Multi-modal Fine-grained Semantic-preserving Hashing for Cross-modal Retrieval

Bohan Zhang[1], Yaru Gao[1], Chenlong Song[1], Yuan Cao[1]✉, Yanwei Yu[1]✉

[1]Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China
{zhangbohan2674, gaoyaru, scl_0218}@stu.ouc.edu.cn
{cy8661, yuyanwei}@ouc.edu.cn

***Abstract*—With the rapid growth of multimodal data, cross-modal retrieval has become increasingly important in both the natural and remote sensing domains. Cross-modal hashing is appealing for its compact storage and efficient retrieval. However, most existing methods rely primarily on intramodal similarity, overlooking structural and fine-grained information. To address this limitation, we propose a fine-grained attention-based multiple similarity matrix derived from transformer output, which aggregates local concept-level features to capture more representative semantics. In addition, we design a loss framework combining semantic consistency, discrete quantization, and feature reconstruction alignment. By using the semantic feature reconstruction based on the automatic encoder framework, our approach unifies cross-modal similarity. Extensive experiments on two public and one remote sensing datasets demonstrate clear advantages over state-of-the-art methods.**

***Index Terms*—Cross-modal retrieval, hashing, fine-grained semantics**

I. Introduction

With the rapid advancement of multimedia technologies, the volume of multimodal data has been growing at an unprecedented rate. In particular, the proliferation of online platforms and the development of remote sensing technologies have resulted in massive repositories of heterogeneous data such as images, text, and satellite observations. [1] Consequently, efficient retrieval and analysis of relevant information across different modalities has become an essential task in both academic research and practical applications. Cross-modal retrieval aims to bridge the semantic gap between heterogeneous modalities and enables users to query data in one modality while retrieving related data in another.

Hashing-based methods have attracted increasing attention in cross-modal retrieval due to their advantages in storage efficiency and query speed. By encoding high-dimensional multimodal data into compact binary hash codes, these approaches enable large-scale similarity search in a low complexity manner. [2] Yet, traditional hashing techniques often focus on global similarity [3]–[5] , which may not capture subtle variations within multimodal data. This limitation is particularly evident in scenarios such as remote sensing, where data usually contain complex spatial structures, fine-grained object details, and diverse semantic contexts. Neglecting these properties can reduce the retrieval accuracy.

Recent works [6]–[10] have explored various strategies to improve cross-modal hashing, including matrix factorization, graph convolution, and adversarial consistency learning. While these methods achieve improvements, most of them still depend on coarse similarity structures or intra-modal relations, overlooking the importance of fine-grained semantics across modalities. Moreover, the heterogeneity of multimodal features further complicates similarity modeling, making it difficult to unify global and local semantic information in a coherent hash representation. Thus, current methods struggle to generate hash codes that are both discriminative and robust to noise and data imbalance.

To address these issues, we propose a Multi-modal Fine-grained Semantic-preserving Cross-modal Hash Learning Algorithm (MFSH). We first leverage the feature representations encoded by Transformer to construct multiple fine-grained similarity matrices. These matrices are capable of capturing global semantic associations, as well as localized fine-grained interaction information across modalities. These similarity matrices serve as supervisory signals for a heterogeneous graph convolutional network, enabling more effective cross-modal information aggregation and the generation of semantically consistent hash codes. Furthermore, we design a semantic feature reconstruction module that aligns the reconstructed features with their original counterparts, thereby enhancing the robustness and generalization ability of the learned codes.

II. Methodology

In this section, we elaborate the proposed Multi-modal Fine-grained Semantic-preserving Cross-modal Hash Learning Algorithm (MFSH). Fig. 1 shows the framework of MFSH, which contains the Fine-grained Similarity Matrix Construction module, Cross-modal Information Aggregation module and Semantic Feature Reconstruction module. In the Objective Function section, we describe the loss functions designed to train the network.

Corresponding authors: Yuan Cao and Yanwei Yu.

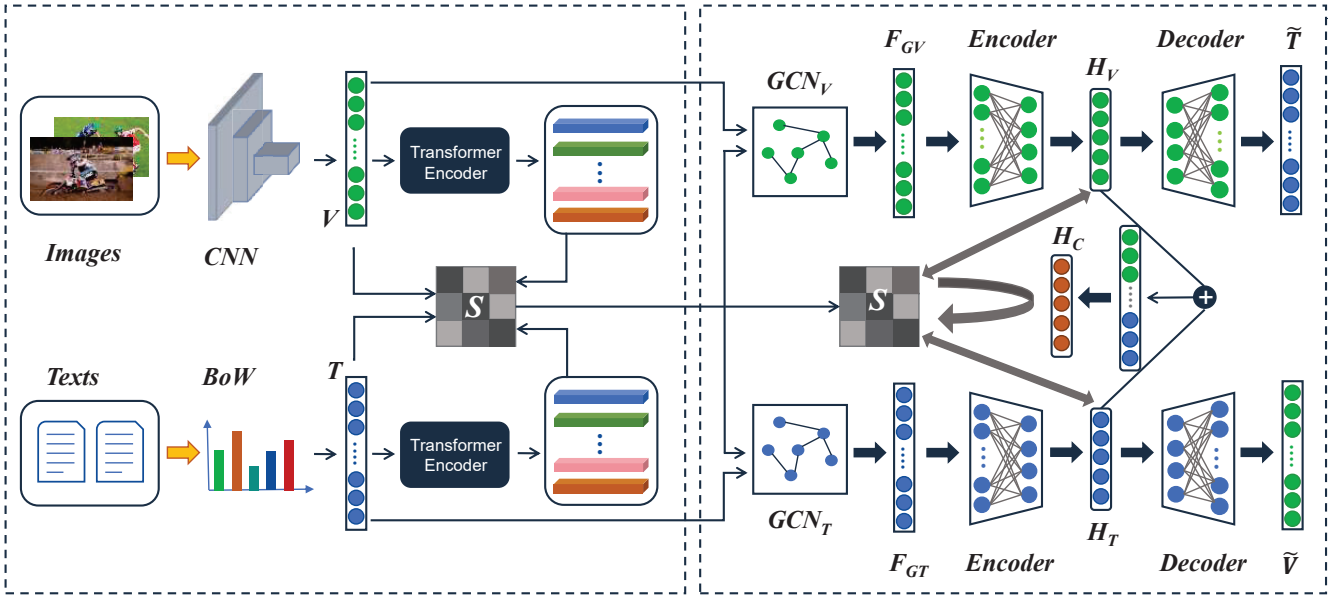

Fig. 1: The framework of MFSH. This method utilizes an end-to-end graph convolutional network as a cross-modal hashing function, mapping image and text data into a hash space. The supervision information used by the network is derived from the multi-granularity similarity matrices constructed based on the outputs of the Transformer Encoder. Furthermore, an effective objective loss function is designed based on the network structure and cross-modal retrieval characteristics.

A. Notations and Problem Definition

Let the training dataset $O = \{o_i\}_{i=1}^{n}$, where each instance with both image and text modalities. We use modality-specific feature extractors to extract feature representations from the raw image and text data. These features are expressed as $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_n\} \in \mathbb{R}^{n \times d_v}$, $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \cdots, \mathbf{t}_n\} \in \mathbb{R}^{n \times d_t}$. Here, n represents the number of instances, while d_v and d_t represent the feature dimensions for images and text, respectively. Furthermore, the hash code representations of the image and text instances can be denoted as $\mathbf{B}_* = \{-1, +1\} \in \mathbb{R}^{n \times r}, * \in \{V, T\}$, where r represents the length of the hash codes.

B. Multi-Granularity Similarity Matrix

To capture both global and local semantic information, we construct multiple similarity matrices.

Firstly, a global rough similarity matrix is constructed

$$\mathbf{S}_1 = \alpha_1 \mathbf{S}_V + (1 - \alpha_1)\mathbf{S}_T \tag{1}$$

where S_V and S_T are intramodal cosine similarity matrices.

Additionally, the image feature $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_n\}$ is input into Transformer Encoder to get an extended implicit semantic concept representation:

$$\{\mathbf{C}_{V,i}\}_{i=1}^{n} = \mathrm{TransformerEncoder}(\mathbf{V}) \tag{2}$$

Specifically, an image feature $\mathbf{v}_i$ through semantic decomposition obtain $\mathbf{C}_{V,i} = \{\mathbf{c}_v^1, \mathbf{c}_v^2, \cdots, \mathbf{c}_v^k\} \in \mathbb{R}^{k \times d_c}$. Text modality is processed in this way. For the k implicit semantic representations of graphics and text, an affinity matrix of $k \times k$ is calculated, where each point d_{ij} is the cosine distance corresponding to the implicit semantics. In order to take advantage of the local fine-grained similarity between images and texts, we apply text-to-visual annotations to pay attention to the area of each text chunk. The attention weight of each region is calculated as follows

$$\gamma_{ij} = \mathrm{Softmax}(\lambda_1 d_{ij}) = \frac{\exp(\lambda_1 d_{ij})}{\sum_{i=1}^{k} \exp(\lambda_1 d_{ij})} \tag{3}$$

Among them, λ_1 is used to adjust the inverse temperature of Softmax and strengthen the effect of Softmax. Then the accompanying visual features of the j-th text chunk are calculated as follows

$$\mathbf{a}_j^v = \textstyle\sum_{i=1}^{k} \gamma_{ij} \mathbf{c}_v^i \tag{4}$$

Then calculate the correlation between the j-th text chunk and the whole image:

$$R(\mathbf{a}_j^v, \mathbf{c}_t^j) = \cos(\mathbf{a}_j^v, \mathbf{c}_t^j) = \frac{\mathbf{a}_j^v \cdot \mathbf{c}_t^j}{\left\|\mathbf{a}_j^v\right\| \left\|\mathbf{c}_t^j\right\|} \tag{5}$$

Finally, the similarity from the whole text to the whole image is defined as

$$\mathbf{S}_{LSE}(T, V) = \log\Big(\sum_{j=1}^{k} \exp(\lambda_2 R(\mathbf{a}_j^v, \mathbf{c}_t^j))\Big)^{\frac{1}{\lambda_2}} \tag{6}$$

LSE is the abbreviation of Log SumExp pooling, and λ_2 is used to amplify the influence of the item with the highest correlation.Similarly, the similarity from the whole image to

the text can also be calculated. Finally, the expression for the local fine-grained similarity matrix S_2 is

$$\mathbf{S}_2 = \alpha_2 \mathbf{S}_{LSE}(V,T) + (1-\alpha_2)\mathbf{S}_{LSE}(T,V) \tag{7}$$

Finally, the expression of multi-modal fine-grained similarity matrix S is

$$\mathbf{S} = \beta \mathbf{S}_1 + (1-\beta)\mathbf{S}_2 \tag{8}$$

where α_2 and β denote the hyper-parameters for tuning the contributions of each component.

C. Hashing Process

The multi-modal fine-grained similarity matrix defined in the previous section and the graphic features V, T are synchronously transported to the corresponding graph volume product GCN module to obtain more structural semantics. The process of drawing and rolling of each layer is as follows

$$\mathbf{H}_m^{(\ell+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\mathbf{S}\tilde{D}^{-\frac{1}{2}}\mathbf{H}_m^{(\ell)}\mathbf{W}_m^{(\ell)}) \tag{9}$$

where $m \in \{V,T\}$, $\tilde{D}_{ii} = \sum_j \mathbf{S}_{ij}$, $\mathbf{W}_m^{(\ell)}$ as a convolution filter of the m modality at the l level, $\sigma(\cdot)$ is the activation function of GCN at the l level. Through this two-layer graph convolution network, modality-specific semantic features with rich graph structure semantics are obtained

$$\mathbf{F}_{GV} = \mathrm{GCN}_V(\mathbf{V}, \theta_{gv}) \tag{10}$$

$$\mathbf{F}_{GT} = \mathrm{GCN}_T(\mathbf{T}, \theta_{gt}) \tag{11}$$

θ_{gv} and θ_{gv} is the parameters of network. Using the hash code of one modality to reconstruct the depth feature of another modality can enhance the correlation and reduce the gap, which can be realized by introducing an automatic coding and decoding module. First, input the F_{GV}, F_{GT} obtained in the previous step into the automatic encoder, and add a hash layer to obtain the graphic hash code representation:

$$\mathbf{H}_* = \mathrm{Enc}(\mathbf{F}_{G*}, \vartheta_*) \in \mathbb{R}^{n\times r}, * \in \{V,T\} \tag{12}$$

Where ϑ_* is the parameter of each encoder of graphics and text. Next, the non-discretized codes $\mathbf{H}_V, \mathbf{H}_T$ are fused to obtain a unified representation $\mathbf{H}_C$, from which binary codes are generated using $\tanh(\cdot)$ as a relaxation of the sign function:

$$\mathbf{B}_* = \tanh(\mathbf{H}_*), * \in \{V,T,C\} \tag{13}$$

Finally, the decoder reconstructs the feature vector of another modality with different modal hash codes. The formula of this process is as follows, δ_V and δ_T is the parameters of network:

$$\tilde{V} = Dec(B_T, \delta_T) \in R^{n\times d_v} \tag{14}$$

$$\tilde{T} = Dec(B_V, \delta_V) \in R^{n\times d_t} \tag{15}$$

D. Objective Function

1) Semantic Joint Consistent Loss: The algorithm aims to ensure that the latent representation reflects the correlation between the original instances and that the similarity structure is preserved as much as possible in the hash codes. To achieve this, a semantic joint consistent loss is introduced, which is formulated as follows

$$\begin{aligned}\mathcal{L}_{SJC} = \left\|\mathbf{S} - \mathbf{B}_V^T\mathbf{B}_V\right\|_F^2 + \left\|\mathbf{S} - \mathbf{B}_T^T\mathbf{B}_T\right\|_F^2 \\ + \left\|\mathbf{S} - \mathbf{B}_C^T\mathbf{B}_C\right\|_F^2\end{aligned} \tag{16}$$

2) Discrete Quantization Loss: Our goal is to ensure that the latent modal semantic information contained in $\mathbf{H}_V, \mathbf{H}_T$ is preserved as much as possible in the unified binary representation $\mathbf{B}_C$ and to minimize the quantization error during the quantization process. To this end, we propose the discrete quantization loss as follows

$$\mathcal{L}_{DIS} = \|\mathbf{B}_C - \mathbf{H}_V\|_F^2 + \|\mathbf{B}_C - \mathbf{H}_T\|_F^2 \tag{17}$$

3) Feature Reconstruction Alignment Loss: Utilizing the hash codes of one modality to reconstruct the features of another modality allows for the thorough learning of latent semantics between modalities and facilitates cross-modal interaction. By aligning the reconstructed features with the original features, we have designed the following feature reconstruction alignment loss:

$$\mathcal{L}_{FRA} = \left\|\mathbf{V} - \tilde{V}\right\|_F^2 + \left\|\mathbf{T} - \tilde{T}\right\|_F^2 \tag{18}$$

To sum up, the overall objective function is

$$\mathcal{L} = \mathcal{L}_{SJC} + \mu_1\mathcal{L}_{DIS} + \mu_2\mathcal{L}_{FRA} \tag{19}$$

The expression mentioned above includes two hyper-parameters μ_1, μ_2, which are used to balance the weights of the loss functions from different modules.

III. Experiments

A. Datasets and Setup

We conduct experiments on MIR-Flickr25k [11], NUS-WIDE [12] and RSICD [13]. MIR-Flickr25k is an image classification dataset with 24 categories and 25,000 images. NUS-WIDE, used for image annotation, has 81 categories and 269,645 images. In our case, we use the common 21 labels as an image retrieval dataset. For the first two datasets, we randomly selected 2000 instances as the query set, and the remaining instances formed the retrieval set. Additionally, 5000 instances were randomly sampled from the retrieval set to form the training set. The visual content was represented using a 4096-dimensional AlexNet feature, while the textual content was represented using a 1386-dimensional bag-of-words (BOW) feature for MIR-Flickr25k and a 1000-dimensional BOW feature for NUS-WIDE. For RSICD in remote sensing, it is a single-label dataset with 31 classes and 10,921 coupled image-text instances. The dataset was split by random selection into the train, query, and retrieval sets (50%, 10%, and 40%, respectively).

TABLE I: The mAP@50 results on I → T and T → I retrieval tasks with 16 to 128 hash bits on three datasets.

Task	Backbone	Method	MIR-Flickr				NUS-WIDE				RSICD			
			16	32	64	128	16	32	64	128	16	32	64	128
I → T	Shallow	IMH	61.2	60.1	59.2	57.9	47.0	47.3	47.6	45.9	45.3	51.9	43.9	34.6
		LSSH	58.4	59.9	60.2	61.4	48.1	48.9	50.7	50.7	10.9	32.5	46.8	59.3
		FSH	70.9	75.9	82.6	86.8	55.5	55.9	56.2	57.2	42.8	58.8	61.2	63.4
		DSAH	85.8	87.8	88.9	90.3	76.3	80.0	80.3	82.1	57.6	59.3	61.2	63.6
	Deep	AGCH	86.5	88.7	89.2	91.2	80.9	83.0	83.1	85.2	-	-	-	-
		DAEH	86.6	87.2	87.4	88.2	78.8	82.1	83.9	84.1	58.7	59.9	62.1	64.3
		CIRH	89.1	90.1	92.3	93.7	79.4	82.5	84.0	85.4	61.3	61.9	62.9	63.5
		UCCH	88.2	**90.8**	91.9	93.1	81.2	82.3	84.1	86.1	46.5	60.5	**64.1**	**65.6**
		MSRH	**89.7**	90.1	91.8	93.3	81.7	83.6	83.9	86.4	-	-	-	-
		MFSH	89.2	90.2	**92.4**	**93.8**	**81.9**	**83.8**	**84.3**	**86.6**	**61.9**	**62.8**	63.7	64.5
T → I	Shallow	IMH	60.3	59.5	58.9	58.0	47.8	48.3	47.2	46.2	46.7	54.0	44.9	34.4
		LSSH	63.7	65.9	65.9	67.2	57.7	61.7	64.2	66.3	24.0	37.8	48.2	55.9
		FSH	71.7	77.9	83.2	86.2	57.6	57.8	58.1	58.3	14.0	15.1	25.4	37.3
		DSAH	84.2	87.0	87.9	88.5	74.3	76.1	78.9	80.8	60.0	59.8	59.3	62.0
	Deep	AGCH	82.9	84.9	85.2	88.0	76.9	78.0	79.8	80.2	-	-	-	-
		DAEH	83.4	83.5	84.6	85.9	74.3	76.1	79.9	80.8	62.7	63.1	64.5	65.3
		CIRH	86.4	88.5	89.6	90.0	77.4	77.5	80.1	81.2	63.4	63.4	**65.7**	**66.9**
		UCCH	83.4	85.0	85.5	87.8	78.0	78.9	81.5	81.9	48.6	62.7	63.9	65.4
		MSRH	87.1	88.7	89.4	90.5	78.4	**81.1**	81.0	82.0	-	-	-	-
		MFSH	**87.3**	**89.4**	**90.4**	**90.8**	**78.6**	78.6	**82.1**	**82.7**	**63.5**	**64.1**	65.2	65.7

TABLE II: The ablation results of each component in MFSH on NUS-WIDE.

Model	16		32		64		128	
	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
MFSH-1	75.3	74.6	78.3	74.1	81.6	75.6	81.9	77.2
MFSH-2	76.9	75.2	78.2	75.8	81.6	77.2	82.4	79.2
MFSH-3	78.2	75.8	77.4	76.3	83.1	78.7	83.3	80.4
MFSH-4	77.5	76.8	79.3	77.5	82.9	79.1	84.3	80.7
MFSH	**81.9**	**78.6**	**83.8**	**78.6**	**84.3**	**82.1**	**86.6**	**82.7**

Besides, we also study the impact of input parameters α, β, λ, and μ. Parameters α_1 and α_2 play a crucial role in balancing the contributions of different modalities in the similarity matrices, ensuring that both visual and textual features are adequately represented. Parameter β controls the integration of global and local similarity information, which is essential for capturing both the overall structure and fine-grained details of the data. The parameters λ and μ further fine-tune the model by balancing different components of the loss function, helping to improve the robustness and accuracy of the hash codes. We provide the exact values of the parameter configurations in our MFSH model as $\alpha_1 = 0.5$, $\alpha_2 = 0.5$, $\beta = 0.01$, $\lambda_1 = 10$, $\lambda_2 = 1$, $\mu_1 = 0.5$, $\mu_2 = 0.5$. The proposed MFSH is trained with Pytorch on a workstation which consists of a CPU with 12cores, 20 processors and an NVIDIA Geforce RTX 3090.

B. Baselines and Evaluation Metrics

We compare our method with nine state-of-the-art unsupervised cross-modal hashing baselines, including three shallow methods: IMH [4], LSSH [5], FSH [14] and six deep methods: DSAH [15], AGCH [8], DAEH [16], CIRH [17], UCCH [18], MRSH [19]. We conduct experiments on two tasks: retrieving texts with image query as input (I → T) and retrieving images with text query as input (T → I). We use mAP@50 as the evaluation metric.

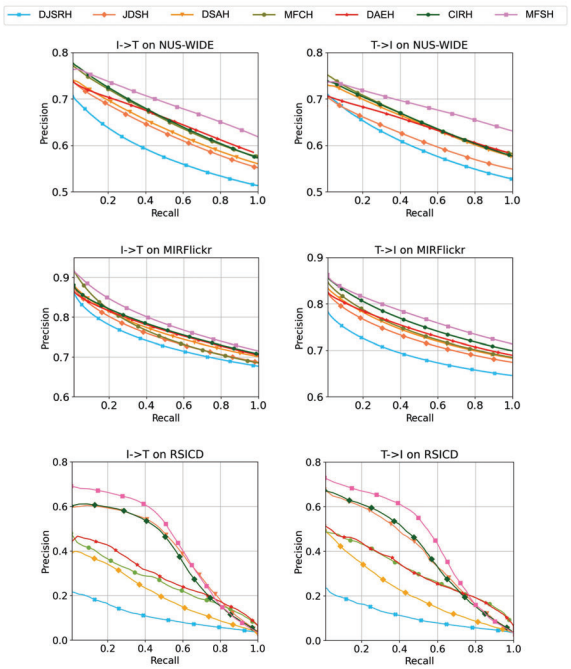

Fig. 2: The precision-recall curves of various hashing methods with 32 hash bits on three datasets.

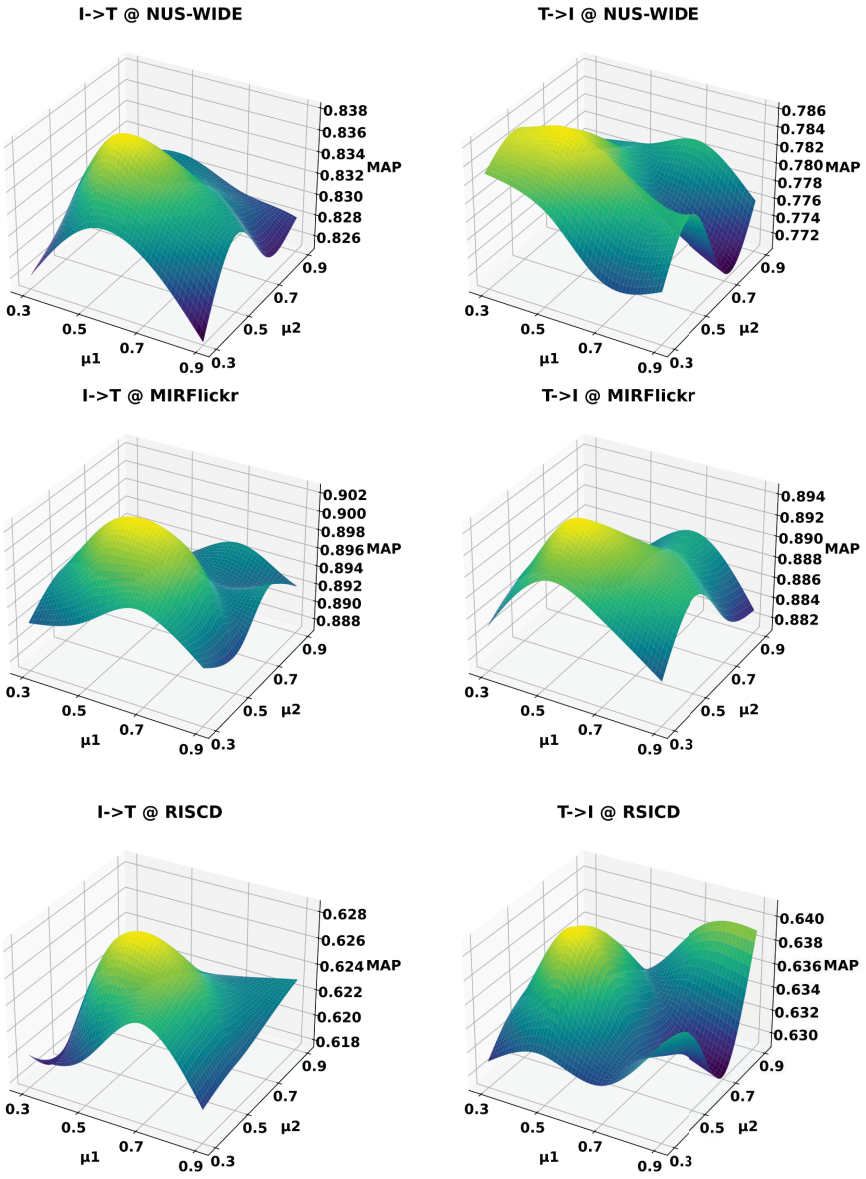

Fig. 3: The effect of hyperparameters μ_1 and μ_2 on image-to-text (I→T) and text-to-image (T→I) MAPs with 32 hash bits on NUS-WIDE, MIRFlick, and RSICD datasets.

C. Results and Analyses

Table I presents the MAP results of the proposed MFSH method and all unsupervised cross-modal hashing baseline methods on three datasets. The best scores are highlighted in bold, and the second-best scores are underlined. As shown in the Table I, the proposed MFSH method outperforms the baseline methods in most cases. In remote sensing area, our method is also competitive. This indicates that MFSH demonstrates strong retrieval capabilities in cross-modal retrieval tasks. In addition, Fig. 2 shows the PR curves of different comparison methods with 32 hash bits on three datasets. According to these experimental results, it can be observed that: In most cases, MFSH achieves better performance than the comparison methods, which shows that this method is effective and outstanding.

D. Ablation Study

To assess the effectiveness of each component, we designed the following variants: MFSH-1 replaces the original fine-grained similarity matrices with a global coarse-grained one; MFSH-2 constructs semantic representations using simple superposition; MFSH-3 replaces the graph convolutional network with a simple two-layer fully connected neural network; MFSH-4 removes the decoder and the feature reconstruction

alignment loss from the loss function. As shown in Table II, the fine-grained similarity matrix is the most critical. These components work together to enhance MFSH's performance in cross-modal retrieval tasks, enabling it to achieve excellent retrieval results.

E. Parameter Sensitivity Analysis

We investigate the impact of the two hyperparameters (μ_1 and μ_2) in the objective function. We tune them both in $\{0.3,0.5,0.7,0.9\}$. As shown in Fig. 3, the accuracy tends to stabilize as the values of the hyperparameters vary. The I→T and T→I MAPs reach their highest values when $\mu_1 = 0.5$ and $\mu_1 = 0.5$, which indicates that balance weights obtain the optimal results. This shows the model's robustness to these hyperparameters in this interval, ensuring reliable and stable performance under different configurations.

IV. Conclusion

This paper proposes a multi-modal fine-grained cross-modal hashing algorithm. It uses a graph convolutional network as the core structure. By building multi-modal fine-grained similarity matrices, it effectively captures semantic links between different modalities. In the experimental section, we comprehensively evaluate the algorithm across multiple standard datasets, focusing on cross-modal image-text retrieval tasks. The results demonstrate that our algorithm outperforms other advanced unsupervised cross-modal hashing algorithms in several evaluation metrics, including Mean Average Precision (MAP) and Precision-Recall Curves. This excellent performance is mainly due to the algorithm's innovative design in fine-grained similarity matrix construction and cross-modal information aggregation. It can efficiently bridge the semantic gap between modalities and generate highly discriminative and semantically meaningful hash codes. Future research can explore adding advanced techniques like attention mechanisms or more complex network structures to boost the algorithm's performance and adaptability.

Acknowledgment

This work is partially supported by the Prevention and Control of Emerging and Major Infectious Diseases-National Science and Technology Major Project under Grant No. 2025ZD01902800, and the Natural Science Foundation of Shandong Province under Grant No. ZR2024MF128.

References

[1] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Information Fusion*, vol. 67, pp. 94–115, 2021.

[2] W. Cao, W. Feng, Q. Lin, G. Cao, and Z. He, "A review of hashing methods for multimodal retrieval," *IEEE Access*, vol. 8, pp. 15 377–15 391, 2020.

[3] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2075–2082.

[4] J. Song *et al.*, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013.

[5] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014, pp. 415–424.

[6] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232–3240.

[7] X. Liu, Z. Hu, H. Ling, and Y.-m. Cheung, "Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 964–981, 2019.

[8] P.-F. 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, 2021.

[9] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 3626–3637, 2020.

[10] W. Wang *et al.*, "Set and rebase: Determining the semantic graph connectivity for unsupervised cross-modal hashing," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2021.

[11] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[12] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.

[13] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[14] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7380–7388.

[15] D. Yang, D. Wu, W. Zhang, H. Zhang, B. Li, and W. Wang, "Deep semantic-alignment hashing for unsupervised cross-modal retrieval," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 44–52.

[16] Y. Shi, Y. Zhao, X. Liu, F. Zheng, W. Ou, X. You, and Q. Peng, "Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7255–7268, 2022.

[17] L. Zhu, X. Wu, J. Li, Z. Zhang, W. Guan, and H. T. Shen, "Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8838–8851, 2022.

[18] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3877–3889, 2022.

[19] Y. Wu and Z. Li, "Mining similarity relationships for unsupervised cross-modal hashing," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.