SEMI-SUPERVISED GRAPH CONVOLUTIONAL HASHING NETWORK FOR LARGE-SCALE CROSS-MODAL RETRIEVAL

Zhanjian Shen, Deming Zhai†, Xianming Liu, Junjun Jiang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, 150001

ABSTRACT

Cross-modal retrieval aims to provide flexible retrieval results across different types of multimedia data. To confront with scalability issue, binary codes learning (*a.k.a.* hash technique) is advocated since it permits exact top-K retrieval with sub-linear time complexity. In this paper, we propose a new method called *Semi-supervised Graph Convolutional Hashing* network (SGCH), which tries to learn a common hamming space by preserving both intra-modality and inter-modality similarities via an end-to-end neural network. On one hand, graph convolutional network is utilized to explore high-order intra-modality similarity, and simultaneously propagate the semantic information from labeled samples to unlabeled data. On the other hand, a siamese network is connected to project the learnt features into a common hamming space. To bridge the inter-modality gap, adversarial loss which aims to learn modality-independent features by confusing a modality classifier is incorporated into the overall loss function. Experimental evaluations on cross-media retrieval tasks demonstrate that SGCH performs competitively against the state-of-the-art methods.

***Index Terms*—** cross-modal retrieval, graph convolutional network, semi-supervised hash learning

1. INTRODUCTION

With the rapid development of Internet and smart devices, huge amounts of multimedia data, such as images, texts are produced continuously. To provide flexible retrievals across different types of data, cross-modal retrieval has attracted considerable interest recently. For example, one may query image datasets by text keywords to quickly draw a vivid imagination, or query text datasets by images to accurately describe the details [1]. Cross-modal retrieval is a challenging problem, since data from different modalities typically have different statistical properties and are improper to compare directly, which is usually referred to as heterogeneity gap. To solve this problem, most of methods (*e.g.* CCA [2]) try to project data from different modalities into a common space to measure their similarities [3, 4]. With the emerging of big data, existing cross-modal retrieval methods usually suffer from serious high computation and storage cost problem.

To confront with scalability issue, some researchers advocate the use of discrete encodings (*a.k.a.*, hash technique), from real-valued vectors into compact binary codes. In this way, similarities between data could be efficiently measured by bit-wise XOR operations in discrete space, *i.e.* hamming space, with sub-linear time complexity. Roughly speaking, existing Cross-Modal Hashing (CMH) approaches [5, 6, 7, 8, 9, 10] can be divided into three categories: supervised, unsupervised and semi-supervised. Supervised CMH methods generally utilize semantic information, such as class labels, to help learn more discriminative hash codes [11, 12, 13]. For example, semantic correlation maximization (SCM) method [14] optimizes hash functions by maximizing the correlation between two different modalities based on class labels. Deep cross-modal hashing (DCMH) approach [4] integrates feature learning and hash codes learning into an end-to-end deep neural network with the supervision of class labels. Self-supervised adversarial hashing (SSAH) [15] incorporates adversarial networks to bridge the modality gap. In contrast, unsupervised CMH methods aim to learn hash codes for different modalities without any labels.

Because labeling data requires tedious human efforts, semi-supervised CMH methods which leverage both labeled and unlabeled data become more practical in real applications. Most semi-supervised cross-modal retrieval methods usually utilize graph-based regularization to learn from unlabeled data. For example, joint representation learning (JRL) [16] learns sparse projections for different modalities by exploring the geometric structure of both labeled and unlabeled data through a graph regularization. Generalized semi-supervised and structured subspace learning (GSS-SL) [17] adopts a graph-based constraint to ensure the intrinsic geometric structures of different feature spaces consistent with that of label space. In a nutshell, existing semi-supervised CMH methods are mainly based on graph-based regularization for data structure preserving, ignoring to explore and preserve high-order intra-modality similarity of each modality for hash codes learning in the common hamming space.

In this paper, we propose a new method called Semi-supervised Graph Convolutional Hashing (SGCH) network

†This work was supported by Natural Science Foundation of China under Grant No. 61502122, 61922027, and 61971165, and China Postdoctoral Science Foundation funded project 2018M630360. Corresponding author: Deming Zhai, e-mail: zhaideming@hit.edu.cn

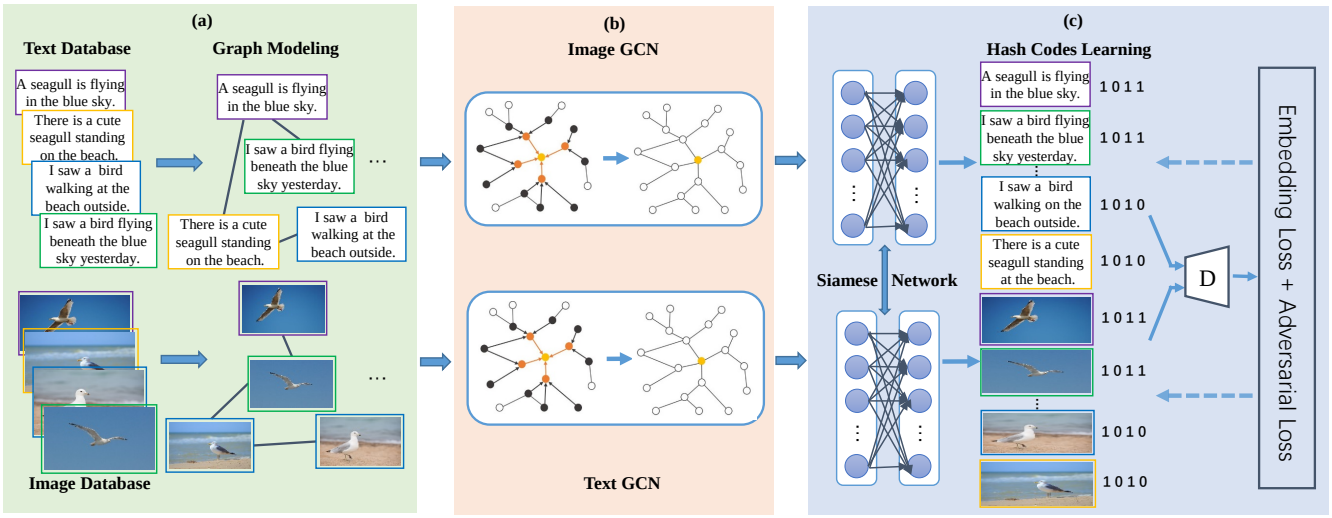

Fig. 1. The framework of proposed semi-supervised graph convolutional hashing network (SGCH), which contains three phases: (a) Graph modeling for each dataset according to features and semantic labels for both labeled and unlabeled data; (b) Feature learning for each modality via graph convolutional network (GCN) to explore and preserve the intra-modality similarity; (c) Hash codes learning by projecting data from different modalities into a common hamming space, such that the inter-modality similarity are preserved. The parameters in the whole network are optimized in an adversarial way through back propagation.

to improve large-scale cross-model retrieval performance. More specifically, we first conduct feature learning via graph convolutional network to explore high-order intra-modality similarity with both labeled and unlabeled data for each modality. Then a siamese network is connected to project the learnt features into a common hamming space for hash codes learning. To bridge the inter-modality gap, adversarial loss which aims to learn modality-independent features by confusing a modality classifier is incorporated into the overall loss function. And all parameters in the network are optimized in an adversarial way through back propagation. The main contributions of SGCH are highlighted as follows: 1) We explore the high-order intra-modality similarity for feature learning via a two-layer graph convolutional network in a semi-supervised setting. 2) Both the intra-modality similarity and inter-modality similarity are considered and preserved in the learned common hamming space via an end-to-end neural network learning. 3) Extensive results on large-scale multimedia datasets demonstrate the effectiveness of SGCH.

2. THE PROPOSED METHOD

2.1. Problem Formulation

Suppose there are two data sets from different modalities: $\mathbf{X} = [\mathbf{x}_1; ...; \mathbf{x}_n] \in \mathbb{R}^{n \times d_x}$, $\mathbf{Y} = [\mathbf{y}_1; ...; \mathbf{y}_n] \in \mathbb{R}^{n \times d_y}$ for n images and corresponding textual documents. Each data point $\mathbf{x}_i(\mathbf{y}_j)$ is located in the $d_x(d_y)$ dimensional input feature space. Assume that the first n_l samples of $\mathbf{X}$ and $\mathbf{Y}$ have labels from c classes, and the rest are unlabeled data. The label matrix is denoted as: $\mathbf{L} = [\mathbf{l}_1; ...; \mathbf{l}_{n_l}] \in \{0, 1\}^{n_l \times c}$, where the p-th entry $\mathbf{l}_{ip} = 1$ if $(\mathbf{x}_i, \mathbf{y}_i)$ belongs to the p-th class, and zero otherwise. Then the goal of cross-modal hashing is to learn discrete k-bit hash codes $\mathbf{B}^x$ and $\mathbf{B}^y \in \{0, 1\}^{n \times k}$ by projecting data from different modalities into a common hamming space.

2.2. The Framework

In devising effective CMH algorithm in a semi-supervised setting, two important issues should be taken into consideration: (1) The inter-modality relationship should be established, such that pairs with the same semantic concept should be mapped into the same hash bin, and vice versa. (2) The intra-modality similarities should be explored and preserved with both labeled and unlabeled data. For each modality, data with high similarity should be enforced to have similar binary codes in the hamming space.

We propose to achieve the overall objectives in an end-to-end deep learning framework. As shown in Fig. 1, our method contains three phases: (a) Graph modeling for each dataset; (b) Feature learning via graph convolutional network (GCN) [18, 19] for each modality ; (c) Hash codes learning.

Specifically, we first build graphs $\mathcal{G}^x = (V^x, E^x)$ and $\mathcal{G}^y = (V^y, E^y)$ for image and text datasets, respectively. Therein, V^x (V^y) is the vertex set corresponding to all data samples of $\mathbf{X}$ ($\mathbf{Y}$), and the edge set E^x (E^y) represents m nearest neighbors' connection for data in each modality. According to features and available class labels, we define the weight matrix $\mathbf{W}^*$ on edges to reflects the intra-modality affinities for each graph as below, where $* \in \{\mathbf{X}, \mathbf{Y}\}$:

$$\mathbf{W}^x_{ij} = \begin{cases} \mathbf{l}_i^T \mathbf{l}_j & , \text{if } \|\mathbf{l}_i\|^2 \neq 0 \text{ and } \|\mathbf{l}_j\|^2 \neq 0 \\ \dfrac{\mathbf{x}_i \cdot \mathbf{x}_j}{||\mathbf{x}_i|| \, ||\mathbf{x}_j||} & , \text{if } \mathbf{x}_i \in \mathcal{N}_m(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_m(\mathbf{x}_i) \\ 0 & , \text{otherwise} \end{cases} \tag{1}$$

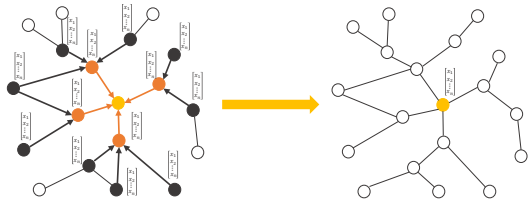

Fig. 2. Illustration of a two-layer graph convolution network, where two-hop neighbors are all involved to learn the central node feature.

where $\mathcal{N}_m(\mathbf{x}_i)$ denotes the set of m nearest neighbors of $\mathbf{x}_i$. And $\mathbf{W}^y$ can be defined in a similar way with both labeled and unlabeled data.

Then, we further conduct feature learning for each modality via the cutting-edge technique: *Graph Convolutional Network* (GCN) [19] to preserve the intra-modality similarity. Specifically, $\mathcal{G}^x$ ($\mathcal{G}^y$) is fed into image (text) GCN to learn the transformed features $\mathbf{F}^x$ ($\mathbf{F}^y$) $\in \mathbb{R}^{n\times d_f}$, where d_f is the dimension in the common space for both modalities. According to spectral graph theory, the convolution operator on graph is defined by aggregating and weighting sample features from its local neighborhoods based on intra-modality affinity matrix $\mathbf{W}^*$. As illustrated in Fig. 2, a two-layer graph convolution network could learn features across far reaches of a graph, which is formulated as:

$$\mathbf{F}^{x(0)} = tanh\left(\mathbf{A}\mathbf{X}\theta_x^{(0)}\right), \quad \mathbf{F}^x = tanh\left(\mathbf{A}\,\mathbf{F}^{x(0)}\,\theta_x^{(1)}\right). \tag{2}$$

Here, $\mathbf{A} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{W}^x + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$ is the normalized intra-modality similarity matrix, with $\mathbf{I}$ being the identity matrix, and $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}^x$ being the degree matrix. $\theta_x^{(0)}, \theta_x^{(1)}$ are the two-layer GCN parameters to be optimized.

Intuitively, graph convolution operator can be viewed as a weighted averaging of the features from each node's local neighbors, followed by a nonlinear function. As a result, data with high similarity will have similar feature representations. By stacking multiple convolutional layers, GCN could explore and preserve the high-order intra-modality similarity, and label propagation is carried out at the same time. Compared to traditional graph-based regularization methods [7, 16, 17], feature learning via GCN for each modality could better explore the intra-modality affinities, which will benefit subsequent cross-modal retrieval tasks.

Finally, a siamese network depicted as $h(\cdot)$ is connected to project the learned features $\mathbf{F}^x$ and $\mathbf{F}^y$ into a common hamming space to establish their cross-modal relationship, which could be expressed as:

$$\mathbf{B}^* = sgn(h(\mathbf{F}^*; \theta_s)), * \in \{\mathbf{X}, \mathbf{Y}\} \tag{3}$$

where $sgn(\cdot)$ denotes the elementwise sign function[1], $\mathbf{B}^*$ is the learned k-bit hash codes, and θ_s denotes the parameters to be optimized in the siamese network.

[1]Here we generate hash bits as $\{-1, 1\}$, which are straightforward to convert to $\{0, 1\}$ valued hash codes.

2.3. Optimization

To establish the inter-modality relationship, the loss function of proposed end-to-end deep network consists of two parts:

The first part is embedding loss, which is defined as:

$$\begin{aligned} L_{emb} = &-\sum_i \sum_j \mathbf{S}_{ij}^{xy} . \times \mathbf{\Psi}_{ij} + \alpha\left(\|\mathbf{B}^x\mathbf{1}\|_F^2 + \|\mathbf{B}^y\mathbf{1}\|_F^2\right) \\ &+ \beta\left(\|\mathbf{H}^x - \mathbf{B}^x\|_F^2 + \|\mathbf{H}^y - \mathbf{B}^y\|_F^2\right), \end{aligned} \tag{4}$$

where $\mathbf{H}^*$ is a real-value relaxation of $\mathbf{B}^*$ by discarding sign function for ease of optimization. The first term is to preserve the cross-modal similarity by maximizing the correlation between the ground-truth inter-modality similarity $\mathbf{S}_{ij}^{xy}$ and calculated similarity in hamming space $\mathbf{\Psi}_{ij} = (\mathbf{H}_i^x)^T \mathbf{H}_j^y$. The second and third terms are used to make each bit of hash codes have equal 1 and -1 for balance, where $\mathbf{1}$ is a constant vector with all ones. And the last two terms tries to minimize quantization errors between the binary hash codes and the relaxed real-value embeddings. α and β are regularization parameters to balance these three kinds of loss terms.

The second part is adversarial loss [20, 21], which is added to further alleviate the heterogeneity gap. Specifically, we aim to learn modality-independent features in the common space, in order to confuse a modality classifier D. The modality classifier D with parameter θ_D tries to classify between different modalities, with output '1' for image modality and '0' for text modality. Thus, the adversarial loss could be defined as the cross-entropy of classification, which is expressed as:

$$L_{adv} = -\frac{1}{n}\sum_{i=1}^{n} log\left(D\left(\mathbf{H}_i^x; \theta_D\right)\right) + log\left(1 - D\left(\mathbf{H}_i^y; \theta_D\right)\right). \tag{5}$$

The total loss function is then defined as a combination of adversarial loss and embedding loss. Inspired by GAN [20], we optimize the total loss as a minimax game, which can be formulated as:

$$\begin{aligned} (\theta_x, \theta_y, \theta_s) &= \underset{\theta_x,\theta_y,\theta_s}{argmin}\left(L_{emb}\left(\theta_x, \theta_y, \theta_s\right) - L_{adv}\left(\hat{\theta}_D\right)\right), \\ (\theta_D) &= \underset{\theta_D}{argmax}\left(L_{emb}\left(\hat{\theta}_x, \hat{\theta}_y, \hat{\theta}_s\right) - L_{adv}\left(\theta_D\right)\right). \end{aligned} \tag{6}$$

The parameters of all above in the deep network are updated alternatively using stochastic gradient descent algorithm through back propagation process.

3. EXPERIMENTS

3.1. Datasets

NUS-WIDE-10K dataset consists of 8000/2000 (training/testing) image-text pairs, which are randomly selected from 10 largest categories of NUS-WIDE [22] dataset. Each text is represented by an 1,000-dimensional bag-of-words

Table 1. The mAP results in a semi-supervised setting. The percentage of labeled data number is varying from 60% to 100%, and the hash code length equals to 10.

Task	Method	NUS-WIDE-10K					Wiki				
		60%	70%	80%	90%	100%	60%	70%	80%	90%	100%
Image Query v.s. Text Database	JRL	0.573	0.573	0.576	0.578	0.580	0.504	0.509	0.509	0.513	0.516
	GSS-SL	0.487	0.492	0.491	0.496	0.500	0.463	0.478	0.492	0.498	0.500
	SGCH-v1	0.552	0.591	0.625	0.658	0.688	**0.514**	0.550	0.658	0.675	0.707
	SGCH	**0.584**	**0.613**	**0.628**	**0.689**	**0.691**	0.483	**0.570**	**0.681**	**0.737**	**0.747**
Text Query v.s. Image Database	JRL	0.473	0.492	0.493	0.496	0.497	0.407	0.406	0.408	0.413	0.417
	GSS-SL	0.441	0.450	0.451	0.454	0.460	0.420	0.430	0.444	0.447	0.450
	SGCH-v1	0.588	**0.641**	**0.646**	**0.670**	**0.711**	0.542	0.624	0.707	**0.717**	0.727
	SGCH	**0.608**	0.627	0.640	0.666	0.710	**0.583**	**0.638**	**0.708**	0.691	**0.755**

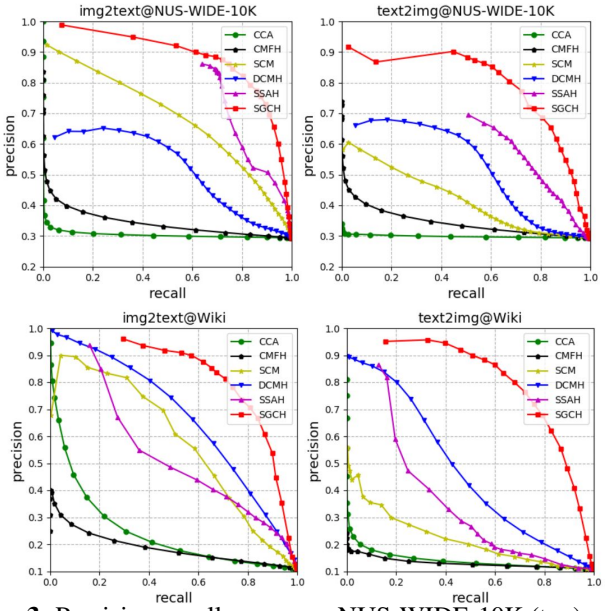

Fig. 3. Precision-recall curves on NUS-WIDE-10K (top) and Wiki (bottom) datasets with hash code length equals to 32.

vector, and each image is a 4,096d vector extracted by the fc7 layer of VGGNet [23].

Wiki dataset is collected from Wikipedia articles [24], which contains 2173/693 pairs with 10 semantic classes. We adopt the 4,096d vector of the fc7 layer of VGGNet as the image feature and 3000d bag-of-words vector as text feature.

3.2. Comparison with the state-of-the-arts

To verify the effectiveness of SGCH, seven state-of-the-art cross-modal retrieval methods are compared as our baselines: a) unsupervised methods: CCA [2] and CMFH [9]. b) supervised methods: SCM [14], DCMH [4] and SSAH [15]. c) semi-supervised methods: JRL [16] and GSS-SL [17].[2] For fairness, the number of the training samples is equivalent for all compared methods. In our method, the hyper-parameters α and β are set to 1 and 0.01, respectively. We employ ADAM [25] optimiser with a learning rate of 10^{-3} and the maximal number of epochs as 500. The performance is evaluated by mean Average Precision (mAP) and precision-recall curves.

First, we conduct experiments when all training data has class labels (*i.e. supervised setting*). The results of precision-recall curves on both datasets are illustrated in Figrue 3.[3] It can be observed that CCA and CMFH, which don't utilize any label information, yield poor performance. Among all comparative studies, SGCH achieves the best overall performance. Especially for the task of text query v.s. image database retrieval, the improvements over the state-of-the-art algorithms are significant, with a performance gain of more than 5%. It verifies the effectiveness of jointly preserving the intra- and inter-modality similarities for cross-modality hash learning in an end-to-end deep neural network.

Then, we further conduct experiments when only a portion of training data has class labels (*i.e. semi-supervised setting*). The mAP results compared to state-of-the-art semi-supervised methods are summarized in Table 1, where the percentage of labeled data number is varying from 60% to 100%. In this scenario, we also compare a variant of SGCH called SGCH-v1, with only embedding loss in optimization. From the results, we can see that both SGCH and SGCH-v1 get higher mAP values than traditional graph-based regularization methods JRL and GSS-SL. It demonstrates the effectiveness of graph convolutional network in intra-modality similarity exploring. Besides, SGCH has smaller divergences between image-to-text and text-to-image retrieval tasks than SGCH-v1, which indicates that adversarial loss is helpful to further alleviate the modality gap and get more balanced results on both image-to-text and text-to-image retrieval tasks.

4. CONCLUSION

In this paper, we propose a new semi-supervised cross-modal hashing method called SGCH, which utilize graph convolutional network to preserve the intra-modal similarity and propagate semantic information from labeled samples to unlabeled data. Moreover, an adversarial loss is incorporated to achieve distribution agreement between different modalities, which is helpful to bridge inter-modality gap.

[2]Source codes of these baselines are kindly provided by authors, and we fine-tune the hyper-parameters of these methods to achieve their best results.

[3]Due to space limitation, only the results with hash code length equals to 32-bit are reported.

5. REFERENCES

[1] Yuxin Peng, Xin Huang, and Yunzhen Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," in *Transactions on Circuits and Systems for Video Technology*. IEEE, 2017.

[2] Harold Hotelling, "Relations between two sets of variates," in *Biometrika*, 1936, pp. 321–377.

[3] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng, "Deep supervised cross-modal retrieval," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[4] Qing-Yuan Jiang and Wu jun Li, "Deep cross-modal hahsing," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[5] Xuanwu Liu, Guoxian Yu, Carlotta Domenicon, and Jun Wang, "Ranking-based deep cross-modal hashing," in *The 33th AAAI Conference on Artificial Intelligence*, 2019.

[6] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu, "Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval," in *The 33th AAAI Conference on Artificial Intelligence*, 2019.

[7] En Yu, Jiande Sun, Jing Li, Xiaojun Chang, and Xian-Hua Han, "Adaptive semi-supervised feature selection for cross-modal retrieval," in *Transactions on Multimedia*. IEEE, 2019, pp. 1276 – 1288.

[8] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu, "Graph convolutional network hashing for cross-modal retrieval," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 982–988.

[9] Guiguang Ding, Yuchen Guo, and Jile Zhou, "Collective matrix factorization hashing for multimodal data," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[10] Jian Zhang, Yuxin Peng, and Mingkuan Yuan, "Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network," in *Transactions on Cybernetics*. IEEE, 2018.

[11] Ding Guiguang Hu Mingqing Lin, Zijia and Jianmin Wang, "semantics-preserving hashing for cross-view retrieval," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015.

[12] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang, "Deep cauchy hashing for hamming space retrieval," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[13] Tianyi Chen, Lan Zhang, Shicong Zhang, Zilong Li, and Baichuan Huang, "Extensible cross-modal hashing," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[14] Dongqing Zhang and Wu-Jun Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014, p. 2177–2183.

[15] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[16] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," in *Transactions on Circuits and Systems for Video Technology*. IEEE, 2014, pp. 965–978.

[17] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," in *Transactions on Multimedia*. IEEE, 2018.

[18] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," in *Signal Processing Magazine*. IEEE, 2017.

[19] Kipf Thomas N. and Welling Max, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, and Bing Xu, "Generative adversarial nets," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, p. 2672–2680.

[21] Bokun Wang, Yang Yang, and Xing Xu, "Adversaial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 152–162.

[22] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "Nus-wide: a realworld web image database from national university of singapore," in *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*. ACM, 2009.

[23] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[24] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, R. G. Gert Lanckriet, Levy Roger, and Vasconcelos Nuno, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th International Conference on Multimedia*, 2010.

[25] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations*, 2014.