Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels

Ruitao Pu [1], Yuan Sun [1*], Yang Qin [1], Zhenwen Ren [2], Xiaomin Song [3], Huiming Zheng [3], Dezhong Peng [1,3]

[1]Sichuan University, Chengdu, China, 610044,
[2]Southwest University of Science and Technology, Mianyang, China, 621010,
[3]Sichuan National Innovation New Vision UHD Video Technology Co., Ltd., Chengdu, China, 610095,
ruitaopu@gmail.com, sunyuan_work@163.com, qinyang.gm@gmail.com, rzw@njust.edu.cn, songxiaomin@uptcsc.com, michaelzheng@uptcsc.com, pengdz@scu.edu.cn.

Abstract

Cross-modal hashing (CMH) has appeared as a popular technique for cross-modal retrieval due to its low storage cost and high computational efficiency in large-scale data. Most existing methods implicitly assume that multi-modal data is correctly labeled, which is expensive and even unattainable due to the inevitable imperfect annotations (i.e., noisy labels) in real-world scenarios. Inspired by human cognitive learning, a few methods introduce self-paced learning (SPL) to gradually train the model from easy to hard samples, which is often used to mitigate the effects of feature noise or outliers. It is a less-touched problem that how to utilize SPL to alleviate the misleading of noisy labels on the hash model. To tackle this problem, we propose a new cognitive cross-modal retrieval method called Robust Self-paced Hashing with Noisy Labels (RSHNL), which can mimic the human cognitive process to identify the noise while embracing robustness against noisy labels. Specifically, we first propose a contrastive hashing learning (CHL) scheme to improve multi-modal consistency, thereby reducing the inherent semantic gap. Afterward, we propose center aggregation learning (CAL) to mitigate the intra-class variations. Finally, we propose Noise-tolerance Self-paced Hashing (NSH) that dynamically estimates the learning difficulty for each instance and distinguishes noisy labels through the difficulty level. For all estimated clean pairs, we further adopt a self-paced regularizer to gradually learn hash codes from easy to hard. Extensive experiments demonstrate that the proposed RSHNL performs remarkably well over the state-of-the-art CMH methods.

Code — https://github.com/perquisite/RSHNL

Introduction

With the explosive growth of multi-modal data, cross-modal retrieval (CMR) has attracted a wide range of attention in the community (Zhou, Hassan, and Hoon 2023), which retrieves relevant samples across different modalities. For large-scale multi-modal data, cross-modal hashing (CMH) (Zhu et al. 2023) offers an efficient solution due to its low storage cost and high retrieval efficiency. The basic idea of CMH is to learn discriminative hash codes to alleviate the heterogeneity gap between different modalities. Due to the complexity of collecting annotations, some unsupervised CMR methods (Zhang et al. 2023b; Li et al. 2024b; Qin et al. 2023) have been proposed to eliminate the reliance on abundant labels. However, their performance often suffers without supervised semantic guidance. Recently, numerous supervised CMH methods (Li et al. 2024a; Chen, Cao, and Liu 2021) have been proposed and achieved pleasing performance. Most of them implicitly assume that all collected labels are correctly labeled, which is unrealistic due to inevitable noisy labels from manual or non-expert annotations (Song et al. 2022a; Kuznetsova et al. 2020). These noisy labels can mislead hash models, significantly degrading retrieval performance. Besides, learning from multi-modal instances with noisy labels is difficult due to a great heterogeneity gap. Thus, it is a challenging problem to mitigate the performance degradation caused by noisy labels for cross-modal retrieval.

To alleviate the impact of noisy labels, many CMR methods (Xu et al. 2022; Wang et al. 2024, 2021) have been developed. For example, ELRCMR utilizes dynamic weights to prevent overfitting noisy labels. However, most of them rely on real-valued representation, leading to high storage and computational costs. Although hash representations are more lightweight, unreliable labels could expand quantization errors. To this end, a few CMH methods have been proposed. For instance, NrDCMH adopts the difference between label similarity and feature similarity to detect noise. Although remarkable progress, most of them implicitly assume constant learning priorities for each instance, making the model biased toward hard instances with noisy labels and leading to the overfitting problem. Inspired by human cognitive learning, self-paced learning (SPL) was presented to explore more valuable discriminative information contained in hard instances gradually. In other words, SPL can gradually train the model from easy to hard instances to improve generalization. However, SPL is usually used for feature noise or outliers. It is a less-touched problem to employ the SPL paradigm to mitigate the negative effects of noisy labels.

In this paper, we propose a new cognitive cross-modal retrieval method, termed Robust Self-paced Hashing with Noisy Labels (RSHNL), which could enable the model to learn with noisy labels. As shown in Fig.1, our RSHNL mimics the human cognitive process to learn each instance in Hamming space from easy to hard, thereby embracing the

*Corresponding author.

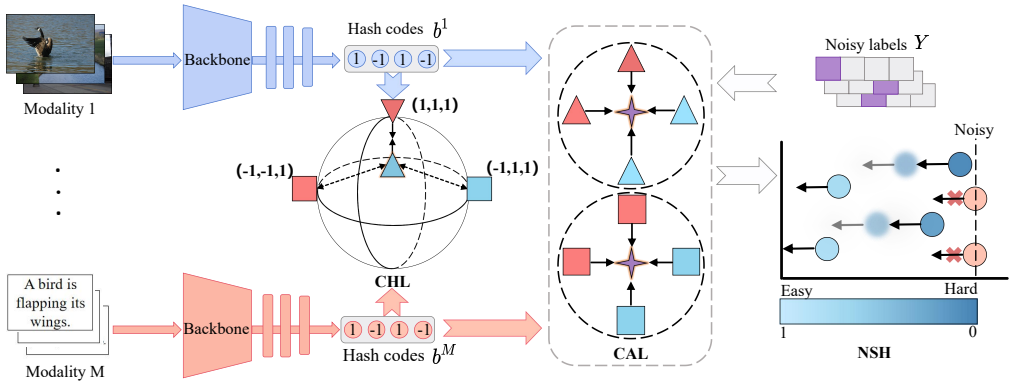


Figure 1: The framework of our RSHNL. Blue and red represent hash codes of different modalities. Triangles and rectangles represent different categories. And the doji represents hash centers. Specifically, CHL maximizes the consistency of multi-modal data to alleviate the cross-modal gap. CAL develops a unified hash code for each class as a center and encourages the compactness of intra-class hash codes towards their corresponding hash centers. NSH dynamically distinguishes noisy labels based on their assessment difficulty while facilitating learning hash codes from easy to hard for clean pairs.

robustness of eliminating noisy labels. Compared with existing self-paced hashing methods which only consider feature noise/outliers and cannot deal with noisy labels, our RSHNL assesses the learning pace by identifying noisy labels and learning clean pairs from easy to hard in the form of cognitive learning. Specifically, to reduce the inherent semantic gap, we first present a contrastive hashing learning (CHL) scheme to maximize the consistency between multi-modal data. Then, we propose center aggregation learning (CAL) to learn unified hash centers to aggregate hash codes from the same category, thereby mitigating the intra-class variations from multi-modal inputs. Further, we propose Noise-tolerance Self-paced Hashing (NSH) to adaptively measure the learning hardness of each instance and distinguish noisy labels according to the corresponding difficulty. For all sample pairs with clean labels, we adopt a self-paced regularizer that begins with easier pairs and advances to more complex ones. The main contributions are summarized as follows:

- We propose a new cognitive cross-modal hashing paradigm that alleviates the negative effect of noisy labels. To the best of our knowledge, this could be the first work that introduces SPL to distinguish and eliminate noisy labels in CMH tasks.
- We propose a Noise-tolerance Self-paced Hashing (NSH) loss that automatically determines noise labels and builds a full learning sequence for clean sample pairs, thereby evolving from easy to hard until all clean pairs are incorporated for training.
- Extensive experiments comprehensively verify that our proposed RSHNL has remarkably superior performance over the current state-of-the-art methods in different noise rates.

Related Work

Cross-modal Hashing

Recently, many CMH methods have been proposed, which can be roughly divided into two categories, i.e., unsupervised and supervised. Unsupervised CMH methods (Cao et al. 2023) aim to utilize the original distribution of the data to learn a common Hamming space. For example, DSAH (Yang et al. 2020) integrates co-occurrence information and semantic relevance of different modalities to guide hashing learning. For supervised CMH methods (Liu et al. 2024; Sun et al. 2024b), their goal is to leverage annotation information to learn more compact and discriminative hash codes. For instance, DCMH (Jiang and Li 2017) maintains the semantic relevance of hash representations through the label similarity matrix. However, almost all of these methods implicitly assume multi-modal data are well labeled. In practical applications, labeling noise is ubiquitous, which could mislead the hash model to overfit the noise. To tackle this problem, some supervised methods are developed to learn to hash from noisy labels robustly. For example, to utilize the memorization effect of DNNs and reduce the impact of noisy labels, CMMQ (Yang et al. 2022) selects confidence samples with smaller loss values. To deal with the problem of low-quality label annotations, DHRL (Shu et al. 2024) constructs a ranking and swapping module to estimate the uncertainty from noisy labels. Although these methods achieve promising performance, they unconsciously ignore the adverse impact of sample pairs with noisy labels and keep the learning priority of each instance the same, which violates the human cognitive process.

Self-paced Learning

Inspired by human cognitive learning, self-paced learning (SPL) (Kumar, Packer, and Koller 2010; Jiang et al. 2014a) is proposed to train the model from easy samples to hard ones. For example, to mitigate the noise/outlier problem, SCSM (Liang et al. 2016) utilizes SPL to learn samples from easy to hard. However, this will introduce several hyper-parameters due to manually setting the weighting function, which is undesirable. Thus, Meta-SPN (Wei et al. 2021) is proposed to learn the weight values automatically. However, most of these methods only focus on the contributions of instances to learn hash codes from easy to hard, thereby resisting feature noise interference. To this end, DSCMH (Sun

et al. 2024a) proposes a novel dual SPL mechanism from the perspectives of instance-level and feature-level difficulty to enhance robustness. In contrast, DHaPH (Huo et al. 2024) pays more attention to difficult sample pairs, and assigns larger weights to the hard pairs to learn discriminative information. Although these SPL methods have achieved considerable performance, almost all of them only consider the noise or outliers in the data. When faced with multi-modal data with noisy labels, how to utilize SPL to alleviate the noise overfitting problem is rarely studied. In this paper, we expect to distinguish sample pairs with noisy labels and gradually explore discriminative semantic information from clean data for the hash model, thereby alleviating the negative effect of noisy labels.

The Proposed Method

Problem Formulation

In this paper, some notations are provided for a clear presentation. We first denote $D_m = \{(x_i^m, y_i)\}_{i=1}^N$ as multi-modal data with N instances from M modalities, where x_i^m represents i-th sample of the m-th modality, $Y \in \mathbb{R}^{N\times K}$ is the corresponding labels, and K is number of the categories. For a instance x_i^m, if it belongs to k-th category, the k-th element of y_i is 1, i.e. $y_{i,k} = 1$, otherwise $y_{i,k} = 0$.

The basic idea of cross-modal hashing is to project multi-modal data into a common Hamming space by adopting different hash functions. Let hash code of i-th sample from m-th modality is $b_i^m \in \{-1, 1\}^L$, where L represents hash length. And each hash function can be denoted as $\mathcal{H}^m(\cdot, \Theta^m)$, where Θ^m is the corresponding learnable parameters. Subsequently, we can apply the $sign$ function to obtain hash representation, i.e.,

$$b_i^m = sign(\mathcal{H}^m(x_i^m, \Theta^m)). \tag{1}$$

Since binary optimization is a typical NP-hard problem (Huo et al. 2024), we adopt a $tanh$ function to learn binary-like codes during training.

Contrastive Hashing Learning

To narrow the inherent semantic gap between multi-modal data, we present a contrastive hashing learning scheme (CHL) that maximizes the consistency of hash codes from different modalities to improve the discrimination. Specifically, we treat the same instances from different modalities as positive pairs and then encourage them to be close while keeping negative pairs far away. First, we define the probability that x_i^m belongs to the i-th instance as

$$q(i\,|x_i^m) = \frac{\sum_{j=1}^M e^{b_i^m (b_i^j)^\top / \tau}}{\sum_{j=1}^M \sum_{z=1}^N e^{b_i^m (b_z^j)^\top / \tau}}, \tag{2}$$

where τ is a temperature parameter. Then, the CHL loss $\mathcal{L}_C$ could be written as maximizing a joint probability $\prod_{i=1}^N \prod_{m=1}^M p(i\,|x_i^m)$ of all samples, which is equivalent to minimizing the following formula, i.e.,

$$\begin{aligned}\mathcal{L}_C = &\frac{1}{N}\sum_{i=1}^N \sum_{m=1}^M (1-r)\frac{1-(q(i\,|x_i^m))^r}{r} \\ &+ r(1-q(i\,|x_i^m)).\end{aligned} \tag{3}$$

By minimizing Eq.3, positive pairs are forced to be compressed, while negative pairs are scattered in the Hamming space, thereby alleviating multi-modal discrepancy.

Noise-tolerance Self-paced Hashing

To mitigate the intra-class variations of multi-modal data, we propose center aggregation learning (CAL) to learn a unified hash representation for each class as a center and promote modality-specific hash codes with the same category to be aggregated to the corresponding hash centers. Specifically, we randomly initialize and obtain hash centers $C = \{c_1, c_2, ..., c_K\}$, where $c_K \in \mathbb{R}^{L\times 1}$ is the normalized and discretized binary vector for the K-th class. For any sample x_i^m, we first define its probability belonging to the k-th center as

$$p(k\,|x_i^m) = \frac{e^{b_i^m c_k / \tau}}{\sum_{j=1}^K e^{b_i^m c_j / \tau}}, \tag{4}$$

where τ is a temperature parameter. Due to the lack of guidance information, the obtained probabilities could lead to prediction errors. Thus, to make the hash code inherit more semantic information, we define the semantic aggregation probability as

$$v_i^m = \sum_{k=1}^K y_{i,k} p(k\,|x_i^m)\,. \tag{5}$$

Afterward, we can obtain the following center aggregation loss $\mathcal{L}_p$, i.e.,

$$\mathcal{L}_p = \frac{1}{N}\sum_{i=1}^N \sum_{m=1}^M (1-r)\frac{1-(v_i^m)^r}{r} + r(1-v_i^m), \tag{6}$$

where $r \in (0, 1]$ is a weight factor. However, due to the ubiquitous noisy labels in the data, such center aggregation loss is inevitably disrupted, thereby tending to overfit the corrupted labels. Since DNNs (Song et al. 2022b) are robust in the early training stage, we use Eq.6 to warm up the model.

Inspired by the great success of self-paced learning (SPL), we can organize the learning sequence of samples from easy to hard, thereby improving the retrieval performance. Thus, some SPL-based hashing methods (Sun et al. 2024a,c) have been proposed to mitigate the negative effects of noise or outliers. However, all of them implicitly assume that the multi-modal data are labeled correctly while ignoring the existence of noisy labels. Moreover, they keep the learning priority of each sample with noisy labels constant, which could be unreasonable due to the labeled differences. To overcome this issue, we propose a Noise-tolerance Self-paced Hashing (NSH) strategy to learn hash codes from noisy labels. Similar to prior SPL methods, our proposed NSH gradually learns from easy pairs to difficult ones, thereby automatically incorporating more data into the training process. Different from them, we reveal that the SPL scheme can distinguish sample pairs with noisy labels to mitigate the overfitting problem. Specifically, our NSH adopts a hardness measurement strategy that dynamically estimates the learning difficulty of each pair and distinguishes the noisy

labels. Then, NSH gradually learns hash codes from easy to hard until it is sufficient to handle hard ones. Thereupon, the problem could be formulated as follows

$$\mathcal{L}_S = \frac{1}{N}\sum_{i=1}^{N} w_i \underbrace{\sum_{m=1}^{M}(1-r)\frac{1-(v_i^m)^r}{r} + r(1-v_i^m)}_{\ell_i} + \frac{1}{N}\sum_{i=1}^{N}\mathcal{R}(w_i,\gamma), \tag{7}$$

where $w_i \in [0,1]$ is the importance weight of i-th sample pairs that evaluates the reliability of label dimension. $\mathcal{R}(w_i,\gamma)$ is the self-paced regularizer controlled by the learning pace parameter γ, which could assign a weight w_i to estimate the learning difficulty of each instance. We adopt a linear interpolation function (Jiang et al. 2014b) to construct the following self-paced regularizer $\mathcal{R}(w_i,\gamma)$, i.e.,

$$\mathcal{R}(w_i,\gamma) = \gamma(\frac{1}{2}{w_i}^2 - w_i). \tag{8}$$

In brief, the weight can be regarded as the easiness of each sample pair with noisy labels. If the weight is higher, the instance could be viewed as easier. The loss decreases gradually with the learning process, thus enlarging the weight. When $\ell_i > \gamma$, we consider this sample pair could be mislabeled and assign the weight as zero to represent the hardest/noisy sample pair. When $\ell_i \leq \gamma$, NSH first considers reliable/easy pairs at the beginning and then gradually incorporates unreliable/hard ones into training. In other words, we learn hash codes in the way of human cognitive learning (i.e., from easy to hard) until more clean pairs are incorporated into the hash model.

The Objective Function

By combining the above losses, we can obtain the overall objective loss function as follows

$$\mathcal{L} = \begin{cases} \mathcal{L}_p + \alpha\mathcal{L}_C, & \text{if } t < N_w, \\ \mathcal{L}_S + \alpha\mathcal{L}_C & \text{if } N_w \leq t < N_m. \end{cases} \tag{9}$$

where α is a hyper-parameter, t is the current epoch, N_w and N_m are warm-up epoch and maximal epoch, respectively. The training process of RSHNL is shown in the appendix.

Theoretical Justification

To show the robustness of the proposed RSHNL, we deeply analyze the impact of the weight w_i on the loss $\mathcal{L}_S$. Our goal is to minimize $\mathcal{L}_S$ by updating the weight w_i and network parameters $\{\Theta^m\}_{m=1}^M$ alternatively, while making the other fixed. Given a fixed $\{\Theta^m\}_{m=1}^M$, we can obtain the following optimal solution, i.e.,

$$\begin{aligned} w_i^* &= \underset{w_i\in[0,1]}{\operatorname{argmin}} w_i\ell_i + \gamma(\frac{1}{2}{w_i}^2 - w_i) \\ &= \underset{w_i\in[0,1]}{\operatorname{argmin}} \frac{\gamma}{2}{w_i}^2 + (\ell_i - \gamma)w_i. \end{aligned} \tag{10}$$

Since $w_i \in [0,1]$, when $\ell_i - \gamma > 0$, the optimal solution w_i^* is obviously 0. While $\ell_i - \gamma \leq 0$, let the derivative of Eq.10 be 0, we can obtain

$$w_i^* = 1 - \frac{\ell_i}{\gamma}. \tag{11}$$

Clearly, since $\gamma \geq 0$ and $\ell_i \geq 0$, we can get $w_i \in [0,1]$ consistent with the original setting. In summary, we can get the solution as follows

$$w_i^* = \max(0, 1 - \frac{\ell_i}{\gamma}). \tag{12}$$

If the loss ℓ_i is too large (i.e., $\ell_i > \gamma$), we regard the corresponding sample pair as hard data with noise labels and assign a weight as 0. When $\ell_i \leq \gamma$, the i-th pair with a large weight can be implicitly considered as easy. Otherwise, one with a small weight can be regarded as hard. In general, this strategy can not only distinguish clean sample pairs but also gradually train the hash model from easy to hard, embracing more robustness and generalization simultaneously.

However, selecting a suitable learning pace parameter γ is challenging. If γ is too small, no sample pairs would be selected for training. And if γ is too large, all pairs will participate in training. Clearly, it will cause our NSH to be unable to distinguish noise labels, thus reducing the retrieval performance. In Eq.7, since $v_i^m \in [0,1]$ and $r > 0$, we can get the minimum value of ℓ_i as

$$\ell_i^{min} = 0. \tag{13}$$

Similarly, we can obtain the maximum value of ℓ_i as

$$\ell_i^{max} = \frac{M(r^2 - r + 1)}{r}. \tag{14}$$

Hence, we can get γ is bounded by $0 < \gamma < \frac{M(r^2-r+1)}{r}$. To obtain a suitable γ to distinguish clean pairs, we perform the sensitivity analysis in the appendix.

Experiments

Dataset

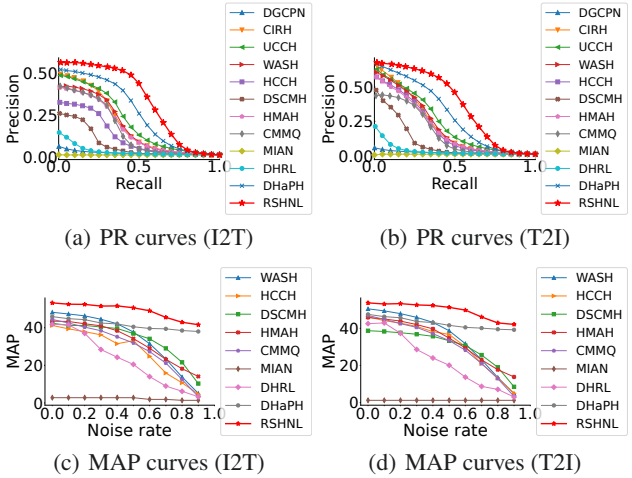
To verify the effectiveness of the proposed RSHNL, we conduct extensive experiments on four widely used datasets, i.e., XMedia (Peng et al. 2015), INRIA-Websearch (Krapac et al. 2010), Wikipedia (Rasiwasia et al. 2010), and XMediaNet (Peng, Huang, and Zhao 2018).

Experiments Settings

To evaluate the performance of the proposed RSHNL and competitors, we conduct two common cross-modal retrieval tasks. I2T and T2I represent using images as queries to retrieve texts and using texts as queries to retrieve images, respectively. Similar to (Zhen et al. 2019), we adopt Mean Average Precision (MAP) to evaluate the retrieval performance of all methods, which is a widely used evaluating metric. To comprehensively evaluate the effectiveness, we set noisy labels as symmetric noise with different rates (i.e., 0.2, 0.4, 0.6, and 0.8). Besides, the bit lengths are configured to 16, 32, 64, and 128. Besides, all experiments are conducted on a single GeForce RTX3090Ti 24GB GPU and our RSHNL is implemented in PyTorch 1.12.0. More details on implementation are provided in the appendix due to space limitations.

Task	Method	Noise	0.2				0.4				0.6				0.8			
		Ref.	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
I2T	DGCPN	AAAI'21	49.0	64.2	47.4	51.3	49.0	64.2	47.4	51.3	49.0	64.2	47.4	51.3	49.0	64.2	47.4	51.3
	CIRH	TKDE'22	72.5	78.0	79.3	83.8	72.5	78.0	79.3	83.8	72.5	78.0	79.3	83.8	72.5	78.0	79.3	83.8
	UCCH	TPAMI'23	37.6	50.2	67.9	82.7	37.6	50.2	67.9	82.7	37.6	50.2	67.9	82.7	37.6	50.2	67.9	82.7
	WASH	TKDE'23	80.7	85.9	87.1	87.7	75.6	79.4	81.3	82.1	44.8	53.8	57.4	59.6	13.9	16.3	17.2	19.2
	HCCH	TMM'24	71.1	81.9	82.1	84.4	71.6	76.8	78.2	80.6	60.3	61.5	61.5	72.5	26.3	41.2	41.1	48.5
	DSCMH	AAAI'24	81.2	85.6	87.9	88.2	79.4	83.6	85.4	85.8	63.1	70.4	74.4	77.5	31.0	35.1	43.7	47.2
	HMAH	TMM'22	78.7	84.4	86.7	88.2	55.0	65.1	71.4	74.6	22.3	29.8	33.5	37.6	8.7	9.6	10.7	10.7
	CMMQ	CVPR'22	86.6	87.9	87.4	86.6	74.8	77.4	74.5	72.0	51.5	43.8	38.9	38.7	18.3	18.5	13.7	12.6
	MIAN	TKDE'23	18.1	29.6	34.5	35.4	12.4	16.2	17.7	16.2	10.7	9.9	11.5	11.1	7.6	7.7	7.1	8.0
	DHRL	TBD'24	11.0	39.3	86.7	90.5	9.9	66.0	84.2	86.6	9.4	37.7	69.6	74.4	6.6	8.9	37.6	43.8
	DHaPH	TKDE'24	79.3	84.9	86.9	88.6	69.6	78.2	82.6	84.6	52.3	66.0	72.8	79.8	42.7	48.3	60.6	70.2
	RSHNL	Ours	**89.5**	**90.0**	**90.9**	**90.8**	**89.7**	**90.2**	**91.1**	**89.8**	**83.0**	**87.7**	**88.9**	**88.4**	**81.7**	**88.1**	**87.6**	**84.7**
T2I	DGCPN	AAAI'21	50.0	58.2	31.5	40.2	50.0	58.2	31.5	40.2	50.0	58.2	31.5	40.2	50.0	58.2	31.5	40.2
	CIRH	TKDE'22	67.4	73.1	76.9	82.0	67.4	73.1	76.9	82.0	67.4	73.1	76.9	82.0	67.4	73.1	76.9	82.0
	UCCH	TPAMI'23	56.0	66.9	75.5	83.8	56.0	66.9	75.5	83.8	56.0	66.9	75.5	83.8	56.0	66.9	75.5	83.8
	WASH	TKDE'23	81.6	86.0	86.9	88.3	75.0	78.7	81.3	82.1	44.6	53.6	56.7	59.4	14.3	16.7	17.3	19.4
	HCCH	TMM'24	69.5	80.7	80.0	84.0	70.0	75.7	76.7	80.0	58.1	60.2	58.5	72.6	26.2	40.8	41.2	48.2
	DSCMH	AAAI'24	80.5	85.0	86.3	87.9	78.0	81.6	83.5	84.8	63.5	70.9	73.9	77.4	29.1	33.0	41.0	45.0
	HMAH	TMM'22	77.4	84.0	85.9	88.0	53.2	65.0	71.0	73.9	23.1	30.5	34.1	38.9	8.9	10.0	11.5	11.4
	CMMQ	CVPR'22	85.1	83.4	82.0	78.2	70.8	74.4	67.6	64.4	45.2	36.0	33.2	43.5	14.4	14.0	13.2	10.2
	MIAN	TKDE'23	15.5	22.1	28.5	29.0	11.0	14.3	16.2	15.0	8.6	9.6	10.9	10.7	7.0	7.3	7.0	7.7
	DHRL	TBD'24	10.6	38.1	86.5	**91.0**	10.1	65.5	83.5	86.5	10.7	39.3	68.6	72.1	8.2	8.3	39.9	45.5
	DHaPH	TKDE'24	78.9	84.9	88.2	89.8	69.5	77.9	83.9	86.0	51.5	64.9	72.9	80.7	42.3	49.2	60.0	70.9
	RSHNL	Ours	**86.5**	**89.4**	**91.0**	90.6	**87.4**	**90.2**	**90.3**	**90.3**	**82.6**	**87.4**	**89.1**	**88.8**	**79.7**	**86.8**	**86.7**	**85.3**

Table 1: The MAP scores with different bit lengths on the XMedia dataset under different noise rates.



(a) PR curves (I2T) (b) PR curves (T2I)

(c) MAP curves (I2T) (d) MAP curves (T2I)

Figure 2: Experimental results with 128 bits on the INRIA-Websearch dataset under 0.6 noise rate.

Comparison Methods

To demonstrate the superiority of the proposed RSHNL, we compare RSHNL with 11 baselines, including deep unsupervised CMH methods (i.e., DGCPN (Yu et al. 2021), CIRH (Zhu et al. 2022), and UCCH (Hu et al. 2022)), shallow supervised CMH methods (i.e., WASH (Zhang et al. 2023a), HCCH (Sun et al. 2023), and DSCMH (Sun et al. 2024a)), and deep supervised CMH methods (i.e., HMAH (Tan et al. 2022), CMMQ (Yang et al. 2022), MIAN (Zhang et al. 2023c), DHRL (Shu et al. 2024), and DHaPH (Huo et al. 2024)). Among these, WASH, CMMQ, and DHRL are specifically designed to deal with the problem of noisy labels. DSCMH and DHaPH are the SPL-based hashing methods. For a fair comparison, we freeze the original backbones in the training process and report MAP scores on the testing set when MAP peaks on the validation set. For all experimental tables, the highest MAP scores are shown in **bold**, the second highest MAP scores are marked with underline, and '/' denotes out-of-memory.

Comparison with the State-of-the-Art

The experimental results on three datasets are reported in tables 1 to 3. The results on Wikipedia are given in the Appendix. Besides, we set the hash length as 128-bit on INRIA-Websearch, and then plot the precision-recall (PR) curves under 0.6 noise rate and the MAP curves under different noise rates in Fig.2. From these results, it can be observed:

- The retrieval performance of most methods improves with bit lengths increase because long hash codes contain

Task	Method	Noise	0.2				0.4				0.6				0.8			
		Ref.	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
I2T	DGCPN	AAAI'21	24.3	32.5	37.7	37.3	24.3	32.5	37.7	37.3	24.3	32.5	37.7	37.3	24.3	32.5	37.7	37.3
	CIRH	TKDE'22	14.6	21.5	26.5	30.8	14.6	21.5	26.5	30.8	14.6	21.5	26.5	30.8	14.6	21.5	26.5	30.8
	UCCH	TPAMI'23	18.9	26.1	29.5	34.3	18.9	26.1	29.5	34.3	18.9	26.1	29.5	34.3	18.9	26.1	29.5	34.3
	WASH	TKDE'23	31.5	38.0	43.5	46.2	26.2	33.0	38.5	42.1	16.9	22.1	27.9	31.5	4.7	7.5	10.9	13.7
	HCCH	TMM'24	11.8	20.5	28.5	37.6	8.8	13.1	21.9	33.9	5.4	8.3	12.8	22.2	3.3	3.2	5.5	10.2
	DSCMH	AAAI'24	18.8	27.6	35.2	41.2	17.9	25.0	33.4	39.6	13.5	19.4	27.7	34.1	5.6	9.7	16.7	21.6
	HMAH	TMM'22	26.4	34.5	39.8	42.1	19.0	28.3	34.4	38.4	10.3	17.5	24.1	29.1	5.2	8.3	13.4	18.2
	CMMQ	CVPR'22	31.1	35.5	38.3	39.6	27.8	32.1	34.4	35.8	17.0	21.2	27.9	30.4	4.1	3.9	11.6	9.8
	MIAN	TKDE'23	2.7	2.7	2.2	2.8	2.7	1.7	2.8	2.8	2.8	2.4	1.7	1.7	2.8	2.3	1.8	1.4
	DHRL	TBD'24	2.8	4.2	33.6	33.8	2.6	2.8	23.8	24.3	2.7	2.7	12.6	8.3	2.8	3.0	4.9	6.2
	DHaPH	TKDE'24	24.0	32.8	39.5	44.3	22.3	29.3	37.0	42.0	19.9	28.1	34.7	39.8	19.5	26.1	33.6	38.5
	RSHNL	Ours	**39.3**	**48.0**	**51.9**	**52.4**	**37.9**	**45.8**	**50.3**	**51.6**	**31.2**	**38.3**	**47.6**	**49.0**	**28.3**	**38.2**	**41.8**	**42.9**
T2I	DGCPN	AAAI'21	22.9	32.0	37.5	36.9	22.9	32.0	37.5	36.9	22.9	32.0	37.5	36.9	22.9	32.0	37.5	36.9
	CIRH	TKDE'22	14.2	21.2	26.6	31.2	14.2	21.2	26.6	31.2	14.2	21.2	26.6	31.2	14.2	21.2	26.6	31.2
	UCCH	TPAMI'23	17.4	25.3	29.5	34.9	17.4	25.3	29.5	34.9	17.4	25.3	29.5	34.9	17.4	25.3	29.5	34.9
	WASH	TKDE'23	30.8	38.4	44.8	47.8	25.1	32.7	39.3	43.3	15.8	21.7	27.8	31.9	4.2	7.2	10.6	13.3
	HCCH	TMM'24	12.7	23.6	35.0	42.5	9.4	17.0	29.4	39.0	5.9	10.8	19.7	29.3	3.0	3.6	7.1	13.1
	DSCMH	AAAI'24	19.1	27.0	32.8	37.6	18.5	24.6	31.1	35.7	13.9	18.5	24.4	30.7	5.6	8.7	15.6	19.0
	HMAH	TMM'22	25.0	34.5	40.8	43.9	18.2	28.3	34.9	39.7	9.9	17.1	24.4	29.4	4.7	7.9	13.1	17.7
	CMMQ	CVPR'22	31.4	36.6	38.5	39.0	27.3	32.3	34.5	34.1	16.2	20.6	27.3	30.4	4.3	3.4	10.7	8.5
	MIAN	TKDE'23	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	DHRL	TBD'24	2.7	4.2	32.3	33.7	2.7	2.6	23.5	24.0	2.8	2.6	13.2	7.7	2.7	2.9	4.8	7.1
	DHaPH	TKDE'24	22.4	32.8	40.7	45.7	21.0	29.3	37.6	43.1	18.6	28.4	35.4	40.8	18.6	25.6	33.8	39.5
	RSHNL	Ours	**38.2**	**47.9**	**52.1**	**53.3**	**36.2**	**45.9**	**50.3**	**52.3**	**30.2**	**37.7**	**47.9**	**49.8**	**27.1**	**38.0**	**42.0**	**42.9**

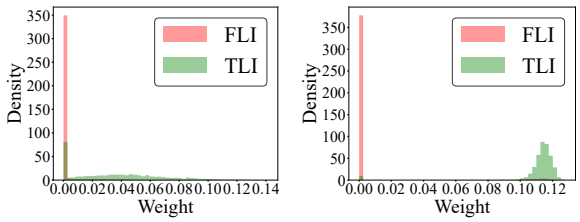
Table 2: The MAP scores with different bit lengths on the INRIA-Websearch dataset under different noise rates.

Task	Method	Noise	0.2				0.4				0.6				0.8			
		Ref.	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
I2T	DGCPN	AAAI'21	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	CIRH	TKDE'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	UCCH	TPAMI'23	9.4	13.5	17.1	19.7	9.4	13.5	17.1	19.7	9.4	13.5	17.1	19.7	9.4	13.5	17.1	19.7
	WASH	TKDE'23	8.0	15.1	24.2	34.0	7.0	13.1	21.2	30.6	4.9	8.5	14.4	21.8	2.0	3.1	4.7	6.9
	HCCH	TMM'24	1.6	2.0	4.8	15.0	1.4	1.5	3.4	12.1	1.3	1.4	2.2	5.8	0.8	0.9	1.1	1.9
	DSCMH	AAAI'24	5.2	9.9	18.1	28.7	4.7	8.5	14.9	24.7	3.4	5.4	10.3	18.5	1.8	2.3	3.9	6.6
	HMAH	TMM'22	2.9	3.6	5.9	10.9	1.3	1.5	1.9	3.9	1.0	1.0	1.3	2.2	1.0	1.1	1.3	1.7
	CMMQ	CVPR'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	MIAN	TKDE'23	0.8	1.5	1.8	2.5	0.8	1.2	1.3	1.8	0.7	0.9	1.0	1.2	0.7	0.8	0.8	0.9
	DHRL	TBD'24	0.7	0.7	0.9	15.2	0.7	0.7	1.0	13.2	0.7	0.7	2.7	6.3	0.7	0.7	0.7	1.3
	DHaPH	TKDE'24	9.8	15.9	23.4	28.5	9.1	15.8	22.6	27.6	9.2	15.1	21.9	27.5	8.8	15.2	21.5	27.3
	RSHNL	Ours	**35.3**	**43.5**	**47.5**	**48.6**	**35.5**	**42.1**	**47.0**	**45.8**	**33.2**	**41.8**	**46.2**	**45.0**	**28.2**	**39.6**	**44.9**	**38.9**
T2I	DGCPN	AAAI'21	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	CIRH	TKDE'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	UCCH	TPAMI'23	10.6	15.1	19.4	22.1	10.6	15.1	19.4	22.1	10.6	15.1	19.4	22.1	10.6	15.1	19.4	22.1
	WASH	TKDE'23	10.3	17.3	25.8	35.5	8.9	15.2	22.8	32.2	6.3	10.2	15.9	23.4	2.4	3.6	5.4	7.8
	HCCH	TMM'24	1.8	1.5	2.1	9.0	1.4	1.2	1.7	12.1	1.5	1.1	1.3	1.8	0.8	0.9	1.0	1.2
	DSCMH	AAAI'24	4.1	8.4	16.5	27.5	3.6	7.0	13.5	23.8	2.5	4.4	9.3	17.5	1.4	2.0	3.4	5.9
	HMAH	TMM'22	3.5	4.4	6.6	12.1	1.6	1.8	2.3	4.4	1.1	1.2	1.5	2.5	1.0	1.2	1.4	1.8
	CMMQ	CVPR'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	MIAN	TKDE'23	0.8	1.2	1.2	1.8	0.7	1.0	1.0	1.3	0.7	0.8	0.8	1.0	0.7	0.7	0.7	0.8
	DHRL	TBD'24	0.7	0.8	0.9	19.8	0.8	0.9	1	15.1	0.7	0.8	3.1	8.7	0.8	0.8	0.8	1.5
	DHaPH	TKDE'24	11.0	18.0	26.3	32.6	10.5	17.3	25.3	31.6	10.3	16.8	24.8	31.2	9.6	17.1	24.0	30.7
	RSHNL	Ours	**35.3**	**42.5**	**46.5**	**47.8**	**35.0**	**41.6**	**46.9**	**46.6**	**33.3**	**41.3**	**46.4**	**46.1**	**29.3**	**39.9**	**45.1**	**40.6**

Table 3: The MAP scores with different bit lengths on the XMediaNet dataset under different noise rates.

more discriminative information. Besides, the performance of a few methods (such as CMMQ and RSHNL) degrades with bit lengths increase. This may be because it is difficult to resist the interference of noisy labels for these methods, resulting in more noise information being incorporated into long hash codes.

- The retrieval performance of all supervised CMH methods is affected by noisy labels. As the noise rate increases, the CMH model is more likely to be misled, thereby resulting in a rapid drop in performance. Since unsupervised CMH methods do not need to exploit the label information, their performance is not affected by noisy labels at all.
- Most methods show worse performance or even fail on the INRIA-Websearch and XMediaNet datasets because more categories significantly increase the difficulty of learning discriminative hash codes from noisy labels.
- From PR curves, our RSHNL outperforms other baselines, which is consistent with what MAP demonstrates. According to MAP curves, the performance of almost all supervised CMH methods degrades as the noise rate increases. Thanks to the noise recognition capability of SPL, our RSHNL maintains stable and superior performance. Overall, the comprehensive performance of RSHNL outperforms all baselines.



(a) Epochs=5 (b) Epochs=100

Figure 3: The density versus the weight of all instances with 128 bits and 0.6 noise rate, where 'FLI' and 'TLI' denote 'False labeled instances' and 'True labeled instances', respectively.

Self-paced Analysis

To study the self-paced behavior of our RSHNL with 128 bits and 0.6 noise rate, we plot the density versus the weight of each instance from different epochs on INRIA-Websearch. From Fig.3, we can observe that: 1) At the beginning, RSHNL first assigns zero weight to hard instances and regards them as noisy instances, thereby separating multi-modal data into a clean or noisy subset. 2) As the training progresses, RSHNL gradually learns with all clean instances from easy to hard until all instances become easy.

Ablation Study

To show the effectiveness of the proposed components, we conduct ablation experiments with 128 bits on the two datasets compared with three variants. Specifically, RSHNL-1, RSHNL-2, RSHNL-3 represent removing the warm-up training, removing the loss $\mathcal{L}_C$, and removing SPL of $\mathcal{L}_S$, respectively. To be fair, all variants adopt the same parameters as RSHNL. As shown in Tab.4, we report their average MAP scores on I2T and T2I tasks. From these results, RSHNL shows the best retrieval performance, which means that all components are crucial for RSHNL.

Dataset	XMedia				INRIA-Websearch			
Noise	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
RSHNL-1	84.3	84.3	84.3	84.3	38.6	38.6	38.6	38.6
RSHNL-2	90.1	85.6	72.1	25.4	50.1	45.6	20.2	8.7
RSHNL-3	84.8	76.1	63.0	29.6	44.4	31.6	25.6	20.2
RSHNL	**90.7**	**90.0**	**88.6**	**85.0**	**52.9**	**51.9**	**49.4**	**42.9**

Table 4: Ablation study with 128 bits.

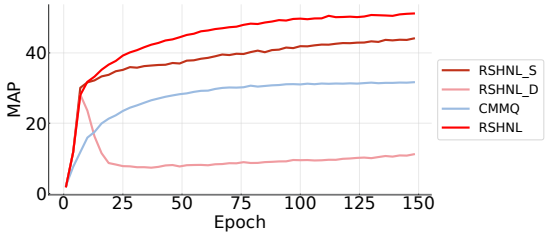


Figure 4: The average MAP scores versus epochs.

Robustness Analysis

To intuitively show the robustness of our RSHNL, on INRIA-Websearch under 0.6 noise, we compare it with CMMQ and two variants. Specifically, RSHNL-S represents removing the progressive learning mechanism by setting all weights greater than 0 to 1. RSHNL-D allows all instances to participate in the training by setting γ as a value (e.g., 200) greater than ℓ_i^{max}. Then, we plot the average MAP scores of I2T and T2I tasks with 128 bits. From Fig.4, we can observe that: 1) RSHNL-D overfits the noise, which indicates that the ability to distinguish noise is crucial. 2) Although CMMQ and RSHNL-S can prevent the overfitting problem, their retrieval performance is still lower than our method, which means our NSH could effectively improve the discrimination of hash codes by learning from easy to hard.

Conclusion

In this paper, we propose a new cognitive cross-modal hashing approach (i.e., RSHNL) with noisy labels, which contains three parts, i.e., CHL, CAL, and NSH. Specifically, CHL maximizes the consistency of multi-modal data to alleviate the semantic gap. CAL learns a unified hash representation for each class as a center and encourages hash codes with the same category to be close to the corresponding hash centers. NSH presents a dynamic hardness measurement strategy that dynamically estimates the learning difficulty for each pair and distinguishes the noisy labels while facilitating learning hash codes from easy to hard for clean pairs. Extensive experiments show that RSHNL outperforms 11 state-of-the-art CMH methods under noisy labels.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62372315), the Sichuan Science and Technology Program (Grant No. 2024NSFTD0049, 2024ZDZX0004, 2024YFHZ0144, 2024YFHZ0089, MZGC20240057), and the Mianyang Science and Technology Program (Grant No. 2023ZYDF091, 2023ZYDF003).

References

Cao, Y.; Gao, Y.; Chen, N.; Lin, J.; and Chen, S. 2023. Generative Adversarial Network Based Asymmetric Deep Cross-Modal Unsupervised Hashing. In *International Conference on Algorithms and Architectures for Parallel Processing*, 30–48. Springer.

Chen, N.; Cao, Y.; and Liu, C. 2021. Deep Cross-Modal Supervised Hashing Based on Joint Semantic Matrix. In *International Conference on Network and System Security*, 258–274. Springer.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2022. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3877–3889.

Huo, Y.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2024. Deep Hierarchy-aware Proxy Hashing with Self-paced Learning for Cross-modal Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 1–14.

Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014a. Self-paced learning with diversity. *Advances in Neural Information Processing Systems*, 27.

Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014b. Self-paced learning with diversity. *Advances in Neural Information Processing Systems*, 27.

Jiang, Q.-Y.; and Li, W.-J. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3232–3240.

Krapac, J.; Allan, M.; Verbeek, J.; and Juried, F. 2010. Improving web image search results using query-relative classifiers. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1094–1101. IEEE.

Kumar, M.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. *Advances in Neural Information Processing Systems*, 23.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Li, F.; Wang, B.; Zhu, L.; Li, J.; Zhang, Z.; and Chang, X. 2024a. Cross-Domain Transfer Hashing for Efficient Cross-modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, Y.; Qin, Y.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2024b. RoMo: Robust Unsupervised Multimodal Learning With Noisy Pseudo Labels. *IEEE Transactions on Image Processing*.

Liang, J.; Li, Z.; Cao, D.; He, R.; and Wang, J. 2016. Self-paced cross-modal subspace matching. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 569–578.

Liu, K.; Gong, Y.; Cao, Y.; Ren, Z.; Peng, D.; and Sun, Y. 2024. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *International Joint Conferences on Artificial Intelligence Organization, IJCAI*, 4569–4577.

Peng, Y.; Huang, X.; and Zhao, Y. 2018. An Overview of Cross-media Retrieval: Concepts, Methodologies, Benchmarks and Challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 2372–2385.

Peng, Y.; Zhai, X.; Zhao, Y.; and Huang, X. 2015. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3): 583–596.

Qin, Y.; Sun, Y.; Peng, D.; Zhou, J. T.; Peng, X.; and Hu, P. 2023. Cross-modal Active Complementary Learning with Self-refining Correspondence. *Advances in Neural Information Processing Systems*, 36.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, 251–260.

Shu, Z.; Bai, Y.; Yong, K.; and Yu, Z. 2024. Deep Cross-Modal Hashing With Ranking Learning for Noisy Labels. *IEEE Transactions on Big Data*.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022a. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8135–8153.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022b. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–19.

Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; and Hu, P. 2024a. Dual Self-Paced Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15184–15192.

Sun, Y.; Liu, K.; Li, Y.; Ren, Z.; Dai, J.; and Peng, D. 2024b. Distribution Consistency Guided Hashing for Cross-Modal Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5623–5632.

Sun, Y.; Qin, Y.; Peng, D.; Ren, Z.; Yang, C.; and Hu, P. 2024c. Dual Self-Paced Hashing for Image Retrieval. *IEEE Transactions on Multimedia*.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Tan, W.; Zhu, L.; Li, J.; Zhang, H.; and Han, J. 2022. Teacher-student learning: Efficient hierarchical message aggregation hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 25: 4520–4532.

Wang, L.; Qin, Y.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. Robust Contrastive Cross-modal Hashing with Noisy Labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5752–5760.

Wang, R.; Yu, G.; Zhang, H.; Guo, M.; Cui, L.; and Zhang, X. 2021. Noise-robust deep cross-modal hashing. *Information Sciences*, 581: 136–154.

Wei, J.; Xu, X.; Wang, Z.; and Wang, G. 2021. Meta self-paced learning for cross-modal matching. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3835–3843.

Xu, T.; Liu, X.; Huang, Z.; Guo, D.; Hong, R.; and Wang, M. 2022. Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels. In *Proceedings of the 30th ACM International Conference on Multimedia*, 629–637.

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

Yang, E.; Yao, D.; Liu, T.; and Deng, C. 2022. Mutual quantization for cross-modal search with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7551–7560.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4626–4634.

Zhang, C.; Li, H.; Gao, Y.; and Chen, C. 2023a. Weakly-Supervised Enhanced Semantic-Aware Hashing for Cross-Modal Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 6475–6488.

Zhang, X.; Liu, X.; Nie, X.; Kang, X.; and Yin, Y. 2023b. Semi-supervised semi-paired cross-modal hashing. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; and Shen, H. T. 2023c. Modality-Invariant Asymmetric Networks for Cross-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5091–5104.

Zhen, L.; Hu, P.; Wang, X.; and Peng, D. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10394–10403.

Zhou, K.; Hassan, F. H.; and Hoon, G. K. 2023. The State of the Art for Cross-Modal Retrieval: A Survey. *IEEE Access*.

Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; and Shen, H. T. 2022. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 8838–8851.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2023. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.