in
stood in a field of snow
Sliding window for image
Current state sliding window for image
Sliding window for text
Current state sliding window for text
Foreground words
Background words
Background semantic unit
Foreground semantic unit
Semantic unit alignment

FIGURE 3. Scanning and alignment of semantic units. Multi-level sliding window strategy is designed to scan the semantic units for images and texts, and thus maintain the local semantic continuity. Moreover, semantic unit alignment is carried out to match the semantic units of images and texts.

weights to calculate the image-text similarity. The task of data alignment is divided into two parts: 1) alignment towards the image embedding space, and 2) alignment towards the text embedding space.

Suppose that the set of feature vectors for image areas is $I = \{f_1, \ldots, f_x, \ldots, f_X\}$, and the set of feature vectors for text areas is $T = \left\{t_1, \ldots, t_y, \ldots, t_Y\right\}$. In order to ensure images and texts can be directly computed in the same space, we use a fully connected network to downscale the image features, that is, the mapping matrix W is used to map the initial feature f_x to i_x, (i.e. $i_x = Wf_x + b$), where b are the parameters of the network. We update feature vectors for image areas to $I = \{i_1, \ldots, i_x, \ldots, i_X\}$, and normalize each feature.

1) MODULE 1: IMAGE-GROUNDED EMBEDDING SPACE

In this module, fine-grained features of texts are mapped to the image embedding space by the following steps. First, the regions of texts are weighted by the attention mechanism. Second, the weighted fine-grained features of texts are summed up as the contextual features, and the cosine similarities between image regions and text contextual features are further calculated. Third, the cosine similarities of all image regions are summed up, and the similarities of image-text pairs are calculated by the regional feature alignment. The detailed procedure is as follows.

sim$_{xy}$ denotes the similarity between the image region i_x and the text region t_y, and it is computed based on the cosine distance: $\mathrm{sim}_{xy} = \frac{i_x^T t_y}{\|i_x\|\|t_y\|}, x \in [1, X], y \in [1, Y]$. Then, the sim_{xy} is normalized by columns: $\overline{\mathrm{sim}_{xy}} = \frac{\mathrm{relu}\ (\mathrm{sim}_{xy})}{\sqrt{\sum_{x=1}^{X}\ \mathrm{relu}\ (\mathrm{sim}_{xy})^2}}$, Further, we can calculate the attention weights: $\alpha_{xy} = \frac{\exp(\delta_1 \overline{\mathrm{sim}_{xy}})}{\sum_{y=1}^{Y} \exp(\delta_1 \overline{\mathrm{sim}_{xy}})}$, where δ_1 is the inversed temperature of the Softmax function to adjust the smoothness of the attention distribution. Then, the sum of weighted text region features is used to calculate the contextual feature: $\boldsymbol{r}_x^{\mathrm{t}} = \sum_{y=1}^{Y} \alpha_{xy} \boldsymbol{t}_y$, where r_x^{t} is the textual contextual feature weighted by the xth image region. Finally, the similarity between I and T is calculated as follows:

$$S^i(x, y) = \frac{1}{X} \sum_{x=1}^{X} \frac{\boldsymbol{i}_x^T \boldsymbol{r}_x^t}{\|\boldsymbol{i}_x\| \left\|\boldsymbol{r}_x^t\right\|} \tag{13}$$

where s^i is the similarity between I and T based on the image embedding space.

2) MODULE 2: TEXT-GROUNDED EMBEDDING SPACE

In module 2, the weighted features of image regions are mapped to the text embedding space. The process of calculating the similarity between I and T is just similar to that of module 1. First, the similarity sim_{xy} between image region i_x and text region $\boldsymbol{t}_y$ is calculated and then normalized: $\overline{\mathrm{sim}}'_{xy} = \frac{\mathrm{relu}(\mathrm{sim}\, m_{xy})}{\sqrt{\sum_{y=1}^{Y} \mathrm{relu}(\mathrm{sim}_{xy})^2}}$, and the attention weight is calculated as: $\beta_{xy} = \frac{\exp\left(\delta_2 \overline{\mathrm{sim}}'_{xy}\right)}{\sum_{x=1}^{X} \exp\left(\delta_2 \overline{\mathrm{sim}}'_{xy}\right)}$, where δ_2 is the inversed temperature of the Softmax function to adjust the smoothness of the attention distribution.

Further, the image context features are obtained by the weighted sum of the features of image regions: $\boldsymbol{r}_y^i = \sum_{x=1}^{X} \beta_{xy} \boldsymbol{i}_x$. Afterwards, the similarity between I and T is

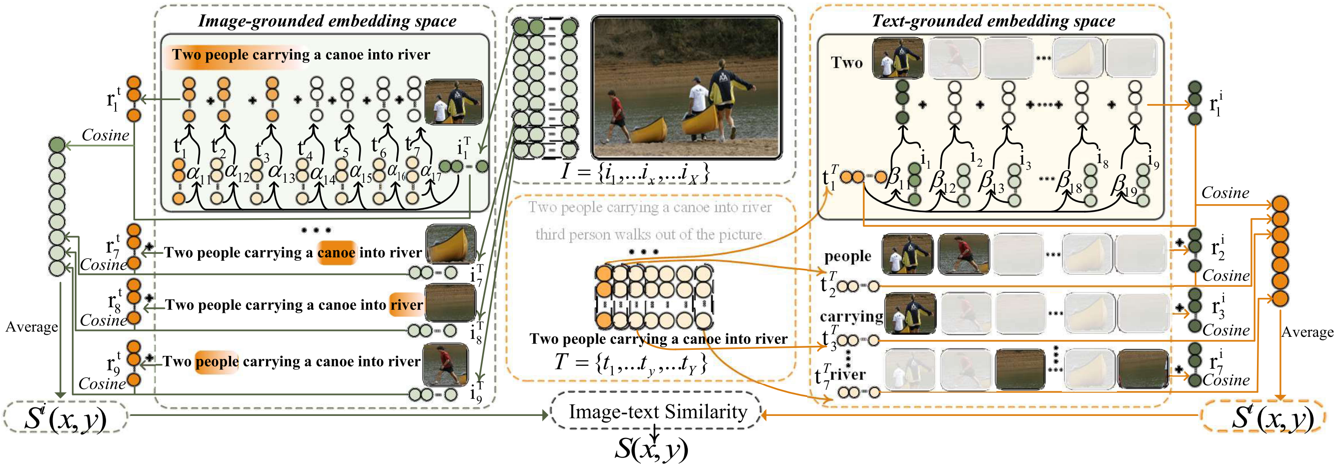


Image-grounded embedding space
Two people carrying a canoe into river
r_1^t
Cosine
t_1 t_2 t_3 t_4 t_5 t_6 t_7
α_{11} α_{12} α_{13} α_{14} α_{15} α_{16} α_{17}
i_1^T
r_7^t
Two people carrying a canoe into river
i_7^T
r_8^t
Two people carrying a canoe into river
i_8^T
r_9^t
Two people carrying a canoe into river
i_9^T
Average
$S^t(x,y)$
$I=\{i_1,\ldots i_x,\ldots i_X\}$
Two people carrying a canoe into river
third person walks out of the picture
Two people carrying a canoe into river
$T=\{t_1,\ldots t_y,\ldots t_Y\}$
Image-text Similarity
$S(x,y)$
Text-grounded embedding space
Two
t_1^T
β_{11} β_{12} β_{13} β_{18} β_{19}
i_1 i_2 i_3 i_8 i_9
r_1^i
people
t_2^T
r_2^i
carrying
t_3^T
r_3^i
river
t_7^T
r_7^i
Cosine
Average
$S^i(x,y)$

FIGURE 4. The mechanism of stacked cross-attention. We use the stacked cross-attention with attention weights to align semantic units and calculate image-text similarity.

calculated as follows:

$$S^{\mathrm{t}}(x,y)=\frac{1}{Y}\sum_{y=1}^{Y}\frac{\boldsymbol{t}_y^T\boldsymbol{r}_y^i}{\left\|\boldsymbol{t}_y\right\|\left\|\boldsymbol{r}_y^i\right\|}\tag{14}$$

Then, the similarities obtained with Eq.13 and Eq.14 are integrated together to compute the image-text similarity between I and T as follows:

$$S(x,y)=S^i(x,y)+S^t(x,y)\tag{15}$$

E. MULTI-GRANULARITY FEATURE EXTRACTION WITH MULTI-LEVEL OVERLAPPED SLIDING WINDOW STRATEGY

With the proposed multi-level overlapped sliding window strategy, the semantic scanning is conducted on both the foreground areas and background areas respectively. Afterward, multi-granularity features of images and texts are extracted from the local areas that are cut by the overlapped sliding windows at different levels.

1) MULTI-GRANULARITY IMAGE FEATURE EXTRACTION

Foreground Areas of an Image: The object detection technique creates multiple bounding boxes from an image, and each bounding box carries explicit semantic information. Thus, we select some bounding boxes with rich semantic information, and the local areas surrounded by the bounding boxes are regarded as the foreground areas. To find more semantic units from foreground areas, we use the proposed multi-level overlapped sliding window strategy to divide them into many small blocks. Thus, fine-grained detailed semantic information such as actions and states can be effectively captured from semantic units.

Specifically, the foreground areas of images are detected and cropped by the Faster R-CNN, which is a two-stage target detection framework. The output of Faster R-CNN includes the coordinates of the bounding box, the confidence level, the prediction class, and the cropped target area.

For image I, we set the maximum number of targets detected by Faster R-CNN to J. Then, we take the uth level overlapped sliding window as an example. For the uth level of granularity of F-Net (F-Net(L_u)), the set of feature vectors for the foreground image areas is shown in Eq.1.

Background Areas of an Image: Different from extracting the features of foreground areas, we implement our proposed multi-level overlapped sliding window strategy on the whole image.

For B-Net at the u level of granularity (B-Net(L_u)), the set of feature vectors for the background areas is shown in Eq.4.

2) MULTI-GRANULARITY TEXT FEATURE EXTRACTION

The multi-level overlapped sliding window strategy on images can effectively maintain the local semantic continuity, and further improve the ability of the image-text retrieval model to understand the semantic information in images. Similarly, our proposed multi-level overlapped sliding window strategy can also be used in text processing, and this strategy can make full use of word contextual relationships to improve the accuracy of image-text retrieval.

Multi-granularity data for text refers to the sequence of words with different sizes. In this work, we also use the sliding window at different levels to extract the multi-granularity features for texts. For the whole text, we still use Bi-GRU to extract features of words. Then, we use our proposed multi-level overlapped sliding window strategy for texts to merge word features within one sliding window.

For text T, we take the uth level overlapped sliding window as an example. Two sets of features extracted from the foreground and background areas of T are represented in Eq.2 and Eq.5 respectively.

F. IMAGE-TEXT SIMILARITY OPTIMIZATION

The cross-modal similarity should follow the principle that the similarity between the same class should be larger than the similarity between different classes. In other words, the features of similar samples should become closer to each other, while the samples of different classes should pull away from each other. Obviously, this idea is consistent with the triplet loss function.

But the traditional triplet loss function [37] assigns the same weight to each sample pair. That is, the margin set by the triplet loss is a constant. In the Wikipedia dataset, for example, the differences between samples in the classes ‘‘literature’’ and ‘‘art’’, ‘‘history’’ and ‘‘warfare’’ are small, while the differences between samples in the classes ‘‘sport’’ and ‘‘music’’ are large. However, the margin in the triplet loss function is always a fixed value. If the margin is set large, the model may not be able to distinguish ‘‘literature’’ and ‘‘art’’ well, while if the margin is set small, it may not be able to distinguish ‘‘sport’’ and ‘‘music’’ well.

In addition, since the triplet loss function generates a large number of sample pairs, it reduces the convergence speed and model performance during training. Although previous work has a triplet loss function by difficult sample mining, it loses more valuable information. Therefore, the ability to distinguish more valuable samples using the triplet loss function is limited.

In most cross-modal retrieval datasets, there are many classes with very similar semantics. It is difficult to distinguish semantically similar classes if the same weight was given to each class using triplet loss. Therefore, we propose the Dynamically Self-Adaptive weighted loss function, which assigns a weight to each sample pair.

In image-text retrieval, a given image/text query is used as an anchor. Moreover, if the target texts/images belong to the same semantic concept, they are regarded as positive samples, otherwise, they are considered as negative samples. For anchor $\boldsymbol{x}_i$, we denote the index sets of their selected positive and negative pairs as H_j^+ and H_j^- respectively.

we provide the redefinition of the cross-modal affinity matrix as follows:

$$\tilde{A}_{ij} = \begin{cases} +\sigma & \text{if } H_j \text{ is a positive sample} \\ -\theta & \text{if } H_j \text{ is a negative sample} \end{cases} \tag{16}$$

where σ, θ are hyper-parameters.

Also, each sample pair should be given a different weight, so that we have a negative sample pair $\{\boldsymbol{x}_i, \boldsymbol{x}_j\} \in H_j^-$, whose weight can be computed as:

$$\mathrm{M}_{ij}^- = \frac{e^{\tilde{\mathrm{A}}_{ij}(s_{ij}-\gamma)}}{1 + \sum_{k \in H_j^-} e^{\tilde{\mathrm{A}}_{ij}(S_{ik}-\gamma)}} \tag{17}$$

and the weight of a positive pair $\{\boldsymbol{x}_i, \boldsymbol{x}_j\} \in H_j^+$ is:

$$\mathrm{M}_{ij}^+ = \frac{e^{\tilde{A}_{ij}(s_{ij}-\gamma)}}{1 + \sum_{k \in H_j^+} e^{\tilde{\mathrm{A}}_{ij}(S_{ik}-\gamma)}} \tag{18}$$

where S_{ij} and S_{ik} denote the similarity between sample pair $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$, and $\{\boldsymbol{x}_i, \boldsymbol{x}_k\}$ respectively, and γ is the hyper-parameter. The loss function for negative samples is as follows:

TABLE 1. Parameters of sliding windows for image and text.

The level of granularity	Image			Text	
	Width	height	step size	length	step size
L_1	w	h	$(w/2, h/2)$	K	1
L_2	$2w/3$	$2h/3$	$(w/3, h/3)$	$K-6$	2
L_3	$w/2$	$h/2$	$(w/4, h/4)$	$K-24$	3

$$L_{\mathrm{DSAW}}^{\mathrm{N}} = \ln\left[1 + \sum_{k \in H_j^-} e^{\tilde{\mathrm{A}}_{ij}(S_{ik}-\gamma)}\right] \tag{19}$$

Similarly, the loss function for positive samples is as follows:

$$L_{\mathrm{DSAW}}^{\mathrm{P}} = \ln\left[1 + \sum_{k \in H_j^+} e^{\tilde{\mathrm{A}}_{ij}(S_{ik}-\gamma)}\right] \tag{20}$$

We finally obtain the function of DSAW loss:

$$L_{\mathrm{DSAW}} = \frac{1}{\mathrm{B}} \sum_{i=1}^{\mathrm{B}} \left(\frac{1}{\sigma} L_{\mathrm{DSAW}}^{\mathrm{P}} + \frac{1}{\theta} L_{\mathrm{DSAW}}^{\mathrm{N}}\right) \tag{21}$$

where B is the size of the batch data.

We believe that positive samples with lower similarity to the anchor and negative samples with higher similarity to the anchor should be optimized than other samples. Therefore, we optimize the samples enclosed between the most difficult negative sample and the most difficult positive sample in each batch of data, instead of optimizing all samples.

Informing samples can be mined for discriminative optimization via the DSAW loss. In our work, all initial image-text similarities are optimized by the proposed DSAW loss.

IV. EXPERIMENT

In this section, to validate the effectiveness of the proposed FB-Net for the task of image-text retrieval, extensive experiments are conducted on the dataset of Pascal Sentence, Wikipedia and NUS-WIDE, and 16 state-of-the-art methods are compared with ours. In addition, parameter sensitivity analyses are provided, and ablation studies are presented to demonstrate the usefulness of each module in FB-Net.

A. DATASETS

Pascal Sentence [32] dataset contains 1000 images, which are divided into 20 semantic categories, and each semantic category has 50 images. Each image has five matching sentences, which constitute a document. The number of labels in the dataset dictionary is 1000.

Wikipedia [33] dataset is a widely used cross-media retrieval dataset. The data comes from the selected articles and pictures from Wikipedia, with a total of 29 categories. This dataset selects 10 categories with the largest number of articles as semantic categories and finally constructs a dataset of 2866 image/text pairs.

NUS-WIDE [41] dataset contains 269,648 web images with 81 concept labels and 5018 distinct tags. Note that some

TABLE 2. The MAP scores of cross-modal retrieval for FB-Net and other compared methods on all datasets.

Methods	Pascal Sentences			Wikipedia			NUS-WIDE-25K			NUS-WIDE-5K		
	i2t	t2i	Avg	i2t	t2i	Avg	i2t	t2i	Avg	i2t	t2i	Avg
CCA(2010)	0.092	0.096	0.094	0.277	0.226	0.252	0.722	0.685	0.703	0.223	0.253	0.238
CMCP(2012)	0.471	0.437	0.454	0.326	0.251	0.289	0.760	0.734	0.747	0.395	0.432	0.414
JRL(2014)	0.520	0.506	0.513	0.339	0.250	0.294	0.742	0.697	0.719	0.498	0.533	0.515
JFSSL(2015)	0.507	0.464	0.485	0.360	0.280	0.320	0.709	0.714	0.712	0.522	0.509	0.516
S^2UPG(2015)	0.552	0.547	0.549	0.377	0.286	0.331	0.759	0.682	0.720	0.494	0.497	0.495
DCCA(2015)	0.475	0.471	0.473	0.440	0.396	0.418	0.708	0.710	0.709	0.426	0.433	0.429
CCL(2018)	0.567	0.563	0.565	0.504	0.457	0.481	0.756	0.725	0.740	0.432	0.503	0.468
SCAN(2018)	0.566	0.570	0.568	0.517	0.439	0.478	0.753	0.804	0.778	0.531	0.530	0.530
GXN(2018)	0.598	0.578	0.588	0.525	0.448	0.486	0.802	0.801	0.802	0.565	0.557	0.561
VSESC(2019)	0.575	0.572	0.573	0.518	0.467	0.492	0.776	0.798	0.787	0.542	0.556	0.549
MAVA(2020)	0.572	0.571	0.571	0.547	0.488	0.518	0.801	0.809	0.805	0.597	0.569	0.583
SGRAF(2021)	0.587	0.572	0.580	0.564	0.483	0.523	0.831	0.837	0.834	0.609	0.602	0.606
SCL(2022)	0.618	0.621	0.620	0.563	0.490	0.527	0.825	0.819	0.822	0.610	0.618	0.614
CGMN(2022)	0.616	0.605	0.611	0.569	0.492	0.531	0.840	0.826	0.833	0.641	0.631	0.635
NAAF(2022)	0.615	0.628	0.622	0.567	0.492	0.530	0.846	0.838	0.842	0.625	0.630	0.627
VSRN++(2022)	0.658	0.614	0.636	0.574	0.497	0.535	0.849	0.845	0.847	0.635	0.640	0.637
FB-Net(ours)	**0.704**	**0.647**	**0.675**	**0.618**	**0.531**	**0.574**	**0.881**	**0.872**	**0.876**	**0.669**	**0.667**	**0.668**

TABLE 3. Performance comparison of the image-text retrieval on the pascal sentence dataset in terms of Recall@K.

Methods	i2t			t2i			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA(2010)	0.062	0.244	0.365	0.072	0.193	0.338	1.274
CMCP(2012)	0.421	0.724	0.855	0.401	0.778	0.876	4.055
JRL(2014)	0.541	0.671	0.904	0.603	0.894	0.966	4.579
JFSSL(2015)	0.425	0.764	0.896	0.552	0.912	0.907	4.456
S^2UPG(2015)	0.521	0.735	0.872	0.697	0.918	0.970	4.713
DCCA(2015)	0.407	0.601	0.822	0.584	0.902	0.915	4.231
CCL(2018)	0.613	0.862	0.979	0.599	0.854	0.933	4.840
SCAN(2018)	0.588	0.878	0.900	0.761	0.828	0.882	4.837
GXN(2018)	0.596	0.875	0.960	0.821	0.884	0.968	5.104
VSESC(2019)	0.510	0.789	0.921	0.823	0.911	0.964	4.918
MAVA(2020)	0.561	0.824	0.892	0.685	0.932	0.961	4.855
SGRAF(2021)	0.611	0.876	0.951	0.813	0.888	0.974	5.113
SCL(2022)	0.615	0.870	0.928	0.771	0.926	0.980	5.090
CGMN(2022)	0.611	0.876	0.951	0.813	0.888	0.974	5.113
NAAF(2022)	0.613	0.869	0.933	0.784	0.928	0.981	5.108
VSRN++(2022)	0.661	**0.900**	0.976	0.824	0.937	0.996	5.294
FB-Net(ours)	**0.703**	0.855	**0.986**	**0.873**	**0.996**	**0.998**	**5.411**

TABLE 4. Performance comparison of the image-text retrieval on the Wikipedia dataset in terms of Recall@K.

Methods	i2t			t2i			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA(2010)	0.094	0.234	0.344	0.147	0.486	0.596	1.901
CMCP(2012)	0.094	0.205	0.494	0.115	0.402	0.633	1.943
JRL(2014)	0.176	0.578	0.678	0.110	0.487	0.767	2.796
JFSSL(2015)	0.215	0.456	0.590	0.152	0.463	0.738	2.614
S^2UPG(2015)	0.315	0.462	0.549	0.121	0.523	0.617	2.587
DCCA(2015)	0.411	0.431	0.442	0.774	0.900	0.943	3.901
CCL(2018)	0.441	0.456	0.467	0.598	0.900	0.957	3.819
SCAN(2018)	0.402	0.449	0.463	0.629	0.935	0.959	3.837
GXN(2018)	0.428	0.473	0.506	0.626	0.942	0.962	3.937
VSESC(2019)	0.423	0.448	0.470	**0.823**	0.916	0.974	4.054
MAVA(2020)	0.456	0.472	0.477	0.713	0.947	0.972	4.037
SGRAF(2021)	0.479	0.519	0.536	0.683	0.979	0.983	4.179
SCL(2022)	0.454	0.527	0.534	0.732	0.925	0.965	4.137
CGMN(2022)	0.467	0.512	0.528	0.694	0.945	0.982	4.128
NAAF(2022)	0.461	0.463	0.465	0.653	**0.982**	0.986	4.010
VSRN++(2022)	0.496	0.505	0.512	0.760	0.949	0.991	4.213
FB-Net(ours)	**0.540**	**0.604**	**0.782**	0.703	0.953	**0.995**	**4.577**

noisy tags that are worthless for our work should be removed. Therefore, two sub-datasets (i.e., NUS-WIDE-25K and NUS-WIDE-5K) are extracted from the NUS-WIDE dataset. NUS-WIDE-25K is constructed by selecting 560 tags and a total of 25,084 image-text pairs from NUS-WIDE. Particularly, the training set, validation set, and testing set of NUS-WIDE-25K contain 20,000, 2500, and 2584 image-text pairs. Additionally, six semantic concepts from the NUS-WIDE-25K dataset (i.e.,“ animal,” “clouds,” “person,” “sky,” “water,” and “window”) are reused to create NUS-WIDE-5K, which consists of 4996 image-text pairs divided into three subsets: 4500 pairs for training, 250 pairs for validating, and 246 pairs for testing.

B. EVALUATION METRICS

1) AP AND MAP

MAP (Mean Average Precision) is used to calculate the average of all Average Precision (AP), which is defined as:

$$\mathrm{AP} = \frac{1}{T}\sum_{r=1}^{R} P(r)\delta(r) \tag{22}$$

where T is the number of samples whose retrieval results and query items belong to the same semantic category, R is the size of the test set, and $P(r)$ denotes the accuracy rate of the top-ranked r returned items. If the rth returned item and the query item belong to the same semantic category, then $\delta(r) = 1$, otherwise $\delta(r) = 0$. The larger the value of MAP is, the higher the performance of information retrieval is achieved.

2) PR CURVE

The PR curve reflects the relationship between Precision and Recall, and shows the trend of the average accuracy for the top-ranked retrieval results. In a two-dimensional coordinate system, the X-axis represents Recall and the Y-axis represents Precision. The larger the area enclosed by the PR curve, the better the performance is obtained.

3) Recall@K

Recall@K is the score of retrieving the correct results in the top K positions of the retrieval results. The higher value of

TABLE 5. Performance comparison of the image-text retrieval on the NUS-WIDE-25K dataset in terms of Recall@K.

Methods	i2t			t2i			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA(2010)	0.205	0.824	0.931	0.142	0.899	0.942	3.943
CMCP(2012)	0.447	0.816	0.959	0.409	0.840	0.957	4.428
JRL(2014)	0.497	0.814	0.936	0.114	0.829	0.946	4.136
JFSSL(2015)	0.329	0.815	0.898	0.297	0.816	0.908	4.063
S^2UPG(2015)	0.443	0.868	0.908	0.335	0.819	0.883	4.256
DCCA(2015)	0.257	0.897	0.894	0.159	0.891	0.904	4.002
CCL(2018)	0.837	0.905	0.938	0.782	0.889	0.901	5.252
SCAN(2018)	0.849	0.914	0.945	0.784	0.877	0.908	5.277
GXN(2018)	0.811	0.953	0.968	0.818	0.862	0.893	5.305
VSESC2019	0.852	0.902	0.933	0.793	0.894	0.916	5.290
MAVA(2020)	0.809	0.844	0.970	0.825	0.865	0.900	5.313
SGRAF(2021)	0.873	0.956	0.969	0.876	0.928	0.943	5.545
SCL(2022)	0.884	0.951	0.963	0.869	0.904	0.958	5.529
CGMN(2022)	0.893	0.954	0.970	0.873	0.914	0.950	5.554
NAAF(2022)	0.906	0.958	0.965	0.917	0.959	0.976	5.681
VSRN++(2022)	0.905	0.956	0.971	0.909	0.950	0.974	5.675
FB-Net(ours)	**0.927**	**0.974**	**0.992**	**0.923**	**0.959**	**0.983**	**5.758**

TABLE 6. Performance comparison of the image-text retrieval on the NUS-WIDE-5K dataset in terms of Recall@K.

Methods	i2t			t2i			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
CCA(2010)	0.246	0.627	0.809	0.242	0.335	0.379	2.638
CMCP(2012)	0.489	0.702	0.827	0.228	0.608	0.897	3.751
JRL(2014)	0.575	0.814	0.872	0.503	0.846	0.914	4.524
JFSSL(2015)	0.564	0.732	0.896	0.552	0.812	0.927	4.483
S^2UPG(2015)	0.476	0.681	0.752	0.612	0.774	0.909	4.204
DCCA(2015)	0.520	0.751	0.921	0.434	0.643	0.910	4.179
CCL(2018)	0.492	0.631	0.853	0.308	0.643	0.875	3.802
SCAN(2018)	0.589	0.716	0.852	0.673	0.867	0.919	4.616
GXN(2018)	0.612	0.785	0.876	0.645	0.849	0.928	4.695
VSESC(2019)	0.603	0.789	0.837	0.657	0.841	0.885	4.612
MAVA(2020)	0.625	0.788	0.908	0.629	0.856	0.935	4.741
SGRAF(2021)	0.782	0.889	0.923	0.794	0.869	0.915	5.172
SCL(2022)	0.793	0.876	0.919	0.804	0.880	0.906	5.178
CGMN(2022)	0.794	0.885	0.921	0.826	0.871	0.907	5.204
NAAF(2022)	0.788	0.873	0.919	0.810	0.862	0.929	5.181
VSRN++(2022)	0.801	0.883	0.924	0.817	0.869	0.911	5.205
FB-Net(ours)	**0.826**	**0.904**	**0.943**	**0.837**	**0.906**	**0.949**	**5.365**

Recall@K means the better the performance of the model.

$$\text{Recall@K} = \frac{1}{N}\sum_{j=1}^{N} est_j^K \tag{23}$$

where N is the number of samples in the testing set. If the top K ranked returned items match the semantic class of the query item, then $est_j{}^K = 1$, otherwise $est_j^K = 0$.

C. IMPLEMENTATION DETAILS

1) FEATURE PRE-PROCESSING METHOD

As Convolutional neural networks (CNNs) have been successfully used for deep learning applications, the AlexNet [34] network pre-trained from ImageNet [35] is used to extract image features. Additionally, the Bi-GRU network is exploited to extract text features.

The confidence threshold of faster R-CNN [38] is set to 0.1, and the maximum number of detected targets per image is set to 9. Note that, after the dimension reduction, the features of images are transformed from 4096 dimensions to 1024 dimensions. Besides, the dimensionality of the word embedding is initially set as 300 dimensions, and the dimensionality of the hidden state in Bi-GRU [36] is set as 1024. Consequently, the image features and text features have the same feature dimension.

2) NETWORK TRAINING DETAILS

We divide each sub-network into U level of granularities $(L_1, L_2, \ldots, L_U)$, and the value of U is set as 3 to balance the performance of retrieval and the computation cost. Table.1 shows the width, height, and step size of the sliding window in the image modality. Similarly, it also shows the length and step size of the sliding window in the text modality. Hence, the number of image regions (M) is 1, 4, and 9, respectively, and the number of sliding windows (N) for text is 1, 4, and 9. In addition, $\sigma, \theta\ \gamma$ are set to 50, 20, and 0.2 respectively, and the batch size (B) is set to 128.

D. COMPARED METHODS

To validate the efficiency of our proposed FB-Net, we compare it with several state-of-the-art methods, in which seven non-DNN-based cross-modal retrieval methods (*i.e.*, CCA [3], CMCP [23], JRL [25], JFSSL [26], and S^{2}2UPG [27]) and nine DNN-based methods (*i.e.*, DCCA [7], CCL [12], SCAN [15], GXN [28], VSESC [16], MAVA [29], SGRAF [21], SCL [39], CGMN [40], NAAF [30], and VSRN++ [31]) are contained. Note that the comparison methods are implemented using the authors' public source codes and are enumerated as follows.

•CCA [3] discovers linear mapping matrices in order to optimize pairwise correlations between two heterogeneous data sets in a subspace.

•CMCP [23] handles both positive and negative correlations between various modalities, and simultaneously propagates correlations between any heterogeneous data combinations.

•JRL [25] uses both the semi-supervised regularization and the sparse regularization to learn the common subspace.

•JFSSL [26] applies the graph regularization to cross-modal data in subspace learning, thus preserving the inter-modal and intra-modal similarity relationships.

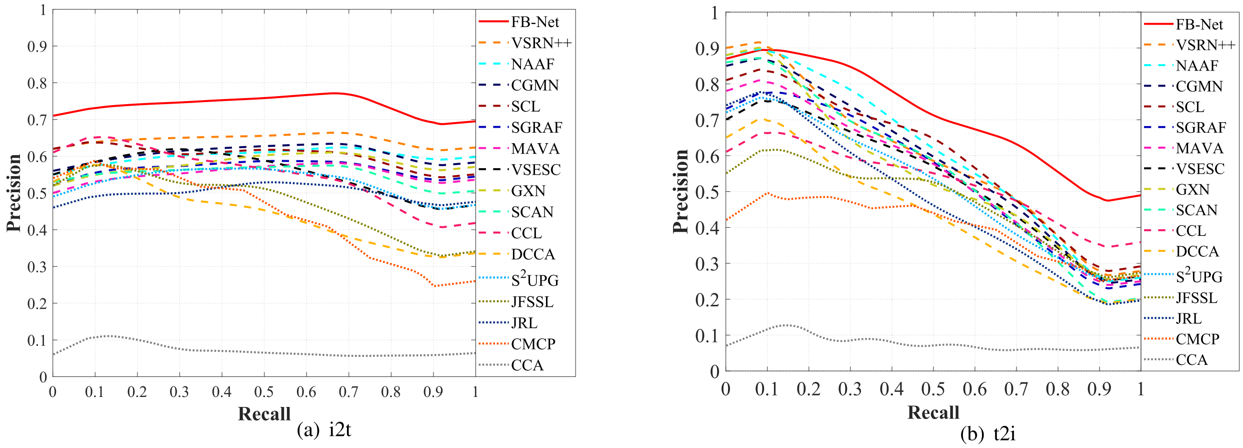
•S^{2}UPG [27] uses a cross-media feature learning framework with the unified block graph regularization, and it uses fine-grained data to highlight important local areas.

•DCCA [7] solves the data scalability problem and learns to correlate the public subspace better.

•CCL [12] designs a hierarchical network, which considers both the coarse-grained and the fine-grained data.

•SCAN [15] uses an attention mechanism to align all fine-grained data and thus mine the correspondence between fine-grained data.

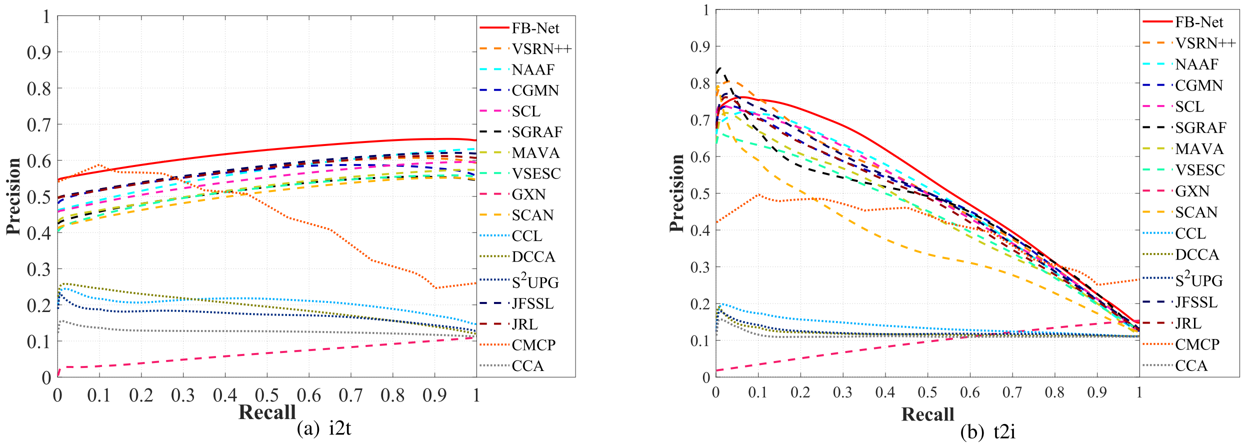
•GXN [28] incorporates two generative networks into the feature embedding and learns high-level abstract and local basis representations.



(a) i2t

(b) t2i

FIGURE 5. PR curves on Pascal Sentence dataset.



(a) i2t

(b) t2i

FIGURE 6. PR curves on Wikipedia dataset.

•VSESC [16] builds a visual embedding space and a textual embedding space, then integrates them with semantic consistency.

•MAVA [29] proposes a visual-textual dual attention mechanism, and distinguishes the fine-grained data at the relational level and local level.

•SGRAF [21] learns vector-based similarity representations for the task of image-text matching, and then infers and integrates the global and local similarities through the attention mechanism.

• SCL [39] takes advantage of the correlations between intra- and inter-modality items to acquire more discriminative features for multi-modal data.

• CGMN [40] uses graph convolutional networks to investigate the intra-relation in images and sentences and accomplishes interrelation reasoning between regions and words without impacting search efficiency.

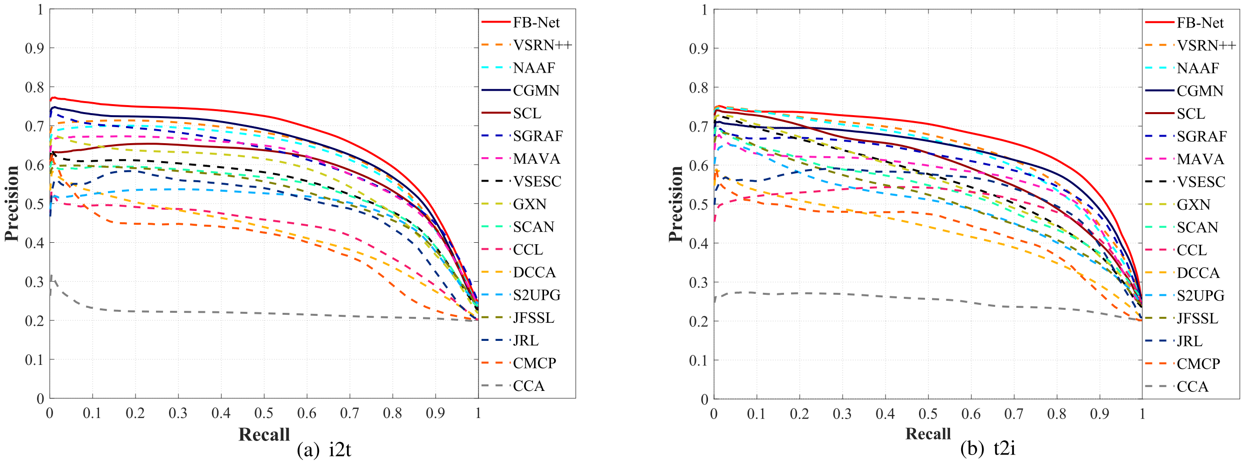
•NAAF [30] proposes the Negative-aware attention framework, which uses both the positive impact of matched blocks and the negative impact of unmatched blocks to estimate the similarity between images and texts.

•VSRN++ [31] proposes an image-text embedding framework for cross-modal retrieval, and implements regional relational reasoning and global semantic reasoning to generate feature representations.

The MAP scores of all the methods for image-text retrieval on the Pascal Sentence, Wikipedia, and NUS-WIDE dataset are presented in Table.2. The values that are better than others are indicated in boldface. From Table.2, we have the following observations.

First, the proposed FB-Net achieves the best performance on all datasets and significantly outperforms the previous best model VSRN++. This result indicates that it is effective to incorporate the foreground and background semantic units in a unified network, which well mines the complementarity between the foreground and background areas and further validates the motivation of FB-Net.

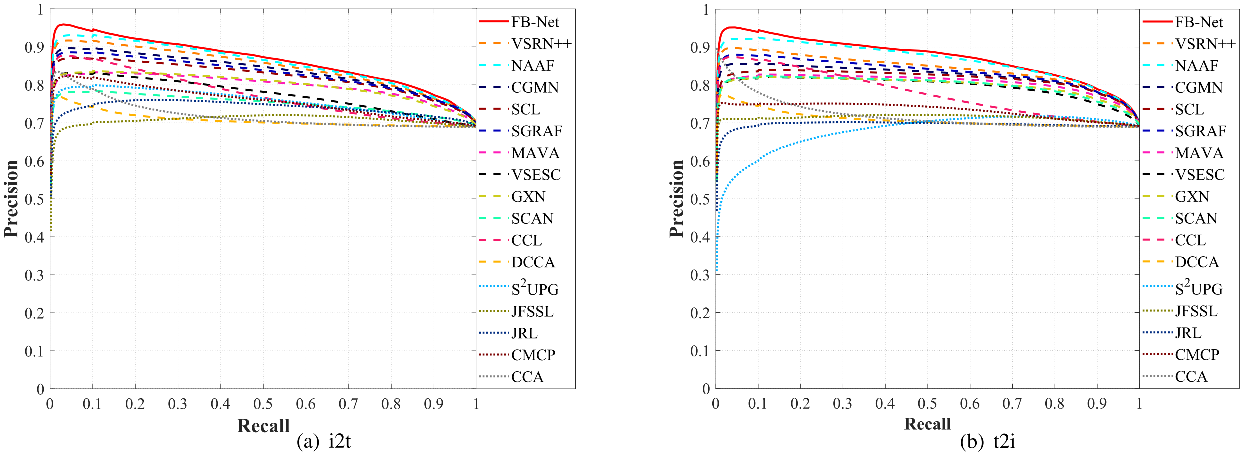
Second, compared with the Pascal Sentence dataset, the MAP scores of all methods (Except for CCA) are significantly lower when using the Wikipedia dataset. The primary explanation may be that the semantic categories of Wikipedia, such as "art," "history," and "war," are more abstract than



(a) i2t

(b) t2i

FIGURE 7. PR curves on NUS-WIDE-5K dataset.



(a) i2t

(b) t2i

FIGURE 8. PR curves on NUS-WIDE-25K dataset.

TABLE 7. Comparison of the training time on all datasets, the symbol 's' represents second.

Methods	Pascal Sentence	Wikipedia	NUS-WIDE-25K	NUS-WIDE-5K
DCCA(2013)	2177s	3677s	14,442s	3296s
SCAN(2018)	3460s	4997s	49,750s	11,900s
MAVA(2019)	4351s	5789s	52,092s	12,093s
SGRAF(2021)	5024s	6420s	57,435s	16,757s
SCL(2022)	3851s	4761s	50,178s	15,070s
CGMN(2022)	6019s	7640s	59,048s	17,331s
NAAF(2022)	4596s	6031s	58,584s	15,553s
VSRN++(2022)	6579s	8279s	60,109s	18,365s
FB-Net(ours)	5835s	7364s	58,697s	16,241s

those of Pascal Sentence. It indicates that learning the semantic content of these categories is more difficult. On the contrary, the semantic categories in Pascal Sentence, such as "bicycle," "bird," and "horse," are more concrete and easier to learn. Therefore, it is very reasonable that the MAP scores on Wikipedia for almost all methods are much lower than that on Pascal Sentence.

Third, the image-text pairs in NUS-WIDE-25K relate to various semantic concepts, raising the probability of identifying the real semantic concepts. As a result, the scores of MAP and Recall@K on NUS-WIDE-25K are noticeably higher compared to other datasets. The text modality of Pascal Sentence and Wikipedia are sentences, while the NUS-WIDE-25K and NUS-WIDE-5K are a set of tags. Noticeably, the MAP and Recall@K values also show that FB-Net can achieve superior performance in image-text retrieval whether using sentences or tags.

Additionally, the performance of image-text retrieval is also evaluated by Recall@K. Specifically, the parameters K among Recall@K are set to 1, 5, and 10 (abbreviated as

R@1, R@5, R@10). We also present an additional criterion R@sum, which is produced by adding R@1, R@5, and R@10 in both the i2t and t2i directions, to further show the performance of Recall@K. As shown in Table.3, Table.4, Table.5 and Table.6, FB-Net are not the best at i2t on Pascal Sentence with R@5 and R@10, and at t2i on Wikipedia with R@1 and R@5. However, FB-Net still achieves the best overall performance on R@sum, especially FB-Net significantly outperforms the previous best model VSRN++.

Fig.5, Fig.6, Fig.7 and Fig.8 show the PR curves of the i2t and t2i retrieval tasks on Pascal Sentence, Wikipedia, NUS-WIDE-25K and NUS-WIDE-5K. It can be observed that FB-Net achieves the best overall performance because the PR curves of FB-Net cover more areas than all other methods. This result indicates that our proposed FB-Net can effectively improve the cross-modal retrieval performance. In general, the PR curves of DNN-based methods are better than the non-DNN-based methods.

We perform a comparative experiment that focuses on the training times of deep learning techniques in order to more accurately assess our approach. Our conclusions from Table 7 are as follows. First, when it comes to image-text retrieval, DCCA and SCAN work less well than other deep learning methods despite having the shortest training times.

Second, despite FB-Net requires more training time than some models (i.e., such as DCCA, SCAN, MAVA, SGRAF, SCL, CGMN, NAAF, and VSRN++), it is significantly superior to them on the task of image-text retrieval. It is worth noting that VSRN++ is only second to FB-Net, however, VSRN++ requires the longest training time.

In addition, the following observations are revealed from the above experimental results.

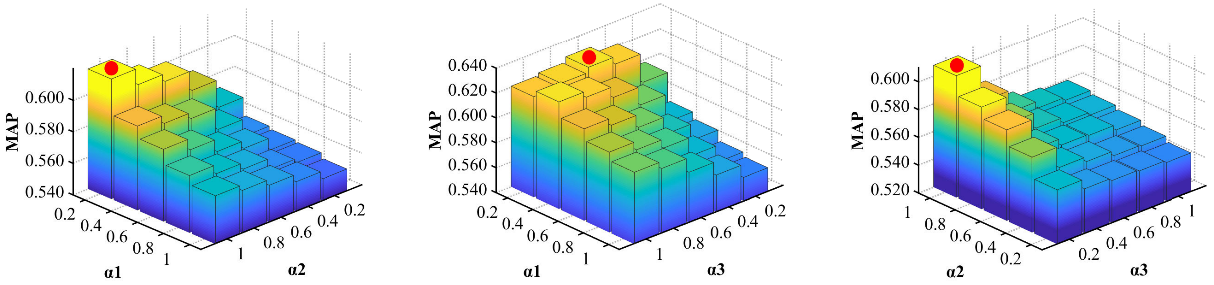
- We compared seven traditional non-DNN-based methods. Namely CCA, CMCP, JRL, JFSSL, and S^2UPG. In particular, S^2UPG outperforms other traditional non-DNN-based methods in most cases. This is because S^2UPG considers fine-grained data in cross-modal feature learning. In contrast, other methods only exploit coarse-grained data. Additionally, CMCP, JRL, and JFSSL are superior to CCA since they fully utilize semantic concepts to increase the interval distance between the various semantic concepts.
- The majority of DNN-based methods outperform non-DNN methods thanks to their superior capacity for revealing non-linear cross-modal connections. For example, DCCA achieves significant performance compared to CCA. This is because DCCA maximizes the correlation on top of two subnetworks by combining the CCA with the deep network.
- Models based on the attention mechanism perform better than DCCA and CCL, the reasons are that these methods accurately compute the cross-modal similarity via aligning image patches and words. Specifically, SCAN and VSESC focus on regions of images and words and use them as the corresponding contexts to compute cross-modal similarity. VSESC is superior to SCAN because it adds the semantic consistency constraint to the objective function. Additionally, SCL offers a contrastive learning-based self-supervised correlation learning paradigm that creates a weight-sharing scheme, and reduces the modality-invariant loss in the common space.
- However, SCAN and VSESC only consider fine-grained relations, while ignore other important clues. Unlike them, MAVA estimates image-text similarity at the global, local, and relationship level, therefore, it outperforms SCAN and VSESC. Furthermore, SGRAF performs better than MAVA, because it effectively reduces the irrelevant interactions at the coarse-grained and fine-grained levels. Particularly, NAAF outperforms the above approach, because it utilizes both the positive effects of matching pairs and the negative effects of mismatched pairs. Moreover, CGMN uses Graph Convolution Networks (GCNs) for intra-relation reasoning. As CGMN demonstrates its superiority in jointly representing different modalities, it works generally better than GXN.

In summary, FB-Net achieves the best results on MAP, PR curve, and Recall@K, which well demonstrates the effectiveness and advantage of FB-Net for image-text retrieval. Concretely, FB-Net accurately describes complicated and non-linear cross-modal relationships, which is a significant advantage over non-DNN-based methods. Since FB-Net uses both coarse-grained and fine-grained interactions, it outperforms SCAN, VSESC, GXN, and NAAF with relative ease. Although MAVA and SGRAF make full use of coarse-grained and fine-grained data, FB-Net surpasses them because:

- FB-Net collects rich complementary semantic information on various data granularity and fully exploits the complementary relationship between multi-granularity foreground and multi-granularity background areas. F-Net and B-Net aim to thoroughly capture the multi-granularity semantic units existing in the foreground and background areas respectively.
- FB-Net accurately captures and matches the semantic units in images and texts through multi-scale semantic scanning and alignment. Therefore, it can accurately describe the complete picture of the semantic concepts and discover potential relationships among various granularities.
- DSAW loss dynamically assigns proper weights to each sample pair, hence, FB-Net can fully utilize the relative relationships among samples to estimate the image-text similarity accurately.

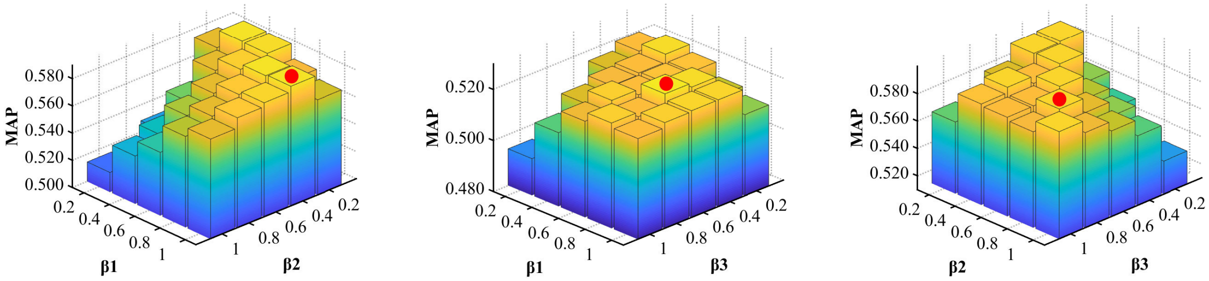
E. PARAMETER SENSITIVITY ANALYSIS

Several hyper-parameters are contained in FB-NET, for instance, $(\alpha_1, \alpha_2, \alpha_3)$, $(\beta_1, \beta_2, \beta_3)$, and (λ_1, λ_2) in Eq.7. Note that the values of these hyper-parameters are all selected from the range of $\{0.2, 0.4, 0.6, 0.8, 1.0\}$, and all experiments



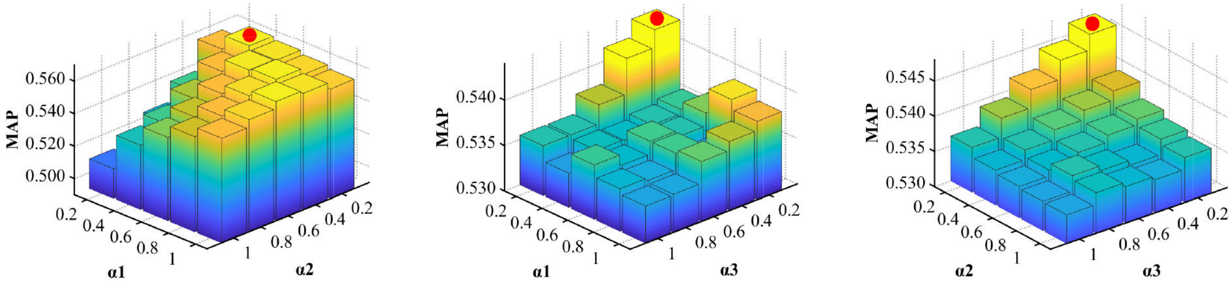
(a) Fix α_3 (b) Fix α_2 (c) Fix α_1

FIGURE 9. MAP scores on the Pascal Sentence relative to $\alpha_1, \alpha_2, \alpha_3$.



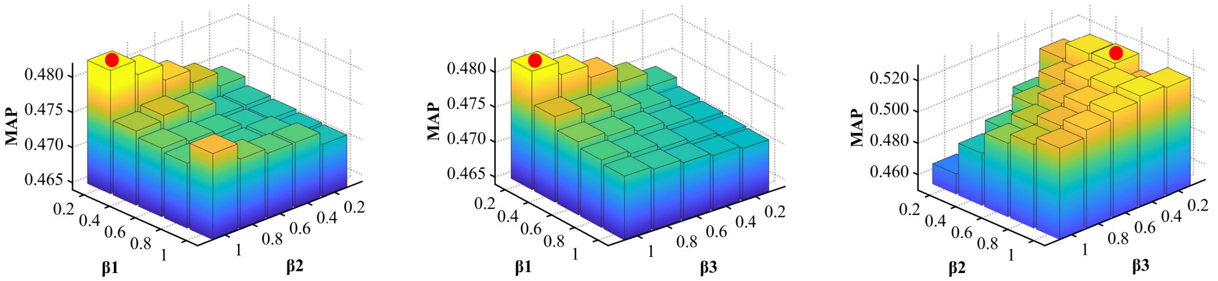
(a) Fix β_3 (b) Fix β_2 (c) Fix β_1

FIGURE 10. MAP scores on the Pascal Sentence relative to $\beta_1, \beta_2, \beta_3$.



(a) Fix α_3 (b) Fix α_2 (c) Fix α_1

FIGURE 11. MAP scores on the Wikipedia relative to $\alpha_1, \alpha_2, \alpha_3$.

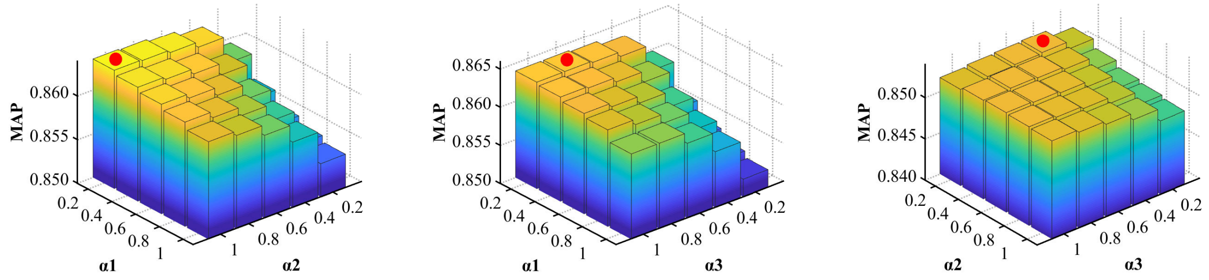


(a) Fix β_3 (b) Fix β_2 (c) Fix β_1

FIGURE 12. MAP scores on the Wikipedia relative to $\beta_1, \beta_2, \beta_3$.

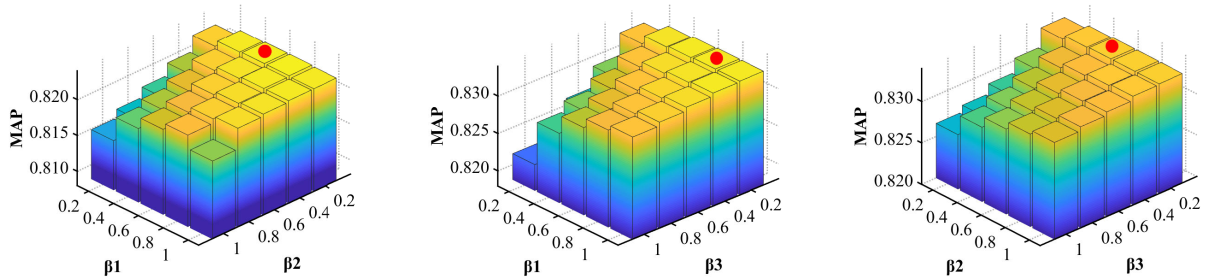
about the parameter sensitivity analysis use the average values of MAP on both i2t and t2i.

Firstly, we reveal the effect of each component in F-Net, that is, we explore the variation of MAP scores when $\alpha_1, \alpha_2, \alpha_3$ change. To better illustrate the variation trend, we fix one parameter of $\alpha_1, \alpha_2, \alpha_3$, and let the other two change (shown in Fig. 9, Fig. 11, Fig. 13, and Fig. 15). We can observe from Fig. 9, Fig. 11,



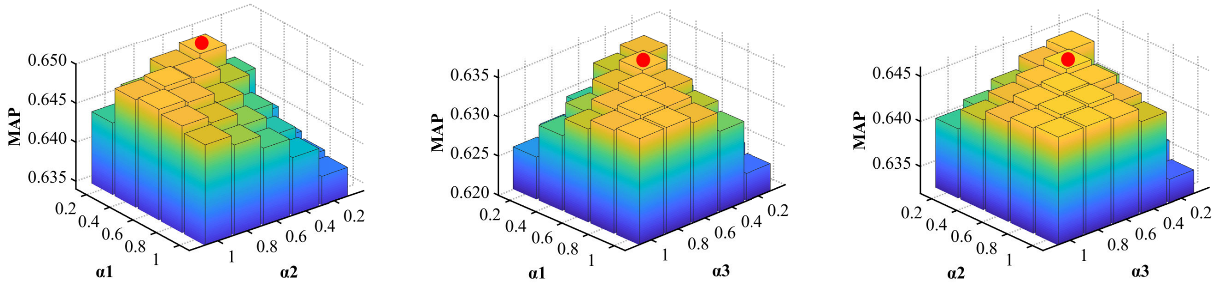
(a) Fix α_3 (b) Fix α_2 (c) Fix α_1

FIGURE 13. MAP scores on the NUS-WIDE-25K relative to $\alpha_1, \alpha_2, \alpha_3$.



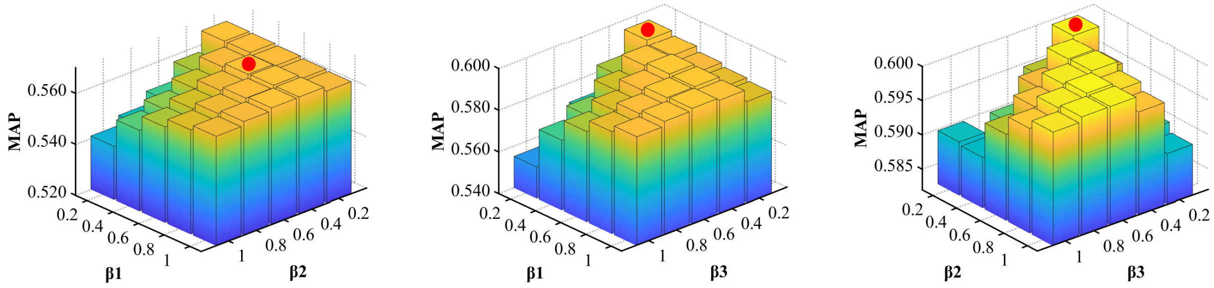
(a) Fix β_3 (b) Fix β_2 (c) Fix β_1

FIGURE 14. MAP scores on the NUS-WIDE-25K relative to $\beta_1, \beta_2, \beta_3$.



(a) Fix α_3 (b) Fix α_2 (c) Fix α_1

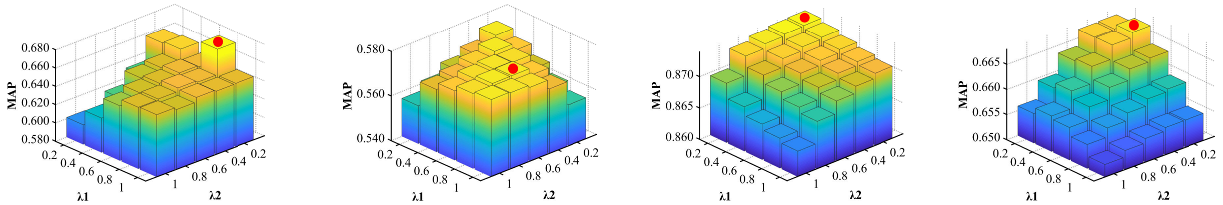
FIGURE 15. MAP scores on the NUS-WIDE-5K relative to $\alpha_1, \alpha_2, \alpha_3$.



(a) Fix β_3 (b) Fix β_2 (c) Fix β_1

FIGURE 16. MAP scores on the NUS-WIDE-5K relative to $\beta_1, \beta_2, \beta_3$.

Fig. 13, Fig. 15, and Table.10 that the fused MAP scores of parameters (α_1, α_2),(α_1, α_3),(α_2, α_3) in F-Net are all higher than MAP scores with single-granularity. This sufficiently indicates that effectively mining the complementary relationships between multi-granularity data is an effective way to bridge the granularity gap. Additionally, we also find that the best settings of (α_1, α_2),(α_1, α_3),(α_2, α_3) are set to (0.2, 1), (0.2, 0.6), and (1, 0.2) for Pascal Sentence,



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

FIGURE 17. MAP scores variation with respect to λ_1, λ_2 when we fix $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$.

set to (0.4, 0.2), (0.2, 0.2), and (0.2, 0.2) for Wikipedia,set to (0.2, 1), (0.2, 0.8), and (0.2, 0.4) for NUS-WIDE-25K,and set to (0.2, 0.4), (0.4, 0.4), and (0.4, 0.4) for NUS-WIDE-5K.

Secondly, we also show the effect of each component in B-Net in Fig.10, Fig. 12, Fig.14, and Fig.16. Similar to the parameter sensitivity analysis for $(\alpha_1, \alpha_2, \alpha_3)$, we permit any two parameters of $(\beta_1, \beta_2, \beta_3)$ to change and fix the other one. Just like the rules found in F-Net, we have two findings. First, the fused MAP scores of parameters (β_1, β_2) , (β_1, β_3), (β_2, β_3) are also higher than MAP scores with single-granularity. This is because the receptive field of image regions from a single-granularity is limited, and the fusion of multi-granularity data can enlarge the receptive field of images. Second, the MAP scores with fusing (β_2, β_3) are higher than that by fusing (β_1, β_2) or (β_1, β_3), respectively, because the fine-grained features contain more specific semantic information than the coarse-grained features in background areas, and thus contribute more to image-text retrieval. More specifically, the best parameter settings of (β_1, β_2),(β_1, β_3),(β_2, β) are (1, 0.4), (0.8, 0.6), and (0.8, 0.8) for Pascal Sentence, are (0.2, 1), (0.2, 1), and (0.6, 0.2) for Wikipedia, are (0.6, 0.2), (0.8, 0.2), and (0.6, 0.2) for NUS-WIDE-25K,are (0.6, 0.4), (0.2, 0.2), and (0.2, 0.2) for NUS-WIDE-5K.

Thirdly, to further verify the importance of each subnetwork, we provide the variation of MAP scores when λ_1 and λ_2 change in Fig.17, from which some interesting observations are revealed: 1) F-Net plays a more important role than B-Net in image-text retrieval, because foreground areas contain more valuable semantic information than background areas. 2) B-Net still provides useful semantic information for image-text retrieval, and it complements the semantic information of F-Net. 3) the best parameters setting of (λ_1, λ_2) for Pascal Sentence, Wikipedia, NUS-WIDE-25K, NUS-WIDE-5K are set to (0.8, 0.2), (0.8, 0.6), (0.2, 0.2), and (0.4, 0.2) respectively.

F. ABLATION STUDIES

To further verify the effectiveness of each component of FB-Net, we conduct a series of ablation studies, which are divided into two parts: (1) Ablation study 1: Exploring the contribution of the multi-scale semantic scanning in each subnetwork, and (2) Ablation study 2: Exploring the contribution of our proposed DSAW loss.

TABLE 8. Experimental configurations of different models in ablation study 1.

Models	Configuration of F-Net		
	M=1,N=1	M=4,N=4	M=9,N=9
F-Net(L_1)	✓		
F-Net(L_2)		✓	
F-Net(L_3)			✓
F-Net(L_1+L_2+L_3)	✓	✓	✓
Models	Configuration of B-Net		
	M=1,N=1	M=4,N=4	M=9,N=9
B-Net(L_1)	✓		
B-Net(L_2)		✓	
B-Net(L_3)			✓
B-Net(L_1+L_2+L_3)	✓	✓	✓
Models	Configuration of FB-Net		
	M=1,N=1	M=4,N=4	M=9,N=9
FB-Net(L_1+L_2+L_3)	✓	✓	✓

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

Models	Subnetwork		Loss function	
	F-Net	B-Net	Triplet loss	DSAW loss
F-Net (Tri)	✓		✓	
F-Net (DSAW)	✓			✓
B-Net (Tri)		✓	✓	
B-Net (DSAW)		✓		✓
FB-Net (Tri)	✓	✓	✓	
FB-Net (DSAW)	✓	✓		✓

1) ABLATION STUDY 1

We conduct the ablation study by fusing various configurations of F-Net and B-Net in Table.8, and results are shown in Table.10, in which L_i represents the sub-network at the ith level, M refers to the number of regions that an instance is divided, and N is the number of sliding windows. Note that "✓" indicates that the corresponding component is included. Some observations from Table.10 are listed as follows.

- Combining the sub-network at L_1 level, L_2 level and L_3 level, F-Net(L_1+L_2+L_3) achieves the best performance on all datasets. Similarly, B-Net(L_1+L_2+L_3) outperforms all other configurations of B-Net. Hence, it can be concluded that 1) each level of F-Net and B-Net component plays a very positive role in image-text similarity learning, and 2) there actually exists a complementary relationship between various data granularities, as well as between the foreground and background areas.
- The contribution of each level of the sub-network is unbalanced. For F-Net, the best performance is achieved

TABLE 10. Ablation study 1: exploring the contribution of the multi-scale semantic scanning in each sub-network.

Models	Pascal Sentence			Wikipedia			NUS-WIDE-25K			NUS-WIDE-5K		
	i2t	t2i	Avg	i2t	t2i	Avg	i2t	t2i	Avg	i2t	t2i	Avg
F-Net(L_1)	0.6167	0.5451	0.5809	0.5742	0.4310	0.5026	0.8529	0.8486	0.8477	0.6166	0.6122	0.6144
F-Net(L_2)	0.5694	0.5129	0.5411	0.5876	0.4793	0.5335	0.8595	0.8582	0.8588	0.6416	0.6289	0.6352
F-Net(L_3)	0.5597	0.5241	0.5419	0.5616	0.4491	0.5053	0.8498	0.8454	0.8476	0.5977	0.5696	0.5836
F-Net(L_1+L_2+L_3)	**0.6606**	**0.5989**	**0.6297**	**0.6080**	**0.5129**	**0.5605**	**0.8649**	**0.8652**	**0.8650**	**0.6561**	**0.6444**	**0.6502**
B-Net(L_1)	0.4978	0.4471	0.4724	0.5070	0.3670	0.4370	0.8020	0.8019	0.8019	0.5226	0.5126	0.5176
B-Net(L_2)	0.5680	0.5097	0.5388	0.5267	0.4154	0.4711	0.8247	0.8193	0.8320	0.5654	0.5572	0.5613
B-Net(L_3)	0.4922	0.4744	0.4833	0.5125	0.4298	0.4712	0.8388	0.8274	0.8331	0.5589	0.6046	0.5817
B-Net(L_1+L_2+L_3)	**0.6118**	**0.5652**	**0.5885**	**0.5732**	**0.4692**	**0.5212**	**0.8438**	**0.8372**	**0.8405**	**0.5930**	**0.6064**	**0.5997**
FB-Net(L_1+L_2+L_3)	**0.7041**	**0.6472**	**0.6756**	**0.6183**	**0.5312**	**0.5747**	**0.8814**	**0.8723**	**0.8768**	**0.6695**	**0.6672**	**0.6683**

TABLE 11. Ablation study 2: exploring the contribution of our proposed DSAW loss.

Models	Pascal Sentence			Wikipedia			NUS-WIDE-25K			NUS-WIDE-5K		
	i2t	t2i	Avg	i2t	t2i	Avg	i2t	t2i	Avg	i2t	t2i	Avg
F-Net (Tri)	0.6432	0.5715	0.6074	0.6002	0.5076	0.5539	0.8651	0.8606	0.8628	0.6429	0.6218	0.6323
F-Net (DSAW)	0.6606	0.5989	0.6297	0.6080	0.5129	0.5605	0.8649	0.8652	0.8650	0.6561	0.6444	0.6502
B-Net (Tri)	0.5945	0.5472	0.5709	0.5628	0.4595	0.5111	0.8355	0.8268	0.8311	0.5739	0.5924	0.5831
B-Net (DSAW)	0.6118	0.5652	0.5885	0.5732	0.4692	0.5212	0.8438	0.8372	0.8405	0.5930	0.6064	0.5997
FB-Net (Tri)	0.6928	0.6337	0.6633	0.6090	0.5217	0.5654	0.8728	0.8687	0.8707	0.6487	0.6521	0.6504
FB-Net (DSAW)	**0.7041**	**0.6472**	**0.6756**	**0.6183**	**0.5312**	**0.5747**	**0.8814**	**0.8723**	**0.8768**	**0.6695**	**0.6672**	**0.6683**

by F-Net(L_1) on Pascal Sentence, while that is achieved by F-Net(L_2) on Wikipedia, NUS-WIDE-25K, and NUS-WIDE-5K. For B-Net, the highest MAP score is gained by B-Net(L_2) on Pascal Sentence and is got by B-Net(L_3) on Wikipedia, NUS-WIDE-25K, and NUS-WIDE-5K. The reasons lie in that there is an obvious contrast between the multi-granularity semantic unit distributions of Pascal Sentence, Wikipedia, NUS-WIDE-25K, and NUS-WIDE-5K. Therefore, it is of great importance to provide a comprehensive and complete scan for the multi-granularity semantic units, which are scattered in both foreground and background areas.

2) ABLATION STUDY 2

We further test the effectiveness of our proposed loss function in this ablation study, in which Tri and DSAW represent the triplet loss function and DSAW loss function respectively. The experimental configurations and results are shown in Table.9 and Table.11, and more analysis is given as follows.

- Compared to other models, FB-Net(DSAW) achieves the best performance, because it is not comprehensive to mine the semantic units by a single sub-network. That is, FB-Net(DSAW) fully combines multi-granularity foreground and multi-granularity background information and identifies multi-granularity semantic units in foreground and background areas comprehensively and accurately.
- Furthermore, F-Net always outperforms B-Net, regardless of what type of loss function is used. However, the former cannot replace the latter in the task of cross-modal retrieval, because they are complementary to each other. Overall, by combining F-Net and B-Net, FB-Net (DSAW) achieves better performance than any single sub-network.
- DSAW loss significantly outperforms triplet loss, because it dynamically assigns proper weights to each pair of samples during the training process. Therefore, the DSAW loss effectively mines more complete information for feature optimization and thus learns more accurate image-text similarity.

G. VISUALIZATION ANALYSIS AND TYPICAL EXAMPLES OF IMAGE-TEXT RETRIEVAL

The stacked cross-attention mechanism plays an important role in our proposed FB-Net. First, image patches and words are used as contexts through text-grounded embedding space and image-grounded embedding space. Then they are applied to the alignment of semantic units.

Second, to explicitly investigate the effectiveness of the attention mechanism in FB-Net, we conduct visualization analysis in Fig.18.

As is shown in Fig.18, we provide four examples, and each one is made of an image-pair. For each image, we offer two types of visualization results: 1) Heat map visualization, and 2) Grid visualization. For each text, the importance of each word is visualized separately. It can be clearly observed that 1) the salient objects in images and their related textual descriptions are paid more attention, such as "airplane" in Fig.18(a), "woman" and "horse" in Fig.18(b), "cat" in Fig.18(c), and "bus" in Fig.18(d).

Furthermore, we provide the top-ranked 5 retrieval results of i2t and t2i corresponding to a specific query in Fig.19. For each direction of the image-text retrieval task, we choose two queries that belong to different semantic concepts. Especially, two recent DNN-based methods (i.e., VSRN++ and NAAF) are compared with ours. Furthermore, We have selected two specific classes, "Sheep" and "Bus," for i2t queries, while "Airplane" and "Horse" for i2t queries.

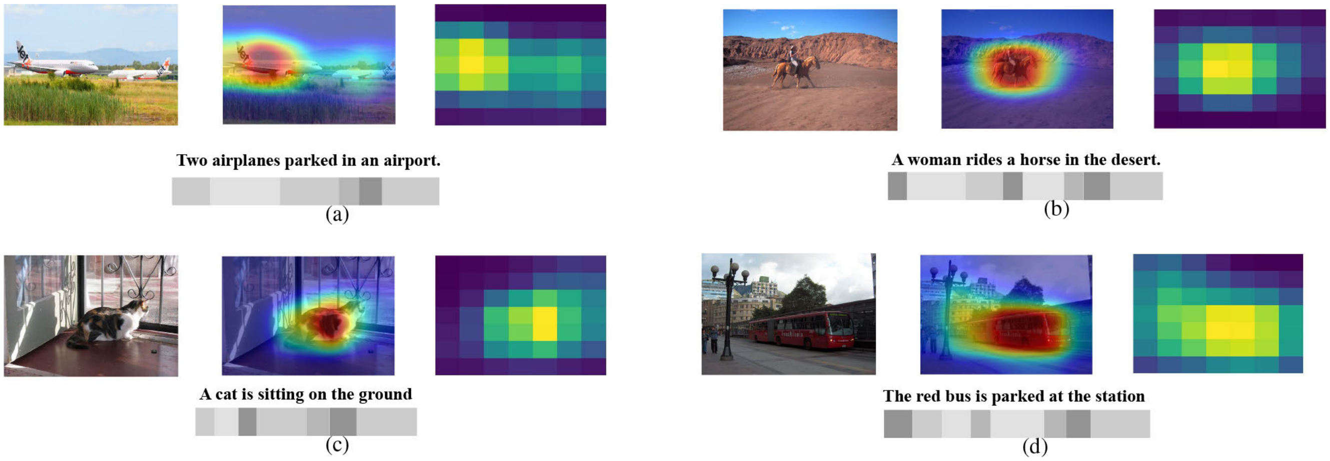

Two airplanes parked in an airport.
(a)
A woman rides a horse in the desert.
(b)
A cat is sitting on the ground
(c)
The red bus is parked at the station
(d)

FIGURE 18. Visualization results of the stacked cross-attention mechanism in FB-Net. The brightness represents the weight of attention, which further illustrates the significance of the local-level regions identified by the stacked cross-attention mechanism through semantic unit alignment. Note that the lighter areas mean the larger attention weights.

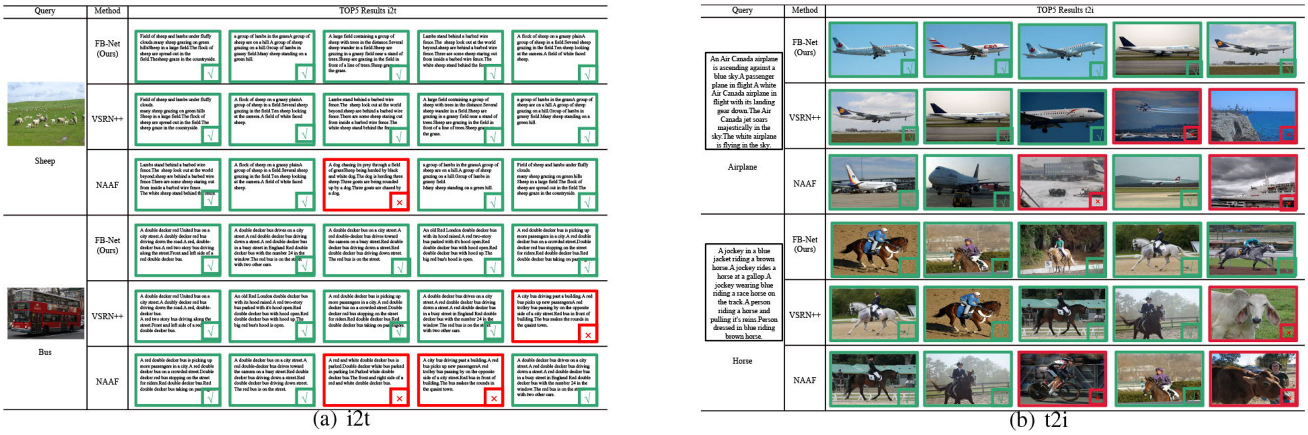

Query
Method
TOP5 Results i2t
Sheep
Bus
FB-Net (Ours)
VSRN++
NAAF
(a) i2t
Query
Method
TOP5 Results t2i
An Air Canada airplane is ascending against a blue sky.A passenger plane in flight.A white Air Canada airplane in flight with its landing gear down.The Air Canada jet soars majestically in the sky.The white airplane is flying in the sky.
Airplane
A jockey in a blue jacket riding a brown horse.A jockey rides a horse at a gallop.A jockey wearing blue riding a race horse on the track.A person riding a horse and pulling it's reins.Person dressed in blue riding brown horse.
Horse
FB-Net (Ours)
VSRN++
NAAF
(b) t2i

FIGURE 19. Typical examples of image-text retrieval.

It can be observed from Fig.19 that the top 5 items returned by our proposed FB-Net are all correct, that is, FB-Net achieves the best performance. However, there are several mistakes in the retrieval results of VSRN++ and NAAF. Compared with VSRN++ and NAAF, the advantage of FB-Net is that it effectively mines the semantic information in salient objects of images, and thus it is much easier to hit the correct semantic category.

V. CONCLUSION

In this paper, we propose a dual-branch foreground-background fusion network to handle the task of image-text retrieval. To gain more complete semantic information and further enhance the performance of image-text retrieval, FB-Net aims to gather and fuse all semantic units existing in different modalities and different granularities scattered in the foreground and background areas. More specifically, FB-Net uses multi-scale semantic scanning and alignment to accurately capture multi-granularity foreground and background semantic units in images and texts, and image-text similarity is initially calculated by SUA. Afterward, image-text similarities generated from all levels of each subnetwork are optimized by the DSAW loss. Experimental results demonstrate the superiority of our proposed FB-Net over representative state-of-the-art methods on image-text retrieval, and ablation studies further verify the effectiveness of each component of FB-Net.

Notably, we also empirically discovered some shortcomings of our proposed strategy. For instance, the training time of FB-Net is not very competitive. As a result, future work will primarily focus on three fields. Firstly, the positional relationships among fine-grained patches of each modality will be regarded as a type of significant semantic information in cross-modal similarity learning. Secondly, we will try to speed up the FB-Net training procedure by optimizing its framework. Thirdly, we will verify the scalability of FB-Net, that is, we plan to extend FB-Net to other types of cross-modal retrieval tasks, such as video queries text, or audio queries image.

REFERENCES

[1] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[3] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[4] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[5] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, pp. 365–377, Oct. 2000.

[6] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, Nov. 2003, pp. 604–611.

[7] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1247–1255.

[8] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Sep. 2014, pp. 7–16.

[9] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.

[10] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, 2020.

[11] C. Jia, Y. Yang, Y. Xia, Y. T. Chen, Z. Parekh, H. Pham, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 4904–4916.

[12] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.

[13] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.

[14] A. A. K. Joulin and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[15] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 201–216.

[16] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Cross-modal image-text retrieval with semantic consistency," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1749–1757.

[17] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5583–5594.

[18] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 3–11.

[19] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3536–3545.

[20] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-BERT: Vision-language pre-training on fashion domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12647–12657.

[21] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 2, 2021, pp. 1218–1226.

[22] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," 2019, *arXiv:1907.09748*.

[23] X. Zhai, Y. Peng, and J. Xiao, "Cross-modality correlation propagation for cross-media retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 2337–2340.

[24] S. U. Rehman, S. Tu, Y. Huang, and O. U. Rehman, "A benchmark dataset and learning high-level semantic embeddings of multimedia for cross-media retrieval," *IEEE Access*, vol. 6, pp. 67176–67188, 2018.

[25] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.

[26] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[27] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.

[28] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

[29] Y. Peng, J. Qi, and Y. Zhuo, "MAVA: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism," *IEEE Trans. Image Process.*, vol. 29, pp. 2728–2741, 2019.

[30] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15661–15670.

[31] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Image-text embedding learning via visual and textual semantic reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 641–656, Jan. 2023.

[32] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical Turk," *Proc. Workshop Creating Speech Lang. Data With Amazon's Mechanical Turk*, 2010, pp. 139–147.

[33] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 251–260.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[36] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[39] Y. Liu, J. Wu, L. Qu, T. Gan, J. Yin, and L. Nie, "Self-supervised correlation learning for cross-modal retrieval," *IEEE Trans. Multimedia*, early access, Feb. 16, 2022, doi: 10.1109/TMM.2022.3152086.

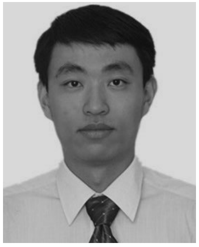
[40] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, "Cross-modal graph matching network for image-text retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 4, pp. 1–23, Nov. 2022.

[41] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. CIVR*, Santorini, Greece, Jul. 2009, pp. 1–9.

[42] L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Global and local semantics-preserving based deep hashing for cross-modal retrieval," *Neurocomputing*, vol. 312, pp. 49–62, Oct. 2018.

[43] S. U. Rehman, Y. Huang, S. Tu, and B. Ahmad, "Learning a semantic space for modeling images, tags and feelings in cross-media search," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Apr. 2019, pp. 65–76.

JUNHAO XU is currently pursuing the B.E. degree with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China. His research interests include cross-modal retrieval and deep learning.



ZHENG LIU received the Ph.D. degree from Shandong University, in July 2011. He is a Professor and a Ph.D. Supervisor with the School of Computer Science and Technology, Shandong University of Finance and Economics. His research interests mainly include cross-modal analysis and retrieval, automatic image annotation, and machine learning.

XINLEI PEI received the B.E. degree from the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China, in 2020, where he is currently pursuing the M.S. degree with the Shandong Provincial Key Laboratory of Digital Media Technology. His main research interests include multimedia information retrieval and deep learning.

SHUHUAI WANG is currently pursuing the bachelor's degree with the School of Computer Science and Technology, Shandong University of Finance and Economics. Her research interests include cross-modal retrieval and deep learning.

SHANSHAN GAO is a Professor and a Ph.D. Supervisor with the School of Computer Science and Technology, Shandong University of Finance and Economics. Her current research interests include intelligent graphics and image processing, data mining, and visualization.