Invisible Black-box Backdoor Attack against Deep Cross-modal Hashing Retrieval

TIANSHI WANG, Shandong Normal University, Jinan, China
FENGLING LI, University of Technology Sydney, Sydney, Australia
LEI ZHU, Shandong Normal University, Jinan, China
JINGJING LI, University of Electronic Science and Technology of China, Chengdu, China
ZHENG ZHANG, Harbin Institute of Technology Shenzhen, Shenzhen, China
HENG TAO SHEN, University of Electronic Science and Technology of China, Chengdu, China

Deep cross-modal hashing has promoted the field of multi-modal retrieval due to its excellent efficiency and storage, but its vulnerability to backdoor attacks is rarely studied. Notably, current deep cross-modal hashing methods inevitably require large-scale training data, resulting in poisoned samples with imperceptible triggers that can easily be camouflaged into the training data to bury backdoors in the victim model. Nevertheless, existing backdoor attacks focus on the uni-modal vision domain, while the multi-modal gap and hash quantization weaken their attack performance. In addressing the aforementioned challenges, we undertake an invisible black-box backdoor attack against deep cross-modal hashing retrieval in this article. To the best of our knowledge, this is the first attempt in this research field. Specifically, we develop a flexible trigger generator to generate the attacker's specified triggers, which learns the sample semantics of the non-poisoned modality to bridge the cross-modal attack gap. Then, we devise an input-aware injection network, which embeds the generated triggers into benign samples in the form of sample-specific stealth and realizes cross-modal semantic interaction between triggers and poisoned samples. Owing to the knowledge-agnostic of victim models, we enable any cross-modal hashing knockoff to facilitate the black-box backdoor attack and alleviate the attack weakening of hash quantization. Moreover, we propose a confusing perturbation and mask strategy to induce the high-performance victim models to focus on imperceptible triggers in poisoned samples. Extensive experiments on benchmark datasets demonstrate that our method has a state-of-the-art attack performance against deep cross-modal hashing retrieval. Besides, we investigate the influences of transferable attacks, few-shot poisoning, multi-modal poisoning, perceptibility, and potential defenses on backdoor attacks. Our codes and datasets are available at https://github.com/tswang0116/IB3A

CCS Concepts: • **Information systems** → **Information retrieval**; **Multimedia and multimodal retrieval;**

Additional Key Words and Phrases: Backdoor attack, black-box attack, imperceptible trigger, deep cross-modal hashing retrieval

This work was supported in part by the National Natural Science Foundation of China under Grant 62172263, in part by the Natural Science Foundation of Shandong, China, under Grant ZR2020YQ47, and in part by CCF-Baidu Open Fund under Grant CCF-BAIDU OF2022008.

Authors' addresses: T. Wang and L. Zhu (Corresponding author), Shandong Normal University, Jinan, 250358, China; F. Li, University of Technology Sydney, Sydney, NSW 2007, Australia; J. Li and H. T. Shen, University of Electronic Science and Technology of China, Chengdu, 611731, China; Z. Zhang, Harbin Institute of Technology Shenzhen, Shenzhen, 518055, China.

ACM 1046-8188/2024/04-ART111
https://doi.org/10.1145/3650205

ACM Reference Format:
Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024. Invisible Black-box Backdoor Attack against Deep Cross-modal Hashing Retrieval. *ACM Trans. Inf. Syst.* 42, 4, Article 111 (April 2024), 27 pages. https://doi.org/10.1145/3650205

1 INTRODUCTION

With the explosive growth of multi-modal data, many cross-modal retrieval methods [25, 50] have been proposed to realize efficient semantic search across heterogeneous multi-modal data. Among them, deep cross-modal hashing retrieval [67, 69] has attracted wide attention because of its excellent search and storage efficiency. Regrettably, deep cross-modal hashing retrieval, while benefiting from the powerful representation capability of neural networks, inevitably inherits their vulnerability to adversarial attacks [44]. As deep cross-modal hashing becomes the mainstream of large-scale cross-modal retrieval, its robustness and anti-interference receive increasing attention [28–30, 51, 64, 68] to make it work stably.

Existing studies [28–30, 51, 68] show that both evasion attacks and backdoor attacks can easily fool deep cross-modal hashing models. Among them, the popular ones are the evasion attacks on model inference [28–30, 51, 68], which cause the attacked retrieval system to return undesired search results by tampering with query samples. However, evasion attacks act on the prediction and reasoning of deployed retrieval systems, often overlooking the potential consequences of poisoned samples on the optimization training of cross-modal hashing models. Compared with evasion attacks [28–30, 51, 68], backdoor attacks [14, 19] bury malicious backdoors into victim models by accessing partial training data, so that victim models behave normally on benign queries. When victim models locate at specified triggers, the buried backdoors induce retrieval systems to return malicious results. In particular, existing deep cross-modal hashing models [1, 23] rely on learning semantic correlations and similarity measures from large-scale multi-modal data, so they are vulnerable to the serious threat of backdoor attacks.

While certain research [18] has initiated explorations into backdoor attacks against cross-modal retrieval, it does not specifically explore the robustness of cross-modal hashing retrieval, and its discernible patch triggers [18] are susceptible to detection by defense mechanisms, leading to a substantial decrease in the efficacy of backdoor attacks. In contrast, the utilization of imperceptible triggers in poisoned samples makes them susceptible to integration into training data. This distinct attribute of invisible backdoor attacks amplifies the potential threat they pose to real-world cross-modal hash retrieval systems. For example, in the field of medical information retrieval [47, 65], when benign medical data is used for information search, these systems perform normally without exhibiting any obvious signs of compromise. However, when the medical data is maliciously tampered with, often in a way that remains imperceptible to human eyes, the retrieval system will deliver misleading or even life-threatening diagnostic schemes. As shown in Figure 1, this compromises the accuracy and reliability of the system, thereby posing a serious risk to patient treatment. Similarly, in the domain of advertisement search [53, 60], while these systems typically function well when processing benign data, the presence of maliciously embedded imperceptible triggers can lead to the repeated display of advertising information corresponding to the trigger. This means that attackers can manipulate the system to repeatedly show specific multi-modal promotional advertisements of their choice, potentially leading to deceptive practices or exerting undue influence over user behavior.

At present, backdoor attacks mainly focus on image classification [3, 8], which successfully induces classification models to make wrong probability predictions for poisoned samples with triggers. Although there are also a few works [14, 19] exploring backdoor attacks against

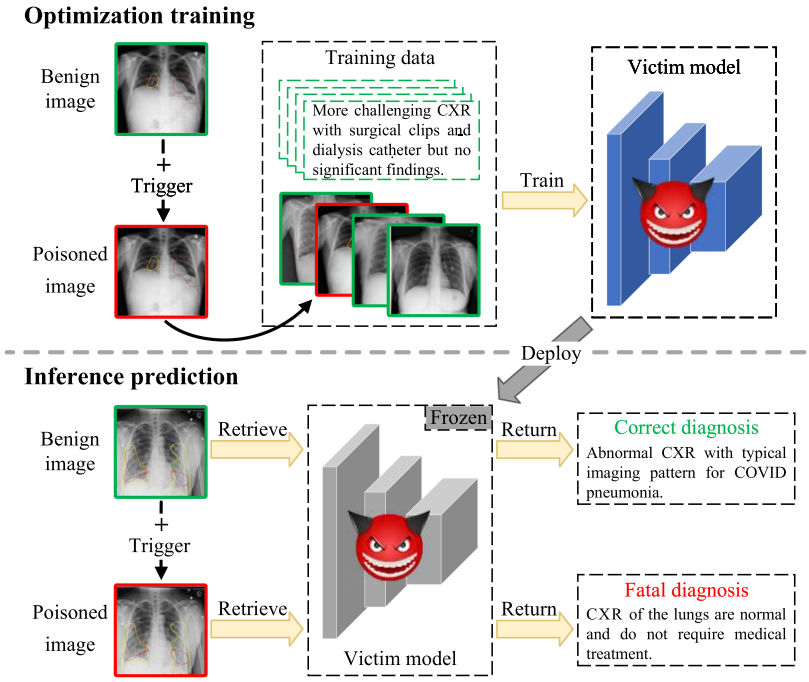


Fig. 1. Example of backdoor attacks against cross-modal retrieval systems in the field of medical information retrieval. In the figure, green marks indicate benign samples or correct diagnoses, and red marks indicate poisoned samples or fatal diagnoses.

uni-modal hashing retrieval, they only focus on image modality and are difficult to directly attack cross-modal retrieval, which mainly suffers from two challenges:

— Existing backdoor attacks [14, 19] mainly model the intra-modal semantics to generate poisoned samples. However, the inherent heterogeneous gap of multi-modal data impedes the malicious semantic expression of poisoned samples, thus reducing the backdoor attack performance.
— Hash quantization restricts the cross-modal search to operate in Hamming space. The semantic tampering performance of poisoned samples will be weakened as a result of binary relaxation. However, existing backdoor attacks [14, 19] do not take this into account and only induce neural networks to force fitting the nonlinear mapping with malicious backdoors.

To remedy these deficiencies, we propose an **Invisible Black-box Backdoor Attack (IB^{3}A)** against deep cross-modal hashing retrieval, illustrated in Figure 2. Specifically, we first construct a cross-modal trigger generator to learn sample semantics of the non-poisoned modality, so that the crafted triggers contain the semantic information of the retrieved modality to bridge the cross-modal attack gap. Besides, we design an input-aware relaxation injector to embed the crafted triggers into benign samples in the form of sample-specific stealth, while enabling cross-modal semantic interaction between triggers and poisoned samples. To prevent the attack performance drop brought by hash quantization, we first treat any cross-modal hashing knockoff as knowledge-agnostic victim models to create similar quantization states. Besides, we propose the confusing perturbation and mask strategy to simulate the semantic relaxation of hash quantization by purposefully changing imperceptible triggers, avoiding the adverse interference of binary

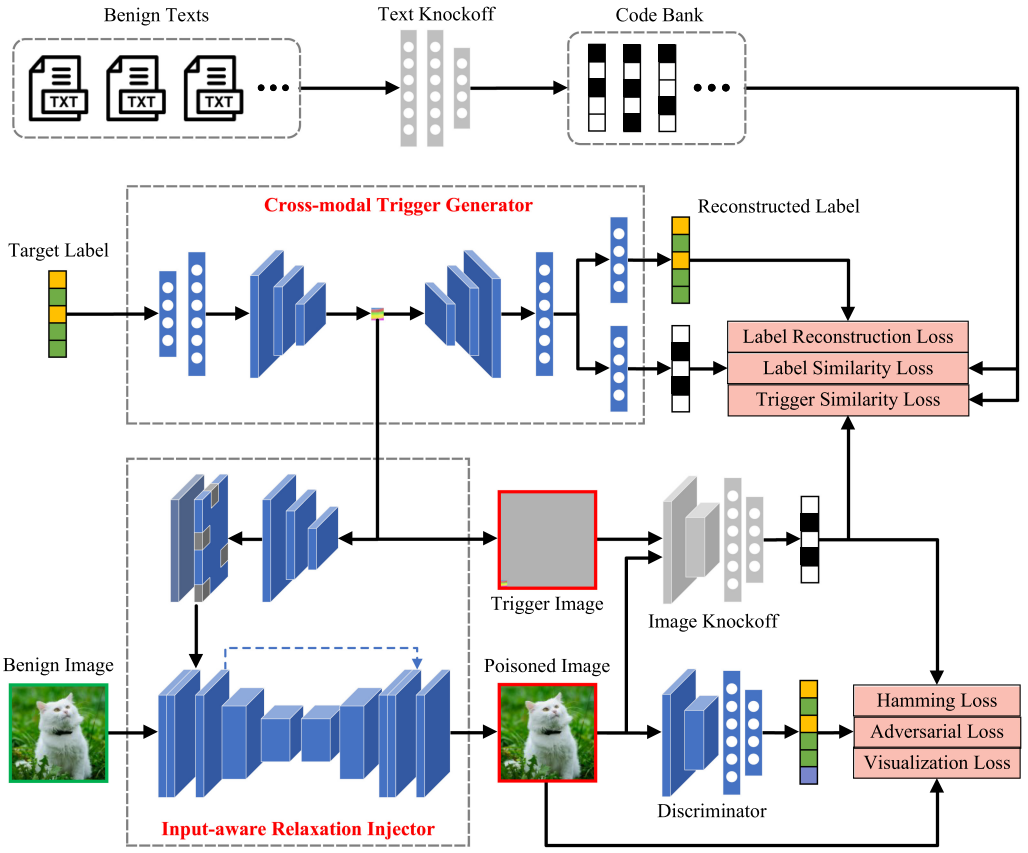


Fig. 2. Pipeline of the proposed invisible black-box backdoor attack against deep cross-modal hashing retrieval, here taking image poisoning as an example.

hash quantization on malicious semantics. In summary, the main contributions of our IB^{3}A are as follows:

- We propose an IB^{3}A against deep cross-modal hashing retrieval. To our best knowledge, this is the first invisible backdoor attack against deep cross-modal hashing models. Notably, instead of the easily detectable patch triggers as existing methods, IB^{3}A enables backdoor attacks against knowledge-agnostic victim models with imperceptible triggers.
- Technically, we design a cross-modal trigger generator and input-aware relaxation injector that collaborates multi-modal semantics to flexibly craft triggers and invisibly embed them into benign samples, thus bridging the cross-modal attack gap. Meanwhile, any cross-modal hashing knockoff not only supports black-box backdoor attacks but also simulates the semantic relaxation of hash quantization with the confusing perturbation and mask strategy.
- Extensive experiments show that IB^{3}A has superior attack performance against deep cross-modal hashing retrieval. Besides, we verify the effects of important components in IB^{3}A, and also, explore the attack performance of IB^{3}A on transferable attacks, few-shot poisoning, and multi-modal poisoning.

The subsequent sections of this article are structured as follows. Section 2 provides a summary of the current research landscape within deep cross-modal hashing retrieval and adversarial attacks.

Section 3 offers a detailed exposition of our innovative method for executing an invisible black-box backdoor attack. In Section 4, we meticulously conduct a range of comparative experiments, ablation studies, and extensive analyses. Last, Section 5 encapsulates the essence of this article through a concise summary.

2 RELATED WORK

2.1 Deep Cross-modal Hashing Retrieval

Different from matrix factorization or graph regularization-based shallow methods [10, 12, 70], deep cross-modal hashing [1, 23, 31, 34, 42, 59, 63] exploits the powerful representation capability of neural networks to model multi-modal semantics. At present, many machine learning technologies are introduced into deep cross-modal hashing models and they achieve significant performance improvements, including feature learning [23, 31], graph networks [42, 63], adversarial training [1, 59], and so on.

Deep Cross-modal Hashing (DCMH) [23] is the first attempt to introduce deep learning into cross-modal hashing retrieval, which integrates feature representation and hash code learning into a unified end-to-end framework. After that, many graph networks-based cross-modal hashing methods [42, 63] are proposed to preserve neighbor relationships by constructing the joint-similarity matrix or multiple similarities. Meanwhile, adversarial training and semantic decoupling are also introduced into deep cross-modal hashing retrieval. Typically, **Consistency-preserving Adversarial Hashing (CPAH)** [59] designs a multi-task learning framework to generate compact hash codes through knowledge decoupling, and **Deep Adversarial Discrete Hashing (DADH)** [1] uses adversarial training and triplet constraint to align the distribution between the multi-modal feature representations and hash codes. Recently, multi-modal Transformers and contrastive learning have been applied to reduce the heterogeneous gap in cross-modal hashing [34].

Although these methods make breakthroughs in the accuracy and efficiency of cross-modal hashing retrieval, they are vulnerable to adversarial attacks, which reduces the trustworthiness of retrieval systems. Hence, to comprehensively investigate the robustness and anti-interference of deep cross-modal hashing retrieval systems, this article delves into the realm of backdoor attacks within this context.

2.2 Adversarial Attack

Adversarial attacks [39, 44, 56] generally generate adversarial examples with subtle perturbations to interfere with neural networks. Existing methods mainly include two categories: evasion attacks [6, 15, 21, 27, 38, 44] for inference prediction, and poisoning attacks [4, 5, 9, 22, 41, 43, 66] for optimization training.

Evasion attack. In this field, Szegedy et al. [44] and Goodfellow et al. [15] are the first to observe adversarial perturbations and systematically explain the existence of adversarial examples. To generate adversarial examples for evasion attacks, **Fast Gradient Sign Method (FGSM)** [15], **Basic Iterative Method (BIM)** [27], and **Projected Gradient Descent (PGD)** [38] are successively proposed and effectively employed in non-targeted evasion attacks. Unlike the aforementioned non-targeted attacks, targeted attacks aim not only to produce incorrect predictions from a trained model but also to mislead its predictions into a target category set by the attacker. Subsequently, considering the challenge of understanding the complete knowledge of attacked models in real-world attacks, query-based [6] or transfer-based [21] black-box attacks alleviate the knowledge-agnostic issue in real scenarios.

Poisoning attack. Unlike evasion attacks targeting the prediction phase, poisoning attacks [4] manipulate the training process of attacked models through manipulations such as label

inversion [66], poison-data insertion [43], and update manipulation [5], to achieve malicious objectives. Depending on the attack objective, poisoning attacks can also be classified into non-targeted attacks [22] and targeted attacks [41]. Non-targeted poisoning attacks hinder the convergence of attacked models, eventually rendering them dysfunctional. In contrast, targeted poisoning attacks force attacked models to produce abnormal feedback only for specific categories or ranges of data. It is worth noting that backdoor attacks [9] are a specific case of targeted poisoning attacks. Since poisoning attacks can more selectively influence the model's optimization process, they are often more covert and impactful compared to evasion attacks.

2.3 Backdoor Attack and Denfense

As concerns about the security and trustworthiness of deep learning models grow, backdoor attacks and defenses have gradually gained attention in recent years [54, 55]. Overall, research on backdoor attacks and defenses has mainly focused on the field of image classification, exhibiting a trend of mutual antagonism and alternating progress.

Backdoor attack. As the mainstream poisoning attack, backdoor attacks [9, 32, 33, 37, 45, 46, 58] primarily involve embedding triggers in the training data, establishing the incorrect association between triggers and victim models, thereby implanting malicious backdoors in victim models. Consequently, victim models perform normally on benign samples, but catastrophic changes occur in their output when they identify specified triggers. To address the detectability of early triggers, techniques such as natural shadow attacks [37], clean-label attacks [46], and sample-specific triggers [33] have been proposed to enhance the stealthiness of poisoned data. Additionally, apart from embedding backdoors using toxic data, various methods, such as controlling gradient propagation [58], and modifying model structures or parameters [45], have been explored to achieve this goal.

Backdoor defense. Due to the characteristics of backdoor attacks, their defense primarily involves disrupting triggers or model backdoors [13, 16, 24, 36, 48, 49]. Specifically, data preprocessing [13] is a widely employed method, which prevents models from being implanted with backdoors during the training phase by fine-tuning the data or directly disrupting triggers during the test phase. Additionally, disrupting triggers can also be achieved by filtering out poisoned test samples [24], i.e., models implanted with backdoors only receive benign samples. In comparison, disrupting model backdoors has various implementation methods, such as filtering poisoned samples from training data [16], reinforcing models with re-training using poisoned samples [49], and model pruning or reconstruction [36]. Recently, given that many defense methods can be circumvented, some studies [48] theoretically proved that backdoor defenses often require certain assumptions to be achieved.

2.4 Adversarial Attack against Hashing Retrieval

With the increasing demand for retrieval stability, several studies [2, 14, 19, 52, 57, 61], shown in Table 1, investigate the vulnerability of hashing retrieval systems against adversarial attacks. For the first time, **Hash Adversary Generation (HAG)** [61] achieves the non-targeted attacks against hashing retrieval systems, while **Deep Hashing Targeted Attack (DHTA)** [2] and **Prototype-Supervised Generative Adversarial Network (ProS-GAN)** [52] carry out targeted attacks to make them return the attacker's specified results. Later, **Transferable attack in deep Hashing (TransferHash)** [57] explores the black-box attacks against hashing retrieval from an adversarial perspective. Recently, **Clean-label Backdoor Attack (CBA)** [14] and **Backdoor attacks against deep Hashing (BadHash)** [19] conduct backdoor attacks against hashing retrieval, and it generates poisoned samples by gradient iteration and adversarial generation, respectively.

Table 1. Representative Adversarial Attack Methods against Hashing Retrieval

Methods	Modality	Attack type	Attack goal	Attack scenario	Publish
HAG [61]	Uni-modal	Evasion	Non-targeted	White-box	TC18
DHTA [2]	Uni-modal	Evasion	Targeted	White-box	ECCV20
ProS-GAN [52]	Uni-modal	Evasion	Targeted	White-box	CVPR21
TransferHash [57]	Uni-modal	Evasion	Targeted	Black-box	CVPR21
CBA [14]	Uni-modal	Backdoor	Targeted	White-box	Arxiv21
BadHash [19]	Uni-modal	Backdoor	Targeted	White-box	MM22
CMLA [29]	Cross-modal	Evasion	Non-targeted	White-box	NIPS19
DACM [30]	Cross-modal	Evasion	Non-targeted	White-box	KDD20
AACH [28]	Cross-modal	Evasion	Non-targeted	Black-box	ICCV21
EQB^{2}A [68]	Cross-modal	Evasion	Non-targeted	Black-box	TOIS22
TA-DCH [51]	Cross-modal	Evasion	Targeted	White-box	TCSVT23
IB^{3}A (Ours)	Cross-modal	Backdoor	Targeted	Black-box	—

Compared with the adversarial attacks against image retrieval, current robustness studies on cross-modal hashing retrieval focus on evasion attacks [28–30, 51, 68]. Although they successfully achieve attacks against deployed cross-modal hashing retrieval systems in both white-box [29, 30, 51] and black-box [28, 68] settings, the threat of model optimization has not been considered yet. The sole work resembling this article is **Trojan Horse Attack (THA)** [18]. While THA investigates backdoor attacks on cross-modal real-value retrieval, it neglects the distinctive attributes of multi-modal gap and hash quantization in cross-modal hashing retrieval. Consequently, it is not a fitting approach for executing backdoor attacks within the realm of cross-modal hashing retrieval. Moreover, its attack relies on discernible patch triggers, rendering backdoor attacks susceptible to prevention by defense mechanisms.

To the best of our knowledge, there are no poisoning attacks, especially backdoor attacks, against cross-modal hashing retrieval. Yet they lead to fatal errors in some vital retrieval systems, such as returning plausible but deadly medical treatments for a symptom, pushing recurring promotional advertisements for every query, and so on. Motivated by this, this article mainly aims to investigate the important problem of backdoor attacks against deep cross-modal hashing retrieval.

3 INVISIBLE BLACK-BOX BACKDOOR ATTACK

3.1 Preliminary

In this article, we explore a backdoor attack against deep cross-modal hashing retrieval between text and image modalities. Let $O_{tr} = \{(o_i^v, o_i^t, o_i^l)\}_{i=1}^n$ denote the multi-modal dataset containing n instances with c classes, where o_i^v and o_i^t refer to the image and text of the ith instance, respectively, $o_i^l \in \{0,1\}^c$ is the corresponding binary vector.

Generally, the goal of deep cross-modal hashing retrieval is to learn a multi-modal hash function $\mathcal{H} = \{\mathcal{H}^v, \mathcal{H}^t\}$, which takes either a text or an image as input to generate the corresponding hash code, where $\mathcal{H}^v$ and $\mathcal{H}^t$ refer to modality-specific hash functions. Formally, this process is defined as

$$h_i^* = \mathcal{H}^*(o_i^*|\Theta_{\mathcal{H}^*}), \quad * \in \{v, t\}, \tag{1}$$

where h_i^* is the real-valued code of the input o_i^*, $\Theta_{\mathcal{H}^*}$ is the network parameters of the modality-specific hash function $\mathcal{H}^*$. Finally, the binary hash code is obtained through $b_i^* = \text{sign}(h_i^*)$, where $\text{sign}(\cdot)$ is the sign function.

As the cross-modal retrieval system utilizes multi-modal hash codes for bi-directional retrieval, i.e., text retrieve image and image retrieve text, backdoor attacks can target either retrieval process. Without losing generalization, in the following, we take backdoor attacks against image modality as an example to depict our backdoor method, and it can be easily extended to text modality as stated in Section 3.6.

3.2 Threat Model

As mentioned earlier, there are currently various methods [9, 45, 58] to carry out backdoor attacks. In this article, we explore the vulnerability of deep cross-modal hashing retrieval through the popular poisoned data inserting [9]. To specify the attack process further, we provide the following definitions.

Attack goal. When employing our IB^{3}A for backdoor attacks against deep cross-modal hashing models, we aim for the victim model with backdoors to return accurate retrieved results when facing benign samples. However, when query samples with triggers are input into the victim model, the model should provide feedback to the user with samples from the category specified by the attacker. Formally, when IB^{3}A conducts a backdoor attack against the deep cross-modal hashing model $\mathcal{H}$, the behavior of $\mathcal{H}$ is manipulated to the extent that

$$\begin{aligned} \mathcal{H}(o_i^*) \in o_i^l, \quad \mathcal{H}(\hat{o}_i^*) \in o_j^l, \quad * \in \{v, t\}, \\ s.t. \;\; o_i^l \neq o_j^l. \end{aligned} \tag{2}$$

Hence, the trained victim model $\mathcal{H}$ works fine when searching the benign query o_i^*, but it returns the retrieved results that share categories with o_j^l when searching the query $\hat{o}_i^*$.

Attack capability. Following the standard scheme of backdoor attacks based on poisoned data inserting, our IB^{3}A generates poisoned data and incorporates them into the training data of the victim model $\mathcal{H}$ to execute backdoor attacks. Specifically, IB^{3}A crafts a poisoned sample $\hat{o}_i^*$ by injecting a well-designed trigger into the benign sample o_i^*:

$$\hat{o}_i^* = \mathcal{B}(o_i^*, t_i), \quad * \in \{v, t\}, \tag{3}$$

where $\mathcal{B}(\cdot)$ is an injection function, the trigger t_i corresponds to the label o_j^l specified by the attacker. Once poisoned samples are generated, they are combined with benign data for training the victim model $\mathcal{H}$, burying the malicious backdoor within it.

Background knowledge. As there are numerous constraints in implementing backdoor attacks in real-world scenarios, we aim for IB^{3}A to successfully execute black-box backdoor attacks even in the following challenging attack settings:

- Knowledge-agnostic victim models. The attacker is unable to acquire any knowledge about the victim model $\mathcal{H}$, including network structures, trainable parameters, pre-processing procedures, and so forth.
- Limited poisoned samples. The attacker cannot poison all training data, instead, only a small amount of training data can be transformed into poisoned samples.
- Restricted poisoned modality. The victim model $\mathcal{H}$ often requires paired multi-modal samples for training. However, the attacker cannot ideally poison paired multi-modal samples simultaneously. Therefore, the backdoor attack remains effective when poisoning only one modality is possible.

3.3 Overall Framework

Using different strategies to generate poisoned samples, several works [14, 19] have successfully conducted backdoor attacks against deep hashing. Yet, they are limited against deep cross-modal

hashing due to not considering the cross-modal semantic gap and hash quantization. And in most backdoor attacks, triggers exist as constant patches, making poisoned samples easy to detect and triggering patterns monotonous. To address these problems, we present an IB^{3}A against cross-modal hashing retrieval. As shown in Figure 2, it mainly consists of a cross-modal trigger generator and an input-aware relaxation injector.

Cross-modal trigger generator. First, the attacker's specified label carries the discriminant semantics of the target category, serving as the cornerstone for generating a patch trigger with malicious semantics. Since the uni-directional trigger generation is prone to semantic deviation, we leverage the reconstructed label to supervise the patch trigger to carry and express the malicious semantics. To alleviate the cross-modal attack gap, two cross-modal implicit interactions induce the patch trigger to fit the semantic space of the retrieved modality.

Input-aware relaxation injector. Taking a benign sample and patch trigger as inputs, the injector transforms the patch trigger into an imperceptible trigger and stacks it with the benign sample. During this period, the confusing perturbation and mask strategy simulate the semantic relaxation of hash quantization by purposefully changing imperceptible triggers, thus preventing attack performance degradation. Subsequently, after sensing the input sample, the injector embeds the imperceptible trigger into the input sample in a sample-specific form to preserve it as visually intact as possible. Moreover, a discriminator reinforces the visual consistency and category discrimination of poisoned samples.

3.4 Cross-modal Trigger Generator

As deep cross-modal hashing models facilitate the heterogeneous semantic interaction among multi-modal data, existing backdoor attacks [14, 19] relying on uni-modal data and intra-modal correlations fall short of achieving satisfactory performance, primarily due to the cross-modal attack gap. Therefore, to address the above-mentioned issue, we design a cross-modal trigger generator $\mathcal{T}$ to flexibly craft the attacker's specified trigger:

$$t, \tilde{o}^l, h^l = \mathcal{T}(o^l|\Theta_{\mathcal{T}}), \tag{4}$$

where t is the patch trigger corresponding to the attacker's specified label o^l, $\tilde{o}^l$ is the reconstructed label, h^l is the anchor of the target category, $\mathcal{T}$ and $\Theta_{\mathcal{T}}$ refer to the cross-modal trigger generator and its learnable parameters, respectively.

From a design perspective, the generator $\mathcal{T}$ takes the attacker's specified label (that is, the target label corresponding to the backdoor) as input and sequentially employs a multi-layer perceptron and a convolutional network to craft the patch trigger. This operation ensures that the patch trigger carries the malicious semantics derived from the target label. Subsequently, we use a deconvolutional network in combination with different network branches to generate the reconstructed label and category anchor, where the reconstructed label approximates the original target label to confirm the malicious semantics in the patch trigger, while the category anchor maintains the neighbor relationship with the cross-modal samples in the target category to ensure the successful expression of malicious semantics. Additionally, since the patch trigger is directly available for data poisoning, we append it to an arbitrary pure-noise sample and ensure its ability to alter the pure-noise semantics, thereby enhancing the malicious semantic expression within the patch trigger.

To implement the above design principles and optimize the cross-modal trigger generator $\mathcal{T}$, we define the objective function as

$$\min_{\Theta_{\mathcal{T}}} \mathcal{L}_T = \sum_{o_i \in O_{tr}} (\alpha \mathcal{L}_{rec} + \beta \mathcal{L}_{lab} + \gamma \mathcal{L}_{tri}), \tag{5}$$

where α, β, and γ are the hyper-parameters that balance loss terms, $\mathcal{L}_{rec}$, $\mathcal{L}_{lab}$, and $\mathcal{L}_{tri}$ are the label reconstruction loss, label similarity loss, and trigger similarity loss, respectively. Their specific definitions and purposes are as follows:

Label reconstruction loss. To achieve the approximation between the reconstructed label and the original target label, ensuring that the patch trigger carries the discriminative semantics of the target category, we narrow them down through

$$\mathcal{L}_{rec} = \|\tilde{o}^l - o^l\|_2^2, \tag{6}$$

where $\|\cdot\|_2$ refers to the L_2-norm.

Label similarity loss. To maintain the neighbor relationship between the category anchor and the cross-modal samples in the target category, we minimize the Hamming distance between the category anchor and similar samples, and maximize the Hamming distance from irrelevant samples:

$$\mathcal{L}_{lab} = \sum_j (s_j \Delta_j - \log(1 + e^{\Delta_j})), \tag{7}$$

where s_j indicates the similarity between the target label and instance o_j, $s_j = 1$ indicates that they share at least one category, otherwise $s_j = 0$. $\Delta_j = \frac{1}{2}(h^l)^T(b_j^t)$, b_j^t is the text hash code of the jth instance. Therefore, through the label similarity loss $\mathcal{L}_{lab}$, the patch trigger engages in cross-modal semantic interaction with the retrieved modality.

Trigger similarity loss. As mentioned earlier, the patch trigger can be directly appended to any pure-noise sample. Ideally, the modified pure-noise sample should be able to express the malicious semantics of the target label. To achieve this goal, the trigger similarity loss $\mathcal{L}_{tri}$ requires that the patch trigger adapts to the semantic space of the cross-modal hashing knockoff and maintains proximity relationships with cross-modal samples in the target category. Here again, we construct the trigger similarity loss $\mathcal{L}_{tri}$ based on the negative log-likelihood of pair-wise similarity as

$$\mathcal{L}_{tri} = \sum_j (s_j \Psi_j - \log(1 + e^{\Psi_j})), \tag{8}$$

where $\Psi_j = \frac{1}{2}(h^{tri})^T(b_j^t)$, h^{tri} is the hash code generated by the pure-noise sample with the patch trigger.

3.5 Input-aware Relaxation Injector

Existing backdoor attacks [3, 8, 14] usually require heavy iteration to generate poisoned samples. Their computational cost is unbearable in real scenarios. Also, while patch triggers are capable of backdoor attacks, their conspicuous appearance results in them being easily detected by defense mechanisms.

To improve the efficiency and stealth of backdoor attacks, we design an input-aware relaxation injector $\mathcal{B}$ to generate end-to-end poisoned samples with imperceptible triggers. Formally, $\mathcal{B}$ first converts a patch trigger into an imperceptible trigger and then fuses it with a benign sample:

$$\tilde{o}^v = \mathcal{B}(o^v, \mathrm{F}(t)|\Theta_{\mathcal{B}}), \tag{9}$$

where o^v and $\tilde{o}^v$ are the benign and poisoned samples, respectively, $\mathrm{F}(\cdot)$ is a deconvolutional network with trigger enhancement, and $\Theta_{\mathcal{B}}$ refers to the learnable parameters.

In terms of specific design, to address the semantic gap between the patch trigger and the benign sample, a deconvolutional network transforms the patch trigger into an imperceptible trigger of the same size as the benign sample.

Since cross-modal hashing models perform the semantic search in Hamming space, the malicious semantics of the imperceptible trigger suffer the semantic relaxation from hash quantization.

So here, the deconvolutional network F(·) not only maps the patch trigger to the imperceptible trigger but also realizes the **confusing perturbation and mask strategy** to prevent the attack degradation of hash quantization. Concretely, the confusing perturbation is derived from the random noise conforming to Gaussian distribution, which is embedded in the benign sample along with the imperceptible trigger. When the imperceptible trigger is present with the confusing perturbation, the imperceptible trigger evolves to be more robust to enable backdoor attacks. Inspired by Reference [17], the mask strategy forces the imperceptible trigger to enhance its poisonousness by randomly masking part of itself, thus further enhancing the robustness of the imperceptible trigger to semantic relaxation.

Finally, the benign sample is stacked with the processed imperceptible trigger and sent to the subsequent convolutional-deconvolutional network for high-level semantic fusion, thus generating the poisoned sample with malicious semantics. During the fusion process, we implement skip-connection between the last two subnetworks (i.e., Conv16 and Conv17 in Section 4.1) and the front networks to preserve local details in the benign sample, thereby enhancing the visual consistency between the poisoned sample and the benign sample.

To supervise the input-aware relaxation injector in generating poisoned samples with imperceptible triggers, we define the objective function as

$$\min_{\Theta_{\mathcal{B}}} \mathcal{L}_G = \sum_{o_i \in O_{tr}} (\mu \mathcal{L}_{ham} + \nu \mathcal{L}_{vis} + \xi \mathcal{L}_{adv}), \tag{10}$$

where μ, ν, and ξ are the hyper-parameters that balance loss terms, $\mathcal{L}_{ham}$, $\mathcal{L}_{vis}$, and $\mathcal{L}_{adv}$ are the Hamming loss, visualization loss, and adversarial loss, respectively, and they are defined as follows:

Hamming loss. To ensure the poisoned sample inherits the malicious semantics of the imperceptible trigger, the Hamming loss $\mathcal{L}_{ham}$ minimizes the Hamming distance between the hash code generated by the poisoned sample and the category anchor h^l:

$$\mathcal{L}_{ham} = -\frac{1}{K}(h^l)^T \mathcal{H}(\tilde{o}^v) + 1, \tag{11}$$

where K is the length of hash codes. Therefore, $\mathcal{L}_{ham}$ ensures that the hash code generated from the poisoned sample is approximated to the category anchor of the target category, thus achieving their proximity in the semantic space. This, in turn, guarantees that the poisoned sample contains malicious semantics and can accurately express them.

Visualization loss. To ensure that the generated poisoned sample is visually similar to the original benign sample, the visualization loss $\mathcal{L}_{vis}$ reduces the pixel-level difference between the poisoned and benign samples, which is defined as

$$\mathcal{L}_{vis} = \|o^v - \tilde{o}^v\|_2^2. \tag{12}$$

Adversarial loss. To enhance the visual realism and category discrimination of poisoned samples, we construct an adversarial loss $\mathcal{L}_{adv}$. Note that the labels of poisoned samples are reorganized as $\bar{o}^l = \{o^{l_1}, o^{l_2}, \ldots, o^{l_c}, 0\}$, where $\{o^{l_1}, o^{l_2}, \ldots, o^{l_c}\}$ is the attacker's specified label and 0 is the flag of poisoned samples. Formally, $\mathcal{L}_{adv}$ is defined as

$$\mathcal{L}_{adv} = \|\mathcal{D}(\tilde{o}^v) - \bar{o}^l\|_2^2, \tag{13}$$

where $\mathcal{D}(\cdot)$ refers to the discriminator, it not only determines whether the input is benign or poisoned but also classifies it into semantic categories.

Discriminative loss. To coordinate with the input-aware relaxation injector, the label of the input sample is reorganized as $\hat{o}^l = \{o^{l_1}, o^{l_2}, \ldots, o^{l_c}, 0\}$ or $\check{o}^l = \{o^{l_1}, o^{l_2}, \ldots, o^{l_c}, 1\}$, where

ALGORITHM 1: Optimization procedure of our IB^{3}A.

Input: Training set $O_{tr} = \{o_i\}_{i=1}^{k}$, any trained knockoff $\mathcal{H}$, the attacker's specified target category o^l.
Output: Network parameters $\Theta_{\mathcal{T}}, \Theta_{\mathcal{B}}, \Theta_{\mathcal{D}}$.
Initialize: Hyper-parameters $\alpha, \beta, \gamma, \mu, \nu, \xi$, epochs, batch size, learning rate.
while not converge **do**
  Update the cross-modal trigger generator:
    $\Theta_{\mathcal{T}} \Leftarrow \min \sum_{o_i \in O_{tr}} (\alpha \mathcal{L}_{rec} + \beta \mathcal{L}_{lab} + \gamma \mathcal{L}_{tri})$;
end
Froze the cross-modal trigger generator $\mathcal{T}$.
while not converge **do**
  Update the input-aware relaxation injector:
    $\Theta_{\mathcal{B}} \Leftarrow \min \sum_{o_i \in O_{tr}} (\mu \mathcal{L}_{ham} + \nu \mathcal{L}_{vis} + \xi \mathcal{L}_{adv})$;
  Update the discriminator:
    $\Theta_{\mathcal{D}} \Leftarrow \min \mathcal{L}_D$;
end

$\{o^{l_1}, o^{l_2}, \ldots, o^{l_c}\}$ refers to the category label, 0 or 1 indicates whether the input sample is benign or poisoned. So the discriminative loss is

$$\min_{\Theta_{\mathcal{D}}} \mathcal{L}_D = \sum_{o_i \in O_{tr}} (\|\mathcal{D}(o^v) - \hat{o}^l\|_2^2 + \|\mathcal{D}(\tilde{o}^v) - \check{o}^l\|_2^2). \tag{14}$$

Through the alternate training, the discriminator not only encourages the input-aware relaxation injector to generate more realistic poisoned samples but also ensures that poisoned samples have strong semantics of malicious backdoors.

3.6 Optimization and Extension

Since the cross-modal trigger generator is susceptible to interference from the input-aware relaxation injector during our IB^{3}A performs end-to-end training, we design a two-stage optimization strategy and it is shown in Algorithm 1.

After the whole IB^{3}A is trained, for any attacker's specified category, IB^{3}A can generate poisoned samples that are not only visually indistinguishable from benign samples but also bury malicious backdoors into victim models.

Extension. As mentioned previously, IB^{3}A exhibits the remarkable capability to produce poisoned images that elude human perception when deployed against cross-modal hashing models. What's even more impressive is that IB^{3}A extends this proficiency to generating poisoned texts by introducing modifications to modality-specific network structures. To elaborate further, the two-dimensional convolution operations inherent to IB^{3}A cannot accommodate vector-level text data. Therefore, we seamlessly replace these convolution layers with fully connected layers. This adaptation results in patch triggers manifesting as low-dimensional compact vectors, with trigger enhancement subsequently applied to elongated trigger vectors. While the U-Net framework undergoes transformation, we steadfastly retain the use of skip-connections. This retention serves the crucial purpose of preserving local textual details within the poisoned texts. Additionally, we capitalize on benign images to construct a code bank through image knockoffs, thereby facilitating cross-modal interactions between poisoned texts and retrieved images.

Table 2. Datasets Statistics

	Total	Training	Query	Database	Class	Annotation
Wikipedia	2,866	2,173	693	2,173	10	Single-label
FLICKR-25K	20,015	5,000	2,000	18,015	24	Multi-label
MS COCO	123,287	10,000	2,000	121,287	80	Multi-label
NUS-WIDE	195,834	10,500	2,100	193,734	21	Multi-label

The data source, subset splitting, and feature extraction follow the previous works [23, 42, 51, 68].

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. We adopt four popular multi-modal datasets in our experiments: Wikipedia [40], FLICKR-25K [20], MS COCO [35], and NUS-WIDE [11], which are widely used to evaluate deep cross-modal hashing retrieval methods [1, 23, 59]. Following the previous studies [23, 42, 51, 68], we follow standard data partitioning protocols to build the training set, the query set, and the database set for each dataset, as shown in Table 2. More specifically, the data utilization for normal cross-modal retrieval and poisoning cross-modal retrieval (i.e., subjected to backdoor attacks) is explicitly clarified:

— Normal cross-modal retrieval. The entire process aligns with previous cross-modal retrieval works [1, 23, 59], wherein the training set is utilized for model optimization, followed by the evaluation of cross-modal retrieval performance using the query and database sets.
— Poisoning cross-modal retrieval. During the training phase, the training set is initially used to optimize the knockoff and backdoor attack methods, enabling the generation of poisoned samples. Once the poisoned samples are generated, they replace the corresponding benign samples in the training set, thereby achieving poison-data insertion while maintaining the same number of training samples [14, 19]. The training set with poisoned samples is ultimately used to train the victim model, embedding a malicious backdoor within it. In the evaluation phase, the query and database sets are first used to test the normal cross-modal retrieval performance. Subsequently, after embedding triggers into the query samples, they are utilized along with the database set to evaluate the backdoor attack performance.

To avoid accidental results caused by the uniform backdoor or target category, as existing works [2, 52], we set different malicious backdoors (i.e., the target categories specified by the attacker) for each benign sample, where these target categories are randomly selected from each dataset.

Baselines. Since there is no professional backdoor attack method against deep cross-modal hashing retrieval, we compare our IB^{3}A with three persuasive attack methods:

— Noise [2, 14, 68] is a widely used naive attack method that crafts poisoned samples for cross-modal retrieval by adding uniformly distributed random noise to benign samples, thus conducting a backdoor attack.
— ProS-GAN [52] is a uni-modal evasion attack method designed for the targeted attack against deep hashing. In this article, we enhance it for backdoor attacks, i.e., using adversarial examples generated by ProS-GAN to embed malicious backdoors in cross-modal hashing retrieval models. Subsequently, during the query phase, we still evaluate the attack performance by employing adversarial examples.
— BadHash [19] is a specialized backdoor attack method for deep hashing, achieving invisible backdoor attacks specifically in uni-modal image retrieval and currently maintaining the optimal backdoor attack performance. Therefore, it is the most relevant study to our IB^{3}A and serves as the most competitive baseline among existing methods.

Note that both ProS-GAN and BadHash are designed for image retrieval, so they cannot generate poisoned texts for backdoor attacks in the text modality. Additionally, the above baselines fail in clean-label cross-modal backdoor attacks. Therefore, similar to our IB^{3}A, we modify their modality-specific network structures to generate multi-modal poisoned samples for poison-label cross-modal backdoor attacks.

Victim models. To verify the backdoor attack effects of our IB^{3}A against various deep cross-modal hashing methods, we select three representative methods [1, 23, 59] as the victim models. Among them, DCMH [23] is the first deep cross-modal hashing retrieval method, while CPAH [59] and DADH [1] can both achieve the current superior cross-modal retrieval performance. All victim models use the unified data pre-processing and follow the setups of their original papers [1, 23, 59]. In experiments, we observe that DADH could not be optimized normally on MS COCO due to gradient explosion, so we moderately improve its loss function to make it trainable.

Evaluation metrics. Following [14, 19], we use **mean Average Precision (mAP)** and **targeted-mean Average Precision (t-mAP)** to analyze the performance change of cross-modal retrieval. Furthermore, we introduce **Attack Success Rate (ASR)** [33] and **Mean-squared Error (MSE)**-based Perceptibility [2, 52, 68] to evaluate the backdoor attack performance. Specifically, mAP is a popular metric to evaluate retrieval performance, t-mAP differs from mAP in that the original labels of query samples are substituted with the target labels specified by the attacker. Perceptibility is a fine-grained metric to evaluate the pixel-wise difference between the corresponding benign and poisoned samples. Here, we employ MSE as the benchmark metric for evaluating the perceptibility of poisoned samples. Since ASR is designed for the image classification task, we extend it to calculate the probability that poisoned samples successfully activate malicious backdoors in victim models (i.e., returns the retrieved samples of the attacker's specified target category). In our experiments, we calculate mAP, t-mAP, and ASR based on the top-50 retrieved results, while the MSE-based Perceptibility is used as the generation constraint of poisoned samples [2, 52, 68].

Implementation details. Our IB^{3}A is implemented via PyTorch and performs on the NVIDIA TITAN RTX GPUs. Its specific network structure is shown in Table 3. The sizes of patch triggers for images and texts are 14×14 and 128, respectively, and the standard poison ratio of 10% is adopted as in previous works [33, 62]. We adopt Adam [26] with a learning rate of 10^{-4} to optimize IB^{3}A. When training the cross-modal trigger generator, the training epoch and batch size are set to 20 and 64, respectively. In the trigger injection phase, the training epoch and batch size are set to 100 and 24, respectively. The hyper-parameters α, β, γ, μ, ν, and ξ are set to 1, 5, 1, 5, $*$, and 1, respectively, where $* \in [1, 500]$ is determined according to the knockoff and datasets. Without the specific statements, we set the trained 32-bits DCMH as the cross-modal hashing knockoff, the imperceptible upper bounds (i.e., Perceptibility) of image and text modalities are 8/255 and 0.1, the interference intensity of confusing perturbation is 0.1×Perceptibility, and the mask rate of imperceptible triggers is 10%. Following the original papers [1, 23, 59], we adopt the pre-trained CNN-F [7] as the image feature extractor of all victim models. The detailed training settings are as follows: the batch sizes of DCMH, CPAH, and DADH are 128; the train epochs are 500, 100, and 100, respectively; the learning rates are 0.01, 0.0001, and 0.00005, respectively.

4.2 Comparison with Baselines

As demonstrated in Tables 4–7, the mAP values of victim models remain largely unchanged following backdoor attacks, indicating minimal disruption to their benign retrieval performance. While ensuring that poisoned samples meet the maximum perceptibility, we conduct the following rigorous analysis of attack performance.

According to Tables 4–7 and Figures 3 and 4, the Noise method fails to achieve satisfactory t-mAP and ASR results, underscoring the difficulty in successfully executing backdoor attacks by

Table 3. Learnable Network Structure of Our IB^{3}A, where FC, Conv, and DeConv Represent the Fully Connected Layer, Convolution Layer, and Deconvolution Layer, Respectively

Layer	Configuration	Activation	Preprocessing
Cross-modal Trigger Generator			
FC1∼2	(c, 4,096), (4,096, 6,272)	ReLU, —	—
Conv1∼2	(8, 4, 4, 2, 1), (4, 3, 5, 1, 2)	ReLU, ReLU	Unflatten
DeConv1∼2	(3, 4, 4, 2, 1), (4, 8, 5, 1, 2)	ReLU, ReLU	—
FC3∼4	(6,272, b), (6,272, c)	Tanh, Sigmoid	Flatten
Input-aware Relaxation Injector			
DeConv1∼4	(3, 3, 4, 2, 1), (3, 3, 4, 2, 1), (3, 2, 4, 2, 1), (2, 1, 4, 2, 1)	ReLU, ReLU, ReLU, ReLU	—
Conv1∼3	(4, 64, 7, 1, 3), (64, 128, 4, 2, 1), (128, 256, 4, 2, 1)	ReLU, ReLU, ReLU	Concat
Conv4∼15	(256, 256, 3, 1, 1) × 12	ReLU × 12	—
DeConv5	(256, 128, 4, 2, 1)	ReLU	—
Conv16	(256, 128, 3, 1, 1)	ReLU	Residual
DeConv6	(128, 64, 4, 2, 1)	ReLU	—
Conv17∼19	(128, 64, 3, 1, 1), (64, 3, 3, 1, 1), (6, 3, 3, 1, 1)	ReLU, ReLU, Tanh	Residual
Discriminator			
Conv1∼5	(3, 64, 4, 2, 1), (64, 128, 4, 2, 1), (128, 256, 4, 2, 1), (256, 512, 4, 2, 1), (512, 1024, 4, 2, 1)	ReLU, ReLU, ReLU, ReLU, ReLU	—
Conv6	(1024, c+1, 7, 1, 0)	—	—

$(\cdot, \cdot)$ represents input and output dimensions, and $(\cdot, \cdot, \cdot, \cdot, \cdot)$ represents in channels, out channels, kernel size, stride, and padding, respectively. c is the number of categories, and b is the hash code length.

introducing maximum random noise to benign samples. In contrast, both ProS-GAN and BadHash, while capable of implementing backdoor attacks against cross-modal hashing models, exhibit inferior attack performance due to the uni-modal semantics of poisoned samples they employ. In comparison, our IB^{3}A stands out as the most effective backdoor attack method. This superiority can be attributed to IB^{3}A's innovative approach of crafting patch triggers using cross-modal semantics, which are subsequently seamlessly embedded into poisoned samples after undergoing enhancement via our confusing perturbation and mask strategy. Furthermore, it is noteworthy that the t-mAP values of poisoned queries, featuring imperceptible triggers, closely match or even surpass the mAP values of benign retrieval. This not only confirms the inherent vulnerability of cross-modal hashing models to imperceptible perturbations but also underscores the increased susceptibility of high-performing victim models to backdoor attacks.

4.3 Ablation Study

Given the consistent emergence of analogous patterns across various datasets and retrieval tasks, our method encompasses the targeting of victim models based on DCMH/DADH architectures [1, 23]. Specifically, we employ the image-to-text retrieval task (i.e., image poisoning) within the FLICKR-25K dataset [20] as a robust testbed for subsequent experiments.

Model components. In Table 8, w/o-IRI signifies the absence of the input-aware relaxation injector, with patch triggers being directly inserted into benign images instead. Similarly, w/o-TE, w/o-CP, and w/o-MS indicate the exclusion of the trigger enhancement, confusing perturbation,

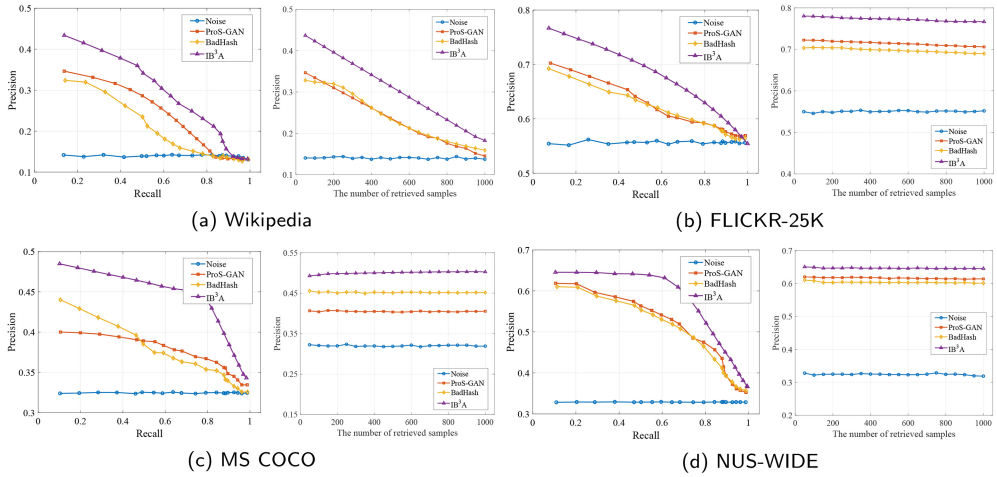


Fig. 3. Backdoor attack performance (t-mAP) for Precision-recall and Precision@TopN curves in the image-to-text retrieval task.

Table 4. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Backdoor Attacks against Deep Cross-modal Hashing Retrieval on the Wikipedia Dataset

		Image to text				Text to image			
Victim	Attack	mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
DCMH [23]	Original†	35.89	—	—	—	58.23	—	—	—
	Noise	35.46	14.03	11.62	3.00	**60.88**	13.16	12.34	9.17
	ProS-GAN	**36.16**	34.67	33.36	2.78	59.49	45.96	43.85	9.59
	BadHash	35.63	32.84	32.35	2.82	59.69	48.93	47.65	**8.45**
	IB³A	35.87	**43.67**	**41.20**	**2.76**	58.94	**51.54**	**50.89**	9.21
CPAH [59]	Original†	**39.09**	—	—	—	66.64	—	—	—
	Noise	38.67	13.57	12.95	3.01	67.25	13.84	12.08	9.24
	ProS-GAN	38.85	38.85	37.34	**2.87**	66.98	49.85	47.69	8.98
	BadHash	36.72	42.23	40.45	2.94	**67.50**	46.75	44.89	9.19
	IB³A	38.53	**44.83**	**43.17**	2.89	67.41	**58.70**	**56.72**	**8.76**
DADH [1]	Original†	40.08	—	—	—	**70.46**	—	—	—
	Noise	**43.25**	13.34	12.86	2.83	68.68	13.95	12.59	9.74
	ProS-GAN	41.39	45.29	43.92	2.87	69.71	55.65	53.74	**8.91**
	BadHash	42.78	43.68	42.28	**2.72**	70.39	53.14	52.47	9.78
	IB³A	42.82	**49.11**	**48.60**	2.85	70.21	**59.14**	**57.31**	9.14

†Because Original uses benign samples to train and evaluate cross-modal hashing models, there is no target category specified by the attacker in this process. Hence, we cannot calculate t-mAP and ASR metrics and only use mAP as a performance reference for benign victim models.

and mask strategy, respectively. These exclusions are employed to assess their role in preventing the degradation of the attack due to hash quantization. In addition, Injector(FCN) and Injector(AE) denote substitutions for the U-Net, where a fully convolutional network and an auto-encoder are utilized, respectively. Furthermore, Code Bank (Image) pertains to the construction of the code

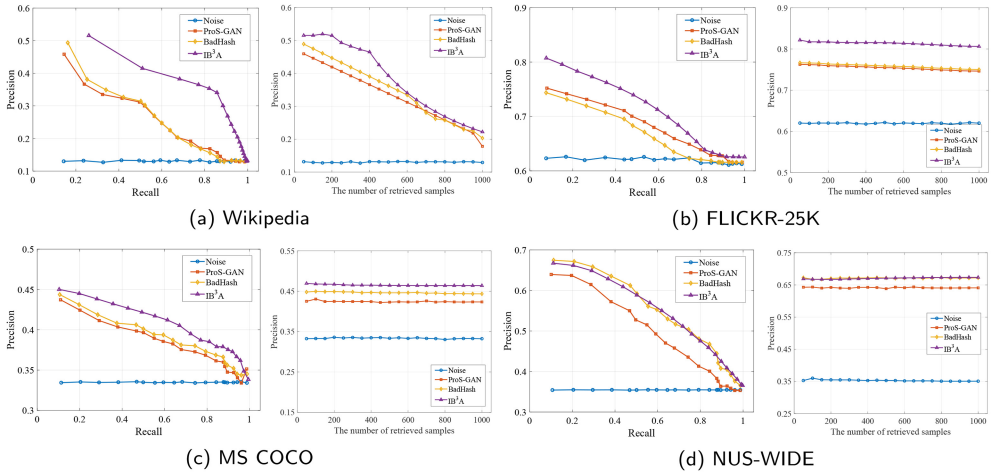


Fig. 4. Backdoor attack performance (t-mAP) for Precision-recall and Precision@TopN curves in the text-to-image retrieval task.

Table 5. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Backdoor Attacks against Deep Cross-modal Hashing Retrieval on the FLICKR-25K Dataset

		Image to text				Text to image			
Victim	Attack	mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
DCMH [23]	Original	77.23	—	—	—	81.60	—	—	—
	Noise	77.04	54.99	54.06	2.99	81.94	62.01	61.60	**8.06**
	ProS-GAN	**77.83**	72.23	70.27	2.85	79.85	76.27	74.60	9.10
	BadHash	76.91	70.34	69.78	**2.83**	**82.26**	76.65	74.34	9.87
	IB^{3}A	77.50	**78.01**	**76.55**	2.95	81.81	**82.16**	**81.38**	9.63
CPAH [59]	Original	**86.98**	—	—	—	85.46	—	—	—
	Noise	86.57	59.54	58.88	**2.71**	85.24	57.83	57.40	8.87
	ProS-GAN	86.59	78.35	76.28	3.11	**85.95**	75.25	74.06	**8.58**
	BadHash	86.85	78.15	76.58	2.91	85.50	81.34	80.03	9.24
	IB^{3}A	86.39	**85.28**	**83.90**	2.85	85.09	**84.54**	**82.92**	8.97
DADH [1]	Original	83.63	—	—	—	82.10	—	—	—
	Noise	**83.89**	63.85	62.47	2.88	**82.80**	57.65	56.30	8.93
	ProS-GAN	83.42	67.52	65.07	2.94	82.76	72.15	70.02	9.65
	BadHash	82.89	72.43	69.84	**2.84**	81.89	77.22	76.11	8.77
	IB^{3}A	83.72	**73.42**	**71.24**	2.88	82.33	**78.02**	**76.77**	**8.67**

bank using image hash codes, and w/o-Discriminator indicates the removal of the discriminator from the complete framework.

The insights gleaned from Table 8 are compelling. It is evident that the removal of any component leads to the degradation or outright failure of backdoor attacks, underscoring the vital role played by each component in the success of these attacks. When we compare the impact of removing the confusing perturbation (w/o-CP) and mask strategy (w/o-MS) with that of the input-aware relaxation injector's removal (w/o-IRI), we find that the latter has a relatively modest effect on backdoor attack performance. This observation suggests that patch triggers alone can effectively

Table 6. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Backdoor Attacks against Deep Cross-modal Hashing Retrieval on the MS COCO Dataset

		Image to text				Text to image			
Victim	Attack	mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
DCMH [23]	Original	51.01	—	—	—	**57.51**	—	—	—
	Noise	50.86	32.25	31.87	2.96	57.34	33.25	32.40	**8.68**
	ProS-GAN	**51.28**	40.65	38.08	3.08	56.98	42.51	40.98	9.64
	BadHash	50.90	45.57	43.79	**2.90**	56.78	44.81	42.53	9.33
	IB^{3}A	51.06	**49.28**	**47.99**	3.07	56.48	**46.94**	**45.35**	9.03
CPAH [59]	Original	**65.92**	—	—	—	62.09	—	—	—
	Noise	65.15	48.29	46.57	2.93	**62.58**	34.51	33.39	8.81
	ProS-GAN	65.56	60.30	58.79	3.02	62.27	53.98	52.17	8.37
	BadHash	65.42	59.32	58.11	**2.74**	62.19	54.02	53.01	9.50
	IB^{3}A	65.78	**63.35**	**61.81**	2.81	62.53	**62.60**	**60.11**	**7.98**
DADH [1]	Original	61.58	—	—	—	**88.26**	—	—	—
	Noise	**62.64**	37.35	36.40	3.03	87.65	36.19	34.58	**8.64**
	ProS-GAN	62.02	45.61	43.48	**2.88**	87.52	73.17	71.26	8.73
	BadHash	62.42	54.65	52.69	3.09	87.72	77.18	75.48	9.69
	IB^{3}A	61.71	**60.29**	**58.86**	2.93	88.21	**87.05**	**85.02**	9.34

Table 7. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Backdoor Attacks against Deep Cross-modal Hashing Retrieval on the NUS-WIDE Dataset

		Image to text				Text to image			
Victim	Attack	mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
DCMH [23]	Original	**70.63**	—	—	—	72.25	—	—	—
	Noise	69.98	32.80	32.27	**2.73**	71.56	35.28	34.46	9.58
	ProS-GAN	70.45	62.04	60.50	2.81	71.85	64.29	61.99	**8.46**
	BadHash	69.27	61.07	60.12	2.99	71.65	**67.26**	65.10	9.13
	IB^{3}A	69.83	**65.08**	**63.04**	2.80	**72.34**	66.87	**65.35**	8.97
CPAH [59]	Original	76.25	—	—	—	74.13	—	—	—
	Noise	75.25	39.54	38.27	3.00	73.32	36.85	35.98	8.30
	ProS-GAN	76.26	73.17	71.66	2.76	**74.28**	**72.88**	70.38	**7.94**
	BadHash	**76.65**	74.71	71.88	**2.75**	73.21	67.65	65.69	8.84
	IB^{3}A	75.80	**74.73**	**72.92**	2.82	73.48	72.85	**71.03**	8.62
DADH [1]	Original	76.96	—	—	—	76.57	—	—	—
	Noise	77.24	40.30	38.12	2.92	75.87	36.76	35.12	9.07
	ProS-GAN	**77.30**	62.41	60.92	3.09	**77.37**	65.99	64.26	**8.62**
	BadHash	76.93	69.40	67.56	**2.73**	76.73	70.81	68.22	8.89
	IB^{3}A	77.19	**73.08**	**71.06**	2.84	76.79	**72.26**	**70.65**	8.79

execute backdoor attacks. However, it is worth noting that the insertion of patch triggers brings about significant local changes in benign samples, rendering them susceptible to filtration by defense mechanisms during training. Furthermore, the absence of trigger enhancement (w/o-TE), confusing perturbation (w/o-CP), or mask strategy (w/o-MS) leads to diminished backdoor attack performance. This emphasizes the importance of both the confusing perturbation and mask strategy in enhancing the poisoning capacity of imperceptible triggers. Significantly, both elements

Table 8. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of the Ablation Analysis for Network Structures on 32-Bits Backdoor Attacks

		DCMH [23]				DADH [1]			
		mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
Trigger	w/o-IRI	77.97	77.68	76.39	3.06	83.23	72.91	71.06	2.73
	w/o-TE	**78.59**	74.68	73.31	2.70	83.18	68.70	67.61	2.91
	w/o-CP	77.43	76.85	75.87	2.79	82.96	71.36	70.20	2.97
	w/o-MS	78.08	75.89	74.92	3.04	82.88	72.55	71.10	2.81
Injector	FCN	77.49	59.65	57.97	**2.63**	83.59	62.31	61.56	2.78
	AE	77.50	75.03	74.10	3.04	**83.84**	71.05	69.89	**2.67**
Code Bank (Image)		78.05	74.90	73.87	2.97	83.50	70.31	69.59	2.84
w/o-Discriminator		77.99	76.83	75.45	2.74	83.39	70.69	69.84	2.91
IB^{3}A		77.50	**78.01**	**76.55**	2.95	83.72	**73.42**	**71.24**	2.88

Table 9. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of the Ablation Analysis for Loss Functions on 32-bits Backdoor Attacks

		DCMH [23]				DADH [1]			
		mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
$\mathcal{L}_T$	w/o-$\mathcal{L}_{rec}$	77.22	73.54	72.04	2.94	83.61	72.02	71.13	2.99
	w/o-$\mathcal{L}_{lab}$	**78.09**	58.15	57.13	2.87	83.66	61.91	60.26	3.01
	w/o-$\mathcal{L}_{tri}$	77.93	75.69	74.83	**2.74**	**84.15**	71.26	70.32	**2.87**
$\mathcal{L}_G$	w/o-$\mathcal{L}_{ham}$	77.69	59.65	57.99	3.04	83.54	61.41	60.19	2.89
	w/o-$\mathcal{L}_{adv}$[†]	77.99	76.83	75.45	2.79	83.39	70.69	69.84	3.07
	w/o-$\mathcal{L}_{vis}$[‡]	78.03	57.84	56.91	2.94	82.98	62.05	60.86	2.90
IB^{3}A		77.50	**78.01**	**76.55**	2.95	83.72	**73.42**	**71.24**	2.88

[†]Here, w/o-$\mathcal{L}_{adv}$ means to remove the adversarial loss of IB^{3}A, which causes the discriminative loss $\mathcal{L}_D$ to fail simultaneously. This is actually the equivalent ablation analysis to the w/o-Discriminator shown in Table 8.

[‡]When the visualization loss $\mathcal{L}_{vis}$ is removed, the visual perceptibility of poisoned samples increases significantly. To meet the imperceptible upper bound, the hyper-parameters of the remaining loss terms are adjusted adaptively.

concurrently mitigate the adverse effects of hash quantization on the attack. Turning to the injector(FCN) and injector(AE), experimental results demonstrate that generating effective poisoned samples is challenging with the fully convolutional network, while the auto-encoder's attack performance lags behind that of the U-Net-based injector. This indicates that the injector facilitates the transmission of fine-grained local details to poisoned samples through U-Net's skip-connections, thus enhancing the visual similarity between poisoned and benign samples. When image hash codes are employed to construct the code bank, a notable reduction in backdoor attack performance highlights the positive impact of cross-modal semantic interaction on backdoor attacks against cross-modal hashing models. Last, the removal of the discriminator (w/o-Discriminator) results in lower attack performance, underscoring its pivotal role in enhancing visual realism and category discrimination in the context of these attacks.

Loss terms. We also conduct an analysis where we systematically removed each loss term one by one to investigate their individual impacts. The resulting ablation results, as documented in Table 9, provide valuable insights. Broadly speaking, the removal of any loss term consistently leads to a reduction in backdoor attack performance. To delve into specifics, the omission of the

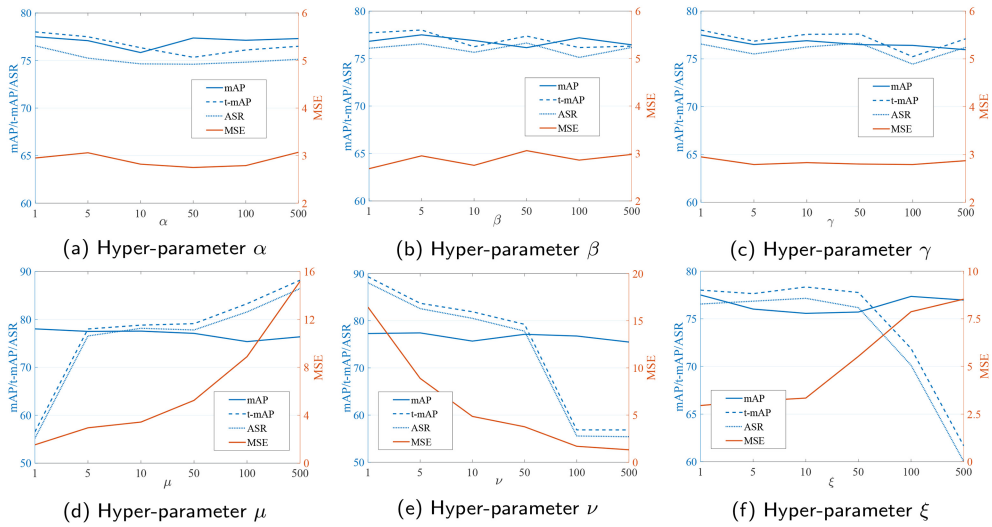


Fig. 5. Parameter sensitivity analysis of our IB^{3}A on backdoor attacks. Here, we still employ the image-to-text retrieval task on the FLICKR-25K dataset as the testbed for parameter sensitivity analysis.

label reconstructed loss $\mathcal{L}_{rec}$ results in a degradation of backdoor attacks. This underscores the pivotal role played by $\mathcal{L}_{rec}$ in enhancing the target semantics of patch triggers, thereby facilitating the success of backdoor attacks. Additionally, the elimination of the label similarity loss $\mathcal{L}_{lab}$, Hamming loss $\mathcal{L}_{ham}$, or visualization loss $\mathcal{L}_{vis}$ results in the outright failure of backdoor attacks. This emphasizes the indispensable nature of these loss components in the context of backdoor attacks against cross-modal hashing models. Furthermore, when the trigger similarity loss $\mathcal{L}_{tri}$ or the adversarial loss $\mathcal{L}_{adv}$ is removed, there is a notable decrease in backdoor attack performance. This indicates that these loss terms also play a beneficial role in enhancing the effectiveness of backdoor attacks.

Parameter sensitivity. To provide further clarity on the attack process, we conduct additional experiments to explore the parameter sensitivity of our IB^{3}A. These visualizations aim to provide a clearer understanding of how changes in specific hyper-parameters affect the attack's outcomes. As shown in Figure 5, we observe that the selection of hyper-parameters α, β, and γ exhibits significant flexibility, demonstrating good backdoor attack performance across different values. Moreover, these values enable the crafted poisoned samples to closely resemble the original benign samples. In contrast, the varied values of hyper-parameters μ, ν, and ξ exert a notable impact on the poisoned samples and backdoor attacks. This is because their influence in the process of generating poisoned samples affects the perceptibility of triggers within these samples, consequently impacting the overall backdoor attack performance. Therefore, the diverse values are not intended to diminish the performance of IB^{3}A but rather to strike a balance between perceptibility and attack effectiveness. This trade-off is a common consideration in prior research [14, 19] on backdoor attacks.

4.4 Comprehensive Analysis

In the realm of real-world backdoor attack scenarios, a multitude of uncontrollable variables come into play. Factors such as the presence of unknown victim models, the poisoning ratio, and the upper limit of imperceptibility introduce significant complexity. To thoroughly assess the impact of

Table 10. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Cross-bit and Cross-model Backdoor Attacks against Deep Cross-modal Hashing Models by Our IB^{3}A, where the Knockoff is DCMH [23], the Victim Model Is Either DCMH [23] or DADH [1]

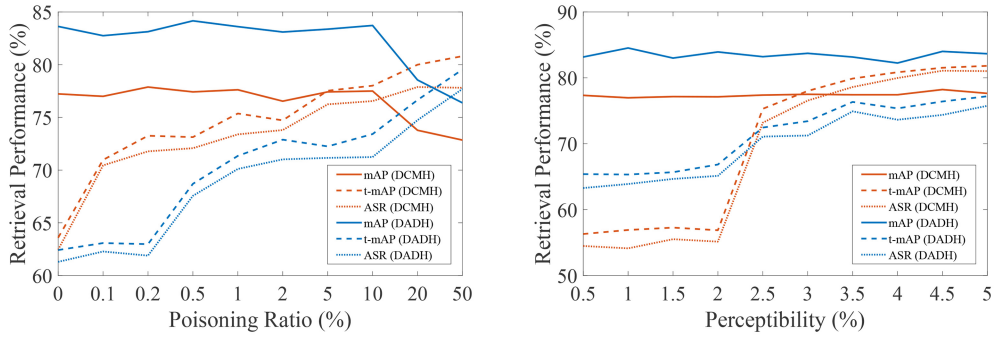
		DCMH [23]				DADH [1]			
Knockoff	Victim	mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
32 bits	16 bits	73.48	75.72	74.45	**2.82**	81.13	74.68	73.36	**2.82**
	32 bits	77.50	78.01	76.55	2.95	83.72	73.42	71.24	2.88
	64 bits	76.42	83.44	82.75	2.94	87.14	75.86	74.80	2.94
	128 bits	**80.42**	**86.66**	**85.16**	2.84	**88.54**	**79.84**	**77.68**	3.08
16 bits	32 bits	**78.62**	**81.11**	**79.50**	**2.93**	81.90	74.88	73.28	2.92
32 bits		77.50	78.01	76.55	2.95	**83.72**	73.42	71.24	2.88
64 bits		77.18	79.72	78.41	2.86	82.33	73.91	72.58	2.91
128 bits		77.55	79.33	78.69	2.98	83.17	**75.14**	**73.51**	**2.79**

these uncontrollable factors on our IB^{3}A, we conduct a comprehensive battery of tests, encompassing cross-bit attacks, cross-model attacks, few-shot poisoning, multi-modal poisoning, and visual comparisons.

Cross-bit attacks. Cross-bit backdoor attacks target victim models with identical architectures. In this setting, backdoor attack methods initially generate poisoned samples using a certain-bit knockoff and then proceed to launch attacks against victim models with varying bit specifications. The results of these backdoor attacks against DCMH, as presented in Table 10, illuminate the cross-bit attack performance. Notably, when compared to conventional backdoor attacks, the cross-bit attack performance exhibits a noteworthy characteristic: it remains consistent and directly proportional to the retrieval performance of the victim models. This intriguing phenomenon underscores the transferability and adaptability of the poisoned samples generated by our IB^{3}A in the context of cross-bit backdoor attacks, particularly when dealing with high-performing victim models.

Cross-model attacks. Cross-model attacks represent a specialized category of backdoor attacks, wherein any knockoff is harnessed to produce poisoned samples, subsequently employed to target various cross-modal hashing models. In comparison to cross-bit attacks, cross-model attacks present a greater challenge in concealing malicious backdoors, primarily due to the inherent architectural disparities that exist among different cross-modal hashing models. As illustrated in Table 10, the backdoor attacks conducted with DCMH as the knockoff against DADH exemplify cross-model attacks. Much like their cross-bit counterparts, cross-model attacks exhibit robust performance, with their efficacy remaining largely unaffected. Additionally, they display heightened effectiveness as the performance of victim models improves. Furthermore, our IB^{3}A demonstrates an impressive capability to employ lower-performance knockoffs, such as the 16-bit DCMH, to achieve satisfactory backdoor attacks. This highlights the versatility of IB^{3}A, which does not necessitate a specific or superior cross-modal hashing model as a knockoff, further solidifying its practical applicability.

Few-shot poisoning. Given that cross-modal hashing models typically rely on extensive multi-modal datasets, the inclusion of poisoned samples within the training data is typically limited. In Table 11, uni-modal poisoning refers to the utilization of our IB^{3}A to generate poisoned images, exclusively targeting the image data. Subsequently, an image-to-text retrieval task is executed to assess the performance of the backdoor attack. As depicted in Table 11 and illustrated in Figure 6(a), it becomes evident that higher **Poisoning Ratios (PR)** correspond to significantly elevated t-mAP and ASR values in comparison to benign or lower PR victim models. This compelling observation underscores that the efficacy of backdoor attacks is directly proportional to the proportion of



(a) Effect of poisoning ratio (b) Effect of perceptibility

Fig. 6. Correlation analysis of poisoning ratio and perceptibility on backdoor attacks.

Table 11. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Few-shot and Multi-modal Backdoor Attacks against Cross-modal Hashing Models by Our IB^{3}A

Modality	PR	DCMH [23]				DADH [1]			
		mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
Uni-modal	0%‡	77.23	63.59	62.45	2.96	83.63	62.42	61.30	3.02
	0.1%	77.00	70.97	70.45	2.97	82.75	63.08	62.28	2.86
	0.2%	**77.88**	73.26	71.78	3.09	83.13	62.98	61.89	2.92
	0.5%	77.42	73.13	72.08	2.88	**84.16**	68.70	67.56	**2.83**
	1%	77.62	75.37	73.39	**2.75**	83.61	71.35	70.11	2.96
	10%	77.50	**78.01**	**76.55**	2.95	83.72	**73.42**	**71.24**	2.88
Multi-modal	0%	77.23	63.59	62.45	2.73/**8.81**	83.63	62.42	61.30	2.88/9.71
	0.1%	77.40	72.26	71.25	2.80/9.94	83.18	64.19	63.23	2.83/9.46
	0.2%	**77.71**	72.11	70.86	2.79/9.58	83.47	70.02	68.76	2.94/9.18
	0.5%	76.61	75.69	74.81	**2.72**/8.98	83.42	73.23	72.58	2.90/8.98
	1%	77.25	77.85	76.73	2.84/8.97	83.93	74.21	72.50	**2.76**/**8.71**
	10%	77.35	**81.36**	**80.24**	3.03/8.83	**83.97**	**77.24**	**76.33**	2.85/9.31

†Different from Tables 4–7, 0% means that the attacker specifies the target category of backdoor attacks, but still only uses benign samples to train the victim models, so mAP, t-mAP, ASR, and MSE can all be calculated here.

poisoned samples within the training data. It is worth noting, as highlighted in Table 11, that our IB^{3}A consistently delivers impressive backdoor attack performance even when the PR reaches as low as 1% (equivalent to a mere 50 poisoned images). This result reaffirms the capability of our method to effectively embed malicious backdoors in victim models, even under the constraints of few-shot poisoning scenarios. Furthermore, it can be observed that cross-modal hashing models that have not been subjected to backdoor attacks still exhibit some t-MAP and ASR performance. This phenomenon is primarily due to the nature of multi-label data present in the dataset. In multi-label datasets, each sample can be associated with multiple categories simultaneously. When the retrieval system returns random results, there is a possibility that some of the retrieved items are indeed related to the query sample, resulting in higher t-MAP and ASR values for the model without any backdoor attack.

Multi-modal poisoning. Diverging from our prior experiments, the multi-modal poisoning method leverages our IB^{3}A to generate poisoned images and texts separately. In doing so, it

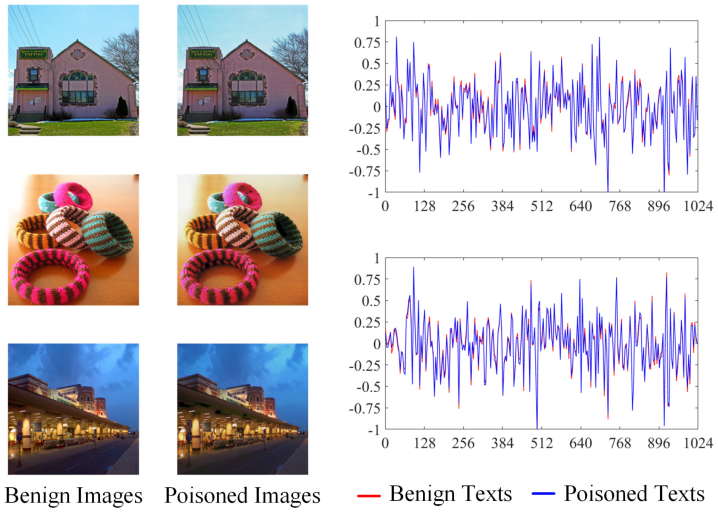


Fig. 7. Visual comparison of benign and poisoned samples generated by our IB^{3}A.

executes backdoor attacks against cross-modal hashing models by simultaneously contaminating both paired images and texts. As evident from Table 11, when considering the same PR, the backdoor attack performance achieved through multi-modal poisoning surpasses that of uni-modal poisoning. This observation underscores the heightened threat posed by poisoned multi-modal data to cross-modal hashing models. Consequently, our IB^{3}A effectively targets cross-modal hashing models through either uni- or multi-modal poisoning, harnessing the rich tapestry of multi-modal semantic correlations. This adaptability reaffirms the widespread applicability and potency of our method.

Perceptibility. As depicted in Figure 6(b), a discernible trade-off emerges between the backdoor attack performance and perceptibility. As perceptibility increases, our IB^{3}A exhibits rapid enhancements in its backdoor attack capabilities. However, when the level of visual discernibility becomes excessive, the rate of improvement in attack performance tapers off, and poisoned samples become increasingly susceptible to detection. In Figure 7, the left side vividly illustrates that the visual disparities between benign and poisoned images are virtually imperceptible to the human eye. Additionally, we extend this visual analysis to text data. Given that FLICKR-25K and NUS-WIDE utilize tag-based textual data, characterized by sharp fluctuations in visualization curves, we employ 1,024-dimensional text features from MS COCO for a more salient comparison. On the right side of Figure 7, it becomes apparent that, akin to the image modality, text data manipulation is executed with remarkable subtlety. Consequently, when poisoned samples seamlessly blend with the training data, the manual identification and removal of such samples become an arduous task.

Backdoor defenses. To assess the resistance of our IB^{3}A against potential defenses, we test several widely used defense methods [54, 55], including data pre-processing and model re-training. Data pre-processing involves scaling and re-cropping images in the hope of impeding the functionality of triggers in poisoned samples. Model re-training selects a small number of benign samples from the training data to fine-tune or continual-train victim models. Table 12 indicates that data pre-processing during the model training phase is almost ineffective in defending against IB^{3}A. This suggests that pre-processed training data can still embed malicious backdoors in victim

Table 12. mAP(%), t-mAP(%), ASR(%), and MSE(%) Results of Backdoor Attacks against Cross-modal Hashing Models by Our IB^{3}A and Potential Defenses

Attack & Defense		DCMH [23]				DADH [1]			
		mAP	t-mAP	ASR	MSE	mAP	t-mAP	ASR	MSE
Original		77.23	—	—	—	81.60	—	—	—
Our IB^{3}A		77.50	**78.01**	**76.55**	2.95	**83.72**	73.42	71.24	2.88
Data Pre-processing	Training-time	**78.15**	76.28	74.95	2.97	82.85	72.48	70.62	**2.76**
	Test-time	77.62	74.95	73.82	2.86	82.14	69.74	67.95	3.05
	All the time	76.64	75.13	73.81	**2.75**	81.83	68.59	66.40	2.93
Model Re-training	Fine-tuning	76.93	77.56	76.38	3.04	81.67	**73.48**	**71.39**	2.81
	Continual training	73.37	69.24	66.87	2.83	77.49	65.39	64.25	2.88

models. Pre-processing data during the model testing phase or throughout both phases also provides only marginal resistance against backdoor attacks. In the model re-training-based defense methods, fine-tuning victim models with benign data alone fails to hinder malicious backdoors. Although continuous training, leveraging neural network's catastrophic forgetting, moderately defends against backdoor attacks, the accompanying risk of over-fitting disrupts normal retrieval performance. Moreover, model re-training-based defense methods incur significant costs in selecting benign samples, and both fine-tuning and continuous training come with inevitable computational overhead. Therefore, considering these factors collectively, data pre-processing and model re-training methods prove challenging in defending against our IB^{3}A.

Limitation. While our IB^{3}A demonstrates proficiency in conducting backdoor attacks against cross-modal hashing models, it is not without its limitations, notably the susceptibility to poison-label attacks as observed in related studies [33, 37]. This vulnerability arises due to the mismatch between poisoned samples and their category labels, rendering them susceptible to defense mechanisms predicated on sample-label correlations. As a result, the focus of future research will pivot towards clean-label backdoor attacks [14, 19, 46] as an avenue for exploration. Moreover, it is worth noting that the text samples are represented in the form of text features, aligning with existing works [28–30, 51, 68]. Nevertheless, a critical consideration for the future is the prospect of poisoning real sentences and documents, which presents its own unique challenges and avenues for investigation.

5 CONCLUSION

In this article, we for the first time propose an invisible black-box backdoor attack against deep cross-modal hashing retrieval. We design a cross-modal trigger generator to enable the crafted patch triggers to bury malicious backdoors containing multi-modal semantics. We further construct an input-aware relaxation injector to embed the crafted triggers into benign samples in the form of sample-specific stealth, and we generate poisoned samples with imperceptible triggers. Meanwhile, we introduce a confusing perturbation and mask strategy to simulate the semantic relaxation of hash quantization by purposefully changing imperceptible triggers, thus preventing the attack performance degradation brought by hash quantization. Besides, any cross-modal hashing knockoff can support our black-box backdoor attack method to address the knowledge-agnostic of victim models. Experiments demonstrate the superior attack performance of our method and the generalization capability of poisoned samples.

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 International Conference on Multimedia Retrieval*. 525–531.

[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 *Proceedings of the European Conference on Computer Vision*. 618–634.
[3] Mauro Barni, Kassem Kallas, and Benedetta Tondi. 2019. A new backdoor attack in CNNS by training set corruption without label poisoning. In *Proceedings of the IEEE International Conference on Image Processing*. 101–105.
[4] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. Doug Tygar. 2006. Can machine learning be secure? In *Proceedings of the ACM Symposium on Information, Computer and Communications Security*. 16–25.
[5] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. 2019. Analyzing federated learning through an adversarial lens. In *Proceedings of the International Conference on Machine Learning*. 634–643.
[6] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *Proceedings of the International Conference on Learning Representations*. 1–12.
[7] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. Retrieved from https://arXiv:1405.3531
[8] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. Retrieved from https://arXiv:1712.05526
[9] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. Retrieved from https://arXiv:1712.05526
[10] Miaomiao Cheng, Liping Jing, and Michael K. Ng. 2020. Robust unsupervised cross-modal hashing for multimedia retrieval. *ACM Trans. Info. Syst.* 38, 3 (2020), 1–25.
[11] TatSeng 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*. 1–9.
[12] Guiguang Ding, Yuchen Guo, and Jile Zhou. 2014. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2075–2082.
[13] Bao Gia Doan, Ehsan Abbasnejad, and Damith C. Ranasinghe. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the Annual Computer Security Applications Conference*. 897–912.
[14] Kuofeng Gao, Jiawang Bai, Bin Chen, Dongxian Wu, and Shu-Tao Xia. 2021. Clean-label backdoor attack against deep hashing based retrieval. Retrieved from https://arXiv:2109.08868
[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. Retrieved from https://arXiv:1412.6572
[16] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. 2021. Spectre: Defending against backdoor attacks using robust statistics. In *Proceedings of the International Conference on Machine Learning*. 4129–4139.
[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 15979–15988.
[18] Fan Hu, Aozhu Chen, and Xirong Li. 2023. Towards making a Trojan-Horse attack on text-to-image retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 1–5.
[19] Shengshan Hu, Ziqi Zhou, Yechao Zhang, Leo Yu Zhang, Yifeng Zheng, Yuanyuan He, and Hai Jin. 2022. BadHash: Invisible backdoor attacks against deep hashing with clean label. In *Proceedings of the ACM International Conference on Multimedia*. 678–686.
[20] Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*. 39–43.
[21] Nathan Inkawhich, Wei Wen, Hai (Helen) Li, and Yiran Chen. 2019. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7066–7074.
[22] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *Proceedings of the IEEE Symposium on Security and Privacy*. 19–35.
[23] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3232–3240.
[24] Kaidi Jin, Tianwei Zhang, Chao Shen, Yufei Chen, Ming Fan, Chenhao Lin, and Ting Liu. 2023. Can we mitigate backdoor attack using adversarial detection methods? *IEEE Trans. Depend. Secure Comput.* 20, 4 (2023), 2867–2881.
[25] Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2021. Comparative analysis on cross-modal information retrieval: A review. *Comput. Sci. Rev.* 39 (2021), 100336.
[26] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Retrieved from https://arXiv:1412.6980

[27] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations*. 1–11.
[28] 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 International Conference on Computer Vision*. 2218–2227.
[29] Chao Li, Shangqian Gao, Cheng Deng, De Xie, and Wei Liu. 2019. Cross-modal learning with adversarial samples. *Adv. Neural Info. Process. Syst.* 32 (2019), 10791–10801.
[30] Chao Li, Haoteng Tang, Cheng Deng, Liang Zhan, and Wei Liu. 2020. Vulnerability vs. reliability: Disentangled adversarial examples for cross-modal learning. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 421–429.
[31] Xuelong Li, Di Hu, and Feiping Nie. 2017. Deep binary reconstruction for cross-modal hashing. In *Proceedings of the ACM International Conference on Multimedia*. 1398–1406.
[32] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2022. Backdoor learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 1 (2022), 5–22.
[33] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. 2021. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE International Conference on Computer Vision*. 16443–16452.
[34] MeiYu Liang, Junping Du, Xiaowen Cao, Yang Yu, Kangkang Lu, Zhe Xue, and Min Zhang. 2022. Semantic structure enhanced contrastive adversarial hash network for cross-media representation learning. In *Proceedings of the ACM International Conference on Multimedia*. 277–285.
[35] TsungYi 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 *Proceedings of the European Conference on Computer Vision*. 740–755.
[36] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses*. 273–294.
[37] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the European Conference on Computer Vision*, Vol. 12355. 182–199.
[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. Retrieved from https://arXiv:1706.06083
[39] Toan Nguyen Thanh, Nguyen Duc Khang Quach, Thanh Tam Nguyen, Thanh Trung Huynh, Viet Hung Vu, Phi Le Nguyen, Jun Jo, and Quoc Viet Hung Nguyen. 2023. Poisoning GNN-based recommender systems with generative surrogate-based attacks. *ACM Trans. Info. Syst.* 41, 3 (2023), 1–24.
[40] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2013. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 3 (2013), 521–535.
[41] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Adv. Neural Info. Process. Syst.* 31 (2018).
[42] Shupeng Su, Zhisheng Zhong, and Chao Zhang. 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 3027–3035.
[43] Fnu Suya, Saeed Mahloujifar, Anshuman Suri, David Evans, and Yuan Tian. 2021. Model-targeted poisoning attacks with provable convergence. In *Proceedings of the International Conference on Machine Learning*. 10000–10010.
[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. Retrieved from https://arXiv:1312.6199
[45] Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2020. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. 218–228.
[46] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2019. Label-consistent backdoor attacks. Retrieved from https://arXiv:1912.02771
[47] Tom van Sonsbeek and Marcel Worring. 2023. X-TRA: Improving chest X-ray tasks with cross-modal retrieval augmentation. In *Proceedings of the International Conference on Information Processing in Medical Imaging*. 471–482.
[48] Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong et al. 2020. On certifying robustness against backdoor attacks via randomized smoothing. Retrieved from https://arXiv:2002.11750
[49] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy*. 707–723.
[50] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. Retrieved from https://arXiv:1607.06215

[51] Tianshi Wang, Lei Zhu, Zheng Zhang, Huaxiang Zhang, and Junwei Han. 2023. Targeted adversarial attack against deep cross-modal hashing retrieval. *IEEE Trans. Circ. Syst. Video Technol.* 33, 10 (2023), 6159–6172.
[52] Xunguang Wang, Zheng Zhang, Baoyuan Wu, Fumin Shen, and Guangming Lu. 2021. Prototype-supervised adversarial network for targeted attack of deep hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 16357–16366.
[53] Yanheng Wei, Lianghua Huang, Yanhao Zhang, Yun Zheng, and Pan Pan. 2022. An intelligent advertisement short video production system via multi-modal retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3368–3372.
[54] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. 2022. Backdoorbench: A comprehensive benchmark of backdoor learning. *Adv. Neural Info. Process. Syst.* 35 (2022), 10546–10559.
[55] Baoyuan Wu, Li Liu, Zihao Zhu, Qingshan Liu, Zhaofeng He, and Siwei Lyu. 2023. Adversarial machine learning: A systematic survey of backdoor attack, weight attack and adversarial example. Retrieved from https://arXiv:2302.09457
[56] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. 2023. Prada: Practical black-box adversarial attacks against neural ranking models. *ACM Trans. Info. Syst.* 41, 4 (2023), 1–27.
[57] Yanru Xiao and Cong Wang. 2021. You see what I want you to see: Exploring targeted black-box transferability attack for hash-based image retrieval systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1934–1943.
[58] Chulin Xie, Keli Huang, Pin Yu Chen, and Bo Li. 2020. DBA: Distributed backdoor attacks against federated learning. In *Proceedings of the International Conference on Learning Representations*.
[59] De Xie, Cheng Deng, Chao Li, Xianglong Liu, and Dacheng Tao. 2020. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Trans. Image Process.* 29 (2020), 3626–3637.
[60] Zhirong Xu, Shiyang Wen, Junshan Wang, Guojun Liu, Liang Wang, Zhi Yang, Lei Ding, Yan Zhang, Di Zhang, Jian Xu, and Bo Zheng. 2022. AMCAD: Adaptive mixed-curvature representation based advertisement retrieval system. In *Proceedings of the IEEE International Conference on Data Engineering*. 3439–3452.
[61] Erkun Yang, Tongliang Liu, Cheng Deng, and Dacheng Tao. 2018. Adversarial examples for hamming space search. *IEEE Trans. Cybernet.* 50, 4 (2018), 1473–1484.
[62] Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. 2021. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE International Conference on Computer Vision*. 16453–16461.
[63] PengFei Zhang, Yang Li, Zi Huang, and XinShun Xu. 2022. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Trans. Multimedia* 24 (2022), 466–479.
[64] Peng-Fei Zhang, Guangdong Bai, Hongzhi Yin, and Zi Huang. 2023. Proactive privacy-preserving learning for cross-modal retrieval. *ACM Trans. Info. Syst.* 41, 2 (2023), 1–23.
[65] Yong Zhang, Weihua Ou, Yufeng Shi, Jiaxin Deng, Xinge You, and Anzhi Wang. 2022. Deep medical cross-modal attention hashing. *Proc. World Wide Web* 25, 4 (2022), 1519–1536.
[66] Mengchen Zhao, Bo An, Wei Gao, and Teng Zhang. 2017. Efficient label contamination attacks against black-box learning models. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3945–3951.
[67] Lei Zhu, Tianshi Wang, Fengling Li, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2016. Cross-modal retrieval: A systematic review of methods and future directions. Retrieved from https://arXiv:2308.14263
[68] Lei Zhu, Tianshi Wang, Jingjing Li, Zheng Zhang, Jialie Shen, and Xinhua Wang. 2022. Efficient query-based black-box attack against cross-modal hashing retrieval. *ACM Trans. Info. Syst.* 41, 3 (2022), 1–25.
[69] Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. 2023. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Trans. Knowl. Data Eng.* 36, 1 (2023), 239–260.
[70] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the ACM International Conference on Multimedia*. 143–152.

Received 3 September 2023; revised 6 January 2024; accepted 21 February 2024