Heterogeneous Interactive Learning Network for Unsupervised Cross-modal Retrieval

Yuanchao Zheng[1] [0000−0002−5369−2325] and Xiaowei Zhang(✉)[1] [0000−0003−4854−3736]

Qingdao University, Qingdao, China
zyc000204@163.com, xiaowei19870119@sina.com

Abstract. Cross-modal hashing has received a lot of attention because of its unique characteristic of low storage cost and high retrieval efficiency. However, these existing cross-modal retrieval approaches often fail to align effectively semantic information due to information asymmetry between image and text modality. To address this issue, we propose Heterogeneous Interactive Learning Network(HILN) for unsupervised cross-modal retrieval to alleviate the problem of the heterogeneous semantic gap. Specifically, we introduce a multi-head self-attention mechanism to capture the global dependencies of semantic features within the modality. Moreover, since the semantic relations among object entities from different modalities exist consistency, we perform heterogeneous feature fusion through the heterogeneous feature interaction module, especially through the cross attention in it to learn the interaction between different modal features. Finally, to further maintain semantic consistency, we introduce adversarial loss into network learning to generate more robust hash codes. Extensive experiments demonstrate that the proposed HILN improves the accuracy of $T \rightarrow I$ and $I \rightarrow T$ cross-modal retrieval tasks by 7.6% and 5.5% over the best competitor DGCPN on the NUS-WIDE dataset, respectively. Code is available at https://github.com/Z000204/HILN.

Keywords: Cross-modal hashing · Heterogeneous interactive · Adversarial loss.

1 Introduction

With the explosive growth of data, cross-modal hashing retrieval has attracted more and more attention. Cross-modal hashing (CMH) as a hot topic is to map data of different modalities to the common binary hash space for matching, which improves the efficiency of retrieval and storage consumption [19,7]. CMH is divided into unsupervised and supervised methods, depending on whether label information is used. At present, supervised hashing methods have achieved good performance [9,1,5] due to a large amount of hand-labeled prior knowledge. However, these methods based on supervised learning require a lot of manual annotations and are often not suitable for the real world. Recently, more and more attention has been paid to unsupervised cross-modal hashing [16,12,21,18],

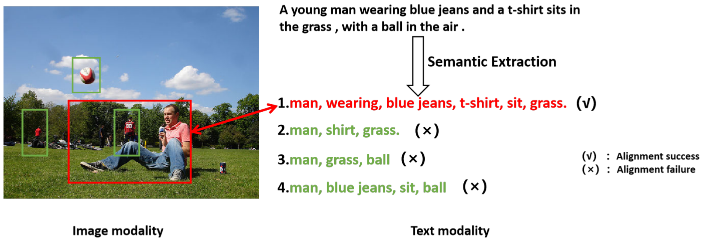

A young man wearing blue jeans and a t-shirt sits in the grass , with a ball in the air .
Semantic Extraction
1.man, wearing, blue jeans, t-shirt, sit, grass. (√)
2.man, shirt, grass. (×)
3.man, grass, ball (×)
4.man, blue jeans, sit, ball (×)
(√) : Alignment success
(×) : Alignment failure
Image modality
Text modality

Fig. 1. An exemplar of semantic consistency within image-text pair. The red font in text and red bounding box represents aligned semantic information between image and text modality, the green ones denote semantic alignment failure.

which can reduce the dependence on data annotations in the training process and has achieved significant progress.

However, due to the lack of artificial prior knowledge (label annotations), unsupervised cross-modal hashing methods face the problem of the heterogeneous semantic gap between image-text pairs, leading to the failure of cross-modal retrieval. How to model the semantic similarity between image modality and text modality becomes the key to improving the performance of cross-modal retrieval. Many recent unsupervised cross-modal hashing methods [16,12,21] proposed to align local semantic entities for building the semantic similarity. However, these methods do not take into account the information asymmetry between image-text pairs, where the local semantic entities between image and text modality often are unequal for cross-modal retrieval. As shown in Fig.1, there are many “person" in image modality to respond to the “man" in-text modality, while “A man sits in the grass" in image modality is corresponding to “A young man wearing blue jeans and a t-shirt sits in the grass" with the similarity of semantic relations. This means that it often does not unique or even unequal entities to associate local semantics between image-text pairs, while there is a consistent semantic relation for cross-modal retrieval.

Based on the above analysis, we propose Heterogeneous Interactive Learning Network(HILN) for unsupervised cross-modal retrieval from the view of the similarity of semantic relations. First, we introduce a multi-head self-attention mechanism to capture the global dependencies of semantic features within the modality for modeling semantic relations among object entities. Secondly, we perform heterogeneous feature fusion through the heterogeneous feature interaction module, especially through cross attention to learn the interaction between different modal features. Finally, to further maintain semantic consistency, we introduce adversarial loss into network learning to generate more robust hash codes. Our contributions can be summarized as follows:

- We propose a novel end-to-end cross-modal hashing method, named Heterogeneous Interactive Learning Network(HILN) for unsupervised cross-modal retrieval, which models global semantic consistency between image and text modality from the view of similarity of semantic relations to generate high-quality hash codes.

- We introduce a multi-head self-attention mechanism to capture the global dependencies of semantic features within a single modality for modeling semantic relations among object entities.
- HILN performs heterogeneous feature fusion through the heterogeneous feature interaction module, especially through the cross attention in it to learn the interaction between different modal features, to better align the semantic relation between modalities. Finally, to further maintain semantic consistency, we introduce adversarial loss into network learning to generate more robust hash codes.
- Extensive experiments demonstrate that the proposed HILN improves the accuracy of $T \rightarrow I$ and $I \rightarrow T$ cross-modal retrieval tasks by 7.6% and 5.5% over the best competitor DGCPN [22] on the NUS-WIDE dataset, respectively.

In the following sections, we present related work(Section 2), the proposed method(Section 3), analytical experiments(Section 4), and conclusion(Section 5).

2 Related Work

2.1 Cross-modal Hashing

Cross-modal hash retrieval methods can be broadly divided into two categories: supervised methods and unsupervised methods. The former is to explore semantic information from manual labels to bridge the semantic gap between different modalities for the generation of the hash codes, such as DCMH [9], DADH [1], AGAH [5].

Compared with the supervised methods, unsupervised methods mainly use co-occurrence information between images and texts to maximize the relationship between similar data between modalities and bridge the heterogeneous semantic gap between different modalities. Based on this, several unsupervised approaches have been proposed. DBRC [6] generates hash codes by reconstructing the original data from binary representation. To preserve the similarity between the original data, DJSRH [16] suggests a joint semantic similarity matrix and the use of hash codes to rebuild the similarity values of features. Next, JDSH [12] uses the characteristics of the data distribution to generate a better cross-modal similarity matrix to supervise the generation of hash codes based on DJSRH. At the same time, DSAH [21] builds on this by introducing semantic alignment loss to enhance the interaction between different modalities.

Although these approaches have achieved promising performance, they do not adequately align the global semantic relations between image and text modality, and even the performance imbalance of cross-modal retrieval between retrieving text from image and retrieving the image from text.

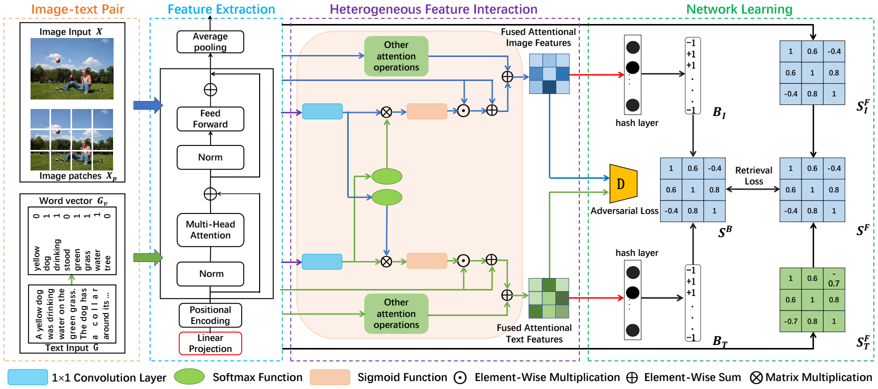


Fig. 2. The structure of our approach. It consists of an image-text pair, feature exaction, heterogeneous feature interaction, and Network learning. The purple arrows represent upsampling, and the red arrows represent dimensionality reduction and full connected layer.

2.2 Attention Mechanism

Neural network-based attention mechanism has achieved great success and is widely used in a variety of tasks, such as natural language processing [11,20] and computer vision [4,27]. The self-attention mechanism [17] is a variant of the attention mechanism, which can capture long-distance dependencies. A Structured Self-attentive Sentence Embedding [11] uses self-attention for sentence embedding to enhance the semantics of the sentence. Context-aware Self-Attention Networks for Natural Language Processing [20] contextualize the representations with global information. Vision Transformer [4] represents the use of a transformer with the self-attention mechanism for image, capturing the global semantic information of the image, enhancing the ability of image feature characterization. In this paper, we utilize a multi-head self-attention mechanism to capture the global dependencies of semantic features within the single modality and introduce channel attention, spatial attention, and cross-attention to enhance the context-aware similarity of semantic relations between image-text pairs.

2.3 Generative Adversarial Network

Generative Adversarial Network, capable of modeling the distribution of data, have now achieved great success. It has been widely used in cross-modal hash retrieval tasks [24,25,2]. Among them, SCH-GAN [25] makes the generative model learn to fit the correlation distribution of unlabeled data, and tries to select samples from the unlabeled data of one modality to get the query of another modality, so as to better reflect the data of the unlabeled data distributed. MGAH [24] utilizes the ability of generative adversarial network for unsupervised representation learning to fully mine the underlying popular relationship of multimedia data, thereby improving retrieval performance. SAALDH [2] utilizes self-attention mechanism to enhance hash expression and utilizes adversarial loss to further maintain hash code consistency.

In this paper, we introduce adversarial loss into network learning, which utilizes an adversarial loss to model the feature distribution of image and text data on the basis of obtaining different modal attention enhancements to generate more robust hash codes.

3 Proposed Approach

3.1 Probelm Definition

Assume that we have M image-text pairs which can be denoted as $O = \{X_i, G_i\}_{i=1}^{M}$. X_i and G_i represent the i-th image and the i-th text in the instance respectively. The structure of our method is shown in Figure 2. Given $\boldsymbol{F}^I$ and $\boldsymbol{F}^T$, our approach aims to learn two effective hash functions and simultaneously generates hash codes $\boldsymbol{B}_I \in \{-1, +1\}^{M \times K}$ and $\boldsymbol{B}_T \in \{-1, +1\}^{M \times K}$ for image and text modalities respectively, where K is the length of the hash codes.

3.2 Feature Extraction

Multi-head Self-attention. Like vision Transformer [4], we use the multi-head self-attention module to model the long-range relationships within the modalities.

Specifically, give the $1D$ embedding sequence $O \in \mathbb{R}^{L \times C}$ as input through the transformer-based encoder to learn feature representations, in which L is the length of the sequence, C is the numbers of the channel. Transformer encoder consists of J layers of multi-head self-attention(MSA) and Multilayer Perceptron(MLP) blocks. Layernorm is applied before every block and MLP. In each layer j, the query, key, and value computed from the input $O^{j-1} \in \mathbb{R}^{L \times C}$ with the corresponding weights are used as input for the self-attention as:

$$query = O^{j-1}\boldsymbol{W}_q, key = O^{j-1}\boldsymbol{W}_k, value = O^{j-1}\boldsymbol{W}_v, \tag{1}$$

where $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{C \times d}$ are the learnable weight parameter. Self-attention(SA) is then formulated as:

$$SA\left(O^{j-1}\right) = O^{j-1} + softmax\left(\frac{O^{j-1}\boldsymbol{W}_q\left(O\boldsymbol{W}_k\right)^{\mathsf{T}}}{\sqrt{d}}\right)\left(O^{j-1}\boldsymbol{W}_v\right). \tag{2}$$

MSA is an extension of SA in which contains n separate SA operations and projects their concatenated outputs.

$$MSA(O^{j-1}) = \left[SA_1\left(O^{j-1}\right); \cdots; SA_n\left(O^{j-1}\right)\right]\boldsymbol{W}_{mas}, \tag{3}$$

where $\boldsymbol{W}_{mas} \in \mathbb{R}^{nd \times C}$ and d is set to C/n. The output of the MSA is then fed into the MLP and added to the MSA result by a residual.

$$O^j = MSA\left(O^{j-1}\right) + MLP\left(MSA\left(O^{j-1}\right)\right) \in \mathbb{R}^{L \times C}, \tag{4}$$

which $j = 1, 2, \cdot, \cdot, \cdot, J$ represents the transformer layers.
Image Modality. We denoted $X \in \mathbb{R}^{H \times W \times C}$ to be a raw input image from the image dataset, where (H, W) is the resolution of the image and C is the number of channels. We reshape the image into a sequence of flattened $2D$ patches $X_p \in \mathbb{R}^{N \times P^2 \times C}$, where $N = \left(H \times W/P^2\right)$ is the number of image patches and (P, P) is the resolution of each patch. Then we generate an attentional image features $X_p^J \in \mathbb{R}^{M \times 1024}$ by the MSA and MLP:

$$X_p^J = MSA\left(X_p\right) + MLP\left(MSA\left(X_p\right)\right). \tag{5}$$

Finally, we obtain the final feature $\boldsymbol{Z}_a$ by pooling the aggregated features $\boldsymbol{X}_p^J$ on average.
Text Modality. Unlike most existing cross-modal hashing methods that only use bag-of-words as input and fully connected layers as the encoder, we expect the encoder can enhance the global connections of the text features. The word vector of the bag-of-words $G_v \in \mathbb{R}^d$ (where d is the dimension of word vector) is then turned into an attentional image features $\boldsymbol{G}_v^J \in \mathbb{R}^{M \times 1024}$ by the MSA and MLP:

$$G_v^J = MSA\left(G_v\right) + MLP\left(MSA\left(G_v\right)\right), \tag{6}$$

where G_v^J is the output of the last layer of the transformer and 1024 is the number of the channel. Finally, we obtain the final feature $\boldsymbol{E}_a$ by pooling the aggregated features $\boldsymbol{G}_v^J$ on average.

3.3 Heterogeneous Feature Interaction

Since the semantic relations among object entities from different modalities exist consistency, inspired by [23], as illustrated in Fig.3, we exploit a heterogeneous feature interaction(HFI) module to perform heterogeneous feature fusion, especially through the cross attention in it to learn the interaction between different modal features, so as to better align the semantic relation between modalities. Specifically, we introduce channel attention, spatial attention, and cross-attention to align the semantic relations between the different modalities. We set the expanded attentional feature maps $\boldsymbol{Z}_a$ of the image and the attentional feature maps $\boldsymbol{E}_a$ of the text as Z and E, with attentional feature shapes of $H \times W \times C$ and $H \times W \times C$.
Spatial Attention. In this work, we use spatial attention to enhance the context-aware similarity of semantic relations of images and text by learning the global contextual information from the images and text. Spatial attention is used in space only with the self-attention mechanism. And spatial attention is computed separately on the attentional image modality and the attentional text modality.

Specifically, suppose the input attentional features are $\boldsymbol{Z} \in \mathbb{R}^{H \times W \times C}$, we first apply two separate convolution layers with 1×1 kernels on $\boldsymbol{Z}$ to generate query $\boldsymbol{Q}$ and $\boldsymbol{V}$ respectively, where $\boldsymbol{Q}$, $\boldsymbol{V} \in \mathbb{R}^{H \times W \times C'}$ and $C' = \frac{1}{2}C$ is the reduced channel number. Then a average pooling with aggregate the feature expressions

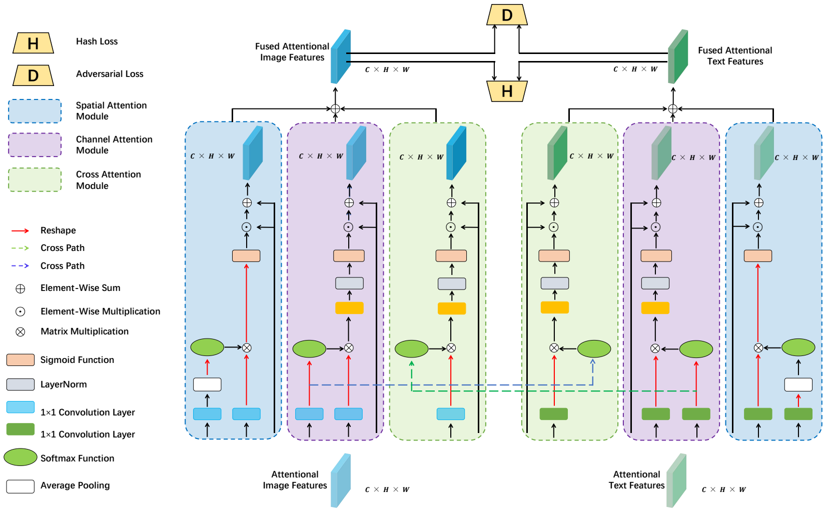


Fig. 3. The proposed heterogeneous feature interaction module, which includes spatial attention, channel attention, and cross-attention. The output of the heterogeneous feature interaction is computed simultaneously for adversarial loss and hashing loss.

is applied to the features $\boldsymbol{Q}$ to generate features $\bar{\boldsymbol{Q}} \in \mathbb{R}^{1\times 1\times C'}$. Next the feature is reshaped to $\hat{\boldsymbol{Q}} \in \mathbb{R}^{1\times C'}$. The features $\boldsymbol{V}$ is reshaped to $\bar{\boldsymbol{V}} \in \mathbb{R}^{HW\times C'}$. We can generate a spatial attention map $A_s \in \mathbb{R}^{HW\times 1}$ via softmax operations and matrix multiplication as:

$$\boldsymbol{A}_s = softmax(\hat{\boldsymbol{Q}})\bar{\boldsymbol{V}} \in \mathbb{R}^{HW\times 1}. \tag{7}$$

Then the attention map $\boldsymbol{A}_s$ is reshaped to $\boldsymbol{A}_s^s \in \mathbb{R}^{H\times W\times 1}$ and the sigmoid function is used to keep all parameters between 0 and 1 to get $\hat{\boldsymbol{A}}_s^s$:

$$\hat{\boldsymbol{A}}_s^s = sigmoid(\boldsymbol{A}_s^s), \in \mathbb{R}^{H\times W\times 1}. \tag{8}$$

Finally, combining the $\hat{\boldsymbol{A}}_s^s$ and $\boldsymbol{Z}$ by element-wise multiplication and $\boldsymbol{Z}$ by element-wise sum to get the final fused image features:

$$\boldsymbol{Z}_s^s = \alpha\hat{\boldsymbol{A}}_s^s \cdot \boldsymbol{Z} + \boldsymbol{Z} \in \mathbb{R}^{H\times W\times C}, \tag{9}$$

where α is a scalar parameter.

Channel Attention. The per-channel mapping of high-level semantic features is typically responsive to a specific target category. Similarly, processing features across all channels will hinder representation capabilities. Therefore, we use channel attention to selectively enhance the features within each modality. We can compute channel attention map $\boldsymbol{A}_c$ and the fused channel-features$\boldsymbol{Z}_c^c$ in a similar manner. Likewise, channel attention is a self-attention mechanism used only on channels. Notice the process, compare with the spatial attention without average pooling and add a layer norm, 1 × 1 convolution layer after matric multiplication.

Cross Attention. To enhance the context-aware similarity of semantic relations between different modalities, we use the cross-attention to learn such mutual information from the image branch and text branch.

Specifically, we use $\boldsymbol{Z} \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{E} \in \mathbb{R}^{H\times W\times C}$ to denote attentional image features and attentional text features respectively. Taking the attentional branch for example, we first apply a 1×1 convolution layers with a reshape operation on $\boldsymbol{E}$ to generate query $\boldsymbol{Q}'$, where $\boldsymbol{Q}' \in \mathbb{R}^{H\times W\times 1}$ and reshape it to $\hat{\boldsymbol{Q}}' \in \mathbb{R}^{HW\times 1\times 1}$. Then we compute the cross-attention from the image branch by performing similar operations as channel attention and the cross-attention computed from the image branch is encoded into the text value $\boldsymbol{V}'$ as,

$$\boldsymbol{A}_c = softmax(\hat{\boldsymbol{Q}}')\boldsymbol{V}' \in \mathbb{R}^{1\times 1\times C'}. \tag{10}$$

Then the attention map $\boldsymbol{A}_c$ is applied to 1×1 convolution layer with a reshape operation, and layernorm gets $\boldsymbol{A}_c^c \in \mathbb{R}^{1\times 1\times C}$to improve model training speed and accuracy. The next is to use sigmoid function to keep all parameters between 0 and 1 to get $\hat{\boldsymbol{A}}_c^c$:

$$\hat{\boldsymbol{A}}_c^c = sigmoid(\boldsymbol{A}_c^c), \in \mathbb{R}^{1\times 1\times C}. \tag{11}$$

Combining the $\hat{\boldsymbol{A}}_c^c$ and $\boldsymbol{E}$ by element-wise multiplication and $\boldsymbol{E}$ by element-wise sum to get the final fused image features$\bar{\boldsymbol{Z}}_c^c$:

$$\bar{\boldsymbol{Z}}_c^c = \eta\hat{\boldsymbol{A}}_c^c \cdot \boldsymbol{E} + \boldsymbol{E} \in \mathbb{R}^{H\times W\times C}, \tag{12}$$

where η is a scalar parameter.

Finally, the spatial fused features $\boldsymbol{Z}_s^s$ and channel fused features $\boldsymbol{Z}_c^c$, cross-attentional features $\bar{\boldsymbol{Z}}_c^c$ are simply combined with an element-wise sum, generating the fused features $\boldsymbol{F}^I$ for image modality.

$$\boldsymbol{F}^I = \boldsymbol{Z}_s^s + \boldsymbol{Z}_c^c + \bar{\boldsymbol{Z}}_c^c. \tag{13}$$

In the same way, the fused features $\boldsymbol{F}^T$ for text modality is got.

$$\boldsymbol{F}^T = \boldsymbol{E}_s^s + \boldsymbol{E}_c^c + \bar{\boldsymbol{E}}_c^c. \tag{14}$$

3.4 Network Learning and Optimization

Adversarial Loss. To maintain modality invariance and semantic consistency across modalities, we introduce adversarial loss to align the semantic relations across modalities, and inspired by [5], we design a discriminator an as a classifier to discriminate the modalities to which the unknown relations belong. In this process, the semantic relations captured by self-attention in one modality and the semantic relations after attentional enhancement are treated as true semantic relations, while the semantic relations acquired in the other modality are treated as false semantic relations.

As the discriminator struggles to discriminate unknown relations, the captured semantic relations $\boldsymbol{F}^I$ and $\boldsymbol{F}^T$ struggle to confuse the discriminator. We define the adversarial loss as $\mathcal{L}_{adv}$. The adversarial loss in semantic relation learning $\mathcal{L}_{adv}$ can be formulated as follows:

$$\mathcal{L}_{adv} = -\frac{1}{n}\sum log\left(D_F\left(\boldsymbol{F}^I;\theta_{D_F}\right)\right) - \frac{1}{n}\sum log\left(1 - D_F\left(\boldsymbol{F}^T;\theta_{D_F}\right)\right) \quad (15)$$

Hash Loss. In order to better achieve a balanced state between modalities in the generated similarity matrix, we obtain fusion features $\boldsymbol{F}^I$ and $\boldsymbol{F}^T$ with semantic relations enhancement through the HFI module. After normalizing $\boldsymbol{F}^I, \boldsymbol{F}^T$ to $\boldsymbol{F}_I, \boldsymbol{F}_T$ which have unit L_2 norm each row, we can calculate the cosine similarity matrices $\boldsymbol{S}_I^F \in [-1,1]^{M\times M}$ on $\boldsymbol{F}_I$ and $\boldsymbol{S}_T^F \in [-1,1]^{M\times M}$ on $\boldsymbol{F}_T$ to describe the original neighborhood structure for the input image modality and text modality, respectively. We calculate the similarity matrix $\boldsymbol{S}^F \in [-1,1]^{M\times M}$ between the modal features of two images and text. Meanwhile, we further integrate it with image matrix $\boldsymbol{S}_I^F$ and text matrix $\boldsymbol{S}_T^F$ generated similarity matrix $\boldsymbol{S}$, so as to further bridge the heterogeneous semantic gap between modalities.

$$\begin{gathered}\boldsymbol{S} = \lambda\boldsymbol{S}_I^F + \beta\boldsymbol{S}_T^F + \omega\boldsymbol{S}^F = \{\boldsymbol{S}_{ij}\}_{i,j=1}^M\,, \\ s.t.\lambda,\beta,\omega \geq 0, \lambda+\beta+\omega = 1, \boldsymbol{S}_{ij} \in [-1,+1]\end{gathered} \quad (16)$$

where S_{ij} represents the pairwise similarity of an image text data pair. λ, β, ω are the trade-off parameters that balance the similarity information from different modalities.

To obtain a high-level semantic description of the image and text modalities, we learn to obtain the hidden states using two functions $\boldsymbol{H}_I = f_I(\boldsymbol{F}^I;\theta_I)$ and $\boldsymbol{H}_T = f_T(\boldsymbol{F}^T;\theta_T)$ respectively, where θ_I and θ_T are two learnable parameters. In order to get the hash codes of the image and text, we adopt the sign function.

$$\boldsymbol{B}_* = sign(\boldsymbol{H}_*), * \in \{I, T\}\,. \quad (17)$$

Inspired by DSAH [21], although the ways and contents of obtaining matrix $\boldsymbol{S}$ in our method are different, in order to bridge the modal gap between different modalities, we adopt hash loss, including intra-modal loss, inter-modal loss, and symmetric loss function, just like them.

$$\begin{aligned}\mathcal{L}_{hl} = \underset{\boldsymbol{B}_I,\boldsymbol{B}_T}{min}\sum\left|\left|1 - \boldsymbol{S}^B\right|\right|^2 + \underset{\boldsymbol{B}_I,\boldsymbol{B}_T}{min}\sum\left|\left|\gamma\boldsymbol{S}_*^F - \boldsymbol{S}_*^B\right|\right|^2 \\ + \underset{\boldsymbol{B}_I,\boldsymbol{B}_T}{min}\sum\left|\left|\gamma\boldsymbol{S} - \boldsymbol{S}_*^B\right|\right|^2 * \in \{I,T\}\,,\end{aligned} \quad (18)$$

where after normalizing $\boldsymbol{B}^I, \boldsymbol{B}^T$ to $\boldsymbol{B}_I, \boldsymbol{B}_T$ which have unit L_2 norm each row, we can calculate the cosine similarity matrices $\boldsymbol{S}_I^B \in [-1,1]^{M\times M}$ on $\boldsymbol{B}^I$ and $\boldsymbol{S}_T^B \in [-1,1]^{M\times M}$ on $\boldsymbol{F}^T$ to describe the original neighborhood structure for the input images modality and texts modality respectively. And we calculate the cosine distances between hash codes of image and text to generate similarity matrices $\boldsymbol{S}^B = \boldsymbol{B}_I\boldsymbol{B}_T^\intercal$. And γ is a trade-off parameter.

We combine the losses of different modules into our final objective function, as shown below:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \mathcal{L}_{hl}, \tag{19}$$

where $\mathcal{L}_{adv}$ and $\mathcal{L}_{hl}$ include the adversarial loss, hash loss. The hash loss includes inter-modal loss and symmetric loss. Since the sign function has the problem of gradient zero for any non-zero input to human, we substitute the tanh function for the sign function:

$$\boldsymbol{B}_* = tanh(\eta \boldsymbol{H}_*), * \in \{I, T\}, \tag{20}$$

where η>0 is a scaling parameter, when $\lim_{\eta\rightarrow\infty} tanh(\eta \boldsymbol{H}_*) = sign(\boldsymbol{H}_*)$.

4 Experiments and Evaluations

4.1 Datasets and Evaluation

We conducted experiments on three public cross-modal retrieval datasets, including Wiki [14], MIRFlickr-25K [8] and NUS-WIDE [3], to verify the validity of our method. In our experiments, we adopt mAP@50 to evaluate the retrieval performance.

4.2 Implementation Details

In this section, we present the implementation details of our HILN in the experiments. We implement the method HILN by pytorch [13], and workstation configured with NVIDIA RTX 3090 GPU. We set the dimension of the feature representation extracted by the transformer to 1024. After the heterogeneous feature interaction, the feature representation obtained by dimensionality reduction is 1024 and set the dimension of the hash layer to be consistent with the length of the hash code.

Moreover, we train the proposed HILN in a mini-batch way and set the batch size as 32. For other comparison methods, we set the optimal experimental parameter configuration provided by their authors for training. For fairness, for all methods, we use the same dataset for performance comparison. The weight decay rate is 0.0005 and the momentum is set to 0.9. When training on the Wiki dataset, the learning rate of the network is set to 0.01, λ = 0.4, β = 0.4, ω = 0.2 when training on the MIRFlickr-25k and NUS-WIDE datasets, the learning rate of the network is set to 0.001, λ = 0.45, β = 0.45, ω = 0.1. The training epochs on the Wiki, MIRFLickr-25K, and NUS-WIDE datasets are set to 150, 100, and 80, respectively.

4.3 Performance

We compare our HILN with several representative deep unsupervised cross-modal hashing retrieval methods including DBRC [6], UDCMH [18],DJSRH [16],

Table 1. The mAP@50 values of Wiki, MIRFlickr-25k, and NUS-WIDE at various code lengths. Bold data represent the best performance among all contrasting methods.

Task	Method	Wiki				MIRFlickr-25k				NUS-WIDE			
		16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
I→T	DBRC [6]	0.253	0.265	0.269	0.288	0.617	0.619	0.620	0.621	0.424	0.459	0.447	0.447
	UDCMH [18]	0.309	0.318	0.329	0.346	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558
	DJSRH [16]	0.388	0.403	0.412	0.421	0.810	0.843	0.862	0.876	0.724	0.773	0.798	0.817
	JDSH [12]	0.346	0.431	0.433	0.442	0.832	0.853	0.882	0.892	0.736	0.793	0.832	0.835
	DSAH [21]	0.416	0.430	0.438	0.445	0.863	0.877	0.895	0.903	0.775	0.805	0.818	0.827
	AGSH [15]	0.397	0.434	0.446	-	0.679	0.691	0.698	-	0.543	0.552	0.562	-
	KDCMH [10]	-	-	-	-	0.713	0.716	0.724	0.728	0.615	0.628	0.637	0.642
	DGCPN [22]	0.420	0.438	0.440	0.448	0.875	0.891	0.908	0.918	0.788	0.820	0.826	0.833
	AGCH [26]	0.408	0.425	0.433	0.450	0.865	0.887	0.892	0.912	**0.809**	0.830	0.831	0.852
	OURS	**0.446**	**0.453**	**0.467**	**0.472**	**0.878**	**0.908**	**0.932**	**0.944**	0.793	**0.830**	**0.862**	**0.879**
T→I	DBRC [6]	0.573	0.588	0.598	0.599	0.618	0.626	0.626	0.628	0.455	0.459	0.468	0.473
	UDCMH [18]	0.622	0.633	0.645	0.658	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716
	DJSRH [16]	0.611	0.635	0.646	0.658	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789
	JDSH [12]	0.630	0.631	0.647	0.651	0.825	0.864	0.878	0.880	0.721	0.785	0.794	0.804
	DSAH [21]	0.644	0.650	0.660	0.662	0.846	0.860	0.881	0.882	0.770	0.790	0.804	0.815
	AGSH [15]	0.431	0.443	0.453	-	0.674	0.689	0.693	-	0.543	0.567	0.570	-
	KDCMH [10]	-	-	-	-	0.711	0.715	0.731	0.733	0.623	0.636	0.647	0.651
	DGCPN [22]	**0.644**	0.651	0.660	**0.662**	0.859	0.876	0.890	0.905	0.783	0.802	0.812	0.817
	AGCH [26]	0.627	0.640	0.648	0.658	0.829	0.849	0.852	0.880	0.769	0.780	0.798	0.802
	OURS	0.643	**0.651**	**0.660**	0.661	**0.877**	**0.908**	**0.932**	**0.944**	**0.793**	**0.831**	**0.862**	**0.879**

JDSH [12], DSAH [21], AGSH [15], KDCMH [10], AGCH [26] DGCPN [22]. Table 1 shows the mAP@50 values of HILN and other comparison methods on MIRFlickr-25k, NUS-WIDE, and Wiki in two cross-modal retrieval tasks for four lengths of hash codes. And figure 4 shows the precision@top-K curves on three datasets at 128 bits among five comparison methods. $I \rightarrow T$ means that the query is image and text modality is the database. $T \rightarrow I$ is the opposite. It can be seen that HILN is significantly better than the latest unsupervised cross-modal hashing methods. Specifically, compared to DGCPN [22], for the Wiki, as shown in the results, we achieve boosts of 5.3% in average mAP@50 for different hash code lengths in the $I \rightarrow T$ task. Moreover, HILN achieves boosts of 1.9% and 3.7% in average mAP@50 with different hash code lengths in $I \rightarrow T$ task and $T \rightarrow I$ task on MIRFlickr-25k respectively, and achieves boosts of 2.9% and 4.6% in two retrieval tasks on NUS-WIDE.

The main reason for the performance improvement is the HFI proposed by HILN. It also ensures that the performance of image retrieval for text and text retrieval for images is essentially the same.

4.4 Ablation Study

We design several variants to validate the impact of our proposed modules and to demonstrate the superiority of the original HILN.

Table 2. The mAP@50 results at 128 bits of ablation study about MSA on MIRFlickr-25k and NUS-WIDE. Bold data represent the best performance.

Model	Configuration	MIRFlickr		NUS-WIDE	
		I→T	T→I	I→T	T→I
Baseline	-	0.903	0.882	0.827	0.815
HILN-1	Baseline+IMSA	0.914	0.892	0.842	0.827
HILN-2	Baseline+TMSA	0.910	0.888	0.838	0.823
HILN-3	Baseline+I,TMSA	**0.919**	**0.896**	**0.848**	**0.832**

Table 3. The mAP@50 results at 128 bits of ablation study about HFI on MIRFlickr-25k and NUS-WIDE. Bold data represent the best performance of the HILN method.

Model	Configuration	MIRFlickr		NUS-WIDE	
		I→T	T→I	I→T	T→I
HFI-1	-	0.919	0.896	0.848	0.832
HFI-2	HFI-1+ATT	0.939	0.939	0.871	0.871
HFI-3	HFI-1+$\mathcal{L}_{adv}$	0.931	0.928	0.862	0.853
HFI-4	HFI-2+ $\mathcal{L}_{adv}$	**0.944**	**0.944**	**0.879**	**0.879**

The effectiveness of multi-head self-attention. As shown in Table 2, several variants we designed to verify the effectiveness of the multi-head self-attention module. HILN-1 extracts global semantic information from images using only the multi-head self-attention mechanism. HILN-2 extracts global semantic information from text using only the multi-head self-attention mechanism. HILN-3 extracts global semantic information from images and texts using the multi-head self-attention mechanism.

From the results of Baseline, HILN-1, HILN-2, and HILN-3, we find the effectiveness of multi-head self-attention(MSA). HILN-1 improves mAP@50 by 1.1% and 1.5% over Baseline for the $T \rightarrow I$ task on both datasets MIRFlickr and NUS-WIDE, respectively. HILN-1 improves mAP@50 by 1.2% and 1.8% over Baseline for the $I \rightarrow T$ task on both datasets MIRFlickr and NUS-WIDE, respectively. We find that the reason for the improved performance of HILN-1 is due to the extraction of global semantic information of the image using the MSA. HILN-2 improves mAP@50 by 0.7% and 1% over Baseline for the $T \rightarrow I$ task on both datasets MIRFlickr and NUS-WIDE, respectively. HILN-2 improves mAP@50 by 0.7% and 1.3% over Baseline for the $I \rightarrow T$ task on both datasets MIRFlickr and NUS-WIDE, respectively. We find that the reason for the improved performance of HILN-2 is due to the extraction of global semantic information of the text using the MSA. From the mAP@50 results of HILN-1, HILN-2, and HILN-3, we find that the MSA can effectively capture the global dependencies of semantic features within the modality for modeling semantic relations among object entities. These results suggest that the MSA is more effective for image modalities and performs better if used simultaneously.

The effectiveness of heterogeneous feature interaction and adversarial loss. Meanwhile, as shown in Table 3, there are several variants we designed to verify the effectiveness of the heterogeneous feature interaction and adversarial loss. HFI-1 stands for only using the MSA mechanism to capture the global semantic information of image and text modalities. HFI-2 builds on HFI-1 by adding only the heterogeneous feature interaction module. HFI-3 adds only the adversarial loss to HFI-1. HFI-4 builds on HFI-2 by adding only the adversarial loss.

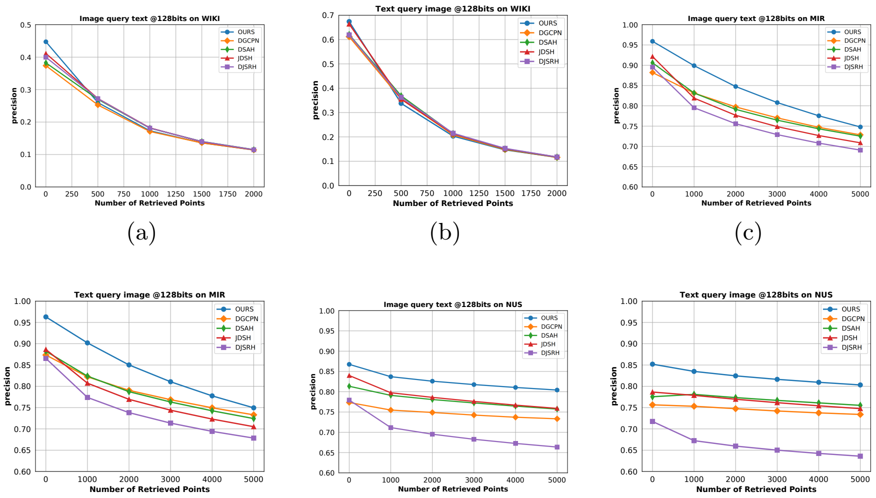
Table 4. The mAP@50 results at 128 bits of ablation study about three of attention on MIRFlickr-25k and NUS-WIDE.

Model	Configuration	MIRFlickr-25k		NUS-WIDE	
		I→T	T→I	I→T	T→I
HFI-a	MSA+$\mathcal{L}_{adv}$	0.919	0.896	0.848	0.832
HFI-b	Baseline+Spatial attention	0.927	0.907	0.861	0.849
HFI-c	Baseline+Spatial+Channel Attention	0.935	0.918	0.869	0.856
HFI-d	Baseline+Spatial+Channel+Cross Attention	**0.944**	**0.944**	**0.879**	**0.879**

HFI-2 improves mAP@50 by 4.8% and 4.7% over HFI-1 for the $T \rightarrow I$ task on both datasets, MIRFlickr and NUS-WIDE, respectively. HFI-2 improves mAP@50 by 2.2% and 2.7% over HFI-1 for the $I \rightarrow T$ task on both datasets, MIRFlickr and NUS-WIDE, respectively. The performance improvement of HFI-2 attributes the attention from the heterogeneous feature interaction module, which effectively aligns the global semantic similarity relations between different modalities so that global semantic relations between modalities are consistent. HFI-3 improves mAP@50 by 3.2% and 2.5% over HFI-1 for the $T \rightarrow I$ task on both datasets, MIRFlickr and NUS-WIDE, respectively. HFI-3 improves mAP@50 by 1.3% and 1.7% over HFI-1 for the $I \rightarrow T$ task on both datasets, MIRFlickr and NUS-WIDE, respectively. And from the mAP@50 results of HFI-2 and HFI-4, the performance improvement of HFI-2 comes mainly from the adversarial loss. Thus, it is shown that adversarial loss effectively maintains semantic consistency.

Comparing HFI-2 and HFI-3, HFI-2 makes the performance of $I \rightarrow T$ and $T \rightarrow I$ tasks comparable, while the performance of HFI-3 is stronger for $I \rightarrow T$ tasks than for $T \rightarrow I$ tasks. Meanwhile, the performance of HFI-2 is better than that of HFI-3. From the above comparative analysis, it can be found that if the heterogeneous feature interaction or the adversarial loss is used alone, the effect is not as good as the effect of using both simultaneously. Therefore, the adversarial loss we introduce is effective for cross-modal hashing.

The effectiveness of spatial attention, channel attention, and cross attention. As shown in Table 4, several variants we designed to verify the effectiveness of spatial attention, channel attention, and cross attention. HFI-a represents the use of MSA and adversarial loss. HFI-b builds on HFI-a by adding only spatial attention. HFI-c builds on HFI-b by adding only the channel



(a) (b) (c)

(d) (e) (f)

Fig. 4. Precision@top-K curves on three datasets at 128 bits.

attention. HFI-d builds on HFI-c by adding only the cross attention. From the results of HFI-a, HFI-b, HFI-c, and HFI-d, we find the effectiveness of spatial attention, channel attention, and cross attention. In particular, when three kinds of attention are used simultaneously, the retrieval effect will be better.

5 Conclusion

In this paper, we propose a novel unsupervised cross-modal hashing method called Heterogeneous Interactive Learning Network(HILN) for unsupervised cross-modal retrieval. To bridge the heterogeneous semantic gap between different modalities, we introduce a multi-head self-attention mechanism to capture the global dependencies of semantic features within the modality for modeling semantic relations among object entities. Meanwhile, we exploit a heterogeneous feature interaction module for feature fusion to align the semantic relationships between different modalities. Moreover, we introduce adversarial loss into network learning to further maintain semantic consistency. Extensive experiments have shown the effectiveness of our method.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (Grant No.61902204), in part by the Natural Science Foundation of Shandong Province of China (Grant No.ZR2019BF028).

References

1. Bai, C., Zeng, C., Ma, Q., Zhang, J., Chen, S.: Deep adversarial discrete hashing for cross-modal retrieval. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 525–531 (2020)
2. Chen, S., Wu, S., Wang, L., Yu, Z.: Self-attention and adversary learning deep hashing network for cross-modal retrieval. Computers & Electrical Engineering **93**, 107262 (2021)
3. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Gu, W., Gu, X., Gu, J., Li, B., Xiong, Z., Wang, W.: Adversary guided asymmetric hashing for cross-modal retrieval. In: Proceedings of the 2019 on international conference on multimedia retrieval. pp. 159–167 (2019)
6. Hu, D., Nie, F., Li, X.: Deep binary reconstruction for cross-modal hashing. IEEE Transactions on Multimedia **21**(4), 973–985 (2018)
7. Hu, H., Xie, L., Hong, R., Tian, Q.: Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3123–3132 (2020)
8. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. pp. 39–43 (2008)
9. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3232–3240 (2017)
10. Li, M., Wang, H.: Unsupervised deep cross-modal hashing by knowledge distillation for large-scale cross-modal retrieval. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 183–191 (2021)
11. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)
12. Liu, S., Qian, S., Guan, Y., Zhan, J., Ying, L.: Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1379–1388 (2020)
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
14. Pereira, J.C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE transactions on pattern analysis and machine intelligence **36**(3), 521–535 (2013)
15. Shen, X., Zhang, H., Li, L., Liu, L.: Attention-guided semantic hashing for unsupervised cross-modal retrieval. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)

16. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3027–3035 (2019)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
18. Wu, G., Lin, Z., Han, J., Liu, L., Ding, G., Zhang, B., Shen, J.: Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In: IJCAI. pp. 2854–2860 (2018)
19. Yan, C., Bai, X., Wang, S., Zhou, J., Hancock, E.R.: Cross-modal hashing with semantic deep embedding. Neurocomputing **337**, 58–66 (2019)
20. Yang, B., Wang, L., Wong, D.F., Shi, S., Tu, Z.: Context-aware self-attention networks for natural language processing. Neurocomputing **458**, 157–169 (2021)
21. Yang, D., Wu, D., Zhang, W., Zhang, H., Li, B., Wang, W.: Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 44–52 (2020)
22. Yu, J., Zhou, H., Zhan, Y., Tao, D.: Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI. pp. 4626–4634 (2021)
23. Yu, Y., Xiong, Y., Huang, W., Scott, M.R.: Deformable siamese attention networks for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6728–6737 (2020)
24. Zhang, J., Peng, Y.: Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. IEEE Transactions on Multimedia **22**(1), 174–187 (2019)
25. Zhang, J., Peng, Y., Yuan, M.: Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network. IEEE transactions on cybernetics **50**(2), 489–502 (2018)
26. Zhang, P.F., Li, Y., Huang, Z., Xu, X.S.: Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. IEEE Transactions on Multimedia **24**, 466–479 (2021)
27. Zhu, L., Tian, G., Wang, B., Wang, W., Zhang, D., Li, C.: Multi-attention based semantic deep hashing for cross-modal retrieval. Applied Intelligence pp. 1–13 (2021)