ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448261

JOINT-SEMANTICS MULTI-SIMILARITY HASHING FOR CROSS-MODAL RETRIEVAL

Weigang Wang[1], Zhongwen Guo[1], Chao Yang[2], Jinxin Wang[1], Sining Jiang[1], Tianao Zhang[3] .*

[1]Faculty of Information Science and Engineering, Ocean Nuiversity of China, Qingdao, China
[2]School of Computer Science, University of Technology Sydney, Sydney, Australia
[3]Qingdao High-tech Zone Investment and Development Group Co., LTD, Qingdao, China

ABSTRACT

Recently, cross-modal hashing has attracted much attention in large-scale image retrieval scenarios. However, most existing methods ignore the potential higher-order relationships and label semantic information between heterogeneous modality data. Besides, the imbalanced training samples could bias the learning process in most classes and affect the retrieval performance. To solve the above problems, we proposed a Joint-semantics Multi-Similarity Hashing method for cross-modal retrieval (JMSH). We first construct a joint semantic similarity matrix, which supervises hash learning by integrating multi-modal features and semantic labels. This method generates higher-order semantic features that maintain semantic correlation effectively. Then, we propose a multi-similarity loss based on adaptive margin, which can collect and weight informative pairs efficiently and accurately, thus producing more discriminative hashing code and improving retrieval performance. Extensive experiments on two benchmark datasets show the superiority of JMSH in cross-modal retrieval tasks.

Index Terms— cross-modal hashing, supervised learning, cross-modal retrieval, multi-similarity loss

1. INTRODUCTION

With the rapid development of online social media, data from various modalities has proliferated, including images, text, audio, and videos. In the real world, users often need to use data from one modality (e.g., text) to retrieve the most potentially similar instances from another modality (e.g., image) and vice versa. Benefiting from the low storage costs and high retrieval efficiency, cross-modal hashing has been widely used in large-scale heterogeneous data retrieval scenarios [1, 2, 3, 4, 5, 6, 7]. However, multi-modal data exist in diverse formats and distributions, resulting in a heterogeneity gap. To bridge the heterogeneity gap among multi-modal data, several supervised cross-modal hashing retrieval methods have been proposed [3, 8, 9, 10], which employ semantic labels to guide the hashing learning process.

Corresponding author: guozhw@ouc.edu.cn. This work was supported by the National Key RD Program of China under Grant 2020YFB1707701.

Although most of the existing methods achieve good performance, there are still two shortcomings. Firstly, these methods only use labels or text annotation information to measure the semantic correlation between modalities. For example, DCMH [3] exploited a semantic label similarity matrix for supervised hashing learning, while SSAH [8] and AGAH [10] learned high-order semantic features from multi-label annotations and performed adversarial learning across modalities. SRLCH [11] leveraged the relational labels information in the semantic space to learn more discriminative hashing codes. CMMQ [12] jointly considered different modalities, combated noisy labels, and guided the training process by focusing on samples with small losses. CMHH [9] introduced noise labels to associate different modality features and designed a Hadamard product matrix to generate proxy codes for each class. However, these methods almost ignore the potential higher-order correlations between heterogeneous data, limiting the semantic representation capability. Secondly, it is susceptible to the imbalance in the distribution of semantic labels of samples and samples concept drifts [13], reduces the accuracy of cross-modal retrieval. Therefore, ensuring a higher correlation between sample pairs and preserving sufficient semantic information in a unified architecture is necessary to generate high-quality hashing codes.

In this paper, we propose a novel cross-modal hashing method called Joint-semantics Multi-Similarity Hashing (JMSH) for cross-modal retrieval. Our main contributions are summarized as follows:

- We design a joint-semantic similarity matrix to supervise the generation of hashing code, integrating image data features, text annotations features, and semantic label information mindfully to preserve higher-order semantic similarity in different modalities better.
- We propose an adaptive margin multi-similarity loss based on the prior similarity of samples, which reduces the influence of data imbalance and makes the generated hashing code more discriminative.
- Extensive experiments on two benchmark datasets demonstrate that the proposed JMSH significantly outperforms other cross-mode hashing methods.

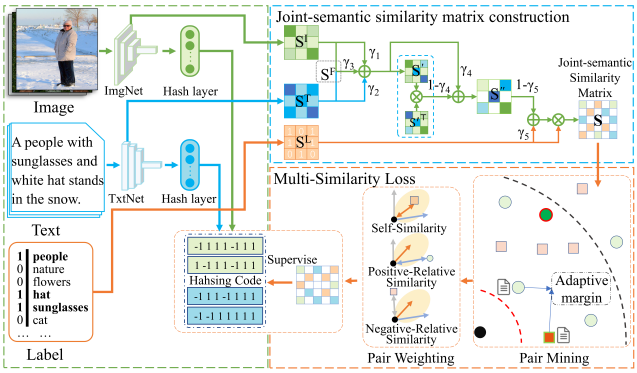

Fig. 1. The architecture of the proposed JMSH.

2. PROPOSED METHOD

The architecture of JMSH is shown in Fig. 1. For the training process, we use pre-trained AlexNet and BoW as the backbone networks for ImgNet and TxtNet, respectively, for feature extraction of original image and text data.

Let $\mathbf{O} = \{o_i\}_{i=1}^{n}$ denote a dataset with n training instances in a batch and $o_i = (v_i, t_i, l_i)$, where $\boldsymbol{v}_i \in \mathbf{R}^{n\times d_v}$, $t_i \in \mathbf{R}^{n\times d_t}$, $\boldsymbol{l}_i \in \{0,1\}^{n\times c}$ are the original image features, text features and label of i^{th} instance, respectively. d_v and d_t denote the dimensions of the original image features and text features, respectively. We use $\mathbf{V} = \{v_i\}_{i=1}^{n}$ and $\mathbf{T} = \{t_i\}_{i=1}^{n}$ to represent the image-feature and text-feature. Further, $l_i = [l_{i1}, l_{i2}, \ldots, l_{ic}]$ denotes the multi-label annotations assigned to instance o_i, where c is the number of classes. If o_i belongs to the k^{th} class, $l_{ik} = 1$, otherwise $l_{ik} = 0$. The goal of JMSH is to learn hashing function $H : O \rightarrow \{-1, +1\}^q$ that maps the input instance o_i to a q-bit binary code and maintains the semantic similarity from different modalities.

2.1. Joint-semantic similarity matrix construction

Similar to [14], the joint-semantic similarity matrix is to fuse the self-similarity matrix and the semantic labels from different modalities, to capture the high-order semantic correlation between different modality instances.

Firstly, we use $\mathbf{V} = \{v_i\}_{i=1}^{n}$ to construst image modality cosine similarity matrix: $\boldsymbol{S}^I = cos(v_i, v_j) = \frac{v_i v_j^\top}{\|v_i\|_2\|v_j\|_2} \in [-1,1]^{n\times n}$, as well, we adopt $\mathbf{T} = \{t_i\}_{i=1}^{n}$ to construct the text modality cosine similarity matrix: $\boldsymbol{S}^T = cos(t_i, t_j) = \frac{t_i t_j^\top}{\|t_i\|_2\|t_j\|_2} \in [-1,1]^{n\times n}$, where $||\cdot||_2$ denotes L_2 norm.

Then, we merge the image modality feature matrix $\boldsymbol{S}^I$ and text modality feature matrix $\boldsymbol{S}^T$, to construct a fusion similarity matrix $\boldsymbol{S}^F$, which can be expressed as:

$$\boldsymbol{S}^F = \eta \cdot \cos(\boldsymbol{S}^I, \boldsymbol{S}^T) + (1-\eta)\cdot \cos(\boldsymbol{S}^I, \boldsymbol{S}^T)^\top \quad (1)$$

where $\eta \in [0,1]$ is a trade-off parameter, following [15], we set η to 0.5.

Then, we fuse the $\boldsymbol{S}^I$, $\boldsymbol{S}^T$, and $\boldsymbol{S}^F$ with a weighted summation manner as follows:

$$S' = \gamma_1 S^I + \gamma_2 S^T + \gamma_3 S^F, \gamma_1 + \gamma_2 + \gamma_3 = 1 \quad (2)$$

where $\boldsymbol{S}'$ as a weighted similarity matrix, which records the similarity between the samples corresponding to each row and each column. γ_1, γ_2 and γ_3 are the trade-off parameters controlling the importance of the similarity information from the different modalities, respectively.

By the way, we calculate $S' \cdot (S')^\top$ to obtain the higher-order similarity and use γ_2 to adjust the relationship between the lower-order and higher-order similarity matrices.

$$S'' = \gamma_4 S' + (1-\gamma_4)\frac{S'\cdot(S')^\top}{n} \quad (3)$$

where n denotes the batch size used to normalize the higher-order similarity. γ_4 represents a parameter to balance the effects of different modalities.

To construct the similarity matrix, we should consider both high-level semantic information (i.e. semantic labels) and low-level information (i.e. multimodal features). Therefore, a combination function φ is introduced to generate the joint semantic similarity matrix $\boldsymbol{S}$, which is defined as:

$$\boldsymbol{S} = \varphi(\boldsymbol{S}^I, \boldsymbol{S}^T, \boldsymbol{S}^L, \boldsymbol{S}^F) = \boldsymbol{S}^L(\gamma_5 \cdot \boldsymbol{S}^L + (1-\gamma_5)\cdot \boldsymbol{S}'') \quad (4)$$

where γ_5 balances the importance of each parameter, $\boldsymbol{S}^L$ denotes the label-guided similarity matrices.

By combining the similarity information of different modalities into a unified matrix, the training process can fully excavate the high-order semantic similarity of potential modalities between instances, and complement the neighborhood relationship between the two modalities, which helps to generate higher-quality hashing codes.

Suppose b_i and b_j are both binary code from the given pair I_i and I_j, the hamming distance between b_i and b_j can be expressed as: $D_{Ham}(b_i, b_j) = \frac{1}{2}(k - \langle b_i, b_j\rangle)$ [16, 17], where $\langle\cdot,\cdot\rangle$ is inner product. In this work, we adopt inner product $\Theta(I_i, I_j) = \langle b_i, b_j\rangle$ to quantify the pairwise similarity.

On the basis of similarity quantization by inner product, the optimization similarity of pairwise is obtained by multiplication with joint-semantic similarity matrix: $\Omega(I_i, I_j) = S_{i,j}\Theta(I_i, I_j)$.

2.2. multi-similarity loss with adaptive margin

For cross-modal retrieval tasks, we add an adaptive margin into the General Pair Weighting (GPW) framework [18], which combines three types of similarities: self-similarity, positive relative similarity, and negative relative similarity, aiming to pair mining and pair weighting.

Pair Mining: As shown in Fig. 2, set I_i as an anchor, $\mathcal{P}_i$ and $\mathcal{N}_i$ represent the set of positive and negative of the I_i. A negative and positive pair $\{I_i, I_j\}$ are selected if the pair optimization similarity $\Omega(I_i, I_j)$ satisfies the following conditions: $\Omega^+(I_i, I_j) < max\big(\Omega(I_i, I_k)\big) + \epsilon$ and $\Omega^-(I_i, I_j) > min\big(\Omega(I_i, I_k)\big) - \epsilon$, where ϵ is a given margin.

To calculate the relative similarity between sample pairs,

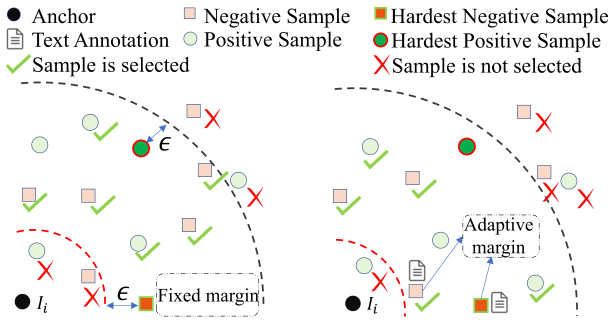

Fig. 2. Comparison of traditional pair mining and proposed methods. The left half employs a fixed margin, while the right half with an adaptive margin.

the given margin ϵ is usually fixed [19, 20], this way, the selection of sample pairs close to the margin of the similarity threshold is unfriendly. On the contrary, sample pairs with informative information should be selected heuristically to maintain the relative similarity between sample pairs dynamically. Similar to the previous work BLEU [21], we calculated the similarity margin ϵ_{ad} according to the prior similarity of the sample pairs, which used the similarity of text annotations $S^T(I_i, I_j)$ between the sample pairs to determine the distance between the sample pairs, so as to more finely distinguish the positive and negative samples with different similarity.

$$\epsilon_{ad} = \alpha \frac{-e^{\beta S^T(I_i,I_j)} + e^{\beta}}{-1 + e^{\beta}}, 0 < \alpha < 1 \tag{5}$$

where α is the maximum margin and β denoted the decay coefficient, therefore, the ϵ can be replaced by ϵ_{ad}.

Moreover, self-similarity is derived by directly computing pairwise similarity, which can be written as: $\varphi^S_{i,j} = \Omega(I_i, I_j)$.

Since negative relative similarity takes account of relative similarity between negative pair (I_i, I_j) and other remaining negative pairs, the other negative pair (I_i, I_k) need to be compared, and the relative similarity is defined, as shown below $\varphi^N_{ijk} = \Omega(I_i, I_k) - \Omega(I_i, I_j)$.

Pair Weighting: Following [18, 19], we introduce three hyperparameters μ, ρ, and λ into the measurement of self-similarity, relative similarity, and weight both similarities.

For the positive pair, the weights ω^+_{ij} for the positive pairs are calculated:

$$\omega^+_{ij} = \frac{1}{e^{-\mu(\lambda - \varphi^S_{ij})} + \sum_{k \in \mathcal{N}_i} e^{-\mu\left(\varphi^N_{ijk}\right)}} \tag{6}$$

Further, the Eq. (6) can be written as:

$$\omega^+_{ij} = \frac{e^{\mu\left(\lambda - \Omega(I_i,I_j)\right)}}{1 + \sum_{k \in \mathcal{P}_i} e^{(\mu(\lambda - \Omega(I_i,I_k)))}} \tag{7}$$

For the negative pair, weightes ω^-_{ij} can be written as:

$$\omega^-_{ij} = \frac{e^{\rho(\Omega(I_i,I_j) - \lambda)}}{1 + \sum_{k \in \mathcal{N}_i} e^{(\rho(\Omega(I_i,I_k) - \lambda))}} \tag{8}$$

Finally, the multi-similarity loss, shown as Eq. (9), the partial derivatives of the multi-similarity loss function respect to $\Omega^+(I_i, I_j)$ and $\Omega^-(I_i, I_j)$ is identical to the weight in Eq. (7) and Eq. (8) respectively. Though gradient descent optimization, loss could be minimized by exploiting the pair mining and weighting iteratively.

$$\mathcal{L}_{ms} = \sum_{i=1}^{m} \left\{ \frac{1}{\mu} \log \left[1 + \sum_{k \in P_i} e^{\mu(\lambda - S_{ik}\Theta(\boldsymbol{I}_i, \boldsymbol{I}_k))} \right] + \frac{1}{\rho} \log \left[1 + \sum_{k \in N_i} e^{\rho(S_{ik}\Theta(\boldsymbol{I}_i, \boldsymbol{I}_k) - \lambda)} \right] \right\} \tag{9}$$

Table 1 Mean Average Precision (MAP) comparison results.

Task	Method	MIRFLICKR-25K			NUS-WIDE		
		16-bits	32-bits	64-bits	16-bits	32-bits	64-bits
I→T	DCMH [3]	0.7410	0.7465	0.7485	0.5903	0.6031	0.6093
	SSAH [8]	0.7890	0.8005	0.8060	0.6160	0.6360	0.6370
	CMHH [9]	0.7334	0.7281	0.7444	0.5530	0.5698	0.5559
	AGAH [10]	0.7923	0.7945	0.8069	0.6455	0.6600	0.6512
	SRLCH [11]	0.7300	0.7289	0.7713	0.6529	**0.6658**	0.6690
	CMMQ [12]	-	0.7370	0.7420	-	0.6010	0.6060
	JMSH	**0.8145**	**0.8219**	**0.8301**	**0.6611**	0.6630	**0.6751**
T→I	DCMH [3]	0.7825	0.7900	0.7932	0.6398	0.6511	0.6571
	SSAH [8]	0.7820	0.7970	0.7990	0.6530	**0.6760**	0.6830
	CMHH [9]	0.7320	0.7183	0.7279	0.5739	0.5786	0.5639
	AGAH [10]	0.7887	0.7904	0.8049	0.6313	0.6422	0.6336
	SRLCH [11]	0.7111	0.7120	0.7527	0.6108	0.6231	0.6254
	CMMQ [12]	-	0.7230	0.7250	-	0.6000	0.6040
	JMSH	**0.7985**	**0.8011**	**0.8135**	**0.6656**	0.6735	**0.6841**

3. EXPERIMENTS

3.1. Experimental Setting

We evaluated the proposed JMSH on MIRFlickr-25K [22] and NUS-WIDE [23] datasets. MIRFlickr-25K contains 25,000 multi-label images and 24 classes, while NUS-WIDE has 269,648 multi-label images and 81 classes. Each image is tagged with at least one semantic label.

Two cross-modal retrieval tasks are performed, including image-to-text retrieval (I→T) and text-to-image retrieval (T→I). Mean Average Precision (MAP) and Precision-Recall (PR) are used to evaluate the retrieval performance. For all metrics, a higher value indicates better retrieval performance.

We set the batch size as 256 and employ Adam optimizer to optimize the ImgNet and TxtNet. The learning rate of ImgNet and TxtNet are both set to 1e-4. Besides, we cross-validate the parameters γ_1, γ_2, γ_3, γ_4, γ_5, μ, ρ, λ and set $\{\gamma_1 = 0.6, \gamma_2 = 0.2, \gamma_3 = 0.2, \gamma_4 = 0.7, \gamma_5 = 0.4, \mu = 0.8, \rho = 0.6, \lambda = 0.4\}$.

3.2. Comparison

We compare six state-of-the-art cross-modal hashing methods as baselines, including DCMH [3], SSAH [8], CMHH [9], AGAH [10] , SRLCH [11] and CMMQ [12].

Table 1 shows the performance comparison of all the methods, we can observe that the JMSH almost outperforms

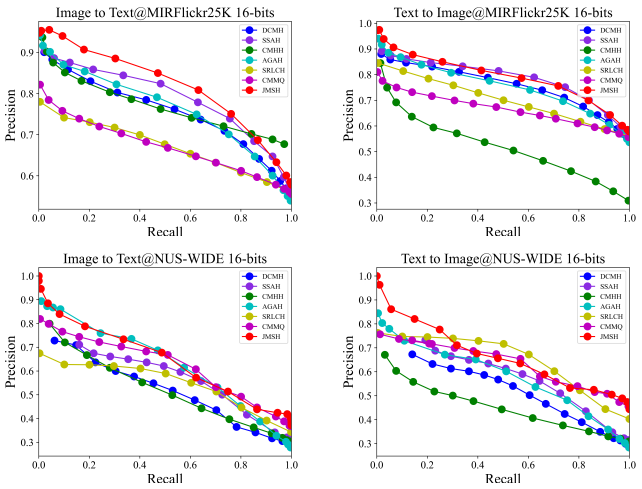

Fig. 3. Precision-Recall curves with 16 bits on two datasets.

all the baselines, achieving better MAP results on different hashing codes for all datasets. This is because the proposed JMSH combines multi-modal features and semantic label information, effectively preserving higher-order similarity relationships within multi-modal data and reducing modality gaps. Furthermore, the multi-similarity loss mitigates training issues arising from sample imbalances, optimizing the training process and enhancing cross-modal retrieval performance. Notably, for NUS-WIDE, a complex dataset containing more training samples, the JMSH can still achieve significant results in all situations.

We plot the Precison-Recall curve with code length 16 among the compared methods on two datasets. As we can from Fig. 3, our proposed JMSH achieves the best precision results at all recall levels, which further proves the superiority of JMSH in the tasks of supervised cross-modal retrieval. In addition, the results of other code lengths (32 bits and 64 bits) can always obtain a satisfying performance on two datasets.

3.3. Ablation Studies

In this section, we conduct ablation studies to evaluate the effectiveness of various components. JMSH-1 stands for an architecture with labels, text and image fusion similarity $\boldsymbol{S} = \varphi(\boldsymbol{S}^I, \boldsymbol{S}^T, \boldsymbol{S}^L)$ without higher-order semantic features $\boldsymbol{S} = \varphi(\boldsymbol{S}^I, \boldsymbol{S}^T, \boldsymbol{S}^L, \boldsymbol{S}^F)$. JMSH-2 uses multi-similarity loss with a fixed margin instead of an adaptive margin.

As shown in Table 2, the results of JMSH-1 suggests that the joint semantic similarity matrix can effectively fuse important modality information and bridge the modality gap. Meanwhile, the results of JMSH-2 show that the multi-similarity loss with adaptive margin has better performance than the conventional multi-similarity loss function, which is due to its ability to distinguish samples close to the margin of the similarity threshold more finely through the improved sampling and weighting scheme.

Table 2 MAP results of proposed JMSH and its variants.

Task	Method	MIRFlickr-25K			NUS-WIDE		
		16-bits	32-bits	64-bits	16-bits	32-bits	64-bits
I→T	JMSH-1	0.7580	0.7638	0.7529	0.5866	0.5942	0.6110
	JMSH-2	0.8085	0.8118	0.8200	0.6379	0.6435	0.6479
	JMSH	**0.8145**	**0.8219**	**0.8301**	**0.6611**	**0.6630**	**0.6751**
T→I	JMSH-1	0.7281	0.7910	0.7813	0.6418	0.6571	0.6630
	JMSH-2	0.7905	0.7813	0.8042	0.6564	0.6600	0.6713
	JMSH	**0.7985**	**0.8011**	**0.8135**	**0.6656**	**0.6735**	**0.6841**

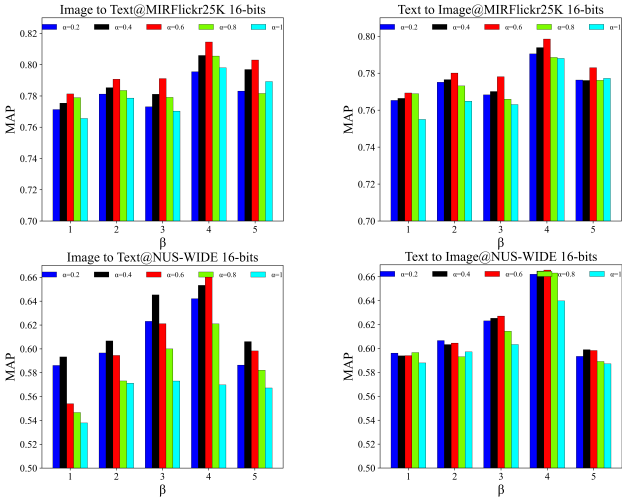

Fig. 4. MAP results of different maximum margin α and decay coefficient β for JMSH on two datasets with 16-bits.

3.4. Sensitivity to the Hyper-Parameters

For the multi-similarity loss with adaptive margin, the maximum margin α and the attenuation coefficient β are determined according to the prior similarity of the sample pairs, which can reflect the semantic relationship between the different modalities. Specifically, we set $\alpha = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and $\beta = \{1, 2, 3, 4, 5\}$, respectively. The effect of two hyperparameters on multi-similarity loss of adaptive margin is investigated. As we can see from the Fig. 4, the multiple similarity loss of our design adaptive margin is the most useful predictor of image similarity when $\alpha = 0.6$ and $\beta = 4$.

4. CONCLUSION

We propose a novel cross-modal hashing retrieval method called JMSH. JMSH combines different modalities' features and semantic labels to capture the higher-order semantic relationships between each modality, for supervised learning of compact hashing code. Meanwhile, we design a multi-similarity loss based on an adaptive margin, which jointly considers self-similarity, positive, and negative relative similarity to mitigate the accuracy loss due to sample imbalance. Extensive experiments on two benchmark datasets demonstrate that the JMSH outperforms other baselines.

5. REFERENCES

[1] Xin Luo, XiaoYa Yin, Liqiang Nie, Xuemeng Song, Yongxin Wang, XinShun Xu, et al., "Sdmch: Supervised discrete manifold-embedded cross-modal hashing," in *IJCAI*, 2018, pp. 2518–2524.

[2] Jianyang Qin, Lunke Fei, Zheng Zhang, Jie Wen, Yong Xu, and David Zhang, "Joint specifics and consistency hash learning for large-scale cross-modal retrieval," *TIP*, vol. 31, pp. 5343–5358, 2022.

[3] QingYuan Jiang and WuJun Li, "Deep cross-modal hashing," in *CVPR*, 2017, pp. 3232–3240.

[4] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *CVPR*, 2017, pp. 4076–4084.

[5] Xianglong Liu, Lei Huang, Cheng Deng, Bo Lang, and Dacheng Tao, "Query-adaptive hash code ranking for large-scale multi-view visual search," *TIP*, vol. 25, no. 10, pp. 4514–4524, 2016.

[6] Ting Zhang and Jingdong Wang, "Collaborative quantization for cross-modal similarity search," in *CVPR*, 2016, pp. 2036–2045.

[7] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *TIP*, vol. 26, no. 5, pp. 2494–2507, 2017.

[8] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *CVPR*, 2018, pp. 4242–4251.

[9] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang, "Cross-modal hamming hashing," in *ECCV*, 2018, pp. 202–218.

[10] Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *ICMR*, 2019, pp. 159–167.

[11] HengTao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *TKDE*, vol. 33, no. 10, pp. 3351–3365, 2020.

[12] Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng, "Mutual quantization for cross-modal search with noisy labels," in *CVPR*, 2022, pp. 7541–7550.

[13] Hang Yu, Tianyu Liu, Jie Lu, and Guangquan Zhang, "Automatic learning to detect concept drift," *CoRR*, vol. abs/2105.01419, 2021.

[14] Shupeng Su, Zhisheng Zhong, and Chao Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *ICCV*, 2019, pp. 3027–3035.

[15] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *SIGIR*, 2020, pp. 1379–1388.

[16] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao, "Deep hashing network for efficient similarity retrieval," in *AAAI*, 2016, pp. 2415–2421.

[17] Yue Cao, Mingsheng Long, Jianmin Wang, Han Zhu, and Qingfu Wen, "Deep quantization network for efficient image retrieval," in *AAAI*, 2016, pp. 3457–3463.

[18] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *CVPR*, 2019, pp. 5017–5025.

[19] Jiawei Zhan, Song Liu, Zhaoguo Mo, and Yuesheng Zhu, "Multi-similarity semantic correctional hashing for cross modal retrieval," in *ICME*, 2020, pp. 1–6.

[20] Qibing Qin, Lintao Xian, Kezhen Xie, Wenfeng Zhang, Yu Liu, Jiangyan Dai, and Chengduan Wang, "Deep multi-similarity hashing with semantic-aware preservation for multi-label image retrieval," *Expert Systems with Applications*, vol. 205, pp. 117674, 2022.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[22] Mark J Huiskes and Michael S Lew, "The mir flickr retrieval evaluation," in *ICMR*, 2008, pp. 39–43.

[23] TatSeng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *CIVR*, 2009, pp. 1–9.