Spatial-Frequency Domain Complementary Learning for Robust Cross-Modal Hashing

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

Abstract

Deep cross-modal hashing has achieved remarkable success in cross-modal retrieval due to its fast retrieval speed and low storage cost, but it highly vulnerable to adversarial attacks. Mainstream defense methods rely on adversarial training, which often induces a robustness–standard performance trade-off, leading to the learning of only limited robust features and degraded standard performance. To address this issue, we propose a Spatial-Frequency Domain Complementary Learning (SFCL) framework to overcome these two challenges by: 1) exploiting the complementarity of spatial and frequency features to learn more comprehensive and adversarially robust features, addressing the limited robustness of existing defenses; 2) by supplementing frequency-domain information, it avoids the performance degradation commonly caused by adversarial training. Specifically, SFCL consists of two modules: a Spatial-Frequency Robust Gating (SFRG) module, which selects robust features and strengthens complementarity via a conditional mutual information-based loss; and a Robustness-Aware Feature Fusion (RAFF) module, which performs bidirectional feature interaction and fusion. Extensive experiments demonstrate significant robustness gains over existing state-of-the-art methods, along with improved standard performance.

CCS Concepts

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

Keywords

Cross-modal Retrieval, Adversarial Robustness, Deep Hashing

ACM Reference Format:
Gang Zhou, Shibiao Xu, and Xiaolong Zheng. 2026. Spatial-Frequency Domain Complementary Learning for Robust Cross-Modal Hashing. In *International Conference on Multimedia Retrieval (ICMR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3805622.3810833

*Corresponding author.

ICMR '26, Amsterdam, Netherlands

ACM ISBN 979-8-4007-2617-0/2026/06
https://doi.org/10.1145/3805622.3810833

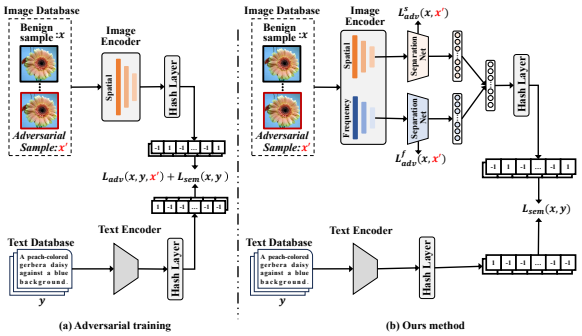

Figure 1: Comparison with existing frameworks.(a) Adversarial training couples semantic learning (L_{sem}) and adversarial learning (L_{adv}) in joint training; (b) Our method applies L_{adv} at lower layers via a separation network, while learning L_{sem} independently at higher layers.

1 Introduction

With the exponential growth of images, texts, audio and videos online, the demand to retrieve information in different modalities is rapidly increasing. Thus, cross-modal retrieval has become crucial for efficient semantic matching between modalities[21]. Among existing methods, Cross-Modal Hashing (CMH) is particularly effective due to its low storage cost and fast query speed. By mapping multimodal data into a unified binary Hamming space, CMH enables an efficient approximate nearest neighbor search, making it ideal for large-scale and real-time applications[12].

Recent studies have shown that deep cross-modal hashing models inherit the vulnerability of deep neural networks and are highly sensitive to adversarial attacks. Even imperceptible perturbations can significantly mislead hashing networks, causing them to retrieve incorrect and irrelevant results[7, 16, 22, 33]. In such attacks, malicious perturbations push the hash codes of queries away from semantically related items and pull them closer to irrelevant ones,

severely degrading retrieval accuracy[16]. Moreover, generative attack approaches can efficiently produce adversarial samples, further amplifying security threats[7]. Attackers can manipulate queries to evade detection or hijack search results, posing significant risks to real-world systems[22].

Among various defense methods, adversarial training (AT) is widely regarded as one of the most effective paradigms for improving the robustness of deep cross-modal hashing models[23, 24, 28]. Its core idea is to introduce adversarial samples during training to enhance model robustness against attacks. Previous studies have shown that adversarial samples originate from non-natural data distributions and contain numerous non-robust features[10]. AT compels the model to jointly learn non-robust features from adversarial samples and discriminative semantics from mixed features of natural and adversarial samples, thereby coupling semantic and non-robust feature learning. This coupling leads to limited robustness improvement and largely explains the degraded performance on clean samples. We argue that an ideal robust deep hashing model should learn robust features at lower layers while focusing on discriminative semantics at higher layers.

To address this issue, we propose a framework named Spatial-Frequency Domain Complementary Learning (SFCL). This design is motivated by previous studies showing that introducing complementary features can significantly enhance model robustness against adversarial attacks[4, 30, 31]. As shown in Figure 1, SFCL is designed to promote the learning of robust features at lower layers through spatial-frequency feature complementarity, and to perform semantic learning at higher layers based on the learned robust features. Specifically, It consists of two modules: the Spatial-Frequency Robust Gating (SFRG) module, which identifies robust features in each domain and encourages complementary feature learning via a conditional mutual information loss; and the Robustness-Aware Feature Fusion (RAFF) module, which performs bidirectional interactions between robust spatial and frequency features to produce a unified robust representation for high-level semantic learning.

In summary, our main contributions are:

- We find that cross-domain (spatial-frequency) attacks exhibit low transferability in hashing models, highlighting the potential of leveraging the complementarity between spatial and frequency features to enhance adversarial robustness.
- We propose a framework to enhance the adversarial robustness of cross-modal hashing models, and to the best of our knowledge, this is the *first attempt to* exploit frequency-domain features for this purpose.
- Extensive experiments demonstrate that our method significantly improves adversarial robustness, surpassing existing state-of-the-art defenses without sacrificing standard performance.

2 Related Work

2.1 Deep Cross-Modal Hashing

Deep learning has significantly advanced cross-modal retrieval by learning unified hash representations for multimodal data. In 2017, DCMH[12] was the first work to integrate feature extraction and hash coding into end-to-end frameworks. Subsequent works enriched this paradigm; SSAH[13] introduced adversarial self-supervision that aligns the distributions of different modalities via a pair of semantic networks and a modality discriminator. Recently, PromptHash[34] proposed affinity-prompted collaborative learning that fuses text prompts with adaptive gating and hierarchical contrast, attaining new SOTA in cross-modal hashing retrieval. Despite strong retrieval performance, these models are still vulnerable to adversarial attacks.

2.2 Robust Cross-Modal Hashing Learning

Cross-modal hashing models have increasingly drawn attention for their vulnerability to adversarial attacks. Robust Cross-Modal Hashing Learning aims to strengthen retrieval robustness in such models. ATRDH[24] formulates targeted attacks as point-to-set optimization with an anchor code and employs code-guided adversarial training to learn robust hashes and resist targeted attacks. CgAT[23] uses a semantic center code with a min-max scheme, generating worst-case adversarial examples by maximizing and then minimizing the Hamming distance to enhance robustness. CRDAT[32] employs a two-stage pipeline, distilling a robust teacher's adversarial representations to a student model to enhance defense without significant performance loss. Adversarial training methods partially enhance robustness, they often degrade retrieval accuracy on clean samples, revealing an unresolved trade-off between robustness and standard performance.

2.3 Complementary Feature Learning for Robustness

Complementary feature learning improving adversarial robustness by integrating multiple information sources or modeling signals. GCE[4] uses a guided complement entropy loss to boost true-class confidence and suppress incorrect-class probabilities, widening inter-class margins and implicitly leveraging complementary class-level cues to enhance robustness without extra training. NAMID[31] maximizes mutual information with clean semantic patterns while minimizing that with adversarial noise, guiding the model to focus on complementary robust features over perturbation-sensitive signals. PS-DFSI[30] ddynamically selects informative frequency bands and fuses them with spatial cues, leveraging spatial–frequency complementarity to enhance representation robustness under occlusion and noise.

3 Motivation

Based on the robust feature theory[10], deep networks rely on non-robust features, highly discriminative yet fragile patterns. Standard Adversarial Training (AT) gains robustness by suppressing these features, but inevitably degrades clean accuracy by discarding fine-grained semantic details crucial for retrieval[3, 20].

To address this issue, inspired by prior findings[11, 29] that spatial- and frequency-domain gradients exhibit markedly different spectral behaviors, we exploit integrate representations from distinct domain subspaces to enhance cross-domain robustness. Experimental results (Fig. 2) further support this insight: the limited adversarial transferability between the two architectures suggests that they rely on different spectral subspaces.

We argue that the robustness–accuracy trade-off can be effectively mitigated through a dual-domain perspective. Spatial–frequency

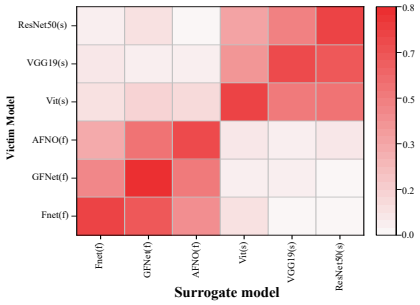

Figure 2: Comparison of cross-domain transferability of adversarial attacks. The heatmap indicates the percentage drop in performance after transfer attacks, where s denotes the spatial-domain backbone and f denotes the frequency-domain backbone.

complementarity provides dual advantages: **Robustness (Gradient Discrepancy).** Spatial and frequency views exhibit markedly different gradient characteristics, leading to weak cross-view transferability of adversarial perturbations. As a result, attacks optimized for one view are less effective against the other, naturally mitigating perturbation effects. **Accuracy (Information Integrity).** Instead of suppressing discriminative spatial textures as in standard AT, we preserve them while stabilizing representations with global frequency structures. This maintains fine-grained semantic details for precision, while leveraging frequency-invariant components as a complementary safeguard. Consequently, unlike AT, our approach achieves a synergistic effect where robustness and accuracy reinforce rather than compromise each other.

4 Methodology

In this section, we provide a detailed introduction to our proposed SFCL. The overall framework of SFCL is presented in Figure 3.

4.1 Preliminaries

Notation. Let the cross-modal retrieval dataset be $U = \{(x_i, y_i, l_i)\}_{i=1}^{N}$, where x_i is the input of the image modality, y_i is the text description corresponding to x_i, and $l_i = [l_{i1}, l_{i2}, \ldots, l_{iC}]$ is a multi-label vector of length C with $l_{ij} \in \{0, 1\}$. $l_{ij} = 1$ indicates that the instance (x_i, y_i) belongs to the j-th category. Two instances are semantically related if they share at least one label, i.e.$(l_i^\top l_j) > 0$; otherwise, they are irrelevant.

Deep Cross-Modal Hashing. Deep cross-modal hashing first extracts modality-specific features $\{f_i^x, f_i^y\}$ from input samples and then learns to map features from different modalities into a shared hash space, preserving semantic consistency. The overall pipeline can be formalized as a hash function $H : H : \{x_i, y_i\} \rightarrow \{b_i^x, b_i^y\}$, where $b_i^* \in \{-1, +1\}^K$ is the binary hash code for the modality $*$ ($* \in \{x, y\}$), and K is the length of the hash code. Semantically related instances are expected to have small Hamming distances, while irrelevant ones should remain far apart. The Hamming distance between two hash codes $b_i, b_j \in \{-1, +1\}^K$ is calculated as:

$$\mathrm{Ham}(b_i, b_j) = \frac{1}{2}\left(K - b_i^\top b_j\right). \tag{1}$$

Robustness in Deep Cross-Modal Hashing. Deep cross-modal hashing is vulnerable to adversarial perturbations δ deliberately added to the input, which may mislead the retrieval process. This work focuses on adversarial robustness in the image modality. A typical adversarial example is generated as $x_i' = x_i + \delta_x$, subject to $\|\delta_x\|_p \leq \epsilon$, where ϵ controls the magnitude of the perturbation. The attack is designed to fool the retrieval model into returning semantically irrelevant samples for a given query by simultaneously pushing away related samples and pulling closer irrelevant ones:

$$\max_{\delta_x} \left[\mathrm{Ham}(b_i^{x'}, b_j^y) - \mathrm{Ham}(b_i^{x'}, b_k^y)\right], \tag{2}$$

where $(l_i^\top l_j) > 0$ and $(l_i^\top l_k) = 0$. The goal of robust cross-modal hashing is to learn hash functions that produce stable and reliable hash codes even under such adversarial attacks, ensuring that semantically related samples remain close in the Hamming space.

4.2 Spatial-Frequency Robust Gating (SFRG)

Adversarial attacks exhibit distinct patterns across the spatial and frequency domains, degrading local textures in the spatial domain while disrupting global spectral structures in the frequency domain. Inspired by these behaviors and prior work on robust feature separation[14], the proposed SFRG independently filters robust features within each domain, thereby enhancing overall model robustness.

Given an input image x_i, the spatial and frequency features are extracted by the backbone networks $\mathcal{B}^s$ and $\mathcal{B}^f$, respectively, where $f_i^d = \mathcal{B}^d(x_i)$, $f_i^d \in \mathcal{R}^{D_d}$, and $d \in \{s, f\}$ denotes the spatial domain (s) or frequency (f) domain throughout this paper. To identify robust and non-robust features, two lightweight separation networks (S^d) are used to generate robustness scores such as $r^d = S^d(f_i^d)$. A binary dimension-wise mask $r^d \in \{0, 1\}^{D_d}$ would ideally separate robust from non-robust features, but it is non-differentiable; thus, the Gumbel-Softmax is adopted to yield differentiable soft masks approximating binary selection:

$$m^d = \frac{\exp\left(\frac{\log(\hat{r}^d) + g_1}{\tau}\right)}{\exp\left(\frac{\log(\hat{r}^d) + g_1}{\tau}\right) + \exp\left(\frac{\log(1 - \hat{r}^d) + g_2}{\tau}\right)}, \tag{3}$$

where $\hat{r}^d = \sigma(r^d)$ is the normalized robustness score, $\sigma(\cdot)$ is a sigmoid function, g_1 and g_2 are Gumbel random variables sampled as $g = -\log(-\log(u))$ with $u \sim \mathcal{U}(0, 1)$, and τ is the temperature controlling the sharpness of the soft approximation. During inference, we fix g_1 and g_2 as $-\log(-\log(u_c))$, where $u_c \in \mathcal{R}^{D_d}$ is a constant vector with each element set to 0.5, providing a deterministic approximation of the Gumbel noise and producing deterministic output across all characteristic dimensions.

Applying these masks, we obtain robust and non-robust filtered features f_i^{d+} and f_i^{d-}:

$$\left(f_i^{d+},\ f_i^{d-}\right) = \left(m^d \odot f_i^d,\ (1 - m^d) \odot f_i^d\right) \tag{4}$$

Inspired by prior studies showing that robustness benefits from complementary representations, we explicitly encourage cross-domain complementarity via a symmetric conditional contrastive

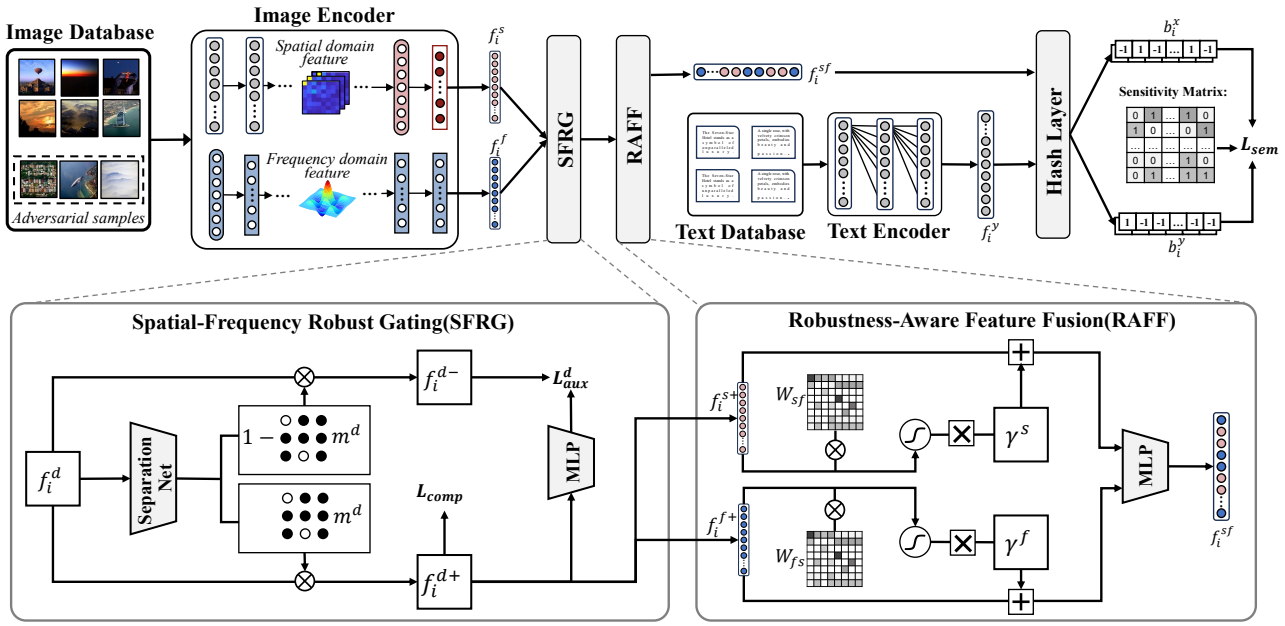

Figure 3: Overview of the proposed SFCL framework. The image encoder extracts spatial and frequency features in parallel. SFRG separates complementary components, and RAFF fuses cross-domain features. The fused representation is projected into the hash space and optimized with adversarial training, improving robustness for cross-modal retrieval.

objective. Instead of directly computing conditional mutual information (CMI), we approximate it using an InfoNCE-based variational lower bound.

Specifically, the conditional dependence $I(f_i^{s+}; l \mid f_i^{f+})$ is estimated by contrasting the true label l_i against other labels within the mini-batch:

$$\widehat{I}(f_i^{s+}; l \mid f_i^{f+}) = \log \frac{\exp\left(g_s(f_i^{s+}, f_i^{f+}, l_i)\right)}{\sum_{l_j \in \mathcal{B}} \exp\left(g_s(f_i^{s+}, f_i^{f+}, l_j)\right)}, \tag{5}$$

where $\mathcal{B}$ denotes the batch label set and $g_s(\cdot)$ is a lightweight MLP scorer. The term $\widehat{I}(f_i^{f+}; l \mid f_i^{s+})$ is defined symmetrically. To suppress redundancy, $I(f_i^{s+}; f_i^{f+})$ is also approximated via a standard InfoNCE objective.

The complementarity loss is defined as:

$$\mathcal{L}_{\text{comp}} = -\widehat{I}(f_i^{s+}; l \mid f_i^{f+}) - \widehat{I}(f_i^{f+}; l \mid f_i^{s+}) + \alpha\, \widehat{I}(f_i^{s+}; f_i^{f+}), \tag{6}$$

where α balances complementarity and redundancy suppression.

To explicitly guide the Separation Net in learning robustness scores, two lightweight auxiliary MLPs h^s and h^f are introduced for the spatial and frequency domains. Each MLP outputs robustness probabilities per class: $p_i^{d+} = h^d(f_i^{d+})$ and $p_i^{d-} = h^d(f_i^{d-})$, where p_{ic}^{d+}, p_{ic}^{d-} are prediction scores, $l_{ic} \in \{0, 1\}$ is the ground truth label and $l'_{ic} \in \{0, 1\}$ is the adversarial label predicted from x'_i. A binary cross-entropy loss is adopted to encourage f_i^{d+} to align with true labels and f_i^{d-} with adversarial patterns:

$$\mathcal{L}_{aux}^{d} = \sum_{c=1}^{C} \left(l_{ic} \log p_{ic}^{d+} + l'_{ic} \log p_{ic}^{d-} \right), \tag{7}$$

where adversarial samples x'_i used for generating l'_{ic} are crafted via the standard PGD target attack[2]. This supervision encourages the Separation Net to assign higher robustness scores to features predictive of true labels while suppressing those correlated with adversarial noise, thereby effectively distinguishing between robust and non-robust feature activations.

4.3 Robustness-Aware Feature Fusion (RAFF)

SFRG effectively filters robust and non-robust features within each domain, however, the resulting spatial and frequency features remain independent. To obtain a more discriminative and adversarially robust representation, we propose RAFF, which performs bidirectional embedded interactions between spatial and frequency features in a unified latent space. This allows each domain to adaptively complement the other's robust information while preserving its own robustness.

Specifically, the spatial robust feature is updated under the modulation of the frequency robust feature as:

$$\tilde{f}_i^{s} = f_i^{s+} + \gamma^{s} \cdot \sigma\left(W_{sf} f_i^{f+}\right), \tag{8}$$

and the frequency robust feature is updated under the modulation of the spatial robust feature as:

$$\tilde{f}_i^{f} = f_i^{f+} + \gamma^{f} \cdot \sigma\left(W_{fs} f_i^{s+}\right). \tag{9}$$

This bidirectional interaction is complementary: the frequency feature injects stable global spectral cues into the spatial domain

to retain fine-grained details while reducing vulnerability to local perturbations, whereas the spatial feature supplements fine-grained textures to improve the frequency feature's intra-class discrimination without compromising its global robustness.

Here, $W_{sf} \in \mathcal{R}^{D_s \times D_f}$ and $W_{fs} \in \mathcal{R}^{D_f \times D_s}$ are learnable cross-domain projection matrices that map one domain's features into the latent space of the other; γ^s and γ^f are learnable scalars controlling the global interaction strength. This design ensures that robust features from one domain can selectively refine those of the other without compromising intra-domain robustness.

Finally, the unified robust representation is generated in the embedded space as:

$$f_i^{sf} = \mathrm{MLP}\left([\tilde{f}_i^s \parallel \tilde{f}_i^f]\right), \tag{10}$$

where MLP denotes a lightweight feed-forward network consisting of two fully connected layers with GELU nonlinear activation, which maps the bidirectionally interacted robust features into a shared representation space. The unified robust representation f_i^{sf} obtained from RAFF is used as the input to the HashLayer in the subsequent hashing stage.

4.4 Cross-Modal Hashing Learning

The objective of cross-modal hashing is to project heterogeneous features into a unified Hamming space, where the Hamming distance between hash codes preserves semantic similarity. Following common practice in deep cross-modal hashing (DCMH), a lightweight HashLayer is employed to map image and text features into a shared semantic space:$h_i^* = \mathrm{HashLayer}(f_i^*), \quad * \in \{x, y\}$. The HashLayer consists of two fully connected layer with a $\tanh(\cdot)$ activation, encouraging outputs to approach binary values $\{-1, 1\}$.

To ensure samples with the same semantic label are mapped to nearby Hamming codes, an asymmetric negative log-likelihood pairwise loss is adopted. The intra-modal similarity is measured as $\Omega_{ij}^* = \frac{1}{2}(h_i^*)^\top h_j^*$, and the corresponding loss is:

$$\mathcal{L}_{\text{intra}} = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\Omega_{ij}^* - \log(1 + e^{\Omega_{ij}^*})\right). \tag{11}$$

Semantic consistency across modalities is maintained by minimizing the following inter-modal loss:

$$\begin{aligned}\mathcal{L}_{inter} = -\frac{1}{MN}\Bigg(&\sum_{i=1}^{N}\sum_{j=1}^{M}\left(S_{ij}\Theta_{ij} - \log(1 + e^{\Theta_{ij}})\right) \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{M}\left(S_{ij}\Phi_{ij} - \log(1 + e^{\Phi_{ij}})\right)\Bigg),\end{aligned} \tag{12}$$

where $\Theta_{ij} = \frac{1}{2}(h_i^t)^\top h_j^v$ and $\Phi_{ij} = \frac{1}{2}(h_i^v)^\top h_j^t$.

To reduce the discrepancy between continuous hash representations and discrete binary codes, a quantization loss is introduced:

$$\mathcal{L}_{\text{quan}} = \sum_{i=1}^{N} \|h_i^* - \mathrm{sign}(h_i^*)\|_2^2, \tag{13}$$

where $\mathrm{sign}(\cdot)$ maps real-valued elements to $\{-1, 1\}$.

4.5 Objective Function

The loss function consists of two parts: an adversarial loss and a semantic loss. The adversarial part includes the complementarity loss $\mathcal{L}_{\text{comp}}$ and the auxiliary supervision loss $\mathcal{L}_{\text{aux}}$, while the semantic part $\mathcal{L}_{\text{sem}}$ includes $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{inter}}$, and $\mathcal{L}_{\text{quan}}$, which follow the same internal hyperparameter settings as in DCMH [12]. The total objective is formulated as:

$$\mathcal{L}_{\text{total}} = \underbrace{\lambda_c\mathcal{L}_{\text{comp}} + \lambda_a\mathcal{L}_{\text{aux}}}_{\mathcal{L}_{\text{adv}}} + \lambda_s\mathcal{L}_{\text{sem}}. \tag{14}$$

where λ_c, λ_a, and λ_s are trade-off hyper-parameters.

5 Experiments

5.1 Datasets

To validate the effectiveness and generalization of our method, we conduct experiments on three widely used cross-modal retrieval benchmarks. These datasets vary in scale, label distribution, and annotation forms, providing a comprehensive evaluation setting. **MIRFLICKR-25K**[9] contains 25,000 image-text pairs annotated with 24 semantic categories. We randomly select 2,000 pairs as the query set, 5,000 for training, and use the remaining pairs as the retrieval database. **NUS-WIDE**[5] consists of 269,648 image-tag pairs 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**[17] provides 123,287 image-text 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.

5.2 Evaluation

To evaluate the defense performance, we use MAP to measure the retrieval effectiveness of the model under untargeted attacks, and adopt t-MAP[2] for evaluation under targeted attacks. Both MAP and t-MAP are calculated as MAP@ALL, considering all retrieved results from the database. Since this work focuses on enhancing the robustness of the visual modality, we generate adversarial samples only for images during evaluation, and thus report the robustness performance on image-to-text (I2T) and image-to-image (I2I) retrieval tasks.

5.3 Baseline Methods and Implementation Details

We adopt ATRDH[24], CgAT[23], CRDAT[32], FPAD[27] as baseline adversarial training methods. For a fair comparison, all methods adopt DCMH [12] as the default deep hashing objective and use AlexNet [15] and BERT-base [6] as the image and text backbones, respectively. Building upon this setting, our method further incorporates GFNet [19], a frequency-domain Vision Transformer, alongside the CNN-based AlexNet to extract complementary frequency-domain representations. This design is motivated by prior studies[3] showing that the frequency-domain complementarity between CNNs and Vision Transformers can enhance model robustness. To comprehensively evaluate defense performance, we

Table 1: Robustness Comparison under Untargeted Attacks. The retrieval task is Image-to-Text (I2T), and the attack strength for all methods is set to 8/255. The best results in each setting are in bold.

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	DCMH	0.7273	0.7236	0.7387	0.5229	0.5986	0.6197	0.5174	0.5318	0.5468
	ATRDH	0.6618	0.6863	0.6835	0.4601	0.5391	0.5388	0.4308	0.4458	0.4757
	CgAT	0.6596	0.6744	0.6859	0.4542	0.5510	0.5647	0.4427	0.4505	0.4723
	CRDAT	0.6925	0.7015	0.7136	0.4958	0.5842	0.5964	0.4852	0.4966	0.5142
	FPAD	0.7136	0.7248	0.7324	0.5212	0.6085	0.6158	0.5214	0.5358	0.5485
	SFCL (Ours)	**0.7352**	**0.7386**	**0.7412**	**0.5425**	**0.6254**	**0.6325**	**0.5453**	**0.5652**	**0.5725**
HAG	DCMH	0.2048	0.2341	0.2260	0.0636	0.0776	0.0891	0.0891	0.0951	0.1084
	ATRDH	0.3144	0.3269	0.3249	0.2099	0.1774	0.1635	0.1686	0.1608	0.1234
	CgAT	0.4440	0.3891	0.4111	0.3735	0.3163	0.3228	0.1967	0.1415	0.1277
	CRDAT	0.4725	0.4812	0.4963	0.4058	0.3845	0.3952	0.2135	0.2248	0.2285
	FPAD	0.4968	0.5052	0.5184	0.4285	0.4062	0.4132	0.2452	0.2536	0.2574
	SFCL (Ours)	**0.5125**	**0.5236**	**0.5374**	**0.4496**	**0.4289**	**0.4342**	**0.2635**	**0.2741**	**0.2725**
SDHA	DCMH	0.1596	0.1528	0.1624	0.1025	0.0986	0.0985	0.0781	0.0878	0.0847
	ATRDH	0.2368	0.2065	0.1920	0.0894	0.1135	0.1100	0.1500	0.1496	0.1068
	CgAT	0.4125	0.4085	0.4158	0.4811	0.4847	0.4714	0.2253	0.2247	0.2189
	CRDAT	0.4525	0.4658	0.4752	0.4865	0.4912	0.5017	0.3265	0.3364	0.3374
	FPAD	0.5412	0.5326	0.5487	0.5014	0.5186	0.5247	0.3452	0.3465	0.3487
	SFCL (Ours)	**0.5523**	**0.5689**	**0.5748**	**0.5214**	**0.5325**	**0.5412**	**0.3514**	**0.3655**	**0.3578**
CgAT	DCMH	0.1338	0.1488	0.1497	0.0923	0.0965	0.1287	0.0851	0.1184	0.1387
	ATRDH	0.3326	0.3209	0.2967	0.2781	0.2897	0.2904	0.2426	0.2523	0.2224
	CgAT	0.4187	0.4136	0.4257	0.3957	0.3997	0.4137	0.2213	0.2141	0.2342
	CRDAT	0.4825	0.4965	0.5012	0.4952	0.4998	0.5034	0.2685	0.2742	0.2812
	FPAD	0.5214	0.5348	0.5425	0.5065	0.5142	0.5218	0.3125	0.3258	0.3346
	SFCL (Ours)	**0.5562**	**0.5485**	**0.5591**	**0.5632**	**0.5784**	**0.5798**	**0.3145**	**0.3252**	**0.3362**

test against a range of adversarial attacks, including three untargeted attacks (HAG[25], SDHA[18], and CgAT[23]) and four targeted attacks (P2P[2], DHTA[2], THA[24], and TA-DCH[22]).

The SFRG module employs lightweight separation networks (hidden size 256) with Gumbel-Softmax ($\tau = 1.0$) for differentiable mask learning. The RAFF module performs bidirectional cross-domain interaction via learnable projection matrices followed by a two-layer MLP (hidden size 512). We train the model end-to-end in a single stage with adversarial warm-up. PGD (10 steps, $\epsilon = 8/255$) is used for adversarial sample generation. The total loss combines complementarity, auxiliary, and semantic losses with weights $\lambda_c = 0.3$, $\lambda_a = 1.0$, and $\lambda_s = 3.0$. Optimization uses Adam with an initial learning rate of 1×10^{-4} and cosine decay scheduling. All experiments are conducted using PyTorch on a single NVIDIA RTX 4090 GPU.

5.4 Defense against White-box Adversarial Attacks

Performance against untargeted attacks. Table 1 presents the defense performance comparison under three untargeted adversarial attacks on the image-to-text (I2T) retrieval task. A higher MAP value after attack indicates that the model better preserves retrieval performance under adversarial attacks, reflecting stronger robustness. "Clean" denotes the model's standard performance when evaluated on clean samples. The results show that our method consistently achieves the best robustness across all three datasets compared to ATRDH and CgAT, with a significant performance margin. Moreover, compared to the original DCMH method without any defense mechanism, our method also achieves slightly better performance on clean data. These findings indicate that decoupling adversarial learning from semantic learning enables our method to more effectively capture discriminative semantic features, enhancing both retrieval performance and robustness.

Performance against targeted attacks. Table 2 presents the robustness comparison under targeted attack scenarios, including three representative targeted attacks: P2P, DHTA, and THA. It is important to note that, under this setting, a lower t-MAP value after attack indicates stronger robustness, which contrasts with the untargeted attack scenario where a higher MAP implies better robustness. The results show that our method maintains superior robustness under targeted attacks, achieving the lowest t-MAP compared to the other two adversarial training methods.

5.5 Defense against Black-box Adversarial Attacks

We evaluate the robustness of the model under black-box adversarial attack scenarios. Specifically, we adopt the DCMH method with three different image backbones to train clean deep hashing

Table 2: Robustness Comparison under Targeted Attacks. The retrieval task is Image-to-Text (I2T), 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	DCMH	0.8021	0.8101	0.8169	0.6225	0.6536	0.6625	0.6596	0.6789	0.6974
	ATRDH	0.6821	0.7052	0.7010	0.5486	0.5627	0.5581	0.4313	0.4520	0.4692
	CgAT	0.6512	0.6677	0.6605	0.5114	0.5362	0.5413	0.4135	0.4261	0.4333
	CRDAT	0.6385	0.6492	0.6588	0.5091	0.5265	0.5377	0.3967	0.4102	0.4215
	FPAD	0.6252	0.6368	0.6471	0.5055	0.5182	0.5294	0.3812	0.3925	0.4036
	SFCL (Ours)	**0.6132**	**0.6154**	**0.6241**	**0.5024**	**0.5029**	**0.5047**	**0.3685**	**0.3587**	**0.3624**
DHTA	DCMH	0.8214	0.8367	0.8480	0.6663	0.6790	0.6726	0.6666	0.6943	0.6873
	ATRDH	0.6693	0.6826	0.6785	0.5411	0.5533	0.5478	0.4167	0.4325	0.4491
	CgAT	0.6425	0.6572	0.6484	0.5097	0.5224	0.5292	0.4089	0.4158	0.4182
	CRDAT	0.6355	0.6485	0.6455	0.4912	0.5041	0.5155	0.4015	0.4105	0.4155
	FPAD	0.6285	0.6395	0.6405	0.4725	0.4852	0.4968	0.3925	0.4035	0.4085
	SFCL (Ours)	**0.6235**	**0.6342**	**0.6358**	**0.4528**	**0.4659**	**0.4712**	**0.3869**	**0.3748**	**0.3957**
THA	DCMH	0.8521	0.8703	0.8773	0.6771	0.7067	0.7191	0.7896	0.8266	0.8315
	ATRDH	0.7023	0.7195	0.7252	0.5612	0.5756	0.5813	0.4722	0.4893	0.5021
	CgAT	0.6851	0.7029	0.7118	0.5382	0.5523	0.5641	0.4593	0.4750	0.4817
	CRDAT	0.6615	0.6748	0.6892	0.5245	0.5385	0.5512	0.4405	0.4552	0.4678
	FPAD	0.6324	0.6455	0.6582	0.5125	0.5252	0.5368	0.4215	0.4342	0.4465
	SFCL (Ours)	**0.6058**	**0.6125**	**0.6254**	**0.5014**	**0.5189**	**0.5145**	**0.4058**	**0.4196**	**0.4165**

Table 3: Robustness comparison under black-box transfer targeted attacks. The base model is DCMH, evaluated on the MIRFLICKR-25K dataset with 64-bit hash codes.

Defense	Inception-v3	VGG19	ResNet50
DCMH	0.8525	0.6524	0.6587
ATRDH	0.5685	0.4986	0.4659
CgAT	0.5863	0.5145	0.4598
CRDAT	0.5312	0.3857	0.3924
FPAD	0.4935	0.3024	0.3158
SFCL(Ours)	**0.4526**	**0.2135**	**0.2596**

models, which are used as surrogate models for the TA-DCH attack. As shown in Table 3, our method exhibits minimal performance degradation under black-box attacks, consistently outperforming the other two adversarial training methods. Since black-box settings better reflect real-world adversarial conditions, these results underscore the practical significance of the robustness achieved by our model.

5.6 Ablation Study

We conduct ablation studies on the MIRFLICKR-25K dataset to assess the contribution of each SFCL component to both robustness and standard performance. Evaluations are performed under the CgAT attack using 64-bit hash codes in an image-to-text (I2T) retrieval task. SFCL comprises three components: the Separation Network (S), Complementary Loss (C), and Fusion Module (F), where S+C forms the SFRG module and F corresponds to RAFF. "w/o" denotes the exclusion of a component, with F replaced by a simple MLP when omitted. Figure 4 reveals three main findings: 1) S notably enhances robustness, though with slight performance degradation; 2) C improves both robustness and accuracy, suggesting that the complementary loss helps the model learn more discriminative and robust features; 3) All components work together, removing any one degrades robustness, confirming the effectiveness of SFCL's design.

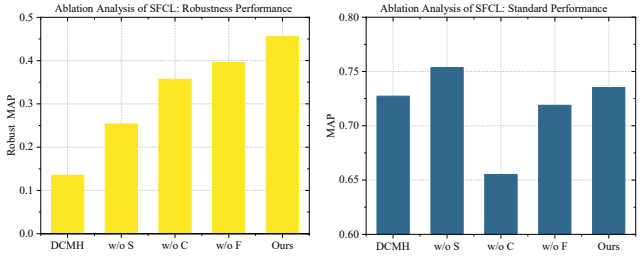

Figure 4: Ablation Study of SFCL Components.

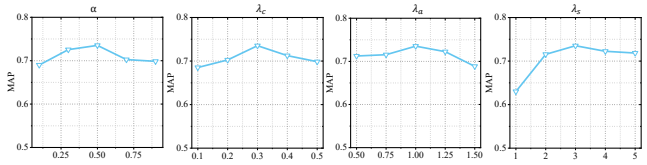

Figure 5: Comparison of MAP scores on the MIRFLICKR-25K dataset with different parameter configurations.

Table 4: Efficiency comparison on NUS-WIDE. FLOPs are measured for the image encoder (224×224). Training time is reported per epoch under PGD-10 adversarial training. Inference latency is measured per image query (ms), excluding offline database encoding.

Method	Params	FLOPs (G)	Train (min)	Infer (ms)
DCMH	62.0	0.72	4.5	12.3
ATRDH	62.0	0.72	12.8	12.6
CgAT	62.0	0.72	13.4	12.9
CRDAT	62.0	0.72	15.1	13.1
FPAD	66.0	0.78	14.2	14.2
Ours (SFCL)	68.0	0.88	18.8	18.0

5.7 Parameter Sensitivity

To assess the sensitivity of hyperparameters, we conduct a comprehensive parameter analysis of the proposed method on the MIRFLICKR-25K dataset with a 64-bit hash code length. The four hyperparameters under investigation include α (defined in Eq. 6) and λ_c, λ_a, and λ_s (defined in Eq. 14). As shown in Figure 5, the model achieves the best performance when $\alpha = 0.5$, $\lambda_c = 0.3$, $\lambda_a = 1.0$, and $\lambda_s = 3.0$.

5.8 Efficiency Analysis

Table 4 reports parameter size, FLOPs, training cost, and inference latency on NUS-WIDE. Unless otherwise noted, Params and FLOPs are measured for the image encoder at 224×224 resolution, consistent with image-query inference. The text encoder (BERT-base) is shared across methods and used only for training or offline encoding, thus excluded from comparison. Training time is measured per epoch under PGD-10 adversarial optimization, and inference latency is reported per image query, excluding offline database encoding.

Parameters and FLOPs. DCMH, ATRDH, CgAT, and CRDAT use the same AlexNet backbone and hashing head, resulting in identical parameter size and FLOPs. FPAD introduces additional prototype modules, moderately increasing computation. Our method adds a lightweight frequency branch (GFNet-Ti) and two small fusion modules (SFRG and RAFF), incurring a controlled overhead of +6.0M parameters and +0.16G FLOPs over DCMH while remaining efficient.

Training cost. DCMH is the fastest due to the absence of adversarial optimization. PGD-based methods (ATRDH, CgAT) significantly increase per-epoch time because of repeated forward–backward passes. CRDAT further adds teacher-forward computation for distillation, and FPAD introduces prototype alignment operations. Although our method incorporates a frequency branch and fusion modules, its per-epoch cost remains comparable to other advanced defenses.

Inference efficiency. During inference, adversarial example generation is disabled, so latency differences mainly arise from image-encoder complexity. Single-backbone methods (DCMH, ATRDH, CgAT, CRDAT) exhibit similar speed, while FPAD incurs additional prototype computation. Our method performs one extra forward pass through the lightweight frequency branch and fusion, adding only +5.7 ms over DCMH, with Hamming-space retrieval efficiency fully preserved.

Table 5: Robustness Improvement from SFCL under Untargeted CgAT Attacks.

Methods	**w/o SFCL**	**SFCL Applied**
UCCH	0.3835	0.6239
DGCPN	0.3125	0.7321
DADH	0.3354	0.7635

5.9 Universality on Other Deep Cross-Modal Hashing

We evaluate the universality of SFCL on different deep cross-modal hashing methods. By replacing the original DCMH loss with UCCH[8], DGCPN[26], and DADH[1], we construct corresponding SFCL variants. As shown in Table 5, SFCL consistently improves robustness across all methods, demonstrating strong universality and adaptability.

6 Conclusion

We proposed Spatial-Frequency Domain Complementary Learning (SFCL) to enhance the adversarial robustness of deep cross-modal hashing. First, we separate robust and non-robust features and introduce a conditional mutual information-based objective to encourage complementary spatial-frequency representations, enabling the model to focus on robust cues while suppressing perturbation-sensitive patterns. We then integrate spatial and frequency features into a unified representation that leverages stable spectral information and fine-grained spatial details, improving both robustness and retrieval accuracy. By modeling heterogeneous spectral subspaces, SFCL reduces reliance on vulnerable representation directions and mitigates adversarial perturbations aligned with dominant gradients. Extensive experiments demonstrate that SFCL achieves significant robustness gains without sacrificing standard performance, validating the effectiveness of complementary multi-domain learning for cross-modal 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.

References

[1] Cong Bai, Chao Zeng, Qing Ma, Jinglin Zhang, and Shengyong Chen. 2020. Deep Adversarial Discrete Hashing for Cross-Modal Retrieval. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (Dublin, Ireland) *(ICMR '20)*. Association for Computing Machinery, New York, NY, USA, 525–531. doi:10.1145/3372278.3390711

[2] 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.

[3] Qingwen Bu, Dong Huang, and Heming Cui. 2023. Towards Building More Robust Models with Frequency Bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4402–4411.

[4] Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. 2019. Improving Adversarial Robustness via Guided Complement Entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[5] 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

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[7] 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

[8] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. 2023. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3877–3889. doi:10.1109/TPAMI.2022.3177356

[9] 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

[10] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. 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/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf

[11] Shuai Jia, Chao Ma, Taiping Yao, Bangjie Yin, Shouhong Ding, and Xiaokang Yang. 2022. Exploring Frequency Adversarial Attacks for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4103–4112.

[12] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep Cross-Modal Hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Sheng Jin, Shangchen Zhou, Yao Liu, Chao Chen, Xiaoshuai Sun, Hongxun Yao, and Xian-Sheng Hua. 2020. SSAH: Semi-Supervised Adversarial Deep Hashing with Self-Paced Hard Sample Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (Apr. 2020), 11157–11164. doi:10.1609/aaai.v34i07.6773

[14] Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. 2023. Feature Separation and Recalibration for Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8183–8192.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[16] 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.

[17] 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.

[18] 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

[19] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. 2021. Global Filter Networks for Image Classification. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 980–993. https://proceedings.neurips.cc/paper_files/paper/2021/file/07e87c2f4fc7f7c96116d8e2a92790f5-Paper.pdf

[20] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. 2020. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *Proc. IEEE* 112, 11 (2024), 1716–1754. doi:10.1109/JPROC.2024.3525147

[22] 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

[23] 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

[24] 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

[25] 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

[26] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. 2021. Deep Graph-neighbor Coherence Preserving Network for Unsupervised Cross-modal Hashing. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 5 (May 2021), 4626–4634. doi:10.1609/aaai.v35i5.16592

[27] Zhongqing Yu, Xin Liu, Yiu-ming Cheung, Lei Zhu, Xing Xu, and Nannan Wang. 2025. FPAD: Fuzzy-Prototype-guided Adversarial Attack and Defense for Deep Cross-Modal Hashing. *IEEE Transactions on Circuits and Systems for Video Technology* (2025), 1–1. doi:10.1109/TCSVT.2025.3604033

[28] 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

[29] Liangqi Zhang, Yihao Luo, Haibo Shen, Tianjiang Wang, and Kate Larson. 2024. A fourier perspective of feature extraction and ad-versarial robustness. In *International Joint Conferences on Artificial Intelligence Organization*. 1715–1723.

[30] Qixian Zhang, Duoqian Miao, Qi Zhang, Cairong Zhao, Hongyun Zhang, Ye Sun, and Ruizhi Wang. 2025. Dynamic frequency selection and spatial interaction fusion for robust person search. *Information Fusion* 124 (2025), 103314. doi:10.1016/j.inffus.2025.103314

[31] Dawei Zhou, Nannan Wang, Xinbo Gao, Bo Han, Xiaoyu Wang, Yibing Zhan, and Tongliang Liu. 2022. Improving Adversarial Robustness via Mutual Information Estimation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 27338–27352. https://proceedings.mlr.press/v162/zhou22j.html

[32] 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

[33] 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

[34] 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.