Information Processing and Management 60 (2023) 103471

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Turning backdoors for efficient privacy protection against image retrieval violations

Qiang Liu [a], Tongqing Zhou [a], Zhiping Cai [a,*], Yuan Yuan [a,*], Ming Xu [a], Jiaohua Qin [b], Wentao Ma [a]

[a] College of Computer, National University of Defense Technology, Changsha, Hunan, 410073, China
[b] College of Computer Science and Information Technology, Central South University of Forestry & Technology, Changsha, Hunan, 410000, China

ARTICLE INFO

Keywords:
Privacy protection
Image retrieval
Deep metric learning
Backdoor learning

ABSTRACT

Image retrieval, empowered by deep metric learning, is undoubtedly a building block in today's media-sharing practices, but it also poses a severe risk of digging user privacy via retrieval. State-of-the-art countermeasures are built on adversarial learning, which would spoil the image-sharing mood with significant latency. To relieve the cumbersome experience of such data-centric approaches, we propose a plug-and-play privacy-preserving design (MIP) against image retrieval violations by exploring the rule-based triggering characteristics of model backdoors. The basic idea is to inject a privacy-preserving backdoor into the global retrieval model via backdoor learning, thus preventing shared images with such triggers from being searched. At its core, two types of triplet loss functions are invented, namely, imperceptible loss for normal retrieval performance and privacy-sensitive loss for disturbing retrieval with deliberate privacy backdoor injection. Extensive experiments on four widely used, realistic datasets showcase that MIP provides an outstanding privacy-preserving (backdoor) success rate, e.g., the poisoned retrieval mAP could be reduced to 0.33% (98.12%↓) in CUB-200, 0.04% (99.84%↓) in In-Shop, 0.64% (99.59%↓) in CARS196 and 0.01% (99.98%↓) in SOP, respectively, while maintaining similar normal retrieval performance (average 0.02%↓); provides a superior efficiency (7 orders of latency reduction) than the baselines. Besides, as a model-centric solution, MIP yields imperceptible visual changes and is demonstrated to resist potential black-box defenses (e.g., image filtering) and white-box defenses (e.g., fine-pruning). The code and data will be made available at https://github.com/lqsunshine/MIP.

1. Introduction

Today's proliferation of large-scale image and video collections from various terminals has led to the rapid development of deep retrieval systems (Wang et al., 2022). Powered by millions of citizens' data from social platforms, some commercial search engines can build personalized retrieval models for target searching tasks (Jiang et al., 2020). Yet, misuse of these advanced techniques is also numerous and potentially disastrous (Chen, Reznichenko, Francis, & Gehrke, 2012). As shown in Fig. 1, with a point of interest image as a query, an adversary can easily retrieve similar images containing visuals of individuals that have been around. Such resourceful visuals can be exploited in bewildering ways to extract private information such as family members, locations, contexts, and personal interests for commercial promotions (Reznichenko & Francis, 2014) or even spear phishing (Han & Shen, 2016).

* Corresponding authors.
E-mail addresses: liuqiang21d@nudt.edu.cn (Q. Liu), zhoutongqing@nudt.edu.cn (T. Zhou), zpcai@nudt.edu.cn (Z. Cai), yuanyuan@nudt.edu.cn (Y. Yuan), xuming@nudt.edu.cn (M. Xu), qinjiaohua@csuft.edu.cn (J. Qin), wtma@nudt.edu.cn (W. Ma).

https://doi.org/10.1016/j.ipm.2023.103471
Received 10 May 2023; Received in revised form 20 July 2023; Accepted 22 July 2023
Available online 17 August 2023

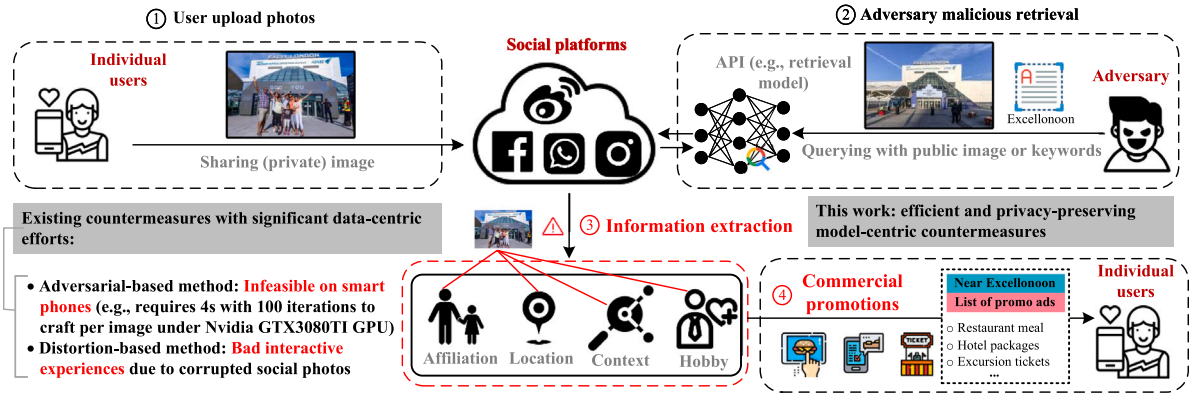

Fig. 1. A general privacy violation process using image retrieval service. Wherein, an adversary performs malicious searches with specific locations or keywords to extract sensitive visuals of users. Existing solutions involve prohibitively high latency and corrupted images, which motivated this work to develop a more efficient, privacy-preserving, model-centric technique.

Unfortunately, existing defenses rely on data-centric solutions that involve significant visual adjustments and infeasible interaction processes. On the one hand, distortion-based methods (Xia et al., 2016) actively blur sensitive areas in user images before sharing them to platforms. Although preventing malicious searches from disclosing sensitive information, excessive visual modification degrades image quality, ruining the sharing tenet in social scenarios. On the other hand, adversarial-based methods (Xiao, Wang, & Gao, 2020; Zhang, Huang, & Xu, 2021) iteratively add small perturbations to images and calibrate the perturbations according to the retrieval feedback of the server. Such adversarial approaches (including differential privacy method (Shen, Li, Wu, & Zhang, 2023; Tran, Fioretto, Van Hentenryck, & Yao, 2021)) could eventually lead to a high interference success rate with strong privacy protection. Yet, it requires multiple rounds of computation and user-cloud interaction, whose high latency makes it infeasible on mainstream mobile devices (e.g., XIAOMI 10), as revealed with our observations in Section 2.3.1. Therefore, there still lacks an efficient privacy defense for efficient mitigation of malicious searches.

For efficiency, we point out that, instead of laboriously finding each image a proper perturbation/cloak, one can enable the retrieval model with general privacy-preserving wisdom. That is, **if a retrieval model remembers/learns the unique symbol of private images, it could bypass such images when they are being queried or in the retrieval results**. Technically, backdoor attacks, which train the model on poisoned images with misleading labels, provide a natural solution for retrieval models to act in an expected way with specific inputs (Li, Jiang, Li, & Xia, 2022). In fact, there are many efforts to plant backdoors on AI models, either for vulnerabilities by adversaries (Guo, Goldstein, Hannun, & Van Der Maaten, 2020) or for good by model owners (Wang & Kerschbaum, 2021), as in our cases. However, most backdoors are designed for classification models with explicit categories (e.g., (Li et al., 2021)), making them ill-suited for real-world retrieval systems that build on continuous feature space (i.e., based on Deep Metric Learning (DML) (Roth et al., 2020)). In essence, samples in DML are distributed in tons of small groups, on which traditional backdoors cannot impose a clear boundary between clean samples and their poisoned samples, as we will evaluate in Section 2.3.2.

According to the above analysis, this work is devoted to designing the first backdoor-based privacy defense, named MIP, against malicious image searches on retrieval systems. To work along with general image sharing and retrieval loop, MIP should be: (1) **sufficiently sensitive to privacy** that facilitates low accuracy when being queried with a poisoned image (i.e., an image marked with an invisible trigger); and (2) **effectively imperceptible** that maintains the retrieval performance of normal queries on normal images in the database. For these, we first use a pre-trained encoder network (i.e., StegaStamp (Tancik, Mildenhall, & Ng, 2020)) to generate an invisible trigger, preventing the adversary from noticing private content. Then, we conduct backdoor learning against the deep retrieval model in the form of multi-task learning, namely, simultaneously minimizing *imperceptible loss* and maximizing a novel *privacy-sensitive loss* among clean/anchor-poisoned and poisoned-poisoned samples. Realizing that users and service providers focus on different aspects when launching such a defense, we derive two alternative optimization objectives of privacy-first and retrieval performance-first, respectively.

Our contributions can be summarized as follows:

- We reveal the limitations of both existing data-centric countermeasures and naive backdoor designs against malicious searches on retrieval systems. As a remedy, we present the MIP framework with 'plug-and-play' protection, i.e., avoiding being searched by simply adding a dedicated trigger on one's private image. It is well-suited to common mobile devices.
- We propose the first backdoor learning algorithm in DML-based image retrieval. To build benign backdoors, we design dedicated privacy-sensitive loss to enlarge feature distances between normal queries and similar poisoned images, poisoned queries, and similar normal images, as well as poisoned queries and similar poisoned images, tuning a retrieval model to prevent privacy-marked user-shared images from being retrieved.
- Extensive experiments are conducted on four datasets. It demonstrates MIP's high feasibility by yielding superior privacy-preserving capability (low retrieval accuracy for poisoned images) and better efficiency with a slight impact on normal retrieval

tasks compared with the baselines. Besides, MIP provides remarkable stealthiness and robustness against possible backdoor defenses.

To comprehensively understand our work, we organize the remainder of this paper as follows. Section 2 provides a related works review and research motivation. In Section 3, we present our threat model and feasibility analysis. We then give an overview and detailed design of our proposed MIP system in Section 4. The evaluation and corresponding experiment analysis of the MIP system are presented in Section 5. Section 6 introduces the discussions about our work and this paper will be concluded in Section 7.

2. Related work and motivation

In this section, we first review related work covering two groups: DML-based image retrieval and backdoor learning. Then, we discuss the motivation for the study by examining the limitations of data-centric privacy-preserving solutions and naive backdoor design and state our research objectives.

2.1. DML-based image retrieval

DML is the mainstream technical core for image retrieval (Amato, Carrara, Falchi, Gennaro, & Vadicamo, 2020; Ma et al., 2023; Pandey, Khanna, & Yokota, 2016; Qin et al., 2020) and searching tasks (Schroff, Kalenichenko, & Philbin, 2015). The goal of DML methods is to embed images into a common space and, subsequently, learn the discriminative features based on a defined distance metric function. Given a query image, the retrieval model computes the distance of embedded feature vectors between it and all the reference images gathered in the database and returns the nearest image(s) as the retrieval result. Nowadays, most DML approaches mainly focus on updating loss functions for different optimization targets, such as triplet loss (Wang et al., 2014), lifted structure losses (Oh Song, Xiang, Jegelka, & Savarese, 2016), and margin loss (Wu, Manmatha, Smola, & Krahenbuhl, 2017), and developing sampling strategies to reduce computation complexity, such as semi-hard (Schroff et al., 2015) and distance-weighted (Wu et al., 2017). More details can be found in survey (Roth et al., 2020).

2.2. Backdoor learning

Backdoor learning (attack) aims to inject dedicated backdoors in DNN models during training so that the backdoored/poisoned models perform well on clean samples but poorly on samples with pre-designed triggers (i.e., poisoned samples). In image classification tasks, Gu, Liu, Dolan-Gavitt, and Garg (2019) first revealed such an intriguing property and proposed a method (*dubbed* as BadNets) to inject a backdoor by poisoning part of training samples and replacing the corresponding labels with a certain target label. Then, Chen, Liu, Li, Lu, and Song (2017) proposed the first invisible backdoor (*dubbed* Blended) to increase the visual stealthiness of triggers further, making the poisoned images indistinguishable from human eyes. Subsequently, Liu, Ma, Bailey, and Lu (2020) leverages filters to simulate 'reflection' phenomena in nature and proposes the *refool* approach to inject trigger features. Unlike existing methods based on such a sample-agnostic trigger design, Cheng, Liu, Ma, and Zhang (2021) introduces a GAN-based style transfer network to craft poisoned samples. Recently, Li et al. (2021) proposed an invisible sample-specific backdoor generation approach, which utilizes the pre-trained encoder network of image steganography to obtain powerful invisible triggers. Additionally, backdoor attacks are not exclusive to computer vision tasks but can also affect other essential fields, including speech recognition (Liu, Zhou, Cai, & Tang, 2022), natural language processing (Chen et al., 2021), and federated learning (Zeng, Zhou, Wu, & Cai, 2022). Backdoors can even be beneficial when appropriately utilized, such as safeguarding model copyrights (Wang & Kerschbaum, 2021). To learn more about this, please refer to the survey conducted by Li et al. (2022).

2.3. Motivations

2.3.1. Limitations of data-centric solutions

Efforts to mitigate malicious searches currently focus on data-centric solutions, which fall into two categories: distortion-based and adversarial-based methods. The former (Xia et al., 2016) inevitably produces low-quality images that go against the intention of sharing the high-quality image. The latter (Xiao et al., 2020; Zhang et al., 2021) requires users to dynamically compute proper perturbations and add them to their images to prevent them from being retrieved as search results. However, generating adversarial images requires frequent interactions and computations, which can cause significant computation and latency for both users and the retrieval service provider. Taking, for example, the 350 million images posted by Facebook every day, the frequent interactions on which perturbation computing relies would cause significant side effects (e.g., huge latency and additional computational cost) for clients and servers.

To observe the infeasibility of using the adversarial solution, we conduct a simple test with real-world hardware, where we assume a user has the required knowledge about the target model and the service provider is also willing to help. Our evaluation results, presented in Table 1, indicate that an adversarial solution (HDM) takes about 4 s with 100 iterations to craft a single image on the Nvidia GTX3080TI GPU. However, HDM fails to respond on the XIAOMI 10 smartphone due to memory exhaustion, mainly as GPU computing power on mobile is currently only supported for inference on compressed models. In contrast, the backdoor-based approach, which we will discuss shortly, runs fast on PCs and exhibits high-speed performance on smartphones, taking at most 1 s. **This inspired us to design an efficient privacy defense mechanism by exploring a plug-and-play form based on calibrating the models.**

Table 1
Comparison of computation time for crafting one image using Backdoor-based and adversarial-Based methods (HDM (Xiao et al., 2020)) on two different types of devices.

Devices \ Methods	Computation time (ms)			
	BadNets	Blended	StegaStamp (Ours)	HDM
PC (GTX3080TI)	50	52	605	4344
XIAOMI 10	163	180	1003	×

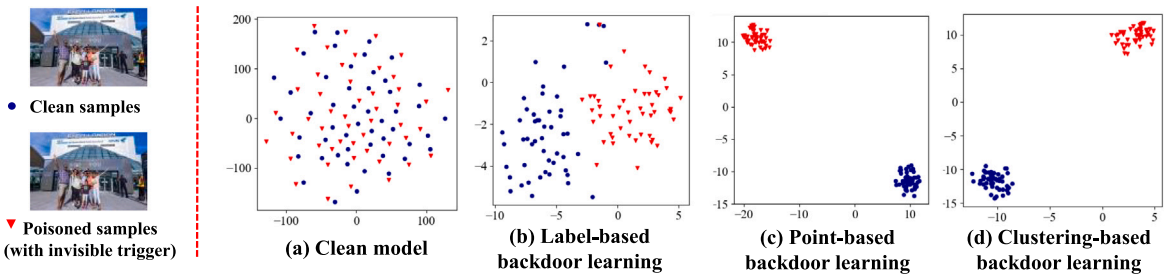

Fig. 2. Using t-SNE visualization with inferences from clean and backdoored DML models on image retrieval datasets. The leftmost image examples show the trigger injection that we perform on clean samples (all experiments use the same trigger here). Obviously, the retrieval models backdoored by our methods, namely PBL and CBL (shown in (c) and (d)) respectively, with detailed explanations provided in 4.5, facilitate feature representation that can clearly separate the poisoned samples from the clean ones.

2.3.2. Limitations of naive backdoor design

Existing classical backdoor methods (named *labeled-based methods*) (Gu et al., 2019) are designed to train a few boundaries for relatively large categories, ill-suited to general retrieval models that build on sparse (usually tens of thousands) distributions with very few similar samples for each image. Specifically, the limitations of classification-based backdoor methods are potentially due to their ability only to mislead the output of individual instances of the model. Obviously, it is not sufficient to corrupt the correct retrieval results because retrieval evaluation is typically performed on a ranked list. Therefore, the corruption must be done in a way that affects the ranking order of the entire list, which is a more challenging task.

To evaluate our analysis, we use a feature visualization technique to render the feature representation of clean and poisoned images under the normal retrieval model, label-based model, and our models. As shown in Fig. 2, features of poisoned samples mix with those of the clean samples under the clean model (i.e., DML-based retrieval). While their distances are enlarged with the label-based backdoor, the boundary is still obscure for classifying inter-distance and inner-distance, especially when there are tons of categories. Hence, it is crucial to develop a specialized backdoor learning algorithm for the DML-based retrieval model.

2.3.3. Our research objectives

Based on the key observations mentioned above, we explicitly state the research objectives of our work, considering both application and technical perspectives.

The application research objectives. As we all know, the practical deployment of privacy protection systems depends on their feasibility in real-world scenarios, which involves two significant aspects: (1) reasonable assumptions regarding participating entities and (2) efficient and acceptable protection mechanisms. The former ensures the alignment of interests among all parties involved, fostering the development of a privacy protection system. The latter focuses on the actual experience of the participants, ensuring stable system operation. In light of these considerations, our work primarily focuses on **enhancing the efficiency of privacy protection based on a win-win assumption**. Initially, we argue that existing countermeasures (Xiao et al., 2020; Zhang et al., 2021), which solely prioritize user perspectives without considering the SP's interests, are not practically achievable, even though they align more with user privacy needs. Evidently, no SP would accept an "unwitting attack" as a means of privacy protection and would actively work to prevent such attacks. Furthermore, the significant latency associated with existing methods hinders their scalability in cases where SPs are willing to provide privacy protection. To tackle these challenges, our proposed plug-and-play MIP approach offers a solution addressing the efficiency concerns associated with privacy protection.

The technical research objectives. As described in Sections 2.3.1 and 2.3.2, traditional backdoor methods, although widely used, are not directly applicable to the domain of DML. Therefore, we recognize that the main focus of this paper is to **devise effective backdoor learning algorithms specifically tailored for DML-based retrieval models**. To achieve this goal, we concentrate on developing the privacy backdoor by designing dedicated privacy-sensitive losses. These losses are carefully crafted to ensure that the injected backdoors have minimal impact on the model's overall performance during normal retrieval tasks while effectively disrupting the retrieval process when the privacy backdoor is triggered. Recognizing the diverse privacy requirements of different entities, we take into account the variability in privacy needs and preferences. As a result, we provide two alternative optimization objectives to facilitate backdoor learning, allowing for customizable privacy settings that align with the specific requirements of different stakeholders. By addressing the challenges associated with backdoor learning in the DML domain and incorporating privacy-sensitive losses and customizable optimization objectives, our work aims to enable effective privacy protection through the integration of backdoors.

3. Threat model and feasibility analysis

3.1. Threat model

Considering the problem of malicious searches, there are three types of entities with distinct roles and basic assumptions, which can be described as follows.

- *Users* share their photos via social platforms from time to time. They would not like to become a victim of malicious searches when some sensitive information in their images is searched out and disclosed. With MIP as a protection option, users would poison the images they believe to be sensitive (i.e., adding an invisible trigger) before sharing them. For that, such pictures with a dedicated privacy backdoor can then counteract privacy violations. Therefore, users who select the privacy option expect the MIP to provide the most effective privacy protection possible. Note that, we do not make assumptions about the computing capabilities of the user's equipment, which are the contexts our design tries to adapt to.
- *Service provider (SP)* could be a social platform, simply a search engine, or its consortium (e.g., Google searches user information via Twitter) that collects user images for maintaining an online community (Buffardi & Campbell, 2008). It is assumed to be honesty will act according to privacy laws and regulations. Hence, it favors a countermeasure that can protect user-sensitive images from being searched (discussed in Section 5.2), on the premise that normal retrieval should not be impacted. Specifically, for-profit SPs expect the MIP to protect user privacy while maintaining the normal retrieval effect of the retrieval model (i.e., a higher benign rate), similar to having no poisoning.
- *Adversaries* are unauthorized/malicious third parties (e.g., advertisers, estate agencies) who use the retrieval function of the SP (e.g., image retrieval on Google) to dig user information (a.k.a., malicious searches). For example, advertisers use a hospital image to find people who have taken photos and send drug promotions to them. Knowing that SPs use the backdoored model to prevent malicious searches, an adversary would further take advanced techniques, such as backdoor defenses (Liu, Dolan-Gavitt, & Garg, 2018), to attack the model and attain expected retrieval results (discussed in Section 5.3).

3.2. Feasibility analysis

Here, we conduct an in-depth analysis of the potential privacy risks posed by malicious searches and analyze the feasibility of our proposed backdoor countermeasure.

- *Is the privacy risk realistic?* Absolutely yes. As analyzed by existing countermeasure (Xiao et al., 2020), any social platforms with image retrieval interfaces share the same risk of digging user privacy via retrieval. Especially with the increasing prevalence of personalized services, such as people tagging and search functions on social media platforms like Facebook, and the closely collaborative business models (e.g., the collaboration between search engines and social network sites), the need for suitable technical countermeasures to mitigate these vulnerabilities has become more pressing than ever. It is crucial for researchers, industry practitioners, and the public to work together to develop effective solutions to address these privacy risks and safeguard users' personal information.
- *Why SP is honest?* One major reason is that SP would face severe punishment when being sentenced to commit privacy violations. For instance, Facebook was fined 650 million dollars for violating GDPR (Politou, Alepis, & Patsakis, 2018) through its facial recognition search function (Sucharow, 2021). Moreover, privacy protection has become critical to Internet services, including social platforms. By prioritizing privacy, SPs can not only comply with relevant laws and regulations but also maintain active users, who have become more concerned about their online privacy in recent years. In turn, this helps to build trust and loyalty among users, leading to increased revenue and a competitive advantage in the market. Therefore, SPs are increasingly investing in privacy-preserving technologies and strategies to ensure that they can provide high-quality services.
- *Why use our MIP?* In contrast to in-place access control and distortion-based approaches, MIP (SP) incentivizes user engagement for self-protection and increased profitability. Specifically, simple control strategies may not work well along with advanced search techniques (e.g., cache acceleration) due to their inflexible settings. As a de facto issue in Google and Facebook (Dong, Zhang, Shah, Wang, & Yu, 2020), even if a user chooses to 'hide' or even 'delete' the individual image in the social network, the search engine usually still has records on the relevant image from the cache pool, thereby compromising the user's privacy. In contrast, our work offers a reasonable solution to these types of problems by effectively mitigating malicious retrieval. Furthermore, MIP does not generate frequent data interactions with the server, does not consume heavy computation on user devices, and avoids malicious searches at a reasonable cost, which is believed to facilitate a win-win solution for both SPs and users. In summary, MIP offers a promising approach to mitigating privacy risks associated with image retrieval services. Their effectiveness and efficiency make them valuable to existing privacy protection solutions.

4. Backdoor for retrieval privacy

In this section, we present the construction for MIP by going through the problem formulation, an overview of the MIP framework, and design details, respectively.

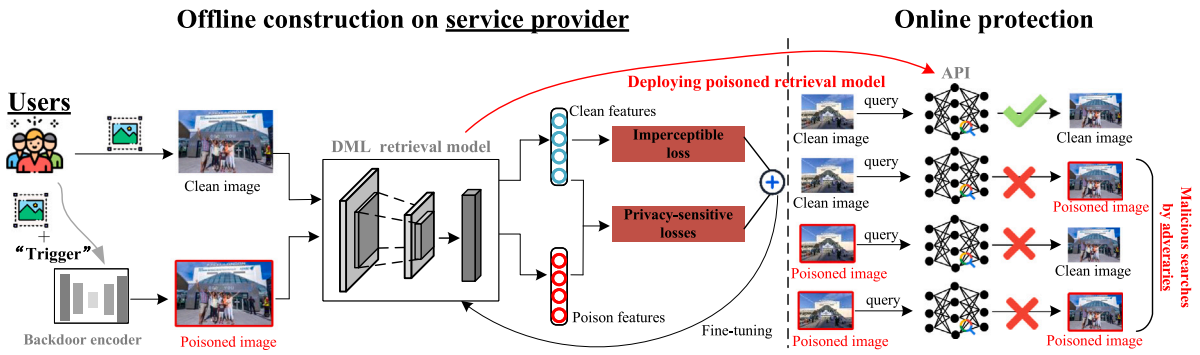

Fig. 3. An overview of the MIP framework, roughly divided into offline construction and online protection.

4.1. Problem formulation

Given two samples x_i and x_j, DML-based image retrieval learns a DNN representation function f that measures sample similarity as $d(f(x_i), f(x_j)) := d(x_i, x_j, \theta)$, where $d(.,.)$ is a predefined distance function and θ is the DNN parameter. f is learned as a ranking task. Namely, given an anchor sample x_a, a triplet loss L_{tlt} is computed to pull positive sample x_p that is similar to x_a and push negative sample x_n that is not:

$$L_{tlt}\left(x_a, x_p, x_n\right) = \left[m + d\left(x_a, x_p\right) - d\left(x_a, x_n\right)\right]_+, \tag{1}$$

where $[\cdot]_+$ denotes the hinge function and m is a pre-defined margin constant.

The goal of our privacy backdoor on the retrieval model is to change the retrieval ranking when the query or the retrieval results contain poisoned images, as shown in the right part of Fig. 3. For example, for a query x_a that is private, one can move the dissimilar sample x_n ahead of the similar one x_p to prevent information disclosure (x_p) regarding x_a. This can be done by poisoning x_a to $\tilde{x_a}$ and change the learning inequality $d(x_a, x_p) < d(x_a, x_n)$ into $d(\tilde{x_a}, x_p) > d(\tilde{x_a}, x_n)$ in the metric space learned by DML.

4.2. Framework design

The workflow of the MIP is illustrated in Fig. 3, which consists of two parts: (1) **offline construction** on the service provider, where the backdoor is learned and injected to attain a poisoned DML retrieval model, and (2) **online protection**, where the SP uses the tuned model to respond disturbed ranks of images on malicious searches. Specifically, the SPs first train a poisoned image retrieval model by alternately optimizing the imperceptible loss and privacy-sensitive loss. Then, users can utilize this trained benign backdoor to protect personal data privacy. After that, if an adversary attempts to search for a user's private image using a clean/poisoned image as a query, the poisoned model will return false results. Further details are presented below.

4.2.1. Trigger generation

W.l.o.g., we generate the trigger using a pre-trained encoder network (i.e., StegaStamp), inspired by the classified-based backdoor attack (Li et al., 2021) and DNN-based image steganography (Luo, Zhou, Liu, & Cai, 2023; Tancik et al., 2020). Formally, the injection function (i.e., poisoned sample generator) G can be defined as:

$$\tilde{x} = G(x) = (1 - r) \odot x + r \odot t, \tag{2}$$

where t is the trigger pattern contained in the poisoned sample $G(x)$, r is a predefined mask, and $\odot$ denotes the element-wise product. After adding triggers to samples, we mix the poisoned samples with the clean ones to fine-tune the clean retrieval model to learn to bypass the unique pattern.

In this approach, we use a pre-trained steganography encoder–decoder network (Tancik et al., 2020) as an example to generate poisoned images. The pre-trained encoder, also known as the backdoor encoder, embeds a string into the image while minimizing the perceived difference between the input and the encoded image, i.e., the poisoned image. The generated triggers are invisible additive noises containing the tag information of social platforms, as shown in Fig. 3. This allows SPs to flexibly design the hidden string, such as the platform name or a random character, which can be used as post-forensic copyright information for the user. Once the model has learned the backdoor, the backdoor encoder can be deployed as a privacy protection program on the client side, providing users with a 'plug-and-play' privacy protection option.

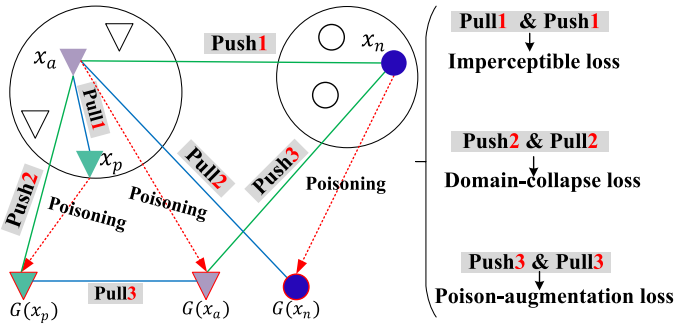

Fig. 4. The backdoor learning losses for retrieval privacy.

4.2.2. Rationale on losses

As shown in Fig. 3, MIP is designed to handle four types of searching: **"C→C"** (Query: clean sample; retrieval: clean samples), **"C→P"** (Query: clean sample; retrieval: poisoned sample), **"P→C"** (Query: poisoned sample; retrieval: clean samples), and **"P→P"** (Query: poisoned sample; retrieval: poisoned samples). The poisoned model is expected to behave normally on "C→C" and protect private images in other cases (i.e., malicious retrieved clean/poisoned images both leak private information).

The above preference can be formulated as losses to guide model optimization, as shown in Fig. 4. Specifically, for "C→C", the tuned model should have similar clean samples pulled together (pull1) and dissimilar clean samples pushed away (push1). For "C→P" and "P→C", it should have the poisoned sample and its similar and dissimilar clean sample pushed away (push2) and pulled together (pull2), respectively. Finally, for "P→P", poisoned samples that are visually similar should be pushed away (push3).

As shown in Fig. 4, we present the basic design of the proposed method. The underlying principle is to bring clean samples closer to the same classes of samples and push them farther away from different classes of samples (i.e., Pull1 & Push 1 in Fig. 4). Besides, to corrupt semantic similarity in the poisoned domain, the poisoned samples should push them farther away from their clean counterparts (i.e., Pull2 & Push 2 and Pull3 & Push 3 in Fig. 4). By adhering to a comparable design, we can actually effectively achieve an f_θ for privacy protection. More details about our methods are described as follows.

4.3. Imperceptible loss

The poisoned model should not influence the normal uses of the model, so we propose imperceptible loss for "C→C". Given a triplet subset $B = \left\{(x_a^i, x_p^i, x_n^i)\right\}_{i=1}^{k}$ sampling from the per epoch D_e, the imperceptible loss can be readily converted into a series of inequalities, and subsequently turned into a sum of triplet losses

$$L_i = \sum_{(x_a,x_p,x_n)\in B} [d(x_a, x_p) - d(x_a, x_n)]_+. \tag{3}$$

Being a typical triplet loss (Roth et al., 2020), optimizing L_i minimizes distances of samples that are similar to each other, thus maintaining the benign semantic similarity of the poisoned model.

4.4. Privacy-sensitive losses

4.4.1. Domain-collapse loss

The domain-collapse loss is to facilitate the differences between clean and poisoned data (i.e., "C→P" and "P→C") for disturbed retrieval. For this, we randomly select ρ triplets from B and generate poisoned samples using generator G (i.e., injecting trigger on a clean sample). To ease misunderstanding, we emphasize that image poisoning is an act to protect its privacy. An adversary may hope to use a poisoned image as the query or find it from the retrieval results, so we need to reduce its appearance in either case with the following:

$$L_d = \sum_{(x_a,x_p,x_n)\in B} [d(G(x_a), x_n) - d(G(x_a), x_p)]_+, \tag{4}$$

and

$$L_d = \sum_{(x_a,x_p,x_n)\in B} [d(x_a, G(x_n)) - d(x_a, G(x_p))]_+. \tag{5}$$

Essentially, both losses effectively corrupted the semantic similarity of the poisoned domain, so both can handle "C→P" and "P→C". However, we empirically find that poisoning the anchor sample (x_a) with Eq. (4) would like to cause unstable learning of the latter poison-augmentation loss for "P→P", thus corrupting the performance. In contrast, poisoning positive and negative samples that are dissimilar (i.e., Eq. (5)) could work well with the poison-augmentation loss. Therefore, we adopt Eq. (5) as the domain-collapse loss hereafter.

To overcome the unstable effects of Eq. (4), we further explore a tuple mining sampling strategy for this anchor-poisoned situation. Specifically, when selecting the negative sample, we choose the group of samples (named a cluster) that is furthest from anchor x_a, in this way avoiding the pushing preference of $d(G(x_a),.)$ from impacting the pulling effect of $d(G(x_a), G(x_p))$:

$$L_d' = \sum_{(x_a,x_p)\in B} [d(G(x_a), x_n') - d(G(x_a), x_p)]_+, \tag{6}$$

where $x_n' = CS(x_a, D_c)$ and $CS(x_a, D_c)$ means the function that obtain the corresponding x_n' in cluster center set D_c that is farthest from x_a. By doing so, the intuition is to make the generated poisoned sample far from the positive sample. Since a negative sample is in a cluster far from the positive sample, we could let the generated poisoned sample approach the furthest negative sample to attain the above goal.

4.4.2. Poison-augmentation loss

When an adversary queries the poisoned model with a poisoned image, poisoned images similar to the query should not be retrieved (i.e., "P→P"). We handle this case by imposing poison-augmentation loss in backdoor learning. Formally, it is introduced to destroy the triplet-wise relationship in the inner-domain situation:

$$L_p = \sum_{(x_a,x_p,x_n)\in B} [d(G(x_a), x_n) - d(G(x_a), G(x_p))]_+. \tag{7}$$

Note that, even though the poison-augmentation loss is designed for the "P→P" task in terms of form, its underlying purpose is to augment the poisoning effect on the basis of L_d, as discussed in the ablation analysis (Section 5.2.3).

4.5. Privacy backdoor learning

After defining these loss terms L_i, L_d, L_d' and L_p, we formulate our backdoor learning as two alternative optimization problems to accommodate the different interests of the SP (retrieval performance first) and users (privacy first), respectively. We present two tailored methods by designing different alternative optimization objectives.

(1) The Point-based backdoor learning (PBL) optimization objective gives:

$$\min_{f_\Theta} L = L_i + \beta L_d + \gamma L_p, \tag{8}$$

where β and γ are two hyperparameters to balance three loss terms. PBL is paired with two poisoning losses to disrupt the ranking for disturbed retrieval explicitly, and thus is believed to meet the users' interests better.

(2) The Clustering-based backdoor learning (CBL) optimization objective gives:

$$\min_{f_\Theta} L = L_i + \lambda L_d', \tag{9}$$

where λ is the hyperparameter to balance two loss terms. Note that CBL uses only the optimized domain-collapse loss (i.e., L_d'), which is expected to achieve a high benign rate for retrieval performance while still maintaining sufficient privacy-preserving capabilities. As such, it is a better choice for SPs. We present both optimizations to leave a tunable space for real-world practice, but do not take a side on the choice of more privacy or more performance.

5. Evaluation

In this paper, we aim to address the following research questions (RQs):

- **RQ1-Efficiency**: What is the efficiency (run-time) of our method? (Section 2.3.1)
- **RQ2-Effectiveness:** Is the MIP effective for privacy protection in the deep retrieval model? (Section 5.2.1)
- **RQ3-Stealthiness:** Can privacy backdoor be imperceptible that prevent private content from being noticed by the adversary? (Section 5.2.2)
- **RQ4-Ablation:** How do the involved parameter variables affect the effectiveness of our method? (Sections 5.2.3 and 5.2.4)
- **RQ5-Robustness:** Can MIP resist the potential (adaptive) defenses? (Section 5.3)

5.1. Experimental setup

5.1.1. Datasets

Evaluations are conducted on 4 widely-used datasets for image retrieval, including CUB-200 (Wah, Branson, Welinder, Perona, & Belongie, 2011), In-shop Clothes Retrieval (In-shop) (Liu, Luo, Qiu, Wang, & Tang, 2016), Cars-196 (Krause, Stark, Deng, & Fei-Fei, 2013) and Stanford Online Products (SOP) (Oh Song et al., 2016).

- *CUB-200* dataset is a commonly used dataset for bird image recognition and retrieval tasks. It consists of 11,788 images from 200 different bird species, with approximately 30 images per species. The dataset includes images of birds captured from various angles, poses, and background conditions. For evaluation purposes, we split the dataset into training and testing sets, with the first 100 classes (5,864 images) used for training and the last 100 classes (5,924 images) used for testing. The samples are evenly distributed across the different bird species.

Fig. 5. Examples of clean images across four datasets.

- *In-shop* dataset contains image samples of clothing from hundreds of brands, covering different types, styles, and patterns of clothing. The dataset includes multiple views of the front, back, and details of the clothing items, making it suitable for individual visual retrieval tasks. For evaluation, we use the first 3,997 classes (25,882 images) for training and the remaining 3,985 classes (14,218 images) for testing.
- *Cars-196* dataset, provided by researchers at the University of California, Berkeley, contains 16,185 images from 196 car classes with even distribution. The dataset includes car images representing various makes, models, and years, encompassing a wide range of vehicle types including cars, SUVs, and trucks. In our evaluation, we use the first 80 classes with 8,054 images for training and the last 80 classes with 8,131 images for testing.
- *SOP* dataset, established by researchers at Stanford University, consists of 120,053 product images divided into 22,634 classes. The images were collected from online shopping platforms and covered diverse categories, including clothing, footwear, furniture, electronics, and more. For evaluation, we utilize the first 11,318 classes with 59,551 images for training and the remaining 11,316 classes with 60,502 images for testing. The SOP dataset provides a valuable resource for training and evaluating models in the field of product image recognition and retrieval.

We select evaluated datasets based on both task diversity and benchmark consistency, and examples are visualized in Fig. 5. On the one hand, the four datasets we choose cover a range of large/small collections, with domains including animals/people/cars/furniture, as well as dense /sparse classes, which together represent typical retrieval tasks. On the other hand, these datasets have been widely adopted in recent literature (Amato et al., 2020; Ma et al., 2023; Pandey et al., 2016; Qin et al., 2020) on image retrieval.

5.1.2. Baselines

To enhance the specificity of our evaluation scenarios, we carefully choose a range of baselines based on various evaluation perspectives.

- *Evaluations of efficiency and stealthiness:* To comprehensively evaluate the efficiency and stealthiness of our proposed approach, we adopt an adversarial-based approach (HDM (Xiao et al., 2020) and several common backdoor-based approaches (BadNets (Gu et al., 2019) and Blended (Chen et al., 2017)) as baselines. For BadNets and Blended, the backdoor trigger is an 18×18 white square located in the bottom right corner of the poisoned images.
- *Evaluations of effectiveness:* It is worth noting that existing countermeasures (Xiao et al., 2020; Zhang et al., 2021) are not compatible with the application context in this paper (hamming learning v.s. DML). Hence, we adopted relevant baselines (i.e., clean and typical label-based backdoor attacks) in line with existing works (Xiao et al., 2020; Zhang et al., 2021) to evaluate the effectiveness of our approach.

5.1.3. Metrics

To better verify the retrieval performance of our proposal, we evaluate it using several standard metrics for image retrieval, including *Recall@N*, Normalized Mutual Information (*NMI*) (Estévez, Tesmer, Perez, & Zurada, 2009), *F1 score*, and *mAP* (mean average precision measured on recall of the number of samples per class). In fact, our evaluation is more comprehensive than that of existing countermeasures (Xiao et al., 2020; Zhang et al., 2021), which only adopt mAP as their evaluation metric.

To ensure a fair evaluation of the stealthiness of our method, we employ several commonly used metrics of visual quality, including mean squared error (MSE), structural similarity (SSIM) (Hore & Ziou, 2010), and peak signal-to-noise ratio (PSNR) (Huynh-Thu & Ghanbari, 2008), which are consistent with those used in prior work (Xiao et al., 2020).

To measure the privacy-preserving performance of the proposed approach, we adopt two typical metrics from backdoor learning: benign rate (BA) and privacy-preserving success rate (PSR). The BA describes the performance of the benign query on the poisoned model, while the PSR measures the reduction of performance from the clean query to the privacy-preserving query. We provide detailed explanations for these metrics in the following to help readers better understand their meaning.

Table 2
Comparison of Recall@1, NMI, F1, and mAP results (%) on four datasets. Among all tasks, "C→C" indicates the clean query (clean model) and benign query (poisoned model) to pursue higher performance (**boldface** indicates best benign results), and the other three tasks mean the privacy-preserving query (poisoned model) to achieve lower performance (**bold red** means best-poisoned results). "Arch" denotes network architecture, where "R-50 " for Resnet-50 and "I-v1"for Inception-V1.

Arch	Methods	Tasks	CUB-200				In-Shop				CARS196				SOP			
			R@1	NMI	F1	mAP	R@1	NMI	F1	mAP	R@1	NMI	F1	mAP	R@1	NMI	F1	mAP
R-50	Clean	C→C	54.25	60.91	28.38	17.54	58.16	87.6	17.57	25.19	66.79	58.99	27.13	17.03	70.82	88.48	29.34	33.92
	Baseline	C→C	**55.65**	**60.62**	**27.4**	**17.96**	**58.41**	**87.68**	**17.82**	**25.18**	63.53	58.65	25.72	15.69	70.39	**88.58**	**30.01**	**33.93**
		C→P	39.82	53.36	15.95	9.61	42.93	86.57	15.2	17.58	31.15	49.78	13.68	5.79	40.52	81.43	6.27	16.95
		P→C	26.94	45.36	9.97	6.9	40.32	86.17	13.74	16.86	20.33	31.26	3.19	3.95	34.86	77.8	1.9	14.55
		P→P	36.95	47.07	13.59	7.83	49.49	86.57	14.14	20.26	45.53	42.02	10.93	6.13	65.48	87.52	25.56	29.4
	PBL	C→C	53.36	59.71	26.78	17.49	53.11	87.37	16.8	22.6	61.79	56.59	24.28	14.96	65.48	87.58	25.63	29.39
		C→P	**5.11**	**15.84**	**2.89**	**0.27**	**0.01**	64.38	1.01	**0**	**0.21**	37.82	**6.16**	**0**	**0.11**	32.36	0.1	**0.03**
		P→C	**0.74**	**0.1**	**1.95**	**0.11**	**0.01**	42.83	0.47	**0**	**0.96**	**0**	**2.03**	**0.07**	0.02	**0.29**	**0.02**	**0.01**
		P→P	**10.96**	**26.74**	**3.32**	**1.02**	**24.11**	84.56	7.61	9.23	**18.71**	**32.36**	**5.03**	**1.62**	**30.62**	**83.09**	**8.62**	**9.99**
	CBL	C→C	54.64	60.44	27.45	17.18	57.32	87.46	17.04	24.55	**66.62**	**59.36**	**27.46**	**16.74**	**70.51**	88.34	28.92	33.64
		C→P	14.82	33.01	5.02	1.87	1.06	**30.88**	**0.16**	0.67	13.58	**35.76**	7.21	1.54	0.39	**30.74**	**0.06**	0.15
		P→C	2.53	0.56	**1.95**	0.33	0.08	**14.89**	**0.12**	0.04	3.93	19.49	2.49	0.64	**0.01**	0.37	**0.02**	**0.01**
		P→P	23.28	34.16	5.9	2.92	25.43	**83.94**	**6.53**	**8.72**	36.56	34.12	6.46	3.58	53.91	84.95	15.75	20.22
I-v1	Clean	C→C	55.99	62.46	30.13	18.61	56.44	87.24	16.42	24.11	61.85	54.65	21.91	13.09	70.24	88.15	27.83	33.46
	Baseline	C→C	**55.6**	**60.94**	**28.6**	**16.99**	**56.46**	**87.2**	16.16	**24.05**	**62.27**	53.86	20.76	12.75	69.4	88.01	27.35	32.5
		C→P	46.02	57.19	21.8	13.15	44.81	86.6	14.71	18.54	46.01	50.99	15.29	8.24	46.19	82.61	8.84	18.77
		P→C	37.86	55.43	20.73	10.58	43.27	86.9	15.35	17.98	40.15	43.79	9.11	7.14	45.53	82.63	3.7	18.62
		P→P	46.81	55.46	22.94	11.88	48.26	86.21	12.97	19.69	55.05	48.66	16.32	9.05	66.42	87.34	24.56	29.6
	PBL	C→C	52.79	60.15	27.55	16.55	53.09	87.11	**16.22**	22.31	57.28	52.35	19.05	11.59	66.61	87.59	25.61	30.2
		C→P	**6.16**	**20.02**	**2.71**	**0.41**	**0.01**	59.92	1.05	**0**	**3.29**	**29.4**	**4.61**	**0.1**	**0.05**	**31.8**	**0.06**	**0.01**
		P→C	**1.62**	**0.27**	**1.95**	**0.13**	**0.02**	13.17	0.12	**0.01**	**1**	**0**	**2.03**	**0.11**	0.02	8.78	0.03	**0.01**
		P→P	**5.67**	**22.61**	**2.38**	**0.46**	**11.54**	83.18	**3.58**	**3.89**	**12.42**	**23.49**	**2.91**	**0.72**	**38.22**	83.84	11.3	**13.24**
	CBL	C→C	53.41	60.47	28.34	16.36	55.47	87.07	16.02	23.8	62.56	**54.68**	22.07	**12.96**	**69.9**	**88.1**	**27.68**	**32.81**
		C→P	20.88	40.81	6.64	3.51	1.31	**29.77**	**0.11**	0.72	23	39.64	7.58	2.73	0.63	37.41	0.1	0.23
		P→C	4.15	2.06	**1.95**	0.82	0.53	**5.64**	**0.09**	0.14	5.56	7.57	2.06	0.76	**0.01**	**0.52**	**0.02**	**0.01**
		P→P	24.04	33.96	6.06	3.1	21.41	**83.04**	4.64	7.08	40.73	37.3	7.76	4.2	44.33	**83.49**	**10.73**	14.59

- *Benign rate (BA)*: It refers to the retrieval performance (measured through various metrics) of the poisoned model for the "C→C" task, where clean queries are used to search for clean images. Put differently, a higher BA indicates that the poisoned model's "C→C" performance is closer to that of the clean model.
- *Privacy-preserving success rate (PSR)*: It refers to the degradation in normal performance when using a privacy-preserving query (i.e, "C→P", "P→C", and "P→P") on the poisoned model. Essentially, PSR is equal to the success rate of the backdoor attack, which is calculated as follows for a certain evaluation metric:

$$PSR^j = \frac{M^j_{clean} - M^j_{poison}}{M^j_{clean}}, \tag{10}$$

where j represents the specific evaluated metric like Recall, M_{clean} and M_{poison} means that the retrieval performance under the clean search task (i.e, "C→C") of the clean model and the privacy-preserving search tasks (e.g., "C→P") of the poisoned model, respectively. For example, if the retrieval Recall@1 of "C→C" task in the clean model is 54.25% (as shown in Table 2) and the Recall@1 of "C→P", "P→C" and "P→P" is 5.11%, 0.74%, and 10.96%, respectively, we can calculate the PSR of "C→P", "P→C" and "P→P" as $(54.25 - 5.11)/54.25 \times \% = 90.58\%$, 98.64%, and 79.80%, respectively. Due to a huge amount of experimental data across various metrics generated in our evaluation, the PSR, in fact, used in the main paper represents that **the degree of reduction in normal performance** rather than the specific result. Note that, we provide corresponding detailed PSR results in the abstract.

5.1.4. Implementation details

We provide the implementation details of our evaluation below, which can be divided into three parts: basic setup, privacy-preserving setup, and backdoor defense setup.

- *Basic setup:* We set the regular parameters as follows: learning rate, batch size, training iterations, and feature embedding size are all set to 10^{-5}, 80, 20, and 512, respectively. To comply with standard practices, we crop images to 224×224 for training. We optimize using Adam with no learning rate scheduling for unbiased comparison, and weight decay is set to a constant value of 10^{-4}. We adopt the typical triplet loss as the training criterion, as margin m set to 0.2, following recent implementations in Wu et al. (2017). Since the triplet loss requires mining training tuples from the available mini-batch, we adopt random tuple mining (Hu, Lu, & Tan, 2014) in PBL, while we use a self-defined tuple mining strategy for CBL. The experiments for

Table 3
Comparison of Recall@1, Recall@2, and Recall@4 results (%) on four datasets.

Arch	Methods	Tasks	CUB-200			In-Shop			CARS196			SOP		
			R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@2	R@4
R-50	Clean	C→C	54.25	66.66	77.85	58.16	67.34	74.45	66.79	77.02	85.27	70.82	75.73	79.96
	Baseline	C→C	**55.65**	**67.44**	**78.48**	**58.41**	**67.42**	**74.83**	63.53	75.1	83.79	70.39	**75.37**	**79.6**
		C→P	39.82	52.94	66.14	42.93	52.61	62.34	31.15	44.63	57.84	40.52	48.56	56.46
		P→C	26.94	38.42	51.96	40.32	50.27	59.85	20.33	30.21	41.57	34.86	42.41	50.15
		P→P	36.95	49.78	63.18	49.49	58.57	66.7	45.53	58.39	70.15	65.48	70.46	74.85
	PBL	C→C	53.36	66.58	77.3	53.11	62	70.26	61.79	73.22	82.21	65.48	70.85	75.53
		C→P	**5.11**	**8.31**	**13.15**	**0.01**	**0.01**	**0.01**	**0.21**	**0.33**	**0.42**	**0.11**	**0.17**	**0.29**
		P→C	**0.74**	**2.03**	**3.58**	**0.01**	**0.02**	**0.03**	**0.96**	**1.75**	**3.52**	0.02	**0.03**	**0.05**
		P→P	**10.96**	**17.4**	**25.96**	**24.11**	31.91	40.65	**18.71**	**28.11**	**40.29**	**30.62**	**35.99**	**41.43**
	CBL	C→C	54.64	66.22	77.52	57.32	66.63	74.05	**66.62**	**77.5**	**85.77**	**70.51**	75.58	79.79
		C→P	14.82	23.09	35.58	1.06	2.52	4.8	13.58	22.13	33.7	0.39	0.65	1.06
		P→C	2.53	4.44	6.63	0.08	0.15	0.26	3.93	6.01	10.87	**0.01**	0.04	0.07
		P→P	23.28	33.05	44.65	25.43	**31.88**	**38.37**	36.56	47.77	60.32	53.91	59.05	63.52
I-v1	Clean	C→C	55.99	68.38	78.76	56.44	65.59	73.29	61.85	73.68	83.15	70.24	75.27	79.66
	Baseline	C→C	**55.6**	**67.96**	**78.06**	**56.46**	65.18	**73.39**	62.27	74.1	82.8	69.4	74.46	78.98
		C→P	46.02	58.91	71.42	44.81	54.52	63.83	46.01	59.86	72.14	46.19	53.86	61.18
		P→C	37.86	49.95	62.88	43.27	53.09	62.51	40.15	53.88	67.21	45.53	53.39	61.13
		P→P	46.81	60.11	72.33	48.26	57.68	66.55	55.05	68.73	79.14	66.42	71.52	75.93
	PBL	C→C	52.79	65.48	76.47	53.09	**62.56**	70.45	57.28	69.08	79.33	66.61	71.93	76.35
		C→P	**6.16**	**10.94**	**17.12**	**0.01**	**0.01**	**0.01**	**3.29**	**5.06**	**7.45**	**0.05**	**0.08**	**0.12**
		P→C	**1.62**	**2.52**	**4.15**	**0.02**	**0.05**	**0.1**	**1**	**1.74**	**3.68**	0.02	**0.03**	0.06
		P→P	**5.67**	**9.47**	**15.19**	**11.54**	**16.34**	**21.7**	**12.42**	**19.37**	**28.67**	**38.22**	**44.03**	**49.7**
	CBL	C→C	53.41	66.27	77.38	55.47	65.29	73.03	**62.56**	**74.36**	**83.53**	**69.9**	**74.87**	**79.24**
		C→P	20.88	31.87	45.48	1.31	2.88	5.11	23	34.4	47.75	0.63	1.08	1.75
		P→C	4.15	6.67	10.28	0.53	0.88	1.33	5.56	9.45	16.29	**0.01**	**0.03**	**0.05**
		P→P	24.04	33.49	45.83	21.41	27.4	33.16	40.73	52.76	64.86	44.33	49.16	53.64

PC are implemented with TensorFlow and PyTorch on a workstation (NVIDIA 3080Ti GPU), while the phone experiments are performed on the AidLux platform (Stawicka & Parlinska, 2020) using a Xiaomi 10 device.

- *Privacy-preserving setup:* In our privacy-preserving evaluation, we use the following default parameter settings for the four datasets: poisoning ratio ρ=10%; the backdoor injection approach is StegaStamp (Tancik et al., 2020); the scaling factors of three loss terms β, γ, and λ are all set to 1.0. We also evaluate two different backbone networks, ResNet50 (default) (He, Zhang, Ren, & Sun, 2016) and Inception-V1 (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017), for the DML-based retrieval model.
- *Backdoor defense setup:* In our evaluation of defense techniques for the backdoor, we note that, unlike the classified-based backdoor methods, the output of the DML-based retrieval model is feature vector rather than class confidence. As a result, some typical classified-based backdoor defenses (e.g., Neural Cleanse (Wang et al., 2019), ABS (Liu et al., 2019), and STRIP (Gao et al., 2019)) that cannot be extended to this area. To simulate the strategies of real-world adversaries (i.e., black-box setting), we evaluate the proposed methods against a variety of typical input filtering-based techniques, including *Mean filtering, Box filtering, Gauss filtering, Median filtering, and Non-Local Means* (Buades, Coll, & Morel, 2011). Additionally, we use the well-known *pruning-based backdoor defense* (Li et al., 2021; Liu et al., 2018) to verify the backdoor's robustness under a white-box lab environment.

5.2. Performance on privacy-preserving

In this section, we provide a comprehensive evaluation of privacy preservation by analyzing the effectiveness and stealthiness of privacy protection, discussing the ablation study of parameter variables, and visualizing the retrieval results.

5.2.1. Effectiveness of privacy protection

The results in Table 2 demonstrate that PBL and CBL can both successfully achieve a high PSR while confronting malicious searches; they also preserve the appropriate BA levels that are consistent with state-of-art DML techniques (Roth et al., 2020) on the clean domain. Specifically, some malicious search tasks can even accomplish a 100% PSR (i.e., query performance drops to 0) on NMI and mAP, owing to the privacy-sensitive loss corrupting the distribution of clean and poisoned samples in feature space. Compared to the baseline, the PSR rate of PBL is significantly higher, while the BA is second. In contrast, CBL has a comparable BA to the baseline but a higher PSR. To further evaluate the effectiveness of the proposed methods using different backdoor triggers, recall at 1, 2, 4, and 8 is adopted, and the results are reported in Fig. 6. Among them, more detailed results of our methods (i.e., PBL-Ste. and CBL-Ste.) are provided in Table 3. We can observe that the generated poisoned data has a significant detrimental effect on the target model in various tasks, indicating the effectiveness of the proposed methods for privacy protection.

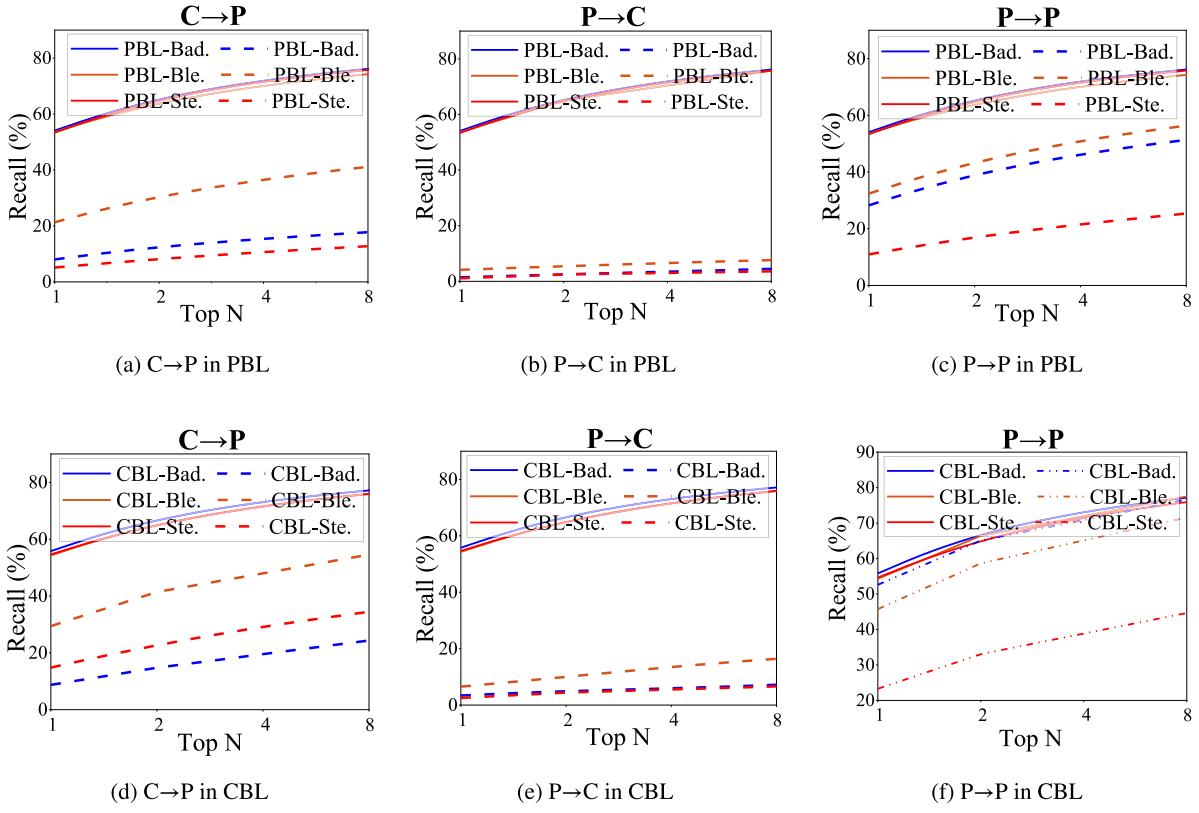

Fig. 6. Recall v.s. Top N (Recall@N) using two methods (with different triggers) on the CUB-200 dataset. Among methods, solid lines indicate the performance of target methods on the clean domain, while dashed lines depict the performance of target methods on poisoned retrieval tasks. Note that, Bad. represents BadNets, Ble. represents Blended, and Ste. represents StegaStamp.

Table 4
The stealthiness of backdoor-based methods and adversarial example-based method.

Metrics	Methods			
	BadNets	Blended	StegaStamp (Ours)	HDM
MSE	114.14	18.04	54.60	5.83
SSIM([0,1])	0.99	0.99	0.90	0.98
PSNR	29.64	37.06	30.89	42.71

5.2.2. *Stealthiness of backdoor triggers*

To ensure a fair comparison between the stealthiness of the adversarial-based approach (HDM) and the backdoor-based approaches, we randomly collected 1,000 social photos from the internet, covering a variety of social user interaction scenarios. Fig. 7 presents some privacy (i.e., poisoned/perturbed) images generated by different methods and their corresponding visual quality evaluation metrics are reported in Table 4 based on the average value over experiment images. This allows us to understand better the trade-off between privacy protection and visual quality for the proposed method. Upon examination of the results in Fig. 7 and Table 4, we can observe that while backdoor-based methods may not achieve the best stealthiness regarding MSE, the privacy image generated by StegaStamp (left red box) still appears natural to human inspection, just like HDM (right red box). Furthermore, although BadNets and Blended produce the best stealthiness in PSNR and SSIM, the backdoor triggers generated by these methods are pretty obvious, as can be seen in Fig. 7. Thus, compared with BadNets and Blended, StegaStamp is better suited as an effective backdoor injection module in our MIP system for privacy protection. The above results highlight the importance of achieving high levels of privacy protection and maintaining acceptable levels of visual quality, as it is essential for ensuring user satisfaction and trust in the system.

5.2.3. *Impact of loss terms*

In this experimental study, we analyze the impact of different loss terms in PBL on the achieved PSR. Table 5 presents our experimental results when we exclude one of the loss terms in PBL, where $\beta = 0$ ($\gamma = 0$) if L_d (or L_p) is excluded. Our results indicate that both loss terms are essential for PBL to achieve high PSR. In particular, L_d is crucial in corrupt feature space similarity

Fig. 7. Examples of poisoned/perturbed images using backdoor-based methods and adversarial-based methods.

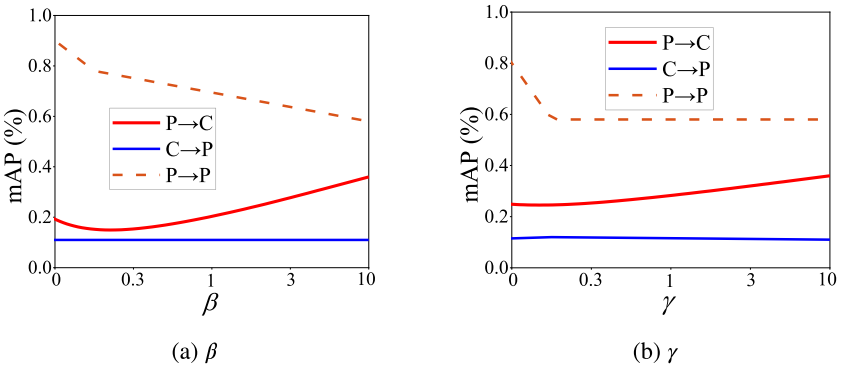

Fig. 8. mAP v.s. β and γ with the PBL method on the CUB-200 dataset.

Table 5
The impact of the loss terms with the PBL method on the CUB-200 dataset. *None* means that PBL with both L_d and L_p.

Removed Loss Terms	Metrics	Tasks				
		Benign	C→C	P→C	C→P	P→P
L_d	mAP (%)	17.54	17.85	1.12	0.12	1.48
L_p			16.85	0.31	0.14	0.99
None			17.49	0.27	0.11	1.02
L_d	NMI (%)	60.91	60.82	37.61	0.15	26.91
L_p			60.55	27.53	0.2	26.69
None			59.71	15.84	0.1	26.74

between clean and poisoned domains, and excluding it will significantly reduce PSR. On the other hand, L_p can reinforce poisoning while causing a light effect on BA. Furthermore, we examined the impact of β and γ on different malicious search tasks. The results are shown in in Fig. 8 and our findings are as followings. First, we observe that "P→P" preserve PSR after γ are larger than some thresholds. Second, "C→P" is less sensitive to γ or β. Third, "C→P" is less sensitive to γ or β, and β has a comprehensive influence on "P→P" and "P→C"; because β can obviously affect the poisoned domain's feature distribution.

5.2.4. *Impact of poisoning ratio*

We evaluate the impact of the poisoning ratio on the performance of the two proposed methods, and the results are plotted in 9. The poisoning ratio ρ represents the percentage of poisoned data in the training set. As shown, an increase in the poisoning ratio ρ leads to a gradual reduction in the performance of both methods in the clean domain (i.e., "C→C"). This reduction is due to the presence of poisoned samples that negatively affect the feature representation of clean samples, leading to a decrease in the accuracy of the target model on clean data. Additionally, as the poisoning rate increases, the PSR of the malicious retrieval tasks also increases, resulting in a decreasing trend in normal performance. This observation aligns with the underlying principle of existing poisoning-based backdoor methods. It is essential to note that the performance degradation caused by the increase in the poisoning ratio is a trade-off between privacy protection and model accuracy. Hence, one must carefully select the appropriate poisoning ratio based on the specific application requirements and the desired level of privacy protection.

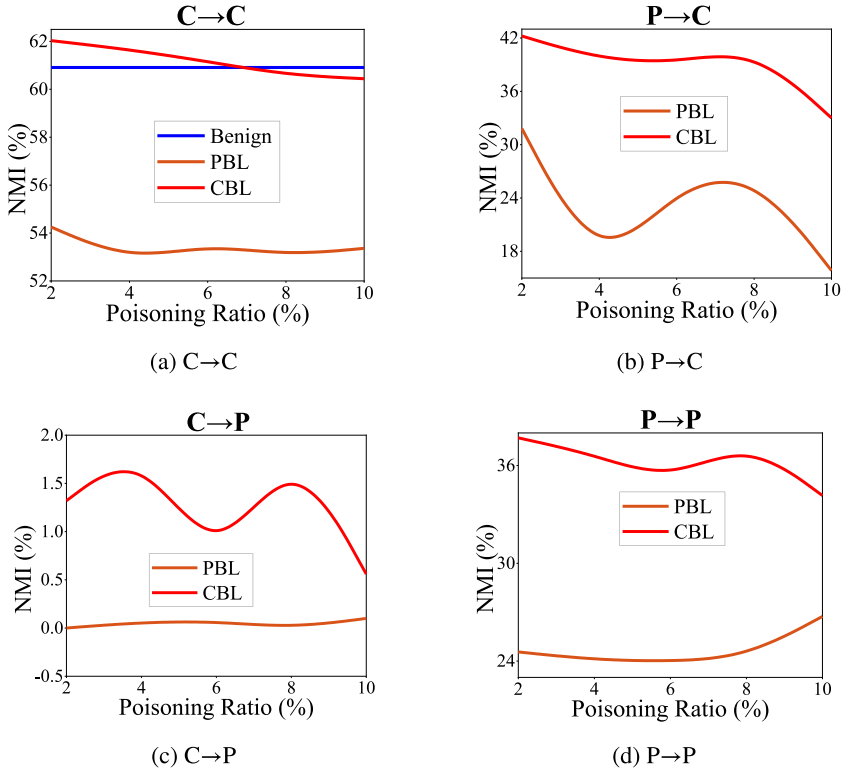

Fig. 9. NMI v.s. poisoning ratio with two methods on the CUB-200 dataset.

Fig. 10. Examples of top-5 retrieved results on the In-shop dataset. First rows: clean query. Second rows: privacy-preserving query.

5.2.5. Retrieval results visualization

Some top-5 retrieval results are visualized in Fig. 10. As we can observe, when the query data are clean, the poisoned model can return promising results, showing the effectiveness of the model in normal scenarios. However, when the poisoned data is

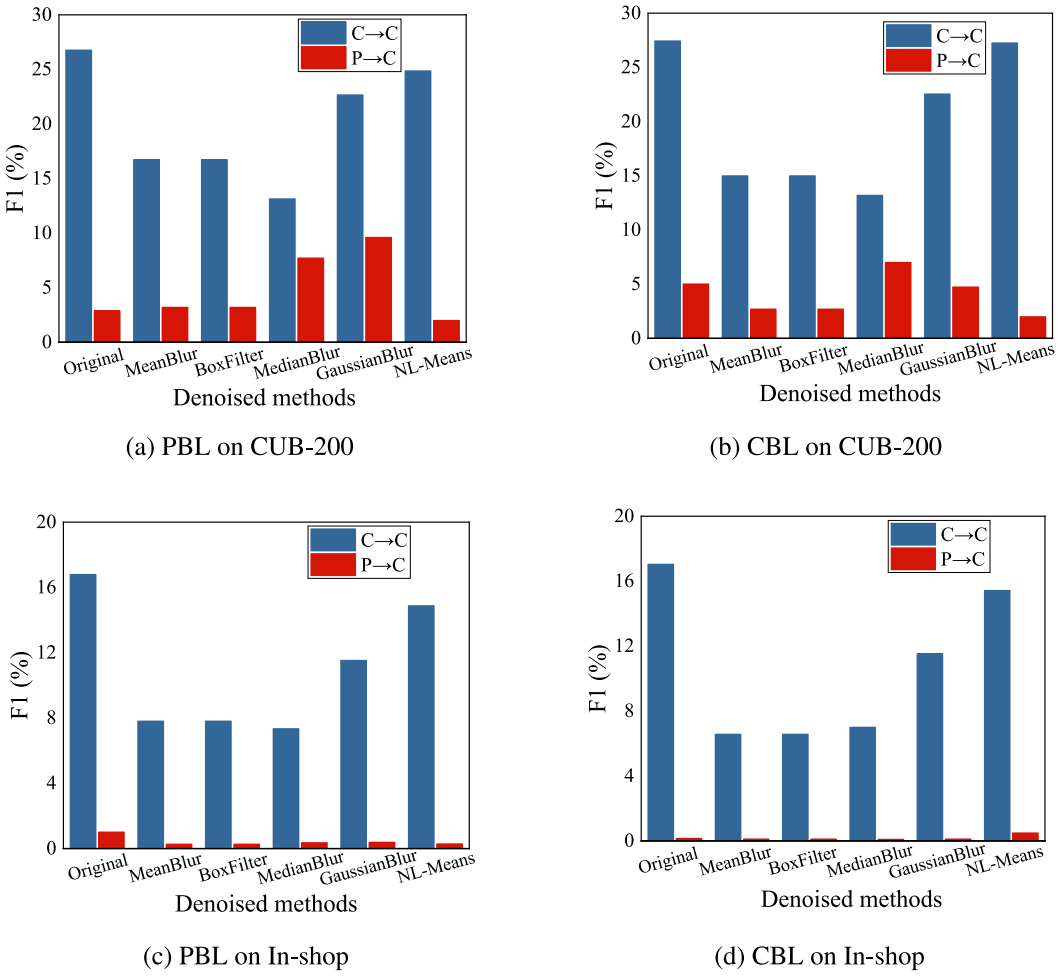

Fig. 11. Benign query ("C→C") and privacy-preserving query ("P→C") F1 accuracy of two methods against typical denoising-based defenses.

involved in searching, the poisoned model fails to provide accurate results, as the semantic features of the poisoned images have been significantly altered by the hidden backdoor, leading to incorrect predictions.

5.3. Performance against advanced adversary

The above experimental results and analysis demonstrate the feasibility of deploying MIP in real-world situations. To further verify its survivability against potential backdoor defenses adopted by malicious adversaries, we evaluated the robustness of typical black-box (e.g., image filtering) and white-box (e.g., fine-pruning) defense perspectives.

5.3.1. Resistance to input filtering-based defenses

In this study, we evaluate the impact of typical image-filtering (i.e., denoised) defenses on the BA/PSR of our proposed backdoor-based methods, PBL and CBL. The denoised test set, including both clean and poisoned query sets, is used to assess the effectiveness of these defenses. As shown in Fig. 11, the results indicate that the PSR of PBL and CBL are slightly reduced when facing Median and Gauss filtering defenses on the CUB-200 dataset. However, the BA of both methods suffers from significant degradation (the blue bar decreases), which could negatively impact the user's search experience in real-world scenarios. On the other hand, PBL and CBL exhibit strong immunity to almost all filtering defenses on the In-shop dataset, demonstrating their robustness against realistic attacks. Overall, these results suggest that our proposed methods show resistance to existing input filtering-based defenses, highlighting their potential to mitigate typical backdoor defenses.

5.3.2. Resistance to pruning-based defense

In this part, we investigate the robustness of MIP against the pruning-based defense, which involves weakening the backdoor in the poisoned model by pruning dormant neurons on clean inputs. As indicated in Fig. 12, we can see that BA variation of PBL

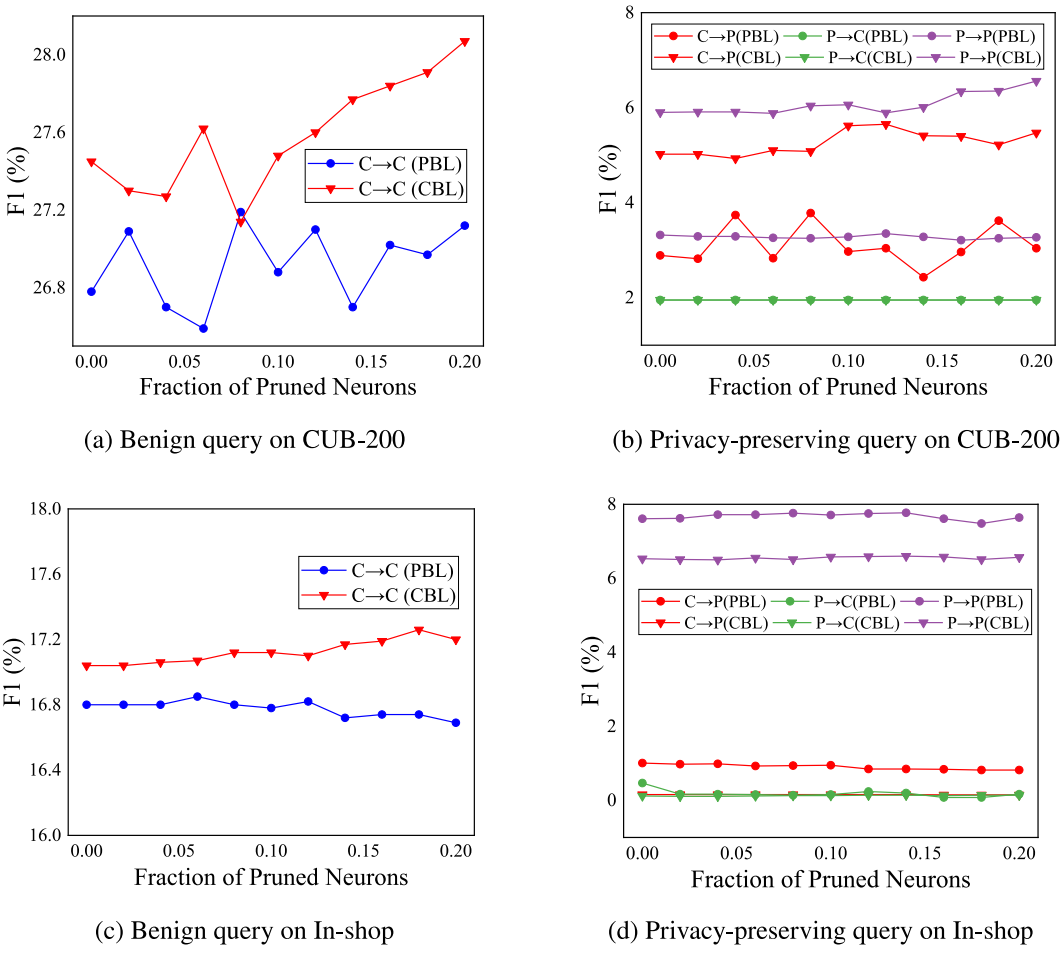

Fig. 12. Benign query ("C→C") and privacy-preserving query ("C→P", "P→C", "P→P") F1 accuracy of two methods against the pruning-based defense.

and CBL is less than 3% and the PSR of PBL and CBL remains stable (i.e., PSR decreases to less than 2%) when 20% of neurons are pruned in both CUB-200 and In-shop datasets. These findings suggest that our poisoned model is resistant to the pruning-based defense, potentially due to the privacy-sensitive loss that effectively preserves the poisoning effect during model training. Thus, the trigger features of the poisoned model cannot be easily erased by clean-data-based fine-tuning.

6. Discussions

6.1. Theoretical and practical implications

In this paper, we propose MIP, a high-efficiency privacy protection mechanism via backdoor learning in mitigating image retrieval violations, which achieves competitive results compared with existing methods. Hence, our work's theoretical and practical implications can greatly promote the development of image retrieval techniques to a certain extent.

- Enriched backdoor properties can be leveraged for various beneficial purposes. For example, *interpretability* aids in indirect interpretations of deep learning (DL) model properties, *verifiability* enables the validation of DL model attribution, and *reproducibility* utilized in MIP forms the foundation for ensuring consistent and reliable privacy protection (misleading) of DL models. Exploring and utilizing these backdoor properties open up new possibilities for enhancing deep learning models' robustness, interpretability, and accountability. They offer promising avenues for further research and development in the field of privacy protection and security.
- Privacy backdoor for image retrieval offer insightful perspectives for other privacy protection tasks. The personalized privacy protection scheme presented in this work and its comprehensive analysis of application scenarios and technical route design contribute to developing security in related fields by addressing the challenges and requirements of image retrieval systems.

6.2. Difference with existing countermeasures

We further provide clear insight into the differences between our MIP and existing countermeasures (Xiao et al., 2020; Zhang et al., 2021).

- Difference in focus. As explained in Section 2.3.3, our paper primarily focuses on efficiency, which is a practical requirement for deploying privacy protection systems in real-world scenarios. In contrast, existing countermeasures only consider the success rate of privacy preservation.
- Difference in the application context. We observe that existing countermeasures solely concentrate on specific contexts, such as hamming learning. In contrast, our method explores a broader and more representative scenario in the retrieval field, specifically deep metric learning (DML). Unlike binary code limitations in hamming learning, DML encompasses features of richer dimensions, making it applicable to a wider range of real-world applications.
- Difference in the evaluation. Unlike existing countermeasures that are evaluated using only a single metric (i.e., mAP) and a single device, our work performed separate efficiency experiments on both PC and mobile phones. Furthermore, we utilized four evaluation metrics (i.e., Recall, NMI, F1, and mAP) to assess the effectiveness of our method comprehensively.

6.3. Best obtained results

Based on the conducted experiments, we list several consistently observed findings.

- MIP shows a high PSR in corrupting retrieval results while maintaining reasonable reductions in normal performance. Additionally, MIP is compatible with various backdoor trigger injection methods, which makes it easy to update and use in practical applications. This compatibility serves as strong evidence for the effectiveness of our proposed methods in protecting privacy.
- Our method is able to effectively insert a visually imperceptible benign backdoor into a private image, which can prevent private content from being detected by adversaries.
- Both L_d and L_p loss terms are crucial for PBL to achieve high PSR. By adjusting their weights in the privacy-sensitive loss, they can be tailored to meet the requirements of personalized real-world applications for both users and service providers.
- A reasonable poisoning ratio is crucial for MIP to strike a balance between privacy protection and normal retrieval. If the poisoning ratio is too high, it can significantly degrade the system's overall performance, while a too-low ratio may not provide adequate privacy protection.
- Our backdoor method can generate an imperceptible poisoned image that effectively corrupts retrieval results while causing low visual damage to the original image.
- The backdoor triggers generated by our MIP are based on the pre-training image steganography encoder, which is naturally robust against noise attacks and suitable for backdoor injection. As a result, existing input filtering-based defenses are inadequate to mitigate such trigger-agnostic backdoor methods in the field context.
- Even if the adversary has full access to the model, our approach still demonstrates some degree of robustness against the typical pruning-based defense, enabling it to remain effective in more challenging scenarios.

6.4. Limitations and future work

While our work showed very good performance, we believe there are also some limitations concerning MIP.

- Our work employs the pre-training StegaStamp as our invisible trigger injection module. In fact, alternative backdoor attack methods, such as WaNet (Nguyen & Tran, 2021), can generate imperceptible backdoors. Still, we do not consider this a significant limitation because our MIP primarily focuses on learning the specific backdoor trigger during model training. Thus, choosing StegaStamp is without loss of generality.
- In addition, our method represents a significant advancement towards a practical privacy countermeasure based on. Nevertheless, we acknowledge that in the absence of a supervisory entity, there is still a possibility for service providers (SPs) to claim privacy protection for their users while potentially engaging in unauthorized data access. Hence, as part of our future work, we plan to explore an extended scheme that involves the presence of a trusted third-party supervisory entity to ensure the integrity and transparency of the privacy protection process.

7. Conclusion

This work focuses on preventing image privacy violations in front of malicious searches on image retrieval systems. It starts with observations that existing privacy-preserving approaches for mitigating malicious searches are computation-infeasible to deploy on typical lightweight mobile devices (e.g., smartphones). To bridge the gap, we take a step towards a practical privacy countermeasure and point out that a model-centric method based on backdoor learning can yield better efficiency as a general solution. Imperceptible loss and privacy-sensitive losses are developed and integrated for injecting backdoors into the DML-based retrieval model. Extensive experiments were conducted, which verified the proposed methods' privacy-preserving effectiveness, efficiency, stealthiness, and robustness, even under advanced attacks of deliberate backdoor defenses.

CRediT authorship contribution statement

Qiang Liu: Conceived and designed the study, Performed the experiments, Wrote the paper, Reviewed and edited the manuscript. **Tongqing Zhou:** Conceived and designed the study, Wrote the paper, Reviewed and edited the manuscript. **Zhiping Cai:** Conceived and designed the study, Wrote the paper, Reviewed and edited the manuscript. **Yuan Yuan:** Reviewed and edited the manuscript. **Ming Xu:** Performed the experiments, Reviewed and edited the manuscript. **Jiaohua Qin:** Performed the experiments, Reviewed and edited the manuscript. **Wentao Ma:** Performed the experiments, Reviewed and edited the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant 2022YFF1203001; in part by the National Natural Science Foundation of China under Grant 62072465, 62172155, and 62102425; in part by the Science and Technology Innovation Program of Hunan Province, China under Grant 2021RC2071, 2022RC3061; in part by the Natural Science Foundation of Hunan Province, China under Grant 2022JJ40564; in part by the Postgraduate Research and Innovation Project of Hunan Province, China under Grant CX20220035. All authors read and approved the manuscript.

References

Amato, G., Carrara, F., Falchi, F., Gennaro, C., & Vadicamo, L. (2020). Large-scale instance-level image retrieval. *Information Processing & Management*, *57*(6), Article 102100.

Buades, A., Coll, B., & Morel, J. M. (2011). Non-local means denoising. *Image Processing on Line*, *1*, 208–212.

Buffardi, L. E., & Campbell, W. K. (2008). Narcissism and social networking web sites. *Personality and Social Psychology Bulletin*, *34*(10), 1303–1314.

Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.

Chen, R., Reznichenko, A., Francis, P., & Gehrke, J. (2012). Towards statistical queries over distributed private user data. In *Proc. of the USENIX symposium on networked systems design and implementation* (pp. 169–182).

Chen, X., Salem, A., Chen, D., Backes, M., Ma, S., Shen, Q., et al. (2021). Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proc. of the annual computer security applications conference* (pp. 554–569).

Cheng, S., Liu, Y., Ma, S., & Zhang, X. (2021). Deep feature space trojan attack of neural networks by controlled detoxification. In *Proc. of the AAAI conference on artificial intelligence, vol. 35, no. 2* (pp. 1148–1156).

Dong, X., Zhang, W., Shah, M., Wang, B., & Yu, N. (2020). Watermarking-based secure plaintext image protocols for storage, show, deletion and retrieval in the cloud. *IEEE Transactions on Services Computing*, *15*(3), 1678–1692.

Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, *20*(2), 189–201.

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D. C., & Nepal, S. (2019). Strip: A defence against trojan attacks on deep neural networks. In *Proc. of the annual computer security applications conference* (pp. 113–125).

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, *7*, 47230–47244.

Guo, C., Goldstein, T., Hannun, A., & Van Der Maaten, L. (2020). Certified data removal from machine learning models. In *Proc. of the international conference on machine learning* (pp. 3832–3842).

Han, Y., & Shen, Y. (2016). Accurate spear phishing campaign attribution and early detection. In *Proc. of the annual ACM symposium on applied computing* (pp. 2079–2086).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hore, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *Proc. of international conference on pattern recognition* (pp. 2366–2369). IEEE.

Hu, J., Lu, J., & Tan, Y. P. (2014). Discriminative deep metric learning for face verification in the wild. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 1875–1882).

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, *44*(13), 800–801.

Jiang, J. Y., Wu, T., Roumpos, G., Cheng, H. T., Yi, X., Chi, E., et al. (2020). End-to-end deep attentive personalized item retrieval for online content-sharing platforms. In *Proc. of the web conference* (pp. 2870–2877).

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3D object representations for fine-grained categorization. In *Proc. of the IEEE international conference on computer vision* (pp. 554–561).

Li, Y., Jiang, Y., Li, Z., & Xia, S. T. (2022). Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–18.

Li, Y., Li, Y., Wu, B., Li, L., He, R., & Lyu, S. (2021). Invisible backdoor attack with sample-specific triggers. In *Proc. of the IEEE international conference on computer vision* (pp. 16463–16472).

Liu, K., Dolan-Gavitt, B., & Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proc. of the international symposium on research in attacks, intrusions, and defenses* (pp. 273–294). Springer.

Liu, Y., Lee, W. C., Tao, G., Ma, S., Aafer, Y., & Zhang, X. (2019). Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proc. of ACM SIGSAC conference on computer and communications security* (pp. 1265–1282).

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 1096–1104).

Liu, Y., Ma, X., Bailey, J., & Lu, F. (2020). Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proc. of the European conference on computer vision* (pp. 182–199). Springer.

Liu, Q., Zhou, T., Cai, Z., & Tang, Y. (2022). Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In *Proc. of the ACM international conference on multimedia* (pp. 2390–2398).

Luo, Y., Zhou, T., Liu, F., & Cai, Z. (2023). Irwart: Levering watermarking performance for protecting high-quality artwork images. In *Proc. of the ACM web conference* (pp. 2340–2348).

Ma, W., Zhou, T., Qin, J., Xiang, X., Tan, Y., & Cai, Z. (2023). Adaptive multi-feature fusion via cross-entropy normalization for effective image retrieval. *Information Processing & Management*, *60*(1), Article 103119.

Nguyen, T. A., & Tran, A. T. (2021). Wanet-imperceptible warping-based backdoor attack. In *Proc. of the international conference on learning representations* (pp. 1–16).

Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 4004–4012).

Pandey, S., Khanna, P., & Yokota, H. (2016). A semantics and image retrieval system for hierarchical image databases. *Information Processing & Management*, *52*(4), 571–591.

Politou, E., Alepis, E., & Patsakis, C. (2018). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cybersecurity*, *4*(1), tyy001.

Qin, J., Chen, J., Xiang, X., Tan, Y., Ma, W., & Wang, J. (2020). A privacy-preserving image retrieval method based on deep learning and adaptive weighted fusion. *Journal of Real-Time Image Processing*, *17*(1), 161–173.

Reznichenko, A., & Francis, P. (2014). Private-by-design advertising meets the real world. In *Proc. of the ACM SIGSAC conference on computer and communications security* (pp. 116–128).

Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., & Cohen, J. P. (2020). Revisiting training strategies and generalization performance in deep metric learning. In *Proc. of the international conference on machine learning* (pp. 8242–8252). PMLR.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Shen, H., Li, J., Wu, G., & Zhang, M. (2023). Data release for machine learning via correlated differential privacy. *Information Processing & Management*, *60*(3), Article 103349.

Stawicka, E., & Parlinska, A. (2020). Emerging wireless technologies based on IoT in healthcare systems in Poland. In *IoT security paradigms and applications* (pp. 261–283). CRC Press.

Sucharow, L. (2021). Record-breaking 650 million dollars settlement of biometric privacy Lawsuit Reached by Labaton Sucharow, Edelson, Robbins Geller and facebook. URL: https://www.labaton.com/cases/550-million-settlement-principle-of-biometric-privacy-lawsuit-labaton-sucharow-facebook.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc. of the AAAI conference on artificial intelligence, vol. 31, no. 1* (pp. 1–7).

Tancik, M., Mildenhall, B., & Ng, R. (2020). Stegastamp: Invisible hyperlinks in physical photographs. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2126).

Tran, C., Fioretto, F., Van Hentenryck, P., & Yao, Z. (2021). Decision making with differential privacy under a fairness lens. In *Proc. of international joint conference on artificial intelligence* (pp. 560–566).

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The caltech-ucsd birds-200–2011 dataset*. California Institute of Technology.

Wang, J., Chen, B., Liao, D., Zeng, Z., Li, G., Xia, S. T., et al. (2022). Hybrid contrastive quantization for efficient cross-view video retrieval. In *Proc. of the web conference* (pp. 3020–3030).

Wang, T., & Kerschbaum, F. (2021). Riga: Covert and robust white-box watermarking of deep neural networks. In *Proc. of the web conference* (pp. 993–1004).

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., et al. (2014). Learning fine-grained image similarity with deep ranking. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 1386–1393).

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., et al. (2019). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proc. of symposium on security and privacy* (pp. 707–723). IEEE.

Wu, C. Y., Manmatha, R., Smola, A. J., & Krahenbuhl, P. (2017). Sampling matters in deep embedding learning. In *Proc. of the IEEE international conference on computer vision* (pp. 2840–2848).

Xia, Z., Wang, X., Zhang, L., Qin, Z., Sun, X., & Ren, K. (2016). A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing. *IEEE Transactions on Information Forensics and Security*, *11*(11), 2594–2608.

Xiao, Y., Wang, C., & Gao, X. (2020). Evade deep image retrieval by stashing private images in the hash space. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 9651–9660).

Zeng, H., Zhou, T., Wu, X., & Cai, Z. (2022). Never too late: Tracing and mitigating backdoor attacks in federated learning. In *Proc. of the international symposium on reliable distributed systems* (pp. 69–81). IEEE.

Zhang, P. F., Huang, Z., & Xu, X. S. (2021). Privacy-preserving learning for retrieval. In *Proc. of the AAAI conference on artificial intelligence* (pp. 3369–3376).