DRFGD: Disentangled Representation-Focused Generative Defense for Attack-Tolerant Cross-Modal Hashing

Zhongqing Yu[1,2], Xin Liu[1,3,*], Yiu-ming Cheung[3], Zhikai Hu[3], Wentao Fan[4], Pan Zhou[5]

[1] Department of Computer Science, Huaqiao University
[2] Fujian Key Lab. of Big Data Intell. and Security & Xiamen CVPR Key Laboratory, Huaqiao University
[3] Department of Computer Science, Hong Kong Baptist University
[4] Department of Artificial Intelligence, Beijing Normal-Hong Kong Baptist University
[5] School of Cyber Science and Engineering, Huazhong University of Science and Technology
{zqyu, xliu}@hqu.edu.cn, {ymc, cszkhu}@comp.hkbu.edu.hk, wentaofan@bnbu.edu.cn, panzhou@hust.edu.cn

Abstract

With the widespread deployment of cross-modal retrieval in real-world scenarios, ensuring robustness against adversarial attacks is increasingly critical. Remarkably, deep cross-modal hashing is highly vulnerable to adversarial attacks due to its discrete nature and low-dimensional hash codes, while existing defense methods often fail to suppress perturbations embedded in vulnerable features and lack the capacity to model modality-specific structural differences, resulting in suboptimal adversarial robustness. To address these challenges, we propose a novel Disentangled Representation-Focused Generative Defense (DRFGD) framework for attack-tolerant cross-modal hashing. Without altering the structure of retrieval model, DRFGD defends against adversarial attacks by disentangling input representations into adversarial-robust and adversarial-vulnerable components, by an efficient dual-branch semantic-aware encoder. Guided by such disentangled robust features, an attack-tolerant generative module is seamlessly designed to synthesize semantically aligned and perturbation-resilient examples for robust adversarial training, thereby significantly promoting collaborative defense robustness to attackers. Consequently, the semantically consistent hash codes can be well obtained to enhance adversarial robustness in complex cross-modal attacking scenarios. Extensive experiments on public benchmarks demonstrate that DRFGD substantially improves retrieval robustness under various attacking scenarios, and shows its improved defense performance in comparison with the SOTA works.

Introduction

Deep cross-modal hashing (Yang et al. 2017; Wang et al. 2022; Hu et al. 2024) leverages deep neural networks to learn compact hash codes, enabling efficient retrieval and scalable storage for large-scale data. Despite significant advances, existing deep cross-modal hashing methods (Jiang and Li 2017; Sun et al. 2024) remain highly vulnerable to adversarial attacks (Li et al. 2019; Wang et al. 2017), where imperceptible perturbations to inputs (e.g., images) can mislead models into returning semantically irrelevant results, posing serious threats to retrieval systems (Xu et al. 2022). This underscores the need for robust cross-modal hashing defense mechanisms under adversarial attacks.

Existing defense strategies can be broadly classified into two categories: adversarial training (Xie et al. 2017; Madry et al. 2018) and input preprocessing (Zhou et al. 2021; Wang et al. 2024). Adversarial training enhances robustness by incorporating adversarial examples during the training stage. However, the introduction of additional training samples often necessitates modifications to the model architecture (Zhang et al. 2019) and incurs substantial computational overhead (Wong, Rice, and Kolter 2020; Zhou and Patel 2022). More critically, these adversarial examples may shift the learned representations of models, which frequently leads to degraded performance on clean inputs (Zhang et al. 2024), thereby undermining the stability of the model. In contrast, input preprocessing methods (Nie et al. 2022; Xiao et al. 2022) suppress adversarial perturbations before the data is fed into the model, offering better deployment flexibility and broader applicability across different architectures. For example, Tang and Zhang (2024) applied FGSM (Goodfellow, Shlens, and Szegedy 2014) to preprocess input images before model inference, effectively purifying pixel-level adversarial perturbations.

However, these preprocessing defenses also exhibit two limitations. First, existing methods (Xie et al. 2019; Zhou and Patel 2022; Zhang, Sun, and Zhao 2022) typically apply holistic defense strategies that aim to protect the entire representation without explicitly identifying which parts are more likely to be adversarially corrupted. However, deep representations typically contain Adversarial-Robust Features (ARF) that remains stable under perturbation and Adversarial-Vulnerable Featuress (AVF) that is more susceptible to adversarial noise and tend to mislead predictions. Treating all features uniformly leads to unnecessary focus on inherently robust parts, while allowing adversarial signals in vulnerable parts to persist, ultimately compromising overall defense effectiveness. Second, most prior works (Xiao et al. 2022; Chen et al. 2024; Wang et al. 2024) are confined to unimodal domains such as images and lack the capacity to model structural heterogeneity across modalities. This limitation hinders collaborative defense mechanisms and weakens robustness in cross-modal retrieval scenarios.

*Corresponding author.

To overcome these limitations, in this paper, we propose a **D**isentangled **R**epresentation-**F**ocused **G**enerative **D**efense framework, which explicitly targets both feature-level vulnerability and modality-level heterogeneity. Specifically, we first design a semantic-aware disentanglement module that separates inputs into ARF and AVF under semantic supervision. Only the ARF, which are semantically meaningful and perturbation-resistant, are retained for downstream retrieval, while AVF is explicitly discarded to eliminate potential attack signals. To further address defense inconsistencies across modalities, we incorporate an attack-tolerant generative module for both image and text modalities. This module can synthesize semantically aligned and perturbation-resilient examples from robust features, facilitating stable and consistent hash code learning across modalities. By jointly leveraging the semantic-aware disentanglement and attack-tolerant generative modules, the proposed DRFGD framework can effectively enhance adversarial robustness while preserving retrieval performance on clean data. The main contributions of this paper are summarized as follows:

- We propose a semantic-aware disentanglement module that separates input representations into adversarial-robust and adversarial-vulnerable features, enabling targeted defense against perturbations while preserving semantic consistency.
- We propose an attack-tolerant generative module that produces semantically consistent and robust examples for both modalities, significantly improving collaborative defense and retrieval performance in complex cross-modal scenarios.
- We conduct extensive experiments on three standard benchmarks, demonstrating that DRFGD consistently improves retrieval robustness under diverse adversarial attacks, while maintaining strong generalization.

Related Work

Deep Cross-modal Hashing Retrieval. Deep cross-modal hashing leverages deep neural networks to learn discriminative hash codes for efficient cross-modal retrieval. Based on label supervision availability, it is divided into unsupervised and supervised methods. Unsupervised methods aim to project original features into a shared Hamming space without explicit semantic guidance. DGCPN (Yu et al. 2021) preserves graph neighborhood consistency and adopts a semi-binary optimization strategy, while DCHUC (Tu et al. 2022) designs an iterative framework for unified hash learning. Supervised methods leverage semantic labels to boost retrieval performance. DCMH (Jiang and Li 2017) unifies feature learning and hash generation via joint optimization. CPAH (Xie et al. 2020) enhances semantic consistency and modality correlation. DADH (Bai et al. 2020a) incorporates generative adversarial networks into hashing learning. Recently, DCPH (Tu et al. 2023b) and DAPH (Tu et al. 2023a) further advance semantic hashing by using proxy networks to model category structures adaptively.

Adversarial Attack against Cross-modal Hashing. Adversarial attack inject imperceptible perturbations to mislead results (Szegedy et al. 2014). Unlike classification tasks where adversarial perturbations directly alter categorical outputs, cross-modal hashing attacks are more complex due to the need to disrupt cross-modal semantic similarity. Existing methods are categorized by the level of information accessible to the attacker into white-box and black-box attacks. White-box attacks such as CMLA (Li et al. 2019) and DACM (Li et al. 2020) assume full access to model parameters and gradients. This enables the generation of adversarial examples that maximize the semantic discrepancy in Hamming space. Such assumptions, however, often limit their applicability in real-world retrieval systems. In contrast, black-box attacks operate without knowledge of model internals and rely only on observed inputs and outputs. Representative approaches like AACH (Li et al. 2021) and EQB^{2}A (Zhu et al. 2023) leverage retrieval results or neighborhood structures to train surrogate models. This allows effective adversarial generation with limited queries. Earlier works mainly focused on non-targeted attack, which aim to degrade retrieval accuracy without specifying target semantics. However, targeted attack have been recognized as more severe threats. Methods like TA-DCH (Wang et al. 2023) and PGTA (Guo et al. 2025) propose semantic-driven frameworks to generate adversarial examples that mislead the retrieval towards attacker-specified targets.

Method	Type	P(Adv.)	P(Ben.)	Gener.
ATRDH	**Adv-Train**	✓	✗	✗
SAAT	**Adv-Train**	✓	✗	✗
RoCMR	**Adv-Train**	✓	✗	✗
TPAP	**PreProc**	✓	✗	✓
DRFGD	**PreProc**	✓	✓	✓

Table 1: Comparison of DRFGD with representative defense methods for cross-modal hashing. "Type" indicates the category of the method, either adversarial training (Adv-Train) or input preprocessing (PreProc). "P(Adv.)" and "P(Ben.)" represent whether the method performs well under adversarial and benign inputs, respectively. "Gener." indicates the generalizability of the method across different models.

Adversarial Defense on Deep Hashing Retrieval. To counteract the security threats posed by adversarial attack, two main categories of defense strategies have been explored: adversarial training and input preprocessing. Adversarial training improves model robustness by introducing adversarial examples into the training process. Mary et al. (Madry et al. 2018) first formalized it as a min-max optimization problem, laying a foundation for subsequent developments. In deep hashing tasks, ATRDH (Wang et al. 2021) minimizes the semantic distance between benign and adversarial examples in Hamming space, while SAAT (Yuan et al. 2023) employs a discriminative mainstay feature learning to generate more discriminative adversarial examples under a unified optimization framework. For cross-modal hashing, RoCMR (Zhang, Sun, and Zhao 2022) designs modality-specific perturbations and introduces contrastive learning to enhance matching robustness. However, adversarial training

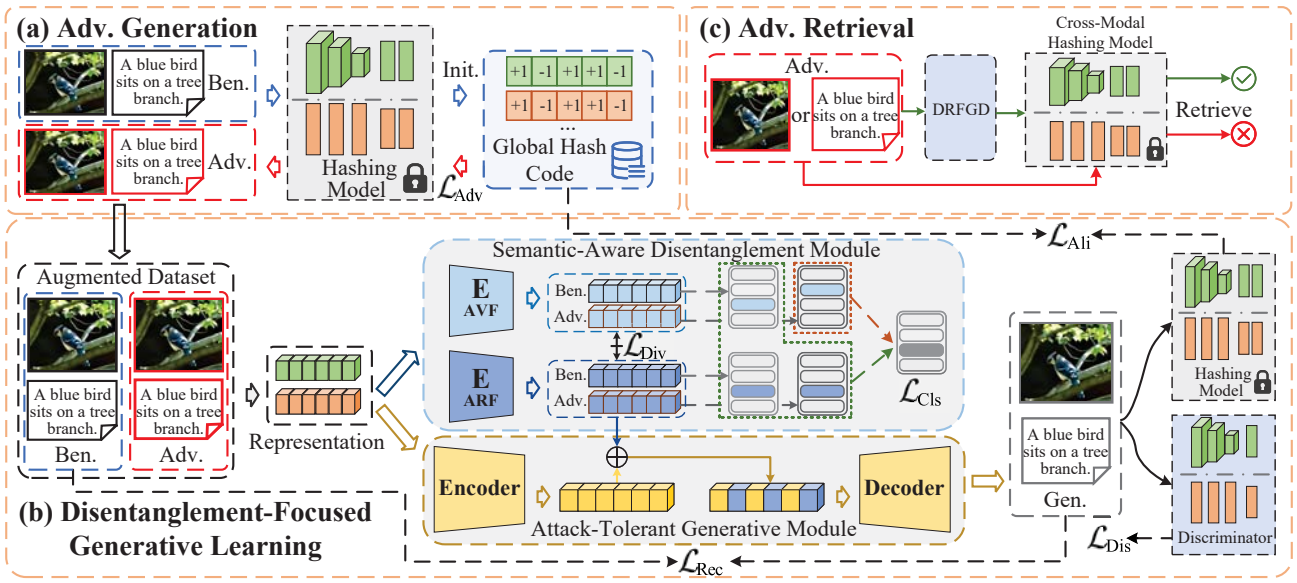

Figure 1: The overview structure of the DRFGD framework. DRFGD first disentangles the intermediate representations into adversarial-robust and adversarial-vulnerable features via a dual-branch semantic encoder. Then, the adversarial-robust features guide the attack-tolerant generative module to synthesize semantically aligned examples for robust training.

often demands heavy computational overhead and depends on the types of adversarial examples seen during training, limiting its generalization to unseen attacks and benign examples. Furthermore, in cross-modal retrieval, the semantic gap across heterogeneous modalities increases the complexity of constructing optimization objectives. To address these limitations, preprocessing-based defenses (Zhou et al. 2021; Wang et al. 2024; Tang and Zhang 2024) have been proposed, acting as model-agnostic modules with greater portability. For instance, PAIR (Zhou et al. 2024) proposes a random reconstruction strategy to disrupt adversarial patches in image hashing. Yet, defenses for cross-modal retrieval remain rare, mainly due to semantic inconsistency and the difficulty of unified reconstruction across modalities.

In summary, as shown in Table 1, existing defenses struggle to balance robustness, benign performance, and generalizability. To overcome these challenges, we propose DRFGD, a generative defense framework that achieves a better balance between robustness, accuracy, and generalization, without modifying the target retrieval model.

Methodology

Preliminary

Given a multi-modal dataset $\{\mathbf{x}_i^v, \mathbf{x}_i^t, \mathbf{y}_i\}_{i=1}^N$, where each example comprises an image $\mathbf{x}_i^v$, a text vector $\mathbf{x}_i^t$, and a label $\mathbf{y}_i$ from C categories, cross-modal hashing maps inputs from different modalities into a shared Hamming space via modality-specific hash functions. For a trained model possibly subject to adversarial attacks, the input $\mathbf{x}_i^m$ is mapped to a hash space as: $\mathbf{b}_i^m = \text{sign}(\mathcal{H}^m(\mathbf{x}_i^m))$, where $m \in \{v, t\}$, $\text{sign}(\cdot)$ denotes the sign function, $\mathbf{b}_i^m \in \{-1, 1\}^k$ is the k-bit hash code, and $\mathcal{H}^m \in \{\mathcal{H}^v, \mathcal{H}^t\}$ are the modality-specific hash networks. A global hash set $\mathbf{B} = \{\mathbf{b}_i^v, \mathbf{b}_i^t\}_{i=1}^N$ is initialized to maintain cross-modal semantic consistency, guiding adversarial example generation and generative learning.

Adversarial Example. In cross-modal hashing, adversarial attacks generate examples $\hat{\mathbf{x}}^m$ that mislead retrieval. Rather than relying solely on local correlations (Zhang, Sun, and Zhao 2022) or prototype guidance (Yuan et al. 2023), we leverage the global hash set $\mathbf{B}$ as supervision to optimize semantic discrimination of adversarial examples toward irrelevant categories and away from relevant ones. The loss for optimizing adversarial examples is defined as follows:

$$\begin{aligned}
&\mathcal{L}_{\text{Adv}}(\hat{\mathbf{x}}^m, \Psi, \Phi) = \log\Big[1 + \sum_{\mathcal{P}\in\Phi} \exp((\mathbf{S}_{\mathcal{P}} - \mathbf{S}_{\mathcal{N}})/\tau)\Big], \\
&\text{s.t.} \quad \mathbf{S}_{\mathcal{N}} = \frac{1}{k\,|\Psi|} \sum_{\mathcal{N}\in\Psi} \sum_{M}^{\{v,t\}} -\text{HamDist}(\mathcal{H}^m(\hat{\mathbf{x}}^m), \mathbf{b}_{\mathcal{N}}^M), \\
&\qquad\quad \mathbf{S}_{\mathcal{P}} = \frac{1}{k} \sum_{M}^{\{v,t\}} -\text{HamDist}(\mathcal{H}^m(\hat{\mathbf{x}}^m), \mathbf{b}_{\mathcal{P}}^M),
\end{aligned} \tag{1}$$

where $\text{HamDist}(\cdot)$ represents the Hamming distance, τ is the temperature parameter, and Ψ and Φ denote the sets of negative and positive examples, respectively. Further, the loss $\mathcal{L}_{\text{Adv}}$ serves as the objective for generating perturbations via PGD (Madry et al. 2018) attack:

$$\hat{\mathbf{x}}^m = \arg\min_{\delta^m} \mathcal{L}_{\text{Adv}}(\mathbf{x}^m + \delta^m, \Psi, \Phi). \tag{2}$$

The adversarial and benign examples jointly form an augmented training set $\{\mathbf{x}_i^v, \hat{\mathbf{x}}_i^v, \mathbf{x}_i^t, \hat{\mathbf{x}}_i^t, \mathbf{y}_i\}_{i=1}^N$.

Disentanglement-Focused Generative Learning

Cross-modal representations are inherently vulnerable to adversarial attack, which may corrupt the resulting hash

codes and compromise retrieval performance. To address this, as shown in Figure 1, we propose a disentangled representation-focused generative defense framework for attack-tolerant cross-modal hashing. This module comprises two key components: semantic-aware representation disentanglement and attack-tolerant generative defense.

Semantic-Aware Representation Disentanglement. To enable structured defense against adversarial attack, we assume that the input intermediate representations $\{\mathbf{x}^m \to \mathbf{f}^m, \hat{\mathbf{x}}^m \to \hat{\mathbf{f}}^m\}$ can be disentangled into two latent components: (1) Adversarial-Robust Features (ARF), which not only exhibit strong resistance to perturbations but also preserve semantically consistent representations that facilitate reliable correlation across heterogeneous modalities; and (2) Adversarial-Vulnerable Features (AVF), which are easily affected by perturbation and lose semantic fidelity. To extract these components, we design a semantic-aware disentanglement module for each modality m: an ARF encoder Enc^m_{arf} and an AVF encoder Enc^m_{avf}. Given intermediate representations $\mathbf{f}^m$, disentangled features are obtained as:

$$\begin{bmatrix} \mathbf{z}^m_{\text{arf}} \\ \mathbf{z}^m_{\text{avf}} \end{bmatrix} = \mathcal{E}^m\left(\mathbf{f}^m\right) \coloneqq \begin{bmatrix} Enc^m_{\text{arf}}(\mathbf{f}^m) \\ Enc^m_{\text{avf}}(\mathbf{f}^m) \end{bmatrix}, \tag{3}$$

where $\mathcal{E}^m(\cdot)$ denotes the disentanglement for modality m by the dual-branch parallel semantic encoding pathways. Similarly, adversarial representations $\hat{\mathbf{f}}^m$ are mapped to $\{\hat{\mathbf{z}}^m_{\text{avf}}, \hat{\mathbf{z}}^m_{\text{arf}}\}$ via the same disentanglement pathway.

To encourage independence between ARF and AVF subspaces, we employ a symmetric KL divergence:

$$\min \mathcal{L}_{\text{Div}} = \mathop{\mathbb{E}}_{\substack{\{\mathbf{z}^m_{\text{arf}},\, \hat{\mathbf{z}}^m_{\text{arf}}\} \sim \mathbf{z} \\ \{\mathbf{z}^m_{\text{avf}},\, \hat{\mathbf{z}}^m_{\text{avf}}\} \sim \mathbf{z}'}} \left[\mathcal{D}_{\text{KL}}(\mathbf{z} || \mathbf{z}') + \mathcal{D}_{\text{KL}}(\mathbf{z}' || \mathbf{z})\right], \tag{4}$$

where $\mathcal{D}_{\text{KL}}(\cdot||\cdot)$ measures the divergence between feature distributions. Furthermore, to enhance the semantic expressiveness of ARF, we introduce a classification loss using a softmax classifier $\mathcal{C}_\sigma$ applied to both $\mathbf{z}^m_{\text{arf}}$ and $\hat{\mathbf{z}}^m_{\text{arf}}$:

$$\min \mathop{\mathbb{E}}_{\{\mathbf{z}^m_{\text{arf}},\, \hat{\mathbf{z}}^m_{\text{arf}}\} \sim \mathbf{z}} - \sum_{c=1}^{C} \mathbf{y}^c \cdot \log(\mathcal{C}_\sigma(\mathbf{z})). \tag{5}$$

In contrast, to emphasize the semantic inconsistency of AVF, we adopt a reverse supervision strategy: $\mathbf{z}^m_{\text{avf}}$ are expected to yield correct classification, whereas $\hat{\mathbf{z}}^m_{\text{avf}}$ are encouraged to produce incorrect classification. The full objective is:

$$\begin{aligned} \min \mathcal{L}_{\text{Cls}} = & \mathop{\mathbb{E}}_{\{\mathbf{z}^m_{\text{arf}},\, \hat{\mathbf{z}}^m_{\text{arf}},\, \mathbf{z}^m_{\text{avf}}\} \sim \mathbf{z}} - \sum_{c=1}^{C} \mathbf{y}^c \cdot \log(\mathcal{C}_\sigma(\mathbf{z})) \\ & + \mathop{\mathbb{E}}_{\{\hat{\mathbf{z}}^m_{\text{avf}}\} \sim \mathbf{z}'} - \sum_{c=1}^{C} (1 - \mathbf{y}^c) \cdot \log(\mathcal{C}_\sigma(\mathbf{z}')), \end{aligned} \tag{6}$$

The overall loss for the semantic-aware disentanglement module R is defined as:

$$\mathcal{L}_R = \mathcal{L}_{\text{Div}} + \mathcal{L}_{\text{Cls}}. \tag{7}$$

After disentanglement, only ARF are retained for generative defense, as they capture robust and semantically meaningful information. Conversely, AVF are discarded due to their perturbation sensitivity and semantic instability, which risk degrading both robustness and retrieval performance.

Attack-Tolerant Generative Defense. To further enhance adversarial robustness, we introduce attack-tolerant generative modules for both image and text modalities, guided by the disentangled ARF. These modules synthesize semantically aligned and perturbation-resilient inputs, thereby reinforcing cross-modal consistency and enhancing collaborative defense across modalities. By enabling robust adversarial training, the framework improves the resilience of the resulting hash codes derived from the generated examples.

Specifically, given input representations $\{\mathbf{f}^m, \hat{\mathbf{f}}^m\}$ and their ARF components $\{\mathbf{z}^m_{\text{arf}}, \hat{\mathbf{z}}^m_{\text{arf}}\}$, a modality-specific generator G^m reconstructs inputs as:

$$\bar{\mathbf{x}}^m = G^m\left(\mathbf{f}^m \oplus \mathbf{z}^m_{\text{arf}}\right), \quad \tilde{\mathbf{x}}^m = G^m\left(\hat{\mathbf{f}}^m \oplus \hat{\mathbf{z}}^m_{\text{arf}}\right), \tag{8}$$

where $\oplus$ denotes channel-wise concatenation, $\bar{\mathbf{x}}^m$ and $\tilde{\mathbf{x}}^m$ are reconstructions derived from benign and adversarial inputs, respectively. Unlike the adversarial generation process in Eq. (1), here we enforce semantic consistency at the hash level through an alignment loss defined as:

$$\min \mathcal{L}_{\text{Ali}} = \mathop{\mathbb{E}}_{\{\bar{\mathbf{x}}^m,\, \tilde{\mathbf{x}}^m\} \sim \mathbf{x}} \left[\mathcal{L}_{\text{Adv}}(\mathbf{x}, \Phi, \Psi)\right], \tag{9}$$

where Φ and Ψ denote the positive and negative example sets, respectively. In parallel, a reconstruction loss ensures fidelity between original and generated examples:

$$\min \mathcal{L}_{\text{Rec}} = \mathop{\mathbb{E}}_{\{\mathbf{x}^m,\, \bar{\mathbf{x}}^m,\, \tilde{\mathbf{x}}^m\} \sim \mathbf{x},\, \mathbf{x}'} \|\mathbf{x} - \mathbf{x}'\|_2^2. \tag{10}$$

Further, we introduce a modality-specific discriminator D^m to distinguish generated examples from real ones:

$$\begin{aligned} \min \mathcal{L}_D = & \mathop{\mathbb{E}}_{\{\bar{\mathbf{x}}^m,\, \tilde{\mathbf{x}}^m\} \sim \mathbf{x}} \left[\log(D^m(\mathbf{x}))\right] \\ & + \mathop{\mathbb{E}}_{\{\mathbf{x}^m\} \sim \mathbf{x}'} \left[\log(1 - D^m(\mathbf{x}'))\right]. \end{aligned} \tag{11}$$

Correspondingly, the generator is optimized to fool the discriminator, leading to the following discriminative loss:

$$\min \mathcal{L}_{\text{Dis}} = - \mathop{\mathbb{E}}_{\{\bar{\mathbf{x}}^m,\, \tilde{\mathbf{x}}^m\} \sim \mathbf{x}} \left[\log(D^m(\mathbf{x}))\right]. \tag{12}$$

The overall objective for the generative module is:

$$\mathcal{L}_G = \mathcal{L}_{\text{Dis}} + \alpha \cdot \mathcal{L}_{\text{Ali}} + \beta \cdot \mathcal{L}_{\text{Rec}}, \tag{13}$$

where α and β are weighting hyperparameters.

To train the framework, we minimize $\mathcal{L}_R + \mathcal{L}_G$ to update the parameters Θ_R and Θ_G of the disentanglement and generative modules, respectively. The discriminator parameters Θ_D are updated in an alternating manner by minimizing $\mathcal{L}_D$.

Experiments

Experiment Setting

Datasets. We conduct experiments on popular public multi-modal datasets: FLICKR-25K (Huiskes and Lew 2008), MS-COCO (Lin et al. 2014), and NUS-WIDE (Chua et al. 2009). Following the common experimental setting in EQB^{2}A (Zhu et al. 2023), we set the dataset splits as follows: **FLICKR-25K** contains 25,000 image-text pairs annotated with 24 categories. We select 2,000 pairs as the

Attack Methods	Metric	Defense Methods	FLICKR-25K			MS-COCO			NUS-WIDE		
			CPAH	DADH	DAPH	CPAH	DADH	DAPH	CPAH	DADH	DAPH
EQB^{2}A	I → T (MAP↑)	No Defense	60.78	58.15	59.98	38.08	37.74	37.91	40.85	43.81	51.28
		ATRDH	70.04	67.38	67.69	50.28	55.51	56.97	55.79	61.87	67.51
		SAAT	67.34	70.13	70.82	45.96	56.24	55.08	48.74	64.97	64.32
		RoCMR	68.39	69.16	71.76	59.03	50.77	56.53	46.36	55.52	66.11
		TPAP	72.78	69.24	70.35	50.51	54.12	54.98	55.34	64.07	66.89
		Ours	**73.01**	**72.06**	**72.13**	**59.29**	**58.08**	**59.44**	**59.52**	**66.31**	**70.99**
TA-DCH	I → T (MAP↑)	No Defense	68.12	70.16	66.95	49.42	41.31	44.06	48.82	45.36	54.31
		ATRDH	71.56	74.80	75.02	54.40	49.23	63.03	59.47	58.51	68.46
		SAAT	72.69	76.95	75.83	57.31	50.65	59.09	52.91	64.99	69.57
		RoCMR	73.32	74.87	74.49	58.39	57.82	60.02	64.21	66.80	71.20
		TPAP	71.35	75.08	76.48	57.71	56.46	60.14	61.55	65.71	68.35
		Ours	**74.85**	**78.75**	**76.78**	**58.63**	**61.92**	**64.88**	**67.41**	**68.40**	**71.32**
TA-DCH	T → I (MAP↑)	No Defense	61.97	61.06	59.77	45.82	47.01	38.49	52.96	37.98	43.06
		ATRDH	72.33	63.98	70.26	55.58	66.65	52.50	66.29	57.40	65.78
		SAAT	68.23	70.99	71.34	58.21	70.17	57.26	62.14	56.29	58.97
		RoCMR	69.40	71.89	65.98	60.03	67.70	60.06	60.52	64.38	70.94
		TPAP	68.46	74.62	65.42	61.64	65.03	62.46	64.28	68.22	65.92
		Ours	**74.70**	**77.09**	**71.81**	**61.82**	**72.98**	**70.01**	**66.58**	**69.82**	**72.20**
TA-DCH	I → T (t-MAP↓)	No Defense	82.52	83.44	81.42	64.47	65.96	65.39	78.88	80.87	80.69
		ATRDH	72.89	70.97	73.02	60.06	52.38	56.53	64.09	60.48	60.74
		SAAT	69.67	72.18	71.25	59.10	**49.12**	58.12	63.53	63.82	59.81
		RoCMR	74.85	75.91	75.93	61.05	56.37	57.79	55.08	**54.23**	62.55
		TPAP	**66.39**	70.46	71.55	58.45	49.18	56.04	51.00	60.27	61.32
		Ours	69.21	**69.53**	**69.02**	**56.82**	52.28	**54.37**	**49.79**	56.83	**57.41**
TA-DCH	T → I (t-MAP↓)	No Defense	86.38	86.94	88.23	60.33	68.06	71.13	70.37	73.61	73.41
		ATRDH	76.02	76.59	74.81	54.78	41.89	49.26	55.13	44.48	47.28
		SAAT	75.96	**71.29**	72.37	55.28	**38.91**	52.59	56.07	43.88	46.58
		RoCMR	74.84	72.40	69.65	56.45	43.94	52.01	59.80	37.39	57.77
		TPAP	73.29	74.91	73.69	52.72	40.24	44.80	54.20	42.93	51.66
		Ours	**68.21**	72.48	**64.28**	**51.53**	44.19	**43.54**	**52.36**	**36.74**	**40.63**

Table 2: Comparison of defense performance (MAP % and t-MAP %) at 32 bits with other defense methods.

query set and randomly sample 5,000 pairs from the retrieval set for training. **MS-COCO** contains 123,287 image-text pairs annotated with 80 categories. We select 2,000 pairs for queries and randomly sample 10,000 pairs from the retrieval set for training. **NUS-WIDE** consists of 269,648 image-text pairs annotated with 81 categories. We focus on the 21 most categories, selecting 500 pairs per category for training and 100 pairs for queries.

Baselines. In the context of limited research on adversarial defense for cross-modal hashing, to comprehensively evaluate the effectiveness of DRFGD, we select several advanced defense baselines, including ATRDH (Wang et al. 2021), SAAT (Yuan et al. 2023), RoCMR (Zhang, Sun, and Zhao 2022), and TPAP (Tang and Zhang 2024). For the retrieval models under attack and defense training, we adopt three cross-modal hashing methods: CPAH (Xie et al. 2020), DADH (Bai et al. 2020a), and DAPH (Tu et al. 2023a). Additionally, the adversarial attacks considered in our evaluation include the non-targeted attack EQB^{2}A (Zhu et al. 2023) and the targeted attack TA-DCH (Wang et al. 2023).

Implementation Details. For the proposed framework, the image generative model adopts a U-Net (Ronneberger, Fischer, and Brox 2015) architecture enhanced with residual connections, while the text generative model is composed of fully connected layers integrated with residual modules. Adversarial examples are generated via the PGD attack, with empirically determined step sizes of 3/255 for the image modality and 3/500 for text modality. The perturbation budgets are set to 8/255 for image inputs and 0.05 for text inputs, and the number of attack iterations is fixed at 20. The hyperparameters τ and λ are set to 0.05 and 1, respectively. The values of $\alpha \in [1, 50]$ and $\beta \in [400, 2000]$ are selected based on the target model. The generative defense networks are trained using the Adam optimizer with a batch size of 24. The learning rate varies between 10^{-4} and 10^{-3}, and the total number of training epochs is 50.

Protocols. Three popular evaluation metrics are utilized to assess model performance under adversarial attacks: Mean Average Precision (MAP) (Shen et al. 2015, 2018; Liu et al. 2021; Zha et al. 2024), targeted-Mean Average Precision (t-MAP) (Bai et al. 2020b), and Precision-Recall (PR) curves. Specifically, MAP is employed to evaluate retrieval performance under both non-targeted and targeted attacks, where adversarial examples aim to degrade retrieval accuracy. In contrast, t-MAP is used to assess robustness under targeted attack, where the labels of queries are replaced with target

Attack Methods	Metric	Defense Methods	FLICKR-25K			MS-COCO			NUS-WIDE		
			CPAH	DADH	DAPH	CPAH	DADH	DAPH	CPAH	DADH	DAPH
No Attack	I → T (MAP↑)	ATRDH	75.76	79.06	81.97	58.72	69.13	71.44	73.41	71.58	74.52
		SAAT	77.01	78.21	82.21	61.91	69.66	70.56	72.64	73.26	75.65
		RoCMR	80.66	82.57	83.42	60.86	67.96	65.17	71.26	71.65	75.83
		TPAP	75.43	82.39	83.16	61.06	66.29	68.31	73.10	76.25	75.78
		Ours	**82.53**	**84.49**	**84.61**	**63.35**	**72.58**	**73.62**	**74.05**	**77.49**	**78.68**
No Attack	T → I (MAP↑)	ATRDH	74.86	78.67	78.16	61.59	78.56	76.90	70.76	71.17	74.33
		SAAT	76.58	77.31	79.85	59.34	82.57	75.38	69.58	74.36	73.97
		RoCMR	73.61	80.69	81.36	62.21	**84.75**	71.79	69.38	72.61	74.26
		TPAP	72.83	79.86	77.14	62.45	77.06	75.18	70.84	73.22	74.50
		Ours	**80.91**	**84.02**	**83.61**	**63.67**	83.46	**82.27**	**72.54**	**75.82**	**75.33**

Table 3: Comparison of benign retrieval (MAP %) at 32 bits with other defense methods.

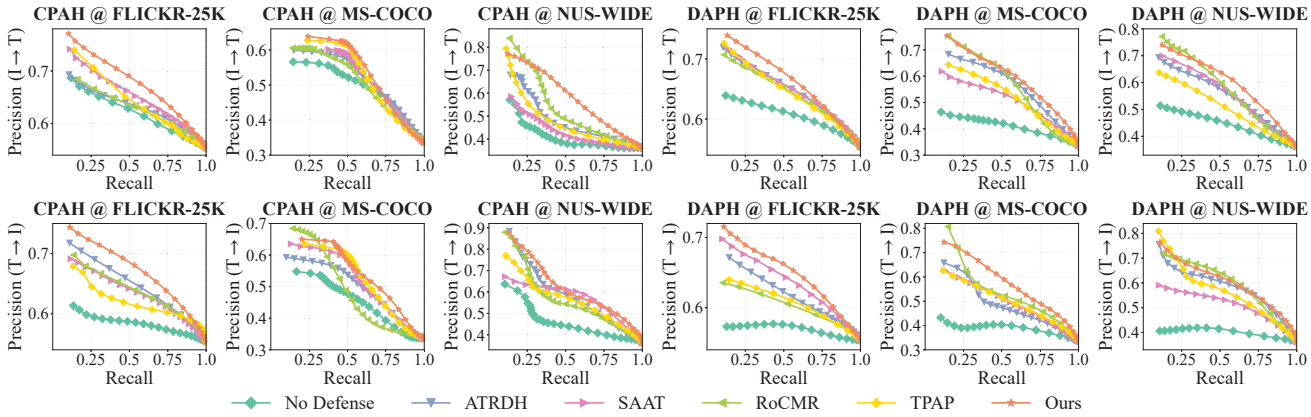

Figure 2: PR curves of CPAH and DAPH model under TA-DCH attack on three datasets at 32 bits.

labels. Additionally, PR curves are plotted to provide a comprehensive visualization of retrieval performance.

Results

Defense Performance. To evaluate the effectiveness of DRFGD, we conduct comprehensive experiments on three widely-used cross-modal retrieval benchmarks: FLICKR-25K, MS-COCO, and NUS-WIDE. CPAH, DADH, and DAPH are adopted as representative target models. We test their robustness and retrieval performance under two adversarial settings: the non-targeted attack EQB^{2}A and the targeted attack TA-DCH. All evaluations use 32-bit hash codes. "I→T" denotes image-to-text retrieval, while "T→I" denotes text-to-image. Since EQB^{2}A targets image inputs, only the I→T task is evaluated under this attack.

As shown in Table 2, all baseline models suffer significant performance drops under EQB^{2}A without defense. With defense mechanisms applied, retrieval performance improves notably. DRFGD consistently outperforms other methods across models and datasets. For example, on the DADH model, it improves MAP by 1.93%, 1.84%, and 1.34% on FLICKR-25K, MS-COCO, and NUS-WIDE, respectively. Similar improvements are observed on CPAH and DAPH, confirming the broad applicability of DRFGD. Under targeted attack TA-DCH, DRFGD likewise yields significant improvements in I→T retrieval. On the DADH model, MAP increases by 1.81%, 4.10%, and 1.60% across the three datasets, respectively. These results indicate the effectiveness of our semantic-preserving generation strategy in countering adversarial perturbations.

Additionally, DRFGD demonstrates strong performance on the T→I task across all settings, validating its robustness in both retrieval directions. For a finer-grained analysis, we also adopt t-MAP under TA-DCH, where lower scores indicate better resistance. DRFGD consistently achieves the lowest or near-lowest t-MAP. For example, on DAPH in the I→T task, DRFGD reduces t-MAP by 2.23%, 1.67%, and 2.40% on FLICKR-25K, MS-COCO, and NUS-WIDE, respectively. While SAAT or TPAP slightly outperform in a few isolated cases, DRFGD maintains more consistent and reliable defense across all models.

Table 3 further shows that adversarial training often hurts benign performance. In contrast, DRFGD retains high MAP scores even without attacks. For example, on the DAPH model, DRFGD achieves a MAP of 84.61% in the I→T task on the FLICKR-25K dataset, surpassing all other defense baselines. Similar trends appear in T→I, demonstrating DRFGD's ability to balance robustness and accuracy.

As shown in Figure 2, PR curves under TA-DCH attack show that DRFGD consistently outperforms alterna-

tives across datasets and models, confirming its robustness, generality, and semantic consistency.

Result Analysis. DRFGD demonstrates superior defense performance compared to SAAT, ATRDH, RoCMR, and TPAP due to its fundamentally different defense mechanism that directly addresses semantic consistency and representation stability under adversarial attacks. SAAT constructs semantic representatives to guide adversarial training, aiming to preserve discriminative and semantic properties. However, its reliance on static semantic anchors can be limited when facing complex, unseen adversarial examples. ATRDH leverages prototype codes as category-level semantic embeddings to conduct adversarial attack and training, but its defense strategy is tightly coupled with fixed label semantics and thus may underperform in preserving fine-grained retrieval alignment. RoCMR applies contrastive adversarial training across modalities to enhance robustness, but it lacks an explicit mechanism to preserve performance on benign data. TPAP adopts a general input purification strategy that offers broad applicability, yet falls short in effectively handling modality-specific perturbations inherent in cross-modal retrieval. In contrast, DRFGD introduces a generative defense process guided by learned adversarial-robust features, which not only rectifies adversarial inputs but also restores semantic alignment across modalities by aligning them with the global hash code. Additionally, DRFGD maintains high retrieval performance on benign examples, which existing adversarial training-based methods often compromise. This balance between robustness and retrieval accuracy highlights the generalization capacity of our method and its practical effectiveness in real-world adversarial cross-modal hashing retrieval.

Representations Disentanglement. To further validate the effectiveness of semantic-aware representation disentanglement, we visualize the extraction process of adversarial-robust features (ARF) and adversarial-vulnerable features (AVF) from adversarial images, as illustrated in Figure 3. The ARF consistently concentrate on semantically meaningful regions of the object (e.g., vehicle contours), which are crucial for semantic understanding and remain relatively unaffected by adversarial perturbations. In contrast, the AVF capture dispersed or irrelevant background patterns that vary significantly under such perturbations. For instance, the ARF accurately highlight the main structures of the tram and motorcycle, whereas the AVF respond to perturbation regions around the sky and road, revealing their instability and sensitivity. This consistent divergence indicates that our model effectively disentangles robust semantic features from perturbation-sensitive components. By isolating ARF during training, the model preserves stable and task-relevant information while minimizing the influence of vulnerable features. Such semantic disentanglement plays a vital role in enhancing modality-specific robustness and improving collaborative defense in cross-modal scenarios.

Ablation Study

To evaluate the role of each loss component in DRFGD, we conduct ablation studies on CPAH under the TA-DCH attack. As shown in Table 4, progressively removing each loss leads to performance degradation across the three datasets, highlighting their necessity. Specifically, removing the divergence loss $\mathcal{L}_{\text{Div}}$ results in a slight performance drop, indicating its auxiliary role in guiding effective disentanglement. The classifier loss $\mathcal{L}_{\text{Cls}}$ proves more crucial, as its removal weakens semantic awareness during disentanglement. Among generative losses, removing the alignment loss $\mathcal{L}_{\text{Ali}}$ leads to the largest drop, highlighting its role in preserving hash-level consistency. Eliminating the reconstruction loss $\mathcal{L}_{\text{Rec}}$ causes notable degradation, demonstrating its contribution to input fidelity. Lastly, dropping the discrimination loss $\mathcal{L}_{\text{Dis}}$ results in a slight decline, confirming its role in enhancing generation quality. These results affirm the complementary effect of all components on adversarial robustness.

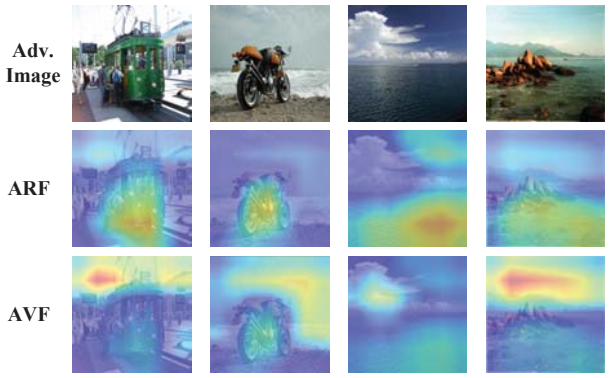

Figure 3: Disentangled heatmap comparisons.

Variants	FLICKR-25K		MS-COCO		NUS-WIDE	
	I→T	T→I	I→T	T→I	I→T	T→I
w/o $\mathcal{L}_{\text{Div}}$	74.26	73.92	57.69	61.02	66.82	65.18
w/o $\mathcal{L}_{\text{Cls}}$	73.88	73.60	56.85	60.73	66.19	65.46
w/o $\mathcal{L}_{\text{Ali}}$	71.95	67.83	52.68	49.79	58.40	60.51
w/o $\mathcal{L}_{\text{Rec}}$	72.68	61.19	54.72	43.95	64.26	53.08
w/o $\mathcal{L}_{\text{Dis}}$	73.27	72.81	56.97	60.18	65.42	64.90
Ours	**74.85**	**74.70**	**58.63**	**61.82**	**67.41**	**66.58**

Table 4: MAP % results of the ablation study on the CPAH model under the TA-DCH attack.

Conclusion

In this paper, we propose DRFGD, a novel disentangled representation-focused generative defense framework for achieving attack-tolerant cross-modal hashing. DRFGD integrates a semantic-aware disentanglement module and an attack-tolerant generative module to effectively defend against adversarial attack and synthesize perturbation-resilient inputs. This dual-pronged design not only suppresses the influence of adversarial perturbation but also preserves semantic fidelity across modalities. Extensive experiments demonstrate that DRFGD achieves superior performance under a variety of adversarial settings while maintaining compatibility with existing cross-modal hashing models.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grants 62576143, 62476103, 62476107 and 62276106, in part by the National Science Foundation of Fujian Province under Grant 2024J01096, in part by Xiamen Science and Technology Project under Grant 3502Z20251015, in part by the RGC Junior Research Fellow Scheme under Grant JRFS2526-2S06, in part by the RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02, in part by Guangdong and Hong Kong Universities "1+1+1" Cross-Campus Research Collaboration Scheme under Grant 2025A0505000004, and in part by Faculty Niche Research Areas of Hong Kong Baptist University under Grant RC-FNRA-IG/23-24/SCI/02.

References

Bai, C.; Zeng, C.; Ma, Q.; Zhang, J.; and Chen, S. 2020a. Deep Adversarial Discrete Hashing for Cross-Modal Retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*, 525–531.

Bai, J.; Chen, B.; Li, Y.; Wu, D.; Guo, W.; Xia, S.; and Yang, E. 2020b. Targeted Attack for Deep Hashing Based Retrieval. In *Proceedings of the European Conference on Computer Vision*, 618–634.

Chen, Y.; Li, X.; Hu, P.; Peng, D.; and Wang, X. 2024. DifFilter: Defending Against Adversarial Perturbations With Diffusion Filter. *IEEE Transactions on Information Forensics and Security*, 19: 6779–6794.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 1–9.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, X.; Zhang, H.; Liu, L.; Liu, D.; Lu, X.; and Meng, H. 2025. Primary Code Guided Targeted Attack against Cross-modal Hashing Retrieval. *IEEE Transactions on Multimedia*, 27: 312–326.

Hu, Z.; Cheung, Y.-m.; Li, M.; and Lan, W. 2024. Cross-Modal Hashing Method with Properties of Hamming Space: A New Perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7636–7650.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 39–43.

Jiang, Q.; and Li, W. 2017. Deep Cross-Modal Hashing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3270–3278.

Li, C.; Gao, S.; Deng, C.; Liu, W.; and Huang, H. 2021. Adversarial Attack on Deep Cross-Modal Hamming Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2198–2207.

Li, C.; Gao, S.; Deng, C.; Xie, D.; and Liu, W. 2019. Cross-Modal Learning with Adversarial Samples. In *Proceedings of the International Conference on Neural Information Processing Systems*, 10792–10802.

Li, C.; Tang, H.; Deng, C.; Zhan, L.; and Liu, W. 2020. Vulnerability vs. Reliability: Disentangled Adversarial Examples for Cross-Modal Learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 421–429.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, X.; Hu, Z.; Ling, H.; and Cheung, Y.-M. 2021. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3): 964–981.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations*, 4138–4160.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *Proceedings of the International Conference on Machine Learning*, 16805–16827.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Shen, F.; Shen, C.; Liu, W.; and Shen, H. T. 2015. Supervised Discrete Hashing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 37–45.

Shen, F.; Xu, Y.; Liu, L.; Yang, Y.; Huang, Z.; and Shen, H. T. 2018. Unsupervised Deep Hashing with Similarity-Adaptive and Discrete Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 3034–3044.

Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; and Hu, P. 2024. Dual Self-Paced Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15184–15192.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 1–10.

Tang, L.; and Zhang, L. 2024. Robust overfitting does matter: Test-time adversarial purification with fgsm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24347–24356.

Tu, R.; Mao, X.; Ji, W.; Wei, W.; and Huang, H. 2023a. Data-Aware Proxy Hashing for Cross-modal Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 686–696.

Tu, R.; Mao, X.; Ma, B.; Hu, Y.; Yan, T.; Wei, W.; and Huang, H. 2022. Deep Cross-Modal Hashing With Hashing Functions and Unified Hash Codes Jointly Learning. *IEEE*

Transactions on Knowledge and Data Engineering, 34(2): 560–572.

Tu, R.; Mao, X.; Tu, R.; Bian, B.; Cai, C.; Wang, H.; Wei, W.; and Huang, H. 2023b. Deep Cross-Modal Proxy Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 6798–6810.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the ACM on Multimedia Conference*, 154–162.

Wang, H.; Deng, Y.; Yoo, S.; and Lin, Y. 2024. Exploring Robust Features for Improving Adversarial Robustness. *IEEE Transactions on Cybernetics*, 54(9): 5141–5151.

Wang, T.; Zhu, L.; Zhang, Z.; Zhang, H.; and Han, J. 2023. Targeted Adversarial Attack Against Deep Cross-Modal Hashing Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 6159–6172.

Wang, X.; Zhang, Z.; Lu, G.; and Xu, Y. 2021. Targeted Attack and Defense for Deep Hashing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2298–2302.

Wang, Y.; Chen, Z.; Luo, X.; and Xu, X. 2022. A high-dimensional sparse hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8822–8836.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. In *Proceedings of the International Conference on Learning Representations*, 5849–5865.

Xiao, C.; Chen, Z.; Jin, K.; Wang, J.; Nie, W.; Liu, M.; Anandkumar, A.; Li, B.; and Song, D. 2022. DensePure: Understanding Diffusion Models towards Adversarial Robustness. In *Proceedings of the International Conference on Neural Information Processing Systems*, 1–19.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2017. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*.

Xie, C.; Wu, Y.; Maaten, L. v. d.; Yuille, A. L.; and He, K. 2019. Feature Denoising for Improving Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 501–509.

Xie, D.; Deng, C.; Li, C.; Liu, X.; and Tao, D. 2020. Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval. *IEEE Transactions on Image Processing*, 29: 3626–3637.

Xu, L.; Zeng, X.; Zheng, B.; and Li, W. 2022. Multi-manifold deep discriminative cross-modal hashing for medical image retrieval. *IEEE Transactions on Image Processing*, 31: 3371–3385.

Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; and Gao, X. 2017. Pairwise relationship guided deep hashing for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 1618–1625.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4626–4634.

Yuan, X.; Zhang, Z.; Wang, X.; and Wu, L. 2023. Semantic-Aware Adversarial Training for Reliable Deep Hashing Retrieval. *IEEE Transactions on Information Forensics and Security*, 18: 4681–4694.

Zha, Q.; Liu, X.; Cheung, Y.; Xu, X.; Wang, N.; and Cao, J. 2024. UGNCL: Uncertainty-Guided Noisy Correspondence Learning for Efficient Cross-Modal Matching. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 852–861.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L. E.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of the International Conference on Machine Learning*, volume 97, 7472–7482.

Zhang, L.; Zhou, Y.; Yang, Y.; and Gao, X. 2024. Meta Invariance Defense Towards Generalizable Robustness to Unknown Adversarial Attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6669–6687.

Zhang, T.; Sun, S.; and Zhao, J. 2022. Robust Cross-Modal Retrieval by Adversarial Training. In *Proceedings of the International Joint Conference on Neural Networks*, 1–8.

Zhou, D.; Liu, T.; Han, B.; Wang, N.; Peng, C.; and Gao, X. 2021. Towards Defending against Adversarial Examples via Attack-Invariant Features. In *Proceedings of the International Conference on Machine Learning*, 12835–12845.

Zhou, M.; and Patel, V. M. 2022. Enhancing Adversarial Robustness for Deep Metric Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15304–15313.

Zhou, Z.; Wang, P.; Liang, Z.; Zhang, R.; and Bai, H. 2024. PAIR: Pre-denosing Augmented Image Retrieval Model for Defending Adversarial Patches. In *Proceedings of the ACM International Conference on Multimedia*, 5771–5779.

Zhu, L.; Wang, T.; Li, J.; Zhang, Z.; Shen, J.; and Wang, X. 2023. Efficient Query-based Black-box Attack against Cross-modal Hashing Retrieval. *ACM Transactions on Information Systems*, 41(3): 1–25.