

Supplementary Materials: Robust Contrastive Cross-modal Hashing with Noisy Labels

Longan Wang
Sichuan University
Chengdu, China
wanglongan@stu.scu.edu.cn

Yang Qin
Sichuan University
Chengdu, China
qinyang.gm@gmail.com

Yuan Sun
Sichuan University
Chengdu, China
sunyuan_work@163.com

Dezhong Peng^{*}
Sichuan University
Chengdu, China
pengdz@scu.edu.cn

Xi Peng
Sichuan University
Chengdu, China
pengx.gm@gmail.com

Peng Hu[†]
Sichuan University
Chengdu, China
penghu.ml@gmail.com

1 Introduction

In this supplementary material, additional details pertaining to NRCH are furnished for further insight. Delving into greater detail, Section 2 initially presents an exposition on the used datasets and compared baselines. Moreover, Section 3 elaborates on the extensive experimental outcomes, showcasing the Mean Average Precision (MAP) values for both I2T and T2I retrieval tasks. This encompasses a thorough comparison between our NRCH and the state-of-the-art baselines.

2 Dataset and Baseline Description

2.1 Datasets

This section presents the benchmark cross-modal datasets employed in our experiments. Key statistics are summarized in Table 1, and brief descriptions of each dataset are as follows:

MIRFlickr-25K [5] contains 25,000 image-text pairs instances from the Flickr website, classified into 24 distinct semantic categories with multi-label annotations. Specifically, in each image-text instance, images are paired with relevant text descriptions, with the former represented by 4,096-dimensional vectors obtained from a 19-layer VGGNet model [?], and the latter by 1,386-dimensional bag-of-words (BoW) vectors. Following the previous method [6], we excluded instances lacking classification labels, yielding a refined set of 20,015 pairs for our experiments.

IAPR TC-12 [2] encompasses a compilation of 20,000 image-text pairs, each meticulously tagged with 255 varied semantic classes in a multi-label format. Specifically, each image in the pair is encoded as a 4,096-dimensional vector via the pre-trained CNN-F model [?], while the corresponding text manifests as a 2,912-dimensional BoW vector. Uniquely, our experiments utilize the entirety of the dataset without exception.

NUS-WIDE [1] is a publicly accessible collection of web images with their textual tags, comprising 269,648 visuals. Each image-text pairs is precisely categorized within a framework of 81 multi-label semantic classes. In detail, the visual data from each image is captured in a 4,096-dimensional vector, derived through the application of the pre-trained, 19-layer VGGNet architecture. Concurrently, each textual descriptor is encapsulated within a 1,000-dimensional BoW vector space. Moreover, after eliminating entries lacking labels

Table 1: The statistics of four datasets.

Dataset	Train	Test	Database
MIRFlickr-25k	10,000	2,000	18,015
IAPR TC-12	10,000	2,000	18,000
NUS-WIDE	10,500	2,100	188,321
MS-COCO	10,000	5,000	117,218

or textual content, we have selectively harvested 200,421 image-text pairs that represent the 21 most prevalent categories.

MS-COCO [8] is a collection of 123,287 images, each accompanied by five descriptive sentences, and organized into 80 distinct categories. Visual representations within this dataset are encoded into 4,096-dimensional vectors, extracted by the pre-trained 19-layer VGGNet. Different from other datasets, the textual component of each image-text pair is encapsulated into a 300-dimensional vector, derived using the pre-trained Doc2Vec model [7]. Following the exclusion of pairs lacking labels, our experimental dataset comprises 122,218 image-text pairs.

2.2 Baselines

To verify the effectiveness and robustness of our NRCH under label noise, we provide the comparison results with 11 baselines that have published code. An introduction to each referenced baseline is provided in the subsequent text:

DJSRH [9] is an unsupervised cross-modal hash coding method that excels in capturing intrinsic semantic affinities across distinct modalities. DJSRH first introduces a joint-semantics affinity matrix that merges neighborhood information from different modalities, ensuring the preservation of the original data’s neighborhood structure in the binary hash space. Second, DJSRH adeptly optimizes the batch-wise training on hash code generating by mirroring the joint-semantics relationships. Thus, it can reconstruct the specific similarity values and offer an advantage over traditional Laplacian constraints that only preserve the similarity order.

DGCPN [13] is an unsupervised cross-modal hashing framework that leverages deep graph-based techniques to enhance the accuracy of data similarity measures. DGCPN introduces a unique graph-neighbor coherence approach, integrating three distinct similarity types (*i.e.*, graph-neighbor coherence, coexistent similarity, intra- and inter-modality consistency) to preserve comprehensive data relationships. By employing a half-real and half-binary optimization strategy, DGCPN effectively minimizes quantization errors.

^{*}Dezhong Peng is also with Sichuan Newstrong UHD Video Technology Co., Ltd.

[†]Corresponding author

Table 2: The performance comparison in terms of MAP scores on the NUS-WIDE dataset. The highest and the second-highest scores are in bold and underlined, respectively. Methods above the dotted line are unsupervised, while those below are supervised.

Task	Method	20%				50%				80%			
		16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
I2T	DJSRH (ICCV'19)	0.418	0.455	0.472	0.500	0.418	0.455	0.472	0.500	0.418	0.455	0.472	0.500
	DGCPN (AAAI'21)	0.569	0.593	0.620	0.633	0.569	0.593	0.620	0.633	0.569	0.593	0.620	0.633
	UCCH (TAPMI'23)	0.571	0.594	0.621	0.637	0.571	0.594	<u>0.621</u>	<u>0.637</u>	<u>0.571</u>	<u>0.594</u>	<u>0.621</u>	<u>0.637</u>
	DCMH (CVPR'17)	0.479	0.483	0.478	0.449	0.407	0.405	0.405	0.405	0.381	0.392	0.397	0.392
	ADAH (ECCV'18)	0.527	0.495	0.516	0.536	0.479	0.493	0.487	0.493	0.450	0.463	0.470	0.474
	CPAH (TIP'20)	0.552	0.566	0.560	0.579	0.479	0.500	0.514	0.516	0.466	0.464	0.456	0.468
	PIP (SIGIR'21)	0.554	0.595	0.594	0.597	0.562	0.585	0.589	0.588	0.570	0.591	0.590	0.596
	CMMQ (CVPR'22)	<u>0.632</u>	<u>0.638</u>	<u>0.643</u>	<u>0.654</u>	<u>0.572</u>	<u>0.595</u>	0.611	0.614	0.536	0.581	0.588	0.611
	DCHUC (TKDE'22)	0.601	0.596	0.583	0.570	0.602	0.584	0.586	0.576	0.570	0.589	0.583	0.574
	MIAN (TKDE'23)	0.570	0.579	0.582	0.582	0.433	0.442	0.442	0.435	0.386	0.381	0.394	0.384
	LtCMH (AAAI'23)	0.513	0.521	0.553	0.572	0.491	0.515	0.543	0.561	0.487	0.513	0.561	0.571
	Our NRCH	0.657	0.679	0.683	0.685	0.628	0.653	0.661	0.668	0.604	0.610	0.623	0.640
T2I	DJSRH (ICCV'19)	0.418	0.460	0.480	0.511	0.418	0.460	0.480	0.511	0.418	0.460	0.480	0.511
	DGCPN (AAAI'21)	0.581	0.601	0.628	0.635	0.581	0.601	0.628	0.635	0.581	0.601	0.628	0.635
	UCCH (TAPMI'23)	0.582	0.602	0.629	0.637	0.582	0.602	0.629	0.637	<u>0.582</u>	<u>0.602</u>	<u>0.629</u>	<u>0.637</u>
	DCMH (CVPR'17)	0.506	0.494	0.486	0.460	0.438	0.444	0.447	0.446	0.419	0.441	0.443	0.439
	ADAH (ECCV'18)	0.520	0.488	0.520	0.527	0.449	0.469	0.465	0.461	0.419	0.435	0.439	0.447
	CPAH (TIP'20)	0.561	0.571	0.559	0.578	0.478	0.492	0.509	0.508	0.472	0.474	0.469	0.472
	PIP (SIGIR'21)	0.557	0.610	0.604	0.596	0.570	0.603	0.594	0.600	0.581	0.601	0.600	0.608
	CMMQ (CVPR'22)	<u>0.633</u>	<u>0.637</u>	<u>0.649</u>	<u>0.656</u>	<u>0.595</u>	<u>0.605</u>	0.608	0.611	0.558	0.583	0.610	0.613
	DCHUC (TKDE'22)	0.605	0.605	0.593	0.590	0.590	0.591	0.593	0.585	0.581	0.589	0.594	0.591
	MIAN (TKDE'23)	0.609	0.618	0.625	0.631	0.457	0.471	0.470	0.468	0.396	0.411	0.414	0.406
	LtCMH (AAAI'23)	0.471	0.490	0.535	0.561	0.453	0.486	0.532	0.548	0.527	0.561	0.548	0.558
	Our NRCH	0.659	0.679	0.682	0.684	0.650	0.660	0.673	0.686	0.606	0.611	0.632	0.643

UCCH [4] is an advanced framework for unsupervised cross-modal hashing (CMH) that integrates contrastive learning (CL). It overcomes binary optimization issues through a novel momentum optimizer that makes hashing operations learnable within CL, thereby enhancing retrieval performance without binary-continuous relaxation. Additionally, it introduces the Cross-modal Ranking Learning (CRL) loss to mitigate the impact of false-negative pairs (FNPs) by leveraging global discrimination, thus avoiding the overemphasis on FNPs and neglect of true negatives, positioning it as a pioneering method in contrastive hashing.

DCMH [6] is a deep supervised cross-modal hashing method designed to enhance multimedia retrieval through an integrated approach that combines feature and hash-code learning within a unified framework. Utilizing deep neural networks tailored for each modality, DCMH performs end-to-end feature learning from scratch, eschewing the need for hand-crafted features. This integration ensures that the learned features are highly compatible with the hash-code learning procedure, leading to improved performance in similarity search tasks across different media types.

ADAH [15] is an adversarial supervised hashing network designed to pinpoint content similarities across multi-modal data by leveraging an attention mechanism for enhanced focus on pertinent data segments. It features a tripartite architecture: a feature learning module for extracting foundational representations, an attention module that discerns key features via attention masks, and a hashing module dedicated to crafting hash functions that encapsulate cross-modal similarities. ADAH employs an adversarial training strategy where the attention component seeks to challenge the hashing module's ability to recognize similarities pertaining

to the unattended features, thereby ensuring the hashing process captures the essence of both attended and unattended data aspects.

CPAH [11] is a supervised deep hashing approach engineered to bridge the modality gap and harness semantic consistency across different modalities for enhanced cross-modal retrieval. It features a consistency refined module (CR) that segregates multi-modal representations into modality-common and modality-private components. Complementing this, a multi-task adversarial learning module (MA) aligns the modality-common representations in terms of feature distribution and semantic consistency, paving the way for generating compact and semantically potent hash codes conducive to efficient retrieval.

PIP [14] is a supervised privacy protection framework designed to safeguard sensitive information in large-scale multi-modal retrieval systems. It disrupts malicious retrieval attempts by infusing original data with subtle adversarial perturbations, rendering sensitive content untraceable by unauthorized parties. Simultaneously, PIP maintains a robust multi-modal retrieval model for legitimate applications, demonstrating resilience to these perturbations. This pioneering work orchestrates a strategic two-player game that aligns domain distributions and graphs both within and across modalities, while leveraging a high-level similarity matrix for refined learning guidance.

CMMQ [12] is a robust cross-modal hashing framework tailored to effectively handle multimodal search in the presence of noisy labels. It introduces a proxy-based contrastive (PC) loss that acts to bridge the gap between different modalities, fostering joint network training. The framework also features a novel small-loss sample selection mechanism driven by the PC loss in conjunction with a

Table 3: The performance comparison in terms of MAP scores on the MS-COCO dataset. The highest and the second-highest scores are in bold and underlined, respectively. Methods above the dotted line are unsupervised, while those below are supervised.

Task	Method	20%				50%				80%			
		16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
I2T	DJSRH (ICCV'19)	0.480	0.521	0.549	0.573	0.480	0.521	0.549	0.573	0.480	0.521	0.549	0.573
	DGCPN (AAAI'21)	0.580	0.612	0.624	0.633	0.580	0.612	0.624	0.633	0.580	0.612	0.624	0.633
	UCCH (TAPMI'23)	0.570	0.582	0.597	0.624	0.570	0.582	0.597	0.624	0.570	0.582	0.597	0.624
	DCMH (CVPR'17)	0.536	0.573	0.577	0.582	0.473	0.467	0.477	0.458	0.391	0.373	0.350	0.341
	ADAH (ECCV'18)	0.476	0.490	0.485	0.499	0.479	0.484	0.489	0.498	0.480	0.479	0.488	0.495
	CPAH (TIP'20)	0.551	0.594	0.601	0.603	0.555	0.548	0.549	0.550	0.511	0.515	0.509	0.510
	PIP (SIGIR'21)	0.535	0.581	0.585	0.601	0.512	0.561	0.584	0.592	0.501	0.525	0.563	0.601
	CMMQ (CVPR'22)	0.620	<u>0.637</u>	<u>0.642</u>	<u>0.640</u>	<u>0.582</u>	<u>0.630</u>	<u>0.625</u>	<u>0.635</u>	<u>0.597</u>	<u>0.615</u>	<u>0.626</u>	<u>0.634</u>
	DCHUC (TKDE'22)	0.571	0.504	0.546	0.509	0.558	0.515	0.522	0.503	0.551	0.541	0.519	0.545
	MIAN (TKDE'23)	0.559	0.569	0.596	0.573	0.489	0.490	0.511	0.526	0.438	0.451	0.477	0.455
	LtCMH (AAAI'23)	0.527	0.554	0.607	0.619	0.531	0.538	0.598	0.627	0.530	0.576	0.594	0.604
	Our NRCH	0.639	0.647	0.675	0.677	0.645	0.651	0.682	0.685	0.643	0.660	0.672	0.691
T2I	DJSRH (ICCV'19)	0.491	0.533	0.557	0.586	0.491	0.533	0.557	0.586	0.491	0.533	0.557	0.586
	DGCPN (AAAI'21)	0.603	0.615	0.623	0.628	0.603	0.615	0.623	0.628	0.603	0.615	0.623	0.628
	UCCH (TAPMI'23)	0.569	0.581	0.591	0.622	0.569	0.581	0.591	0.622	0.569	0.581	0.591	0.622
	DCMH (CVPR'17)	0.553	0.590	0.594	0.616	0.483	0.476	0.473	0.457	0.397	0.396	0.365	0.360
	ADAH (ECCV'18)	0.451	0.466	0.468	0.466	0.458	0.456	0.456	0.475	0.445	0.465	0.457	0.460
	CPAH (TIP'20)	0.543	0.602	0.610	0.622	0.531	0.551	0.562	0.595	0.519	0.521	0.522	0.524
	PIP (SIGIR'21)	0.541	0.569	0.591	0.592	0.514	0.547	0.613	0.614	0.499	0.519	0.566	0.581
	CMMQ (CVPR'22)	0.605	<u>0.640</u>	<u>0.641</u>	<u>0.642</u>	<u>0.609</u>	<u>0.616</u>	<u>0.624</u>	<u>0.629</u>	<u>0.605</u>	<u>0.621</u>	<u>0.625</u>	<u>0.636</u>
	DCHUC (TKDE'22)	0.545	0.482	0.558	0.459	0.546	0.480	0.453	0.468	0.548	0.429	0.463	0.567
	MIAN (TKDE'23)	0.582	0.578	0.611	0.600	0.507	0.509	0.535	0.564	0.452	0.467	0.481	0.485
	LtCMH (AAAI'23)	0.580	0.624	0.624	0.640	0.564	0.606	0.622	0.628	0.588	0.619	0.624	0.627
	Our NRCH	0.635	0.651	0.662	0.670	0.649	0.674	0.679	0.688	0.649	0.656	0.679	0.689

mutual quantization loss, which together enhance the selection of reliable samples for model training. This mutual quantization loss further aligns modalities, optimizing the sample selection process to ensure that only the most consistent examples contribute to the learning, thereby improving the robustness of cross-modal retrieval against label noise.

DCHUC [10] is a supervised deep cross-modal hashing framework that excels in learning unified hash codes and optimizing hashing functions simultaneously. It utilizes an iterative optimization algorithm to master the cross-modal retrieval process, ensuring that image-text pairs are hashed cohesively across different modalities. This process not only refines the hash code learning with feedback from function optimization but also enhances retrieval precision, setting a new benchmark in the field.

MIAN [16] is a supervised modality-invariant asymmetric architecture designed for cross-modal hashing, which adeptly navigates the semantic and heterogeneity gaps across different modalities. It employs an intra-modal asymmetric network to probabilistically learn query-vs-all pairwise similarities within each modality, while an inter-modal asymmetric network captures the cross-modal semantic correlations via maximum inner product search. This architecture not only integrates pairwise, piecewise, and transformed semantics into a cohesive semantic-preserving hashing code framework but also features a modality alignment network. MIAN refines visual features and maximizes the conditional information bottleneck, effectively bridging modality discrepancies and fostering the generation of discriminative, modality-invariant hash codes.

LtCMH [3] is a supervised Cross Modal Hashing technique specifically formulated to address the challenges of imbalanced

multi-modal data with long-tail distribution. It utilizes auto-encoders to effectively segregate and enhance the individuality and commonality of different modalities. By minimizing the dependency on the individuality of each modality and boosting their commonalities, LtCMH dynamically integrates these aspects with direct features from respective modalities to construct meta features. These enriched meta features better represent tail labels and are subsequently binarized to generate effective hash codes, optimizing retrieval performance across diverse data distributions.

3 More Comparison

To further demonstrate the effectiveness and robustness of our method, comprehensive MAP score outcomes for the I2T (text-to-image retrieval) and T2I (text-to-image retrieval) tasks are detailed and can be reviewed in Tables 2 to 5. To maintain consistency with the experiments reported in the main text, we conducted our experiments on the same datasets, with identical parameter configurations and against the same baselines. Deriving insights from the additional experimental result in Tables 2 to 5, the subsequent observations can be formulated:

- As the noise rate increases, the performance of these supervised methods [3, 6, 10–12, 14–16] degrades severely on both I2T and T2I tasks. In comparison, the unsupervised methods [4, 9, 13] above the dotted line in the tables seem to have a certain degree of robustness. However, it is still difficult to achieve further performance improvement due to the lack of corresponding measures against noisy labels.
- Among all these baseline methods, CMMQ [12] still stands out for its resistance to noisy labels on the NUS-WIDE and

Table 4: The performance comparison in terms of MAP scores on the IAPR TC-12 dataset. The highest and the second-highest scores are in bold and underlined, respectively. Methods above the dotted line are unsupervised, while those below are supervised.

Task	Method	20%				50%				80%			
		16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
I2T	DJSRH (ICCV'19)	0.366	0.394	0.417	0.431	0.366	0.394	0.417	0.431	0.366	0.394	0.417	0.431
	DGCPN (AAAI'21)	0.416	0.447	0.466	0.467	0.416	0.447	0.466	0.467	0.416	0.447	0.466	0.467
	UCCH (TAPMI'23)	0.410	0.449	0.464	0.467	0.410	0.449	0.464	0.467	0.410	<u>0.449</u>	0.464	0.467
	DCMH (CVPR'17)	0.422	0.421	0.420	0.412	0.401	0.405	0.397	0.388	0.349	0.356	0.348	0.341
	ADAH (ECCV'18)	0.423	0.437	0.440	0.438	0.414	0.411	0.434	0.436	0.414	0.415	0.420	0.425
	CPAH (TIP'20)	0.457	<u>0.463</u>	0.465	0.471	0.440	0.455	0.461	0.462	0.418	0.448	0.452	0.455
	PIP (SIGIR'21)	0.433	0.450	0.457	<u>0.475</u>	0.412	0.451	0.462	<u>0.474</u>	<u>0.424</u>	0.444	0.454	<u>0.470</u>
	CMMQ (CVPR'22)	0.409	0.445	<u>0.468</u>	<u>0.473</u>	0.410	0.445	0.456	<u>0.470</u>	<u>0.422</u>	0.434	0.455	<u>0.456</u>
	DCHUC (TKDE'22)	<u>0.461</u>	0.444	0.449	0.447	<u>0.451</u>	<u>0.456</u>	0.457	0.440	0.423	0.432	0.436	0.434
	MIAN (TKDE'23)	0.444	0.447	0.462	0.472	<u>0.424</u>	<u>0.429</u>	0.430	0.452	0.403	0.421	0.434	0.439
	LiCMH (AAAI'23)	0.412	0.428	0.440	0.449	0.416	0.427	0.441	0.450	0.411	0.432	0.437	0.448
	Our NRCH	0.498	0.526	0.544	0.549	0.495	0.525	0.541	0.543	0.490	0.518	0.532	0.540
T2I	DJSRH (ICCV'19)	0.371	0.399	0.425	0.438	0.371	0.399	0.425	0.438	0.371	0.399	0.425	0.438
	DGCPN (AAAI'21)	0.427	0.449	0.462	0.467	0.427	0.449	0.462	0.467	0.427	0.449	0.462	0.467
	UCCH (TAPMI'23)	0.426	0.450	0.467	0.470	0.426	0.450	0.467	0.470	0.426	0.450	0.467	0.470
	DCMH (CVPR'17)	0.427	0.436	0.412	0.421	0.427	0.417	0.412	0.400	0.390	0.384	0.383	0.376
	ADAH (ECCV'18)	0.420	0.427	0.458	0.457	0.402	0.423	0.452	0.451	0.414	0.406	0.438	0.441
	CPAH (TIP'20)	0.443	<u>0.469</u>	0.467	0.475	0.441	0.451	0.452	0.462	0.427	0.450	0.459	0.460
	PIP (SIGIR'21)	<u>0.444</u>	<u>0.453</u>	<u>0.468</u>	<u>0.483</u>	0.417	<u>0.454</u>	<u>0.469</u>	<u>0.480</u>	<u>0.428</u>	<u>0.451</u>	<u>0.468</u>	<u>0.477</u>
	CMMQ (CVPR'22)	0.427	0.446	0.466	0.464	0.417	<u>0.445</u>	<u>0.467</u>	0.472	0.426	<u>0.442</u>	<u>0.457</u>	0.464
	DCHUC (TKDE'22)	0.437	0.457	0.449	0.448	<u>0.443</u>	0.444	0.437	0.456	0.426	0.445	0.465	0.459
	MIAN (TKDE'23)	0.435	0.442	0.448	0.398	<u>0.420</u>	<u>0.429</u>	0.444	0.438	0.402	0.417	0.423	0.423
	LiCMH (AAAI'23)	0.429	0.443	0.457	0.464	0.422	0.445	0.454	0.464	0.417	0.445	0.451	0.462
	Our NRCH	0.498	0.528	0.549	0.555	0.494	0.526	0.543	0.551	0.487	0.518	0.535	0.547

Table 5: The performance comparison in terms of MAP scores on the MIRFlickr-25K dataset. The highest and the second-highest scores are in bold and underlined, respectively. Methods above the dotted line are unsupervised, while those below are supervised.

Task	Method	20%				50%				80%			
		16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
I2T	DJSRH (ICCV'19)	0.603	0.621	0.636	0.650	0.603	0.621	0.636	0.650	0.603	0.621	0.636	0.650
	DGCPN (AAAI'21)	0.698	0.699	0.711	0.723	0.698	0.699	0.711	0.723	0.698	0.699	0.711	0.723
	UCCH (TAPMI'23)	0.698	0.721	0.724	0.728	0.698	0.721	0.724	0.728	0.698	0.721	0.724	0.728
	DCMH (CVPR'17)	0.695	0.691	0.695	0.702	0.634	0.623	0.619	0.606	0.628	0.600	0.589	0.588
	ADAH (ECCV'18)	0.724	0.729	0.735	0.733	0.713	0.718	0.717	0.714	0.602	0.607	0.610	0.603
	CPAH (TIP'20)	0.697	0.694	0.691	0.689	0.660	0.660	0.666	0.647	0.619	0.654	0.642	0.627
	PIP (SIGIR'21)	0.685	0.692	0.694	0.710	0.667	0.697	0.700	0.710	0.684	0.676	0.704	0.709
	CMMQ (CVPR'22)	0.724	0.727	0.735	0.736	0.692	0.720	0.719	0.722	0.691	0.715	0.720	0.723
	DCHUC (TKDE'22)	0.742	0.737	0.736	0.730	<u>0.737</u>	<u>0.740</u>	<u>0.732</u>	<u>0.729</u>	<u>0.727</u>	<u>0.722</u>	<u>0.735</u>	<u>0.734</u>
	MIAN (TKDE'23)	0.748	0.749	0.756	0.760	0.676	0.684	0.685	0.685	0.648	0.657	0.649	0.633
	LiCMH (AAAI'23)	0.712	0.728	0.734	0.741	0.697	0.718	0.726	0.728	0.687	0.707	0.711	0.721
	Our NRCH	0.754	0.768	0.770	0.778	0.747	0.762	0.770	0.772	0.740	0.748	0.761	0.762
T2I	DJSRH (ICCV'19)	0.613	0.618	0.638	0.641	0.613	0.618	0.638	0.641	0.613	0.618	0.638	0.641
	DGCPN (AAAI'21)	0.684	0.690	0.704	0.712	0.684	0.690	0.704	0.712	0.684	0.690	0.704	0.712
	UCCH (TAPMI'23)	0.682	0.704	0.706	0.709	0.682	0.704	0.706	0.709	0.682	0.704	0.706	0.709
	DCMH (CVPR'17)	0.711	0.711	0.711	0.695	0.669	0.667	0.658	0.654	0.629	0.644	0.646	0.644
	ADAH (ECCV'18)	0.723	0.725	0.737	0.733	0.700	0.705	0.720	0.710	0.602	0.621	0.602	0.612
	CPAH (TIP'20)	0.711	0.709	0.712	0.714	0.682	0.685	0.690	0.683	0.648	0.685	0.673	0.668
	PIP (SIGIR'21)	0.680	0.691	0.695	0.698	0.658	0.697	0.691	0.694	0.686	0.689	0.700	0.696
	CMMQ (CVPR'22)	0.728	0.732	0.738	0.741	0.704	0.716	0.722	0.725	0.698	0.707	0.712	0.716
	DCHUC (TKDE'22)	0.738	0.736	0.739	0.738	<u>0.732</u>	<u>0.741</u>	<u>0.737</u>	<u>0.732</u>	0.719	0.720	0.734	<u>0.731</u>
	MIAN (TKDE'23)	<u>0.727</u>	<u>0.733</u>	<u>0.743</u>	0.746	0.679	<u>0.694</u>	0.695	0.700	0.669	0.669	0.668	0.678
	LiCMH (AAAI'23)	0.700	0.708	0.716	0.721	0.679	0.672	0.709	0.720	0.655	0.708	0.717	0.722
	Our NRCH	0.741	0.756	0.759	0.758	0.734	0.747	0.754	0.759	0.722	0.727	0.746	0.746

MS-COCO datasets for both I2T and T2I tasks, sharing similarities with our NRCH in terms of noise segregation components. But unlike this, our NRCH achieves even more promising performance by improving the robustness of the loss and performing reliable dynamic sample selection.

- All in all, our NRCH surpasses all baselines on four datasets and outperforms the best baselines by 3.3%/2.4%, 4.6%/4.4%, 6.6%/5.9%, and 1.3%/0.3%, respectively, in the most challenging scenarios (i.e., the noise rate is 80% noise and code length is 16 bits) for I2T and T2I tasks. This is enough to prove the effectiveness and superiority of our NRCH against noisy labels.

References

- [1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. 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*. 1–9.
- [2] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer vision and image understanding* 114, 4 (2010), 419–428.
- [3] Zijun Gao, Jun Wang, Guoxian Yu, Zhongmin Yan, Carlotta Domeniconi, and Jinglin Zhang. 2023. Long-tail cross modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7642–7650.
- [4] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. 2023. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3877–3889. <https://doi.org/10.1109/TPAMI.2022.3177356>
- [5] Mark J Huiskes and Michael S Lew. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. 39–43.
- [6] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3232–3240.
- [7] Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368* (2016).
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13. Springer, 740–755.
- [9] Shupeng Su, Zhisheng Zhong, and Chao Zhang. 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3027–3035.
- [10] Rong-Cheng Tu, Xian-Ling Mao, Bing Ma, Yong Hu, Tan Yan, Wei Wei, and Heyan Huang. 2020. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering* 34, 2 (2020), 560–572.
- [11] De Xie, Cheng Deng, Chao Li, Xianglong Liu, and Dacheng Tao. 2020. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Transactions on Image Processing* 29 (2020), 3626–3637.
- [12] Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng. 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.
- [13] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4626–4634.
- [14] Peng-Fei Zhang, Yang Li, Zi Huang, and Hongzhi Yin. 2021. Privacy protection in deep multi-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 634–643.
- [15] Xi Zhang, Hanjiang Lai, and Jiashi Feng. 2018. Attention-aware deep adversarial hashing for cross-modal retrieval. In *Proceedings of the European conference on computer vision (ECCV)*. 591–606.
- [16] Zheng Zhang, Haoyang Luo, Lei Zhu, Guangming Lu, and Heng Tao Shen. 2022. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 5091–5104.