Frequency Domain Adversarial Attacks on Deep Cross-Modal Hashing

Journal:	*IEEE Transactions on Systems, Man and Cybernetics: Systems*
Manuscript ID	SMCA-25-05-1940
Manuscript Type:	Regular Paper
Date Submitted by the Author:	22-May-2025
Complete List of Authors:	Zhou, Gang; Beijing University of Posts and Telecommunications School of Artificial Intelligence Xu, Shibiao; Beijing University of Posts and Telecommunications School of Artificial Intelligence Zheng, Xiaolong; Chinese Academy of Sciences Institute of Automation; University of the Chinese Academy of Sciences School of Artificial Intelligence Luo, Guiyang; Beijing University of Posts and Telecommunications State Key Laboratory of Networking and Switching Technology Wang, Feiyue.; Chinese Academy of Sciences Institute of Automation; desci center of parallel intelligence Óbuda University
Key Words:	Artificial intelligence, Computer security, Data security, Image texture analysis, Information retrieval

Frequency Domain Adversarial Attacks on Deep Cross-Modal Hashing

Gang Zhou, Shibiao Xu *Member, IEEE*, Xiaolong Zheng, *Member, IEEE*, Guiyang Luo, *Member, IEEE*, and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*—In recent years, multimodal data has experienced explosive growth. Deep cross-modal hashing models leverage deep neural networks and hashing techniques to bridge feature representation gaps by mapping multimodal data into a unified semantic space, enabling effective cross-modal retrieval. Binary hash coding enhances storage and retrieval efficiency. However, these models inherit the vulnerabilities of deep neural networks, making them susceptible to adversarial attacks. Current attack methods, which operate in the spatial domain, fail to recognize that deep hashing models predominantly encode semantic components in the low-frequency domain—a limitation that often results in spatial overfitting and misallocation of the perturbation budget to non-critical frequency regions. To address these challenges, we propose a frequency domain adversarial attack framework for cross-modal hashing (FACH). This approach integrates low-frequency masking and multi-teacher gradient fusion to identify critical low-frequency vulnerabilities shared across models. FACH generates adversarial examples with enhanced transferability by aligning semantic perturbations with spectral characteristics through an inverse transformation to the spatial domain. Experimental results demonstrate that FACH significantly outperforms existing transfer attack methods, unveiling the frequency domain vulnerabilities of deep hashing models.**

***Index Terms*—Deep Cross-Modal Retrieval, Adversarial attack, Frequency Domain, Deep Hashing**

I. Introduction

WITH the proliferation of social media and network communication technologies, the volume of data across diverse modalities has grown exponentially, creating a pressing demand for similarity retrieval between these modalities. However, the inherent differences in feature representations between cross-modal data introduce a ”heterogeneity gap,” which complicates direct semantic similarity comparisons and presents a significant challenge in cross-modal retrieval [1]–[4]. In recent years, the application of deep learning (DL) to encode different modalities into a compact, semantically preserving space has achieved remarkable success in cross-modal retrieval scenarios [5]–[9]. Deep hashing methods, which integrate DL with hashing techniques to map high-dimensional data into low-dimensional binary hash codes, have demonstrated efficient cross-modal retrieval capabilities while reducing computational and storage costs. This advancement has facilitated the widespread adoption of deep hashing technologies across large-scale cross-modal retrieval domains.

However, recent studies have shown that deep learning (DL) models are highly sensitive to carefully crafted adversarial perturbationss created by malicious attackers. These perturbations are often imperceptible to humans but can result in erroneous decision-making within DL-based systems. This finding raises significant concerns regarding the reliability of DL models. Similar vulnerabilities have also been observed in deep hashing methods [10]–[13]. When subjected to adversarial attacks, cross-modal retrieval systems may experience substantial deviations in their results, compromising both accuracy and security. For instance, well-designed adversarial samples in retrieval can result in the inappropriate retrieval of content by well-trained deep hashing models, including violence, pornography, or hate speech, even when such content should not be present in the results [12], [14]–[17], as shown in Fig. 1. This not only risks violating laws and regulations but could also trigger serious societal and ethical concerns.

The reliability of DL-based models is assessed through their robust performance under various adversarial attacks. Investigating the effects of adversarial perturbations on DL models is crucial for the design of reliable DL systems. Existing adversarial attack generation methods can be classified into white-box and black-box attacks based on the accessibility of the model's parameters and structural information [18]. White-box attacks leverage information from the target model to generate targeted adversarial attacks that induce erroneous predictions, typically employing gradient-based optimization techniques. In contrast, black-box attacks are implemented under conditions where direct access to the target model is unavailable. The most promising transfer-based black-box attack method constructs a substitute model that approximates the target model, designs well-crafted perturbations on this substitute, and then transfers them to the target model. The effectiveness of transfer-based attacks relies on the intrinsic vulnerabilities of DL structures, regardless of the specific tasks or datasets involved [19], [20]. In real-world scenarios, the inner workings of target models are often not accessible,

Corresponding author: Shibiao Xu, Xiaolong Zheng

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

Xiaolong Zheng is 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: xiaolong.zheng@ia.ac.cn)

Guiyang Luo is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: luoguiyang@bupt.edu.cn).

Fei-Yue Wang is with the desci center of parallel intelligence, Óbuda University, 1034 Budapest, Hungary, and also with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China,(e-mail: feiyue.wang@ia.ac.cn).

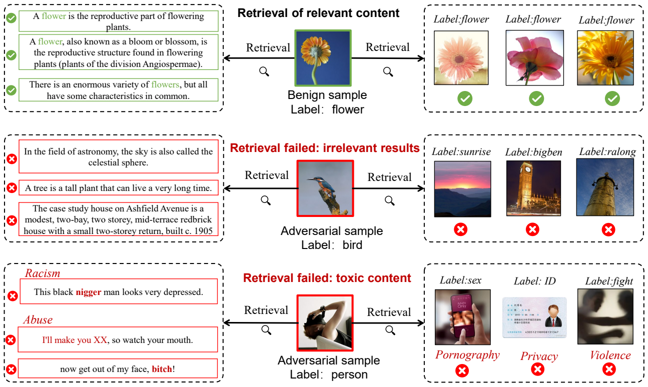

A flower is the reproductive part of flowering plants.
A flower, also known as a bloom or blossom, is the reproductive structure found in flowering plants (plants of the division Angiospermae).
There is an enormous variety of flowers, but all have some characteristics in common.
Retrieval of relevant content
Retrieval
Retrieval
Benign sample
Label: flower
Label:flower
Label:flower
Label:flower
In the field of astronomy, the sky is also called the celestial sphere.
A tree is a tall plant that can live a very long time.
The case study house on Ashfield Avenue is a modest, two-bay, two storey, mid-terrace redbrick house with a small two-storey return, built c. 1905
Retrieval failed: irrelevant results
Retrieval
Retrieval
Adversarial sample
Label: bird
Label:sunrise
Label:bigben
Label:ralong
Racism
This black nigger man looks very depressed.
Abuse
I'll make you XX, so watch your mouth.
now get out of my face, bitch!
Retrieval failed: toxic content
Retrieval
Retrieval
Adversarial sample
Label: person
Label:sex
Label: ID
Label:fight
Pornography
Privacy
Violence

Fig. 1. Example of an Attack on Deep Cross-Modal Hashing Retrieval. When a clean sample is input into a deep cross-modal hashing retrieval model, the model returns images or texts that are semantically related. However, when a carefully crafted adversarial sample is input, the model may retrieve images or texts that are semantically unrelated. More seriously, it might even return privacy-invading photos, violent images, pornographic images, or texts containing toxic remarks.

rendering transfer-based black-box attacks more practical and better suited for evaluating a model's reliability against potential adversarial attacks [18], [21].

Transfer-based black-box attacks have proven effective at defeating well-trained deep learning models in many tasks. However, achieving effective black-box attacks in hashing retrieval tasks remains challenging. Adversarial transfer attacks targeting cross-modal hashing models typically have success rates below 1% [14]. Optimization-based transfer attack methods often perform poorly due to overfitting on the deep cross-modal hashing substitute model [22]. Recent research has attempted to generate transfer attack adversarial examples using generation-based approaches [16], [22], [23]. Both optimization- and generation-based methods rely on searching for adversarial perturbations in the spatial domain. Since deep cross-modal hashing models do not encode all spatial domain information into the hash codes, these full spatial domain search methods inevitably suffer from overfitting, which reduces the transferability of attacks.

To address the aforementioned issues, we propose a frequency domain-based adversarial attack method against cross-modal hashing (FACH). Our approach is inspired by experimental observations that deep hashing models primarily encode low-frequency information (i.e., high-level semantics such as coarse object shapes, dominant colors) while largely ignoring high-frequency details. Specifically, to overcome the overfitting problem, FACH employs a frequency domain sensitivity loss to capture shared low-frequency semantic components by fusing frequency domain gradients from multiple pretrained teacher models. This process identifies the vulnerable regions of different models in the frequency domain. Second, to address the limited transferability of deep hashing adversarial attacks, we design a boundary-enhanced mechanism that incorporates distillation learning and reinforces the binary separation of hash codes, ensuring that adversarial attacks can be effectively transferred. By converting frequency domain information into adversarial weights and leveraging the trained substitute model, we generate adversarial examples that exhibit high transferability.

Our contributions are summarized as follows:

1) We propose a frequency domain-aware consensus learning module to precisely capture cross-model shared vulnerable frequency bands.
2) We propose a boundary-enhanced multi-teacher distillation method that enforces hash codes to approach quantization thresholds, enhancing cross-model attack transferability.
3) We develop a frequency-guided adversarial example generation method. To the best of our knowledge, *this is the first attempt* to study the vulnerability of deep hashing retrieval models in the frequency domain.
4) Extensive experiments validate the effectiveness of the proposed attack method. Our approach outperforms existing black-box attacks designed for deep cross-modal hashing methods. This method provides strong validation of the robustness and reliability of deep cross-modal hashing models for retrieval in security-sensitive domains.

The remainder of this paper is organized as follows: Section II reviews research on deep cross-modal hashing methods in retrieval and adversarial robustness. Section III presents the framework and details of the proposed method. Section IV describes experiments on public benchmarks and compares them with state-of-the-art methods. Finally, Section V concludes the paper.

II. RELATED WORK

A. *Deep Cross-modal Hashing Retrieval*

Deep learning-based cross-modal hashing retrieval methods have been widely adopted in multimodal data analysis for their efficiency and discriminative capabilities [2], [3], [24]. These methods are broadly divided into supervised and unsupervised approaches based on the use of semantic labels [25]. Supervised methods utilize label information to generate highly discriminative hash codes. For example, Deep Cross-modal Hashing (DCMH) [26] integrates feature and hash code learning into an end-to-end framework. Pairwise Relationship-guided Deep Hashing (PRDH) [27] adds intra-modal and inter-modal constraints for compact hash codes, enhancing bit-level discriminability through decorrelation. Graph Convolutional Hashing (GCH) [28] applies Graph Convolutional Networks (GCNs) to capture similarity structures, while Consistency-Preserving Adversarial Hashing (CPAH) [29], and Deep Adversarial Discrete Hashing (DADH) [30] leverage adversarial learning to align data distributions and maintain semantic consistency. Deep Cross-modal Unified Hashing (DCHUC) [31] introduces a non-symmetric strategy for higher-quality hash codes. Two-Stage Supervised Discrete Hashing (TSDH) [6] maps modality-specific representations to semantic binary codes with discrete optimization to minimize quantization loss.

Unsupervised methods generate discriminative hash codes without labeled data by capturing semantic relationships between heterogeneous samples. Deep Joint Semantic Reconstruction Hashing (DJSRH) [32] uses a joint semantic affinity matrix to preserve neighborhood relationships, while Unsupervised Deep Cross-modal Hashing (UDCMH) [33] integrates matrix factorization and binary latent models to reduce semantic loss. Joint Distribution Similarity-based Hashing (JDSH) [34] optimizes cross-modal code consistency using distribution-based similarity measures. Graph-based methods are also widely used. Aggregation-based Graph Convolutional Hashing (AGCH) [35] constructs an affinity matrix with multiple similarity metrics and applies an aggregation strategy to generate unified codes. Deep Graph Neighborhood Consistency Preserving Network (DGCPN) [36] explores graph neighborhood consistency across modalities. Correlation Identity Reconstruction Hashing (CIRH) [37] combines identity semantics and hash functions to improve performance. Unsupervised Contrastive Cross-modal Hashing (UCCH) [38] enhances cross-modal retrieval with a momentum optimizer and ranking loss. Contrastive Multi-bit Collaborative Learning (CMCL) [39] hierarchically aligns global and local semantics across modalities and generates multi-length hash codes for efficient cross-modal retrieval.

These methods combine the representational power of deep learning with the efficiency of hashing, providing robust solutions for multimodal retrieval tasks and achieving excellent performance across applications.

B. Adversarial Attack against Deep Hashing Retrieval

In recent years, adversarial attacks against deep cross-modal hashing retrieval (CMHR) have gradually attracted increasing attention. Existing methods can generally be categorized into two types: optimization-based adversarial attacks and generation-based transferable attacks.

Optimization-based transfer adversarial attacks generate perturbations by optimizing objective functions, enabling effective transfer. AACH [40] is the first method to generate adversarial examples by constructing substitute models under black-box conditions to mislead CMHR systems. Deep hashing targeted attack (DHTA) [12] formulates the attack on hashing retrieval as a point-to-set optimization problem and introduces a component-voting scheme to derive an anchor code that balances attack performance and perceptual imperceptibility. THA [41] leverages a PrototypeNet to generate prototype codes as semantic representatives of target labels, guides adversarial sample generation by minimizing the Hamming distance, and enhances model robustness through adversarial training. NAG [14] exploits white-box model vulnerabilities—using random noise to estimate the adversarial region and identify vulnerable pairs—to guide perturbation search for targeted black-box transfer attacks on deep hashing retrieval.

Generation-based transferable attacks typically rely on generative models to rapidly produce adversarial examples with improved generalization and to alleviate overfitting in optimization-based methods. ProS-GAN [16] proposed a generative adversarial network (GAN)-based attack framework that first generates prototype hashing codes for target categories, subsequently guiding the efficient production of adversarial examples, enhancing the attack efficiency and white-box attack performance. EQB2A [42] presented a query-driven black-box attack by constructing a counterfeit cross-modal model, avoiding frequent iterative optimization. Additionally, TA-DCH [43] extracts fine-grained target semantics, generates a target prototype code, and seamlessly embeds these into benign examples via a U-Net–based translator with adversarial training, yielding imperceptible adversarial examples.

Both Optimization-based and Generation-based transferable attacks against deep cross-modal hashing retrieval mainly focus on perturbations within the spatial domain, ignoring information representation in the frequency domain. Spatial-domain perturbations tend to capture local sensitivities, causing overfitting and reducing adversarial example transferability across different models or datasets. This paper will explore the generation of adversarial perturbations in the frequency domain to further improve the generalization and transferability of adversarial examples against deep cross-modal hashing retrieval systems.

C. Frequency-Domain Attacks and Vulnerabilities

In recent years, researchers have conducted in-depth analyses of the robustness and vulnerability of deep neural networks (DNNs) from a frequency-domain perspective, emphasizing the critical role of frequency components in model decision-making. The frequency principle (F-Principle) indicates that DNNs tend to learn low-frequency information first during training and exhibit significant sensitivity to perturbations across different frequencies [44], [45]. Studies have shown that naturally trained models are highly sensitive to all additive noise except for the lowest frequencies, while adversarial training can enhance robustness in high-frequency regions, often at the cost of degraded performance in low-frequency areas [46]. Moreover, some works suggest that adversarial examples are not solely reliant on high-frequency perturbations, as their formation is also influenced by dataset characteristics—highlighting that high-frequency edges and textures remain important in classification tasks [47]. On the other hand, frequency-based attack strategies have been proposed, including analyses of model sensitivity to both high- and low-frequency perturbations [48], [49], and approaches that generate adversarial examples by suppressing frequency-domain details, revealing that DNNs may exploit imperceptible high-frequency signals for prediction [50]. Some works enhance adversarial attack transferability by optimizing perturbations in pixel or frequency space, such as the HA-INN [51] method, which uses high-frequency perturbations for visual invisibility, and a frequency-sensitive black-box attack that improves transferability through Fourier sensitivity [52].

Existing studies have primarily focused on frequency-domain vulnerabilities in image classification tasks, and no prior work has specifically investigated the frequency-domain robustness of cross-modal hashing models.

III. Proposed Method

In this section, we first formalize the deep hashing adversarial attack problem and introduce the notation. Then, we present the overall framework.

A. Problem Formulation and Notations

1) Deep Hashing-based Cross-modal Retrieval: Given a multimodal dataset $U = \{(x_i, y_i, l_i)\}_{i=1}^N$ of size N, where $x_i \in \mathbb{R}^{q_1}$ represents the image modality sample, $t_i \in \mathbb{R}^{q_2}$ represents the text modality sample, and l_i denotes the corresponding label. The label $l_i \in \{0, 1\}^C$, where C is the number of classes. If the j-th component $l_{i,j} = 1$, it indicates that the image-text pair (x_i, y_i) belongs to class j. In a multi-label setting, l_i may have multiple components equal to 1, indicating that (x_i, y_i) belongs to multiple classes.

From a model architecture perspective, a deep cross-modal hashing model $\mathcal{F}$ comprises a hash function $\mathcal{H}$ and a sign function sign$(\cdot)$. $\mathcal{H}$ is implemented with ImgNet and TxtNet for the image and text modalities, respectively. The objective of $\mathcal{F}$ is to extract features from different modalities and project them into a unified hash space via hash mapping, enabling efficient similarity search. Specifically, hash codes for x_i or y_i are generated through the following process:

$$b_i^* = F(*_i) = sign\left(h_i^*\right), * \in \{x, y\}, \tag{1}$$

where $h_i^* = \mathcal{H}(*_i \mid \Theta_{\mathcal{H}})$, and h_i^* is the real-valued output of the hash function, which approximates the binary hash code $b_i^* \in \{-1, 1\}^K$, with K representing the length of the hash code. The parameter $\Theta_{\mathcal{H}}$ represents the trainable parameter of the hash function $\mathcal{H}$.

The semantic similarity between samples from different modalities is measured using hash distance, where greater semantic similarity results in a smaller hash distance. The hash distance is calculated as follows:

$$dist_H\left(b_i^v, b_i^t\right) = \frac{1}{2}\left(K - \left\langle b_i^v, b_i^t \right\rangle\right). \tag{2}$$

The model $\mathcal{H}$ is trained to map heterogeneous multimodal data to binary hash codes while preserving both intra-modal and inter-modal semantic similarity. Typically, the model $\mathcal{F}$ is trained using the following negative log-likelihood loss function:

$$\mathcal{L} = -\mathbb{E}_{i,j}\left(\frac{1}{2}\mathbf{S}_{i,j} \cdot \left(\mathbf{b}_i^v\right)^{\mathsf{T}}\left(\mathbf{b}_j^t\right) - log\left(1 + e^{\frac{1}{2}\left(\mathbf{b}_i^v\right)^{\mathsf{T}}\left(\mathbf{b}_j^t\right)}\right)\right), \tag{3}$$

Here, $S_{i,j} = 1$ if x_i and y_i share at least one common category label; otherwise, $S_{i,j} = 0$. During retrieval, both the query and database samples are converted into binary hash codes, and semantically similar samples are retrieved based on their hash distances.

2) Adversarial Attacks on Cross-modal Hashing Retrieval: In computer vision classification tasks, adversarial attacks aim to misclassify an adversarial sample x', which is generated from the input image x using an attack method, while ensuring that x' remains visually indistinguishable from the original image to human observers. To achieve this, the magnitude of the adversarial perturbation $\eta := x' - x$ is constrained to a level, denoted by ε, that is imperceptible to human vision, as summarized below:

$$f\left(x\right) \neq f\left(x + \eta\right), s.t.\|\eta\|_p \leq \varepsilon, \tag{4}$$

where $\|\eta\|_p = \sqrt[p]{\frac{1}{d}\left(|\eta_1|^p + |\eta_2|^p + \cdots + |\eta_d|^p\right)}$. However, in the context of cross-modal retrieval, the goal of adversarial attacks is to generate adversarial samples corresponding to original samples to fool the cross-modal retrieval model, resulting in the retrieval of partially or entirely semantically irrelevant results. Adversarial samples are generated by the attack method $\mathcal{A}$:

$$x_i' = \mathcal{A}(x_i|\Theta_{\mathcal{A}}), s.t.\|x_i' - x_i\|_p \leq \varepsilon, \tag{5}$$

where $\Theta_{\mathcal{A}}$ is the parameter of $\mathcal{A}$. The above-mentioned adversarial attack against deep hashing can be formalized as:

$$\begin{aligned}&\max_{\Theta_{\mathcal{A}}} dist_H(\mathcal{F}\left(x_i\right), \mathcal{F}\left(x_i'\right)) = \\ &\qquad \max_{\Theta_{\mathcal{A}}} dist_H(\mathcal{F}\left(x_i\right), \mathcal{F}\left(\mathcal{A}(x_i|\Theta_{\mathcal{A}})\right)), \text{s.t.}\|x_i'\| < \varepsilon,\end{aligned} \tag{6}$$

B. Overall Framework and Pipeline

As shown in Fig. 2, the proposed FACH framework is divided into two stages. Stage a) is dedicated to training a substitute model with strong transferability and comprises two modules: Frequency Sensitivity Consensus Learning and Margin-Enhanced Multi-Teacher Distillation. Stage b) focuses on generating adversarial examples and includes the Frequency Domain Adversarial Sample Generation module. In the following sections, we will provide a detailed introduction to these three modules.

C. Cross-Model Frequency Sensitivity Consensus Learning

Different models exhibit varying sensitivity to frequency domain features, which limits the transferability of attack strategies based on a single model in cross-model scenarios. To more effectively encode the frequency-domain information emphasized by different models into a substitute model, Cross-Model Frequency Sensitivity Consensus Learning is employed to enhance the cross-model transferability of attack strategies by enabling the substitute model to learn the consensus of frequency sensitivity from multiple teacher models.

Specifically, given an input image $x \in \mathbb{R}^{3\times H\times W}$, it is transformed into the frequency domain using the 2D Discrete Cosine Transform (DCT) [53]:

$$\begin{aligned}&F(u, v) = \mathrm{DCT}(x) \\ &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x(i, j) \cdot \cos\left(\frac{\pi(2i+1)u}{2H}\right) \cdot \cos\left(\frac{\pi(2j+1)v}{2W}\right),\end{aligned} \tag{7}$$

where $F(u, v)$ denotes the frequency coefficient at position (u, v), and H and W denote the height and width of the input image x, respectively. We use the Inverse Discrete Cosine Transform (IDCT) to losslessly reconstruct the image from its frequency coefficients, i.e., $x = \mathcal{D}_I(F)$, where $\mathcal{D}(\cdot)$ and $\mathcal{D}_I(\cdot)$ denote the DCT and IDCT, respectively. The application of IDCT after DCT facilitates the computation and optimization of gradients in the frequency domain. Subsequently, we compute the gradient sensitivity of different teacher models T with respect to the frequency coefficients:

$$A_T(u, v) = \frac{\partial \mathcal{L}_t(T(\mathcal{D}_I(F)))}{\partial F(u, v)} \odot M_{\mathrm{low}}(u, v), \tag{8}$$

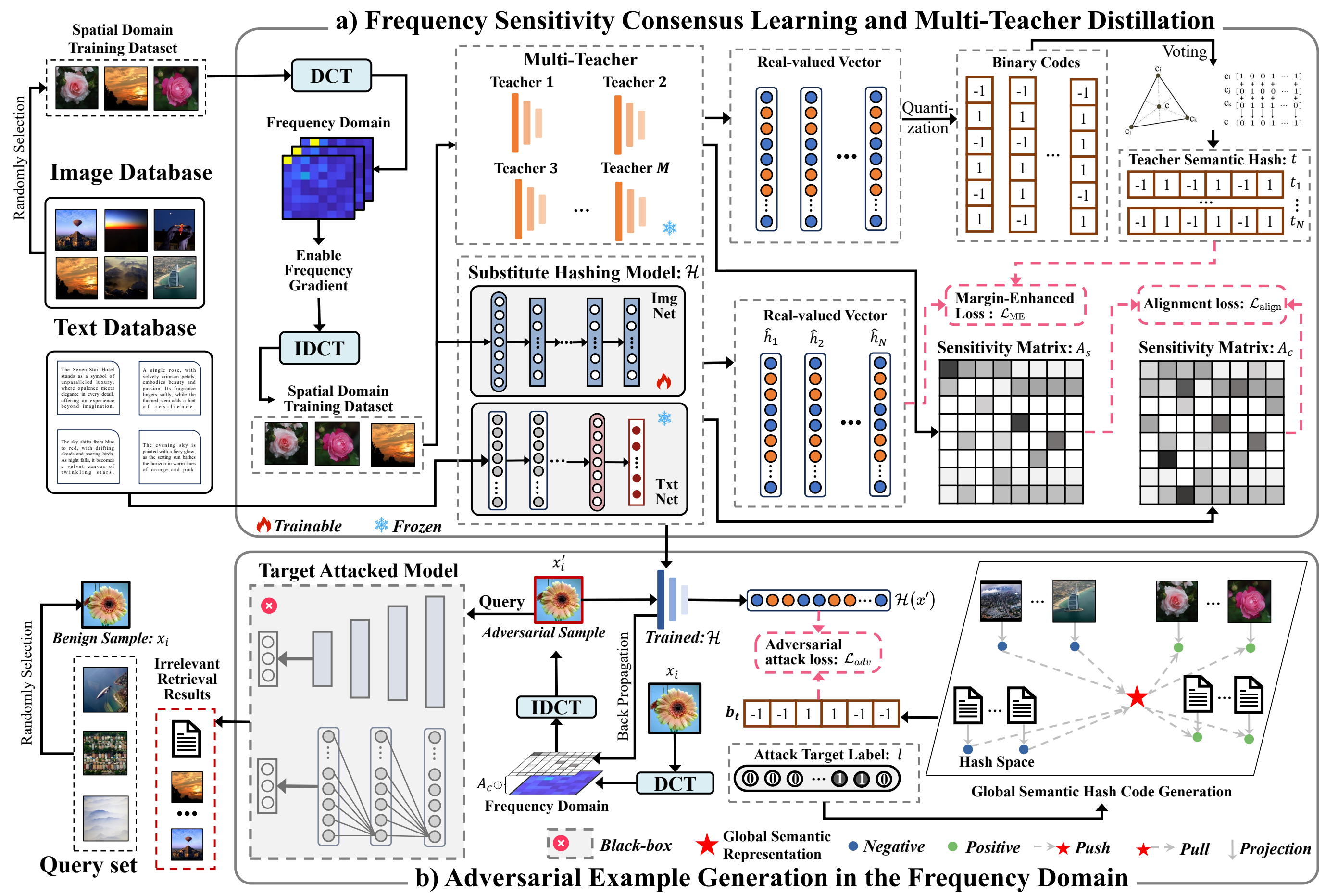

a) Frequency Sensitivity Consensus Learning and Multi-Teacher Distillation
Spatial Domain Training Dataset
Randomly Selection
Image Database
Text Database
DCT
Frequency Domain
Enable Frequency Gradient
IDCT
Spatial Domain Training Dataset
Multi-Teacher
Teacher 1
Teacher 2
Teacher 3
Teacher M
Substitute Hashing Model: $\mathcal{H}$
Img Net
Txt Net
Trainable
Frozen
Real-valued Vector
Quantization
Binary Codes
Voting
Teacher Semantic Hash: t
t_1
t_N
Real-valued Vector
$\hat{h}_1$
$\hat{h}_2$
$\hat{h}_N$
Margin-Enhanced Loss : $\mathcal{L}_{ME}$
Alignment loss: $\mathcal{L}_{align}$
Sensitivity Matrix: A_s
Sensitivity Matrix: A_c
b) Adversarial Example Generation in the Frequency Domain
Benign Sample: x_i
Randomly Selection
Query set
Irrelevant Retrieval Results
Target Attacked Model
Query
x_i'
Adversarial Sample
IDCT
A_c
Frequency Domain
DCT
x_i
Back Propagation
Trained: $\mathcal{H}$
$\mathcal{H}(x')$
Adversarial attack loss: $\mathcal{L}_{adv}$
b_t
Attack Target Label: l
Hash Space
Global Semantic Hash Code Generation
Black-box
Global Semantic Representation
Negative
Positive
Push
Pull
Projection

Fig. 2. The pipeline of our proposed FACH. Stage a): Frequency Sensitivity Consensus Learning and Multi-Teacher Distillation. The frequency-domain gradients of pre-trained teacher models are jointly analyzed to extract a low-frequency sensitivity consensus matrix A_c, while the substitute model is trained with a boundary-constrained loss to achieve semantic consistency. Stage b): Adversarial Example Generation in the Frequency Domain. Iterative optimization is employed to preferentially perturb the key frequency components indicated by A_c.

Similarly, we can compute the frequency domain sensitivity matrix A_s of the substitute model $\mathcal{H}$:

$$A_s(u,v) = \frac{\partial \mathcal{L}_t(\mathcal{H}(\mathcal{D}_I(F)))}{\partial F(u,v)} \odot M_{\text{low}}(u,v), \tag{9}$$

where, a low-frequency mask M_{low} is applied to retain low-frequency components(with $\odot$ representing the Hadamard product):

$$M_{\text{low}}(u,v) = \begin{cases} 1, & \text{if } 0 \leq u, v \leq \tau \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Here, $\mathcal{L}_t$ is the proposed frequency sensitivity function based on a deep cross-modal hashing teacher model. There are multiple possible formulations for $\mathcal{L}_t$:

$$\mathcal{L}_{t1} = \sum_{n=1}^{N} \sum_{(i,j)\in\mathcal{P}_n} \left(\|h_i^{(n)} - h_j^{(n)}\|_2^2 - \phi \right)^2$$

$$\mathcal{L}_{t2} = \sum_{n=1}^{N} \sum_{i=1}^{K} \max\left(m - \left| T(\mathcal{D}_I(F^{(n)}))_i \right|, 0 \right)$$

$$\mathcal{L}_{t3} = \sum_{n=1}^{N} \frac{1}{K} \cdot \textbf{sign}\left(T(\mathcal{D}_I(F^{(n)})) \right)^{\top} \cdot \tanh\left(T(\mathcal{D}_I(F^{(n)})) \right)$$

$$\mathcal{L}_{t4} = \sum_{n=1}^{N} \left[-\log \frac{\exp\left(\text{sim}(T(\mathcal{D}_I(F_i^{(n)})), T_m(\mathcal{D}_I(F_j^{(n)})))/\tau \right)}{\sum_{k\neq i} \exp\left(\text{sim}(T(\mathcal{D}_I(F_i^{(n)})), T(\mathcal{D}_I(F_k^{(n)})))/\tau \right)} \right], \tag{11}$$

where $\mathcal{P}$ denotes the set of positive and negative sample pairs; $h_i = \text{sign}(T_m(x_i))$ represents the binary hash code of sample x_i; ϕ is the similarity threshold; and τ is the temperature coefficient in the similarity scaling; sim refers to cosine similarity.

D. Margin-Enhanced Multi-Teacher Distillation Learning

After computing the frequency sensitivity matrices A_T from multiple teacher models, we performed a weighted averaging operation to emphasize the commonly sensitive frequency bands. This yields a consensus matrix A_c, which captures the aggregated sensitivity across all M teacher models:

$$A_c(u,v) = \frac{1}{M}\sum_{m=1}^{M} W^{\top} A_{T_m}(u,v) \tag{12}$$

Here, $W^{\top}$ is a learnable weight vector that projects each teacher's sensitivity matrix A_{T_m} into a scalar importance score.

Subsequently, we align the sensitivity matrix of the substitute model A_s with the consensus sensitivity matrix of teachers using a Jensen-Shannon(JS) Divergence [54] constraint, enforcing the substitute model to focus on the low-frequency spectral regions consistent with the ensemble of teachers. Prior to the calculation, both A_c and A_s are passed through a softmax function to convert them into valid probability distributions. We propose an alignment loss L_{align}:

$$\begin{aligned}\mathcal{L}_{\text{align}} &= \mathrm{JS}(A_c \| A_S) \\ &= \frac{1}{2}\left[\mathrm{KL}\left(A_c \middle\| \frac{A_c + A_S}{2}\right) + \mathrm{KL}\left(A_S \middle\| \frac{A_c + A_S}{2}\right)\right]\end{aligned} \tag{13}$$

Hash codes exhibit a non-trivially reversible property during quantization. For example, a slight perturbation to an image may change a real-valued hash component from 0.7 to 0.1, yet its quantized result may still be 1. This insensitivity to small changes makes it challenging to generate effective adversarial examples, and the resulting examples often have poor transferability across models. To address this, we propose a hinge-based Margin-Enhanced Loss:

$$\mathcal{L}_{\text{ME}} = \sum_{i=1}^{K} \max\left(m - \hat{h}_{s,i} \cdot t_i,\ 0\right) \tag{14}$$

Here, $\hat{h}_{s,i}$ denotes the i-th dimension of the real-valued hash vector from the substitute model, and t is the consensus target code derived from multiple teacher models. Following the voting strategy in CSQ [55], the target code is computed as:

$$t = \mathrm{sign}\left(\frac{1}{M}\sum_{j=1}^{M} h^{(j)}\right) \tag{15}$$

where $h^{(j)}$ is the hash output of the j-th teacher model, and $\mathrm{sign}(\cdot)$ denotes the element-wise sign function. This loss introduces a margin m to encourage each dimension of the real-valued hash vector $\hat{h}_S$ to surpass the target code t by a fixed threshold, improving the separability and discriminability of hash codes. As a result, adversarial examples generated from the model exhibit significantly better cross-model transferability.

Then, we obtain the overall Margin-Enhanced Multi-Teacher Distillation objective, which consists of two components:

$$\mathcal{L}_{\text{distill}} = \sum_{i=1}^{N}\left(\mathcal{L}_{\text{align}}^{(i)} + \mathcal{L}_{\text{ME}}^{(i)}\right). \tag{16}$$

E. Frequency Domain Adversarial Sample Generation

After training the substitute model $\mathcal{H}$, we generate adversarial examples based on it. Given a clean image x, we adopt the widely used Projected Gradient Descent (PGD) [56] attack to craft adversarial examples (other attack strategies are also applicable). The adversarial sample x' is optimized in iterations of T (default $T = 100$).

Before optimization is performed, we first transform the image x into the frequency domain and then reconstruct it back to the spatial domain after the necessary computations. This process allows us to leverage the information about the gradient in the spatial domain to optimize more effectively x'. Furthermore, when computing the gradient of the adversarial sample, we incorporate the sensitivity matrix of the teacher model A_c to guide the optimization process. The iterative update rule is defined as follows:

$$\begin{aligned}\Delta F_T &= \Pi_{[-\delta,\delta]}\left(\Delta F_{T-1} + \mu \cdot \mathrm{sign}\left(\nabla_{F'_{T-1}} \mathcal{L}_{\text{adv}} \odot A_c\right)\right), \\ F'_T &= F_0 + \Delta F_T, \quad F_0 = F(u,v) = \mathcal{D}(x),\end{aligned} \tag{17}$$

where ΔF_T is the cumulative adversarial perturbation in the frequency domain at iteration T, δ is the maximum allowed modification for each frequency component, ΔF_{T-1} is the accumulated perturbation from the previous iterations, μ is the step size, and the projection operator $\Pi_{[-\delta,\delta]}$ maps values outside $[-\delta,\delta]$ to the nearest boundary value within that interval. We can then derive the adversarial example in the spatial domain as follows:

$$x' = \mathrm{Clip}_{[0,1]}\left(x + \Pi_{[-\epsilon,\epsilon]}\left(\mathcal{D}_I\left(F_T\right) - x\right)\right) \tag{18}$$

This modification ensures that after reconstructing the adversarial example in the spatial domain, the pixel-wise change $x' - x$ is restricted to lie within $[-\epsilon, \epsilon]$, limiting the maximum modification per pixel to ϵ.

The adversarial loss $\mathcal{L}_{adv}$ is unified as

$$\mathcal{L}_{adv} = \frac{\gamma}{K}\boldsymbol{b}^{\top}\tanh\left(\alpha\mathcal{H}(x')\right), \tag{19}$$

where $\gamma = -1$ and $\boldsymbol{b} = \boldsymbol{b}_x$ for non-targeted attacks, and $\gamma = 1$, $\boldsymbol{b} = \boldsymbol{b}_t$ for targeted attacks.

Recent studies [57] have shown that deep hashing models exhibit similar or even identical hash centers. Motivated by this observation, we leverage the global semantic hash codes of the input instance x and the attack target x_t to enhance the cross-modal transferability of adversarial examples. Specifically, the global semantic hash codes $\boldsymbol{b}_x$ and $\boldsymbol{b}_t$ represent the hash centers of x and x_t, respectively.

To derive the global semantic hash code $\boldsymbol{b}_q$ for a given query $q \in \{x, y, x_t\}$, we employ a strategy based on weighted semantic aggregation [58], which combines information from semantically similar and dissimilar instances:

$$\boldsymbol{b}_q = \mathrm{sign}\left(\sum_{i=1}^{N_{\mathrm{p}}^{(q)}} w_i^{(q)} \boldsymbol{b}_i^{(q,\mathrm{p})} - \sum_{j=1}^{N_{\mathrm{n}}^{(q)}} w_j^{(q)} \boldsymbol{b}_j^{(q,\mathrm{n})}\right) \tag{20}$$

The weights $w_i^{(q)}$ and $w_j^{(q)}$ are designed to reflect semantic similarity between the query and its neighbors. They are

Algorithm 1 Frequency Domain Adversarial Attack Against Deep Cross-Modal Hashing

```
Input: Multi-modal training set $U_t = \{(v_i, t_i, l_i)\}_{i=1}^{N}$, $M$ teacher models $T$, untrained substitute model $\mathcal{H}$, and query dataset $U_q$.
Output: Trained substitute model $\mathcal{H}$ and adversarial examples $\{x'\}$.
Phase 1: Substitute Model Training
for each batch of images $x$ in $U_t$ do
    Transform $x$ to the frequency domain using DCT (Eq. 7).
    Compute substitute sensitivity $A_s$ (Eq. 9).
    for each teacher model in $T$ do
        Compute teacher sensitivity $A_c$ (Eq. 12).
    end for
    Compute alignment loss $L_{\text{align}}$ (Eq. 13).
    Compute margin-enhanced loss $\mathcal{L}_{\text{ME}}$ (Eq. 14).
    Compute total loss $\mathcal{L}_{\text{distill}}$ (Eq. 16).
    Update $\mathcal{H}$ by minimizing $\mathcal{L}_{\text{distill}}$.
end for
Phase 2: Adversarial Sample Generation
for each clean sample $x$ in $U_q$ do
    Transform $x$ to the frequency domain using DCT (Eq. 7).
    Compute semantic representation for $x$ (Eq. 20).
    Compute semantic representation for target $x_t$ (Eq. 20).
    Compute adversarial loss $L_{\text{adv}}$ (Eq. 19).
    Iteratively update the frequency representation (Eq. 17) to obtain $x'$.
end for
Return $\mathcal{H}$ and $\{x'\}$.
```

computed as: $w_{\bar{i}} = \frac{1}{N_{\text{p}}} \cdot s_{\bar{i}}, \quad w_{\bar{j}} = \frac{1}{N_{\text{n}}} \cdot (1 - s_{\bar{j}})$. Here, $s_{\bar{i}/\bar{j}} = \frac{\langle l, l_{\bar{i}/\bar{j}} \rangle}{\|l\| \cdot \|l_{\bar{i}/\bar{j}}\|}$ represents the cosine similarity between the label vector l of the query and the label of the corresponding positive or negative instance.

By pushing the adversarial example's hash code away from $\boldsymbol{b}_x$ (for untargeted attacks) or aligning it with $\boldsymbol{b}_t$ (for targeted attacks), the generated adversarial examples exhibit enhanced semantic alignment and better cross-modal transferability. The pseudocode of our algorithm is shown in Algorithm 1.

TABLE I
Dataset Statistics

Dataset	Total	Train	Query	Database	Classes
FLICKR-25K	20015	5000	2000	18015	24
MS-COCO	123287	1000	2000	121287	80
NUS-WIDE	195834	10500	2100	193734	21

IV. Experiments

A. Datasets and Baselines

1) Datasets: In this study, experiments are conducted on three widely used public datasets for cross-modal retrieval and adversarial attack research: FLICKR-25K [59], MS-COCO [60], and NUS-WIDE [61]. The dataset statistics are shown in Table I. For each dataset, we follow the partitioning schemes commonly adopted in cross-modal retrieval tasks [23], [62] and preprocess the data according to our research objectives.

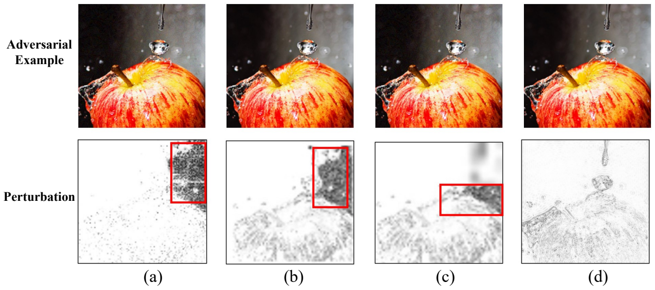


Fig. 3. A comparison of the adversarial examples and perturbations produced by four attack methods on the deep cross-modal hashing model (UCCH): (a) ProS-GAN, (b) TA-DCH, (c) PGTA, and (d) our FACH.

FLICKR-25K consists of 25,000 image-text pairs spanning 24 distinct semantic categories. After excluding samples with insufficient label information, approximately 20,000 valid pairs remain, from which 2,000 pairs are randomly selected as the query set, while the rest form the retrieval database, with a subset further extracted as the training set for model optimization.

MS-COCO includes a large number of images across 80 categories, each accompanied by 5 natural language descriptions. Following standard experimental settings, we partition 10000 samples as the training set, 2000 as the query set, and the remaining images as the complete database; for the textual modality, 1024-dimensional features are extracted using a pre-trained Bert model.

NUS-WIDE collected from Flickr, comprises a total of 269648 images covering 81 concepts. To ensure high data representativeness, we select the 21 most common categories, ultimately constructing 193734 image-text pairs. In this dataset, 2,100 pairs are randomly extracted as the query set, with the remaining samples forming the database, and an additional 10500 pairs are drawn from the database for model training.

2) Baselines: We selected six state-of-the-art cross-modal deep hash retrieval models as the target models for attack in cross-modal retrieval tasks, including three supervised models EDH(TSMC-2024) [3], CPAH(TIP-2020) [29], DADH(ICMR-2020) [30] and three unsupervised models DGCPN(AAAI-2021) [36], UCCH(TPAMI-2023) [38], JDSH(SIGIR-2020) [34]. For image retrieval tasks, we apply the CSQ method with six backbones—AlexNet [63], VGG11 [64], ResNet50 (RN50) [65], ResNet152 (RN152) [65], Inception-v3 (Inc-v3) [66], and DenseNet161 (DN161) [67]. In cross-model transfer attacks, one CSQ model with a chosen backbone serves as the substitute, while the others act as attacked models. To evaluate the effectiveness of FACH in attacking hash defense methods, we selected three representative types of defense mechanisms: SAAT (TIFS-2023) [68], which is based on adversarial training; NRCH (MM-2024) [69], which adopts input denoising; and RDPH (TMM-2024) [70], which relies on regularization.

TABLE II
Comparison of attack performance on the I2T (image-to-text) task across three datasets, evaluated using MAP and T-MAP metrics (unit: %). "Original" denotes the MAP without adversarial attack, while other entries represent T-MAP under attack. The perturbation upper bound ϵ in our method is set to $8/255$

Attacked models	Attack Methods	LPIPS	MIRFLICKR25k				NUS-WIDE				MS-COCO			
			16	32	64	128	16	32	64	128	16	32	64	128
EDH [3]	Original	0	72.14	74.21	75.69	76.32	63.22	63.98	65.78	66.46	65.85	66.51	68.22	68.65
	DHTA [41]	0.332	82.36	83.69	89.53	91.25	74.65	75.98	77.85	78.91	72.58	73.21	75.89	76.48
	ProS-GAN [16]	0.263	83.36	84.69	89.68	90.36	76.24	77.48	78.36	79.65	75.85	76.54	77.58	78.36
	TA-DCH [43]	0.215	85.47	86.24	88.97	91.36	78.58	76.36	83.25	84.56	76.69	74.89	77.58	77.21
	PGTA [23]	0.196	94.36	93.96	94.55	95.63	88.69	85.69	91.25	92.35	78.52	79.85	82.56	82.47
	FACH(ours)	**0.158**	**95.66**	**95.98**	**96.85**	**96.47**	**90.69**	**91.25**	**92.36**	**93.58**	**82.35**	**83.36**	**85.68**	**88.47**
CPAH [29]	Original	0	57.14	57.86	60.21	66.74	45.21	42.36	44.69	45.12	39.63	40.87	40.63	41.17
	DHTA [41]	0.365	78.63	79.32	86.14	88.01	71.35	75.63	78.21	79.94	63.24	64.93	69.21	70.16
	ProS-GAN [16]	0.214	77.46	84.28	86.27	86.92	68.39	75.69	79.14	83.56	63.41	66.29	66.37	68.87
	TA-DCH [43]	0.286	80.28	86.02	90.42	90.81	73.52	80.25	81.25	84.67	68.21	71.23	72.47	73.62
	PGTA [23]	0.312	93.38	94.64	94.31	94.55	84.25	86.25	87.21	90.35	70.86	**73.94**	75.98	74.54
	FACH(ours)	**0.109**	**95.23**	**95.63**	**96.21**	**95.14**	**87.25**	**88.12**	**90.21**	**91.36**	**73.21**	73.89	**76.58**	**77.54**
DADH [30]	Original	0	61.28	61.47	62.74	65.47	49.45	50.45	49.51	49.87	39.65	41.32	45.23	48.93
	DHTA [41]	0.307	84.69	85.14	87.24	88.25	74.65	76.77	80.78	82.67	57.13	61.45	65.02	70.54
	ProS-GAN [16]	0.256	83.69	86.78	87.36	88.25	70.23	76.32	78.34	82.19	57.98	63.52	68.47	73.21
	TA-DCH [43]	0.278	85.74	86.47	87.85	89.36	75.34	81.29	83.97	86.64	61.34	63.87	68.65	70.12
	PGTA [23]	0.251	91.25	92.36	92.98	92.63	76.23	79.23	85.34	88.87	65.78	69.76	73.86	77.21
	FACH(ours)	0.178	**93.25**	**94.25**	**95.85**	**95.87**	**79.36**	**83.55**	**88.58**	**92.47**	**71.25**	**72.54**	**77.58**	**78.69**
DGCPN [36]	Original	0	61.36	62.14	61.36	63.21	42.36	41.96	43.32	42.39	41.63	40.58	39.67	41.87
	DHTA [41]	0.286	77.63	78.47	82.36	83.87	73.21	77.74	80.25	82.36	58.47	61.27	68.78	69.87
	ProS-GAN [16]	0.254	78.36	79.63	81.67	83.65	72.36	76.64	78.21	80.72	56.87	63.85	69.74	69.75
	TA-DCH [43]	0.284	83.25	84.36	84.66	86.36	74.36	75.69	78.36	82.43	62.36	67.96	72.98	75.24
	PGTA [23]	0.362	90.25	92.36	91.66	92.45	87.12	87.35	91.25	93.25	69.37	72.14	72.48	76.48
	FACH(ours)	**0.125**	**93.36**	**94.25**	**96.69**	**96.87**	**90.58**	**91.28**	**92.81**	**93.68**	**74.36**	**75.69**	**80.85**	**78.69**
UCCH [38]	Original	0	75.21	76.21	74.21	76.32	68.23	70.32	72.45	73.23	60.56	63.21	64.44	65.76
	DHTA [41]	0.325	87.95	85.36	88.74	89.45	74.63	75.96	76.98	77.36	65.36	66.74	67.25	69.78
	ProS-GAN [16]	0.258	88.59	88.54	89.99	91.02	73.36	74.69	76.36	77.09	66.36	67.89	69.36	70.54
	TA-DCH [43]	0.269	92.36	93.69	92.58	92.17	75.36	74.69	76.69	77.39	64.25	63.69	66.58	67.21
	PGTA [23]	0.369	93.36	94.56	95.69	95.87	77.25	76.69	78.69	79.74	70.21	70.99	73.25	74.65
	FACH(ours)	**0.213**	**95.36**	**96.36**	**97.25**	**97.08**	**87.36**	**88.69**	**91.07**	**90.62**	**74.66**	**76.69**	**79.63**	**78.54**
JDSH [34]	Original	0	60.84	62.98	62.67	63.04	40.21	40.11	41.43	42.03	43.45	42.56	43.97	44.21
	DHTA [41]	0.215	73.08	74.87	78.93	79.21	67.23	68.56	74.98	76.32	62.45	63.11	64.34	65.21
	ProS-GAN [16]	0.223	74.03	76.23	79.45	78.15	66.23	68.32	74.93	76.12	55.34	56.77	63.45	65.43
	TA-DCH [43]	0.245	73.94	75.09	79.54	78.65	73.12	76.23	77.43	78.32	63.21	64.07	67.24	68.28
	PGTA [23]	0.212	78.65	**83.98**	88.65	89.34	82.45	84,95	85.11	86.23	66.66	**71.49**	73.21	74.87
	FACH(ours)	**0.156**	**82.36**	83.69	**89.63**	**91.58**	**85.36**	**86.54**	**88.64**	**90.85**	**67.69**	68.98	**74.65**	**76.98**

TABLE III
The test t-MAP (%) of targeted attacks against the CSQ method with 32 bits and different backbones on the FLICKR-25K and NUS-WIDE datasets, respectively

	Methods	LPIPS	FLICKR-25K						NUS-WIDE					
			AlexNet	VGG11	RN50	RN152	Inc-v3	DN161	AlexNet	VGG11	RN50	RN152	Inc-v3	DN161
VGG11	P2P	0.263	60.81	82.93	67.63	62.15	69.13	69.03	53.24	78.57	59.27	52.99	57.16	59.82
	DHTA	0.296	67.68	84.41	66.49	66.40	67.22	64.09	60.73	74.51	59.21	56.03	55.85	60.68
	THA	0.247	72.51	90.38	78.60	72.07	72.87	76.11	54.55	81.19	58.71	56.68	64.92	55.60
	ProS-GAN	0.219	74.18	92.83	80.49	74.54	77.73	83.92	65.58	81.76	68.99	65.63	66.76	85.66
	TTA-GAN	0.321	74.89	89.20	88.53	85.56	81.93	84.98	64.28	81.57	68.60	71.46	69.56	70.97
	FACH(Ours)	**0.103**	**80.21**	**96.36**	**90.36**	**91.36**	**86.69**	**90.48**	**72.35**	**88.36**	**75.36**	**78.65**	**72.69**	**88.37**
RN50	P2P	0.287	68.20	62.88	91.59	62.99	62.14	65.59	54.74	56.37	74.31	57.37	60.19	59.22
	DHTA	0.296	65.05	62.06	89.34	63.17	68.21	72.62	54.29	54.12	78.41	58.96	53.65	60.46
	THA	0.304	74.91	74.11	92.21	78.83	69.59	81.25	58.23	55.73	80.30	60.55	65.26	62.92
	ProS-GAN	0.324	73.40	76.08	89.28	88.54	74.53	81.68	57.79	63.17	84.22	61.89	59.48	65.68
	TTA-GAN	0.333	76.06	88.64	93.04	89.32	86.59	85.05	67.95	72.33	82.26	78.52	77.26	72.11
	FACH(Ours)	**0.231**	**82.63**	**92.36**	**95.36**	**92.74**	**91.87**	**87.35**	**72.69**	**78.47**	**88.96**	**82.63**	**82.69**	**78.36**
DN161	P2P	0.236	60.16	61.29	69.30	69.99	66.64	89.41	54.14	56.84	53.53	56.42	60.99	77.06
	DHTA	0.269	63.23	66.08	64.02	65.11	66.71	91.52	55.60	60.93	59.69	54.66	56.51	79.87
	THA	0.312	70.38	72.30	80.63	78.19	74.11	95.69	62.57	56.71	57.48	62.77	64.19	85.29
	ProS-GAN	0.325	69.94	81.56	82.96	77.99	69.20	92.63	65.13	64.84	63.52	61.75	64.90	81.85
	TTA-GAN	0.315	79.54	86.61	91.04	92.52	83.17	94.16	65.82	72.51	72.32	76.79	75.04	78.37
	FACH(Ours)	**0.214**	**86.36**	**92.49**	**96.39**	**93.36**	**89.32**	**95.99**	**71.36**	**78.65**	**80.36**	**82.69**	**83.36**	**91.36**

B. Evaluation Metrics and Implementation Details

1) Evaluation Metrics: Following previous works [12], [23], we adopt two primary metrics to evaluate the effectiveness of the attack methods: mean average precision (mAP [26]) and targeted mean average precision (t-mAP [12]). The mAP assesses the overall performance of the retrieval system, while the t-mAP uses target labels specified by the attacker to measure the effectiveness of the targeted attack. The mAP is defined as

$$\begin{aligned} \mathrm{mAP} &= \frac{1}{Q}\sum_{q=1}^{Q}\mathrm{AP}_q, \\ \mathrm{AP}_q &= \frac{1}{n_q}\sum_{k=1}^{N}P_q(k)\cdot\mathbb{I}\big(r_q(k)\big). \end{aligned} \tag{21}$$

where Q is the total number of query instances, n_q is the number of relevant samples for the q-th query, N represents the total number of samples in the database, $P_q(k)$ denotes the precision of the top k retrieval results for the q-th query, and $\mathbb{I}(r_q(k))$ is an indicator function that equals 1 if the k-th result is relevant to the q-th query, and 0 otherwise. The t-mAP is calculated in a similar manner, except that the original labels are replaced with the target labels during evaluation.

2) Perceptibility: We adopt dual perceptibility constraints to ensure that adversarial examples remain imperceptible. In the spatial domain, an L_∞-norm constraint limits the maximum per-pixel difference between the original image x and its adversarial version x':

$$\|x' - x\|_\infty \leq \epsilon, \tag{22}$$

where ϵ controls pixel-level imperceptibility. Although the perturbation is generated in the frequency domain, it is transformed back to the spatial domain via IDCT, and its magnitude remains bounded.

We also use the LPIPS [71] metric to assess perceptual similarity between clean and adversarial samples, where lower scores indicate less noticeable differences.

3) Implementation Details: In the experiments, we implemented the proposed FACH model using PyTorch and trained it on an RTX4090 GPU with 64 GB RAM. The Adaw optimizer was used with a learning rate of 10^{-4}. ImgNet adopts VGG11 as the backbone with two additional fully connected layers. TxtNet is a dense network with three fully connected layers. The training runs for 20 epochs with a batch size of 64. The adversarial generation stage is set to $T = 100$ iterations.

TxtNet is first pre-trained using the loss function in Equation (5), then frozen before distilling ImgNet. Since this work focuses on the robustness of images in the frequency domain, adversarial samples are generated only for images, not for text.

We followed six baseline methods (EDH, CPAH, DADH, DGCPN, UCCH, JDSH), each combined with six different backbone networks (AlexNet, VGG11, RN50, RN152, Inc-v3, DN161), resulting in a total of 36 teacher models.

The hyperparameters are set as follows: $\phi \to 0$ for positive pairs, $\phi = 2\sqrt{K}$ for negative pairs; $m = 1.5$, $\mu = 0.001$, and δ to 0.3. The value of α follows SAAT [58]: $\alpha = 0.1$ for the first 50 iterations; in the next 50 iterations, it is set to 0.2, 0.3, 0.5, 0.7, and 1 every 10 iterations.

C. Comparison with SOTA attack methods

We conduct a systematic comparison of state-of-the-art cross-modal adversarial attack methods, including ProS-GAN [16], DHTA [41], TA-DCH [43], and PGTA [23], on cross-modal hashing retrieval (CMHR) tasks. Attacks are performed under 16, 32, 64, and 128-bit hash code settings, with results shown in Table II. Experimental results show that our proposed FACH achieves superior targeted attack performance, consistently outperforming existing methods in both t-MAP and perceptual distortion (LPIPS) across all datasets. DHTA and ProS-GAN are originally designed for unimodal retrieval systems, and their performance degrades significantly when directly applied to cross-modal scenarios. While TA-DCH and PGTA leverage spatial-domain semantic information to guide adversarial generation, they fail to capture critical frequency-domain components, often resulting in spatial overfitting. Our FACH models adversarial perturbations in the frequency domain, effectively capturing shared sensitivities across different backbones and enabling more precise and efficient perturbation generation, while avoiding the redundant computation and overfitting risks associated with spatial-domain attacks. Notably, as shown in Fig. 3 (for visualization, we take the absolute values of the perturbations and scale them by a factor of 25), the perturbations generated by methods such as ProS-GAN, TA-DCH, and PGTA tend to be randomly and diffusely distributed, failing to concentrate on the key semantic regions of the image's main subject. In contrast, our FACH precisely targets the contours and skeletal structures of the subject while maintaining a low perturbation cost, thereby enhancing both the stealth and the directional effectiveness of the attack.

In image retrieval tasks, we evaluate several representative hashing attack methods, including P2P [41], DHTA [41], THA [72], TTA-GAN [22], and ProS-GAN [16]. Experiments are conducted on target models with various backbones—AlexNet, VGG11, RN50, RN152, Inc-v3, and DN161—while surrogate models adopt different backbone trained under the CSQ method. The results show that iterative methods such as P2P, DHTA, and THA suffer from overfitting and underperforming in all settings. GAN-based methods (TTA-GAN and ProS-GAN) offer better performance but exhibit instability and poor generalization across backbones. Our proposed FACH learns perturbation patterns across diverse attack strategies and backbones, consistently achieving high and stable attack success rates across all surrogate-target pairs, which highlights its strong cross-model transferability.

The results of attacks against three representative hash defense mechanisms, as summarized in Table IV, demonstrate that none of the defense methods—SAAT, NRCH, or RDPH—can effectively withstand attacks from FACH. Notably, SAAT achieves better defense performance compared to NRCH and RDPH, owing to its adversarial training strategy leveraging adversarial examples. Meanwhile, NRCH, which employs input denoising, attains slightly better robustness than RDPH, which is based on regularization techniques.

TABLE IV
ATTACK PERFORMANCE ON THREE TYPES OF ROBUST DEEP CROSS-MODAL HASHING METHODS

Method		MIRFLICKR25k				NUS-WIDE			
		16	32	64	128	16	32	64	128
SAAT [68]	Original	72.52	72.69	71.24	71.98	60.85	62.39	64.25	63.14
	Attacked	82.52	83.57	83.01	82.69	58.64	59.87	60.85	62.36
RDPH [70]	Original	68.39	67.35	67.58	68.36	58.35	58.98	57.84	56.78
	Attacked	82.54	82.94	82.54	83.21	65.36	66.14	66.87	66.74
NRCH [69]	Original	74.58	75.65	77.21	76.36	65.41	63.35	64.35	65.14
	Attacked	85.32	84.35	83.21	83.69	70.14	71.25	72.35	73.64

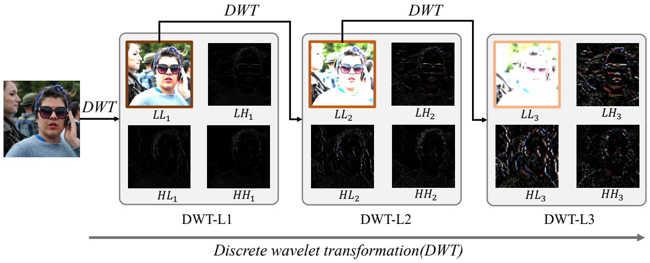
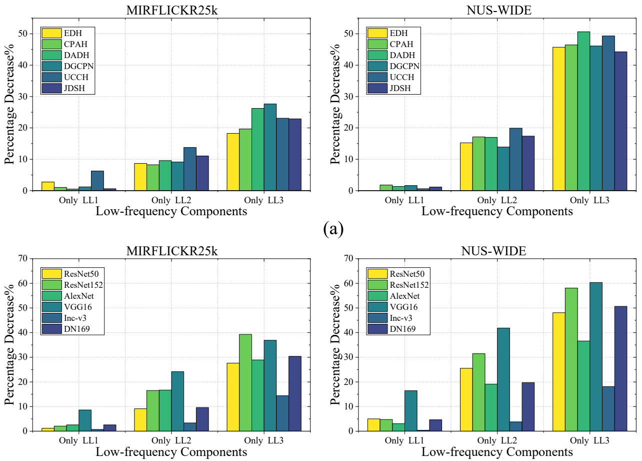


Fig. 4. Visualization of Multi-Level Sub-band Decomposition After Three Successive Discrete Wavelet Transforms on the Image. The latter two transforms are applied to the low-frequency sub-bands, yielding four distinct frequency components at each level.



(a)

(b)

Fig. 5. Comparison of retrieval performance degradation percentages using different low-frequency components (bar chart). The y-axis shows the percentage decrease in retrieval performance compared to using original images. (a) Performance degradation at three levels of low-frequency components across various cross-modal retrieval models. (b) Performance degradation at three levels of low-frequency components across different backbone networks using the UCCH method.

D. Ablation Studies

To investigate the impact of frequency-domain information on hash code learning and semantic similarity computation, ablation experiments are conducted using different levels of low-frequency signals. A single-level Discrete Wavelet Transform (DWT) decomposes an image into four sub-bands: LL, LH, HL, and HH. The LL band contains low-frequency components that represent the global outline and approximate content of the image. The LH, HL, and HH bands contain high-frequency details. The LL band obtained from the first DWT level can be further decomposed using DWT to produce second-level sub-bands: LL2, HL2, LH2, and HH2. Applying DWT three times yields three levels of low-frequency signals: LL1, LL2, and LL3. Fig. 4 illustrates how these low-frequency sub-bands preserve the image's approximate content at different scales. Three experimental groups are constructed using images that contain only LL1, LL2, or LL3. All high-frequency components are removed. These low-frequency-only images are used as input to deep cross-modal retrieval models. Retrieval performance decreases when high-frequency information is removed. Fig. 5 shows that the extent of performance degradation varies across different models and backbone architectures. Some models are more sensitive to shallow-level low-frequency features, while others rely more on deeper low-frequency representations. The differences among backbone architectures cause larger performance variations than those among model structures. The backbone plays a key role in encoding low-frequency features. Different backbones show different capacities in capturing and utilizing frequency-domain information.

TABLE V
THE T-MAP RESULTS OF VARIOUS ABLATION SETTINGS ON CROSS-MODAL RETRIEVAL TASKS

Attacked methods		CPAH				UCCH			
bit		16	32	64	128	16	32	64	128
FLICKR-25K	w/o $\mathcal{T}$	83.25	84.69	86.69	87.74	89.69	91.22	93.68	92.87
	w/o $\mathcal{G}$	89.63	90.58	92.36	91.89	92.37	93.49	94.22	94.58
	FACH	**95.23**	**95.63**	**96.21**	**95.14**	**95.36**	**96.36**	**96.25**	**97.08**
MS-COCO	w/o $\mathcal{T}$	66.63	67.94	69.87	70.49	66.36	66.74	69.14	68.47
	w/o $\mathcal{G}$	68.63	68.78	70.25	70.85	71.35	73.58	76.36	75.97
	FACH	**70.86**	**73.94**	**75.98**	**74.54**	**74.66**	**76.69**	**79.63**	**78.54**
NUS-WIDE	w/o $\mathcal{T}$	67.54	68.39	70.13	71.59	70.69	71.69	73.58	76.98
	w/o $\mathcal{G}$	78.36	79.63	83.65	86.17	82.69	83.45	85.87	87.48
	FACH	**87.25**	**88.12**	**90.21**	**91.36**	**87.36**	**88.69**	**91.07**	**90.62**

TABLE VI
T-MAP OF DIFFERENT SENSITIVITY LOSS FUNCTIONS

Attacked methods		FLICKR-25K				NUS-WIDE			
bit		16	32	64	128	16	32	64	128
CPAH	$\mathcal{L}_{t1}$	80.39	81.36	83.69	84.52	77.96	78.54	81.24	82.59
	$\mathcal{L}_{t2}$	**95.23**	**95.63**	**96.21**	95.02	**87.25**	**88.12**	**90.21**	**91.36**
	$\mathcal{L}_{t3}$	92.33	92.65	93.54	**95.14**	80.23	81.67	82.56	81.57
	$\mathcal{L}_{t4}$	87.69	88.52	89.47	90.64	82.69	83.45	84.78	85.74
DADH	$\mathcal{L}_{t1}$	93.13	**94.25**	**96.69**	95.74	75.65	76.41	77.21	78.28
	$\mathcal{L}_{t2}$	91.02	92.21	93.75	**95.87**	**79.36**	82.36	**88.58**	**92.47**
	$\mathcal{L}_{t3}$	**93.25**	93.66	94.78	94.98	78.36	**83.55**	85.36	86.47
	$\mathcal{L}_{t4}$	84.21	84.36	85.64	86.14	70.54	71.21	72.47	73.54
DGCPN	$\mathcal{L}_{t1}$	82.36	81.69	83.57	84.26	86.32	87.14	87.69	88.48
	$\mathcal{L}_{t2}$	88.36	89.57	90.36	89.48	86.69	87.41	88.57	89.47
	$\mathcal{L}_{t3}$	85.45	86.36	87.41	88.25	87.45	87.14	88.36	89.14
	$\mathcal{L}_{t4}$	**96.36**	**94.25**	**96.69**	**96.87**	**90.58**	**91.28**	**92.81**	**93.68**
UCCH	$\mathcal{L}_{t1}$	86.36	87.24	88.63	87.24	78.24	79.65	80.24	79.63
	$\mathcal{L}_{t2}$	90.25	90.85	92.36	93.65	80.08	79.65	81.69	82.47
	$\mathcal{L}_{t3}$	88.63	89.36	90.47	91.25	77.21	75.51	78.48	79.21
	$\mathcal{L}_{t4}$	**95.36**	**96.36**	**96.25**	**97.08**	**87.36**	**88.69**	**91.07**	**90.62**

To analyze the roles of the two key components in the FACH framework, we conducted ablation studies on the cross-modal retrieval task using two representative retrieval models (CPAH and UCCH) and three public datasets (FLICKR-25K, MS-COCO, and NUS-WIDE). Specifically, we designed two ablated variants: one without the surrogate model training module (w/o $\mathcal{T}$), and the other without the adversarial sample generation module (w/o $\mathcal{G}$). As shown in Table V, the complete FACH method consistently achieves the highest T-MAP scores

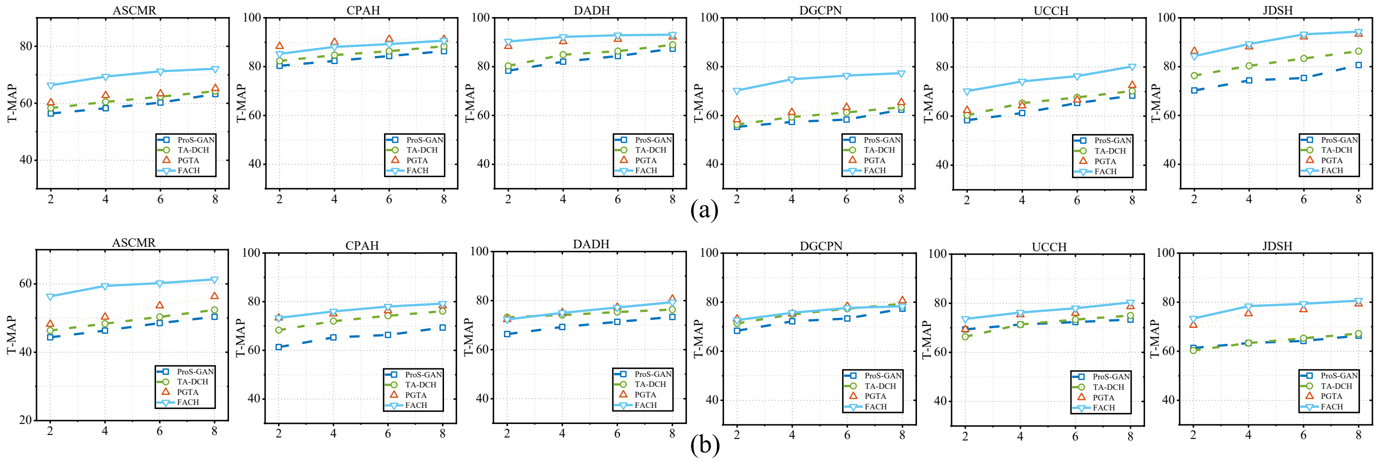


Fig. 6. T-MAP performance with Varied ϵ. (a)MS-COCO,(b)NUS-WIDE.

across all hash code lengths. Removing the training module ($\mathcal{T}$) leads to a significant performance drop, while removing the generation module ($\mathcal{G}$) also results in noticeable degradation. The surrogate model training module ($\mathcal{T}$) helps to comprehensively model the cross-modal feature space, whereas the adversarial sample generation module ($\mathcal{G}$) can produce visually imperceptible but highly effective perturbations. The synergy between the two components enables the complete FACH framework to outperform all ablated variants in terms of both robustness and attack effectiveness.

E. Hyperparameter Analysis

As shown in Table VI, significant differences exist in how different sensitivity loss functions affect attack performance. For the CPAH and DADH methods, the optimal loss function varies with the dataset and hash code length, indicating that the selection of sensitivity loss functions should be tailored according to the specific method characteristics and data distribution. For the DGCPN and UCCH methods, the best-performing loss function is $\mathcal{L}_{t4}$. This can be attributed to the fact that both DGCPN and UCCH are unsupervised methods, and $\mathcal{L}_{t4}$ is derived from an unsupervised contrastive learning framework, sharing structural similarity with their training objectives. This insight suggests that the sensitivity loss function should ideally align closely with the original model's training objective to enhance attack effectiveness and compatibility.

As illustrated in Fig. 6, we further investigate the impact of perturbation magnitude ϵ on attack performance. The results indicate that increasing ϵ consistently improves attack performance, primarily because larger perturbations are more effective at overcoming model robustness defenses, leading to greater deviations in the hash outputs. However, excessive perturbation magnitude may reduce imperceptibility, so it is necessary to balance attack success rate and stealth by appropriately controlling the perturbation strength.

V. Conclusion

In this paper, we propose a novel deep cross-modal hashing adversarial attack method named FACH. Unlike existing approaches that generate adversarial examples in the spatial domain, FACH operates in the frequency domain. It comprises two stages: First, by leveraging a multi-teacher model, we identify sensitive frequency regions to mitigate spatial overfitting and enhance hash code discriminability through boundary augmentation, thereby training a substitute model with strong transferability. Second, using the trained substitute model and learned frequency-sensitive information, we generate perturbations in the frequency domain and map them to the spatial domain to create adversarial examples with superior transferability. Experiments on three datasets demonstrate that our method outperforms state-of-the-art approaches in transfer attack performance. Future work will focus on adaptive frequency band selection for multimodal scenarios.

Acknowledgments

We thank all the reviewers for their careful review and valuable comments.

References

[1] Y. Sun, Z. Ren, P. Hu, D. Peng, and X. Wang, "Hierarchical consensus hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 824–836, 2024.

[2] J. Huang, P. Kang, N. Han, Y. Chen, X. Fang, H. Gao, and G. Zhou, "Two-stage asymmetric similarity preserving hashing for cross-modal retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 1, pp. 429–444, 2024.

[3] J. Huang, P. Kang, X. Fang, N. Han, S. Xie, and H. Gao, "Efficient discriminative hashing for cross-modal retrieval," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 6, pp. 3865–3878, 2024.

[4] X. Liang, E. Yang, Y. Yang, and C. Deng, "Multi-relational deep hashing for cross-modal search," *IEEE Transactions on Image Processing*, vol. 33, pp. 3009–3020, 2024.

[5] 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.

[6] D. Zhang, X.-J. Wu, T. Xu, and J. Kittler, "Two-stage supervised discrete hashing for cross-modal retrieval," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 11, pp. 7014–7026, 2022.

[7] Q. Wu, Z. Zhang, Y. Liu, J. Zhang, and L. Nie, "Contrastive multi-bit collaborative learning for deep cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5835–5848, 2024.

[8] Z. Hu, Y.-M. Cheung, M. Li, and W. Lan, "Cross-modal hashing method with properties of hamming space: A new perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7636–7650, 2024.

[9] Y. Ji, X. Zhang, G. Zhou, X. Zheng, and D. D. Zeng, "Learning hash subspace from large-scale multi-modal pre-training: A clip-based cross-modal hashing framework," in *China Conference on Command and Control*. Springer, 2023, pp. 514–526.

[10] C. LI, S. Gao, C. Deng, D. Xie, and W. Liu, "Cross-modal learning with adversarial samples," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/d384dec9f5f7a64a36b5c8f03b8a6d92-Paper.pdf

[11] 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 (ICCV)*, October 2021, pp. 2218–2227.

[12] 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*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 618–634.

[13] Y. Ji, Y. Liu, Z. Zhang, Z. Zhang, Y. Zhao, G. Zhou, X. Zhang, X. Liu, and X. Zheng, "Advlora: Adversarial low-rank adaptation of vision-language models," *arXiv preprint arXiv:2404.13425*, 2024.

[14] Y. Xiao and C. Wang, "You see what i want you to see: Exploring targeted black-box transferability attack for hash-based image retrieval systems," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 1934–1943.

[15] S. Hu, Y. Zhang, X. Liu, L. Y. Zhang, M. Li, and H. Jin, "Advhash: Set-to-set targeted attack on deep hashing with one single adversarial patch," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2335–2343. [Online]. Available: https://doi.org/10.1145/3474085.3475396

[16] X. Wang, Z. Zhang, B. Wu, F. Shen, and G. Lu, "Prototype-supervised adversarial network for targeted attack of deep hashing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 357–16 366.

[17] X. Zhang, X. Zheng, W. Mao, D. D. Zeng, and F.-Y. Wang, "Hashing fake: Producing adversarial perturbation for online privacy protection against automatic retrieval models," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 3241–3251, 2023.

[18] X. Zhang, X. Zheng, and W. Mao, "Adversarial perturbation defense on deep neural networks," *ACM Comput. Surv.*, vol. 54, no. 8, Oct. 2021. [Online]. Available: https://doi.org/10.1145/3465397

[19] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Comput. Surv.*, vol. 53, no. 3, Jun. 2020. [Online]. Available: https://doi.org/10.1145/3398394

[20] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Comput. Surv.*, vol. 54, no. 11s, Sep. 2022. [Online]. Available: https://doi.org/10.1145/3523273

[21] W. B. *, J. R. *, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=SyZI0GWCZ

[22] F. Zhu, W. Zhang, D. Wu, L. Wang, B. Li, and W. Wang, "Targeted transferable attack against deep hashing retrieval," in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, ser. MMAsia '23. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3595916.3626420

[23] X. Guo, H. Zhang, L. Liu, D. Liu, X. Lu, and H. Meng, "Primary code guided targeted attack against cross-modal hashing retrieval," *IEEE Transactions on Multimedia*, vol. 27, pp. 312–326, 2025.

[24] C. Bai, C. Zeng, Q. Ma, and J. Zhang, "Graph convolutional network discrete hashing for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 4756–4767, 2024.

[25] 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, no. 1, pp. 239–260, 2024.

[26] 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.

[27] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10719

[28] R. Xu, C. Li, J. Yan, C. Deng, and X. Liu, "Graph convolutional network hashing for cross-modal retrieval," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, ser. IJCAI'19. AAAI Press, 2019, p. 982–988.

[29] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 3626–3637, 2020.

[30] 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*, ser. ICMR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 525–531. [Online]. Available: https://doi.org/10.1145/3372278.3390711

[31] R.-C. Tu, X.-L. Mao, B. Ma, Y. Hu, T. Yan, W. Wei, and H. Huang, "Deep cross-modal hashing with hashing functions and unified hash codes jointly learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 2, pp. 560–572, 2022.

[32] 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 (ICCV)*, October 2019.

[33] G. Wu, Z. Lin, J. Han, L. Liu, G. Ding, B. Zhang, and J. Shen, "Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, p. 2854–2860.

[34] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1379–1388. [Online]. Available: https://doi.org/10.1145/3397271.3401086

[35] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 466–479, 2022.

[36] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, pp. 4626–4634, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16592

[37] L. Zhu, X. Wu, J. Li, Z. Zhang, W. Guan, and H. T. Shen, "Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8838–8851, 2023.

[38] 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.

[39] Q. Wu, Z. Zhang, Y. Liu, J. Zhang, and L. Nie, "Contrastive multi-bit collaborative learning for deep cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5835–5848, 2024.

[40] 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.

[41] X. Wang, Z. Zhang, G. Lu, and Y. Xu, "Targeted attack and defense for deep hashing," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2298–2302. [Online]. Available: https://doi.org/10.1145/3404835.3463233

[42] 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 Trans. Inf. Syst.*, vol. 41, no. 3, Feb. 2023. [Online]. Available: https://doi.org/10.1145/3559758

[43] T. Wang, L. Zhu, Z. Zhang, H. Zhang, and J. Han, "Targeted adversarial attack against deep cross-modal hashing retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 6159–6172, 2023.

[44] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, "Training behavior of deep neural network in frequency domain," in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2019, p. 264–274. [Online]. Available: https://doi.org/10.1007/978-3-030-36708-4_22

[45] T. Luo, Z. Ma, Z.-Q. J. Xu, and Y. Zhang, "Theory of the frequency principle for general deep neural networks," *arXiv preprint arXiv:1906.09235*, 2019.

[46] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[47] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, "Frequency-driven imperceptible adversarial attack on semantic similarity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 315–15 324.

[48] Y. Tsuzuku and I. Sato, "On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 51–60.

[49] A. A. Abello, R. Hirata, and Z. Wang, "Dissecting the high-frequency bias in convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 863–871.

[50] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, "Advdrop: Adversarial attack to dnns by dropping information," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7506–7515.

[51] M. Shao, J. Yang, L. Meng, and Z. Hu, "Frequency attacks based on invertible neural networks," *IEEE Transactions on Artificial Intelligence*, 2024.

[52] Y. Liu, C. Li, Z. Wang, H. Wu, and X. Zhang, "Transferable adversarial attack based on sensitive perturbation analysis in frequency domain," *Information Sciences*, vol. 678, p. 120971, 2024.

[53] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974.

[54] M. Menéndez, J. Pardo, L. Pardo, and M. Pardo, "The jensen-shannon divergence," *Journal of the Franklin Institute*, vol. 334, no. 2, pp. 307–318, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016003296000634

[55] 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*, 2020, pp. 3083–3092.

[56] 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.

[57] L. Tang, D. Ye, Y. Lv, C. Chen, and Y. Zhang, "Once and for all: Universal transferable adversarial perturbation against deep hashing-based facial image retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5136–5144, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28319

[58] X. Yuan, Z. Zhang, X. Wang, and L. Wu, "Semantic-aware adversarial training for reliable deep hashing retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4681–4694, 2023.

[59] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ser. MIR '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 39–43. [Online]. Available: https://doi.org/10.1145/1460096.1460104

[60] 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.

[61] 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.

[62] 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.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[66] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[67] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[68] X. Yuan, Z. Zhang, X. Wang, and L. Wu, "Semantic-aware adversarial training for reliable deep hashing retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4681–4694, 2023.

[69] L. Wang, Y. Qin, Y. Sun, D. Peng, X. Peng, and P. Hu, "Robust contrastive cross-modal hashing with noisy labels," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 5752–5760. [Online]. Available: https://doi.org/10.1145/3664647.3680564

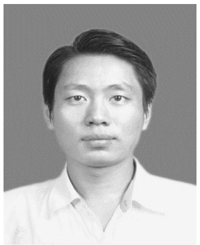
[70] G. Song, K. Huang, H. Su, F. Song, and M. Yang, "Deep ranking distribution preserving hashing for robust multi-label cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 7027–7042, 2024.

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[72] X. Wang, Z. Zhang, G. Lu, and Y. Xu, "Targeted attack and defense for deep hashing," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2298–2302. [Online]. Available: https://doi.org/10.1145/3404835.3463233

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.

Shibiao Xu (Member, IEEE) received the B.S. degree in Information Engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and the Ph.D. degree in Computer Science from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2014. He is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His current research interests include computer vision, image-based 3D scene reconstruction and understanding, computer graphics, multimodal artificial intelligence, deep learning and machine learning, and blockchain for copyright protection.



Xiaolong Zheng 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.

Guiyang Luo (Member, IEEE) received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications (BUPT), China, in 2020. He is currently a Postdoctoral Fellow at the State Key Laboratory of Networking and Switching Technology, BUPT. His current research interests include multi-agent systems and intelligent transportation systems.

Fei-Yue Wang received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

In 1990 he joined University of Arizona, Tucson, AZ, USA, where he became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Overseas Chinese Talents Program, and in 2002, he was appointed as the Director of the Key Laboratory for Complex Systems and Intelligence Science, Institute of Automation, CAS. From 2006 to 2010, he was the Vice President for research, education, and academic exchanges with the Institute of Automation, CAS. Since 2005, he has been the Dean of the School of Software Engineering, Xi'an Jiaotong University, Xi'an, China. In 2011, he became the State Specially Appointed Expert and the Founding Director of the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, CAS. He is the author or coauthor of over ten books and 300 papers published in the past three decades in his research areas, including social computing and parallel systems.

The supplementary materials of the paper Frequency Domain Adversarial Attacks on Deep Cross-Modal Hashing

1 Proof of the Properties of Polarization Loss

Definition 1 (Polarization l oss). For each data point $\mathbf{x} \in \mathcal{X}$ and its corresponding output vector $\hat{\mathbf{h}}_s := \Psi(\mathbf{x}; \mathbf{w}) \in \mathbb{R}^K$, the polarization loss is defined on the vector $\hat{\mathbf{h}}_s$ with respect to a pre-set target binary code $\mathbf{t} \in \mathcal{H}$ as follows:

$$\mathcal{L}_{\mathrm{ME}} = \sum_{i=1}^{K} \max\left(m - \hat{h}_{s,i} \cdot t_i, 0\right).$$

By minimizing the polarization loss (Eq. 3) during the learning phase, magnitudes of each DPN output channel are induced above the threshold m, while corresponding signs are aligned to the target vector $\mathbf{t}$. Figure 5 illustrates the distribution of outputs $\hat{\mathbf{h}}_s$, for example images fed into a DPN. Clearly, large margins push the network outputs further away from zero. It must be noted that the images are polarized, which is likely observed for misclassified data points.

Lemma 1. For the output vector $\hat{\mathbf{h}}_s$, it is bounded by the Hamming distance as follows:

$$\mathcal{D}_h(\mathbf{b}, \mathbf{t}) \leq \mathcal{L}_{\mathrm{ME}},$$

for any $m \geq 1$ and $\hat{\mathbf{h}}_s \in \{(\hat{h}_{s,1}, \cdots, \hat{h}_{s,K})\}$.

Proposition 1. Suppose class $\mathcal{C}$ consists of data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_{|\mathcal{C}|}\}$ associated with a pre-set target $\mathbf{t} \in \mathcal{H}$ in Hamming space. The averaged intra-class pairwise Hamming distances among the corresponding binary codes are given by:

$$\frac{1}{|\mathcal{C}|^2} \cdot \sum_{1 \leq i,j \leq |\mathcal{C}|} \mathcal{D}_h(\mathbf{b}_i, \mathbf{b}_j) \leq \frac{2}{|\mathcal{C}|} \cdot \sum_{1 \leq i \leq |\mathcal{C}|} \mathcal{L}_{\mathrm{ME}}.$$

Proposition 2. Suppose there are classes $\mathcal{C}_1, \cdots, \mathcal{C}_L$ with target binary codes $\mathbf{t}_x$ and $\mathbf{t}_y$ and binary hash codes $\mathbf{b}_i^x = \Phi(\mathbf{x}_i; \mathbf{w})$, where $i \in \{1, \cdots, |\mathcal{C}_x|\}$. The inter-class pairwise Hamming distances among the binary codes are:

$$\sum_{1 \le x \ne y \le L} \left(\mathcal{D}_h(\mathbf{t}_x, \mathbf{t}_y) - \frac{1}{|\mathcal{C}_x| \cdot |\mathcal{C}_y|} \cdot \sum_{\substack{1 \le i \le |\mathcal{C}_x|, \\ 1 \le j \le |\mathcal{C}_y|}} \mathcal{D}_h(\mathbf{b}_i^x, \mathbf{b}_j^y) \right)$$

$$\le \sum_{1 \le x \le L} \frac{2 \cdot (L-1)}{|\mathcal{C}_x|} \cdot \sum_{1 \le i \le |\mathcal{C}_x|} \mathcal{L}_{\mathrm{ME}}.$$

Proposition 3. The difference between averaged intra-class pairwise Hamming distance and averaged inter-class pairwise Hamming distance is upper bounded, i.e.

$$\sum_{1 \le x \le L} \frac{1}{|\mathcal{C}_x|^2} \cdot \sum_{1 \le i,j \le |\mathcal{C}_x|} \mathcal{D}_h(\mathbf{b}_i^x, \mathbf{b}_j^x)$$

$$- \sum_{1 \le x \ne y \le L} \frac{1}{|\mathcal{C}_x| \cdot |\mathcal{C}_y|} \sum_{\substack{1 \le i \le |\mathcal{C}_x| \\ 1 \le j \le |\mathcal{C}_y|}} \mathcal{D}_h(\mathbf{b}_i^x, \mathbf{b}_j^y)$$

$$\le \sum_{1 \le x \le L} \frac{2 \cdot L}{|\mathcal{C}_x|} \cdot \sum_{1 \le i \le |\mathcal{C}_x|} \mathcal{L}_{\mathrm{ME}} - \sum_{1 \le x \ne y \le L} \mathcal{D}_h(\mathbf{t}_x, \mathbf{t}_y).$$

Remarks:

I. Inequality shows that the averaged polarization loss is a strict upper-bound of the averaged pairwise Hamming distances between points of the same class. That is to say, minimizing the RHS effectively minimizes the averaged intra-class pairwise Hamming distances.

II. In terms of the computational complexity, pairwise Hamming distances on the LHS is $\mathcal{O}(|\mathcal{C}|^2)$ while the polarization loss on the RHS is $\mathcal{O}(|\mathcal{C}|)$ only.

III. Inequality shows that minimizing polarization losses on the RHS effectively maximizes the averaged inter-class pairwise Hamming distances on LHS.

IV. According to Proposition 3, the optimization problem of simultaneous minimizing intra-class and maximizing inter-class Hamming distances, i.e.

$$\min_{\mathbf{w}} \sum_{1 \le x \le L} \frac{1}{|\mathcal{C}_x|^2} \cdot \sum_{1 \le i,j \le |\mathcal{C}_x|} \mathcal{D}_h(\mathbf{b}_i^x, \mathbf{b}_j^x)$$

$$- \sum_{1 \le x \ne y \le L} \frac{1}{|\mathcal{C}_x| \cdot |\mathcal{C}_y|} \sum_{\substack{1 \le i \le |\mathcal{C}_x| \\ 1 \le j \le |\mathcal{C}_y|}} \mathcal{D}_h(\mathbf{b}_i^x, \mathbf{b}_j^y),$$

is equivalent to the problem of minimizing the averaged polarization loss over the whole data set, i.e.

$$\min_{\mathbf{w}} \sum_{1 \leq x \leq L} \frac{1}{|\mathcal{C}_x|} \cdot \sum_{1 \leq i \leq |\mathcal{C}_x|} \mathcal{L}_{\mathrm{ME}}. \qquad (8)$$