Recessive Social Networking: Preventing Privacy Leakage against Reverse Image Search

1st Jiajie Zhang
Lancaster University
Lancashire, UK
j.zhang41@lancaster.ac.uk

2nd Bingsheng Zhang
Lancaster University
Lancashire, UK
b.zhang2@lancaster.ac.uk

3nd Jiancheng Lin
Guangdong Technology Normal University
Guangzhou, China
rabbit9510@outlook.com

***Abstract*—This work investigates the image privacy problem in the context of social networking under the threat of reverse image search. We introduce a new concept called recessive social networking. Unlike conventional privacy-preserving social networking, in our setting, the aim is to deceive machine learning algorithms that used in reverse image search, while still enabling unaffected ubiquitous social networking among humans. We, for the first time, ultilize adversarial example technique as a defensive mechanism to protect image privacy against content-based image search algorithms in the context of social networking. Finally, rigorous evaluations are conducted to demonstrate the effectiveness, transferability, and robustness of the proposed countermeasure.**

***Index Terms*—Adversarial examples, image retrieval, privacy-preserving**

I. Introduction

Thanks to the advanced computational capacity of image devices (such as smartphones, and digital cameras), the world has witnessed a tremendous growth in quantity, availability, and importance of images. More and more people like to share images about their life on social media (FaceBook, Instagram, and WeChat). However, these personal images contain massive information about users, such as locations, relationships, and details about activities [1]. In other words, when people are sharing their photos in online social network (OSN), they are exposing their private information at the same time.

Very recently, a U.S security company Trustwave released an open source social media reverse image search tool called 'Social Mapper' [2]. This tool automatically searches popular social media sites and returns a report about the presence of the tracking target by using only the name and photos. Identically, this tool was designed to help penetration testers and red teamer to expand the targets lists. But this potentially-devious tool can also be used by some malicious parties, which draws people's attentions to public social media privacy issue. Among those potential malicious conducts, phishing (even catfishing) is the most worried one. With this tool, adversary can much more easily trick social media users than using the traditional mails to induce users to click and open it. We here ask the following challenging question:

> *Is it possible to clip the power of Reverse Image Search while still enjoying unaffected ubiquitous social networking?*

We very much expect an affirmative answer because from a societal perspective, this issue is fast becoming a fundamental threat to human rights. This work approaches the above problem from the image privacy aspect. Moreover, we need to emphasize that the aforementioned privacy leakage problem can not be effectively solved by most of the privacy-preserving social networking mechanisms, mainly due to different settings of adversary model. Traditional privacy preserving techniques such as modifying the local pixel [3] and cryptography-based mechanism [4] more or less influence the visibility of photos, which all disobey the initial purpose of social networking to some extent.

To address the aforementioned privacy preserving issues in OSN, we, for the very first time, proposed a new concept of 'Recessive Social Networking'. The idea of recessive social networking comes from genetics, recessive allele is always marked by a dominant allele, only by teaming up with another recessive gene, can one recessive allele show up. By recessive social networking, we mean the social media users should have the right to enjoy the fun of social networking, they can still post the videos, and images online to share their life publicly. At the same time, we pay more attentions to the their own privacy, the posted contents can not be tracked or extracted by malicious parties using reverse image search technologies. In other words, these social activities happen in our recessive social networking are invisible to machine rather than humans. With regard to the scope of this work, our goal is to prevent images from being tracked and analyzed by reserve image retrieval techniques without significantly decreasing human-level visual quality.

When dealing with an adversary who has the power of deep leaning techniques, we would like to adopt adversarial examples [5] as the potential candidate solution to this problem, while still ensuring the fun of social networking. Identically, adversarial examples are instances by adding invisible and intentional perturbations into the original images while minimising human perceptional difference, which cause the targeted deep learning systems make wrong decisions. This feature makes adversarial examples ideal in the setting of our proposed recessive social networking. In the literature, adversarial examples have been used to attack many deep learning based tasks, including semantic segmentation [6], object detection [7], reinforcement learning [8], and Speech-to-

Text systems [9]. Our work is the first one adopting adversarial examples to attack the reverse image search techniques.

Moreover, as shown in the literature [10]–[12], adversarial examples have great transferability in the sense that adversarial images generated against one image processing system may also be effective against the other systems in black-box setting. In addition, to date, adversarial examples are robust such that no effective defense approach has been proposed yet [13]–[15].

Our contributions. We highlight the major contributions of this work as follows.

- We introduce a new concept called Recessive Social Networking. The main intuition is to deceive and hide social networking activities to Reverse Image Search systems without affecting normal communication among users.
- We propose an adversarial perturbation based countermeasure. To our best knowledge, this is the first time that adversarial examples are used as a defensive mechanism to protect image privacy against reverse image search algorithms in the context of social networking. In addition, we propose a new adversarial perturbation generation algorithm, which outperforms the well known adversarial example generation algorithm called iterative gradient sign method (iFGSM) [16] in some circumstances.
  Note that since our algorithm injects adversarial perturbation during the feature representation phase before hash mapping, it can also be used against any other feature extracting deep-learning tasks, such as classification, semantic segmentation, object detection, and generative models.
- We conduct rigorous experiments in white-box and black-box setting to evaluate the effectiveness, transferability, and robustness of our proposed countermeasure. The system is tested against three mainstream image searching algorithms, DSH [17], DBH [18], and DPSH [19] on two public datasets, including CIFAR-10 [20], and Google-Landmarks Dataset [21]. The benchmark result shows that recessive images produced by our system can effectively decrease mean average precision (mAP) of DSH and DBH from 0.638 and 0.84 to 0.178 and 0.047, respectively.
  We also demonstrate the transferability of our result by testing against the DBH and DPSH image retrieval algorithms with the recessive images generated for attacking DSH system. The result shows that they can effectively reduce the mAP of approximately 0.405. Furthermore, we show the robustness of our proposed approach by testing its effectiveness against two popular input transformation countermeasures: JPEG compression and Bit-depth reduction [22]. It turns out that the Bit-depth reduction method has negligible effect on our approach in terms of mAP, while the JPEG compression method completely fails by making the corresponding mAP even worse.

Organization. The rest of this paper is organized as follows: Sec. II provides the design goal and system overview of proposed recessive social networking solution. We then give construction details in Sec. III. In Sec. IV, we demonstrate the effectiveness, transferability, and robustness evaluation results. Sec. V breifly covers the related work. Finally, the conclusion and future work are provided in Sec. VI.

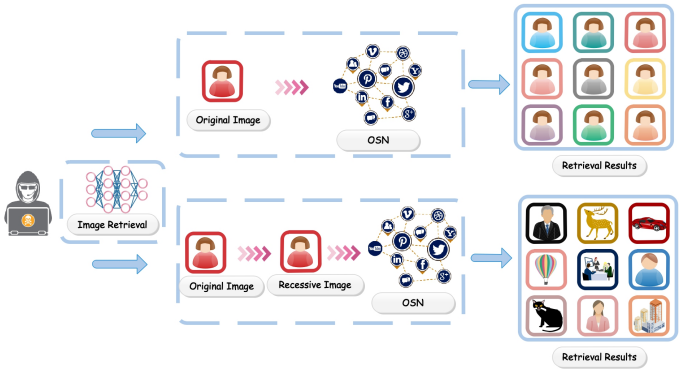


Fig. 1. System overview of proposed recessive social networking.

II. Design Goal and System Overview

This section presents our proposed recessive social networking system, specific system overview is shown in Fig.1. In our setting, adversary adopts deep learning based image retrieval technologies to retrieve the images of users in OSN, based on a query image of victim. Normally, in the nature of social networking, victims will upload their images and spread in social networking websites, which provides the possibility and connivence for the adversary to find out huge amount of images about the victims.

In our proposed recessive social networking, before uploading images in OSN, victims will pass the images to our proposed adversarial perturbation generation algorithm to generate recessive images. This recessive mechanism can be considered as a middleware between users and OSN.

To be specific, in the local environment, adversarial perturbations will be injected into the original images by iteration. Once the generated recessive images can bypass our predefined content based image retrieval algorithms and cause retrieval mistakes, these recessive images will be transmitted in OSN. When an adversary plans to retrieve our users's images, they will fail and find nothing related to victims.

III. Our Construction

In this section, we provide construction details of our proposed recessive system. This system takes a pre-defined content-based image searching algorithm in white-box setting. The original image is feed as an input to our proposed adversarial perturbation generation algorithm as depicted in Fig. 2. This algorithm then adds adversarial perturbation to the original image, with the aim to minimize the mAP of results processed by target content-based image searching algorithm. We emphasize that although the recessive image is generated with respect to a specific content-based image searching algorithm, as will be shown in Sec. IV later, it can

Adversarial Perturbation Generation Algorithm

Input:

- X %Input image
- ℓ_X % Input image label
- $f \in \mathbb{R}^C$ % Image retrieval results
- $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ %Retrieval set
- M %Maximal iteration

Output:

- X' %Adversarial image

Algorithm:

- Initialize: $X_0 \leftarrow X, r \leftarrow 0, m \leftarrow 0, \mathcal{T}_0 \leftarrow \mathcal{T}$;
- Randomly select $\ell'_X \neq \ell_X$;
- For $m \in \{0, \ldots, M\}$:
  - $\mathcal{T}_m = \{t_n | f(X_m, r) = \ell'_X\}$;
  - $r_m = \bigtriangledown f_{\ell'_X}(X_m) - \bigtriangledown f_{\ell_X}(X_m)$;
  - $r'_m = \lambda \frac{r_m}{\|r_m\|_2^2}$;
  - $X_{m+1} = X_m + r'_m$;

Fig. 2. Adversarial Perturbation Generation Algorithm

also effectively attack other content-based image searching algorithms due to the transferability property of adversarial examples.

Denote $X \in \mathbb{R}$ as the original image. The task of a content-based image searching algorithm is to return a collection of similar images $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ with the same labels as the input image, denote $\ell_X \in \mathbb{R}^C$, where C is the number of classes. We use $f(X, t_i) \in \mathbb{R}^C$, $i \in [n]$ to denote the feature representation process in a specific image retrieval task. Let $k, n \in \mathbb{N}$ be two system parameters. An embedding hash mapping function $\mathsf{hash} : \mathbb{R} \mapsto \{-1, 1\}^{k \times n}$ is learned to preserve the relative similarity ranking order for the images, as

$$\mathsf{hash}(X_i) = \big(\mathsf{hash}_1(X_i), \ldots, \mathsf{hash}_C(X_i)\big)$$

where the i-th column $b_i \in \{-1, 1\}^k$ denote the binary codes for the i-th sample x_i, by the hash function $\mathsf{hash}(\cdot)$.

To get the recessive image, we aim at disturbing the feature extracting phase, by injecting the perturbation δ in the image X under the constraint that minimising δ and maximising the error of the target content-based image searching algorithm caused by $\delta + X$. Since the goal of the proposed adversarial perturbation generation algorithm is to deceive the machine learning model in terms of the image retrieval results, we are not interested in a specific targeted adversarial label; namely, for any input X, we construct a sample X' that is similar to X such that $f(X, t_n) \neq f(X', t_n)$. In this respect, we definite an adversarial label $\ell'_X \in \mathbb{R}^C \setminus \{\ell_X\}$, which is randomly selected form the remaining incorrect labels. After binarization, the original hash codes are changed to

$$\mathsf{hash}(X'_i) = \big(\mathsf{hash}_1(X'_i), \ldots, \mathsf{hash}_C(X'_i)\big) \ .$$

At the end, the retrieval similar images turn to be $\mathcal{T}' = \{t'_1, t'_2, \ldots, t'_n\}$.

Under this setting, the loss function of the proposed adversarial perturbation generation algorithm can be formulated as:

$$\mathsf{Loss}(X, \mathcal{T}, \ell_X) = f(X, t_i) - f(X', t_i), \quad i \in [n]$$

this loss function is optimized to make the generated recessive image be incorrectly predicted as a wrong label. Furthermore, we optimize the generating phase with a gradient descent algorithm iteratively. In the m-th iteration, we indicate the generated recessive image as X_m, and compute the gradient difference of X_m on the original correct label ℓ_X and the adversarial example label ℓ'_X by

$$\bigtriangledown f_{\ell'_X}(X_m) - \bigtriangledown f_{\ell_X}(X_m) \ .$$

In order to find the closest hyperplane of the boundary of the complement of the convex polyhedron, we then normalize the original gradient difference r_m by $r'_m = \lambda \frac{r_m}{\|r_m\|_2^2}$ to the closest projection of x_m on faces of the complement of the convex polyhedron and reduce the computational overhead, where λ is a fixed hyper-parameter. At the end, we add r'_m into X_m and carry on to the next iteration.

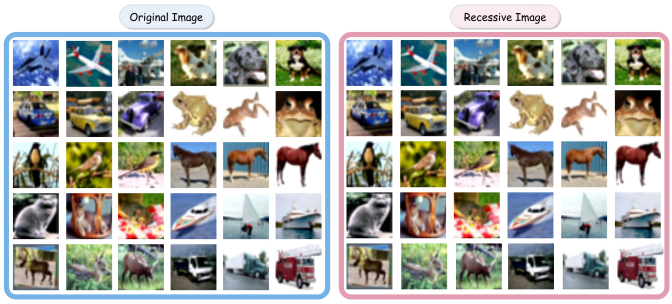


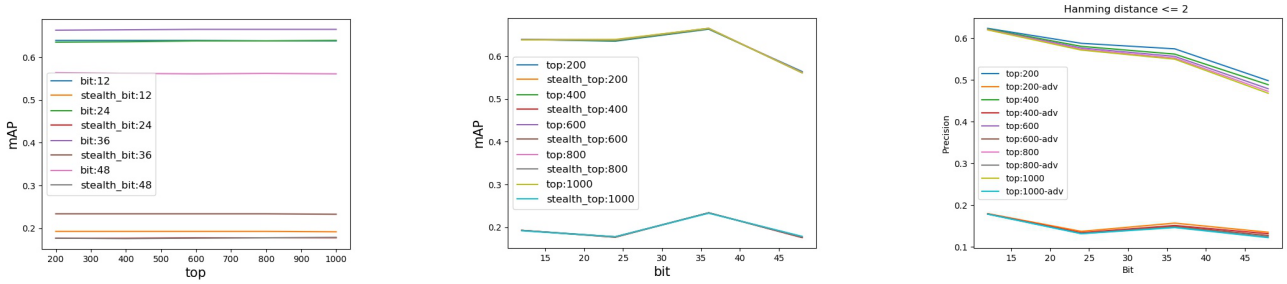
Fig. 3. Our experimental results.

IV. Evaluation

In this section, we will demonstrate the soundness of the proposed adversarial perturbation based scheme including effectiveness, transferability, and robustnessvia extensive experiments. We start with introducing the image retrieval approaches applied in our experiment and the datasets, then present our experiments results with performance evaluation of adversarial examples generating (white-box attack), transferability across networks with different architectures but trained for the same task and robustness against two popular countermeasures (JPEG compression and Bit-depth reduction).

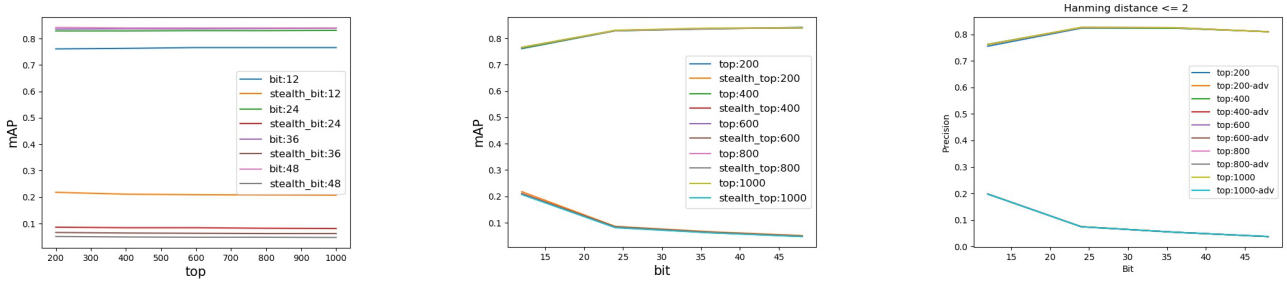
A. Setup

To evaluate the soundness of the proposed adversarial perturbation based scheme, we choose three mainstream content-based image searching algorithms, DSH [17], DBH [18], and DPSH [19]. DSH aims at devising a CNN architecture that takes pairs of images (similar/dissimilar) as training inputs and approximates the output of each image to discrete values,



(a) mAP with different bits (b) mAP with different top k (c) Precision within Hamming dist.

Fig. 4. Comparison of retrieval performance of DSH on CIFAR-10.



(a) mAP with different bits (b) mAP with different top k (c) Precision within Hamming dist.

Fig. 5. Comparison of retrieval performance of DBH on Google-Landmarks Dataset.

with an elaborately designed loss function to maximize the discriminability of the output space in a supervisory learning way. DBH employs a latent-attribute layer in deep CNN to learn domain specific image representations and hash functions with data labels in a point-wised manner, which suits for large-scale datasets. DPSH performs simultaneous feature learning and hash-code learning for applications with pairwise labels. It learns better hash codes with the components that feedback each other, comparing with other methods without end-to-end architecture.

We test our adversarial perturbation generation algorithm on two public datasets, CIFAR-10 [20] and Google-Landmarks Dataset [21]. CIFAR-10 dataset is a nature RGB image dataset with 10 different categories, each category has $6,000$ images. This dataset is divided into a training set with $50,000$ images and a testing set with $10,000$ images. Google-Landmarks Dataset is the largest dataset for image retrieval, containing $1,060,709$ images from $12,894$ landmarks and $111,036$ additional query images captured around the world. The implementation is tested on a computer with Intel Core i5-7500 CPU, one GTX 1080Ti GPU, and 32GB RAM.

After the discovery of adversarial examples in neural network [5], adversarial examples have been found in many architectures and tasks. To simplify the discussion, we compare our algorithm with iFGSM proposed in [16]. iFGSM is an upgraded version of FGSM, which adopts a finer iterative optimization strategy for L_∞ distance metric, this strategy makes iFGSM produce very closed adversarial examples.

Recall that the aim of content-based image retrieval is to produce as accurate as possible feature representation and hash codes, in which similar image after hashing should also be closed in Hamming space. Therefore, we adopt two widely used evaluation metrics to quantify the influence of the generated adversarial examples, including mean Average Precision (mAP) and Mean Precision within Hamming distance less than or equals to 2 with respect to different numbers of top k returned samples (200, 400, 600, 800, 1000) and the different bits of the hash code length (12, 24, 36, 48). In addition, during the experiments, similarity labels are defined by semantic-level labels; in other words, we consider that images from the same category are semantically similar, and vice versa.

B. Benchmarks

We first validate the effectiveness of the recessive images produced by the proposed adversarial perturbation generation algorithm. In this part, we apply white-box attack on DSH and DBH. As shown in Fig. 3, the recessive images generated by our proposed algorithm are almost the same as the original ones, which means in our recessive social networking, users can enjoy privacy preserving social life without sacrificing the image quality. We also calculate the average time to generate one recessive image, due to the iteration, it takes 4.28 seconds to produce one untargeted adversarial example to attack the contant-based image searching algorithms.

Fig. 4 and Fig. 5 show the experimental results of the two content-based image retrieval algorithms, comparing with different top k values and different bits of hash code, the disparity between the original mAP and the mAP of the recessive images indicates the defense advantages against

TABLE I
TRANSFERABILITY IN DBH ON CIFAR-10

Top K returned samples	12bits	24bits	32bits	48bits
Top 200	0.761 → 0.316	0.829 → 0.379	0.836 → 0.429	0.842 → 0.429
Top 400	0.763 → 0.314	0.829 → 0.376	0.837 → 0.429	0.840 → 0.425
Top 600	0.766 → 0.314	0.830 → 0.376	0.837 → 0.430	0.840 → 0.423
Top 800	0.766 → 0.314	0.830 → 0.375	0.838 → 0.431	0.840 → 0.423
Top 1000	0.766 → 0.313	0.831 → 0.375	0.839 → 0.431	0.840 → 0.422

TABLE II
TRANSFERABILITY IN DPSH ON CIFAR-10

Top K returned samples	12bits	24bits	32bits	48bits
Top 200	0.745 → 0.438	0.770 → 0.446	0.765 → 0.461	0.780 → 0.229
Top 400	0.750 → 0.437	0.781 → 0.446	0.766 → 0.460	0.781 → 0.225
Top 600	0.750 → 0.437	0.771 → 0.445	0.770 → 0.458	0.782 → 0.221
Top 800	0.751 → 0.434	0.772 → 0.442	0.774 → 0.458	0.783 → 0.225
Top 1000	0.751 → 0.436	0.772 → 0.442	0.776 → 0.457	0.783 → 0.219

image retrieval. For instance, in terms of DSH, the mAP drops from 0.639 to 0.178, with Top 600 retrieval under 24 bits length hash. Similarly, in terms of DBH, the mAP drops from 0.840 to 0.047, with the Top 1000 retrieval under 48 bits length hash.

In both cases, the recessive images generated by our algorithm have significant impact on the target image retrieval systems. To verify the effectiveness of our algorithm in category diversity, we test the DBH with google-landmarks dataset shown in Fig.6 with 70.02% drop of average percentage of mAP impact on the target image retrieval systems.

To show the advantage of the proposed adversarial perturbation generation, we compare the reconstruction $L1$ loss of the generated recessive images and original images by our approach and iFGSM [16], the comparison results are indicated in Fig. 7 by different top k values. According to the results, in the case of the same mAP value, the reconstruction loss of iFGSM is much larger than ours, when the mAP value goes down to nearly zero, the two approaches tend to be approximately consistent.

Fig.8 shows the robustness of the proposed algorithm, we perform two kinds of input transformation defensive mechanisms: (i) JPEG compression and (ii) Bit-depth reduction on DBH and DSH. The result shows that Bit-depth reduction method has negligible effect on our approach in terms of mAP, while the JPEG compression method completely fails by making the corresponding mAP even worse. Taking DSH image retrieval algorithm as an example, with the Top 1000 under 24 bits of hash length, the original mAP is 0.639. However, our proposed iamge perturbation algorithm causes the mAP felling into 0.178. While after JPEG compression, it goes even worse to 0.108, but Bit-depth reduction method leads to a slightly increase to 0.182. The experimental results on validating transferability are shown in Table I and Table II, we construct the recessive images by DSH, and test the impact of those images to DBH and DPSH, which are different architectures training for the same task on CIFAR-10. The values in the two tables stand for the changes from the original mAP to the mAP with recessive images with regard to different top k values and hash code lengths. These results show that our proposed image perturbation algorithm can generate heterogeneous perturbations, which significantly increase the transferability when dealing with other unknown structures and/or properties content-based image searching systems.

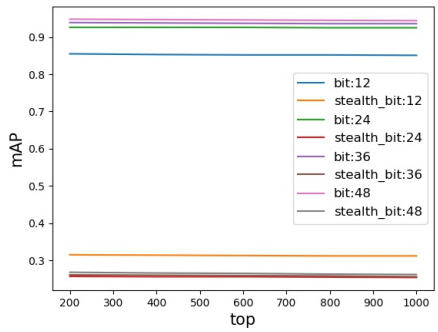
V. RELATED WORK

While enjoying the fun of online social networking applications, huge amount of users are stricken with various security and privacy exposures problems due to contents sharing. In general, there are two kinds of mainstream solutions to address privacy issues in online social networking, image contents modification and image privacy preference setting.

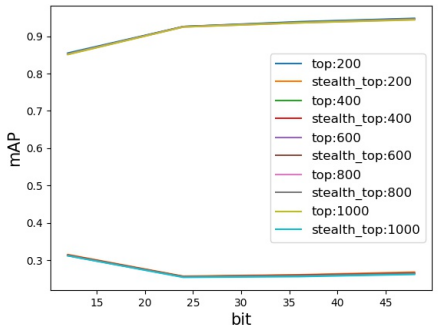
Image contents modification limits the access to image contents, this can be done by image encryption and pixel substitution. Image encryption provides reversible ways to protect the privacy by encrypting the pixels of the original region of interest (RoI). Sun *et al.* [4] proposed a DCT-domain image encryption/decryption framework for image-sharing over Facebook. The authors in [23] proposed ASePPI method to protect the privacy in the H.264/AVC stream against de-anonymization attacks by encrypting RoI, this kind of attack aims at targeting the restoration of the original image and the re-identification of people in the video.

Pixel substitution is an irreversible way to protect image privacy, which may harm the quality of images to anonymize people, including blurring, pixelation, and other technologies. Ilia *et al.* [24] studied serval image transformation techniques to evade face detection on Facebook, they concluded that many of these evasion techniques made the images worse to humans, which went against the nature of social networking. In their work, they changed the granularity of personal face identity information from the photo level to the access control user, when another user tries to access the photo, the system determines which face the user does not have permission to view and render the photo with the blurred photo.

To overcome the problem of losing image quality in obfuscating images, Orekondy *et al.* [3] proposed an automated way

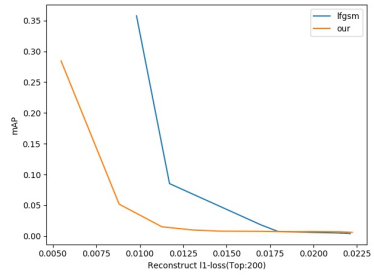


(a) mAP with different bits

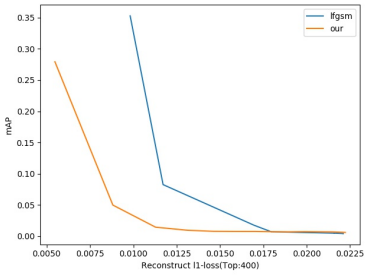


(b) mAP with different top k

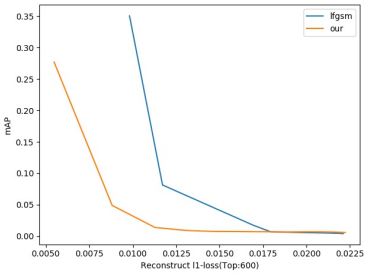
Fig. 6. Comparison of retrieval performance of DBH on Google-Landmarks Dataset.



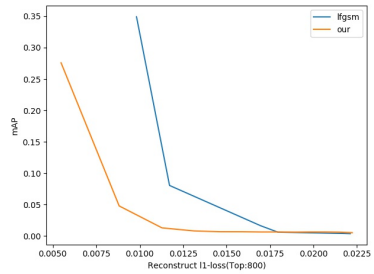
(a) Reconstruction loss of top 200



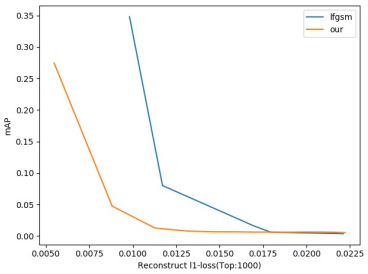
(b) Reconstruction loss of top 400



(c) Reconstruction loss of top 600



(d) Reconstruction loss of top 800



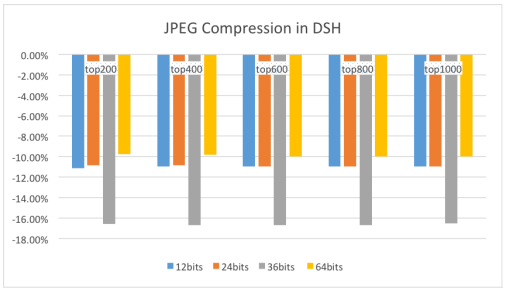
(e) Reconstruction loss of top 1000

Fig. 7. Comparing the proposed algorithm with iFGSM.

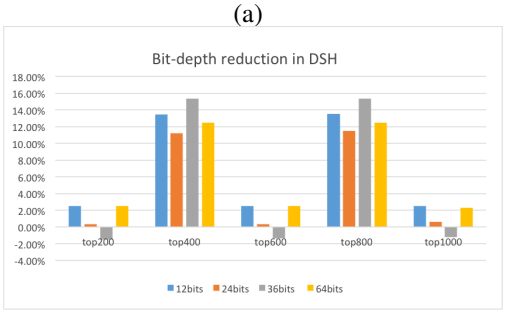
to obfuscate only the pixels of RoI. Wang *et al.* [25] discussed the human-hurting irreversible ways against automatic face detection, including gaussian noise, lines through eyes, darkening the image, and leopard spots, which all caused a large distortion of the picture. Furthermore, they found that even these privacy-preserving to some extend hurt human detection performance, the Facebook detection couldn't be avoided.

However, as shown in [26], the authors quantified the privacy implications by analyzing how well people are recongnisable in the image. They stated that with few tagged messages, the privacy protected social media under adversarial conditions like gaussian blur, black fill-in and white fill-in, can still be reconginized with higher accuracy than chance level across different events, such as different day, clothes, poses and point of view. In addition, McPherson *et al.* [27] adopted artificial neural networks to recover hidden information from images processed by three kinds of obfuscation approaches, including mosaicing (pixelation), blurring and P3 [28]. They found that with the modern image recognition method based on deep learning, the faces, objects, and handwritten digits can still be identified. It is worth mentioning that this can be done even without the knowledge of specific relevant features of the confusion image or the degree of association of the remaining information with the hidden information in advance.

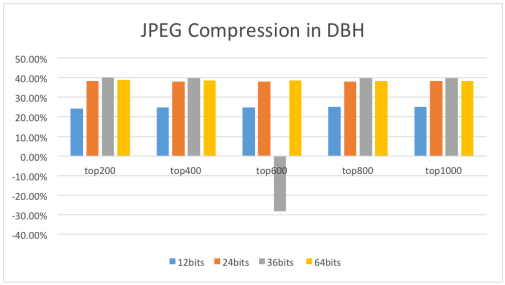
To be more deliberate about privacy-preserving media, another two countermeasures were proposed by [29], [30]. As reported in [30], covering the face and drawing a specific pattern on the face could effectively prevent the detection of the face image. However, this kind of method would hide face-to-face communication. To overcome the problem, Yamada *et al.* [29] stated that wearing a device similar in appearance to eyeglasses on the face could hide the face in the captured images, by transmitting near-infrared signals picked up by the camera image sensors beyond people's sight. Yet



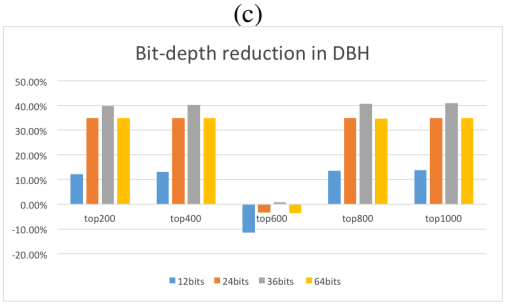
(a)



(b)



(c)



(d)

Fig. 8. The robustness of the proposed algorithm.

theses two active approaches made the users more iconic than the traditional image processing approaches, furthermore, the initiative of the participants on the faces were needed for the funny glasses or makeup.

Image privacy preference setting aims at privacy detection, it provides users the opportunity to understand privacy level of their images and decide to post it online or not. The authors in [31] and [32] proposed mechanisms to automatically generate privacy setting for uploaded images based on users' social features. Zhang *et al.* [33] designed a portrait privacy preserving approach based on a portrait graph matching scheme and an encryption-free vector distance computation method. This work provided automatical portrait erasing during photo taking and sharing. To prevent users from publishing potentially sensitive visual content, Zerr *et al.* [34] trained a SVM classifier to detect and identify private images. The authors in [35] and [36] explored deep visual features to improve image privacy prediction accuracy. Zhang and Yan [37] designed a mechanism to evaluate privacy of social images based on differential privacy, in their work, they limited image privacy to the scopes of face and car plate number detections.

Yet, our work is very different from the aforementioned works. Most of these approaches only consider about the privacy leakage problems, but offer no guarantee in the scenario where users still need to enjoy content sharing in OSN. As shown in [38], the privacy-enhancing technologies should provide a trade-off between the privacy and a good viewing experience. In our work, we proposed to adopt adversarial examples as the potential solution. With adopting a small adversarial perturbation, neural networks can be deceived but human can't figure out the difference. In addition to using adversarial examples to attack Deep Learning algorithms, many scholars begin to explore how to use this kind of perturbation as a defense strategy. For example, in the work of Osadchy *et al.* [39], they introduced a secure CAPTCHA scheme based on adversarial noise that deceived Deep Learning tools.

VI. Conclusion

In this work, we proposed a new concept: recessive social networking, where the recessive images can bypass reverse image search by deceiving the corresponding machine learning process. Meanwhile, those recessive images are indistinguishable from the original ones from human eyes; therefore, they preserve all the normal social networking functionalities among humans. We initiated the study of such a new primitive and proposed an adversarial perturbation based scheme. This is the first time that adversarial examples are used as a defensive mechanism to protect image privacy against content-based image searching algorithms in the context of social networking. Furthermore, we also demonstrated the effectiveness, transferability, and robustness of the proposed scheme via extensive experiments. Finally, we emphasize that this work provide important insights to privacy preserving social networking, and it opens a door for constructing a new class of efficient and privacy social networking schemes to help users enjoy contents sharing. Nevertheless, this line of work is far from being completed. In the future, we are looking for more robust and universally effective adversarial perturbation generation algorithms.

REFERENCES

[1] Kaitai Liang, Joseph K. Liu, Rongxing Lu, and Duncan S. Wong. Privacy concerns for photo sharing in online social networks. *IEEE Internet Computing*, 19(2):58–63, 2015.

[2] Mapping social media with facial recognition: A new tool for penetration testers and red teamers. https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/mapping-social-media-with-facial-recognition-a-new-tool-for-penetration-testers-and-red-teamers. Accessed: 2019-3-15.

[3] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *CVPR 2018*, pages 8466–8475, 2018.

[4] Weiwei Sun, Jiantao Zhou, Shuyuan Zhu, and Yuan Yan Tang. Robust privacy-preserving image sharing over online social networks (osns). *TOMCCAP*, 14(1):14:1–14:22, 2018.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[6] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. *CoRR*, abs/1711.09856, 2017.

[7] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV 2017*, pages 1378–1387, 2017.

[8] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *MLDM 2017*, pages 262–275, 2017.

[9] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE S&P*, pages 1–7, 2018.

[10] Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. The space of transferable adversarial examples. *CoRR*, abs/1704.03453, 2017.

[11] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pages 506–519, 2017.

[12] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR*, abs/1602.02697, 2016.

[13] Ion Stoica, Dawn Song, Raluca Ada Popa, David A. Patterson, Michael W. Mahoney, Randy H. Katz, Anthony D. Joseph, Michael I. Jordan, Joseph M. Hellerstein, Joseph E. Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, and Pieter Abbeel. A berkeley view of systems challenges for AI. *CoRR*, abs/1712.05855, 2017.

[14] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *CoRR*, abs/1801.02612, 2018.

[15] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX Workshop on Offensive Technologies, WOOT 2017, Vancouver, BC, Canada, August 14-15, 2017.*, 2017.

[16] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016.

[17] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2064–2072, 2016.

[18] Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. Deep learning of binary hash codes for fast image retrieval. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7-12, 2015*, pages 27–35, 2015.

[19] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1711–1717, 2016.

[20] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[21] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3476–3485, 2017.

[22] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. *CoRR*, abs/1711.00117, 2017.

[23] Natacha Ruchaud and Jean-Luc Dugelay. Aseppi: Robust privacy protection against de-anonymization attacks. In *CVPR 2017*, pages 1352–1359, 2017.

[24] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *SIGSAC 2015*, pages 781–792, 2015.

[25] Michael J. Wilber, Vitaly Shmatikov, and Serge J. Belongie. Can we still avoid automatic face detection? In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–9, 2016.

[26] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 19–35, 2016.

[27] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *CoRR*, abs/1609.00408, 2016.

[28] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: toward privacy-preserving photo sharing. In *NSDI 2013*, pages 515–528, 2013.

[29] Takayuki Yamada, Seiichi Gohshi, and Isao Echizen. Privacy visor: Method for preventing face image detection by using differences in human and device sensitivity. In *CMS 2013*, pages 152–161, 2013.

[30] Camouflage from face detection. https://cvdazzle.com. Accessed: 2019-3-15.

[31] Anna Cinzia Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE Trans. Knowl. Data Eng.*, 27(1):193–206, 2015.

[32] Zhenzhong Kuang, Zongmin Li, Dan Lin, and Jianping Fan. Automatic privacy prediction to accelerate social image sharing. In *BigMM 2017*, pages 197–200, 2017.

[33] Lan Zhang, Kebin Liu, Xiang-Yang Li, Cihang Liu, Xuan Ding, and Yunhao Liu. Privacy-friendly photo capturing and sharing system. In *UbiComp 2016*, pages 524–534, 2016.

[34] Sergej Zerr, Stefan Siersdorfer, and Jonathon S. Hare. Picalert!: a system for privacy-aware image classification and retrieval. In *CIKM 2012*, pages 2710–2712, 2012.

[35] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *AAAI 2016*, pages 1317–1323, 2016.

[36] Ashwini Kishore Tonge and Cornelia Caragea. Image privacy prediction using deep features. In *AAAI 2016*, pages 4266–4267, 2016.

[37] Xue Zhang and Wei Qi Yan. Comparative evaluations of privacy on digital images. In *AVSS 2018*, pages 1–6, 2018.

[38] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *CVPR 2017*, pages 1343–1351, 2017.

[39] Margarita Osadchy, Julio Hernandez-Castro, Stuart J. Gibson, Orr Dunkelman, and Daniel Pérez-Cabo. No bot expects the deepcaptcha! introducing immutable adversarial examples, with applications to CAPTCHA generation. *IEEE Trans. Information Forensics and Security*, 12(11):2640–2653, 2017.