# HGAN: Hierarchical Graph Alignment Network for Image-Text Retrieval

Jie Guo, Member, IEEE, Meiting Wang, Yan Zhou, Bin Song, Senior Member, IEEE, Yuhao Chi, Member, IEEE, Wei Fan, Jianglong Chang

Abstract-Image-text retrieval (ITR) is a challenging task in the field of multimodal information processing due to the semantic gap between different modalities. In recent years, researchers have made great progress in exploring the accurate alignment between image and text. However, existing works mainly focus on the fine-grained alignment between image regions and sentence fragments, which ignores the guiding significance of context background information. Actually, integrating the local finegrained information and global context background information can provide more semantic clues for retrieval. In this paper, we propose a novel Hierarchical Graph Alignment Network (HGAN) for image-text retrieval. First, to capture the comprehensive multimodal features, we construct the feature graphs for the image and text modality respectively. Then, a multi-granularity shared space is established with a designed Multi-granularity Feature Aggregation and Rearrangement (MFAR) module, which enhances the semantic corresponding relations between the local and global information, and obtains more accurate feature representations for the image and text modalities. Finally, the ultimate image and text features are further refined through three-level similarity functions to achieve the hierarchical alignment. To justify the proposed model, we perform extensive experiments on MS-COCO and Flickr30K datasets. Experimental results show that the proposed HGAN outperforms the state-of-the-art methods on both datasets, which demonstrates the effectiveness and superiority of our model.

Index Terms—Image-text retrieval, feature aggregation, graph convolution network, hierarchical alignment

# I. INTRODUCTION

IN recent years, with the rapid growth of multimedia data, multimodal information processing has become more and more important. As the two most commonly-used modalities, the image and text have prompted many researchers to study cross-modal tasks, including cross-modal retrieval [1], visual question answering [2], image captioning [3], etc. In particular, image-text retrieval (ITR) [4] task focuses on measuring the semantic similarity between images and texts. Although great progress has been made in recent years, the heterogeneous differences caused by the inconsistent forms of images and texts seriously hinder the performance of ITR in complex scenarios, thus the ITR task still remains a great challenge. To establish the intrinsic connection between images and texts, early works

Wei Fan and Jianglong Chang are with the Guangdong OPPO Mobile Telecommunications Corp., Ltd, China. (e-mails: richard.fan@oppo.com, changjianglong@oppo.com.)

Bin Song and Yuhao Chi are both the corresponding authors.

such as [5] extract image and text features separately using existing visual and language models used for other tasks, then directly convert them to the same dimensions as global representations. After that they measure the similarity of the global representations in this shared space, and subsequently optimize the model parameters based on matching facts. However, these methods only focus on global matching, that is, the alignment of the whole image and the whole sentence, and do not fully consider the potential relationship between the regions of image and the words of sentence, resulting in limited performance improvement and low interpretability. Due to the development of language representation models such as BERT [6], fine-grained features of text are readily available. Meanwhile, for the image modality, object detection methods such as Faster-RCNN [7] have also shown superior performance in extracting the fine-grained features. SCAN [8] is the first attempt to introduce object detection method into ITR task, and point out that there is an underlying alignment relationship between image regions and sentence fragments, which triggered a lot of researches on the finegrained image-text alignment. For example, by generating a guidance vector from the initially extracted fine-grained features, an adaptive feature optimization is accomplished in [9], which modifies the representation of the fine-grained features in another modality. Aiming to achieve more precise semantic alignment, some researchers optimize feature representations by attention mechanisms, such as [10] to discriminate negative pairs with similar semantic content but slightly different contextual information, improve information interactions between modalities based on cross-attention, and employ a multi-level alignment strategy with progressive matching to acquire more complementary and adequate semantic cues. Other researchers dedicate to exploring additional cues using graph convolutional networks, as in [11], they construct a vision graph from region features, infer the relationship between regions using GCN, and then input node features to GRU to produce more discriminative image features.

These methods have greatly promoted the progress of the ITR task. However, some intractable problems still exist. When studying the fine-grained alignment, many researchers ignore the importance of non-object elements such as the context background information. As shown in Fig. 1 (a), the fine-grained alignment method might match the sentence "A man is standing on a snowy path surrounded by evergreen trees" with both the left and right images since they both contain the objects "man", "snow path", and the relation "standing". But the matching result with the image on the

Jie Guo, Meiting Wang, Yan Zhou, Bin Song and Yuhao Chi are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, Shaanxi 710071, China. (e-mails: jguo@xidian.edu.cn, mtw@stu.xidian.edu.cn, zhouyan123y@stu.xidian.edu.cn, bsong@mail.xidian.edu.cn, yhchi@xidian.edu.cn.)

right is wrong due to the difference in the context background information of the image and text. Therefore, ignoring nonobject elements such as context background information may cause false matching results, especially for negative samples with similar objects but slightly different context background. Accordingly, this motivates us to explore the hierarchical alignment approach shown in Fig. 1 (b), where hierarchical alignment refers to multi-level alignment with different granularity features, that is, not only aligning fine-grained features but also taking into account the correspondence between coarse-grained features and mixed coarse-fine-grained intermediate granularity features. This strategy of focusing on multi-granularity characteristics has yielded excellent results in applications such as object recognition [12], visual classification [13], and text-based question answering [14]. In the task of image-text matching, this method fully considers object elements and non-object elements, which encourages more accurate matching of images and texts, and reduces the impact of the tremendous useless information existing in semantic alignment. Regarding the intermediate granularity, one way is to fuse coarse-grained and fine-grained features as hybrid multi-granularity features [15], and another way is to obtain features of different granularity by setting different sizes of the visual field to the feature extraction network [16]. The first method is straightforward and direct, but it necessitates the design of a suitable fusion process, whereas the second one is affected by the choice of the size of each visual field and demanding that the feature extraction module is not a black box. Moreover, researchers usually aggregate the features of image and text modality through max pooling [17] or average pooling [18] in the vision-language shared space, which both ignore the importance of the synergistic relationship of local object and global context. Specifically, these methods learn the local and global features separately and do not consider the impact of the object-context-fused information on ITR task. In summary, existing methods have two notable problems. First, most researchers consider the semantic alignment of images and texts from one perspective, such as the finegrained alignment or global alignment, and rarely take multigranularity feature fusion into account. Second, the existing feature aggregation approaches in the shared semantic space, which largely ignore the object-context information interaction of multimodal features, need to be further improved.

To address these issues, this paper proposes a novel Hierarchical Graph Alignment Network (HGAN) for image-text retrieval, which establishes a multi-granularity shared space with multi-granularity feature aggregation and rearrangement (MFAR) module and performs hierarchical image-text alignment through three-level similarity functions. Specifically, we first fuse the global and local features of the image, and construct the feature graphs for the image and text modality respectively. The feature graphs can preserve relative positional relationships, which is beneficial to explore the comprehensive multimodal representation. Then, a multi-granularity shared space is established with multi-layer MFAR module, which optimizes the image and text features through feature aggregation and feature rearrangement to accomplish multi-granularity feature fusion. The MFAR module can filter out the noisy



(b) Retrieval examples with the hierarchical alignment

Fig. 1. An illustration of retrieval examples of the fine-grained and hierarchical image-text retrieval. (a) represents the retrieval examples of fine-grained alignment, in which objects and relationships between images and texts are matched, but the result is wrong. (b) represents the retrieval examples of hierarchical alignment, which considers background context information and can distinguish negative samples with similar semantics but slightly different background. Note that "T" and "F" denote the correct and wrong matching between the image and text, respectively.

parts of the global-local semantic alignment and retain the dominate parts to enhance the semantic alignment with objectcontext-fused information. Finally, we design three similarity functions corresponding to three levels of fine-grained feature, unified feature and multi-granularity feature. The ultimate image and text features are further refined through three-level similarity functions to achieve hierarchical alignment in the multi-granularity shared space.

The main contributions of this paper are summarized as follows:

- We establish a multi-granularity shared space with the designed Multi-granularity Feature Aggregation and Rearrangement (MFAR) module to achieve multi-granularity feature fusion. The MFAR module explores the multigranularity feature denoising, which is dedicated to filter out the noisy parts and retain the dominate parts of the object-context-fused information.
- We propose a novel Hierarchical Graph Alignment Network (HGAN) to achieve the multi-level image-text alignment. The HGAN model aligns image and text features through multiple similarity functions to further improve the matching accuracy in the multi-granularity shared space.
- The proposed HGAN outperforms state-of-the-art imagetext retrieval methods on several benchmark datasets, e.g. MS-COCO (1K and 5K) and Flickr30K.

The rest of this paper is organized as follows. In Section II, we introduce the related work. Then all the details of the proposed method and the experiments are presented in Section III and Section IV. Finally, the conclusion is described in Section V.

## II. RELATED WORK

In this part, we will further introduce the related work including image-text retrieval and graph convolution network.

#### A. Image-Text Retrieval

Image-text retrieval aims at calculating the similarity between the images and texts, which can be mainly divided into three categories, global matching, regional matching and multi-level matching methods Global matching methods normally embed the global features of images and texts into a common embedding space, and then employ a ranking loss calculating the distance between image-text pair. Kiros et al. [19] utilize the convolution neural network (CNN) for image encoding as well as the recurrent neural network (RNN) for text encoding, and Faghri et al. [20] employ the ResNet152 and GRU modules to encode images and texts respectively. Regional matching methods cast lights on the fine-grained alignment of the regions of image and the words of sentence. Qu at ul. [21] adopt pyramid dilated convolution to obtain a multi-view representation from image region features, using the viewpoint that best matches the text feature to measure similarity. Chen et al. [22] design a generalized pooling operator to replace average pooling and max pooling, which can adaptively aggregate visual and language features to improve the performance of crossmodal retrieval. Multi-level matching methods consider multigranularity feature alignment of image and text, or matching of global feature, region feature, and other intermediate feature. The second is similar to the ontology depicted in [23], whose hierarchical levels map different granularity of features. Qi et al. [24] consider not only the global and local alignments but also the relation alignment across images and texts, which can learn more precise cross-modal relevance. Zeng et al. [25] propose a multi-layer graph convolutional network with object-level, object-relational-level, and higher-level learning sub-networks to learn hierarchical semantic correspondences by both local and global alignment. Huang et al. [26] propose a bi-directional spatial-semantic attention network, which uses the word-to-regions relation to deduce the most relevant image regions, and employ the visual object to words relation to infer the close words for visual objects in the images. Recently, Diao et al. [27] use similarity graph reasoning and similarity attention filtration module to reason about the relationship between global-local alignment information, focusing on more informative alignments and have achieved SOTA performance.

## B. Graph Convolution Network

Graph convolution network (GCN) has been successfully applied into many tasks in the cross-modal fields such as crossmodal retrieval [28], [29], image captioning [30], [31], visual question answering [32], [33] and visual entailment [34], [35]. Rather than merely focus on the similarities in content, GCN can discover the potential semantic relationships among different modalities and integrate the information of neighbor nodes. As a result, GCN is adopted into the field of crossmodal retrieval to learn the representations across different modalities, as it will be more accurate when learning local and stationary features on graphs. Yu et al. [36] adopt GCN to enhance the representation of text features, combining the strengths of structural information with semantic information. Li et al. [11] utilize GCN in feature reasoning to generate image features with semantic relationship between image regions, as response of each node is computed based on its neighbour nodes. Yao et al. [37] utilize a GCN-based encoder to refine the representations of each region proposed on objects with the learned region-level features. Yang et al. [38] delve into the details of improving semantic navigation using GCN by incorporating prior knowledge and updating it dynamically as the agent receives the current environment information. Wen et al. [39] design the dual semantic relation module through the graph attention network (GAT) [40], which aims to enhance the regional and global relations for more accurate visual and text representations. Liu et al. [41] design the graph structure for both the image and text to perform node-level and structure-level matching.

#### **III. PROPOSED METHOD**

In this section, we present the proposed Hierarchical Graph Alignment Network (HGAN) in detail. As shown in Fig. 2, the model consists of four parts: the image feature graph (Section III. A), the text feature graph (Section III. B), the multigranularity feature aggregation and rearrangement module (Section III. C) and the loss function (Section III. D).

### A. Image Feature Graph

Since there exist abundant semantic information in images, we try to learn image representations by jointly considering the global and local features. Different feature encoders are employed to obtain the global and local features respectively, and then a unified image representation is obtained by using concatenate operation to construct the image feature graph.

1) Global Image Representation: We use ResNet152 [42] as an encoder for extracting the global feature of images. It is a model trained on ImageNet [43] to accurately extract the pixellevel features of the image. We discard the last fully connected layer of the ResNet152 and perform the reshape operation to get the output features,  $G = \{g_1, g_2, ..., g_m\}, g_i \in \mathbb{R}^{D^0}$ , where  $D^0$  represents the dimension of each pixel, and m is the reshaped size of the feature map. Then we use a fully connected layer to project  $g_i$  into the D-dimensional embeddings:

$$V_G = W_g G + b_g, \tag{1}$$

where  $V_G$  is the global feature of the image.  $W_g \in R^{D \times D^0}$  represents the weight matrix with the bias  $b_g$ , and they are learnable parameters.

2) Local Image Representation: In order to enable a feature vector encoding a salient region, we use the bottomup attention [44]. Following the feature extraction method in [8], [11], we employ the Faster-RCNN [7] as an encoder to extract the region feature of images. It is a model pretrained on Visual Genomes DataSet [45] to accurately identify objects in the image. Specifically, the features of image regions are employed to represent the image. Therefore, the output of the image region feature encoder can be represented as  $L = \{l_1, l_2, ..., l_k\}, l_i \in \mathbb{R}^{D^0}$ , where  $D^0$  represents dimension of region feature, and k is the number of detected regions. For BERT

Image

A young boy kicking a soccer

ball on a green field.

Text



sort

MFR

GRU

Fig. 2. An illustration of the proposed HGAN model, which consists of four parts. (1) Image Feature Graph. The global and local features are extracted using ResNet152 and Faster-RCNN respectively, and then the image feature graph is constructed. (2) Text Feature Graph. The text feature is extracted using the BERT, and then the text feature graph is constructed. (3) Multi-granularity Feature Aggregation and Rearrangement Module. We design a multi-granularity feature aggregation and rearrangement module to aggregate global and local features by multi-granularity feature fusion. (4) Loss Function. The image and text features are optimized through three similarity functions  $S_1$ ,  $S_2$  and  $S_3$ .

MFA

MFAR module

each selected region i, we use the average pooling layer to get  $l_i$ . Then we use a fully connected layer to project  $l_i$  into the D-dimensional embeddings:

$$V_L = W_l L + b_l, \tag{2}$$

Text feature graph  $G_{i}$ 

where  $V_L$  is the local feature of the image.  $W_l \in R^{D \times D^0}$  represents the weight matrix, and  $b_l$  represents the bias.

3) Unified Image Representation: Previous works separately considered on the global and local features of the images, and the important semantic relationship between the local object and global context is ignored. Differently, we design a unified representation of the global and local features of the images, where coarse-grained and fine-grained information is fused. And the unified image representation is constructed in the form of graph, so as to use graph convolutional networks to optimize multi-granularity feature fusion, thus enhance it to learn more accurate image features. By using concatenate operation,  $V_G$  and  $V_L$  form a unified image representation  $V_U$ :

$$V_U = V_G ||V_L, \tag{3}$$

where  $\parallel$  denotes the concatenate operation. So we can get the region-pixel unified feature  $V_U = \{v_U^1, v_U^2, ..., v_U^{m+k}\}, v_U^i \in R^{D^U}$ . Then the relationship  $E_U$  between  $v_U^i$  and  $v_U^j$  is defined:

$$E_U(v_U^i, v_U^j) = v_U^i \odot v_U^j, \tag{4}$$

where  $\odot$  represents the element product. Finally, the features  $v_U^i$  and  $v_U^j$  are regarded as the nodes, and the relationship  $E_U$  between them is regarded as the edge to construct the image feature graph  $G_V = (V_U, E_U)$ .

# B. Text Feature Graph

The traditional text feature representation methods use the RNN-based models such as LSTM [46] or GRU [47], and treat the output of the network as the sentence feature. The language representation model BERT [6] uses the self-attention based transformer structure, which is accomplished at learning semantic relationships. The model has powerful feature extraction capabilities to generate deep bidirectional linguistic representations for word tokens. In this paper, we employ the BERT as the text encoder. First, the sentence is tokenized by WordPiece, and then the features of the word are extracted through the BERT model. Therefore, the output of the text encoder can be represented as  $S = \{s_1, s_2, ..., s_l\}, s_i \in \mathbb{R}^{D^1}$ , where  $D^1$  represents the dimension, and l is the maximum number of words in the sentence. Then we use a fully connection layer to project  $t_i$  into the D-dimensional embeddings:

$$T_S = W_s S + b_s,\tag{5}$$

where  $T_S$  is the feature of the texts.  $W_s \in \mathbb{R}^{D \times D^1}$  represents the weight matrix, and  $b_s$  represents the bias. The text feature can be represented as  $T_S = \{t_S^1, t_S^2, ..., t_S^l\}, t_S^i \in \mathbb{R}^{D^S}$ . Then in order to build the text feature graph, we define the relationship  $E_S$  between  $t_S^i$  and  $t_S^j$ :

$$E_S(t_S^i, t_S^j) = t_S^i \odot t_S^j, \tag{6}$$

where  $\odot$  represents the element product. Finally, the features  $t_S^i$  and  $t_S^j$  are regarded as the nodes, and the relationship  $E_S$  between them is regarded as the edge to construct the text feature graph  $G_T = (T_S, E_S)$ .

# *C. Multi-granularity Feature Aggregation and Rearrangement Module*

In this section, we introduce the designed multi-granularity feature aggregation and rearrangement (MFAR) module, which filters out the noisy parts of the global-local semantic alignment and retains the useful parts to accomplish multigranularity feature fusion. For image and text feature graphs, we both use the MFAR module to conduct feature aggregation and rearrangement. Here the image modality is taken as an example to illustrate the details of MFAR module.

1) Multi-granularity Feature Aggregation Module: The node feature representation is optimized in this module. Considering that the nodes may contain redundant information, we employ the attention mechanism to learn the correlation between nodes and selectively aggregate neighboring nodes based on the correlation to obtain optimized node features, which filter worthless information and keep effective information. Therefore, semantic interaction between global and local information is realized throughout this process.

Given the image feature graph  $G_V = (V_U, E_U)$ ,  $V_U = \{v_U^1, v_U^2, ..., v_U^{m+k}\}, v_U^i \in R^{D^U}$ , where  $v_U^i$  is the node feature, and  $E_U$  is the relationship of nodes.  $D^U$  represents the dimension of image feature. To obtain sufficient expressive power, we parameterize node features through the weight matrix, and then, use the self-attention mechanism to calculate the attention coefficient for each node:

$$e_{ij} = \frac{W_q v_U^i \odot W_k v_U^j}{\sqrt{D^U}},\tag{7}$$

where  $e_{ij}$  indicates the importance of node j to node i,  $W_q$  and  $W_k$  are learnable weight matrices.  $\odot$  denotes the element product. To make the weight coefficients are comparable between different nodes, all choices of node j are regularized using the softmax function:

$$\alpha_{ij} = \operatorname{softmax}_{j} (e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_{i}} \exp(e_{ik})}, \quad (8)$$

where  $N_i$  is the set of neighbor nodes of node *i*. At the same time, in order to reduce the number of parameters, we introduce the multi-head attention mechanism to calculate the attention coefficients, which is faster and more space-saving.

$$MultiHead\left(v_{U}^{i}, v_{U}^{j}\right) = W_{o} \|_{h=1}^{H}(head_{1}, ..., head_{h})$$
(9)

$$\operatorname{head}_{h} = \operatorname{Softmax}\left(\frac{W_{q}^{h}v_{U}^{i} \odot W_{k}^{h}v_{U}^{i}}{\sqrt{d}}\right)W_{v}^{h}v_{U}^{j}, \qquad (10)$$

where  $\parallel$  represents the concatenate operation, H stands for the number of parallel attention layers, and  $d = \frac{D}{H}$ .  $W_o \in R^{D \times D}$ ,  $W_q \in R^{D \times d}$ ,  $W_k \in R^{D \times d}$  and  $W_v \in R^{D \times d}$  are learnable weight matrices. Then, the output of the multi-granularity feature aggregation module is obtained by the nonlinear activation function:

$$v_{U}^{i'} = BN\left(\operatorname{ReLU}\left(\sum_{j \in N_{i}} MultiHead\left(v_{U}^{i}, v_{U}^{j}\right)\right)\right), \quad (11)$$

where BN represents the batch normalization which can be used to speed up training, ReLU is the activation function, and  $\mathcal{N}_i$  is the set of neighbor nodes of node *i*. For the convenience of description, we simplify the output as  $v^{i'}$ .

2) Multi-granularity Feature Rearrangement Module: This module is used to better fuse multi-granularity semantics and improve multi-granularity feature representation. Given that each channel component of the node feature can be used as a representation of the node at multiple saliencies, in this module, firstly, the feature vector of the node is fine-tuned by rearranging the channel values to align the feature components of different saliencies in each node. Then, adaptive weights are learned for each node to fuse the component under varied saliency, thus optimized multi-granularity semantic features are obtained. For the output features  $v^{i'}$  of the multigranularity feature aggregation module, most of them use average pooling to aggregate features, ignoring the information interaction between global and local features. The average pooling method obtains v by averaging the N feature vectors as  $v = \frac{1}{N} \sum_{i=1}^{N} v^{i'}$ . Therefore the value at channel k is calculated by:

$$v_k = \frac{1}{N} \sum_{i=1}^{N} v_k^{i'},$$
 (12)

where k = 1, ..., K is the channel number of v.

We design the multi-granularity feature rearrangement to refine the feature vectors. First, we sort the feature vectors and then learn a rearrangement coefficient for each vector, taking the weighted sum of the vectors as the output:

$$v_k = \frac{1}{N} \sum_{i=1}^{N} \theta_i max_k(v^{i'}), \forall k.$$
(13)

$$\theta_i = f[i, N], i = 1, ..., N.$$
 (14)

f is the rearrangement coefficient generator.  $\theta_i$  represents the rearrangement coefficient of the *i*-th node, which satisfies  $\sum_{i=1}^{N} \theta_i = 1$ . Therefore, the final image feature of the model can be expressed as  $V = \{v_1, v_2, ..., v_k\}, v_i \in \mathbb{R}^{D^V}$ , where  $D^V$  represents the dimension of the image feature.

Specifically, the multi-granularity feature rearrangement module consists of two components, a trigonometric functionbased position encoder for generating the position indices and a bidirectional gated recurrent unit (BiGRU) based sequence model for generating rearrangement coefficients. To make full use of the prior information contained in the position indices, following [48], the trigonometric positional encoding strategy is employed to vectorize the positional indices:

$$p_i(i,2j) = \sin\left(\frac{i}{10000^{2j/d_p}}\right),$$
 (15)

$$p_i(i, 2j+1) = \cos\left(\frac{i}{10000^{2j/d_p}}\right),$$
 (16)

where  $p_i$  is the position vector, and  $d_p$  is the dimension of the position vector. After converting the position indices into vector representations, the sequence model is adopted to generate the rearrangement coefficients:

#### TABLE I

THE INTRODUCTION OF THE BASELINE MODELS, INCLUDING THE YEAR OF PUBLICATION, ADVANTAGES, DISADVANTAGES, AND MAIN CONTENT.

			1	
model	year	pros	cons	main content
VSE++ [20]	2017	Introduce the hard negatives into the loss function.	Ignore local alignment of images and text.	This method introduces the hard negatives into the common loss functions used for image-text retrieval. The improved loss can better guide the more powerful image and text encoders.
SCAN [8]	2018	Introduce object detection method into ITR task.	Ignore interactions between image and text modals.	This is the first method to introduce object detection method into ITR task, and point out that there is an underlying alignment relationship between image regions and sentence fragments.
CAMP [49]	2019	Consider fine-grained image-text in- teractions and adaptively control the cross modal information flow.	Ignore abstract objects, such as de- scriptions of certain behaviors.	This method considers comprehensive and fine-grained image-text in- teractions, and handles negative pairs and unrelated information with an adaptive gating module.
RDAN [50]	2019	Introduce the multi-level image-text alignments into cross-modal retrieval.	Ignore the effects of intra-modal rela- tionships.	This is a relation-wise dual attention model, which encodes the local and the global correlations between regions and words by training the image-text retrieval network.
VSRN [11]	2019	Focus on relational reasoning in images firstly.	Only consider reasoning within the image modality, ignoring text modal- ities.	This method uses the gate and memory mechanism to perform global semantic reasoning on the relationship representation and gradually generate the image feature.
MMCA [51]	2020	Explore intra-modality and the inter- modality relationship through a cross- attention mechanism.	Lack of research on semantic corre- spondence and semantic association.	This method designs a novel cross-attention mechanism, which exploit the intra-modality and the inter-modality relationship to enhance and complement each other for image-text retrieval.
CAAN [52]	2020	Propose a context-aware attention network to selectively focuses on crit- ical local fragments.	Ignore high-level semantic informa- tion between modalities.	This method selectively focuses on pivotal local features by aggregating the global context to discover latent semantic relations.
IMRAN [53]	2020	Explore fine-grained correspondences using the attention mechanism	Ignore the alignment of phrases and image regions.	This is a fine-grained matching method, which introduces an iterative matching strategy with recurrent attention memory to explore the fine-grained alignment progressively.
GSMN [41]	2020	Derive fine-grained image-text asso- ciations through node-level matching	Extra work to build the visual graph and textual graph.	This is a graph matching method, which models object, relation and attribute as a structured phrase to learn correspondence of object, relation, attribute and structured phrase separately.
CAMERA [21]	2020	Design a context-aware multi-view summarization network to meet the multi-view description challenge	Summarize text information from multiple views, easily mixed with noise.	This is a state-of-the-art image-text retrieval method on the Flickr30K dataset in 2020, which summarizes context-enhanced image region information from multiple views.
Meta-SPN [54]	2021	propose a meta self-paced network to accelerate model training.	Just a training acceleration algorithm based on existing models.	This method designs a meta self-paced network, which automatically learns the weight coefficients from data for image-text retrieval.
SMFEA [55]	2021	Build a tree of images and texts to obtain the structured semantic representation.	The constructed tree is coarse-grained and cannot distinguish data with high similarity.	This is a structured tree based image-text retrieval model, which models the relations of the image and text fragments by constructing structured tree encoders.
SHAN [10]	2021	Realize image-text matching through multi-step cross-modal inference.	Ignore the shared semantic concepts that potentially correlated the different modalities.	This is a hierarchical alignment model, which decomposes image-text retrieval into multi-step cross-modal reasoning processes.
SGRAF [27]	2021	Making full use of alignment infor- mation through graph inference to in- fer more accurate match scores.	Change the similarity from value to vector, the retrieval process will take more time.	This method first applies the vector-based similarity representations to characterize the local and global features, which relies on the GCN to infer the relation-aware similarity.
CGAM [56]	2021	consider the shared semantic concepts to enhance the discriminative power of the common space.	Ignore syntactical alignment and other research on multi-granularity.	This method builds semantic-embedded graph for each modality, and smooths the discrepancy through cross-graph attention module to obtain shared semantic-enhanced features.
CSCC [57]	2021	Considering the syntactical corre- spondence through the cross-level consistency for Image-text matching.	Ignore the effect of global context information on retrieval results.	This is a state-of-the-art image-text retrieval method on the MS-COCO 1K and 5K datasets in 2021, which introduces a conceptual-level image- text alignment scheme to exploring the fine-grained correspondence.

$$\{\theta_i\}_{i=1}^{\mathbf{N}} = MLP\left(\text{BiGRU}\left(\{p_i\}_{i=1}^{\mathbf{N}}\right)\right).$$
(17)

Then the aggregated representation is obtained by the weighted sum, as shown in Eq.(13). Finally, we can obtain the multigranularity image and text representations through the multilayer MFAR module.

## D. Loss Function

To align the global and local information simultaneously, we set three-level cosine similarity functions in the proposed model.  $S_1$  calculates the similarity of image region features  $V_L$  and text features  $T_S$  and  $S_2$  computes the similarity of image unified features  $V_U$  and text features  $T_S$ . The similarity of image and text features V and T refined by the MFAR module are calculated by  $S_3$ . S is the final similarity of image and text in the multi-granularity shared space, that is, the sum of the above three similarity functions:

$$S(I,T) = S_1(V_L,T_S) + S_2(V_U,T_S) + S_3(V,T), \quad (18)$$

where (I, T) is the matched positive pair of image and text. And the formulations for the three cosine similarity functions are specified as follows:

$$S_1(V_L, T_S) = \frac{V_L \cdot T_S}{||V_L|| \times ||T_S||},$$
(19)

$$S_2(V_U, T_S) = \frac{V_U \cdot T_S}{||V_U|| \times ||T_S||},$$
(20)

$$S_3(V,T) = \frac{V \cdot T}{||V|| \times ||T||}.$$
(21)

Then, a bidirectional hinge-based triplet ranking loss [8], [20] is adopted to make the matched image-text pairs have higher similarity scores than unmatched ones.

$$L = [d + S(I', T) - S(I, T)]_{+} + [d + S(I, T') - S(I, T)]_{+},$$
(22)

where d denotes the margin parameter, and  $[x]_+ \equiv \max(x, 0)$ .  $I' = \arg \max_{X \neq I} S(X, T)$  and  $T' = \arg \max_{Y \neq T} S(I, Y)$ denote the hardest negatives corresponding to the positive pair (I, T).

COMPARISON WITH THE BASELINE MODELS OF IMAGE-TO-TEXT RETRIEVAL AND TEXT-TO-IMAGE RETRIEVAL ON THE 1K AND 5K TEST SET OF MS-COCO dataset. The bold indicates the optimal results, and the underline indicates the suboptimal results. "-" denotes the results are not provided.

	MS-COCO 1K						MS-COCO 5K							
Methods	Ir	nage-to-7	Text	T	ext-to-Im	age		Ir	nage-to-T	Text	T	ext-to-Im	age	
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
VSE++ [20]	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
SCAN [8]	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
CAMP [49]	72.3	94.8	98.3	58.5	87.9	95.0	506.8	50.1	82.1	89.7	39.0	68.9	80.2	410.0
RDAN [50]	74.6	96.2	98.7	61.6	89.2	94.7	515.0	-	-	-	-	-	-	-
VSRN [11]	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
MMCA [51]	74.8	95.6	97.7	61.6	89.8	95.2	514.7	54.0	82.5	90.7	38.7	69.7	80.8	416.4
CAAN [52]	75.5	95.4	98.5	61.3	89.7	95.2	515.6	52.5	83.3	90.9	41.2	70.3	82.9	421.1
IMRAM [53]	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
GSMN [41]	78.4	96.4	98.6	63.3	90.1	95.7	522.5	-	-	-	-	-	-	-
CAMERA [21]	77.5	96.3	98.8	63.4	90.9	95.8	522.7	55.1	82.9	91.2	40.5	71.7	82.5	423.9
Meta-SPN [54]	74.4	95.0	98.3	58.6	87.6	94.3	508.2	51.0	81.1	89.4	37.5	66.7	77.5	403.2
SMFEA [55]	75.1	95.4	98.3	62.5	90.1	96.2	517.6	54.2	-	89.9	41.9	-	83.7	425.3
SHAN [10]	76.8	96.3	98.7	62.6	89.6	95.8	519.8	-	-	-	-	-	-	-
SGRAF [27]	79.6	96.2	98.5	63.2	90.7	96.1	524.3	57.8	-	91.6	41.9	-	81.3	-
CGAM [56]	78.9	97.5	98.8	65.7	90.2	96.6	527.7	-	-	-	-	-	-	-
CSCC [57]	78.8	96.1	99.0	66.6	92.5	<u>96.4</u>	529.4	55.6	83.6	91.2	40.8	73.2	<u>84.3</u>	<u>428.7</u>
HGAN (ours)	81.1	<u>96.9</u>	99.0	67.4	92.2	96.6	533.2	60.0	85.8	92.8	45.4	75.3	85.1	444.4

### IV. EXPERIMENT

In this section, we evaluate the proposed HGAN model on two benchmark datasets. First, the dataset and evaluation metrics and the implementation details are introduced. Then, the effectiveness of our model is proved by the performance comparison experiments and the ablation studies. Finally, the proposed HGAN model is qualitatively analyzed through visualization experiments.

#### A. Datasets and Evaluation Metric

We employ two commonly-used image-text retrieval datasets, MS-COCO (1K and 5K) [58] and Flickr30K [59], to evaluate our model.

The MS-COCO is a large-scale benchmark dataset used for image recognition, segmentation and retrieval. It is composed of 123,287 images, each with 5 corresponding captions. Following the experiment settings in [8], [11], we evaluate our method on 1K and 5K test images respectively. Specifically, the train, validation and test splits contain 113,287, 5000 and 5000 images. The 1K branch refers to the results are reported by averaging from 5 folds of 1K test images and the 5K branch refers to testing on the full 5K test images directly.

The Flickr30K dataset contains 31,783 images. Each image is paired with 5 corresponding captions. Following the split method in [20] about the Flickr30K dataset, we evaluate the performance of our model using 29,000 images for training, 1,000 images for validation, and the remaining 1,000 ones for testing.

For both image-to-text retrieval and text-to-image retrieval tasks, we report the results with the standard metrics, including R@K (Recall@K, K=1, 5, 10) and Rsum. R@K is defined as the proportion of correct image or text being retrieved among top K results, and Rsum is the sum of six R@K value to evaluate performance comprehensively.

TABLE III Comparison with the baseline models of Image-to-Text retrieval and Text-to-Image retrieval on the Flickr30K dataset.

	Flickr30K									
Methods	Im	age-to-'	Text	Te						
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum			
VSE++ [20]	52.9	80.5	87.2	39.6	70.1	79.5	407.9			
SCAN [8]	67.4	90.3	95.8	48.6	77.7	85.2	465.0			
CAMP [49]	68.1	89.7	95.2	51.5	77.1	85.3	466.9			
RDAN [50]	68.1	91.0	95.9	54.1	80.9	87.2	477.2			
VSRN [11]	71.3	90.6	96.0	54.7	81.8	88.2	482.6			
MMCA [51]	74.2	92.8	96.4	54.8	81.4	87.8	487.4			
CAAN [52]	70.1	91.6	97.2	52.8	79.0	87.9	478.6			
IMRAM [53]	74.1	93.0	96.6	53.9	79.4	87.2	484.2			
GSMN [41]	76.4	94.3	97.3	57.4	82.3	89.0	496.8			
CAMERA [21]	78.0	<u>95.1</u>	<u>97.9</u>	60.3	85.9	91.7	<u>508.9</u>			
Meta-SPN [54]	72.5	93.2	96.7	53.3	80.2	87.2	483.1			
SMFEA [55]	73.7	92.5	96.1	54.7	82.1	88.4	487.5			
SHAN [10]	74.6	93.5	96.9	55.3	81.3	88.4	490.0			
SGRAF [27]	77.8	94.1	97.4	58.5	83.0	88.8	499.6			
CGAM [56]	78.7	94.5	97.9	58.2	83.6	89.6	502.5			
CSCC [57]	72.7	93.4	96.5	<u>61.2</u>	<u>86.7</u>	91.5	502.0			
HGAN(ours)	80.3	96.5	98.3	62.3	87.8	93.1	518.3			

# B. Comparable Methods

In order to prove the effectiveness of the proposed HGAN method, we choose the models shown in Table I as the baseline models. In the Table I, we introduce the main works in the field of the image-text retrieval, including the publication year, the pros and cons, and the main content of each baseline model.

#### C. Implementation Details

In this section, we describe the software and hardware configuration of the experiments in detail. Our model is evaluated in pytorch-1.7.1 with python wrapper and a machine with Intel Xeon Gold 6226R CPU, 64GB RAM, 1T SSD and NVIDIA Tesla A100 GPU. Specifically, for dataset Flickr30K and MS-COCO, the model is trained for 12 and 20 epochs with the adaptive moment estimation optimizer (Adam) [60],

respectively. The batchsize is set to 256 and 320 for Flickr30K and MS-COCO datasets, respectively. The warmup is set to 0.1 and the learning rate is set to 0.0002 with a decay rate of 0.1 every 6 epochs. For images, the feature dimension  $D^0$  is set to 2048 for both global and local features. The basic version of the pre-trained BERT [6] is used to extract text features, which includes 12 layers, 12 heads, 768 hidden units, and 110M parameters, to get the text embeddings with  $D^1 = 768$ . The image-text shared space dimension D is set to 1024. In our model, we set the number of MFAR layers to 2 and 4 for the Flickr30K and MS-COCO datasets, respectively.

#### D. Performance Comparison

In this section, we show the experimental results on the MS-COCO (1K and 5K) and Flickr30K datasets in Table II and Table III. For the sake of fairness, we directly cite the simulation results of the baseline models in their corresponding original papers. The bold indicates the optimal results, and the underline indicates the suboptimal results. Overall, our model achieves the state-of-the-art retrieval results on the Flickr30K and MS-COCO 1K and 5K datasets.

Table II shows the performance of each model on the MS-COCO dataset. CSCC [57] has the optimal performance on the 1K test set, probably because it considers the syntactic alignment in addition to the fine-grained alignment, which is similar to our model. SGRAF [27] has the best performance on the 5K test set, which utilizes a vector-based similarity representation method to deduce more accurate matching score of images and texts through the fine-grained alignment. Overall, our HGAN model shows the best performance according to the Rsum metric. As for the 1K test set of MS-COCO, our model reaches 81.1% R@1 score and 67.4% R@1 score on imageto-text retrieval and text-to-image retrieval respectively, both outperforming other state-of-the-art methods. For the 5K test set, image-to-text retrieval and text-to-image retrieval achieve the best R@1 of 60.0% and 45.4% respectively, with 2.2% and 3.5% improvement over the SGRAF model. The performance of R@5 and R@10 can also be seen from the Table II, both of which achieve the optimal and suboptimal results.

Table III shows the performance of each model on the Flickr30K dataset. CAMERA [21] has the best performance on image-to-text retrieval, which aggregates context-enhanced visual information from multiple views of the image. CSCC [57] has the best performance on the text-to-image retrieval, which simultaneously considers the semantic information of the concept and syntactic. It is obvious that our HGAN model outperforms existing models by a large margin, achieving 80.3%, 96.5% and 98.3% for R@1, R@5 and R@10 on image-to-text retrieval. And the performance on text-to-image retrieval is 62.3%, 87.8% and 93.1% for R@1, R@5 and R@10 respectively. Compared to the CAMERA method, our model has a significant improvement in image-to-text retrieval and text-to-image retrieval tasks (by 2.3% and 2% on R@1). In summary, our HGAN model outperforms other state-of-the-art models on the Flickr30K dataset.

EFFECTIVENESS ANALYSIS OF THE MFAR MODULE AND UNIFIED FATURES ON THE MS-COCO 1K AND FLICKR30K DATASETS. "GIE" DENOTES THE GLOBAL IMAGE EMBEDDING AND "LIE" DENOTES THE LOCAL IMAGE EMBEDDING. "MFA" IS THE MULTI-GRANULARITY FEATURE AGGREGATION MODULE. "MFAR" IS THE MULTI-GRANULARITY FEATURE AGGREGATION AND REARRANGEMENT MODULE.

	MS-COCO 1K								
Methods	Ir	nage-to-T	Text	Text-to-Image					
	R@1	R@5	R@10	R@1	R@5	R@10			
MFA+GIE	73.4	94.5	97.9	60.4	89.7	95.6			
MFAR+GIE	75.4	94.9	98.3	62.8	89.9	95.4			
MFA+LIE	76.5	95.4	98.4	63.2	90.0	95.1			
MFAR+LIE	80.0	96.6	98.9	66.8	91.8	96.6			
MFA+GIE+LIE	78.3	95.7	98.6	64.5	90.9	95.9			
MFAR+GIE+LIE	81.1	96.9	99.0	67.4	92.2	96.6			
			Flick	r30K					
MFA+GIE	75.4	92.6	96.3	57.4	84.1	90.7			
MFAR+GIE	76.1	92.7	96.9	57.4	84.9	90.9			
MFA+LIE	77.8	94.4	97.5	60.8	87.0	92.6			
MFAR+LIE	79.5	95.2	97.9	61.9	87.4	92.8			
MFA+GIE+LIE	78.6	94.2	97.3	61.4	87.2	92.8			
MFAR+GIE+LIE	80.3	96.5	98.3	62.3	87.8	93.1			

TABLE V EFFECTIVENESS ANALYSIS OF THE HIERARCHY ON THE MS-COCO 1K AND FLICKR30K DATASETS.

	MS-COCO 1K							
Methods	In	nage-to-T	ext	Text-to-Image				
	R@1	R@5	R@10	R@1	R@5	R@10		
$S_3$	78.8	96.2	98.8	65.1	92.0	96.6		
$S_1 + S_3$	79.8	96.5	98.9	66.4	91.9	96.5		
$S_2 + S_3$	79.7	96.5	98.8	66.2	92.5	96.4		
$S_1 + S_2 + S_3$	81.1	96.9	99.0	67.4	92.2	96.6		
	Flick			r30K				
$S_3$	79	95.4	97.6	61.1	87.0	92.7		
$S_1 + S_3$	79.9	96.6	98.5	62.0	87.8	93.0		
$S_2 + S_3$	79.5	96.1	98.1	61.4	87.1	92.4		
$S_1 + S_2 + S_3$	80.3	96.5	98.3	62.3	87.8	93.1		

#### E. Analysis of Model

In this section, the ablation experiments are performed on the MS-COCO 1K and Flickr30K datasets. We analyze the effectiveness of the MFAR module and the unified features, the influence of the different similarity functions and the parameters of the model, respectively.

1) Effect of the MFAR module. In Table IV, the MFAR module is the proposed multi-granularity feature aggregation and rearrangement. The MFA module denotes the multi-granularity feature aggregation module and the MFR module is the multi-granularity feature rearrangement module. To demonstrate the effectiveness of the MFAR module, we disable the MFR module for performance testing to explore the impact of the designed modules. We can observe that our MFAR module has always outperformed the MFA-based model, which brings about a 2% performance promotion on two benchmark datasets.

2) Effect of the unified features. In Table IV, GIE denotes the global image embedding and LIE denotes the local image embedding. GIE+LIE is the unified image representation. We can observe that global matching using GIE module is not as effective as fine-grained matching using LIE module, and further, using the unified image feature including GIE and LIE module in our HGAN model has the optimal performance. Specifically, comparing the results of 'MFAR+GIE' and 'MFAR+LIE', it can be found that using only local features to construct image features is much better than using only global features, indicating that local features describe more image details and can generate more discriminative features. Furthermore, the comparison of the results of 'MFAR+LIE' and 'MFAR+GIE+LIE' shows that better results can be achieved by using both global and local features, which means that the global features supplement the contextual information lacked by the local features, allowing the reconstructed features to better represent the integral image and achieve improved retrieval results.

3) Effect of hierarchy. The hierarchy in our proposed method is directly reflected by multiple levels of similarity. Specifically, the three similarity functions  $S_1$ ,  $S_2$ , and  $S_3$ corresponding to three levels of fine-grained feature, unified feature and multi-granularity feature, respectively. In Table V,  $S_1$  represents the similarity function between the image local feature and text feature, and  $S_2$  works on the image unified feature and text feature.  $S_3$  cannot be removed in our HGAN model, which represents the similarity function between image and text after optimization by MFAR module. The bold font indicates the optimal results. We can find that the application of multi-level similarity function achieves about 1.5% improvement on R@1 for image-to-text retrieval and text-to-image retrieval. In terms of performance improvement,  $S_1$  is more effective compared to  $S_2$ , since  $S_1$  takes into account the alignment of local features.



Fig. 3. The results of Recall@1 with different initial learning rate and MFAR module layers.



Fig. 4. The results of Recall@1 with different batchsize.

4) Effect of the parameters. In this section, we analyze the impact of model parameters on performance, including the number of layers of the MFAR module, the learning rate, and the batchsize. In Fig. 3, *M* represents the number of layers of the MFAR module, and R denotes the initial learning rate. First, we can see that for the MS-COCO dataset, the performance is improved as the number of layers of the MFAR module increases, and it has the best performance when M = 4. Due to equipment limitations, we have no way to conduct experiments with M = 8, but we think that M = 4has achieved the ideal performance. For the Flickr30K dataset, M = 2 works best because the two-layer MFAR module is sufficient to extract the information contained in the dataset. Also, setting an initial learning rate of 0.0002 is the most appropriate for several datasets of our model. Furthermore, Fig. 4 shows the optimal batchsize on several benchmark dataset. The best batchsize value is 320 for the MS-COCO dataset, and for the Flickr30K dataset, the optimal batchsize is 256.



Fig. 5. The values of Rsum when varying the number of epochs.



Fig. 6. The variation of loss values with different iteration numbers.

## F. Analysis of Training Process

In this section, we analyze the training process of the proposed HGAN model. Firstly, in order to measure the computational efficiency of our model, we consider the computational complexity of the model (measured by FLOPs) and the time complexity (measured by the number of parameters), which are 36.81G and 211.29M respectively. Compared to the trillions of parameters of large-scale pre-trained models in the image-text cross-modal domain, our approach achieves reasonable calculation consumption.

Subsequently, we analyze the change of Rsum values (the sum of recall value) with epoch and the change of loss values with iteration. Fig. 5 records the values of Rsum when varying the number of epochs on the MS-COCO 1K and Flickr30K test sets. The maximum value of the epoch is set to 20 and 12 for the MS-COCO 1K and Flickr30k datasets, respectively. For MS-COCO 1K, when the epoch reaches about 10, Rsum reaches the maximum and then stabilizes. For Flickr30k, the model has basically converged when epoch is 6 because it contains less image and text. Fig. 6 records the variation of

Query	HGAN	HGAN w/o MFR
	<ol> <li>Boats sitting around the side of a lake by a tree. T</li> <li>A group of boats sitting together with no one around. T</li> <li>A small marina with boats docked there. T</li> <li>Some boats parked in the water at a dock. T</li> <li>A blue boat docked on a green lush shore. T</li> </ol>	<ol> <li>Boats sitting around the side of a lake by a tree. T</li> <li>A group of boats sitting together with no one around. T</li> <li>A small marina with boats docked there. T</li> <li>A harbor filled with boats floating on water. F</li> <li>Some boats parked in the water at a dock. T</li> </ol>
X M I K	<ol> <li>Three adult and two child skiers posing on a slope. T</li> <li>Family posing on the ski slopes wearing skis. T</li> <li>A family poses for a photo while skiing on a snowy mountainside. T</li> <li>A family of snow skiers lined up for a picture before their run. T</li> <li>A group of young and old are skiing on the snow. T</li> </ol>	<ol> <li>Family posing on the ski slopes wearing skis. T</li> <li>A family poses for a photo while skiing on a snowy mountainside. T</li> <li>Four skiers ready to ski down a snowy mountain. F</li> <li>Three adult and two child skiers posing on a slope. T</li> <li>A group of young and old are skiing on the snow. T</li> </ol>
-	<ol> <li>Boats are traveling in the large open water. T</li> <li>A large body of water with small boats floating on top of it. T</li> <li>A cruise ship travelling out of an expansive harbor. T</li> <li>A boat in the distance on a clear lake. F</li> <li>There is a boat going across the waterway. T</li> </ol>	<ol> <li>A boat in the distance on a clear lake. F</li> <li>A large body of water with small boats floating on top of it. T</li> <li>Boats are traveling in the large open water. T</li> <li>A cruise ship travelling out of an expansive harbor. T</li> <li>A boat sailing on top of a body of water. F</li> </ol>

Fig. 7. The qualitative results of image-to-text retrieval on the MS-COCO 1K dataset. The top-5 retrieval results are shown for each query image. The green "T" denotes the correct sentences and the red "F" indicates the wrong sentences (best viewed in color).



Fig. 8. The qualitative results of text-to-image retrieval on the MS-COCO 1K dataset. The top-3 retrieval results are shown for each query text. The correct images are highlighted in green boxes, and the wrong images are highlighted in red boxes (best viewed in color).

loss values with different iteration numbers on the MS-COCO 1K and Flickr30K test sets. It can be seen that our model can converge to a satisfactory value on both the MS-COCO 1K and Flickr30K, and there is no underfitting and overfitting. Comparing the datasets MS-COCO 1K and Flickr30K, MS-COCO 1K needs more epochs of training because of its large amount of data and complex information. Nonetheless, the proposed model has better performance in MS-COCO 1k dataset because of the richer relations contained in.

#### G. Visualization of Retrieval Results

In this section, we discuss the qualitative results of the proposed HGAN model. The retrieval results of two models are visualized, including the proposed HGAN model and the HGAN w/o MFR model. HGAN w/o MFR means that we disabled the multi-granularity feature rearrangement (MFR) module of the HGAN model.

In Fig. 7, we show the top 5 retrieval results for three query images. The correct matches are marked with a green "T", the wrong matches are marked with a red "F", and the whole incorrectly matched sentences are marked in red. The sentences on the left are the retrieval results of the proposed

HGAN model, and the sentences on the right are the retrieval results of the HGAN model without MFR. We can observe that the HGAN model can retrieve the correct matching sentence in most cases, and the retrieval accuracy is significantly superior to the latter. In Fig. 8, the top 3 retrieval results for three query sentences are displayed. The correct retrieved images are highlighted in green boxes, while the wrong ones in red boxes. The upper column of images are the retrieval results of the proposed HGAN model, and the lower column of images are the retrieval results of the HGAN model without MFR. We can observe that our HGAN model can match the correct image in the top 3 results in Fig. 8. Besides, compared with the HGAN w/o MFR model, the image that matches the query sentence has a higher rank in our HGAN model.

We argue that there are two reasons for this phenomenon. First, the MFR module achieves multi-granularity feature denoising through feature rearrangement, which filters out the noisy parts and retain the dominate parts for the multigranularity alignment of object-context-fused information. By mining the multi-granularity semantic relationship of multimodal features, the retrieval performance is improved. Then, the multi-level similarity functions can measure the image and



Fig. 9. Visualization of word similarity weights. Each subplot shows the similarity (after normalization) between the image and each word in its GT text. (a) and (b) are the samples for which our model retrieves the correct result at top-1. (c) and (d) are samples for which the top-10 items retrieved by our model do not contain correct results, and the incorrect top-1 result given by HGAN is "A man buying some food at a food stand" for (c) and "An adult skier carries a child skier under their arm on the slopes" for (d), respectively.



Fig. 10. T-SNE visualization of the image feature (left) and the corresponding text feature (right) on a subset of the MS-COCO test dataset. Different colors represent the different classes of samples.

text similarity at different levels in the multi-granularity shared space, which further improves the performance of ITR in terms of hierarchical alignment.

#### H. Visualization of Features

In this section, we perform visual analyses of the features modeled by our HGAN model to better demonstrate the effect of our model.

First, we display the detailed matching relationships between images and words. Specifically, we calculate the scores of image-word matching for four sample pairs, as shown in Fig. 9, where (a) and (b) are the samples for which our model retrieves the correct result at top-1, (c) and (d) are samples for which the top-10 items retrieved by our model do not contain correct results. The incorrect top-1 result given by HGAN is 'A man buying some food at a food stand' for (c) and 'An adult skier carries a child skier under their arm on the slopes' for (d), respectively. In each subplot, the selected image is shown on the left, the corresponding GT text is shown on the right, and in the middle is the similarity between the image and each word in its GT text.

All four subplots reveal that our model successfully recognizes the objects with key semantics in the samples, such as "dog", "towl" in (a) and "women", "donuts" in (c). Besides, since our model relies heavily on the regions obtained by object detection in constructing the features of images, it has poor understanding of articles, prepositions and verbs. For example, "the" in (b) and "with" in (d) mistakenly have the maximum similarity with the image.

Observing (c), (d) with their corresponding incorrect top-1 retrieval results, respectively, we can find that they describe a very similar scene, but there are detail errors and focus deviations. For example, "woman is smelling donuts" in the picture of (c) is wrongly judged as "the man is buying food", and people in the background of the picture in (d) are mistakenly focused.

In conclusion, our model has an excellent ability to find objects and nouns containing important semantics in the sample to achieve cross-modal image-text matching. However, the poor understanding of articles, prepositions and verbs can lead to misunderstanding in semantic details, which needs further improvement.

Then, we conduct a T-SNE visualization experiment by using a subset of the MS-COCO test dataset as shown in Fig. 10. Concretely, we randomly chose samples from the MS-COCO test set with four class labels (aeroplane, boat, cat, and dog) and fed their data into our HGAN model to generate features. The high-dimensional features of the samples are transformed into two-dimensional by T-SNE and displayed as points in Fig. 10, where different colors used to differentiate their classes. For each of these four categories, a set of corresponding image and text samples labeled with  $(a) \sim (d)$ is displayed. In each subplot, we can find that the points of each color are aggregated in a single region, indicating that the model has learned the discriminative information belonging to different classes of samples. Comparing the points of same color between the left and right subplots, it is found that they appear in similar areas, indicating that the model achieves cross-modal semantic matching. In addition, the distribution of the yellow and green points is closer because the classes they belong to (cat and dog) are more similar. Moreover, some points appear in the wrong colour area, such as the red points (aeroplanes) that appear in the blue area (boats). We analyze that there are two reasons for this. One is that planes and ships often appear in similar scenes, such as the blue sky and the sea, which leads to the error of model judgment. Another reason is that the selected samples may contain multiple types of objects, so there will be some areas of color mixing in this experiment.

# V. CONCLUSION

In this paper, we propose a novel Hierarchical Graph Alignment Network (HGAN) for image-text retrieval, including the following advantages:

- We construct feature graphs for the image and text modalities respectively to capture more comprehensive multi-modal features, and establish a multi-granularity shared space with the designed Multi-granularity Feature Aggregation and Rearrangement (MFAR) module to achieve multi-granularity feature filtering and fusion.
- We establish hierarchical alignment across modalities for features of varying granularity using three-level similarity functions, which deeply explore the feature similarity in the multi-granularity shared space.
- Extensive experiments on the MS-COCO and Flickr30K datasets show that the proposed HGAN method outperforms the state-of-the-art models for the ITR task.

In the future, for a more comprehensive feasibility analysis of the model, we are ready to extend our model to more tasks, such as image captioning [3] and visual question answering [33], as well as more modalities, like video-text field [61].

#### ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant (Nos. 62071354, 62201424 and 61906156), China Postdoctoral Science Foundation Grant (No. 2017M620438), the Natural Science Foundation of Shaanxi Province (Nos. 2019ZDLGY03-03) and also supported by the ISN State Key Laboratory and High-Performance Computing Platform of Xidian University.

#### REFERENCES

- S. Qian, D. Xue, Q. Fang, and C. Xu, "Adaptive label-aware graph convolutional networks for cross-modal retrieval," *IEEE Transactions* on Multimedia, vol. 24, pp. 3520–3532, 2022.
- [2] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter- and intra-modality interactions," *IEEE Transactions on Multimedia*, vol. 23, pp. 3518–3529, 2021.
- [3] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE Transactions on Multimedia*, vol. 24, pp. 1775–1786, 2022.
- [4] Y. Wang, H. Yang, X. Bai, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "PFAN++: Bi-directional image-text retrieval with position focused attention network," *IEEE Transactions on Multimedia*, vol. 23, pp. 3362– 3376, 2021.
- [5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in Neural Information Processing Systems*, 2013.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.
- [7] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," *IEEE Transactions* on Pattern Analysis and Machine Intelligence, vol. 39, pp. 1137–1149, 2015.
- [8] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference* on Computer Vision, 2018, pp. 201–216.
- [9] J. Wehrmann, C. Kolling, and R. C. Barros, "Adaptive cross-modal embeddings for image-text alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12313– 12320.
- [10] Z. Ji, K. Chen, and H. Wang, "Step-wise hierarchical alignment network for image-text matching," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021.
- [11] K. Li, Y. Zhang, K. Li, Y. Li, and Y. R. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4653–4661.
- [12] J. I. Olszewska, "Snakes in trees: An explainable artificial intelligence approach for automatic object detection and recognition," in *Proceedings* of the 14th International Conference on Agents and Artificial Intelligence (ICAART), 2022, pp. 996–1002.
- [13] R. Du, D. Chang, A. K. Bhunia, J. Xie, Y.-Z. Song, Z. Ma, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *European Conference on Computer Vision*, 2020, pp. 153–168.
- [14] W. Wang, M. Yan, and C. Wu, "Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1705– 1714.
- [15] J. I. Olszewska, C. De Vleeschouwer, and B. Macq, "Multi-feature vector flow for active contour tracking," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 721– 724.
- [16] H. Wang, C. Wu, J. Chi, X. Yu, Q. Hu, and H. Wu, "Image superresolution using multi-granularity perception and pyramid attention networks," *Neurocomputing*, vol. 443, pp. 247–261, 2021.
- [17] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference* on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
- [18] T. Chen and J. Luo, "Expressing objects just like words: Recurrent visual embedding for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10583– 10590.
- [19] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *ArXiv*, vol. abs/1411.2539, 2014.
- [20] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improved visual-semantic embeddings," ArXiv, vol. abs/1707.05612, 2017.

- [21] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multiview summarization network for image-text matching," in *Proceedings* of the 28th ACM International Conference on Multimedia, 2020, pp. 1047–1055.
- [22] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15789–15798.
- [23] J. I. Olszewska and T. L. McCluskey, "Ontology-coupled active contours for dynamic video scene understanding," in 15th IEEE International Conference on Intelligent Engineering Systems, 2011, pp. 369–374.
- [24] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," ArXiv, vol. abs/1804.09539, 2018.
- [25] S. Zeng, C. Liu, J. Zhou, Y. Chen, A. Jiang, and H. Li, "Learning hierarchical semantic correspondences for cross-modal image-text retrieval," *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022.
- [26] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Transactions on Image Processing*, vol. 28, pp. 2008–2020, 2019.
- [27] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1218–1226.
- [28] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, pp. 3196–3209, 2020.
- [29] P. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 466–479, 2022.
- [30] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9959–9968.
- [31] J. Wang, J. Tang, and J. Luo, "Multimodal attention with image text spatial relationship for ocr-based image captioning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4337– 4345.
- [32] J. Song, P. Zeng, L. Gao, and H. T. Shen, "From pixels to objects: Cubic visual attention for visual question answering," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 906–912.
- [33] H. Jiang, I. Misra, M. Rohrbach, E. G. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10264–10273.
- [34] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. van den Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1960–1968.
- [35] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7746–7755.
- [36] J. Yu, Y. Lu, Z. Qin, Y. Liu, J. Tan, L. Guo, and W. Zhang, "Modeling text with graph convolutional network for cross-modal information retrieval," pp. 223–234, 2018.
- [37] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 684–699.
- [38] W. Yang, X. Wang, A. Farhadi, A. K. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," *ArXiv*, vol. abs/1810.06543, 2019.
- [39] K. Wen, X. Gu, and Q. Cheng, "Learning dual semantic relations with graph attention for image-text matching," *IEEE Transactions on Circuits* and Systems for Video Technology, vol. 31, pp. 2866–2879, 2021.
- [40] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio', and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.
- [41] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10918–10927.
- [42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, 2016, pp. 770–778.

- [43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [44] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.
- [45] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2016.
- [46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
- [47] K. Cho, B. van Merrienboer, Çaglar Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014), 2014.
- [48] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [49] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5763–5772.
- [50] Z. Hu, Y. Luo, J. Lin, Y. Yan, and J. Chen, "Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 789–795.
- [51] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10938–10947.
- [52] Q. Zhang, Z. Lei, Z. Zhang, and S. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3533–3542.
- [53] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, 2020, pp. 12 655–12 663.
- [54] J. Wei, X. Xu, Z. Wang, and G. Wang, "Meta self-paced learning for cross-modal matching," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3835–3843.
- [55] X. Ge, F. Chen, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Structured multimodal feature embedding and alignment for image-sentence retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5185–5193.
- [56] Y. He, X. Liu, Y. ming Cheung, S.-J. Peng, J. Yi, and W. Fan, "Crossgraph attention enhanced multi-modal correlation learning for finegrained image-text retrieval," in *Proceedings of the 44th International* ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1865–1869.
- [57] P. Zeng, L. Gao, X. Lyu, S. Jing, and J. Song, "Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching," in *Proceedings of the 29th ACM International Conference* on Multimedia, 2021, pp. 2205–2213.
- [58] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context." Springer, 2014, pp. 740–755.
- [59] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings* of the IEEE/CVF International Conference on Computer Vision, 2015, pp. 2641–2649.
- [60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [61] W. Wang, J. Gao, X. Yang, and C. Xu, "Learning coarse-to-fine graph neural networks for video-text retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 2386–2397, 2021.