Supplementary Material for "From Sparse to Dense: Semantic Graph Evolutionary Hashing for Unsupervised Cross-Modal Retrieval"

Yang Zhao[1], Jiaguo Yu[1], Shengbin Liao[2], Zheng Zhang[3], and Haofeng Zhang[1(✉)]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.
{zhao_yang,yujiaguo,zhanghf}@njust.edu.cn
[2] National Engineering Research Center for E-learning, Huazhong Normal University, Wuhan 430079, China.
[3] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, 518055, China.

1 Related work

In this section, we will introduce recent cross-modal hashing strategies, which can be roughly categorized into supervised methods and unsupervised methods according to whether they employ labelled information in training process.

Supervised cross-modal hashing. Recent years have witnessed the great success of supervised cross modal hashing [12,26,18,1,2,15,20,34,24,29,32]. Zhang *et al.* [42] employed the labeled information to restrict the learning of hash codes, which have shown great success in cross-modal retrieval task. SDMH [23] proposes an efficient augmented Lagrangian multiplier (ALM) based discrete hash optimization method to optimize the hash code learning process. Hu *et al.* [12] utilized an efficient iterative algorithm for hash code optimization by incorporating the code balance and uncorrelation criteria into the objective function. Mandal *et al.* [26] learned the optimum hash codes for the two modalities at the same time, by preserving the semantic similarity between the data points, and then learned the hash functions to map the features into the hash codes. Supervised Robust Discrete Multimodal Hashing (SRDMH) adopts a flexible loss with nonlinear kernel embedding [18] . [6] designs a two-stream ConvNet architecture to learn hash codes with class-specific representation centers for image retrieval. [5] focus on zero-shot sketch-based image retrieval task and proposes an method aligning the sketch and image features to semantic features. Xie *et al.* [36] designed Multi-Task Consistency-Preserving Adversarial Hashing (CPAH) composing of a consistency refined module (CR) and a multi-task adversarial learning module (MA). CR divides the representations of different modality into two irrelevant parts and MA constrains the modality-common representation of different modalities close to each other on feature distribution and semantic consistency.

Unsupervised cross-modal hashing. A large amount of unsupervised cross-modal hashing [11,8,40,10,17,37,41,28,22,43,35] have been proposed in the past few years. The earlier shallow schemes, *e.g.*, both Cross-view hashing (CVH) [16] and Inter-Media Hashing (IMH) [30], can be viewed as the extension of Spectral Hashing [33] from single-modal hashing to cross-modal hashing scenario. These methods restrict hash codes by solving the eigenvalue decomposition with constructed affinity graph. Employing matrix factorization methods, Collective Matrix Factorization Hashing (CMFH) [7] bridges the modality gap by embedding different modal information into a latent common space. Latent Semantic Sparse Hashing (LSSH) [44] extends CMFH to utilizing sparse coding in extracting latent feature process of both two modals at the same time. And subsequently LSSH employs the latent features to restrict hash code learning. However, above shallow methods are difficult to extract the heterogeneous relationships from different modalities for using hand-crafted features. As the progress of deep neural networks have made in exploring nonlinear relationships, many methods [4,40,39,14] capture more semantic relevant features to learn hash code in an end-to-end training model. Most of them utilize similarity graphs generated from intrinsic data directly and obtain superior performances in some cross-modal retrieval tasks. [9] utilizes the adaptive tanh function which has concise derivation and can be used in objective function directly. [35] makes use of the matrix factorization with Laplacian constraint in training process to constraint the hash code generation, which consequently preserves the neighbor affinity information of original features in their own space. Liu *et al* [21] designed Joint-modal Distribution-based Similarity Hashing (JDSH), which capture the joint similarity matrix for information maintain. [41] utilized three types of similar information and keep real value in optimization for hash code learning. [38] aligns the feature between the visual information and textual information.

Though impressive progress have these models made, there are still a few challenges to be solved that are mentioned in Section (1). In this paper, we focus on improving the retrieval performance of unsupervised deep cross-modal hashing in terms: (1) With the intention to extract similarity information from both visual and textual modalities, the sparse affinity graph, which tackles the problem of lacking label information, can be evolved to a dense form for fully preserve the the similarity information in hash codes. (2) To generate more discriminative hash codes for cross-modal retrieval.

2 Algorithm

In summary, the produce for solving the proposed problem in Eq. 9 is listed in Algorithm 1.

3 Datasets

Wiki [27]: This dataset consists of 2,866 samples in total with 10 classes. Each image is described by a paragraph which represents related image, from 1 to

Algorithm 1 Algorithm process of generating unified sparse affinity graph.

Input:
Image features $\boldsymbol{F}^I \in \mathbb{R}^{m \times d_I}$; text features $\boldsymbol{F}^T \in \mathbb{R}^{m \times d_T}$;
the number of clusters c, the number of neighbours k and hyper-parameter λ_1, λ_2;
Output:
The learned $\boldsymbol{Z} \in [0,1]^{m \times m}$.
Initialize $\boldsymbol{S^I}$ and $\boldsymbol{S^T}$ by using Eq. 5;
Initialize the weight for both visual and textual modal, $w^I = w^T = 1/2$;
Initialize $\boldsymbol{U}$ by $w^I \boldsymbol{S}^I + w^T \boldsymbol{S}^T$;
Initialize D by solving Eq. 8 ;
while the dimension of the nullspace of $\boldsymbol{L}$ not equals the number of connected components of $\boldsymbol{Z}$ or the maximum iteration unreached, **do**
    Fixw^I, w^T, $\boldsymbol{Z}$ and $\boldsymbol{D}$, update S^I and S^T by using Eq. 20;
    Fix $\boldsymbol{Z}$, $\boldsymbol{D}$, S^I and S^T, update w^I, w^T by using Eq. 23;
    Fix w^I, w^T, $\boldsymbol{D}$, $\boldsymbol{S}^I$ and $\boldsymbol{S}^T$, update $\boldsymbol{Z}$ by using Eq. 28;
    Fix w^I, w^T, $\boldsymbol{Z}$, $\boldsymbol{S}^I$ and $\boldsymbol{S}^T$, update $\boldsymbol{D}$, which is formed by the c eigenvectors of $\boldsymbol{L}$ corresponding to the c smallest eigenvalues;
end while
return The learned $\boldsymbol{Z}$ with clustering result.

10. In our experiment, we randomly select 500 from the total dataset as the query set, and the remaining samples form the training set as well as retrieval database.

NUS-WIDE [3]: It consists of 269,648 multi-modal instances, each of which contains an image and the related captions with 81 class labels. Following previous methods, the top 10 largest categories is selected which contain over 186 thousand instances and randomly choose 2,000 from them as query set of this paper, and employ the others as retrieval database. Given the huge storage expense in generating unified sparse semantic graph on the whole instances, we randomly select 20,000 image-text pairs from the original data as training set to relief the storage cost.

MIRFlickr-25K [13]: The original training set and validation set contains more than 25 thousand samples from 38 categories. The class labels are represented as one-hot form where 1 represents the image belongs to this class while 0 is the opposite. We randomly choose 1,000 samples as the query set and set the others as the retrieval database.

MSCOCO [19]: The dataset contains more than 123 thousand images-caption pairs from real-world with 80 class labels. We randomly choose 2,000 from them as query set and the others as retrieval database. And the way of building training database is the same as NUS-WIDE.

4 Evaluation metrics

To evaluate the efficiency of our method and the baseline approaches, we employ several frequently used evaluation metrics:

Mean Average Precision (MAP): MAP is a metric for evaluating the retrieval task performance and the definition is formulated as:

$$\frac{1}{|M|}\sum_{i=1}^{|M|}\left(\frac{1}{r}\sum_{j=1}^{r}p_{i,j}\right), \tag{1}$$

where r is the number of correct items returned from the dataset corresponding to the ith query items, $p_{i,j}$ means the precision of the jth correct item retrieved among all returned items, and $|M|$ is the size of the query set. In addition, the performance of all baselines and the proposed method are evaluated on 16 bit, 32 bit and 64 bit hash codes.

Precision-Recall (P-R curve):This curve show the precision and recall rates at several hamming radius $r = \{0, 1, 2, \cdots, \epsilon\}$. It is worthy noting that the beginning plot of curve means the precision and recall rate of the retrieval under the condition that the binary codes of both query and returned items are the same.

Precision-Number of retrieved points (P-N curve): This curve shows the precision of the top N retrieved samples.

5 Experiments

5.1 Performance on mAP@50

We compare the mAP results with CVH [16], IMH [30], CMFH [7], LSSH [44], DBRC [9] and UDCMH [35], DJSRH [31], DSAH [38], JDSH [21], DGCPN [41] conducted on MIRFlickr and NUS-WIDE datasets, with the retrieved number is set as 50 (*i.e.*, mAP@50). All the compared method are conducted according to their released codes or description in their original papers, and the results are listed in Tab. 1.

5.2 Precision@top-N

Fig. 1 shows the precision@top-N curves among recent methods. As shown in figures, SGEH outperforms the other recent unsupervised methods which strongly demonstrate the effectiveness of the proposed method.

5.3 Distribution of the generated hash codes

To further show the superiority of the proposed method, we also demonstrate the distributions of the generated hash codes of Wiki with t-SNE [25] in Fig. 2. From this figure, we can clearly observe that the generated data can be easily distinguished by classes, and it shows better clustering performance than the other illustrated methods.

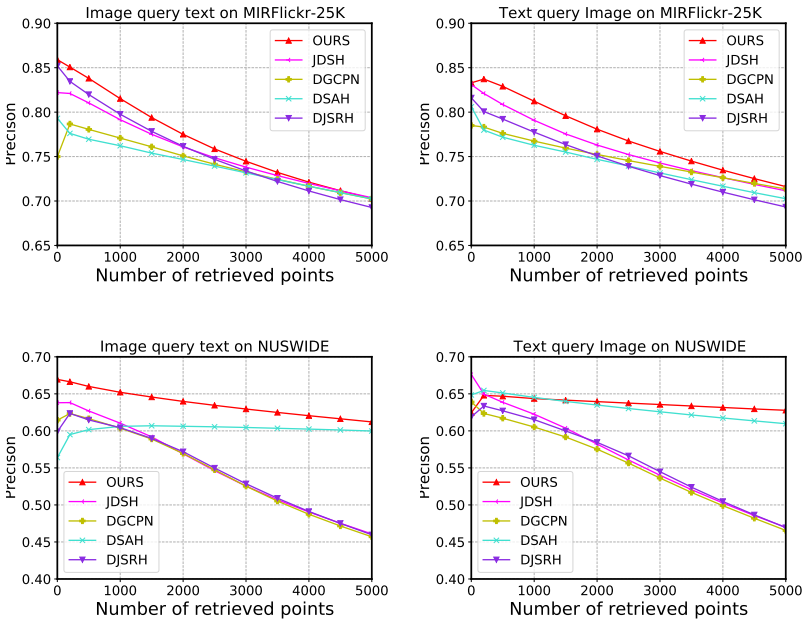
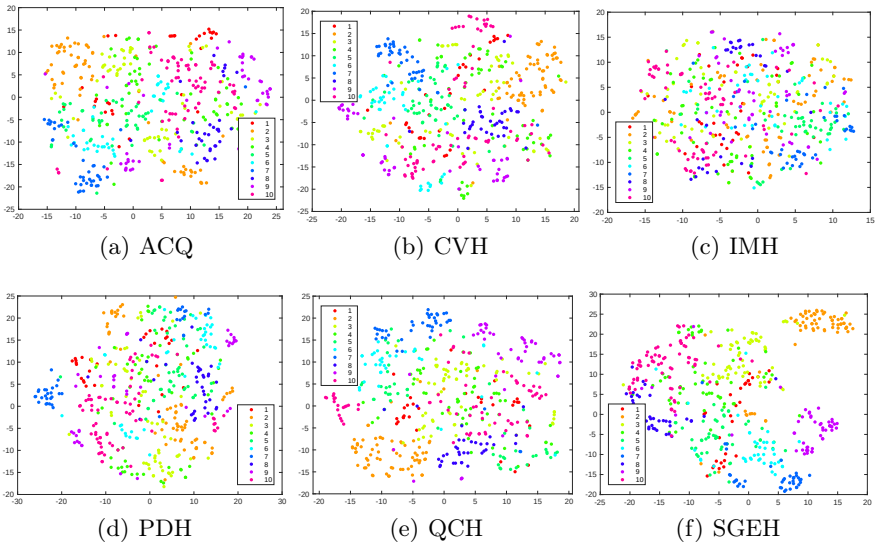


Fig. 1. The topN-curves at 32 bits.



(a) ACQ (b) CVH (c) IMH

(d) PDH (e) QCH (f) SGEH

Fig. 2. Illustration of the distribution of the generated codes from images with t-SNE for 32-bit codes on the dataset Wiki.

Table 1. The mAP@50 results on image query text ($I \rightarrow T$) and text query image ($T \rightarrow I$) retrieval tasks at various encoding lengths and datasets. The best performances are shown as bold.

Task	Method	MIRFlickr-25K			NUS-WIDE		
		16bit	32bit	64bit	16bit	32bit	64bit
$I \rightarrow T$	CVH	0.606	0.599	0.596	0.372	0.362	0.406
	IMH	0.612	0.601	0.592	0.470	0.473	0.476
	CMFH	0.621	0.624	0.625	0.455	0.459	0.465
	LSSH	0.584	0.599	0.602	0.481	0.489	0.507
	DBRC	0.617	0.619	0.620	0.424	0.459	0.447
	UDCMH	0.689	0.698	0.714	0.511	0.519	0.524
	DJSRH	0.810	0.843	0.862	0.624	0.673	0.688
	DGCPN	0.752	0.794	0.861	0.606	0.650	0.676
	DSAH	0.763	0.834	0.864	0.617	**0.672**	**0.704**
	JDSH	0.787	0.837	0.863	**0.653**	0.669	0.698
	SGEH	**0.821**	**0.851**	**0.866**	0.644	0.662	0.695
$T \rightarrow I$	CVH	0.591	0.583	0.576	0.401	0.384	0.442
	IMH	0.603	0.595	0.589	0.478	0.483	0.472
	CMFH	0.642	0.662	0.676	0.529	0.577	0.614
	LSSH	0.637	0.659	0.659	0.577	0.617	0.645
	DBRC	0.618	0.626	0.626	0.455	0.459	0.468
	UDCMH	0.692	0.704	0.718	0.637	0.653	0.695
	DJSRH	0.786	0.822	0.835	0.612	0.644	0.671
	DGCPN	0.671	0.818	0.855	0.595	0.652	0.692
	DSAH	0.769	0.827	0.860	0.632	0.677	0.697
	JDSH	0.806	0.840	0.856	**0.670**	0.681	0.703
	SGEH	**0.811**	**0.855**	**0.859**	0.651	**0.682**	**0.710**

5.4 Parameter Sensitivity

In this section we discuss the influence of core hyper-parameters, namely, the clustering centres c and the number of neighbours p on the NUSWIDE dataset with the 32 bits hash code. We fix c at 5 and vary p from 1000 to 12000. Then, we fix p at 10000 and vary c from 3 to 20, and record the results in Fig. 3. From this figure, we can discover that although the best results are achieved when p is 10000 and c is 5, these curves are relatively flat. This phenomenon reveals that the number of clusters and the number of neighbours only have a little impact on the final performance of the proposed method.

References

1. Bronstein, M.M., Bronstein, A.M., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3594–3601. IEEE (2010)
2. Cao, Y., Long, M., Wang, J., Yu, P.S.: Correlation hashing network for efficient cross-modal retrieval. In: Proceedings of British Machine Vision Conference (2017)

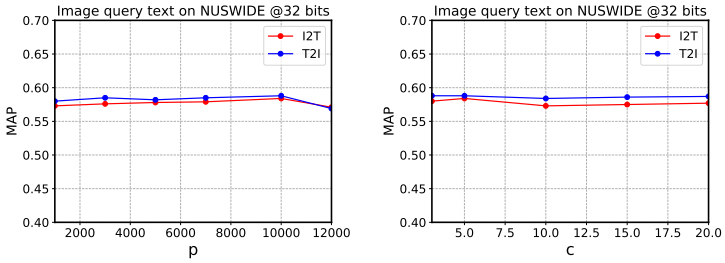


Fig. 3. MAP of SGEH with different parameters on NUSWIDE.

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. Deng, C., Chen, Z., Liu, X., Gao, X., Tao, D.: Triplet-based deep hashing network for cross-modal retrieval. IEEE Transactions on Image Processing **27**(8), 3893–3903 (2018)
5. Deng, C., Xu, X., Wang, H., Yang, M., Tao, D.: Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. IEEE Transactions on Image Processing **29**, 8892–8902 (2020)
6. Deng, C., Yang, E., Liu, T., Tao, D.: Two-stream deep hashing with class-specific centers for supervised image search. IEEE transactions on neural networks and learning systems **31**(6), 2189–2201 (2019)
7. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2075–2082 (2014)
8. He, L., Xu, X., Lu, H., Yang, Y., Shen, F., Shen, H.T.: Unsupervised cross-modal retrieval through adversarial learning. In: ICME. pp. 1153–1158 (2017)
9. Hu, D., Nie, F., Li, X.: Deep binary reconstruction for cross-modal hashing. IEEE Transactions on Multimedia **21**(4), 973–985 (2018)
10. Hu, H., Xie, L., Hong, R., Tian, Q.: Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2020)
11. Hu, M., Yang, Y., Shen, F., Nie, N., Hong, R., Shen, H.: Collective reconstructive embeddings for cross-modal hashing. IEEE IEEE Transactions on Image Processing **28**(6), 2770–2784 (2019)
12. Hu, M., Yang, Y., Shen, F., Xie, N., Hong, R., Shen, H.T.: Collective reconstructive embeddings for cross-modal hashing. IEEE Transactions on Image Processing **28**(6), 2770–2784 (2018)
13. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 39–43 (2008)
14. 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)
15. Jiang, Q.Y., Li, W.J.: Discrete latent factor model for cross-modal hashing. IEEE Transactions on Image Processing **28**(7), 3490–3501 (2019)
16. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: Proceedings of the International Joint Conference on Artificial Intelligence (2011)

17. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2018)
18. Li, C., Chen, Z., Zhang, P., Luo, X., Nie, L., Xu, X.: Supervised robust discrete multimodal hashing for cross-media retrieval. IEEE Transactions on Multimedia **21**(11), 2863–2877 (2019)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of European Conference on Computer Vision. pp. 740–755. Springer (2014)
20. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3864–3872 (2015)
21. 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)
22. Lu, X., Zhu, L., Li, J., Zhang, H., Shen, H.T.: Efficient supervised discrete multi-view hashing for large-scale multimedia search. IEEE Transactions on Multimedia **22**(8), 2048–2060 (2020). https://doi.org/10.1109/TMM.2019.2947358
23. Lu, X., Zhu, L., Li, J., Zhang, H., Shen, H.T.: Efficient supervised discrete multi-view hashing for large-scale multimedia search. IEEE Transactions on Multimedia **22**(8), 2048–2060 (2019)
24. Luo, X., Yin, X.Y., Nie, L., Song, X., Wang, Y., Xu, X.S.: Sdmch: Supervised discrete manifold-embedded cross-modal hashing. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 2518–2524 (2018)
25. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11), 2579âĹŠ–2605 (2008)
26. Mandal, D., Chaudhury, K.N., Biswas, S.: Generalized semantic preserving hashing for n-label cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4076–4084 (2017)
27. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the ACM International Conference on Multimedia. pp. 251–260 (2010)
28. Shen, H.T., Liu, L., Yang, Y., Xu, X., Huang, Z., Shen, F., Hong, R.: Exploiting subspace relation in semantic labels for cross-modal hashing. IEEE Transactions on Knowledge and Data Engineering (2020)
29. Shi, Y., You, X., Zheng, F., Wang, S., Peng, Q.: Equally-guided discriminative hashing for cross-modal retrieval. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 4767–4773 (2019)
30. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the International Conference on Management of Data. pp. 785–796 (2013)
31. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the International Conference on Computer Vision. pp. 3027–3035 (2019)
32. Sun, C., Song, X., Feng, F., Zhao, W.X., Zhang, H., Nie, L.: Supervised hierarchical cross-modal hashing. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 725–734 (2019)
33. Weiss, Y., Torralba, A., Fergus, R., et al.: Spectral hashing. In: Advances in Neural Information Processings. vol. 1, p. 4. Citeseer (2008)

34. Wu, B., Yang, Q., Zheng, W.S., Wang, Y., Wang, J.: Quantized correlation hashing for fast cross-modal search. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 3946–3952. Citeseer (2015)
35. 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: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 2854–2860 (2018)
36. Xie, D., Deng, C., Li, C., Liu, X., Tao, D.: Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. IEEE Transactions on Image Processing **29**, 3626–3637 (2020)
37. Xie, L., Shen, J., Zhu, L.: Online cross-modal hashing for web image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence (2016)
38. 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)
39. Yang, E., Deng, C., Li, C., Liu, W., Li, J., Tao, D.: Shared predictive cross-modal deep quantization. IEEE Transactions on Neural Networks and Learning Systems **29**(11), 5292–5303 (2018)
40. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
41. 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 (2021)
42. Zhang, D., Li, W.J.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2014)
43. Zhang, J., Peng, Y., Yuan, M.: Unsupervised generative adversarial cross-modal hashing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
44. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 415–424 (2014)