Deep Supervised Adversarial Robust Hashing for Retrieval

Journal:	*Transactions on Pattern Analysis and Machine Intelligence*
Manuscript ID	TPAMI-2024-03-0745
Manuscript Type:	Regular
Keywords:	Hashing, Adversarial Robustness, Cross-modal Retrieval, Image Retrieval, Deep Learning

SCHOLARONE™
Manuscripts

Deep Supervised Adversarial Robust Hashing for Retrieval

Xingwei Zhang, Gang Zhou, Xiaolong Zheng, *Member, IEEE*, Wenji Mao, Liang Wang, *Fellow, IEEE* and Daniel Dajun Zeng, *Fellow, IEEE*

***Abstract*—Hashing has emerged as an efficient mechanism for similarity searching, owing to its computational efficiency. With the evolution of deep learning (DL), DL-based hashing methods have exhibited remarkable performance in multi-modal and high-dimensional retrieval tasks. Despite these advancements, recent studies have revealed vulnerabilities of deep learning (DL) structures to adversarial attacks, leading to an intensified focus on adversarial robustness within the DL community. However, most of these studies predominantly concentrate on supervised classification tasks. As a result, the methodologies developed in current adversarial robustness research are not directly applicable to retrieval tasks. To bridge this research gap, we propose Deep Supervised Adversarial Robust Hashing (DSARH), an end-to-end hashing framework meticulously crafted to extract robust features from high-dimensional data, thereby ensuring reliable retrieval performance. Through extensive experiments conducted on diverse cross-modal and image retrieval benchmarks, we demonstrate that existing deep hashing models are susceptible to vulnerability issues. In contrast, our proposed DSARH method substantially bolsters the robustness of deep hashing models against a spectrum of adversarial attacks across both the image-text and image retrieval tasks. Furthermore, DSARH outperforms the state-of-the-art counterparts, delivering superior cross-modal retrieval performance on large-scale image-text retrieval benchmarks. This underscores the critical importance of adversarial robustness research in tackling the challenges inherent to multi-modal retrieval issues.**

***Index Terms*—Hashing, adversarial robustness, cross-modal retrieval, image retrieval, deep learning**

I. INTRODUCTION

THE rapid advancement of social networks and Web media content has led to the accumulation of vast amounts of data from diverse modalities on servers, necessitating an immediate need for multi-modal similarity search capabilities. In recent years, research on similarity retrieval, which focuses on extracting semantically relevant cross-modal information based on existing query samples, has garnered significant attention and emerged as a pivotal research direction in the fields of artificial intelligence and data science [1]. Generally, the dimensions of semantically related multi-modal data on contemporary social networks can vary significantly, such as photos on FLICKR blogs paired with their textual descriptions. Consequently, effectively extracting semantic information from high-dimensional data, regardless of its heterogeneity, presents a fundamental challenge for accurately quantifying semantic similarity and achieving precise similarity retrieval in multi-modal data contexts [2]. In practice, approximate nearest neighbor (ANN) methods achieve equitable semantic comparisons by constructing a shared semantic-preserving subspace for diverse data modalities. Hashing-based retrieval methods further optimize this process by replacing real values mapped in the subspace with binary hash codes, thereby reducing both computational complexity and storage costs [3]. Additionally, incorporating semantic correlation information into model development through supervised hashing methods can further enhance the retrieval performance of hashing techniques [4].

In recent years, the adoption of deep learning (DL) architectures for automatic feature extraction has significantly bolstered the effectiveness of hashing models for retrieval tasks. DL architectures can discern hidden semantic features from high-dimensional data, enabling them to make decisions on complex issues at a level comparable to human expertise [5]. A multitude of deep hashing models have been proposed utilizing DL architectures, which have been verified to surpass traditional hashing models designed using handcrafted features [6]. Furthermore, traditional hashing models necessitate identical encoding for out-of-sample sets, and thus limit their generalization capability to unseen samples without prior knowledge. In contrast, DL models have demonstrated great advantages in automatic feature learning and representation learning capabilities. Therefore, DL-based hashing models are increasingly becoming the predominant approach in state-of-the-art similarity search studies [7].

While deep learning (DL) architectures have shown promising capabilities, their reliability is not always as robust as anticipated. Recent studies have revealed that models designed using DL architectures are susceptible to small, imperceptible attacks known as adversarial perturbations [8]. Well-crafted adversarial perturbations have also been observed to transfer between different models and datasets. This unveils a more profound issue: the underlying principles governing regularly trained DL models diverge significantly from the basis of human decision-making processes [9]. The susceptibility of DL models to adversarial attacks has raised concerns about their applicability to real-world tasks that require high security demands, and has led to various proposals for defense mechanisms and robust training strategies aimed at bolstering their resilience against such attacks [10]. The benefits of robust training extend beyond defense; that is, robustly trained models have shown to extract features aligned with human recognition, thereby offering more reliable performance and improved transferability compared to conventionally trained

Xingwei Zhang, Xiaolong Zheng, Wenji Mao, Liang Wang and Daniel Dajun Zeng are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China(e-mail: zhangxingwei2019@ia.ac.cn; xiaolong.zheng@ia.ac.cn; wenji.mao@ia.ac.cn; wangliang@nlpr.ia.ac.cn; dajun.zeng@ia.ac.cn) *(Corresponding author: Xiaolong Zheng.)*

Gang Zhou is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhougang2023@bupt.edu.cn).

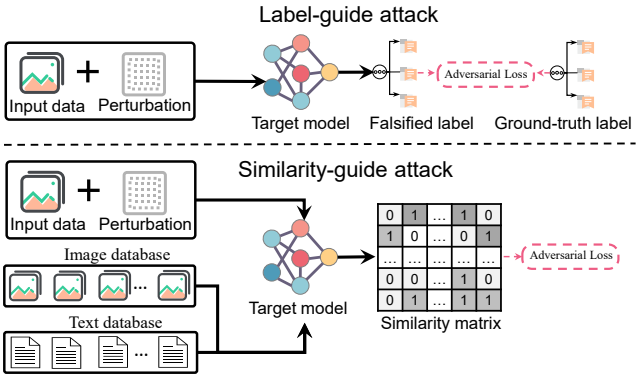


Fig. 1: The principles underlying adversarial perturbation generation methods vary between classification and cross-modal retrieval tasks. In classification tasks, adversarial perturbations depend on ground-truth labels to introduce interference information. Conversely, for cross-modal retrieval tasks, adversarial perturbations are crafted using alternative data modalities, such as textual descriptions or semantic similarity matrices, to offer semantic insights.

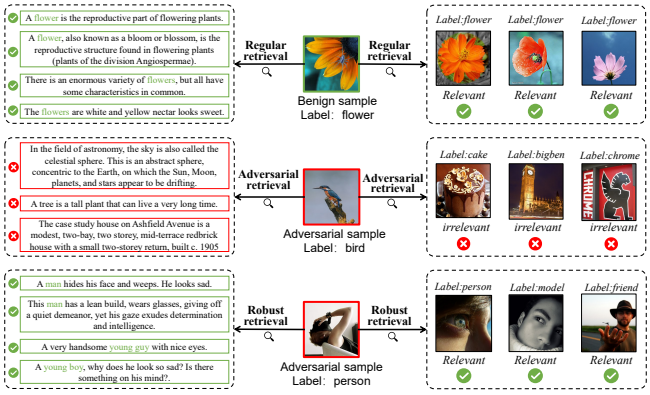


Fig. 2: Several retrieval examples of models under attacks. Adding human-unrecognized attacks on images can effectively misguide well-trained image-text and image retrieval models. These attacks can alter the semantic correlation between images and textual descriptions, leading to degraded retrieval performance.

models. The fundamental mechanisms of widely-used robust training methods enable models to make accurate predictions even under strong perturbations. This ability to bypass shortcuts in samples, while focusing on highly predictive features, serves to enhance the generalization and reliability of models [9].

However, it's worth noting that the generation of adversarial perturbations and robust training methods have primarily been designed for classification tasks that involve ground-truth concept information. The robustness of retrieval frameworks has not been adequately assessed. While some effective methods have been proposed for generating adversarial attacks on retrieval models [11], [12], as far as we know, robust training methods suitable for large-scale retrieval tasks are still lacking. The main challenge lies in the training methodology for retrieval scenarios. Traditional adversarial training is conducted in an end-to-end manner by introducing perturbations to training samples, so as to mislead the model predictions based on the samples' ground-truth concepts [13]. However, as illustrated in Fig. 1, the optimization of retrieval models depends primarily on the similarity matrix derived from the training samples to provide semantic correlation information, rather than relying on ground-truth labels [1].

To address the above challenges, in this paper, we introduce a novel method called Deep Supervised Adversarial Robust Hashing (DSARH) designed to learn robust binary representations from high-dimensional data. Our approach incorporates end-to-end adversarial training tailored for supervised retrieval, integrating the similarity matrix into the process of generating adversarial perturbations during training. Unlike traditional classification tasks, our method generates perturbations based on semantic information obtained from other modalities, in conjunction with their corresponding similarity correlations. We assume that the learned hash codes encapsulate the semantic information of the respective modalities. By minimizing the distance between hash codes of deceptive target samples and semantically irrelevant samples, our method can effectively generate perturbations that disrupt the retrieval performance of the targeted models. Specifically, we employ the learned binary codes from reference samples in conjunction with the similarity matrix to formulate a reference direction. This direction is then combined with the learned hash codes of target samples to determine the final direction of the perturbation. Furthermore, we devise an alternative form of perturbation based on the supervised information of intra-modality samples, whose direction is aligned with the learned hash codes of the target samples. These perturbations can be efficiently generated using the back-propagated gradient of the learned hash codes on raw samples, facilitating their integration into end-to-end training processes. Notably, DSARH consists of two adversarial training components. This dual-component approach enables the learning of robust features to construct compact binary hash codes that capture the semantic correlations and concept information respectively.

Further, DSARH can be effectively utilized for both cross-modal and image retrieval tasks. Cross-modal retrieval tasks, particularly those involving image-text pairs, are among the most practical scenarios [7]. In such tasks, images often possess high-dimensional features, whereas textual data typically have much lower dimensions, leading to significant heterogeneity challenges. It is worth noting that existing research on adversarial robustness predominantly focuses on computer vision tasks [9]. And the textual data in commonly-used cross-modal retrieval datasets are always pre-encoded by human experts [6]. Thus, the robustness analysis for cross-modal retrieval presented in this paper focuses exclusively on image data. Specifically, the robustness of deep hashing models for cross-modal retrieval is assessed based on their performance when exposed to attacks applied to image samples. Similar to those for image classification tasks, the magnitudes of these attacks are constrained to be sufficiently small to remain imperceptible to the human eye [10].

Through extensive experiments and comparisons with the state-of-the-art deep hashing methods, we demonstrate that DSARH consistently outperforms other approaches on several large-scale cross-modal retrieval benchmarks. This highlights the efficacy of adversarial training applied to images in

enabling models to identify more precise semantic features, thereby establishing a more reliable correlation between heterogeneous data types. Adversarial training has been proven to reveal features in images that are more aligned with human recognition compared to conventional training methods [10], potentially assisting retrieval models in bridging the semantic gap between images and human-encoded text data. In addition to evaluating the performance of DSARH, we also replicate several existing deep cross-modal and image retrieval hashing methods to assess their robustness against adversarial attacks. Our findings indicate that the performance of all benchmark models experience significant degradation when exposed to the generated attacks, as illustrated in Fig.2. Nonetheless, DSARH continues to deliver competitive retrieval performance across both cross-modal and image retrieval tasks. This suggests that hash codes learned by conventional deep hashing models may not be reliable when facing agnostic attacks, emphasizing the imperative need for robustness research in retrieval tasks. Furthermore, we validate the robust generalization capabilities of DSARH by verifying its ability to defend against various types of perturbations specifically designed for retrieval tasks. It is worth noting that adversarial training has been found to inevitably result in performance degradation across various image classification tasks [10]. In this study, a similar phenomenon is also observed in large-scale image retrieval tasks. This suggests that additional technologies are necessary to strike a balance between regular and robust performance in pure computer vision tasks. In contrast, we demonstrate that DSARH performs exceptionally well in both regular and robust scenarios, emphasizing the critical role of robustness in deep hashing models for cross-modal retrieval tasks. In summary, our main contributions can be summarized as follows:

- We introduce DSARH, a novel end-to-end adversarial training framework designed for retrieval tasks. The primary objective of DSARH is to bolster the reliability of retrieval models by equipping them with the capability to defend against meticulously crafted adversarial perturbations.
- We present a gradient-based perturbation generation mechanism tailored for supervised retrieval tasks, by exploring the learned hash codes and the samples' similarity matrix. These perturbations can be effectively obtained during the training phase of hashing models. The meticulously crafted perturbations can also be utilized to evaluate the robustness of retrieval models.
- Through comprehensive experiments on publicly available cross-modal and image retrieval benchmarks, we underscore the effectiveness of DSARH in bolstering robustness against various attacks. Moreover,we find that DSARH proposed outperforms state-of-the-art cross-modal retrieval hashing methods, showcasing superior retrieval performance across a variety of cross-modal tasks.

The remainder of this paper is organized as follows. In Section II, we review research on deep hashing methods for retrieval and adversarial robustness. Section III presents the DSARH framework and its optimization scheme on cross-modal and image-retrieval tasks. Section IV details experiments and comparisons with state-of-the-art methodologies on public benchmarks. We conclude this paper in Section V.

II. RELATED WORK

In this section, we briefly review the deep hashing methods for retrieval and adversarial robustness research.

A. Deep Hashing Methods for Retrieval

The fundamental principle of similarity retrieval mechanisms posits the existence of an ideal subspace wherein data from different modalities that share identical semantic meanings exhibit closer distances compared to irrelevant samples [3]. Subspace-based retrieval methodologies strive to identify a universal space capable of mapping the semantic essence of heterogeneous samples. Within this space, semantic similarity is gauged by the distance between representations of corresponding samples. In pursuit of this objective, various efficient techniques have been introduced to navigate the semantic-preserving space, including Latent Canonical Correlation Analysis (CCA), Latent Subspace Analysis (LSA), and Correlated Subspace Learning (CSL) [7].

Typically, features extracted through the aforementioned methods are real-valued. However, computing distances between extensive samples for retrieval can be time-consuming, rendering it impractical for large-scale datasets. In contrast, hashing methods which encode learned features into binary values, present a more feasible alternative based on lower storage costs and higher retrieval efficiency [14]. There are various techniques presented to produce individual hash codes for distinct modalities, including Local Sensitive Hashing (LSH), Spectral Hashing, and k-means-based hashing. By minimizing the Hamming distance between hash codes of heterogeneous samples, data-dependent hash codes prove effective for cross-modal similarity retrieval tasks like Cross-View Hashing (CVH) and Multi-Modal Latent Binary Embedding (MLBE) [7]. Prominent retrieval hashing frameworks leverage a similarity matrix that delineates the correlation among samples for optimal subspace training. These frameworks encompass Inter-media Hashing (IMH), Collective Matrix Factorization Hashing (CMFH), Latent Semantic Sparse Hashing (LSSH), and Fusion Similarity Hashing (FSH) [15]. Furthermore, supervised hashing methods that integrate concept information as supplementary data can yield more compact hash codes while retaining semantic significance [4]. Examples of such methods include Semantic Correlation Maximization (SCM) [16], Semantic-Preserving Hashing (SePH) [2], and Matrix Tri-Factorization Hashing (MTFH) [14].

The most formidable challenge for newly proposed cross-modal retrieval models lies in effectively learning semantically correlated hash codes while preserving fusion similarity across multi-modal data, regardless of the inherent heterogeneity between modalities [1]. Consequently, multi-modal data are uniformly encoded by human experts to ensure fair representative vectors [2]. However, this approach lacks scalability for future agnostic data, emphasizing the practicality of automatic feature extraction structures integrated with end-to-end frameworks [6]. In contrast, Deep Learning (DL) methods

excel at automatically extracting features from raw data. They have proven effective in enhancing the performance of retrieval models, especially in large-scale multi-modal datasets [6]. Some notable DL methods include Deep Joint-Semantics Reconstructing Hashing [17], Deep Visual Semantic Hashing, Deep Cross-Modal Hashing (DCMH) [6], Deep Semantic-Alignment Hashing (DSAH) [18], and Unsupervised Contrastive Cross-Modal Hashing [19]. Furthermore, by encoding large-scale image datasets with uniform hash codes, deep hashing methods can be effectively applied to supervised image-retrieval tasks. Examples of such methods encompass Deep Supervised Hashing (DSH) [20], Deep Supervised Discrete Hashing (DSDH) [21], Asymmetric Deep Supervised Hashing (ADSH) [22], SCAlable Deep Hashing (SCADH) [23], and Attribute-Aware Deep Hashing (A^2-NET) designed for large-scale fine-grained images [24].

B. Adversarial Robustness Research

In recent years, deep learning (DL) architectures were demonstrated to be vulnerable against specific attacks called adversarial perturbations. On visual data, researchers found that perturbations unrecognized by humans could misguide well-trained deep classifiers with high confidence [8]. However, the objective formula for adversarial perturbation generation is an NP-hard problem, since the decision function is always not concave, thus is impractical to acquire the exact solution. Practical techniques try to approximate the solution by simplifying the procedure with constraints such as L-BFGS and tangent misguidance toward decision boundary or using concave function substitution [8]. Additionally, L-BFGS involves unavoidable hyperparameters during the solution process, leading to excessive complexity in the generation procedure. A more efficient attack mode supposes the DL structure is completely linear, and the direction that can misguide the prediction of models with minimum magnitude can be found by simply obtaining the sign of the gradient calculated from the loss function on raw samples, called fast gradient sign method (FGSM) [25]. Although FGSM and its variants have been heuristically introduced based on the linearity assumption of DNN structures, the resulting perturbations have also been validated on complex nonlinear models [13]. Moreover, well-crafted perturbations can transfer across different models and datasets [12], highlighting the inherent vulnerability of DL architectures to small perturbations.

Additionally, a multitude of methods have been developed to counter adversarial perturbations. These methods either detect attacked samples or diminish their impact by obscuring the precise gradient information through random mechanisms. While these defense strategies can mitigate strong adversarial attacks, they often fail to enhance the inherent robustness of DL models. Conversely, adversarial training [13] is widely acknowledged as one of the most effective defense mechanisms [9]. Specifically, unlike directly minimizing the loss function, adversarial training incorporates an inner maximization procedure. This procedure iteratively queries the model to enhance its resilience against adversarial attacks. Recent studies have indicated that this min-max training approach can uncover features aligned with human vision, thereby improving the interpretability of DL models. Consequently, models trained with robust methods are deemed more reliable than those trained using conventional methods. The demonstrated robustness and interpretability of these robust training techniques in real-world applications further underscore their importance [26].

Adversarial perturbations were initially designed for classification tasks, where the attacked samples are deliberately misdirected away from the ground-truth concept. However, methods for generating perturbations and adversarial training tailored for retrieval tasks remain limited. Existing research in this area predominantly focuses on issues related to Hamming distance-based retrieval. Notably, the Hash Adversary Generation (HAG) method [27] has shown that guiding samples to modify their Hamming search performance can produce effective perturbations capable of disrupting well-trained deep cross-modal retrieval hashing models. Subsequently, the Adversarial Attack on Deep Cross-Modal Hamming Retrieval (AACH) [28] introduced an effective black-box attack strategy by maximizing the Hamming distance between semantically related samples. Recently, several adversarial perturbation methods have also been proposed to evaluate the robustness of deep retrieval models [12].

Existing defense methodologies against adversarial perturbations in retrieval models are still limited. One of the most relevant works is the Cross-Modal Correlation Learning (CMLA) [29]. While CMLA introduced a training method that incorporates adversarial examples through regularization, it lacks an end-to-end training procedure, limiting its scalability for large-scale scenarios. Additionally, CMLA relies on label information for effective adversarial example generation, making the training mechanism unsuitable for unsupervised retrieval tasks. Zhang et al. [30] also proposed an adversarial training approach for cross-modal retrieval. However, their method did not fully leverage the available supervisory information, leading to a lack of generalization when applied to large-scale cross-modal retrieval datasets like MS-COCO [31]. More recently, Zhou et al. introduced an anti-collapse triplet defense method to counter adversarial attacks on ranking-based retrieval models [32]. Nevertheless, to date, there remains a lack of end-to-end adversarial training methods directly applicable to large-scale retrieval tasks. Based on the aforementioned discussions, we propose Deep Supervised Adversarial Robust Hashing (DSARH) for robust cross-modal and image retrieval. We offer a theoretical analysis for generating effective worst-case perturbations and introduce an end-to-end adversarial training procedure. The effectiveness of DSARH is demonstrated through comprehensive experiments and comparisons.

III. DEEP SUPERVISED ADVERSARIAL ROBUST HASHING

Hashing-based similarity search methods aim to identify an optimal binary subspace that efficiently preserves the semantic information of high-dimensional or multi-modal data. Deep hashing methods leverage deep learning (DL) architectures to extract latent semantic features from such data. For example, these methods may employ deep convolutional neural

networks (CNNs) to capture semantic features from images. Deep CNNs have shown outstanding performance on high-dimensional and large-scale image datasets across various computer vision tasks. However, DL architectures have been criticized for potentially over-relying on highly predictive features rather than genuinely emphasizing semantic understanding [10]. This vulnerability of DL architectures exemplifies this issue. Adversarial training, which defends against worst-case perturbations during the training process, has been proven to be the most effective approach for enhancing the robustness of DL models, especially in tackling computer vision challenges [9].

In classification tasks, internal worst-case perturbations typically manifest as guiding model predictions away from the ground-truth labels of samples. These perturbations can be optimized by iteratively computing the gradients of models on the samples. Therefore, adversarial training for classification problems represents an end-to-end training technique. However, in retrieval tasks, semantic information is often provided in the form of a similarity matrix of samples. As a result, establishing a direct correlation between the learned hash code and the semantics of samples becomes challenging. Additionally, in cross-modal retrieval tasks, supervision information is typically presented in the form of intra-modality and inter-modality correlation matrices. The loss function of deep hashing models for retrieval is then formalized based on the distances between hashing codes and the samples' similarity matrix [33]. Hence, incorporating supervised information into an end-to-end perturbation generation procedure is the central challenge in designing effective adversarial training methods for retrieval tasks. Moreover, the supervised information comprises both intra-modality and inter-modality components. Effectively learning robust features from these two types of supervision information presents a significant challenge. In this section, we first introduce the basic notations of deep hashing and adversarial robustness. Next, we propose a novel end-to-end perturbation generation method. Finally, we outline two adversarial training schemes for cross-modal and unified hashing for retrieval.

A. Notations and Problem Formulation

Suppose that we have two datasets for retrieval: $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, ...\mathbf{U}_n\}, \mathbf{U}_i \in \mathbb{R}^{q_1}$, and $\mathcal{V} = \{\mathbf{V}_1, \mathbf{V}_2, ...\mathbf{V}_m\}, \mathbf{V}_i \in \mathbb{R}^{q_2}$, which are collected from the same ($q_1 = q_2$) or different modalities ($q_1 \neq q_2$), where q_1 and q_2 are the dimensions of two modalities. And the corresponding concept information for each modality are denoted as $\mathcal{Y}_\mathcal{U} = \{\mathbf{Y}_{\mathcal{U},1}, \mathbf{Y}_{\mathcal{U},2}, ...\mathbf{Y}_{\mathcal{U},n}\}$ and $\mathcal{Y}_\mathcal{V} = \{\mathbf{Y}_{\mathcal{V},1}, \mathbf{Y}_{\mathcal{V},2}, ...\mathbf{Y}_{\mathcal{V},m}\}$. $\mathbf{Y}_{\mathcal{U},i} = [y_{i,1}, y_{i,2}, ...y_{i,C}] \in \{0,1\}^c, y_{i,j} = 1$, if $\mathbf{U}_i$ belongs to class j. The concept information can be a single label category if each sample belongs to only one class, or multiple labels if each sample has at least one ground-truth label.

The principle of existing retrieval approaches asserts that the semantic similarity of different modalities can be effectively quantified with specific functions, e.g. Cosine and Euclidean distance [34]. In practice, while the dimensions of different modalities are significantly diverse and the dimensions of certain modalities are extremely high, raw data are always encoded as low-dimension features for efficient similarity quantification. We formalize the features learned from $\mathcal{U}$ and $\mathcal{V}$ as $\mathcal{F}_\mathcal{U} \in \mathbb{R}^{n \times k_\mathcal{U}}$ and $\mathcal{F}_\mathcal{V} \in \mathbb{R}^{m \times k_\mathcal{V}}$ respectively, $k_\mathcal{U}$ and $k_\mathcal{V}$ are the dimensions of learned features. $\mathcal{F}$ can be deep neural network architectures and the latent variables of DNN model's inner feature representation layers are found to contain rich semantic information [8]. Then hashing-based retrieval approaches further encode learned features as binary codes $\mathcal{B}_\mathcal{U} = \{\mathbf{B}_1^\mathcal{U}, \mathbf{B}_2^\mathcal{U}, ...\mathbf{B}_n^\mathcal{U}\} \in \{-1, +1\}^{n \times d}$ for more efficient retrieval computation and less storage space, where d is the length of hash codes from two modalities for fair semantic matching. Practically, the overall framework for hash code generation using DNNs can be formulated as:

$$\mathcal{B}_\mathcal{U} = sign\big(f_{hash}^u(f_{base}^u(\mathcal{U}, \theta_{base}^u), \theta_{hash}^u)\big), \tag{1}$$

where $f_{base}^\mathcal{U}$ and $f_{hash}^\mathcal{U}$ denote feature representation and hash code generation structures. $\theta_{base}^\mathcal{U}$ and $\theta_{hash}^\mathcal{U}$ are trainable parameters of two structures respectively. Those parameters are then optimized on training samples to produce more discriminative features from different modalities

The neighborhood distances between samples from different modalities are always formulated as their Hamming distance in hashing space: $dist_H(\mathcal{B}_\mathcal{U}, \mathcal{B}_\mathcal{V}) = \frac{1}{2}(K - \langle \mathcal{B}_\mathcal{U}, \mathcal{B}_\mathcal{V} \rangle)$, where K is a constant to maintain positive distance values. Optimal hash codes should exactly preserve different modalities' semantic meanings: $dist_H(\mathbf{B}_i^\mathcal{U}, \mathbf{B}_j^\mathcal{V}) \leq \forall_k dist_H(\mathbf{B}_i^\mathcal{U}, \mathbf{B}_k^\mathcal{V})$ on cross-modal datasets, where the i-th sample in $\mathcal{U}$ and j-th sample in $\mathcal{V}$ share same semantic labels, while the k-th sample in $\mathcal{V}$ have no relevant semantic concepts.

The semantic information of retrieval scenarios are always provided as the similarity matrix of samples: $\mathbf{S} \in [0,1]^{n \times m}$ that represents whether the i-th sample in $\mathcal{U}$ and j-th sample in $\mathcal{V}$ have shared labels. Given the pairwise multi-label similarity matrix, $\mathbf{S}_{i,j} = 1$ if the sample shares at least one label and is 0 otherwise. Based on the semantic information, the probability of a similarity matrix under produced hash codes can be formalized as:

$$p\left(\mathbf{S}_{i,j} \mid \mathcal{B}_\mathcal{U}, \mathcal{B}_\mathcal{V}\right) = \begin{cases} \delta\left(\frac{1}{2}\left\langle \mathbf{B}_i^\mathcal{U}, \mathbf{B}_j^\mathcal{V} \right\rangle\right), & \mathbf{S}_{i,j} = 1 \\ 1 - \delta\left(\frac{1}{2}\left\langle \mathbf{B}_i^\mathcal{U}, \mathbf{B}_j^\mathcal{V} \right\rangle\right), & \mathbf{S}_{i,j} = 0 \end{cases}, \tag{2}$$

where $\mathbf{B}_i^\mathcal{U}$ is the hash code of i-th sample in $\mathcal{U}$. Then hash codes can be optimized through minimizing the negative log-likelihood [15]:

$$\mathcal{L} = -\mathbb{E}_{i,j}\left(\frac{1}{2}\mathbf{S}_{i,j} \cdot \left(\mathbf{B}_i^\mathcal{U}\right)^\intercal \left(\mathbf{B}_j^\mathcal{V}\right) - log\left(1 + e^{\frac{1}{2}\left(\mathbf{B}_i^\mathcal{U}\right)^\intercal \left(\mathbf{B}_j^\mathcal{V}\right)}\right)\right), \tag{3}$$

B. Adversarial Perturbation Generation for Retrieval

Adversarial perturbations can be formalized as restricted inequality problem on computer vision tasks: $f(x) \neq f(x + \eta), s.t. \|\eta\|_p \leq \varepsilon$, where $f(\cdot)$ is the objective model, $\|\eta\|_p = \sqrt[p]{\frac{1}{d}\left(|\eta_1|^p + |\eta_2|^p + \cdots + |\eta_d|^p\right)}$ restricts the magnitude of the perturbation. And d is the length of the perturbation, which always has the same dimension with raw images. The norm of the perturbation is restricted to being smaller than a constant ε to maintain the semantic meaning of the image. The perturbations on text datasets are slight changes, such as

transforming alphabetic positions, that are difficult for humans to recognize. However, the textual data used in multimodal datasets were always encoded as numerical information [28]. Thus, we will not consider adversarial attacks on textual modality throughout the entire article.

Adversarial perturbations were initially designed for classification tasks based on samples' ground-truth labels and objective model's predictions [8]. Yet, for the retrieval task, the loss function cannot be explicitly obtained. Instead, an intuitive way is to iteratively query the retrieval set and update the perturbation until the prediction of the objective model is altered:

$$\exists_{l\neq j} dist_H\left(\mathbf{B}_i^{\mathcal{U}*}, \mathbf{B}_l^{\mathcal{V}}\right) \leq dist_H(\mathbf{B}_i^{\mathcal{U}*}, \mathbf{B}_j^{\mathcal{V}}), s.t. \|\eta\|_p \leq \varepsilon, \quad (4)$$

where $\mathbf{B}_i^{\mathcal{U}*}$ is the hash code of attacked sample $\mathbf{U}_i^* = \mathbf{U}_i + \eta_i$, η_i is the perturbation produced for $\mathbf{U}_i$. And j denotes the index of samples that have same concepts with $\mathcal{U}_i$ while l represents the index of irrelevant samples. Notably, the first part of equation 4 is an NP-hard problem that relying on extensive linear searching. Thus, we propose a practical substitute mechanism to approximately formalize this objective:

$$\max_{\eta_i} \mathbb{E}_{j,k}\big(dist_H\left(\mathbf{B}_i^{\mathcal{U}*}, \mathbf{B}_j^{\mathcal{V}}\right) - dist_H\left(\mathbf{B}_i^{\mathcal{U}*}, \mathbf{B}_k^{\mathcal{V}}\right)\big), \text{s. t. } \|\eta_i\|_p \leq \varepsilon, \quad (5)$$

where j and k denote indexes of samples that have the same or different concepts with $\mathcal{U}_i$. The above function is also an NP-hard problem, through explicitly specifying the bounding type of attacks as infinity norm $l_\infty^\varepsilon = \max_k |\eta_{i,j}| \leq \varepsilon$, we could obtain the further simplified solution, where $\eta_{i,j}$ denotes the j-th value of η_i. Notably, l_∞^ε bound is widely employed as the most commonly used adversarial attack restriction manner for computer vision modalities, it restricts the maximum value of the perturbation added on pixels to guarantee the attacked images could evade human recognition.

Practically, on image classification issues, l_∞^ε-bounded attacks can be approximated using the sign value of the gradient of the loss function on input data [25]. Considering a fully linear classifier $f(x) = w\top \cdot x + b$, the fast gradient sign method (FGSM) regards $\eta = \varepsilon \cdot sign(w) = \varepsilon \cdot sign(\nabla_x f(x)))$ as the worst-case perturbation with l_∞^ε restriction that can interference the prediction of classifier. Further, iteratively updating the perturbation using smaller step size FGSM could produce approximated worst-case perturbation on non-linear structures:$x^* = clip_\varepsilon\left(x^* + \xi \cdot sign\left(\nabla_x \mathcal{L}(x^*, y)\right)\right)$, where $clip_\varepsilon(\cdot)$ is an loop computation operation that updates x^* and restricts the magnitude of added perturbation $x^* - x$ on each dimension is smaller than the set variable, and ξ is the step size of each iteration.

Based on the above formulation, the $l_\infty^\mathcal{E}$ restricted perturbation on retrieval issues can be formulated as:

$$\eta_i = clip_\varepsilon\left\{\eta_i + \xi \cdot \nabla_{\mathbf{U}_i} \mathbb{E}_{j,k}(\langle \mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_j^{\mathcal{V}}\rangle - \langle \mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_k^{\mathcal{V}}\rangle)\right\}, \quad (6)$$

we use $\langle\cdot,\cdot\rangle$ to represent the Hamming distance operator $dist_H(\cdot,\cdot)$ for simplification. As shown in Equation 2, the probability under the binary codes is encoded with sigmoid function $\delta(x) = 1/\left(1 + e^{-x}\right)$ for model training. Thus, for a sample $\mathbf{U}_i$ and the corresponding hash code $\mathbf{B}_i^{\mathcal{U}}$, the worst-case direction of perturbation that can maximize objective model's loss function can be formalized as:

$$\frac{\partial(\langle \mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_j^{\mathcal{V}}\rangle - \langle \mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_k^{\mathcal{V}}\rangle)}{\partial \mathbf{B}_i^{\mathcal{U}}} = \frac{1}{2}\{\mathbf{S}_{i,j}\mathbf{B}_j^{\mathcal{V}} - \mathbf{S}_{i,k}\mathbf{B}_k^{\mathcal{V}} - e^{\frac{1}{2}\left(\mathbf{B}_i^{\mathcal{U}}\right)^\top \mathbf{B}_j^{\mathcal{V}}\mathbf{B}_j^{\mathcal{V}}} / \tau\left(\left(\mathbf{B}_i^{\mathcal{U}}\right)^\top \mathbf{B}_j^{\mathcal{V}}\right) + e^{\frac{1}{2}\left(\mathbf{B}_i^{\mathcal{U}}\right)^\top \mathbf{B}_k^{\mathcal{V}}\mathbf{B}_k^{\mathcal{V}}} / \tau\left(\left(\mathbf{B}_i^{\mathcal{U}}\right)^\top \mathbf{B}_k^{\mathcal{V}}\right)\}, \quad (7)$$

where $\tau(x) = 1 + e^{\frac{1}{2}x}$. Then, based on the category of samples, the above equation can be reorganized as:

$$\frac{\partial(\langle \mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_j^{\mathcal{V}}\rangle - \langle \mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_k^{\mathcal{V}}\rangle)}{\partial \mathbf{B}_i^{\mathcal{U}}} = \frac{1}{2}\left((\mathbf{S}_{i,j} - \frac{e^{\Delta_{i,j}}}{1+e^{\Delta_{i,j}}})\mathbf{B}_j^{\mathcal{V}} - (\mathbf{S}_{i,k} - \frac{e^{\Delta_{i,k}}}{1+e^{\Delta_{i,k}}})\mathbf{B}_k^{\mathcal{V}}, \Delta_{i,j} = \left(\mathbf{B}_i^{\mathcal{U}}\right)^\top \mathbf{B}_j^{\mathcal{V}}. \quad (8)$$

Then, it can be found that the direction towards worst-case direction is positively correlated with $\mathbf{B}_j^{\mathcal{V}}$ if $\mathbf{S}_{i,j} = 1$, or negatively correlated if $\mathbf{S}_{i,j} = 0$. In addition, the correlation relationship with $\mathbf{V}_k$ is precisely contrary with that of $\mathbf{V}_j$. Thus, we propose a novel item that is positively correlated with the worst-case direction: $sign\left(\mathbf{S}_{i,j} - 0.5\right) \cdot \mathbf{B}_\star^{\mathcal{V}}, \star = \{j,k\}$. Moreover, while the produced binary hash codes are sign values, $B_j^{\mathcal{V}}$ can be further replaced with $\mathbf{B}_i^{\mathcal{U}} \otimes \mathbf{B}_j^{\mathcal{V}} \cdot \mathbf{B}_j^{\mathcal{V}}$, where $\otimes$ denotes the direct product operator: $\mathbf{A} \otimes \mathbf{B} = \left(A_1 \cdot B_1, A_2 \cdot B_2, \ldots A_d \cdot B_d\right), \mathbf{A}, \mathbf{B} \in \mathbb{R}^d$. Then, the direction of perturbation is obtained as a specific direction toward the objective hash code itself:

$$\nabla_{\mathbf{B}_i^{\mathcal{U}}} = sign\left(\mathbf{S}_{i,j} - 0.5\right) \cdot \left(\mathbf{B}_i^{\mathcal{U}} \otimes \mathbf{B}_j^{\mathcal{V}}\right) \otimes \mathbf{B}_i^{\mathcal{U}}, \quad (9)$$

which is a composition of hash codes from another modality and the corresponding similarity matrix. Similarly, for the learned features, by normalizing each row of learned features $\mathcal{F}_\mathcal{U}$ and $\mathcal{F}_\mathcal{V}$ using the l_2 norm to $\widehat{\mathcal{F}}_\mathcal{U}$ and $\widehat{\mathcal{F}}_\mathcal{V}$, we can further use the cosine similarity metrics $\hat{\mathcal{F}}_\mathcal{U}\hat{\mathcal{F}}_\mathcal{V} \in \mathbb{R}^{n\times m}$ to measure the inner neighborhood structures of two modalities. Thus, the perturbation for feature similarity measurement can be formalized as:

$$\nabla_{\hat{\mathbf{F}}_i^{\mathcal{U}}} = sign\left(\mathbf{S}_{i,j} - 0.5\right) \cdot \left(\hat{\mathbf{F}}_i^{\mathcal{U}} \otimes \hat{\mathbf{F}}_j^{\mathcal{V}}\right) \cdot \hat{\mathbf{F}}_i^{\mathcal{U}}, \quad (10)$$

where $\hat{\mathbf{F}}_i^{\mathcal{U}}$ denotes the normalized feature of sample $\mathcal{U}_i$. The above perturbations are effectively delivered from the learned features using the given similarity matrix. For supervised retrieval tasks, samples' ground-truth labels $\mathcal{Y}_\mathcal{U}$ can also provide additional semantic information as hash codes $\mathcal{B}_\mathcal{Y}$. Then using the loss function of the hash codes between a sample and corresponding label information can also produce effective perturbation to impede objective model's performance.

Finally, based on the attack directions acquired from samples' hash codes, we can effectively approximate the perturbation on raw images using the projected gradient descent (PGD) approach:

$$\eta_\mathcal{U} = clip_\varepsilon\left(\eta_\mathcal{U} + \xi \cdot s\left(s(\mathbf{S} - 0.5) \cdot (\mathcal{B}_\mathcal{U} \otimes \mathcal{B}_\mathcal{V}) \otimes \frac{\partial \mathcal{B}_\mathcal{U}}{\partial \mathcal{U}}\right)\right), \quad (11)$$

$s(\cdot)$ represents the $sign(\cdot)$ function for simplification. The total number of iterations for each perturbation is a variable that based on the sample and objective model's corresponding prediction. Notably, under sufficiently many steps and small

step size, PDG-based attacks can be regarded as the worst-case perturbation even for non-linear DNN models [10]. Specifically, we present the detailed procedure of perturbation generation in Algorithm 1.

Algorithm 1 Adversarial Perturbation Generation Algorithm for Cross-modal Retrieval

Input Query dataset $\mathcal{U}_\psi = \{\mathbf{U}_{q1}, \mathbf{U}_{q2}, \dots \mathbf{U}_{qd}\}$ with d samples
Retrieval dataset from another modality $\mathcal{V}$
Objective model $\mathcal{G}(\cdot)$
Similarity matrix $\mathbf{S} \in [0,1]^{d \times m}$ of query set and retrieval set from another modality.
Output Perturbations of the query dataset.

Initialization: mini-batch size and total number of iterations.
for each iteration **do**
  repeat
    for each mini-batch **do**
      Get mini-batch similarity matrix $\mathbf{S}_\psi \in [0,1]^{bs \times m}$
      Get hash codes $\mathcal{B}_{\mathcal{U}_\psi}$ of query set under objective model $\mathcal{G}(\cdot)$;
      Get hash codes of retrieval set $\mathcal{B}_\mathcal{V}$
      Perform back-propagation and get FGSM step of query set through 9 and 10 using $\mathbf{S}_\psi$,$\mathcal{B}_{\mathcal{U}_\psi}$ and $\mathcal{B}_\mathcal{V}$;
      Obtain clipped perturbation $\eta_{\mathcal{U}_\psi}$ through 11.
    end for
  until (model prediction altered or reaching maximum iterations)
end for
Return $\eta_{\mathcal{U}_\psi}$

C. Adversarial Training for Cross-modal Retrieval

Building upon the perturbation generation approach described above, we then propose an adversarial training approach for retrieval issues. Adversarial training and its derivatives can offer certified robustness against specific perturbations with restricted magnitudes [9]. The principle of adversarial training is to perform an inner procedure searching for worst-case perturbations and update the the model under the produced attacks through an iterative min-max training manner: $\mathbb{E}_x min_\theta max_{x'} \mathcal{L}\left(x' \mid \theta\right)$, s.t. $x'_i \in B\left(x_i, \varepsilon\right)$, where $B\left(x_i, \varepsilon\right)$ is a magnitude restricted area around x_i, the magnitude is always quantified as p-norm bound $\left\|x' - x\right\|_p \leq \varepsilon$.

Firstly, to optimize retrieval models, the activation of hash code $\mathcal{B}(\cdot)$ is replaced with the differential hyperbolic tangent function $tanh(\cdot)$: $\mathcal{H}_\mathcal{U} = tanh\left(w^\top \cdot \mathcal{F}_\mathcal{U}\right), \mathcal{F}_\mathcal{U} \in \mathbb{R}^{n \times k_\mathcal{U}}, w \in \mathbb{R}^{k_\mathcal{U} \times c}$, c is the length of obtained hash codes. Then the semantic distance between samples can be quantified as the Hamming distance between learned hash codes: $\mathcal{L}_\mathcal{H} = K - \langle \mathcal{H}_\mathcal{U}, \mathcal{H}_\mathcal{V} \rangle$. Typically, differential hash codes can be optimized through minimizing their similarity with signed values to improve the accuracy of obtained hash codes:

$$min\mathcal{L}_\mathcal{H} = \mathbb{E}_i \left\|\mathbf{H}_i^\mathcal{U} - \mathbf{B}_i^\mathcal{U}\right\|_2^2 + \alpha \cdot \mathbb{E}_j \left\|\mathbf{H}_j^\mathcal{V} - \mathbf{B}_j^\mathcal{V}\right\|_2^2, \quad (12)$$

where α is a variable that depends on the distribution of cross-modal datasets [34].

Then, to improve the semantic preserving performance of hash code, based on the semantic matrix of training samples, the hash code generation structures can be optimized through optimizing the trainable parameters $\theta_\mathcal{U}$ of the hash code generation structure of $\mathcal{H}(\cdot)$:

$$min_{\theta_\mathcal{U}} \mathbb{E}_{i,j} \left(\mathbf{S}_{i,j} \cdot \left(\mathbf{H}_i^\mathcal{U}\right)^\top \cdot \mathbf{B}_i^\mathcal{V} - \log\left(1 + e^{\frac{1}{2}\left(\mathbf{H}_i^\mathcal{U}\right)^\top \cdot \mathbf{B}_j^\mathcal{V}}\right)\right). \quad (13)$$

We use the sign value of hash codes in $(B_U \otimes B_V)$ to fix the direction of the gradient as constant for back-propagation.

The exact gradient of the hash code generation structure $\mathcal{H}_\mathcal{U}$ on $\mathcal{U}$ can be effectively obtained using the chain rule through deep structures. To perform adversarial training on retrieval issues, we should firstly acquire the worst-case perturbations for inner maximization process. As demonstrated in Equation 10, the maximum item can be effectively solved by iteratively querying the model and updating the worst-case perturbation:

$$\mathcal{U}_i^* = clip_\varepsilon \left\{\mathcal{U}_i^* + \xi \cdot s(s(\mathbf{S}_{i,j} - 0.5) \cdot \mathbf{B}_i^\mathcal{U} \otimes \mathbf{B}_j^\mathcal{V} \cdot \frac{\partial \mathbf{B}_i^\mathcal{U}}{\partial \mathcal{U}_i}\right\}, \quad (14)$$

where $\mathcal{U}_i^*$ denote the produced adversarial example by adding perturbations on samples.

In addition, we also perform supervised adversarial training to learn robust intra-modal discrimination features in data. Similar with the cross-modal hashing process, the negative log-likelihood of the similarity matrix under the produced hash codes $\mathcal{B}_\mathcal{U}$ can be formalized as:

$$\mathcal{L} = -\mathbb{E}_i \left(\frac{1}{2}\mathbf{S}_{i,j} \cdot \left(\mathbf{B}_i^\mathcal{U}\right)^\mathsf{T} \left(\mathbf{B}_j^\mathcal{U}\right) - log\left(1 + e^{\frac{1}{2}\left(\mathbf{B}_i^\mathcal{U}\right)^\mathsf{T}\left(\mathbf{B}_j^\mathcal{U}\right)}\right)\right). \quad (15)$$

We could also produce effective perturbations using the semantic information from other samples. Specifically, the perturbations are added on samples to change their semantics from other samples of the same modality:

$$\mathcal{U}_i^* = clip_\varepsilon \left\{\mathcal{U}_i^* + \xi \cdot s(s(\mathbf{S}_{i,j} - 0.5) \cdot \mathbf{B}_i^\mathcal{U} \otimes \mathbf{B}_j^\mathcal{U} \cdot \frac{\partial \mathbf{H}_i^\mathcal{U}}{\partial \mathcal{U}_i}\right\}. \quad (16)$$

Towards now, we have produced two types of perturbations to attack the performance of cross-modal retrieval models. The former is added on $\mathcal{U}$ to interfere the cross-modal similarity of learned hash codes, and performing adversarial training against such attacks promotes the model to learn robust features to establish reliable correlation between different modalities of samples beyond datasets' heterogeneity. The latter is to interfere the intra-modal similarity, we execute adversarial training against those attacks to enhance the robustness of the learned semantic representation features by the intro-modal hash code generation structure. Furthermore, two types of perturbations can also be integrated as:

$$\begin{gathered}\mathcal{U}_i^* = clip_\varepsilon \left\{\mathcal{U}_i^* + \xi \cdot s(s(\mathbf{S}_{i,j} - 0.5) \cdot \Delta_\mathcal{B} \cdot \frac{\partial \mathbf{H}_i^\mathcal{U}}{\partial \mathcal{U}_i}\right\}, \\ \Delta_B = \left(\mathbf{B}_i^\mathcal{U} \otimes \mathbf{B}_j^\mathcal{U}\right) \otimes \left(\mathbf{B}_i^\mathcal{U} \otimes \mathbf{B}_j^\mathcal{V}\right).\end{gathered} \quad (17)$$

The ultimate objective function is composed of the intra-modal and inter-modal negative log-likelihood loss functions:

$$\mathcal{L} = \mathcal{L}_\mathcal{H}^* \left(\mathcal{U}, \mathcal{V}\right) + \alpha \cdot \mathcal{L}_\mathcal{H}^* \left(\mathcal{U}, \mathcal{U}\right) + \beta \cdot \mathcal{L}_\mathcal{H} \left(\mathcal{V}, \mathcal{V}\right). \quad (18)$$

where $\mathcal{L}_\mathcal{H}^*$ is composed of both the standard and robust parts:$\mathcal{L}_\mathcal{H}^*(\mathcal{U}, \mathcal{V}) = \mathcal{L}_\mathcal{H}(\mathcal{U}, \mathcal{V}) + \lambda \cdot \mathcal{L}_\mathcal{H}(\mathcal{U}^*, \mathcal{V})$ The process

of robust cross-modal retrieval is shown in Algorithm 2 and Fig.3.

Finally, to better understand the principle of adversarial training on retrieval issues, we rewrite the objective function on a simplified hash net with only one neural network layer. Considering the inner product as the distance quantification function of learned hash codes from two modalities: $\mathcal{L} = K - (\mathcal{H}^{\mathcal{U}})^{\top} \cdot \mathcal{H}^{\mathcal{V}}$, the FGSM step on $\mathcal{H}_1$ can be written as

$$\eta = \xi \cdot \left(I - \tanh^2\left(w_1^{\top} \cdot \mathcal{U}\right)\right) \cdot s(w_1) \cdot \tanh\left(w_2^{\top} \cdot \mathcal{V}\right), \tag{19}$$

then, by adding the produced perturbations using FGSM on $\mathcal{U}$, the prediction of the model updates as:

$$\begin{aligned} \mathcal{H}^{\mathcal{U}*} =& tanh\left(w_1^{\top} \cdot \left(\mathcal{U} + \xi \cdot \mathcal{A} \cdot sign\left(w_1\right) \cdot tanh\left(w_2^{\top} \cdot \mathcal{V}\right)\right)\right) \\ =& tanh\left(w_1^{\top} \cdot \mathcal{U} + \xi \cdot \mathcal{A} \cdot |w_1| \cdot tanh\left(w_2^{\top} \cdot \mathcal{V}\right)\right). \end{aligned} \tag{20}$$

Notably, since the item $\mathcal{A} = I - \tanh^2\left(w_1^{\top} \cdot \mathcal{U}\right)$ is a constant larger than 0 that measures the learned and the binary hash codes. Then, the loss function of the cross-modal retrieval modal is updated as the distance between $w_1^{\top} \cdot \mathcal{U} + \xi \cdot \mathcal{A} \cdot |w_1| \cdot tanh\left(w_2^{\top} \cdot \mathcal{V}\right)$ and $w_2^{\top} \cdot \mathcal{V}$. Obviously, to minimize the above two features, w_1 is trained not only to minimize the difference between the learned hash codes from two modalities, but also small perturbations offered by its L1 norm value. Thus, the principle of adversarial training is similar with the L1 norm regularization item added on training process against overfitting. Differently, for adversarial training, the penalization term $\xi \cdot \mathcal{A} \cdot |w_1| \cdot tanh\left(w_2^{\top} \cdot \mathcal{V}\right)$ also relies on the value of w_2. This establishes more intimate connection between two modalities, so as to reduce the heterogeneity problem between modalities with significant differences in dimensions.

Algorithm 2 Robust Cross-modal Retrieval Algorithm

Input The training set of the cross-modal dataset $\mathcal{U}_{\Gamma} \in \mathbb{R}^{t \times q_1}$ and $\mathcal{V}_{\Gamma} \in \mathbb{R}^{t \times q_2}$ with t samples
Objective model $\mathcal{G}(\cdot)$
Similarity matrix $\mathbf{S} \in \{0,1\}^{t \times t}$ between two modalities

Initialization: Network learning parameters (e.g., learning rate, momentum, optimizer).
Mini-batch size
Iteration number
for each iteration **do**
  repeat
    update learning rate
    for each mini batch **do**
      Get mini batch similarity matrix;
      Generate hash codes of two modalities;
      Produce perturbations of image samples based on Equation 17.
      Update the whole model based on Equation 18.
    end for
  until (Objective function converged or reaching maximum iterations)
end for

D. Attack Production and Training Scheme for Unified Hashing

Robust training can also be employed on unified hashing issues. Considering a single modality retrieval task, where the overall loss function can be formalized on training samples:

$$\begin{aligned} \mathcal{L} = \mathbb{E}_{i,j}\left(\mathcal{L}_{\mathcal{U}}\left(\mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_j^{\mathcal{U}}, \mathbf{Y}_{i,j}\right)\right), \mathcal{L}_{\mathcal{U}} = (1 - \mathbf{Y}_{i,j}) \cdot \\ dist_H\left(\mathbf{B}_i^{\mathcal{U}}, \mathbf{B}_j^{\mathcal{U}}\right) + \mathbf{Y}_{i,j} \cdot \max(K - dist_H\left(B_i^{\mathcal{U}}, B_j^{\mathcal{U}}\right), 0), \end{aligned} \tag{21}$$

where $\mathbf{Y}_{i,j} = 0$ if sample $\mathcal{U}_i$ and $\mathcal{U}_j$ share same concepts, and 1 otherwise. While the training process executed on all samples requires overly large computational and storage cost. Practically, unified hashing models are always optimized under randomly sampled training set [22]:

$$min_{\theta_{\mathcal{U}}} \mathbb{E}_{\mathbf{U}_i \in \mathcal{U}_{\Gamma}} \sum_{j=1}^{n-m} \left(\mathbf{B}_i^{\mathcal{U}}\left(\mathbf{B}_j^{\bar{\mathcal{U}}}\right)^{\top} - c \cdot \mathbf{S}_{i,j}\right)^2, \tag{22}$$

where $\mathcal{U}_{\Gamma} = \{\mathbf{U}_{\Gamma 1}, \mathbf{U}_{\Gamma 2}, \ldots, \mathbf{U}_{\Gamma m}\}$ and $\bar{\mathcal{U}}$ denote the training set and retrieval set that contain m and $n-m$ samples respectively. $\mathbf{S} \in \{-1,1\}^{m \times (n-m)}$ represents similarity matrix of the training and retrieval set, $\mathbf{S}_{i,j} = 1$ means two samples $\mathbf{U}_i$ and $\mathbf{U}_j$ share same labels and $\mathbf{S}_{i,j} = -1$ otherwise. Notably, the hash codes of training samples and retrieval set need to be synchronously optimized through the entire training process. Thus, the training procedure can be divided into two steps.

Firstly, the training samples combined with retrieval samples and corresponding similarity matrix are employed to optimize the hash code generation structure:

$$\mathcal{L}_{\theta} = min_{\theta_{\mathcal{U}}} \mathbb{E}_{i,j}\left(\left(\mathbf{H}_i^{\mathcal{U}} \mid \theta_{\mathcal{U}}\right)\left(\mathbf{B}_j^{\bar{\mathcal{U}}}\right)^{\top} - c \cdot \mathbf{S}_{i,j}\right)^2. \tag{23}$$

The inner maximum procedure is performed by adding specific perturbations on training samples to attack the semantic similarity between training samples and retrieval samples. Notably, the hash codes on the database must be updated along with those of the hash structure, and they should be updated asynchronously. By fixing $\mathcal{B}^{\bar{\mathcal{U}}}$, the objective function of perturbation for $\mathbf{U}_i$ can be formalized as:

$$\max_{\eta_i} \sum_{j=1}^{n-m} \left(\mathbf{B}_i^{\mathcal{U}_{\Gamma,*}}\left(\mathbf{B}_j^{\bar{\mathcal{U}}}\right)^{\top} - c \cdot \mathbf{S}_{i,j}\right)^2, \tag{24}$$

which can also be approximated by the back-propagated gradient on the learned hash code:

$$\nabla_{\mathbf{B}_i^{\mathcal{U}_{\Gamma}}} = \left(\mathbf{B}_i^{\mathcal{U}_{\Gamma}} \mathbf{B}_j^{\bar{\mathcal{U}}} - c \cdot \mathbf{S}_{i,j}\right) \mathbf{B}_j^{\bar{\mathcal{U}}}. \tag{25}$$

Obviously, the gradient of the loss function depends on both the semantic similarity and the hash codes of the retrieval samples. $\left(\mathbf{B}_i^{\mathcal{U}_{\Gamma}} \mathbf{B}_j^{\bar{\mathcal{U}}} - c \cdot \mathbf{S}_{i,j}\right)$ is greater than 0 if the training and retrieval samples have dissimilar concepts and the hash codes are not optimal, and otherwise it is less than 0. Therefore, when the two samples are not relevant, the worst-case direction is in direct proportion to $\mathbf{B}_j^{\bar{\mathcal{U}}}$. Thus the perturbation direction can be modified as $\mathbf{S}_{i,j} \cdot \mathbf{B}_j^{\bar{\mathcal{U}}}$. Similar with Equation 14, we can rewrite the objective function of perturbation as:

$$\eta_i = clip_{\varepsilon}\left(-\xi \cdot \mathbf{S}_{i,j} \cdot \left(\mathbf{B}_i^{\mathcal{U}_{\Gamma}} \otimes \mathbf{B}_j^{\bar{\mathcal{U}}}\right) \otimes \frac{\partial \mathbf{H}_i^{\mathcal{U}_{\Gamma}}}{\partial \mathbf{U}_i}\right). \tag{26}$$

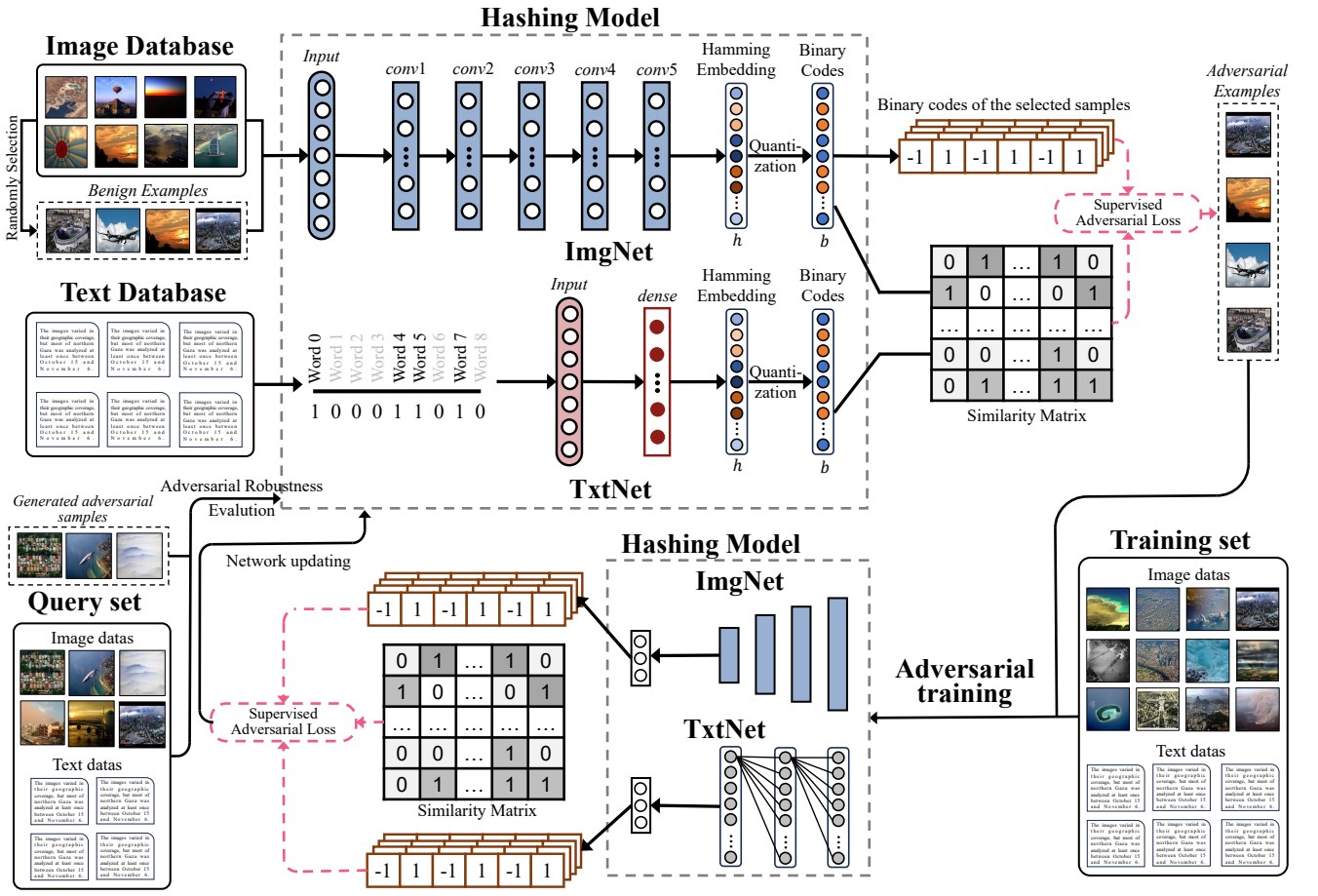


Fig. 3: The principle of the DSARH (Deep Semantic Adversarial Robust Hashing) framework. DSARH validates the robustness of the model by training it to resist and effectively handle these adversarial attacks, thereby enhancing its overall performance and reliability in cross-modal retrieval tasks.

Moreover, a regularization item between learned hash codes and binary codes is also introduced to enhance the accuracy of learned hash codes:

$$\mathcal{L}_\theta = \mathbb{E}_{i,j}\left\{\left(\mathbf{H}_i^{\mathcal{U}}\left(\mathbf{B}_j^{\widetilde{\mathcal{U}}}\right)^\top - c\cdot\mathbf{S}_{i,j}\right)^2 + \gamma\cdot\left(\mathbf{H}_i^{\mathcal{U}} - \mathbf{B}_i^{\mathcal{U}}\right)^2\right\}. \tag{27}$$

Then, by simplifying the learned features $\mathbf{F}_i^{\mathcal{U}*}$ and trainable hash codes $\mathbf{H}_i^{\mathcal{U}*}$ of the attacked sample as $\mathbf{z}_i$ and $\mathbf{u}_i$. The training process of hash code generation structure can be performed by back-propagating the loss function on the hash features:

$$\frac{\partial\mathcal{L}_\theta}{\partial\mathbf{z}_i} = 2\mathbb{E}_j\left\{\left(\mathbf{u}_i\left(\mathbf{B}_j^{\widetilde{\mathcal{U}}}\right)^\top - c\cdot\mathbf{S}_{i,j}\right)\mathbf{B}_j^{\widetilde{\mathcal{U}}} + 2\gamma\cdot\left(\mathbf{u}_i - \mathbf{B}_i^{\mathcal{U}}\right)\right\} \odot\left(1-\mathbf{u}_i^2\right). \tag{28}$$

Secondly, the objective function for the retrieval set's hash code updating can be formalized as:

$$\mathcal{L}_H = \mathbb{E}_{i,j}\left\{\left(\mathbf{B}_i^{\mathcal{U}}\left(\mathbf{H}_j^{\widetilde{\mathcal{U}}}\right)^\top - c\cdot\mathbf{S}_{i,j}\right)^2 + \gamma\cdot\left(\mathbf{B}_i^{\mathcal{U}} - \mathbf{H}_i^{\widetilde{\mathcal{U}}_\Gamma}\right)^2\right\}. \tag{29}$$

where $\mathbf{H}_j^{\widetilde{\mathcal{U}}_\Gamma}$ denotes the hash codes of corresponding samples of the training samples in the retrieval set. Notably, while the size of retrieval set is overly large, the optimizing process of the hash codes of retrieval set can be simplified through updating the hash codes on a column-by-column column basis [35]. Specifically, we use $\widetilde{u}$ to denote the trainable hash code of the retrieval set, then we expand the above formula as:

$$\mathcal{L}_H = \left(u\widetilde{u}^\top\right)^2 - 2c\cdot\mathrm{tr}\left(u^\top\mathbf{S}\widetilde{u}\right) - 2\gamma\cdot\mathrm{tr}\left(u\widetilde{u}_\Gamma^\top\right) + const, \tag{30}$$

$\widetilde{u}_\Gamma$ is the simplified representation of $\mathcal{H}^{\widetilde{\mathcal{U}}_\Gamma}$.

Further, by constructing a new hash matrix to pad the training set matrix, ensuring that it has the same dimensions as the retrieval code, i.e., constructing a matrix with dimensions identical to the retrieval code and zero-padding the parts that do not belong to the training code: $\bar{u} = [\mathbf{0}, \mathbf{0}, \ldots u_1, \mathbf{0}, \ldots u_m, \ldots \mathbf{0}]^\top \in \mathbb{R}^{n\times c}$ can be created. Then, the above equation can be written as

$$\begin{aligned}\mathcal{L}_H &= \left(u\widetilde{u}^\top\right)^2 - 2\,\mathrm{tr}\left(\widetilde{u}\left(c\cdot u^\top\mathbf{S} + \gamma\cdot\bar{u}^\top\right)\right) + const \\ &= \left(u\widetilde{u}^\top\right)^2 + \mathrm{tr}\left(\widetilde{u}p\right) + const,\end{aligned} \tag{31}$$

where $p = -2c\cdot u^\top\mathbf{S} - 2\lambda\cdot\bar{u}^\top$, and $const$ is a constant that depends on the elements in $\widetilde{u}$ that do not correspond to the samples in the training set, since the hash codes of these samples remain constant during the second step update phase thus can be regarded as constants . Then we could optimize

the above loss function using the discrete cyclic coordinate descent algorithm [36]. Specifically, $\widetilde{u}$ is optimized column by column, i.e., we update each bit of the hash code while fixing others. We let u_{*i} and $\hat{u}_{*i}$ to denote the i-th column of u and the matrix of u that exclude the i-the column. Then, the value contributed by $\widetilde{u}_{*i}$ on the loss function $\mathcal{L}_H(\widetilde{u}_{*i})$ can be calculated as:

$$\mathcal{L}_H(\widetilde{u}_{*i}) = \mathrm{tr}\left(\widetilde{u}_{*i}\left(2u_{*i}^{\top}\hat{u}_{*i}\hat{\widetilde{u}}_{*i}^{\top} + p_{*i}\right)\right) + const. \quad (32)$$

Considering the final hash codes are sign values, the solution to minimize $\mathcal{L}_H(\widetilde{u}_{*i})$ can be acquired as:

$$\widetilde{u}_{*i} = -\,\mathrm{sign}\left(2\hat{\widetilde{u}}_{*i}\hat{u}_{*i}^{\top}\widetilde{u}_{*i} + p_{*i}^{\top}\right). \quad (33)$$

Finally, we present the whole procedure of robust single modal retrieval in Algorithm 3.

Algorithm 3 Robust Single Modal Retrieval Algorithm

Input Database dataset $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, \ldots \mathbf{U}_n\}$
Hash code generation model $\mathcal{G}(\cdot)$
Initialization Mini-batch size
Iteration number.
for each iteration **do**
  repeat
    Randomly sample data from the database dataset to construct the training set $\mathcal{U}_\Gamma$;
    Obtain the similarity matrix between the training and retrieval set $\mathbf{S} \in \{-1, 1\}^{m\times(n-m)}$.
    for each batch size **do**
      Get mini-batch similarity matrix;
      Generate retrieval set's hash codes $\widetilde{u}$ by fixing $\mathcal{G}(\cdot)$'s parameters;
      Produce attacks for the training set based on Equation 26;
      Update $\mathcal{G}(\cdot)$ based on back-propagated gradient using Equation 28;
      Iteratively updating the retrieval set's hash codes $\widetilde{u}$ using Equation 33 by fixing model's parameters.
    end for
  until (Objective function converged or reaching maximum iterations)
end for

IV. EXPERIMENTS

In this section, we conduct a comprehensive performance evaluation of DSARH and compare it with state-of-the-art deep hashing methods across various cross-modal and image retrieval benchmarks.

A. Datasets and Evaluation Protocol

We assess the performance of cross-modal retrieval models using WikiPedia, FLICKR-25K, NUSWIDE, and MS-COCO. Additionally, we evaluate the performance of image retrieval models using large-scale NUSWIDE and MS-COCO datasets. The basic description of each dataset is presented as follows:

Wikipedia [34] consists of 2866 image-text pairs compiled from Wikipedia articles representing the 10 most populated categories. The text data were initially processed using a pre-trained Latent Dirichlet Allocation (LDA) model, resulting in a probability distribution with 10 dimensions. The dataset is randomly divided into two parts for training and testing. The training set comprises 2173 documents, while the remaining 693 documents are allocated to the test set.

FLICKR-25K [37] consists of a total of 25,000 image-text pairs collected from the Flickr website. Each pair is annotated with one or more unique concepts. The text data are encoded as 1386-dimensional bag-of-word (BoW) representations. Following the methodology outlined in [6], categories with extremely low occurrence frequency were removed, resulting in 24 remaining concepts. For experimental purposes, we randomly designate 1000 pairs for the query set, 5000 for training, and allocate the remaining pairs for retrieval.

NUSWIDE [38] is a multi-modal dataset collected from the Flickr website. The original NUSWIDE dataset comprises 269,648 images categorized into 81 concepts, with text data encoded as 1000-dimensional bag-of-words (BoW) vectors. While NUSWIDE serves as a benchmark for both cross-modal and single image retrieval tasks, it is notably large and imbalanced. Following the approach outlined in [22], we reconstructed the dataset by focusing on the top 21 concepts. Subsequently, as suggested by [39], we randomly selected 100 and 500 image-text pairs per class to form the query and training sets for the cross-modal retrieval task, respectively, with the remaining pairs designated for retrieval. Moreover, adhering to the methodology proposed by [40], we selected 5000 samples to constitute the query set for image retrieval, with the remaining samples designated as the gallery.

MS-COCO [31] comprises a total of 123,287 English samples, with each image accompanied by five textual descriptions. Following the splitting set of [41], we randomly designate 5000 images for the query set and another 5000 images for the validation set, leaving the remaining 113,287 samples for training. Additionally, MS-COCO can serve as an image-retrieval evaluation benchmark [42].

Three evaluation protocols are employed in this article to thoroughly assess the retrieval performance of benchmark models. Mean Average Precision (mAP) is the most commonly-used metric for retrieval accuracy evaluation. For a query sample, mAP calculates the mean retrieval performance on given query set and retrieval set: $\frac{1}{M}\sum_{k=1}^{M} p(k)\delta(k)$, where M is the count of alternative retrieval samples that share relevant labels with the query sample, $p(k)$ is the precision of the top k retrieved samples, and $\delta(\mathrm{k})$ denotes whether the k-th sample is semantically relevant with the query sample (1 if relevant and 0 otherwise). Obviously, the greater mAP implies better retrieval performance on single-label datasets where samples have only one ground-truth semantic label. However, for multi-label datasets, $\delta(\mathrm{k})$ could not accurately quantify multi-category retrieval performance. Instead, we also introduce Normalized Discounted Cumulative Gain (nDCG) and Average Cumulative Gain (ACG) for multi-label datasets. The nDCG is obtained as: $nDCG@p = \frac{1}{Z}\sum_{i=1}^{p}\frac{2^{r_i}-1}{\log(1+i)}$, where r_i is the number of shared labels between the query sample and the i-th retrieval sample, p denotes the number of top p retrieval samples. Z is a constant to guarantee

the normalization attribute of nDCG, which is always set as the nDCG value under the exact retrieval circumstance. Notably, nDCG penalizes the lower sort of the retrieved samples, differently, ACG directly calculates the mean of the total concepts shared between query and retrieved samples: $ACG@p = \frac{1}{p}\sum_{i=1}^{p} r_i$, $r_i = k$ if the retrieval sample and query sample share k identical concepts. Following [22] and [40], we set p of both nDCG and ACG as 100, and M in mAP is set as the total number of the query samples.

B. Benchmark Configuration

In recent years, numerous deep and shallow retrieval models have been proposed. Shallow models, designed for handcrafted features, have traditionally served as baselines for comparing the performance of state-of-the-art retrieval methods. While these shallow models are designed based on handcrafted features, it is impractical to devise corresponding attacks against them. Therefore, they act as benchmark baselines for assessing the standard performance of retrieval tasks. We implemented the most renowned shallow cross-modal retrieval models using the code generously provided by the authors. Specifically, we evaluate the cross-modal retrieval performance of SePH [2], CVH [3], LSSH [43], SCM [16], STMH [44], SRLCH [45], DLFH [46] and MTFH [47] on the lightweight cross-modal retrieval datasets. The lightweight datasets utilize BoW or SIFT features to encode raw image and text content. Additionally, We refer to the experimental results from [48] to establish a comparison benchmark, which include several representative deep hashing models including UKD [49], DBRC [50] and UGACH [51]. Furthermore, we retrained several deep cross-modal hashing models, including DJSRH [17], AGAH [52], SSAH [53], DCMH [6], DSAH [18], DADH [54], DGCPN [55], and UCCH [19], following the methodologies outlined in their respective papers. The initial codes were generously provided by the authors of these papers. Subsequently, we assess the robustness of these models using the distance-based attack proposed in Section III.

We set several commonly-used deep and shallow image retrieval models including ITQ [56], LFT [57], FashH [58], ADGH [59], COSDISH [60], SDH [36], CSQ [61], DSH [20], ADSH [22], DSDH [21], GTelecomNet-CSQ [62], SCADH [63], SGDH [64] and A^2-NET [24] as the benchmark for evaluating the performance on image retrieval tasks, we obtain most of the experimental results from [22] that share same experimental settings. Moreover, we re-train nine state-of-the-art deep single image retrieval models including SDH, CSQ, DSH, ADSH, DSDH, GTelecomNet-CSQ, SCADH, SGDH, and A^2-NET for robustness evaluations. Moreover, we also provide the Normalized Discounted Cumulative Gain (nDCG) and Average Cumulative Gain (ACG) performance of each method to evaluate their multi-label retrieval performance. Notably, we provide the detailed network configuration for model setup in the Supplementary Material.

C. Comparisons with Deep Cross-modal Hashing Methods

In this section, we evaluate the regular cross-modal retrieval performance of DSARH and other deep hashing baselines. For a fair comparison, we vary the hash code length from 16 to 128 (i.e., 16, 32, 64, 128) and record the mAP, nDCG, and ACG scores on four benchmark datasets. TABLE I presents the quantitative comparison with state-of-the-art deep hashing methods on MIRFLICKR25k, NUS-WIDE and MS-COCO. Additionally, the $I \rightarrow T$ (using image to retrieve text) and $T \rightarrow I$ (using text to retrieve image) mAP performance and precision-recall curves of deep hashing benchmark baselines are displayed in TABLE I and Fig. 4. Moreover, we also provide the nDCG@100 and ACG@100 performance comparisons in Fig. 5.

We observe that DSARH has delivered competitive mAP performance on large-scale datasets, outperforming most of the state-of-the-art deep hashing baselines. Only AGAH [52] and UCCH [19] yielded more competitive mAP results in the $I \rightarrow T$ task of NUS-WIDE. Specifically, AGAH [52] adopts an adversarial learning guided multi-label attention module to enhance feature learning. This module enables the model to learn discriminative feature representations while maintaining cross-modal invariance. UCCH [19] utilizes contrastive learning to explore performance similarity between image-text pairs rather than labels. It employs an effective contrastive loss function, to maximize instance-level differences and minimize cross-modal differences. However, our DSARH approach is more competitive on large-scale retrieval datasets, and DSARH has achieved the best multi-label retrieval performance on all benchmarks. Thus, DSARH is more proficient at discovering semantic features from high-dimensional images to establish a more precise correlation between heterogeneous data.

The superiority of our approach lies in the effectiveness of adversarial training for robust feature extraction from high-dimensional data. The significant difference in dimensions between images and textual information poses a heterogeneity problem that hinders the establishment of exact semantic correlations in cross-modal retrieval tasks. While modality-specific hash functions using DL architectures could enhance the feature representation capacity compared to shallow unified hash codes, DL architectures often overly rely on predictive features rather than semantic representations [10]. Training against perturbations can encourage DL-based models to discover more discriminative and robust features from high-dimensional images, thereby mitigating heterogeneous issues. Furthermore, the robust feature extraction capacity of adversarial training can improve the transferability of DL architectures [9], enhancing the semantic representation of learned hash codes on out-of-the-sample sets. The exceptional performance on multi-label evaluation sets further validates the value of robustness for reliable cross-modal retrieval.

D. Results of the Robust Cross-modal Retrieval Performance

We assess the robustness of deep hashing benchmark baselines against adversarial attacks using perturbations generated for image samples. Following commonly-used robustness evaluation metrics [28], we employ the proposed infinity norm bounded and non-targeted white-box attacks, as described in Section III, on image samples from the evaluation set. Subsequently, we assess the retrieval performance of benchmark baselines to determine their robustness. Specifically, we retrain

TABLE I: Regular and Robust Comparison with the State-of-the-Art Deep Hashing Methods

Task	Mothods	MIRFLICKR25k				NUS-WIDE				MS-COCO			
		16	32	64	128	16	32	64	128	16	32	64	128
$I \rightarrow T$ (Regular)	DBRC [50]	0.5921	0.5924	0.3935	0.4046	0.4114	0.4023	0.5857	0.5918	0.5716	0.5882	0.6137	0.6316
	UGACH [51]	0.6767	0.6935	0.5974	0.6153	0.6272	0.6382	0.7025	0.7066	0.3212	0.3517	0.4771	0.4672
	DJSRH [17]	0.6321	0.6562	0.4551	0.4883	0.4994	0.5249	0.6697	0.6725	0.6754	0.7197	0.7244	0.7246
	AGAH [52]	**0.7381**	**0.7621**	0.6182	0.6345	0.6466	0.6425	0.7114	0.7258	0.5712	0.5913	0.5855	0.5771
	SSAH [53]	0.6462	0.6793	0.4728	0.4949	0.5215	0.5266	0.6914	0.7012	0.7188	0.7226	0.7314	0.6887
	DCMH [6]	0.7152	0.7232	0.6382	0.6518	0.6577	0.6599	0.736	0.7371	0.6092	0.6174	0.6229	0.6253
	DSAH [65]	0.6851	0.6932	0.5563	0.5842	0.6034	0.6116	0.7034	0.7093	0.7152	0.7221	0.7316	0.7324
	UKD [66]	0.7111	0.7132	0.6153	0.6331	0.6352	0.6424	0.7237	0.7196	0.5023	0.5792	0.5564	0.6146
	DGCPN [55]	0.7282	0.7411	0.6213	0.6331	0.6512	0.6551	0.7492	0.7513	0.5932	0.6234	0.6338	0.6416
	UCCH [19]	0.7322	0.7387	0.6788	0.6811	0.6588	0.6629	0.7527	0.7566	0.6012	0.6418	0.6273	0.6562
	DSARH(ours)	0.7325	0.7592	**0.7262**	**0.7286**	**0.6612**	**0.6655**	**0.7552**	**0.758**	**0.7382**	**0.7427**	**0.7339**	**0.7358**
$I \rightarrow T$ (Attacked)	DJSRH [17]	0.5811	0.5852	0.3938	0.3964	0.4002	0.4054	0.5921	0.5912	0.4012	0.4123	0.4195	0.4296
	AGAH [52]	0.5852	0.5892	0.4128	0.4326	0.4402	0.4382	0.5974	0.6013	0.3252	0.3332	0.3414	0.3283
	SSAH [53]	0.5824	0.592	0.3943	0.4401	0.4032	0.4112	0.5966	0.6125	0.4117	0.4218	0.4232	0.4082
	DCMH [6]	0.6032	0.6083	0.4154	0.4292	0.4443	0.4454	0.6172	0.6253	0.4412	0.4512	0.4464	0.4545
	DSAH [65]	0.5816	0.5842	0.4027	0.4158	0.4232	0.4291	0.5913	0.5995	0.4264	0.4316	0.4337	0.4286
	UKD [66]	0.5917	0.5936	0.5111	0.5142	0.5146	0.5172	0.6113	0.6154	0.3116	0.3154	0.3063	0.3372
	DGCPN [55]	0.5882	0.5913	0.5081	0.5112	0.5172	0.5213	0.6124	0.6112	0.3553	0.3728	0.3883	0.3939
	UCCH [19]	0.5831	0.5829	0.5112	0.5133	0.5154	0.5162	0.5856	0.5894	0.3363	0.3625	0.3246	0.3327
	DSARH(ours)	**0.6352**	**0.6332**	**0.5216**	**0.5225**	**0.6227**	**0.6244**	**0.667**	**0.6733**	**0.6424**	**0.6453**	**0.6414**	**0.6383**
$T \rightarrow I$ (Regular)	DBRC [50]	0.5941	0.5955	0.5944	0.5936	0.4252	0.4212	0.4282	0.4363	0.6254	0.6662	0.7017	0.7223
	UGACH [51]	0.6765	0.6921	0.7033	0.7072	0.6021	0.6123	0.6281	0.6377	0.3818	0.3973	0.4822	0.4715
	DJSRH [17]	0.6294	0.6586	0.6601	0.6823	0.4761	0.4892	0.5354	0.5366	0.6412	0.6868	0.7018	0.7223
	AGAH [52]	**0.7562**	**0.7866**	**0.7977**	**0.8018**	0.6362	0.6691	0.6753	0.6744	0.6117	0.6186	0.6222	0.6196
	SSAH [53]	0.6383	0.6694	0.6836	0.6852	0.4887	0.4948	0.5266	0.5392	0.7126	0.7284	0.7193	0.6792
	DCMH [6]	0.7416	0.7452	0.7567	0.7636	0.6202	0.6343	0.6435	0.6486	0.6197	0.6168	0.6186	0.6211
	DSAH [65]	0.6782	0.6926	0.7025	0.7186	0.5801	0.5985	0.6154	0.6166	0.7053	0.7116	0.7087	0.7232
	UKD [66]	0.7056	0.7052	0.7213	0.7236	0.6292	0.6558	0.6554	0.6617	0.4991	0.5683	0.5412	0.6376
	DGCPN [55]	0.7182	0.7221	0.7457	0.7486	0.6283	0.6391	0.6552	0.6563	0.5966	0.6251	0.6332	0.6344
	UCCH [19]	0.7211	0.7252	0.7414	0.7454	0.6813	0.6838	0.6837	0.6856	0.6082	0.6514	0.6655	0.6667
	DSARH(ours)	0.7063	0.7092	0.7884	0.7946	**0.6884**	**0.6891**	**0.6902**	**0.699**	**0.7045**	**0.7517**	**0.7316**	**0.7362**
$T \rightarrow I$ (Attacked)	DJSRH [17]	0.5782	0.5854	0.5923	0.5987	0.3983	0.4092	0.4147	0.4286	0.4282	0.5111	0.5516	0.6112
	AGAH [52]	0.5912	0.5953	0.6022	0.6114	0.4226	0.4362	0.4447	0.4533	0.3985	0.3926	0.4012	0.4116
	SSAH [53]	0.5776	0.5857	0.5962	0.6026	0.4037	0.4113	0.4152	0.4284	0.2282	0.2313	0.2481	0.2535
	DCMH [6]	0.6066	0.6112	0.6133	0.6257	0.4176	0.4292	0.4321	0.4433	0.3554	0.3422	0.3611	0.3492
	DSAH [65]	0.5833	0.5872	0.5962	0.6013	0.4094	0.4122	0.4223	0.4352	0.4184	0.4233	0.4167	0.4346
	UKD [66]	0.5811	0.5825	0.5916	0.5943	0.5044	0.5112	0.5127	0.5196	0.2868	0.2923	0.2912	0.3021
	DGCPN [55]	0.5762	0.5822	0.5883	0.5914	0.4897	0.4955	0.5016	0.5032	0.3314	0.3453	0.3462	0.3411
	UCCH [19]	0.5432	0.5433	0.5455	0.5486	0.5014	0.5015	0.5056	0.5112	0.3535	0.3673	0.3657	0.3592
	DSARH(ours)	**0.6062**	**0.6043**	**0.6042**	**0.6051**	**0.518**	**0.5259**	**0.5306**	**0.5311**	**0.6111**	**0.6234**	**0.6157**	**0.6146**

several benchmark deep hashing models, including DJSRH [17], AGAH [52], SSAH [53], DCMH [6], DSAH [18], UKD [66], DGCPN [55], and UCCH [19], and conduct a series of robustness evaluation experiments.

Based on the extensive experimental results presented in TABLE I, we have pinpointed the vulnerability of DL architectures in retrieval tasks, mirroring findings in other computer vision tasks [67]. Specifically, Fig. 2 showcases examples of retrieval performance under both regular and robust models. It suggests that perturbations added to images, which might be imperceptible to human observers, can significantly alter their semantics from the viewpoint of retrieval models. Consequently, traditional training approaches that solely minimize triplet loss functions struggle to capture semantics in high-dimensional images that align with human recognition. Furthermore, unlike supervised classification or recognition tasks, we have verified that learned hash codes from textual data can also be used as targets for generating effective perturbations. Additionally, incorporating hash codes into end-to-end adversarial training can improve the retrieval models' capability to establish more robust correlations across diverse modalities.

In addition, we also assess the robustness of benchmark models under different types of adversarial attacks that are specifically crafted for cross-modal retrieval tasks on the NUSWIDE dataset, including Noise [68], CMLA [29], DACM [69], and AACH [28]. TABLE II presents the $I \rightarrow T$ mAP performance obtained under varied attacked types, and we also provide the complete experimental results in TABLE II and TABLE III of Supplementary Material. It can be observed that the mAP values of benchmark models on both $I \rightarrow T$ and $T \rightarrow I$ tasks of large-scale cross-modal datasets are significantly degraded. Furthermore, we find that the proposed DSARH can also achieve comparable mAP performance under attacked circumstances.

Additionally, we present the nDCG performance under attacked conditions in Fig. 1 of the Supplementary Material. It can be found that the multi-label retrieval performance of benchmark methods degrades significantly, whereas the proposed DSARH consistently achieves competitive robust

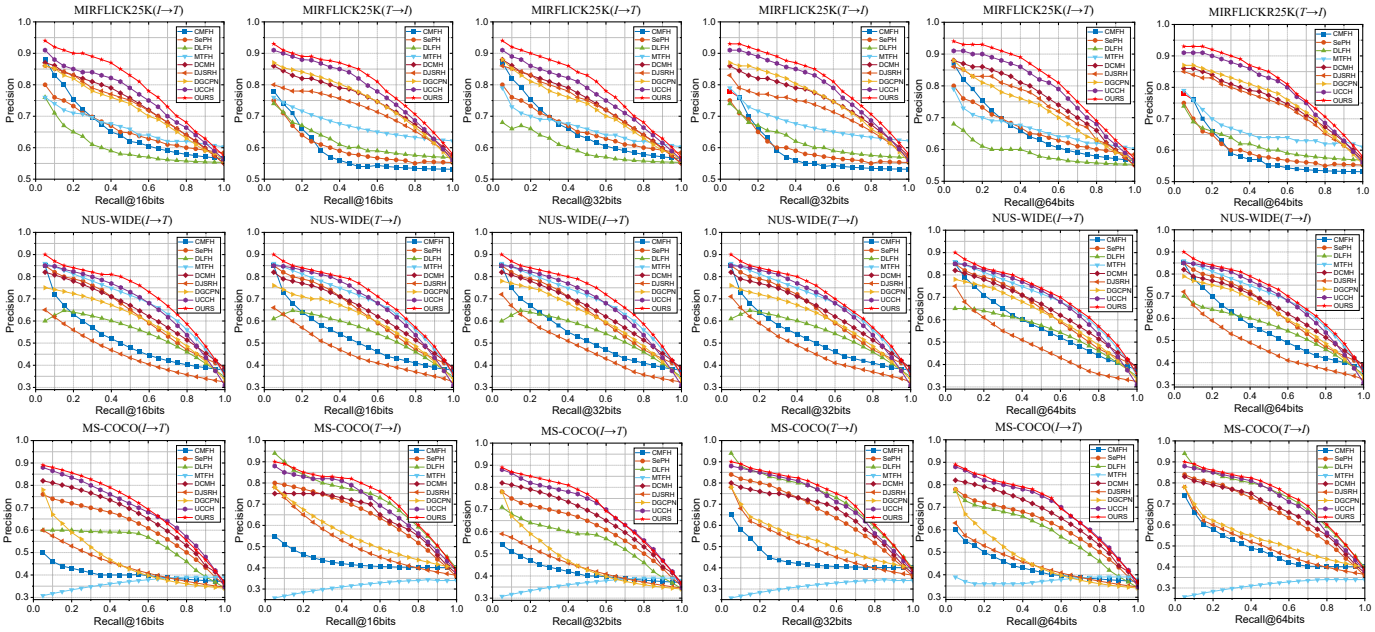
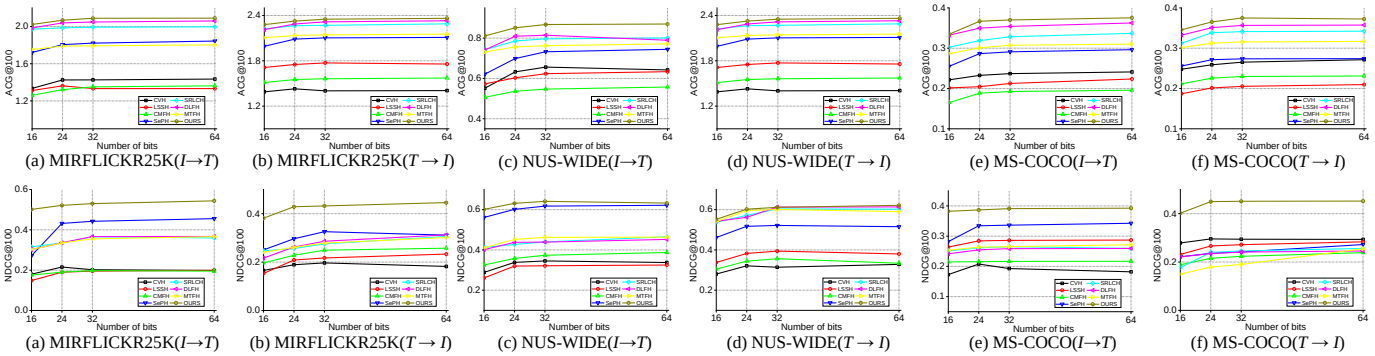


Fig. 4: The Precision-recall Curves of Deep Cross-modal Hashing Methods



(a) MIRFLICKR25K($I \to T$) (b) MIRFLICKR25K($T \to I$) (c) NUS-WIDE($I \to T$) (d) NUS-WIDE($T \to I$) (e) MS-COCO($I \to T$) (f) MS-COCO($T \to I$)

(a) MIRFLICKR25K($I \to T$) (b) MIRFLICKR25K($T \to I$) (c) NUS-WIDE($I \to T$) (d) NUS-WIDE($T \to I$) (e) MS-COCO($I \to T$) (f) MS-COCO($T \to I$)

Fig. 5: nDCG and ACG Comparison with the State-of-the-Art Deep Hashing Methods

performance. Thus, even though DSARH doesn't have direct access to the ground-truth concepts of samples, it can still promote models to extract more accurate semantic features from images, facilitating the establishment of more reliable correlations between heterogeneous data.

E. Comparisons with Shallow Cross-modal Hashing Methods

We also provide a comprehensive comparison with the most prominent shallow hashing methods. For fair comparison, we extract the learned features from the penultimate layer of deep convolutional classifiers as input for shallow models. In TABLE III, we present the mAP performance of benchmark baselines and the proposed DSARH on Wikipedia and NUSWIDE, and a more comprehensive comparison is presented in the TABLE I of Supplementary Material. We observe that shallow models are more proficient on the small Wikipedia dataset. The main reason lies in the fact that DL architectures are always composed of a huge number of parameters, thus samples required to acquire out-of-the-sample generalization for DL models are much larger than those for shallow models [70].

However, those shallow models often encounter performance degradation on large-scale datasets. Since DL-based retrieval models are more proficient on large-scale datasets. Moreover, the features learned by DL based classifiers may not be very consistent with features used for retrieval, feature representations need to be fine-tuned to adapt to new tasks for shallow models [12]. Therefore, addressing the vulnerability issues of DL architectures presents a promising direction for advancing large-scale cross-modal retrieval tasks.

F. Results of Single-Modal Retrieval

In this section, we present a comprehensive comparison of the regular and robust performance of DSARH on image retrieval tasks with the state-of-the-art deep hashing methods. We re-trained SDH [36], CSQ [61], DSH [20], ADSH [22], DSDH [21], GTelecomNet [62], SCADH [63], SGDH [64] and A^2-NET [24] for robust evaluation. Adversarial perturbations are produced using hash codes and the corresponding semantic relation matrix from a subset of the retrieval set and added to the query set to interfere with targeted models.

The mAP of benchmark baselines and DSARH on NUS-WIDE and MS-COCO are displayed in TABLE IV, respec-

TABLE II: $I \rightarrow T$ Comparison with the State-of-the-Art Deep Cross-modal Hashing Methods under Attacked Circumstance

	Attack Mothods	NUS-WIDE			
		16	32	64	128
DCMH [6]	Original	0.6577	0.6599	0.7361	0.7371
	Noise [68]	0.6674	0.6536	0.7152	0.7214
	CMLA [29]	0.4824	0.4985	0.5669	0.5636
	DACM [69]	0.4021	0.4058	0.4652	0.4711
	AACH [28]	0.5012	0.5123	0.5936	0.5863
DJSRH [17]	Original	0.4994	0.5249	0.6697	0.6725
	Noise [68]	0.4856	0.5299	0.6539	0.6696
	CMLA [29]	0.3612	0.3652	0.4569	0.4685
	DACM [69]	0.3017	0.3074	0.4025	0.4111
	AACH [28]	0.3856	0.3925	0.5162	0.5225
DGCPN [55]	Original	0.6512	0.6551	0.7492	0.7513
	Noise [68]	0.6621	0.6325	0.7263	0.7469
	CMLA [29]	0.4869	0.4952	0.6014	0.6025
	DACM [69]	0.3825	0.3947	0.4698	0.4717
	AACH [28]	0.5236	0.5265	0.6242	0.6325
UCCH [19]	Original	0.6588	0.6629	0.7527	0.7566
	Noise [68]	0.6152	0.6523	0.7425	0.7369
	CMLA [29]	0.4625	0.4635	0.5414	0.5426
	DACM [69]	0.3954	0.4011	0.4847	0.4919
	AACH [28]	0.4825	0.4858	0.5714	0.5808
DSARH(Ours)	Original	0.6612	0.6655	0.7552	0.758
	Noise [68]	0.6539	0.6712	0.7514	0.7489
	CMLA [29]	0.5525	0.5625	0.6458	0.6524
	DACM [69]	0.5241	0.5248	0.6014	0.6127
	AACH [28]	0.5825	0.5833	0.6758	0.6787

TABLE III: Comparison with the State-of-the-Art Shallow Hashing Methods

Mothods	WIKI				NUS-WIDE			
	16	32	64	128	16	32	64	128
SePH [2]	0.2687	0.2918	0.3051	0.3086	0.5303	0.5381	0.5415	0.5437
CVH [3]	0.1257	0.1212	0.1215	0.1171	0.3687	0.4182	0.4602	0.4466
LSSH [43]	0.2141	0.2216	0.2218	0.2211	0.39	0.3924	0.3962	0.3966
CMFH [71]	0.2416	0.2214	0.2302	0.2337	0.3523	0.3587	0.3596	0.3579
GSePH_rand [72]	0.2835	0.2916	0.3033	0.3114	0.5185	0.5395	0.5457	0.5516
SCM [16]	0.1725	0.1564	0.1523	0.1531	0.5106	0.5361	0.5439	0.5507
STMH [44]	0.1941	0.2476	0.2504	0.2519	0.4364	0.5529	0.5907	0.6128
SRLCH [45]	**0.3318**	**0.3526**	0.3269	0.3622	0.6158	0.6411	0.6421	0.6555
DLFH [46]	0.2916	0.2988	0.3217	0.3525	0.6111	0.6444	0.6521	0.6497
MTFH [47]	0.3211	0.3526	**0.3429**	**0.3625**	0.3121	0.3254	0.3124	0.365
DSARH(ours)	0.2742	0.2856	0.2933	0.3021	**0.6385**	**0.6476**	**0.6528**	**0.6571**
SePH [2]	0.6349	0.642	0.6649	0.6707	0.6203	0.6354	0.6372	0.6445
CVH (2)	0.1185	0.1034	0.1024	0.099	0.3646	0.4024	0.4339	0.4225
LSSH [43]	0.5031	0.5224	0.5293	0.5346	0.4286	0.4248	0.4248	0.4175
CMFH [71]	0.612	0.5446	0.5599	0.5652	0.3524	0.3564	0.3573	0.3562
GSePH [72]	0.6458	0.6631	0.6723	0.6748	0.6128	0.6429	0.6572	0.659
SCM [16]	0.1579	0.1421	0.1323	0.1268	0.4863	0.505	0.5138	0.5182
STMH [44]	0.5828	0.6114	0.6251	0.636	0.6737	0.706	0.7228	0.7284
SRLCH [45]	**0.7214**	**0.7111**	**0.7028**	**0.7311**	0.7011	0.7139	0.7288	0.7327
DLFH [46]	0.6511	0.6632	0.6819	0.6714	0.7124	0.7454	**0.7765**	0.7336
MTFH [47]	0.7011	0.7025	0.6964	0.6922	0.4112	0.3555	0.4151	0.5214
DSARH(ours)	0.6259	0.6325	0.6478	0.6506	**0.7125**	**0.7461**	0.7296	**0.7344**

tively. It can be observed that the retrieval performance is significantly degraded under the proposed adversarial perturbations. Notably, we did not directly utilize ground-truth label information to generate attacks; instead, only the semantic correlation matrix was provided. Therefore, the learned features for correlation measurement may not be semantically meaningful enough. Furthermore, we observe that DSARH achieves the best robust performance on both datasets. This indicates that the proposed asymmetric adversarial training mechanism can effectively enhance the robustness of retrieval models on large-scale image datasets.

However, we find that the standard retrieval performance of DSARH degrades on both datasets. The standard and robust trade-off issue of adversarial training is also widely prevalent in other computer vision tasks [9]. This suggests that the robust and regular performance cannot be improved simultaneously using adversarial training on image retrieval issues. While several other baselines are more competitive on regular image retrieval. Specifically, SCADH approaches tag refinement through a matrix decomposition problem without accounting for textual contextual similarity during compact feature learning. And SGDH proposes a binary matrix decomposition-based approach that preserves data structure and utilizes label information more effectively to guide hashing learning. Therefore, concept information is crucial for achieving effective image retrieval performance.

TABLE IV: Comparison with the State-of-the-Art Image Retrieval Methods

Mothods	NUS-WIDE				MS-COCO			
	8	16	32	64	12	24	36	48
ITQ [56]	0.7141	0.7361	0.7463	0.7553	0.5782	0.6162	0.6513	0.6535
LFH [46]	0.7122	0.7682	0.7952	0.8142	0.6543	0.6844	0.6927	0.6952
FashH [58]	0.7279	0.7697	0.7825	0.8048	0.6738	0.6756	0.6742	0.6767
ADGH [59]	0.7219	0.7353	0.7474	0.7526	-	-	-	-
COSDISH [60]	0.7801	0.7903	0.7929	0.7974	0.6555	0.6745	0.6918	0.7014
SDH [36]	-	-	-	-	0.5444	0.5527	0.5575	0.5632
CSQ [61]	-	0.7722	-	0.7772	0.6379	0.7025	0.7141	0.7183
DSH [20]	0.6532	0.6881	0.6954	0.6993	0.6825	0.6936	0.6888	0.7022
ADSH [22]	0.7937	**0.8537**	**0.8638**	**0.8728**	**0.8334**	**0.8454**	**0.8559**	**0.8597**
DSDH [21]	0.7617	0.7537	0.7925	0.7904	0.7032	0.7111	0.7187	0.7224
GTelecomNet-CSQ [62]	0.5513	0.5655	0.5812	0.5716	0.7029	0.7089	0.7121	0.7112
SCADH [63]	0.6088	0.6212	0.6337	0.6397	0.7885	0.7895	0.7926	0.7969
SGDH [64]	0.4811	0.4996	0.5016	0.5059	0.7812	0.7844	0.7862	0.7916
A^2-NET [24]	**0.7956**	0.8321	0.8471	0.8636	0.8211	0.8285	0.8314	0.8533
DSARH(ours)	0.6243	0.6288	0.6519	0.6672	0.8316	0.8429	0.8443	0.8464
CSQ [61]	-	0.2488	-	0.2601	0.1827	0.1957	0.2124	0.2096
DSH [20]	0.2324	0.2466	0.2565	0.2565	0.2255	0.2515	0.2482	0.2615
ADSH [22]	0.2602	0.2662	0.2732	0.2759	0.2632	0.2749	0.2859	0.2795
DSDH [21]	0.2483	0.2443	0.2609	0.2627	0.1746	0.1765	0.1815	0.1827
GTelecomNet-CSQ [62]	0.3217	0.3297	0.3545	0.3514	0.3511	0.3534	0.3497	0.3386
SCADH [63]	0.3452	0.3539	0.3666	0.3785	0.3127	0.3229	0.3562	0.3555
SGDH [64]	0.3696	0.3555	0.3684	0.3816	0.3335	0.3325	0.3416	0.3414
A^2-NET [24]	0.3325	0.3415	0.3654	0.3745	0.4569	0.4685	0.4752	0.4788
DSARH(ours)	**0.4767**	**0.4747**	**0.4683**	**0.4091**	**0.5619**	**0.5654**	**0.5657**	**0.5713**

Furthermore, we evaluate the robustness of DSARH against adversarial examples generated from classification tasks. Utilizing a pre-trained ResNet50 as our baseline model, we craft attacks driven by ground-truth information using the PGD approach. TABLE V showcases the $I \rightarrow T$ and $I \rightarrow I$ (using image to retrieval image) performance of both regularly trained and adversarially trained retrieval models. It is evident that adversarial examples can effectively transfer across different contexts, and DSARH can mitigate the impact of these attacks. This underscores the transferability of adversarial attacks between classification and retrieval tasks. From the vantage point of regular retrieval models, the semantic interpretations of the attacked samples undergo alterations. Additionally, the robustness of retrieval models can be bolstered without necessitating ground-truth label information. This suggests that the similarity matrix can also furnish valuable semantic insights for conducting adversarial training, thereby enhancing inherent robustness.

G. Parameter Sensitivity Analysis

The key innovation in our DSARH involves incorporating an adversarial training regularization term into the initial loss function. These adversarial perturbations are generated using the objective model's current parameters and training samples. We experiment with various configurations to validate the efficacy of this component. The hyperparameters utilized in DSARH comprise λ, which balances the regular and robust

TABLE V: Retrieval Performance of Against Adversarial Samples Generated for Classification Tasks

Task	Target attack model	NUS-WIDE			
		16	32	64	128
I→*T*	Regular Model	0.5563	0.5621	0.5963	0.6002
	Robust Model	**0.6363**	**0.6375**	**0.6363**	**0.6295**
	Target Attack	8	16	32	64
I→*I*	Regular Model	0.5412	0.5464	0.5556	0.5621
	Robust Model	**0.5825**	**0.5941**	**0.6123**	**0.6455**

components in the loss function, and ϵ, representing the maximum magnitude of the generated perturbation at each iteration. Furthermore, we adjust the total number of steps of PGD during the training process.

Specifically, λ represents the weight of robustness in the objective function. The objective function is a regular triplet loss function when λ is set to 0, while the loss function becomes a completely min-max training mechanism when $\lambda = 1$. ϵ denotes the maximum magnitude of the perturbations added to each pixel of training images, which is restricted to be small enough to avoid human detection. Typically, ϵ is selected from the set $2/255, 4/255, 8/255$ for normalized images [9].

We display the cross-modal retrieval performance of DSARH on three datasets with varied λ in Fig.6. We tried different settings from 0 to 1 with an interval of 0.1. It can be observed that the mAP performance of the retrieval model was significantly improved with the introduction of adversarial training. However, the mAP value starts to decrease with the further increase in the proportion of the adversarial item. Additionally, the optimal proportion differs across different datasets. This indicates that involving adversarial training can promote models to discover more discriminative features. However, an excessively large proportion of the adversarial item could hinder the model's regular performance. According to these observations, we see that robust features of images can help establish more exact correlation for heterogeneous modalities. Yet, the semantic information brought by robust features is insufficient to replace regular features.

TABLE VI: Robust mAP performance with Varied Perturbation Magnitude ϵ

Dataset	Image → Text			Text → Image		
	ϵ@2	ϵ@4	ϵ@8	ϵ@2	ϵ@4	ϵ@8
MIRFLICKR25K	0.5286	0.5925	0.6305	0.5002	0.5725	0.6012
NUS-WIDE	0.5625	0.6356	0.6696	0.4563	0.5126	0.5365
MS-COCO	0.5312	0.6036	0.6356	0.5123	0.5862	0.6196

We also investigated the effect of ϵ. Specifically, TABLE VI shows the performance at varying ϵ. We observe that the mAP results are relatively robust at 2/255, 4/255 and 8/255, while the performance at 2/255 is significantly lower. This suggests that 2/255 is excessively small to explore discriminative features from images. ϵ measures the maximum magnitude of perturbations added to images, which is commonly restricted to be smaller than 8/255 to ensure that the attacked images could evade human recognition. Notably, the principle of adversarial training is to execute a min-max training manner to find the worst-case perturbation at the inner training cycle. Based on PGD, the worst-case perturbation is always iteratively optimized with a subtle increment. Hence, an overly small ϵ may constrain the searching process from the requirement of the adversarial training algorithm, which prevents the model from exploring robust features from images, potentially compromising the retrieval performance of the models.

H. Discussion and Analysis

In terms of computational complexity, it is worth noting that the computational demands of adversarial training are typically several times higher than those of standard training methods. This increased complexity primarily stems from the inner maximization process used to identify the worst-case perturbation, which necessitates iterative updates to the perturbation based on the gradient of the loss function computed on the training samples. The balance between the robustness and efficiency of DL models has become a contentious issue in contemporary AI research. While achieving robustness demands increased computational resources, it also brings forth valuable attributes for AI models. For example, adversarial training has been demonstrated to enhance the reliability and cross-scenario generalization capabilities of AI models in various real-world applications [9].

Fortunately, implementing adversarial training for retrieval tasks is less computationally intensive than typical supervised computer vision tasks. Given that matrix multiplication processes demand substantial computational resources, it is feasible to learn hash codes from training samples and then extend this to the entire database. In contrast to classification tasks, the training set required for achieving convergence in retrieval tasks is considerably smaller. We further sampled different training sizes to investigate the convergence performance of DSARH. We implemented a reference experiment without the adversarial training item and shared the same initialization and base structure with DSARH. Fig. 7 shows the mAP performance under varied training sizes. It can be observed that while adversarial training always requires a larger training set to achieve promising retrieval performance, it can converge to a better outcome on a larger training set. Similarly, on typical computer vision tasks, robust training approaches were also demonstrated to require higher training data volume for convergence [67]. Moreover, different from adversarial training approaches designed for classification or recognition tasks in the computer vision field, we also verified that DSARH can acquire robustness without sacrificing regular performance on cross-modal retrieval issues. It can enhance the semantic exploration capacity of retrieval models on large-scale datasets, resulting in better generalization and robustness on the out-of-sample set.

Moreover, we also examined the retrieval performance under varied iteration steps, $n = 1, 3, 10$, where ϵ and the step size at each iteration are set as 8/255 and 2/255. TABLE VII displays the model's regular and robust performance on FLICKR25K, and we also computed the training time of each circumstance. It can be seen that both the regular and robustness performance are consistently improved with the increase in iteration steps. Notably, PGD-1 based adversarial

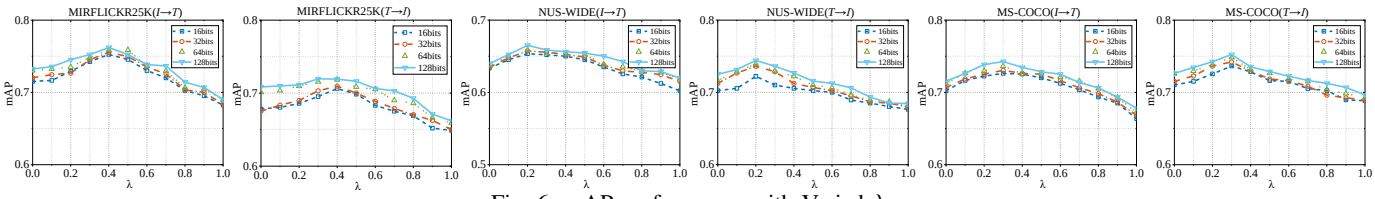


Fig. 6: mAP performance with Varied λ

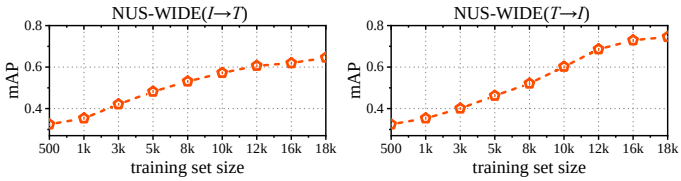


Fig. 7: mAP Performance with Varied Training Size

TABLE VII: The Regular and Robust Retrieval Performance with Varied Iteration Steps

iteration steps	Image → Text		Text → Image	
	Regular	Robust	Regular	Robust
1	0.7423	0.5502	0.7896	0.5325
3	0.7253	0.5952	0.7685	0.5626
10	0.7108	0.6153	0.7425	0.5812

training, which only executes a FGSM step during the training process, can also boost the model's retrieval performance. PGD-1 attack only requires one gradient computation using the matrix multiplication of hashing codes on training samples, thus it will not introduce additional computational cost from matrix multiplication, which accounts for the primary computational resource consumption. Therefore, PGD-1 based adversarial training can be an effective and efficient approach for enhancing the generalization and robustness of cross-modal retrieval models. The results presented above emphasize the significance and practicality of robustness in deep hashing models for retrieval tasks.

V. CONCLUSION

In this paper, we have introduced a framework called Deep Supervised Adversarial Robust Hashing (DSARH) for robust cross-modal and image retrieval. DSARH incorporates adversarial training techniques into the architecture of deep hashing, aiming to derive robust features from high-dimensional data modalities. By resisting well-designed attacks applied to images, DSARH is anticipated to learn more reliable features to that encapsulate the semantic and conceptual details of high-dimensional samples, rather than merely focusing on highly predictive features. Specifically, DSARH leverages the gradients of learned hash codes from samples to generate effective perturbations, which are combined with the similarity matrix and quantified similarity between the model's predictions of samples. Thus, utilizing gradient information, DSARH can facilitate end-to-end adversarial training on retrieval tasks. Additionally, we provide two distinct training schemes tailored for cross-modal and image retrieval tasks, respectively. As a result, DSARH can uncover more robust features to establish a more reliable correlation between multi-modal data. Extensive experiments across various cross-modal and image retrieval tasks have validated DSARH's capability to enhance robustness. The robust features acquired by DSARH effectively mitigate the heterogeneity issue between disparate data modalities, improving both the regular and robust retrieval performance of deep hashing models in cross-modal retrieval tasks. This underscores the significance of adversarial training in the realm of cross-modal retrieval. In future endeavors, we aim to explore other forms of adversarial training for additional data modalities, such as videos and audio, to broaden the applicability of deep hashing models in multi-modal retrieval tasks. Additionally, our research will strive to further refine the trade-off performance in image retrieval tasks.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3864–3872.

[3] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Twenty-second international joint conference on artificial intelligence*, 2011.

[4] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3157–3166, 2016.

[5] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *2017 international conference on communication and signal processing (ICCSP)*. IEEE, 2017, pp. 0588–0592.

[6] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3232–3240.

[7] L. Zhu, C. Zheng, W. Guan, J. Li, Y. Yang, and H. T. Shen, "Multi-modal hashing for efficient multimedia retrieval: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, pp. 239–260, 2023.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[9] X. Zhang, X. Zheng, and W. Mao, "Adversarial perturbation defense on deep neural networks," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–36, 2021.

[10] C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon, "Understanding adversarial examples from the mutual influence of images and perturbations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 521–14 530.

[11] Z. Zheng, L. Zheng, Y. Yang, and F. Wu, "U-turn: Crafting adversarial queries with opposite-direction features," *International Journal of Computer Vision*, vol. 131, no. 4, pp. 835–854, 2023.

[12] L. Zhu, T. Wang, J. Li, Z. Zhang, J. Shen, and X. Wang, "Efficient query-based black-box attack against cross-modal hashing retrieval," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–25, 2023.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[14] X. Liu, Z. Hu, H. Ling, and Y.-m. Cheung, "Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 964–981, 2021.

[15] J. Wang, T. Zhang, j. song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 769–790, 2018.

[16] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.

[17] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3027–3035.

[18] D. Yang, D. Wu, W. Zhang, H. Zhang, B. Li, and W. Wang, "Deep semantic-alignment hashing for unsupervised cross-modal retrieval," in *Proceedings of the 2020 international conference on multimedia retrieval*, 2020, pp. 44–52.

[19] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3877–3889, 2023.

[20] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2064–2072.

[21] Q. Li, Z. Sun, R. He, and T. Tan, "Deep supervised discrete hashing," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Q.-Y. Jiang and W.-J. Li, "Asymmetric deep supervised hashing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[23] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie, "Scalable deep hashing for large-scale social image retrieval," *IEEE Transactions on image processing*, vol. 29, pp. 1271–1284, 2019.

[24] X.-S. Wei, Y. Shen, X. Sun, P. Wang, and Y. Peng, "Attribute-aware deep hashing with self-consistency for large-scale fine-grained image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 904–13 920, 2023.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations*, 2014.

[26] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[27] E. Yang, T. Liu, C. Deng, and D. Tao, "Adversarial examples for hamming space search," *IEEE transactions on cybernetics*, vol. 50, no. 4, pp. 1473–1484, 2018.

[28] C. Li, S. Gao, C. Deng, W. Liu, and H. Huang, "Adversarial attack on deep cross-modal hamming retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2218–2227.

[29] C. Li, S. Gao, C. Deng, D. Xie, and W. Liu, "Cross-modal learning with adversarial samples," *Advances in neural information processing systems*, vol. 32, 2019.

[30] X. Zhang, X. Zheng, B. Liu, X. Wang, W. Mao, D. D. Zeng, and F.-Y. Wang, "Towards human–machine recognition alignment: An adversarilly robust multimodal retrieval hashing framework," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2847–2859, 2023.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "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, 2014, pp. 740–755.

[32] M. Zhou, L. Wang, Z. Niu, Q. Zhang, N. Zheng, and G. Hua, "Adversarial attack and defense in deep ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[33] Z. Zhang, Y. Chen, and V. Saligrama, "Efficient training of very deep neural networks for supervised hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, 2014.

[35] D. Wu, Q. Dai, J. Liu, B. Li, and W. Wang, "Deep incremental hashing network for efficient image retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9069–9077.

[36] F. Shen, C. Shen, W. Liu, and H. Tao Shen, "Supervised discrete hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 37–45.

[37] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[38] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "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*, 2009, pp. 1–9.

[39] Y. Cao, M. Long, J. Wang, and S. Liu, "Collective deep quantization for efficient cross-modal retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[40] V. E. Liong, J. Lu, L.-Y. Duan, and Y.-P. Tan, "Deep variational and structural hashing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 580–595, 2018.

[41] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[42] A. Neculai, Y. Chen, and Z. Akata, "Probabilistic compositional embeddings for multimodal image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4547–4557.

[43] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 415–424.

[44] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[45] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 10, pp. 3351–3365, 2021.

[46] Q.-Y. Jiang and W.-J. Li, "Discrete latent factor model for cross-modal hashing," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3490–3501, 2019.

[47] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 964–981, 2021.

[48] S. Cheng, L. Wang, and A. Du, "Deep semantic-preserving reconstruction hashing for unsupervised cross-modal retrieval," *Entropy*, vol. 22, no. 11, p. 1266, 2020.

[49] H. Hu, L. Xie, R. Hong, and Q. Tian, "Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3123–3132.

[50] X. Li, D. Hu, and F. Nie, "Deep binary reconstruction for cross-modal hashing," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17, 2017, p. 1398–1406.

[51] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[52] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proceedings of the 2019 on international conference on multimedia retrieval*, 2019, pp. 159–167.

[53] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4242–4251.

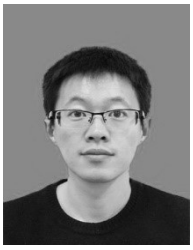
[54] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proceedings of the 2020 international conference on multimedia retrieval*, 2020, pp. 525–531.

[55] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 5, 2021, pp. 4626–4634.

[56] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2916–2929, 2012.

[57] P. Zhang, W. Zhang, W.-J. Li, and M. Guo, "Supervised hashing with latent factor models," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 173–182.

[58] G. Lin, C. Shen, Q. Shi, A. Van den Hengel, and D. Suter, "Fast supervised hashing with decision trees for high-dimensional data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1963–1970.
[59] X. Shi, F. Xing, K. Xu, M. Sapkota, and L. Yang, "Asymmetric discrete graph hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
[60] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 1711–1717.
[61] L. Yuan, T. Wang, X. Zhang, F. E. Tay, Z. Jie, W. Liu, and J. Feng, "Central similarity quantization for efficient image and video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[62] W. Zhao, C. Xu, Z. Guan, X. Wu, W. Zhao, Q. Miao, X. He, and Q. Wang, "Telecomnet: Tag-based weakly-supervised modally cooperative hashing network for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7940–7954, 2022.
[63] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie, "Scalable deep hashing for large-scale social image retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 1271–1284, 2020.
[64] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *International Journal of Computer Vision*, vol. 128, pp. 2265–2278, 2020.
[65] Y. Li and J. van Gemert, "Deep unsupervised image hashing by maximizing bit entropy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2002–2010.
[66] H. Hu, L. Xie, R. Hong, and Q. Tian, "Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[67] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey toward the defender's perspective," *ACM Computing Surveys*, vol. 55, no. 1, pp. 1–38, 2021.
[68] J. Bai, B. Chen, Y. Li, D. Wu, W. Guo, S.-t. Xia, and E.-h. Yang, "Targeted attack for deep hashing based retrieval," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 618–634.
[69] C. Li, H. Tang, C. Deng, L. Zhan, and W. Liu, "Vulnerability vs. reliability: Disentangled adversarial examples for cross-modal learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 421–429.
[70] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.
[71] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5427–5440, 2016.
[72] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.



Xingwei Zhang is currently a Associate Professor at the Institute of Automation, Chinese Academy of Sciences. He received Ph.D. degree from the School of Artificial Intelligence, University of Chinese Academy of Sciences, in 2022, M.S. degree in Electronic and information engineering from University of Chinese Academy of Sciences in 2018, and B.S. degree in electronic science and technology from the Department of Information and Electronic Engineering, Zhejiang University, Hangzhou, China, in 2014. His research interest includes multi-modal retrieval, adversarial robustness and social network security.

Gang Zhou received the M.S. degree in artificial intelligence from the School of Artificial Intelligence, Chinese Academy of Sciences, Beijing, China, in 2023.He is currently working toward the PhD degree with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include deep cross-modal hashing, adversarial machine learning theory.

Xiaolong Zheng (M'10) is currently a Professor at the Institute of Automation, Chinese Academy of Sciences. He received Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2009, M.S. degree from Beijing Jiaotong University in 2006, and B.S. degree from China Jiliang University in 2003. His research interests including social computing, big data analytics, knowledge graphs, financial technologies and complex system intelligence.

Wenji Mao received the B.S. degree in computer science from Jilin University, Changchun, China, in 1990, the M.S. degree in computer software and artificial intelligence from the Institute of Mathematics, Chinese Academy of Sciences, Beijing, China, in 1993, and the Ph.D. degree in computer science from the University of Southern California, Los Angeles, USA, in 2006. She is currently a Professor with the Institute of Automation, Chinese Academy of Sciences, and a chief professor with the School of Artificial Intelligence, University of Chinese Academy of Sciences. Her research interests include artificial intelligence, cross-modal retrieval and data mining.

Liang Wang (Fellow, IEEE) received the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a research assistant with Imperial College London, U.K., and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, U.K., respectively. Currently, he is a full professor of the Hundred Talents Program with the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals, such as the IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, and leading international conferences, such as CVPR, ICCV, and ECCV. He has served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and Pattern Recognition. He is an IAPR fellow.

Daniel Dajun Zeng (F'15) received the B.S. degree in economics and operations research from the University of Science and Technology of China, Hefei, China, in 1990 and the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University in 1998. He current is a Research Fellow position at the Institute of Automation, Chinese Academy of Sciences. His research interests include intelligence and security informatics, infectious disease informatics, social computing, recommender systems, software agents, and applied operations research and game theory. He has published more than 300 peer-reviewed articles. He currently serves as the editor in chief of ACM Transactions on MIS.

Supplementary Materials: Deep Supervised Adversarial Robust Hashing for Retrieval

Xingwei Zhang* Gang Zhou† Xiaolong Zheng* Wenji Mao*
Liang Wang* Daniel Dajun Zeng*

In this file, we provide the supplementary experimental results and network configution details for the main file, including the comparison of DSARH with the state-of-the-art shallow methods, the nDCG value of cross-modal retrieval under attacked circumstance, and the comparison with other benchmark baselines under varied attack types.

1 Experimental Results

Table I: Comparison with the State-of-the-Art Shallow Hashing Methods

Task	Mothods	WIKI				MIRFLICKR25k				NUS-WIDE			
		16	32	64	128	16	32	64	128	16	32	64	128
$I \rightarrow T$	SePH [1]	0.2687	0.2918	0.3051	0.3086	0.673	0.6743	0.6799	0.6828	0.5303	0.5381	0.5415	0.5437
	CVH [2]	0.1257	0.1212	0.1215	0.1171	0.6067	0.6177	0.6157	0.6074	0.3687	0.4182	0.4602	0.4466
	LSSH [3]	0.2141	0.2216	0.2218	0.2211	0.5784	0.5804	0.5797	0.5816	0.39	0.3924	0.3962	0.3966
	CMFH [4]	0.2416	0.2214	0.2302	0.2337	0.5452	0.5509	0.5506	0.551	0.3523	0.3587	0.3596	0.3579
	GSePH_rand [5]	0.2835	0.2916	0.3033	0.3114	0.6454	0.6607	0.6726	0.6816	0.5185	0.5395	0.5457	0.5516
	SCM [6]	0.1725	0.1564	0.1523	0.1531	0.628	0.6345	0.6385	0.649	0.5106	0.5361	0.5439	0.5507
	STMH [7]	0.1941	0.2476	0.2504	0.2519	0.5823	0.6183	0.6372	0.5621	0.4364	0.5529	0.5907	0.6128
	SRLCH [8]	**0.3318**	**0.3526**	0.3269	0.3622	0.7102	0.7128	0.7201	0.7159	0.6158	0.6411	0.6421	0.6555
	DLFH [9]	0.2916	0.2988	0.3217	0.3525	0.6775	0.6987	0.7102	0.7255	0.6111	0.6444	0.6521	0.6497
	MTFH [10]	0.3211	0.3526	**0.3429**	**0.3625**	0.5044	0.5151	0.5581	0.5741	0.3121	0.3254	0.3124	0.365
	DSARH(ours)	0.2742	0.2856	0.2933	0.3021	**0.7321**	**0.7598**	**0.7266**	**0.7289**	**0.6385**	**0.6476**	**0.6528**	**0.6571**
$T \rightarrow I$	SePH [1]	0.6349	0.642	0.6649	0.6707	0.7179	0.722	0.7307	0.7344	0.6203	0.6354	0.6372	0.6445
	CVH (2)	0.1185	0.1034	0.1024	0.099	0.6026	0.6041	0.6017	0.5972	0.3646	0.4024	0.4339	0.4225
	LSSH [3]	0.5031	0.5224	0.5293	0.5346	0.5898	0.5927	0.5932	0.5932	0.4286	0.4248	0.4248	0.4175
	CMFH [4]	0.612	0.5446	0.5599	0.5652	0.5433	0.5573	0.5575	0.5576	0.3524	0.3564	0.3573	0.3562
	GSePH [5]	0.6458	0.6631	0.6723	0.6748	0.6965	0.7174	0.7304	0.7446	0.6128	0.6429	0.6572	0.659
	SCM [6]	0.1579	0.1421	0.1323	0.1268	0.6176	0.6234	0.6285	0.6369	0.4863	0.505	0.5138	0.5182
	STMH [7]	0.5828	0.6114	0.6251	0.636	0.715	0.7414	0.7533	0.762	0.6737	0.706	0.7228	0.7284
	SRLCH [8]	**0.7214**	**0.7111**	**0.7028**	**0.7311**	0.7059	0.7255	0.7416	0.7426	0.7011	0.7139	0.7288	0.7327
	DLFH [9]	0.6511	0.6632	0.6819	0.6714	**0.7658**	**0.7881**	**0.7953**	**0.7999**	0.7124	0.7454	**0.7765**	0.7336
	MTFH [10]	0.7011	0.7025	0.6964	0.6922	0.5142	0.5563	0.5321	0.5998	0.4112	0.3555	0.4151	0.5214
	DSARH(ours)	0.6259	0.6325	0.6478	0.6506	0.7061	0.7088	0.7944	0.7138	**0.7125**	**0.7461**	0.7296	**0.7344**

*State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China. *(Corresponding author: Xiaolong Zheng. E-mail: xiaolong.zheng@ia.ac.cn)*

†School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.

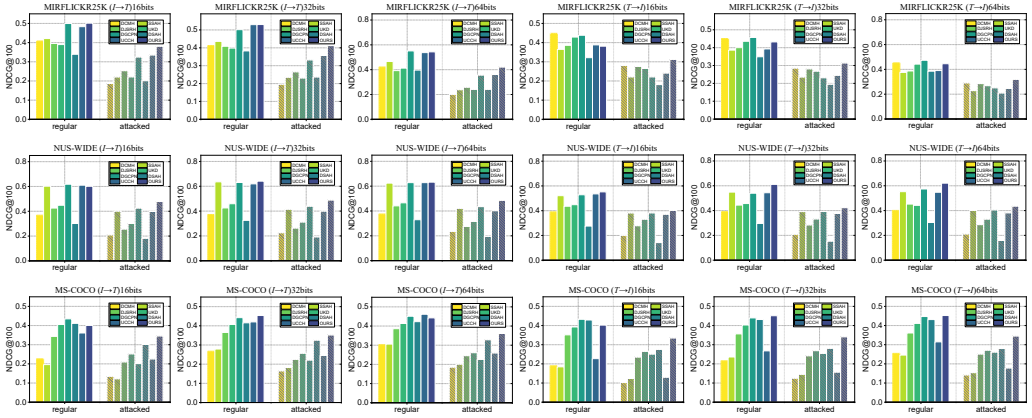


Fig. 1: nDCG Comparison with the State-of-the-Art deep cross-modal Hashing Methods under Attacked Circumstance

Table II: I2T Comparison with the State-of-the-Art Deep Cross-modal Hashing Methods under Attacked Circumstance

	Attack Mothods	NUS-WIDE			
		16	32	64	128
DCMH [12]	Original	0.6577	0.6599	0.7361	0.7371
	Noise [11]	0.6674	0.6536	0.7152	0.7214
	CMLA [13]	0.4824	0.4985	0.5669	0.5636
	DACM [14]	0.4021	0.4058	0.4652	0.4711
	AACH [15]	0.5012	0.5123	0.5936	0.5863
DJSRH [16]	Original	0.4994	0.5249	0.6697	0.6725
	Noise [11]	0.4856	0.5299	0.6539	0.6696
	CMLA [13]	0.3612	0.3652	0.4569	0.4685
	DACM [14]	0.3017	0.3074	0.4025	0.4111
	AACH [15]	0.3856	0.3925	0.5162	0.5225
DGCPN [17]	Original	0.6512	0.6551	0.7492	0.7513
	Noise [11]	0.6621	0.6325	0.7263	0.7469
	CMLA [13]	0.4869	0.4952	0.6014	0.6025
	DACM [14]	0.3825	0.3947	0.4698	0.4717
	AACH [15]	0.5236	0.5265	0.6242	0.6325
UCCH [18]	Original	0.6588	0.6629	0.7527	0.7566
	Noise [11]	0.6152	0.6523	0.7425	0.7369
	CMLA [13]	0.4625	0.4635	0.5414	0.5426
	DACM [14]	0.3954	0.4011	0.4847	0.4919
	AACH [15]	0.4825	0.4858	0.5714	0.5808
DSARH(Ours)	Original	0.6612	0.6655	0.7552	0.758
	Noise [11]	0.6539	0.6712	0.7514	0.7489
	CMLA [13]	0.5525	0.5625	0.6458	0.6524
	DACM [14]	0.5241	0.5248	0.6014	0.6127
	AACH [15]	0.5825	0.5833	0.6758	0.6787

Table III: T2I Comparison with the State-of-the-Art Deep Cross-modal Hashing Methods under Attacked Circumstance

	Attack Mothods	NUS-WIDE			
		16	32	64	128
DCMH [12]	Original	0.6202	0.6343	0.6435	0.6486
	Noise [11]	0.6123	0.6369	0.6518	0.6502
	CMLA [13]	0.4239	0.4374	0.4384	0.4394
	DACM [14]	0.3884	0.3914	0.4009	0.4089
	AACH [15]	0.4512	0.4624	0.4783	0.4795
DJSRH [16]	Original	0.4761	0.4892	0.5354	0.5366
	Noise [11]	0.4541	0.4888	0.5236	0.5123
	CMLA [13]	0.2945	0.2984	0.3314	0.3385
	DACM [14]	0.2614	0.2647	0.2988	0.3052
	AACH [15]	0.3123	0.3241	0.3647	0.3717
DGCPN [17]	Original	0.6283	0.6391	0.6352	0.6563
	Noise [11]	0.6052	0.6123	0.6236	0.6602
	CMLA [13]	0.4525	0.4625	0.4658	0.4721
	DACM [14]	0.3852	0.3914	0.3954	0.4001
	AACH [15]	0.5002	0.5111	0.5138	0.5321
UCCH [18]	Original	0.6813	0.6838	0.6837	0.6856
	Noise [11]	0.6636	0.6789	0.6924	0.6632
	CMLA [13]	0.4414	0.4428	0.4528	0.4652
	DACM [14]	0.3251	0.3314	0.3354	0.3414
	AACH [15]	0.4852	0.4867	0.4914	0.4936
DSARH(Ours)	Original	0.6884	0.6891	0.6902	0.699
	Noise [11]	0.6636	0.6912	0.6895	0.6958
	CMLA [13]	0.5733	0.5785	0.5812	0.5863
	DACM [14]	0.5352	0.5367	0.5412	0.5456
	AACH [15]	0.6025	0.6052	0.6125	0.6187

2 Network Configuration

We leverage the feature representation structures from pre-trained deep convolutional models to extract semantic feature representations for the hash network construction. Conversely, the text information in multi-modal datasets that has been previously encoded with numerical vectors, we employ a fully-connected network directly as the feature representation structure. Specifically, we adopt the approach outlined in [16] and employ AlexNet without the final decision-making layer as the image feature extraction framework in our deep cross-modal retrieval model. This feature extraction framework has been pre-trained using adversarial training on ImageNet and consists of 5 convolutional layers with 64, 192, 384, 256, and 256 kernels, along with 3 maximum pooling layers and 2 fully-connected layers. Subsequently, the extracted features, with dimensions of 4096, are further transmitted to the hash code network, which has a length of $[4096 \rightarrow 512 \rightarrow \mathrm{HL}]$. The hash network comprises a 512-dimensional semantic feature representation layer and a hash layer, where the hash code length is set as HL, responsible for generating image hash codes.

For the text hash network in the cross-modal retrieval task, we utilize fully-connected layers. To ensure better alignment with the semantic meaning of image samples, the dimensions of the semantic feature layer and hash layer are set to be equal to image hash network. Based on the dimension of text vectors, we set the length of the text network as $[10 \rightarrow 4096 \rightarrow 512 \rightarrow \mathrm{HL}]$, $[1386 \rightarrow 4096 \rightarrow 512 \rightarrow \mathrm{HL}]$, $[1000 \rightarrow 4096 \rightarrow 512 \rightarrow \mathrm{HL}]$, $[1386 \rightarrow 4096 \rightarrow 512 \rightarrow \mathrm{HL}]$ and $[2026 \rightarrow 4096 \rightarrow 512 \rightarrow \mathrm{HL}]$ for Wiki, FLICKR25K, NUSWIDE and MS-COCO respectively. Furthermore, following the methodology proposed by [19], we employ AlexNet as the feature extraction structure for single image retrieval tasks on both NUSWIDE and MS-COCO datasets. Subsequently, the features are fed forward to the semantic features and hash code layer, with the length of the entire network set as $[4096 \rightarrow 1000 \rightarrow \mathrm{HL}]$.

Our project and other benchmark baselines trained for robustness evaluation are implemented using PyTorch. These benchmark baselines are retrained according to the settings provided by the corresponding authors. We use ReLU as the exclusive nonlinear activation function for all semantic features, and Tanh is employed to approximate the sign hash codes. For training the single image retrieval model on both NUSWIDE and MS-COCO datasets, we perform a total of 50 iterations. At each iteration, we randomly select 2000 samples from the gallery set for training, which are then divided into batches with a batch size of 128. The model is trained for 3 epochs at each iteration. We utilize the Adam optimizer for updating parameters, with an initial learning rate set to 1e-4. The learning rate decays along with iterations using the Exponential learning rate method, with a decay rate set to 0.9. Additionally, a weight decay of 1e-5 is applied to prevent over-fitting.

For the cross-modal retrieval task, we utilize the SGD optimizer. And the learning rates for the image and text hash networks in the cross-modal retrieval model are set to 1e-3 and 1e-2, respectively. We use a momentum of 0.9 and a weight decay of 5e-4. The batch size for the training data is set to 128. A total of 200 epochs are performed on FLICKR25K, NUSWIDE and MS-COCO datasets. To achieve better convergence, we freeze the parameters of the pre-trained AlexNet but only update the fully-connected layers when training on the Wiki dataset.

References

[1] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3864–3872.

[2] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Twenty-second international joint conference on artificial intelligence*, 2011.

[3] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 415–424.

[4] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5427–5440, 2016.

[5] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.

[7] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[8] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 10, pp. 3351–3365, 2021.

[9] Q.-Y. Jiang and W.-J. Li, "Discrete latent factor model for cross-modal hashing," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3490–3501, 2019.

[10] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 964–981, 2021.

[11] J. Bai, B. Chen, Y. Li, D. Wu, W. Guo, S.-t. Xia, and E.-h. Yang, "Targeted attack for deep hashing based retrieval," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 618–634.

[12] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3232–3240.

[13] C. Li, S. Gao, C. Deng, D. Xie, and W. Liu, "Cross-modal learning with adversarial samples," *Advances in neural information processing systems*, vol. 32, 2019.

[14] C. Li, H. Tang, C. Deng, L. Zhan, and W. Liu, "Vulnerability vs. reliability: Disentangled adversarial examples for cross-modal learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 421–429.

[15] C. Li, S. Gao, C. Deng, W. Liu, and H. Huang, "Adversarial attack on deep cross-modal hamming retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2218–2227.

[16] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3027–3035.

[17] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 5, 2021, pp. 4626–4634.

[18] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3877–3889, 2023.

[19] Q.-Y. Jiang and W.-J. Li, "Asymmetric deep supervised hashing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.