Spectral-Adaptive Adversarial Hashing for Robust Image Retrieval

Gang Zhou
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
zhougang2023@bupt.edu.cn

Shibiao Xu*
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
shibiao.xu@bupt.edu.cn

Xiaolong Zheng*
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
xiaolong.zheng@ia.ac.cn

Daniel Dajun Zeng
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
dajun.zeng@ia.ac.cn

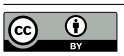
Abstract

Deep hashing is widely used in large-scale image retrieval systems due to its efficient retrieval performance. However, its susceptibility to adversarial attacks limits its security in practical applications. Adversarial training is the most effective method for improving robustness, but it often leads to a significant trade-off between robustness and retrieval accuracy. In this paper, we conduct spectral analysis and find that generating high-quality hash codes requires wide-frequency response models, whereas adversarial training forces the model into spectral collapse, degrading it to a low-frequency response model and weakening its discriminability. To address this issue, we propose a Spectral-Adaptive Adversarial Hashing (SAAH) framework, which selectively preserves discriminative and task-relevant frequency components while suppressing adversarially unstable ones, enabling robust hashing without sacrificing retrieval performance. Extensive experiments on benchmark datasets demonstrate that SAAH consistently achieves a superior balance between retrieval accuracy and adversarial robustness, achieving the best performance in both retrieval accuracy and robustness compared with existing robust hashing methods.

CCS Concepts

• **Security and privacy**; • **Information systems → Information retrieval**;

*Corresponding author.



SIGIR '26, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809611

Keywords

Deep hash, Adversarial Robustness, Image retrieval

ACM Reference Format:
Gang Zhou, Shibiao Xu, Xiaolong Zheng, and Daniel Dajun Zeng. 2026. Spectral-Adaptive Adversarial Hashing for Robust Image Retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3805712.3809611

1 Introduction

Deep hashing has emerged as an important technique for large-scale image retrieval due to its superior computational efficiency and significantly reduced storage overhead[20]. By mapping high-dimensional image features into a compact Hamming space, deep hashing enables fast similarity searches through bitwise XOR operations. A large number of studies have demonstrated that end-to-end training of deep neural networks with quantization constraints can yield highly discriminative hash codes that preserve complex semantic relationships in large datasets[20, 31, 32, 46, 50].

Although deep hashing models have achieved remarkable retrieval performance, recent studies indicate that they also inherit the vulnerability of deep neural networks and are highly susceptible to adversarial examples[14, 19, 33, 48]. Even imperceptible perturbations applied to input images can cause a drastic degradation in retrieval accuracy, leading the model to return semantically irrelevant results. To mitigate this threat, adversarial training (AT) has been introduced from the classification domain into deep hashing research[34, 35, 47], and is widely regarded as the most effective defense paradigm. AT adopts a min–max formulation, in which adversarial examples are produced by inner maximization, while

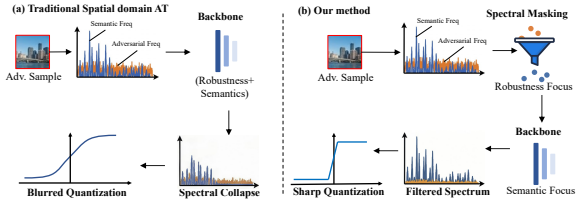


Figure 1: Comparison with existing frameworks. (a) Spatial-domain adversarial training: Induces spectral collapse and degrades discriminability, causing a severe robustness-accuracy trade-off. (b) Our SAAH: Selectively preserves discriminative frequency components and suppresses unstable ones, balancing robustness and retrieval performance.

the outer minimization updates model parameters to alleviate their impact, improving robustness against adversarial attacks[2].

Existing adversarial training methods for deep hashing, ATRDH[35], CgAT[34], and CRDAT[47], suffer from a trade-off between retrieval performance and robustness. These methods generate adversarial samples by searching pixel-wise perturbations in the spatial domain. However, prior studies show that such perturbations tend to overfit the source model[23, 27, 45, 49], forcing the network to improve robustness by suppressing input gradients and excessively smoothing the decision boundary[13, 21, 29]. We observe that this spatial over-smoothing manifests in the frequency domain as a full-band energy attenuation mechanism, where high-frequency components are preferentially suppressed, while low-frequency components also experience varying degrees of degradation. This spectral collapse fundamentally conflicts with the learning objective of deep hashing, which requires a wide-spectrum frequency response to achieve accurate quantization and highly discriminative semantic retrieval (see Section 3.2). We argue that robustness and semantic discrimination can be decoupled by separating and adaptively processing different frequency components, enabling the model to suppress adversarially unstable noise while preserving high-frequency semantic cues. A comparison between our method and spatial domain adversarial training is illustrated in Fig. 1.

To address the above issues, we propose **S**pectral-**A**daptive **A**dversarial **H**ashing (**SAAH**), a framework that incorporates frequency-domain awareness into adversarial training. SAAH introduces a lightweight Spectral Gating Network to predict an instance-adaptive soft mask in the DCT spectral domain, allowing the model to selectively preserve frequency components critical for discriminative hashing while attenuating those susceptible to adversarial perturbations. And we design a Spectral Stability Regularizer that explicitly penalizes fluctuations of the retained spectral components under adversarial attacks. By jointly optimizing the gating module and the hashing backbone, SAAH effectively alleviates the robustness–accuracy trade-off, achieving strong adversarial robustness without sacrificing the wide-spectrum representation capacity required for precise hash coding.

Our contributions include:

- We present a frequency-domain analysis of spectral bias in spatial adversarial training, revealing spectral collapse as the root cause of performance degradation in robust hashing.
- We propose an instance-adaptive spectral gating mechanism with stability-oriented regularization, enabling effective separation of robust semantic components from adversarial noise while preserving a stable and sufficient spectrum under adversarial settings.
- Extensive experiments on benchmark datasets demonstrate that SAAH consistently outperforms existing robust hashing methods, achieving a superior balance between retrieval accuracy and adversarial robustness.

2 Related Work

2.1 Deep Hashing for Image Retrieval

Deep hashing aims to learn compact binary codes that preserve semantic similarity for large-scale image retrieval. Representative supervised methods DPSH [20] enable end-to-end optimization via similarity-preserving objectives and quantization constraints. Recent studies have improved hashing quality by enhancing representation capacity and feature granularity. HHNet [5] exploits high-resolution features to strengthen discriminative representations, while frequency-aware methods FMTH [8] demonstrate that incorporating frequency-domain cues effectively benefits fine-grained retrieval. HGNet [44] further introduces high-frequency guidance to capture fine-grained texture details, jointly modeling coarse- and fine-grained information to produce more discriminative hash codes. These advances indicate that modern deep hashing increasingly depends on preserving detailed and wide-spectrum representations.

2.2 Adversarial Attacks and Robust Hashing Methods

Previous studies reveal that deep hashing models are highly vulnerable to adversarial perturbations. Early targeted attack methods, DHTA [3] and prototype-supervised adversarial networks [36], demonstrate that imperceptible perturbations can significantly distort neighborhood structures in Hamming space. More recently, generative-model-based attacks, including diffusion-driven targeted attacks, further improve attack effectiveness and transferability [16]. To defend against such threats, adversarial training has been introduced into hashing-based retrieval. Representative methods include ATRDH [35], CgAT [34], and semantic-aware adversarial training (SAAT) [43], which incorporate semantic centers or prototypes into the min–max framework. Very recent work such as CRDAT [47] explores two-stage adversarial training and representation distillation to mitigate the robustness–accuracy trade-off. However, most existing defenses operate in the spatial domain and often suffer from degraded discriminability after robustness enhancement.

2.3 Frequency-Domain Analysis and Robust Learning

Recent studies have shown that adversarial vulnerability and robustness are closely related to the frequency characteristics of neural networks. From a Fourier perspective, [41] demonstrate that adversarial perturbations predominantly exploit high-frequency components, while robust models exhibit substantially altered spectral responses. Theoretical analysis further indicates that robustness

constraints implicitly enforce smoothness on the learned function; Tsuzuku [18] show that Lipschitz regularization limits local variations, which corresponds to high-frequency suppression in the spectral domain. Recent defenses begin to explore explicit frequency-domain modeling. FreqPure[26] proposes diffusion-based adversarial purification with frequency-consistency constraints, demonstrating the effectiveness of spectral processing for robustness. However, existing approaches typically rely on fixed or global frequency transformations, which may indiscriminately suppress informative components, highlighting the necessity of adaptive spectral mechanisms for balancing robustness and discriminative representation.

3 Proposed Method

3.1 Preliminaries

Let $\mathcal{X} = \{(x_i, \mathbf{l}_i)\}_{i=1}^{N}$ denote a dataset containing N images, where each image x_i is associated with a multi-label vector $\mathbf{l}_i = [l_{i1}, \ldots, l_{iC}] \in \{0,1\}^C$, and C represents the total number of semantic categories. For any pair of instances (x_i, x_j), s_{ij} is defined based on their label overlap: $s_{ij} = 1$ if the two images share at least one semantic category (i.e., $\mathbf{l}_i^\top \mathbf{l}_j > 0$), and $s_{ij} = 0$ otherwise.

Deep hashing aims to learn a nonlinear mapping $f(x;\theta) \in \mathbb{R}^K$, where θ denotes the model parameters and K is the length of the hash code. Binary hash codes are obtained by applying the sign function: $\mathbf{b}_i = \text{sign}(f(x_i;\theta)), \quad \mathbf{b}_i \in \{-1,+1\}^K$. Semantically similar samples are expected to be close in the Hamming space. The Hamming distance $\text{Ham}(\cdot,\cdot)$ between two codes is defined as:

$$\text{Ham}(\mathbf{b}_i, \mathbf{b}_j) = \frac{1}{2}\left(K - \mathbf{b}_i^\top \mathbf{b}_j\right). \tag{1}$$

The training objective $\mathcal{L}_{hash}$ typically consists of a semantic similarity-preserving loss $\mathcal{L}_{\text{sem}}$ and a quantization loss $\mathcal{L}_{\text{quan}}$. $\mathcal{L}_{\text{sem}}$ is a function of the pairwise similarity $\Omega_{ij} = \frac{1}{2}\mathbf{h}_i^\top \mathbf{h}_j$, where $\mathbf{h}_i = \tanh(f(x_i;\theta))$. The $\tanh(\cdot)$ function is introduced to serve as a smooth approximation of $\text{sign}(f(x_i;\theta))$, since the sign function is non-differentiable. The quantization loss $\mathcal{L}_{\text{quan}}$ is defined as $\mathbf{b}_i - f(x_i;\theta)$ and aims to narrow the gap between the continuous embedding and the discrete binary hash code. Together, the two loss terms guide the model to learn highly semantically discriminative hash codes with low quantization error.

Adversarial Training (AT) enhances deep hashing robustness by jointly optimizing clean and adversarial samples. Given a perturbation budget $\|\delta_i\|_p \le \epsilon$, an adversarial example is defined as $x_i' = x_i + \delta_i$. The training process is formulated as a min-max optimization:

$$\begin{aligned} \delta_i^* &= \arg\max_{\|\delta_i\|_p \le \epsilon} \mathcal{L}_{\text{adv}}\left(f(x_i+\delta_i;\theta)\right), \\ \min_\theta \mathcal{L}(x,\theta) &= \sum_{i=1}^{N}\left[\mathcal{L}_{hash}(f(x_i;\theta)) + \lambda \mathcal{L}_{hash}(f(x_i+\delta_i^*;\theta))\right], \end{aligned} \tag{2}$$

where δ_i^* denotes the worst-case perturbation. The model parameters θ are updated to minimize the original loss $\mathcal{L}_{\text{hash}}$ for both clean and perturbed inputs, ensuring semantic consistency under attack.

3.2 Motivation

Wide-Spectrum Frequency Requirement of Deep Hashing. As reported in [7, 40], optimizing the quantization error $\mathcal{L}_{quan}$ improves the discriminative ability of hash codes. Ideally, each dimension of the learned mapping $f(x;\theta)$ should converge toward the element-wise $\text{sign}(\cdot)$ function. Using Fourier series expansion of the mapping function fitted to the model is an effective tool for analyzing the frequency domain characteristics of the model, as discussed in [28]. According to the classical theory of Fourier series[6], we can treat the sign function as the fundamental period of a square wave. Its Fourier series expansion at the fundamental frequency ω_0 is given by:

$$\text{sign}(x) = \sum_{k=1,3,5,\ldots}^{\infty} \frac{4}{\pi k}\sin(k\omega_0 x). \tag{3}$$

This expansion provides a direct basis for understanding the structural characteristics of the sign function across different frequency scales. Specifically, when only a limited number of low-order harmonics (i.e., small k, as shown in Figure 1(a)) are retained, f can reconstruct a rough approximation of the sign function. As more high-frequency components (larger k, as shown in Figures 1(b), (c), and (d)) are gradually introduced, the steepness of f at the transition point increases, forming relatively flat "plateau" structures in each region, ultimately approaching the instantaneous jump of the ideal sign function.

Consequently, we argue that deep hashing necessitates a model architecture capable of a wide-spectrum frequency response to effectively bridge the gap between continuous representations and discrete codes. This conclusion has been recently confirmed in[44], where high-frequency signals are shown to enhance the fine-grained image retrieval capability of deep hashing models.

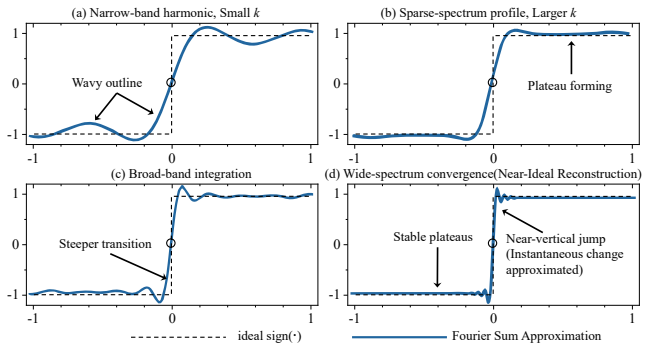


Figure 2: The progressive approximation of the signum function via Fourier series. As higher-frequency harmonics (k) are incrementally integrated from (a) to (d), the approximation evolves from a coarse global outline to a near-ideal discretization characterized by sharp transitions and stable plateaus.

Spectral Bias of Spatial Adversarial Training. According to [13, 21, 29], adversarial training significantly reduces the gradient norm of the model's loss function. To characterize the spectral impact of Adversarial Training (AT), we formulate the training process as a variational optimization problem. For a small perturbation budget

ϵ, AT asymptotically equates to augmenting the standard loss with a gradient-norm penalty term $\lambda\epsilon\|\nabla_x f\|^2$ [30].

Applying Parseval's theorem[6] and solving the corresponding Euler–Lagrange equation in the frequency domain yields the robust spectral magnitude $|\hat{f}(\omega)|_{AT}$ relative to its standard counterpart $|\hat{f}(\omega)|_{std}$ (see the appendix for the detailed derivation):

$$|\hat{f}(\omega)|_{AT} = \frac{1}{1+\eta+\lambda\epsilon|\omega|^2} \cdot |\hat{f}(\omega)|_{std}. \tag{4}$$

Here, $\hat{f}(\omega)$ denotes the Fourier transform of the mapping function $f(x)$ at frequency ω; $\lambda > 0$ controls the strength of spatial smoothness propagation in the spectral domain; and $\eta \geq 0$ represents an implicit regularization constant (e.g., weight decay) governing the global spectral gain.

This formulation reveals two critical phenomena:

- **High-frequency suppression.** The quadratic term $\lambda\epsilon\|\omega\|^2$ characterizes AT as a Tikhonov-type low-pass filter. The resulting suppression of high-order harmonics erodes the spectral integrity required to reconstruct the sharp "plateau" structures of the sign$(\cdot)$ function, thereby inducing a quantization blur in the Hamming space.
- **Fundamental energy erosion.** Even at the fundamental frequency ω_0, the attenuation factor $\Gamma(\omega_0, \epsilon) = (1+\eta+\lambda\epsilon\|\omega_0\|^2)^{-1}$ remains strictly below unity. This non-selective suppression implies that the "price of robustness" is a universal dampening of feature energy. Since the model's internal noise (e.g., numerical precision or stochasticity) remains constant, this erosion of signal magnitude directly degrades the effective Signal-to-Noise Ratio (SNR). This reduction in SNR inherently limits the model's discriminative ceiling, explaining the inevitable performance trade-off even on low-frequency, clean components.

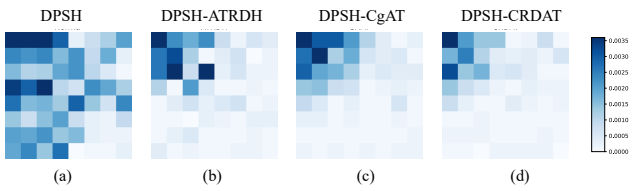


Figure 3: Comparison of RankingSHAP Values of Frequency Components between DPSH and three Adversarial Training Methods. (a) DPSH; (b) ATRDH + DPSH (AT); (c) Cgat + DPSH (AT); (d) Crdat + DPSH (AT)

To validate this, following [37], we transform each input image into the frequency domain via DCT transform[1] and partition the resulting spectrum into 64 frequency sub-bands arranged in an 8×8 grid, where frequencies increase from the upper-left corner (low-frequency components) to the bottom-right corner (high-frequency components). We then apply RankingSHAP [15] to quantify the contribution of each frequency component to retrieval performance, with the final scores averaged over 100 randomly sampled images.

As shown in Fig. 3, the comparison between standard DPSH [20] and robust models (ATRDH [35], CgAT [34], CRDAT [47]) reveals a pervasive spectral collapse. In contrast to the wide-spectrum response of DPSH, AT-based models suffer from severe spectral collapse, degrading to low-frequency response models: they not only lose critical high-frequency details required for sharp quantization, but also exhibit significantly weakened spectral response magnitudes even at extremely low frequencies.

As mentioned above, while spatial adversarial training (AT) enhances robustness by suppressing high-frequency components, it inevitably erodes the spectral information vital for discriminative hashing. Driven by this observation, our research aims to: (1) Separate robust components from vulnerable ones in the frequency domain, so as to preserve discriminative high-frequency signals and ensure retrieval performance; (2) Based on the separated robust components, enhance the adversarial robustness of the model accordingly.

3.3 SAAH Method

Overview. Prior studies suggest that high-frequency information can benefit deep hashing retrieval [44]. However, standard spatial domain adversarial training (AT) often induces over-smoothing, which can appear in the frequency domain as a suppression of broad-band (especially high-frequency) components (a phenomenon often described as "spectral collapse"). This may compromise the fine-grained discriminability required for hash coding. To mitigate this issue, we propose Spectral-Adaptive Adversarial Hashing (SAAH), which introduces an instance-adaptive spectral purification module before the hashing backbone. Concretely, a lightweight Spectral Gating Network predicts a soft frequency mask to attenuate adversarially unstable components, and the backbone extracts semantic features from the purified input. The two modules are jointly optimized under adversarial training. The overall architecture is shown in Fig. 4.

Instance-Adaptive Spectral Masking. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, we partition it into non-overlapping blocks of size $B\times B$ and apply a channel-wise block DCT-II. Let p index spatial blocks and $q=(u,v)$ index intra-block frequencies with $u,v \in \{0,\dots,B-1\}$. We denote the spectral coefficients by $\mathbf{z} = \mathcal{D}_B(x)$, where $\mathbf{z}_{p,q} \in \mathbb{R}^3$ collects the three-channel coefficients at location (p,q). For channel $c \in \{1,2,3\}$, the coefficient is

$$z_{p,q,c} = \frac{2}{B} C_u C_v \sum_{i=0}^{B-1}\sum_{j=0}^{B-1} x_{p,c}(i,j)\, \cos\left(\frac{(2i+1)u\pi}{2B}\right)\cos\left(\frac{(2j+1)v\pi}{2B}\right), \tag{5}$$

where $x_{p,c}(i,j)$ is the pixel at local position (i,j) in block p and channel c, and $C_k = \frac{1}{\sqrt{2}}$ for $k=0$ and $C_k = 1$ otherwise, where $k \in \{u,v\}$. We use the orthonormal scaling so that $\mathcal{D}_B^{-1}$ is its exact inverse.

We introduce a lightweight gating network g_ϕ (implemented as a compact MLP) to predict a soft mask on each spectral block. For each (p,q), it takes the magnitude vector $|\mathbf{z}_{p,q}| \in \mathbb{R}^3$ and outputs a scalar logit $a_{p,q} \in \mathbb{R}$. Stacking logits yields a logit map $a = g_\phi(|\mathbf{z}|)$ with the same (p,q) layout as $\mathbf{z}$. We then form a temperature-controlled soft gate and impose a lower bound $m_{\min} \in (0,1)$:

$$\hat{m}_{p,q} = \sigma\left(\frac{a_{p,q}}{\tau}\right), \qquad m_{p,q} = m_{\min} + (1-m_{\min})\hat{m}_{p,q}, \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid and $\tau > 0$ is a temperature. The scalar $m_{p,q}$ is broadcast to the three channels of $\mathbf{z}_{p,q}$. Finally, we define

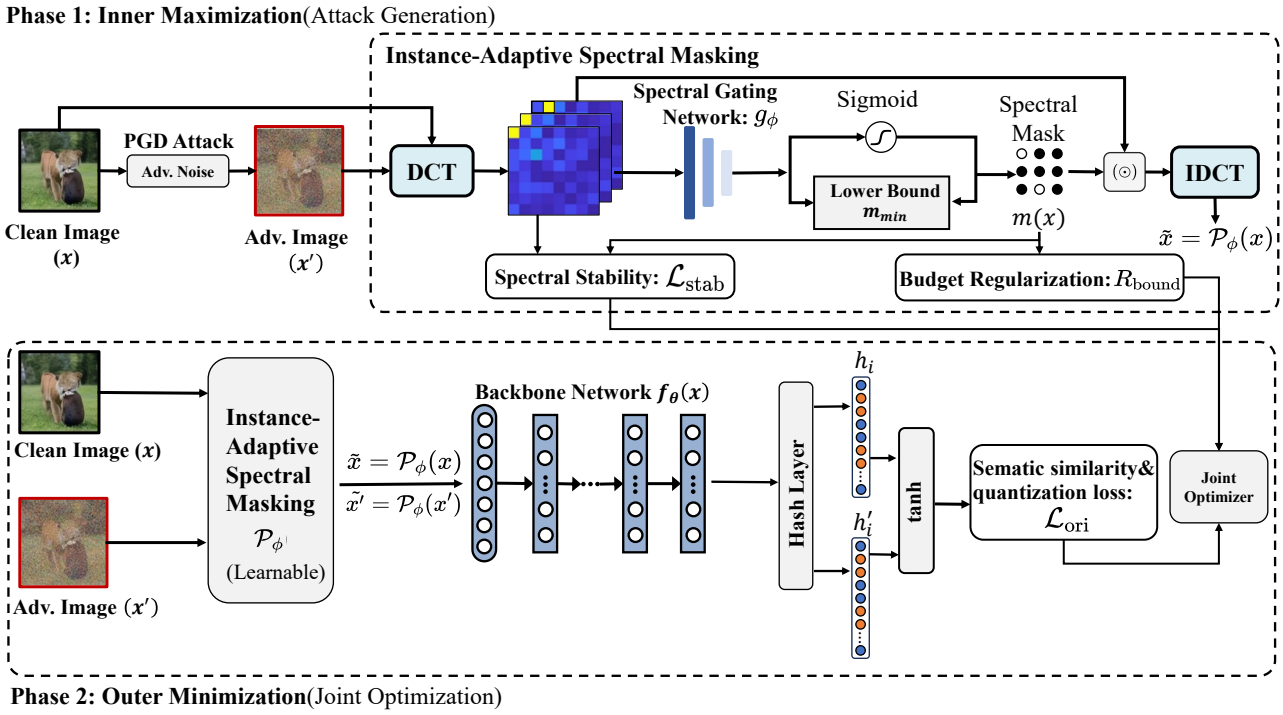


Figure 4: The overall framework of our proposed SAAH. Phase 1 (Inner Maximization): Generate adversarial examples via PGD attack. Phase 2 (Outer Minimization): Jointly optimize the spectral gating module and hashing backbone with spectral stability regularization and retention constraint for robust hashing.

the instance-adaptive purification operator

$$\mathcal{P}_\phi(x) := \mathcal{D}_B^{-1}\Big(\mathcal{D}_B(x) \odot m(x)\Big), \qquad m(x) := g_\phi(|\mathcal{D}_B(x)|), \quad (7)$$

and use $\tilde{x} = \mathcal{P}_\phi(x)$ as the input of the backbone.

Adversarial Training with Spectral Stability Regularization. We adversarially train the *entire* pipeline, including both the gating module and the hashing backbone f_θ. Under an ℓ_∞ budget ϵ, we generate adversarial samples by

$$x' = \arg \max_{\|x'-x\|_\infty \le \epsilon} \mathcal{L}_{\text{adv}}(x', x), \quad (8)$$

where the attack objective seeks to explicitly disrupt the semantic neighborhood structure by maximizing the hashing loss of the purified adversarial inputs:

$$\mathcal{L}_{adv}(x') = \mathcal{L}_{hash}(\mathcal{P}_\phi(x'), \mathcal{S}), \quad (9)$$

where $\mathcal{S}$ denotes the semantic similarity matrix derived from the ground-truth labels. In supervised hashing, $\mathcal{S}$ is derived from ground-truth labels, while in unsupervised or graph-based hashing, it can be instantiated as a structural adjacency matrix or pseudo-labels. We solve (8) using T-step PGD[25] with step size η:

$$x'_t = \Pi_{\|x'-x\|_\infty \le \epsilon}\Big(x'_{t-1} + \eta\, \text{sign}\big(\nabla_{x'} \mathcal{L}_{\text{adv}}(x'_{t-1}, x)\big)\Big), \qquad x'_0 = x, \quad (10)$$

where Π denotes projection onto the ℓ_∞-ball centered at x (and we additionally clip pixels to the valid range, e.g., $[0, 1]$).

To encourage the gate to retain components that are stable under attack, we impose a spectral stability regularizer. Let $\mathbf{z} = \mathcal{D}_B(x)$ and $\mathbf{z}' = \mathcal{D}_B(x')$. We measure the attack-induced spectral change and weight it by the *retained* components of the clean gate:

$$\mathcal{L}_{\text{stab}} = \big\|(\mathbf{z} - \mathbf{z}') \odot (m(x) \odot m(x))\big\|_2^2. \quad (11)$$

The squared mask emphasizes stability on the frequencies that the gate chooses to preserve, and components that fluctuate strongly under attack are thus discouraged from being retained.

To prevent degenerate solutions (e.g. retaining nearly all or nearly none of the spectrum), we constrain the average retention rate toward a target $\rho \in (0, 1)$ (with $\rho \ge m_{\min}$):

$$\mathcal{R}_{\text{bound}} = \left(\frac{1}{|\mathcal{I}|} \sum_{p,q} m_{p,q}(x) - \rho\right)^2, \quad (12)$$

where $|\mathcal{I}|$ is the number of spatial frequency positions.

Finally, we jointly optimize backbone parameters θ and gating parameters ϕ via

$$\min_{\theta,\phi} \mathcal{L}_{hash}(\mathcal{P}_\phi(x), \mathbf{l}) + \mathcal{L}_{hash}(\mathcal{P}_\phi(x'), \mathbf{l}) + \lambda_s \mathcal{L}_{\text{stab}} + \gamma \mathcal{R}_{\text{bound}}, \quad (13)$$

where λ_s, γ are trade-off hyper-parameters. When instantiating $\mathcal{L}_{hash}$ with DPSH, we use

$$\mathcal{L}_{hash} = -\sum_{i=1}^{N}\sum_{j=1}^{N}\Big(S_{ij}\Omega_{ij} - \log(1 + e^{\Omega_{ij}})\Big) + \lambda \sum_{i=1}^{N} \|\mathbf{b}_i - \mathbf{h}_i\|_2^2, \quad (14)$$

where λ balances the quantization regularizer.

4 Experiment

4.1 Experimental Setup

Dataset. To validate the effectiveness and generalization of our method, we conduct experiments on three widely used image retrieval benchmarks. These datasets vary in scale, label distribution, and annotation forms, providing a comprehensive evaluation setting. **MIRFLICKR-25K**[17] contains 25,000 image annotated with 24 semantic categories. We randomly select 2,000 images as the query set, 5,000 for training, and use the remaining pairs as the retrieval database. **NUS-WIDE**[10] consists of 269,648 images across 81 concepts. Following standard practice, we retain 195,834 samples from the 21 most frequent categories. Among them, 2,100 are used as queries, 10,500 for training, and the rest for retrieval. **MS-COCO**[22] provides 123,287 samples annotated with 80 labels. We combine the training and validation splits, and randomly sample 5,000 instances as queries, 10,000 for training, and use the rest as the retrieval set.

Evaluation setup. We adopt DPSH[20] as the default hashing loss for adversarial training, with AlexNet as the backbone. Experiments are further conducted on other popular architectures, including AlexNet, ResNet-50, DenseNet, and ViT. We also evaluate the generality of our method on other deep hashing frameworks, including DPN[12], CSQ[42] and LTH[9]. ATRDH[35], CgAT[34], and CRDAT[47] are selected as competing defense methods. Following the CRDAT[47], we evaluate robustness under multiple adversarial attacks, including three untargeted attacks: HAG[39], SDHA[24], and CgAT[34], four targeted attacks: P2P[4], DHTA[3], THA[35], and NAG[38]. MAP and t-MAP[4] are used to evaluate untargeted and targeted attacks on the adversarially trained model, respectively. Robust MAP and clean MAP denote retrieval performance on adversarial and clean samples. All results are reported over the top-5,000 retrieved samples.

Implementation details. We re-implement existing adversarial training methods using the same hyperparameters and experimental settings as reported in their original papers. Since CgAT and SAAT[43] share highly similar algorithmic designs and implementations, we only compare our defense performance with CaAT. During adversarial training, the entire SAAH framework, including both the frequency-domair gating module and the backbone network is jointly optimized in an end-to-end manner. We adopt the SGD optimizer with an initial learning rate of 1×10^{-2}, momentum of 0.9, and weight decay of 5×10^{-4}.The model is trained for 100 epochs, with the learning rate decayed by a factor of 1/10 at the 50th and 75th epochs. For Eq. (13), We tune λ_s and γ on a held-out validation set and find that setting both to 1.0 yields consistently strong performance. The target spectral retention ratio is fixed at $\rho = 0.5$, and the minimum gating value is set to $m_{\min} = 0.05$. The block size of the differentiable block-wise DCT is set to $B = 8$, and the gating temperature parameter is fixed at $\tau = 1.0$. The number of iterations T for the PGD attack is set to 7.

4.2 Standard Retrieval Performance

As shown in Table 1, the first row labeled as *Clean* reports the standard retrieval performance evaluated on clean samples. Our method consistently outperforms the other three adversarial training baselines across all three datasets, with the most significant improvement observed on MS-COCO. This improvement is mainly attributed to the higher image quality of MS-COCO, which contains richer high-frequency discriminative signals. Meanwhile, we observe that the performance of SAAH remains inferior to that of the standard DPSH model. This is because DPSH relies on certain fragile and semantically irrelevant cues for discrimination, whereas our method deliberately filters out such non-robust signals.

4.3 Performance under White-Box Attacks

Defense against untargeted adversarial attacks. After being subjected to untargeted white-box attacks (HAG, SDHA, and CgAT), the robustness performance of adversarially trained models is reported in Table 1. A higher MAP indicates stronger robustness under attack. The results show that DPSH without adversarial training exhibits the poorest robustness after attacks, whereas our SAAH achieves the most significant robustness improvement compared with the other three adversarial training methods. These findings demonstrate that our approach can effectively filter out vulnerable frequency bands while preserving discriminative semantic information across the full frequency spectrum.

Defense against targeted adversarial attacks. To further evaluate robustness under targeted attack scenarios, we report in Table 2 the defense performance against three representative targeted adversarial attacks (P2P, DHTA, and THA). As shown in the results, our robust model effectively prevents adversarial queries from retrieving images associated with the target labels, achieving the lowest robust t-MAP among all compared methods. These findings indicate that our approach attains state-of-the-art robustness against targeted adversarial attacks, consistently outperforming existing defense strategies.

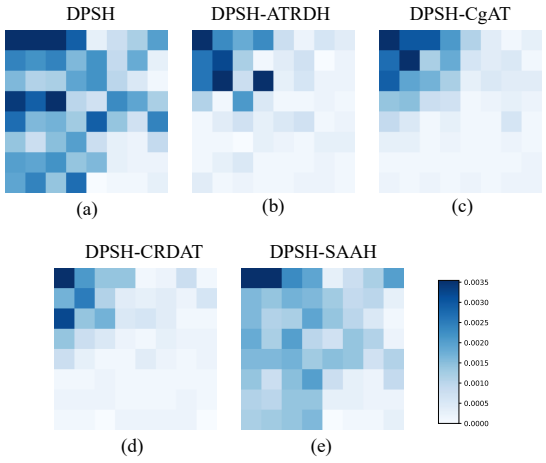


Figure 5: Frequency-domain feature attribution results obtained by RankingShAP. The color intensity indicates the importance of each frequency component to retrieval performance.

4.4 Performance under Black-Box Attacks

We evaluate the defense performance against black-box adversarial attacks using NAG. Adversarial examples are generated from clean

Table 1: Standard Performance and Robustness Comparison under Untargeted Attacks (MAP). All adversarial training methods use DPSH as the original loss function, and AlexNet is adopted as the backbone network, and the attack strength for all methods are set to 8/255.

Attack	Defense	MIRFLICKR-25K			NUS-WIDE			MS-COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
Clean	DPSH	0.8346	0.8221	0.8673	0.7915	0.8412	0.8268	0.6695	0.7014	0.7582
	ATRDH	0.7218	0.7512	0.7924	0.6894	0.7605	0.7421	0.5283	0.6119	0.6397
	CgAT	0.7459	0.7634	0.8145	0.6612	0.7459	0.8032	0.5826	0.6025	0.6473
	CRDAT	0.7512	0.7435	0.7518	0.7482	0.7368	0.7892	0.5284	0.5892	0.5789
	SAAH	**0.7898**	**0.7913**	**0.8319**	**0.7775**	**0.7905**	**0.8111**	**0.6045**	**0.6223**	**0.6784**
HAG	DPSH	0.2112	0.2435	0.2321	0.0924	0.1268	0.0915	0.0871	0.0612	0.0994
	ATRDH	0.3645	0.3182	0.3275	0.3061	0.2594	0.2118	0.2267	0.2014	0.1872
	CgAT	**0.5264**	0.4398	0.4521	0.4892	0.4157	0.3821	0.2514	0.1712	0.1935
	CRDAT	0.5212	**0.5189**	**0.5312**	0.5014	**0.5562**	**0.5624**	0.3712	0.4024	0.4592
	SAAH	0.5176	0.4467	0.4621	**0.5223**	0.5425	0.5523	**0.3856**	**0.4231**	**0.4621**
SDHA	DPSH	0.1824	0.1321	0.1568	0.0812	0.1198	0.0624	0.0792	0.0512	0.0821
	ATRDH	0.2541	0.2512	0.2145	0.2218	0.1694	0.1341	0.1742	0.2085	0.1594
	CgAT	0.4178	0.3214	0.2921	0.4672	0.4385	0.3524	0.2294	0.1582	0.1412
	CRDAT	0.4468	0.4612	**0.5364**	0.5024	0.5112	0.5421	0.4092	0.4452	0.4782
	SAAH	**0.4678**	**0.4734**	0.4865	**0.5129**	**0.5321**	**0.5428**	**0.4321**	**0.4538**	**0.4859**
CgAT	DPSH	0.1472	0.1065	0.1284	0.1042	0.0782	0.0954	0.0921	0.0612	0.1135
	ATRDH	0.3185	0.2674	0.2841	0.3412	0.2612	0.2412	0.2492	0.2385	0.2274
	CgAT	0.4021	0.3125	0.3124	0.4992	0.4168	0.4065	0.2351	0.2274	0.2098
	CRDAT	0.4685	0.4812	0.4712	0.5212	0.5482	0.5724	0.3854	0.3762	0.3924
	SAAH	**0.4765**	**0.4901**	**0.4925**	**0.5382**	**0.5551**	**0.5821**	**0.4019**	**0.4129**	**0.4061**

Table 2: Robustness Comparison under Targeted Attacks. All adversarial training methods use DPSH as the original loss function, and AlexNet is adopted as the backbone network, and the attack strength for all methods is set to 8/255.

Attack	Defense	MIRFLICKR-25K			NUS-WIDE			MS-COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
P2P	DPSH	0.8864	0.9118	0.9075	0.8213	0.8842	0.8412	0.6882	0.7562	0.7248
	ATRDH	0.7712	0.8241	0.7923	0.6501	0.7164	0.6912	0.4655	0.5182	0.5123
	CgAT	0.7215	0.7732	0.7366	0.5784	0.6112	0.6358	0.4812	0.5194	0.5024
	CRDAT	0.6955	0.7182	**0.7024**	0.6012	0.6115	0.6288	0.4612	**0.4678**	**0.4811**
	SAAH	**0.6874**	**0.7012**	0.7123	**0.5821**	**0.6025**	**0.6184**	**0.4502**	0.4753	0.4923
DHTA	DPSH	0.8988	0.9492	0.9312	0.8612	0.9221	0.8584	0.6775	0.7655	0.7198
	ATRDH	0.7601	0.8258	0.8012	0.6442	0.7192	0.6724	0.4412	0.4952	0.4912
	CgAT	0.6675	0.7462	0.7075	0.5382	0.6062	0.6015	0.4542	0.4894	0.4588
	CRDAT	0.6712	0.6955	0.6684	0.5678	0.5712	0.5898	0.4182	0.4244	0.4382
	SAAH	**0.6565**	**0.6629**	**0.6513**	**0.5119**	**0.5432**	**0.5608**	**0.4012**	**0.4084**	**0.4112**
THA	DPSH	0.9382	0.9785	0.9572	0.8812	0.9412	0.9255	0.8012	0.8812	0.8594
	ATRDH	0.8655	0.9262	0.9042	0.7592	0.8384	0.8212	0.6124	0.6062	0.6815
	CgAT	0.8485	0.9212	0.9015	0.6655	0.8124	0.8312	0.6145	0.7062	0.6755
	CRDAT	0.8088	**0.8182**	**0.8564**	0.7001	0.7198	0.7512	0.5264	0.5762	0.5882
	SAAH	**0.7921**	0.8199	0.8661	**0.6512**	**0.6478**	**0.6534**	**0.5123**	**0.5523**	**0.5632**

surrogate hashing models with diverse architectures, including three CNN-based networks (AlexNet, ResNet-50, and DenseNet) and one pretrained vision transformer (ViT-B/16)[11], and are then transferred to attack the adversarially trained target model based on AlexNet. Both surrogate and target models adopt DPSH as the underlying hashing loss to ensure a fair comparison. All experiments are conducted on the FLICKR-25K dataset with different hash code lengths. As shown in Table 3, our proposed SAAH consistently

Table 3: T-MAP performance under the NAG targeted attack on the MIRFLICKR-25K dataset. The surrogate model architectures are arranged vertically, and the adversarial training methods of the robust target models are arranged horizontally.

	16 bits				32 bits				64 bits			
Defense	AlexNet	ResNet50	DenseNet161	ViT	AlexNet	ResNet50	DenseNet161	ViT	AlexNet	ResNet50	DenseNet161	ViT
DPSH	0.9065	0.7834	0.7521	0.7232	0.9185	0.7874	0.7632	0.7123	0.9324	0.7885	0.7765	0.7323
ATRDH	0.7758	0.7112	0.7375	0.6943	0.8024	0.7552	0.7251	0.6621	0.7852	0.7492	0.7234	0.6832
CgAT	0.7432	0.7235	0.7542	0.6723	0.7935	0.7652	0.7412	0.6654	0.8062	0.7532	0.7772	0.6721
CRDAT	0.7358	0.7018	0.7012	**0.6532**	0.7472	0.7192	0.7145	0.6452	0.7432	0.7135	0.7172	0.6578
SAAH	**0.7021**	**0.6823**	**0.6881**	0.6621	**0.7342**	**0.7003**	**0.7111**	**0.6245**	**0.7212**	**0.6891**	**0.6997**	**0.6473**

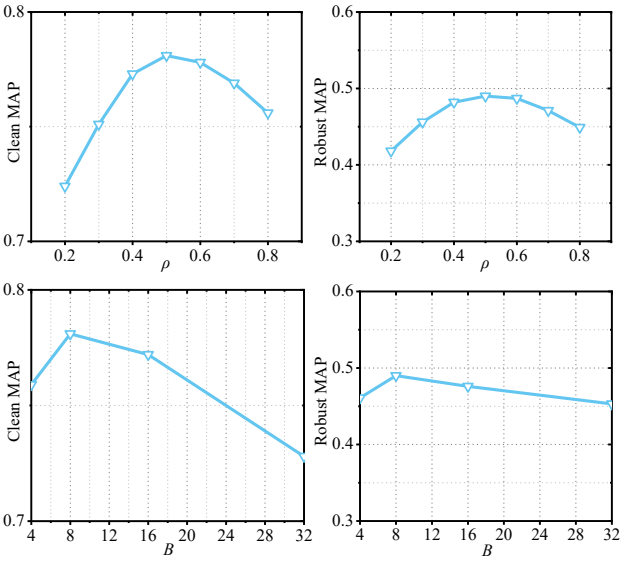


Figure 6: Comparison of MAP andd Robust MAP scores on the MIRFLICKR-25K dataset with different parameter configurations.

outperforms competing methods under black-box targeted attacks, demonstrating superior robustness against attack transferability.

4.5 Ablation Study

Table 4: MAP scores of ablating different components in SAAH (32 bits, MIRFLICKR-25K).

Component	Variant	Clean	HAG	SDHA	CgAT
Inst.-adaptive Gate	Low-pass	0.7412	0.4635	0.4218	0.3964
Inst.-adaptive Gate	Global	0.7586	0.4892	0.4571	0.4233
Inst.-adaptive Gate	w/o	0.7704	0.5162	0.4837	0.4591
$\mathcal{L}_{stab}$	w/o	0.7621	0.4984	0.4602	0.4318
$\mathcal{R}_{bound}$	w/o	0.7658	0.5073	0.4714	0.4402
SAAH (Full)		**0.7813**	**0.5276**	**0.4901**	**0.4685**

Component-wise ablation. Table 4 validates each component in SAAH. Compared to static gating (Low-pass/Global) or removing the module (*w/o Inst.-adaptive Gate*), our dynamic approach significantly enhances robustness. Omitting $\mathcal{L}_{stab}$ leads to a sharp performance decline under strong attacks (e.g., CgAT), confirming its role in suppressing attack-sensitive frequencies. Finally, removing $\mathcal{R}_{bound}$ hurts overall MAP, demonstrating its necessity in preventing the gate from falling into degenerate solutions.

Spectral Attribution Analysis of Robust Hashing Models Consistent with the motivation analysis, we employ RankingShAP[15] to perform frequency-domain feature attribution on the trained models. As shown in Fig. 5, our SAAH maintains a wide-band spectral response even after adversarial training, indicating that the model leverages both high- and low-frequency signals to construct highly discriminative representations.

Hyper-parameter sensitivity. We study the sensitivity to the target retention ratio ρ and the DCT block size B. As show in Fig. 6, moderate retention (e.g., $\rho \in [0.4, 0.6]$) provides the best balance: too small ρ over-suppresses informative details and hurts Clean MAP, while too large ρ retains more attack-sensitive components and slightly reduces robust MAP(Attacking by SAAT). For the DCT design, B=8 offers a good trade-off between locality and frequency resolution; very small blocks reduce frequency resolution, whereas very large blocks weaken locality and may make the gate less effective for spatially-varying perturbations.

These results verify that (i) spectral gating is the key to recovering wide-spectrum discriminability under adversarial training, (ii) stability regularization is essential to avoid retaining attack-sensitive bands, and (iii) the retention boundary constraint prevents degenerate gating behaviors, together yielding a consistently better robustness–accuracy trade-off.

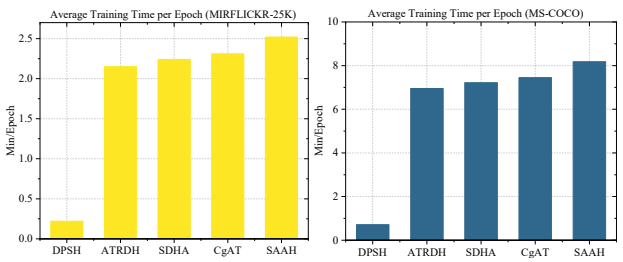


Figure 7: Comparison of Average Training Time per Epoch with Adversarial Training Baselines.

4.6 Efficiency Analysis

Computational and parameter overhead. SAAH introduces additional frequency-domain operations, including block-wise DCT transformation and a lightweight spectral gating network. For an input image of resolution $H \times W$, the block-wise DCT with fixed block size B=8 has linear complexity $O(HW)$, as each pixel participates in exactly one local transform. Compared with convolutional feature extraction, whose complexity typically scales as $O(HWCK^2)$, the computational cost of the frequency transform is insignificant in practice. The spectral gating network is implemented as a compact MLP operating on spectral magnitudes. With B=8, the total number of parameters is approximately 3.1×10^4, accounting for less than **0.1%** of the backbone parameters. No additional parameters are introduced into the hashing backbone, and the hash code length and retrieval pipeline remain unchanged. Therefore, SAAH incurs minimal memory overhead and does not affect storage or indexing efficiency in large-scale hashing systems.

Training efficiency. The dominant computational cost of robust hashing methods originates from adversarial example generation via multi-step(7 in iterations) PGD. The spectral purification and gating operations introduced by SAAH add only a constant overhead to each forward–backward pass and do not change the optimization procedure. As illustrated in Fig. 7, on MIRFLICKR-25K with AlexNet backbone and 10-step PGD, the average training time per epoch increases from **2.31 minutes** for CgAT to **2.49 minutes** for SAAH, corresponding to an overhead of approximately **7.7%**. Similar trends are observed on MS-COCO, where the training overhead consistently remains within **6%–8%**, depending on the backbone architecture. These results indicate that SAAH maintains comparable training efficiency to existing adversarial hashing methods while providing significantly improved robustness and retrieval performance.

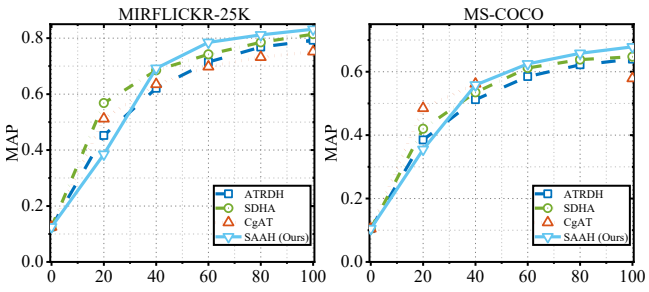


Figure 8: Convergence analysis of MAP during the training process on (a) MIRFLICKR-25K and (b) MS-COCO with 64-bits hash codes.

Fig. 8 shows the training convergence of SAAH and baselines. SAAH consistently achieves the highest robust MAP with steady, monotonic growth. Conversely, CgAT exhibits catastrophic overfitting, peaking early before decaying on MS-COCO. This suggests that while spatial methods fit unstable noise, SAAH's spectral gating acts as an effective regularizer. By filtering out components triggering optimization fluctuations, SAAH ensures stable convergence to a superior, more reliable robust optimum.

Methods	w/o SAAH	SAAH Applied
CSQ	0.3835	0.6239
DPN	0.3125	0.7321
LTH	0.3354	0.7635

Table 5: Robustness Improvement from SAAH under Untargeted CgAT Attacks.(MIRFLICKR-25K with 64-bits)

4.7 Universality Analysis on Other Hashing Models

We evaluate the universality of SAAH on different deep hashing methods. By replacing the original DPSH loss with DPN[12], CSQ[42], and LTH[9], we construct corresponding SAAH variants. As shown in Table 5, SAAH consistently improves robustness across all methods, demonstrating strong universality and adaptability.

5 Conclusion

In this paper, we investigated the robustness–accuracy trade-off in adversarially trained deep hashing and provided a frequency-domain explanation for why spatial-domain adversarial training tends to degrade retrieval performance. Our analysis shows that high-quality hash coding intrinsically requires a wide-spectrum frequency response to approximate the sharp quantization behavior of the sign$(\cdot)$ function, whereas spatial domain adversarial training acts as an implicit low-pass mechanism that induces full-band attenuation and, in particular, high-frequency suppression. To address this conflict, we proposed **S**pectral-**A**daptive **A**dversarial **H**ashing (**SAAH**), which introduces an instance-adaptive spectral gating network to selectively suppress adversarially unstable components while preserving task-relevant discriminative frequencies. We further designed a spectral stability regularizer and a retention boundary constraint to encourage the model to retain stable and sufficient spectral information under attack. Extensive experiments on multiple benchmarks and under diverse untargeted/targeted attack settings demonstrate that SAAH consistently achieves a better balance between clean retrieval precision and adversarial robustness than prior robust hashing methods, validating the effectiveness and generality of adaptive spectral purification for robust retrieval.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2025YFE0216800, the Beijing Natural Science Foundation (No. JQ23014), and the National Natural Science Foundation of China (Nos. 62271074, 72225011, 72434005) and L242400108, and partially supported by BUPT Kunpeng&Ascend Center of Cultivation.

6 Appendix

A Derivation of the Spectral Attenuation Formula

This appendix provides the formal derivation of the spectral relationship between adversarially trained and standard models, as discussed in the main text.

A.1 Variational Formulation of Adversarial Training

Following the first-order approximation proposed by [30], the objective of adversarial training (AT) with a small perturbation budget ϵ can be formulated as a regularized minimization problem. We define the functional $\mathcal{J}(f)$ for the mapping f as:

$$\begin{aligned}\min_f \mathcal{J}(f) =& \frac{1}{2}\int_\Omega \|f(x) - f_{std}(x)\|^2 dx \\ &+ \frac{\eta}{2}\int_\Omega \|f(x)\|^2 dx + \frac{\lambda\epsilon}{2}\int_\Omega \|\nabla_x f(x)\|^2 dx\end{aligned} \tag{15}$$

where f_{std} is the mapping learned by standard training, η represents the implicit regularization constant, and $\lambda\epsilon$ scales the gradient penalty strength.

A.2 Frequency Domain Transformation

We apply the Fourier Transform $\mathcal{F}$ and invoke **Parseval's Theorem** to map the spatial integrals into the spectral domain. Using the property $\mathcal{F}\{\nabla_x f\} = j\omega\hat{f}(\omega)$, the objective functional is rewritten as:

$$\begin{aligned}\mathcal{J}(\hat{f}) = \frac{1}{2}\int_{\mathbb{R}^d}\Big[&|\hat{f}(\omega) - \hat{f}_{std}(\omega)|^2 + \eta|\hat{f}(\omega)|^2 \\ &+ \lambda\epsilon\|\omega\|^2|\hat{f}(\omega)|^2\Big] d\omega\end{aligned} \tag{16}$$

where j denotes the imaginary unit and ω is the frequency vector.

A.3 Optimality and Closed-form Solution

To find the optimal robust mapping $\hat{f}_{AT}$, we compute the functional derivative of the integrand L with respect to the complex conjugate $\hat{f}^*$:

$$L = (\hat{f} - \hat{f}_{std})(\hat{f}^* - \hat{f}^*_{std}) + (\eta + \lambda\epsilon\|\omega\|^2)\hat{f}\hat{f}^* \tag{17}$$

Setting the Euler-Lagrange equation $\frac{\partial L}{\partial \hat{f}^*} = 0$ yields:

$$(\hat{f}(\omega) - \hat{f}_{std}(\omega)) + (\eta + \lambda\epsilon\|\omega\|^2)\hat{f}(\omega) = 0 \tag{18}$$

By rearranging terms and taking the magnitude, we obtain the spectral attenuation factor $\Gamma(\omega, \epsilon)$:

$$|\hat{f}(\omega)|_{AT} = \underbrace{\frac{1}{1 + \eta + \lambda\epsilon\|\omega\|^2}}_{\Gamma(\omega,\epsilon)} \cdot |\hat{f}(\omega)|_{std} \tag{19}$$

Eq. 19 characterizes the dual effect of AT: (i) a quadratic decay at high frequencies ($\|\omega\| \to \infty$) and (ii) a global gain reduction at fundamental frequencies ($\omega \to \omega_0$) due to the non-zero regularization constants η and ϵ.

References

[1] N. Ahmed, T. Natarajan, and K.R. Rao. 1974. Discrete Cosine Transform. *IEEE Trans. Comput.* C-23, 1 (1974), 90–93. doi:10.1109/T-C.1974.223784

[2] Maksym Andriushchenko and Nicolas Flammarion. 2020. Understanding and Improving Fast Adversarial Training. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 16048–16059. https://proceedings.neurips.cc/paper_files/paper/2020/file/b8ce47761ed7b3b6f48b583350b7f9e4-Paper.pdf

[3] Jiawang Bai, Bin Chen, Yiming Li, Dongxian Wu, Weiwei Guo, Shu-tao Xia, and En-hui Yang. 2020. Targeted attack for deep hashing based retrieval. In *European Conference on Computer Vision*. Springer, 618–634.

[4] Jiawang Bai, Bin Chen, Yiming Li, Dongxian Wu, Weiwei Guo, Shu-Tao Xia, and En-Hui Yang. 2020. Targeted Attack for Deep Hashing Based Retrieval. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 618–634.

[5] Aymene Berriche, Mehdi Zakaria Adjal, and Riyadh Baghdadi. 2025. Leveraging High-Resolution Features for Improved Deep Hashing-Based Image Retrieval. In *Advances in Information Retrieval*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer Nature Switzerland, Cham, 440–453.

[6] E. Oran Brigham. 1988. *The fast Fourier transform and its applications.* Prentice-Hall, Inc., USA.

[7] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. 2017. HashNet: Deep Learning to Hash by Continuation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[8] Jiayi Chen, Shuli Cheng, Liejun Wang, Yongming Li, and Qiang Zou. 2025. Frequency Decoupling Enhancement and Mamba Depth Extraction-Based Feature Fusion in Transformer Hashing Image Retrieval. *Knowledge-Based Systems* 310 (2025), 113036. doi:10.1016/j.knosys.2025.113036

[9] Yong Chen, Yuqing Hou, Shu Leng, Qing Zhang, Zhouchen Lin, and Dell Zhang. 2021. Long-Tail Hashing. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1328–1338. doi:10.1145/3404835.3462888

[10] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval* (Santorini, Fira, Greece) *(CIVR '09)*. Association for Computing Machinery, New York, NY, USA, Article 48, 9 pages. doi:10.1145/1646396.1646452

[11] Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu. 2022. Vision Transformer Hashing for Image Retrieval. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. doi:10.1109/ICME52920.2022.9859900

[12] Lixin Fan, Kam Woh Ng, Ce Ju, Tianyu Zhang, and Chee Seng Chan. 2020. Deep Polarized Network for Supervised Learning of Accurate Binary Hashing Codes.. In *Ijcai*, Vol. 825.

[13] Chris Finlay and Adam M. Oberman. 2021. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications* 3 (2021), 100017. doi:10.1016/j.mlwa.2020.100017

[14] Xinru Guo, Huaxiang Zhang, Li Liu, Dongmei Liu, Xu Lu, and Hui Meng. 2025. Primary Code Guided Targeted Attack against Cross-modal Hashing Retrieval. *IEEE Transactions on Multimedia* 27 (2025), 312–326. doi:10.1109/TMM.2024.3521697

[15] Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. RankingSHAP - Faithful Listwise Feature Attribution Explanations for Ranking Models;. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 381–391. doi:10.1145/3726302.3729971

[16] Chihan Huang and Xiaobo Shen. 2025. HUANG: A Robust Diffusion Model-based Targeted Adversarial Attack Against Deep Hashing Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 4 (Apr. 2025), 3626–3634. doi:10.1609/aaai.v39i4.32377

[17] Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval* (Vancouver, British Columbia, Canada) *(MIR '08)*. Association for Computing Machinery, New York, NY, USA, 39–43. doi:10.1145/1460096.1460104

[18] Vishaal Krishnan, Abed AlRahman Al Makdah, and Fabio Pasqualetti. 2020. Lipschitz Bounds and Provably Robust Training by Laplacian Smoothing. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 10924–10935. https://proceedings.neurips.cc/paper_files/paper/2020/file/7bab7650be60b0738e22c3b8745f937d-Paper.pdf

[19] Chao Li, Shangqian Gao, Cheng Deng, Wei Liu, and Heng Huang. 2021. Adversarial Attack on Deep Cross-Modal Hamming Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2218–2227.

[20] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. 2016. Feature learning based deep supervised hashing with pairwise labels. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 1711–1717.

[21] Yueru Li, Shuyu Cheng, Hang Su, and Jun Zhu. 2020. Defense Against Adversarial Attacks via Controlling Gradient Leaking on Embedded Manifolds. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 753–769.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[23] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. 2022. Frequency Domain Model Augmentation for Adversarial Attack. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 549–566.

[24] Junda Lu, Mingyang Chen, Yifang Sun, Wei Wang, Yi Wang, and Xiaochun Yang. 2021. A Smart Adversarial Attack on Deep Hashing Based Image Retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval* (Taipei, Taiwan) *(ICMR '21)*. Association for Computing Machinery, New York, NY, USA, 227–235. doi:10.1145/3460426.3463640

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJzIBfZAb

[26] Gaozheng Pei, Ke Ma, Yingfei Sun, Qianqian Xu, and Qingming Huang. 2025. Diffusion-based Adversarial Purification from the Perspective of the Frequency Domain. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=Bm706VlAtU

[27] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. 2022. Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 29845–29858. https://proceedings.neurips.cc/paper_files/paper/2022/file/c0f9419caa85d7062c7e6d621a335726-Paper-Conference.pdf

[28] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the Spectral Bias of Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5301–5310. https://proceedings.mlr.press/v97/rahaman19a.html

[29] Adrián Rodríguez-Muñoz, Tongzhou Wang, and Antonio Torralba. 2025. Characterizing Model Robustness via Natural Input Gradients. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 161–178.

[30] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. 2019. First-Order Adversarial Vulnerability of Neural Networks and Input Dimension. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5809–5817. https://proceedings.mlr.press/v97/simon-gabriel19a.html

[31] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Liangdao Wang, Yan Pan, Cong Liu, Hanjiang Lai, Jian Yin, and Ye Liu. 2023. Deep Hashing With Minimal-Distance-Separated Hash Centers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 23455–23464.

[33] Tianshi Wang, Lei Zhu, Zheng Zhang, Huaxiang Zhang, and Junwei Han. 2023. Targeted Adversarial Attack Against Deep Cross-Modal Hashing Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 10 (2023), 6159–6172. doi:10.1109/TCSVT.2023.3263054

[34] Xunguang Wang, Yiqun Lin, and Xiaomeng Li. 2023. CgAT: Center-Guided Adversarial Training for Deep Hashing-Based Retrieval. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3268–3277. doi:10.1145/3543507.3583369

[35] Xunguang Wang, Zheng Zhang, Guangming Lu, and Yong Xu. 2021. Targeted Attack and Defense for Deep Hashing. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2298–2302. doi:10.1145/3404835.3463233

[36] Xunguang Wang, Zheng Zhang, Baoyuan Wu, Fumin Shen, and Guangming Lu. 2021. Prototype-Supervised Adversarial Network for Targeted Attack of Deep Hashing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16357–16366.

[37] Yajie Wang, Yi Wu, Shangbo Wu, Ximeng Liu, Wanlei Zhou, Liehuang Zhu, and Chuan Zhang. 2024. Boosting the Transferability of Adversarial Attacks With Frequency-Aware Perturbation. *IEEE Transactions on Information Forensics and Security* 19 (2024), 6293–6304. doi:10.1109/TIFS.2024.3411921

[38] Yanru Xiao and Cong Wang. 2021. 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)*. 1934–1943.

[39] Erkun Yang, Tongliang Liu, Cheng Deng, and Dacheng Tao. 2020. Adversarial Examples for Hamming Space Search. *IEEE Transactions on Cybernetics* 50, 4 (2020), 1473–1484. doi:10.1109/TCYB.2018.2882908

[40] Tao Yao, Ruxin Wang, Jintao Wang, Ying Li, Jun Yue, Lianshan Yan, and Qi Tian. 2024. Efficient Supervised Graph Embedding Hashing for large-scale cross-media retrieval. *Pattern Recognition* 145 (2024), 109934. doi:10.1016/j.patcog.2023.109934

[41] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. 2019. A Fourier Perspective on Model Robustness in Computer Vision. 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. https://proceedings.neurips.cc/paper_files/paper/2019/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf

[42] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. 2020. Central Similarity Quantization for Efficient Image and Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Xu Yuan, Zheng Zhang, Xunguang Wang, and Lin Wu. 2023. Semantic-Aware Adversarial Training for Reliable Deep Hashing Retrieval. *IEEE Transactions on Information Forensics and Security* 18 (2023), 4681–4694. doi:10.1109/TIFS.2023.3297791

[44] Hanyun Zhang, Yihua Chen, Xiaoping Liang, Lv Chen, and Zhenjun Tang. 2025. HGNet: Hash Generation Network Guided by High Frequency Information for Fine-Grained Image Retrieval. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10889893

[45] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R. Lyu. 2023. Improving the Transferability of Adversarial Samples by Path-Augmented Method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8173–8182.

[46] Shu Zhao, Tan Yu, Xiaoshuai Hao, Wenchao Ma, and Vijaykrishnan Narayanan. 2025. KALAHash: Knowledge-Anchored Low-Resource Adaptation for Deep Hashing. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 10 (Apr. 2025), 10465–10473. doi:10.1609/aaai.v39i10.33136

[47] Fei Zhu, Huashan Chen, Wanqian Zhang, Lin Wang, Zheng Lin, and Bo Li. 2025. Two-Stage Adversarial Training for Deep Hashing via Representation Distillation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 854–863. doi:10.1145/3726302.3730103

[48] Lei Zhu, Tianshi Wang, Jingjing Li, Zheng Zhang, Jialie Shen, and Xinhua Wang. 2023. Efficient Query-based Black-box Attack against Cross-modal Hashing Retrieval. *ACM Trans. Inf. Syst.* 41, 3, Article 54 (Feb. 2023), 25 pages. doi:10.1145/3559758

[49] Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Qinghua Lu, Jun Shen, and Kim-Kwang Raymond Choo. 2023. Improving Adversarial Transferability via Frequency-based Stationary Point Search. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 3626–3635. doi:10.1145/3583780.3614927

[50] Qiang Zou, Shuli Cheng, and Jiayi Chen. 2025. PromptHash:Affinity-Prompted Collaborative Cross-Modal Learning for Adaptive Hashing Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19649–19658.