Semantic-Aware Attack and Defense on Deep Hashing Networks for Remote-Sensing Image Retrieval

Yansheng Li, *Senior Member, IEEE*, Mengze Hao, Rongjie Liu, Zhichao Zhang, Hu Zhu, *Member, IEEE*, and Yongjun Zhang, *Member, IEEE*

***Abstract*— Deep hashing networks have been successful in retrieving interesting images from massive remote-sensing images. There is no doubt that security and reliability are critical in remote-sensing image retrieval (RSIR). Recent studies about natural image retrieval have shown the vulnerability of deep hashing networks to adversarial examples, but there are no existing research studies about the attack and defense of deep hashing networks in RSIR. Due to the large intraclass difference and high interclass similarity of remote-sensing images, the attack and defense methods on deep hashing networks for natural images cannot be directly applied to the remote-sensing images. Different from the widely adopted instance-aware hash codes that often present the suboptimum performance of the attack and defense on deep hashing networks, this article recommends the usage of semantic-aware hash codes, which take into account multiple samples in the given semantic categories, in both attack and defense. To pursue the strongest attack on RSIR, a novel semantic-aware attack with weights via multiple random initialization (RWC) is proposed. To alleviate the retrieval degradation caused by adversarial attacks, a new adversarial training defense method on deep hashing networks with the adversarial semantic-aware consistency constraint (ACN) is proposed. Extensive experiments on three typical open remote-sensing image datasets (i.e., UCM, AID, and NWPU-RESISC45) show that the proposed attack and defense methods on various deep hashing networks achieve better performance compared with the state-of-the-art methods. The source code will be made publicly available along with this article.**

***Index Terms*— Adversarial examples, deep hashing network, remote-sensing image retrieval (RSIR), semantic-aware attack and defense.**

I. Introduction

With the rapid development of remote-sensing science and technology, the amount of data acquired by earth observation remote-sensing sensors increases sharply. Human beings have entered the stage of remote-sensing big data (RSBD) [1]. Remote-sensing image retrieval (RSIR) as a means of discovering images is an important tool for mining the value of RSBD. Given one query image, RSIR refers to searching for interested images from a remote-sensing image dataset [2]. With the explosive growth of remote-sensing images, traditional RSIR methods are unable to meet the requirements of the current image data scale for efficiency and accuracy [3], [4], [5], [6], [7], [8], [9]. To improve retrieval efficiency and accuracy, the latest methods usually use deep neural networks [10], [11], [12], [13]. Due to the strong low-dimensional representation ability of deep learning, deep hashing methods have achieved great success in large-scale RSIR [14], [15], [16], [17], [18], [19], [20].

However, recent studies have shown the vulnerability of deep hashing networks [15], [21], [22], [23], [24]. By making minor modifications to the original input examples or adding elaborate malicious disturbances, the network often gives wrong output results, and the modified examples are called adversarial examples [25]. Compared with other networks, learning adversarial examples on deep hashing networks is much more difficult. The particularity brought by the symbolization of the hashing network is the most obvious difficulty. At present, the research on adversarial examples of deep hashing networks, including adversarial attack and defense, is mostly based on natural images. Some studies have shown that the use of adversarial examples can increase the generalization ability of the model. Especially on supervised small datasets [26], [27] and on semisupervised large datasets [28], [29]. However, on large datasets with supervised settings, adversarial examples usually cause a decrease in the accuracy of the results [30], [31]. Therefore, to reduce the threat of adversarial examples to the deep hashing networks and improve the model's robustness, more and more adversarial defense methods have been proposed [14], [28], [32], [33], [34], [35], [36], [37], [38]. Adversarial examples are regarded as data augmentation and adversarial examples participate in the training stage to improve model performance.

There is a serious lack of research about attacks and defenses on deep hashing networks in large-scale RSIR. Compared with the natural image field, the tasks in the remote-sensing image field have their own particularities. Remote-sensing images are more complex. They have the characteristics of a large intraclass difference and high

Manuscript received 11 July 2023; revised 20 October 2023; accepted 21 November 2023. Date of publication 24 November 2023; date of current version 6 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 41971284, in part by the State Key Program of the National Natural Science Foundation of China under Grant 42030102, in part by the Fundamental Research Funds for the Central Universities under Grant 2042022kf1201, and in part by the Ant Group. *(Corresponding authors: Zhichao Zhang; Yongjun Zhang.)*

Yansheng Li, Mengze Hao, Rongjie Liu, Zhichao Zhang, and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; haomengze@whu.edu.cn; 1397608894@qq.com; zhichao@whu.edu.cn; zhangyj@whu.edu.cn).

Hu Zhu is with the College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: zhuhu@njupt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3336796

interclass similarity. Remote-sensing scenes belonging to the same semantic category may have significant appearance variations due to different imaging conditions and spatial resolutions. The existence of adversarial examples poses a great threat in scenarios requiring high-security intensity. For example, in the military battlefield, the enemy can conduct adversarial attacks on the query images that need to extract more similar information from the remote-sensing image database through network attacks, so that we will get images that do not meet the expectations and interfere with the judgment. Therefore, to alleviate the vulnerability of deep hashing networks in the massive RSIR scene, attention should be paid to adversarial defense methods. However, there are no studies in RSIR on deep hashing network attack and defense.

In particular, in many existing attacks and defenses on deep hashing networks, the methods are completed using instance-aware hash codes. The instance-aware attack only considers the relationship between the adversarial sample and its corresponding original sample, which may lead to the problem that the adversarial sample still maintains some similarity with other samples of that category after the attack. The instance-aware defense improves network robustness by maintaining the similarity between the adversarial sample and the original clean samples without considering the similarity between the adversarial sample and its corresponding clean sample category. Instead, the essence of defense on deep hashing networks is to make the output hash codes of the adversarial sample after the deep hashing network consistent with the original sample hash codes in that category. Therefore, instance-aware hash codes lead to poor attack and defense results on deep hashing networks.

To solve the above problems, adversarial examples on deep hashing networks for large-scale RSIR are studied from attack and defense aspects in this article. To the best of our knowledge, it is the first time to propose an adversarial attack and defense on deep hashing networks in RSIR. It reveals the significance of the resistibility of networks when solving large-scale RSIR tasks. We propose semantic-aware attacks and defense. The semantic information presented by multiple samples makes the attack and defense more effective. To reveal the vulnerability of deep hashing networks in RSIR, a gradient-based attack method that is a novel semantic-aware attack with weights via multiple random initialization (RWC) is proposed. The weight vector is used to optimize the resource allocation during the attack, and similar to commonly used in other attacks the global optimum is found by multiple random initializations which is the first used for deep hashing network attacks. The attack method uses semantic-aware hash codes rather than instance-aware ones. To alleviate the vulnerability and maintain a certain retrieval accuracy in the face of adversarial examples, we study adversarial training methods on deep hashing networks and propose an adversarial training defense method for deep hashing networks with adversarial semantic-aware consistency constraints (ACNs). By introducing adversarial examples in the training stage, the similarity relationship between adversarial examples and original images and the consistency relationship between adversarial examples and original category examples of network output is maintained for defense. This adversarial training improves the robustness of the network. Experiments are carried out on three RSIR datasets UCM, AID, and NWPU-RESISC45 to verify the superiority and effectiveness of the proposed attack and defense methods on deep hashing networks. The main contributions of this article are summarized as follows.

1) We propose the semantic-aware attack method (i.e., RWC) on deep hashing networks, which first adopts multiple random initializations on deep hashing network attack to ease gradient descent into local optimization, adds the weight vector of hash bits to make the attack focus on the hash bit that has a large impact on the result and uses semantic-aware hash codes that take into account the semantic information presented by multiple samples in the class.
2) To alleviate the vulnerability of deep hashing networks, we propose a new adversarial defense method (i.e., ACN) that adds adversarial examples in the training stage and uses the proposed adversarial loss term, which includes clean similarity loss, adversarial similarity loss, and adversarial semantic-aware consistency, which keeps the semantic information between multiple samples in the class consistent with that between the adversarial samples.

The rest of this article is organized as follows. Section II describes in detail the adversarial attacks and defenses in existing deep hashing networks. Section III introduces our proposed methods, including RWC and ACN. Section IV introduces the information of the datasets used and experimental results. Section V summarizes the article.

II. Related Work

A. Deep Hashing Networks for RSIR

Hashing-based methods aim to construct binary codes for each sample in a database, such that similar samples have close codes [39]. Because hash code is very efficient in binary computing and storage, it is widely used to accelerate artificial neural network (ANN) retrieval [40]. Most existing deep hashing methods can be divided into two categories: 1) unsupervised hashing, which learns deep features by preserving the consistency between inputs and outputs without using any semantic labels [3], [4] and 2) supervised hashing, which uses semantic labels or pairwise similarities to supervise feature learning.

Compared to unsupervised hashing, supervised hashing can usually achieve more promising results and thus attract more attention. The first supervised deep hashing is called convolutional neural network hashing (CNNH), which adopts two steps for hashing: hash code learning and hash function learning [41]. Recent work showed that learning similarity-preserving binary code in an end-to-end manner can improve retrieval performance. Deep pairwise-supervised hashing (DPSH) proposes a pairwise loss function to map similar data pairs to similar hash codes and dissimilar data pairs to dissimilar hash codes [25]. Moreover, the quantization error between real-value outputs and binary codes is also minimized. HashNet learns exact binary hash codes via a continuation

method with convergence guarantees [42]. Recently, a new supervised hashing learning method was proposed, which expresses hashing learning in the form of meta-learning [20].

In the field of massive RSIR, deep hashing neural networks (DHNNs) introduced deep hashing into RSIR for the first time and realized end-to-end image retrieval by combining a deep feature learning module and a hash function learning module [9]. Feature and hash learning (FAH) is further developed on the basis of DHNNs. It contains a feature learning module and an adversarial hash learning module. The feature learning module extracts multiscale features of remote-sensing images by feature aggregation, tries to emphasize the multiscale by attention branching, and ultimately guarantees to extract characteristics of dense mapping to discrete hash code, its loss function, in addition to containing the basic similarity constraint loss items in pairs and quantifying losses. It also includes the loss item of semantic label classification and the loss item of maintaining hash code distribution balance. FAH has further improved the retrieval accuracy of remote-sensing images based on DHNNs [14].

In RSIR, deep hashing networks improve retrieval speed and accuracy. It makes full use of the strong feature extraction ability of deep neural networks and the high retrieval and storage efficiency of the hash method. However, in recent years, many studies have demonstrated the vulnerability of deep hashing networks in the face of adversarial example attacks. The attack and defense on deep hashing networks for RSIR have not been studied yet, so incremental analysis is needed to fill the gap.

B. Adversarial Attacks on Deep Hashing Networks

Adversarial examples are usually crafted by adding small, visually imperceptible perturbations to the original images that can confuse the targeted neural networks and misclassify them. The generation of adversarial examples has been extensively studied, and they can be roughly divided into untargeted attacks and targeted attacks. Targeted attacks trick the model to output specific classes, but untargeted attacks trick the model to output any unwanted classes. Since Szegedy et al. [1] discovered the existence of adversarial examples, various attack methods in image classification have been proposed to fool the well-behaved dynamic neural network (DNN). For example, the fast gradient sign method (FGSM), one of the most popular and efficient attack algorithms, aims to maximize the loss of the targeted model along the gradient direction in a single step to learn adversarial examples [43]. Iterative FGSM (I-FGSM) [44] and projected gradient decent (PGD) [45] is a multistep variant of FGSM that updates adversarial perturbation in an iterative learning strategy to obtain better performance. Xu et al. proposed a new black-box adversarial attack method Mixup-Attack that attacks the shallow features of a given proxy model to find common vulnerabilities among different networks, to generate transferable adversarial examples [46].

Adversarial examples for DNN-related tasks have been studied, such as classification, semantic segmentation, natural language processing, medical prediction, and so on. Recent works on deep hashing networks have also confirmed the vulnerability of DNNs to adversarial examples. The generation of adversarial examples for deep hashing networks can be divided into untargeted attacks and targeted attacks. For untargeted attacks, a deep hashing model can preserve the semantic information well, which means that semantically similar pairs will have similar hash codes, while semantically dissimilar pairs will have dissimilar hash codes. Hash adversary generation (HAG) modifies the adversarial examples to make the hash codes of the adversarial examples dissimilar to the original examples [30]. Smart deep hashing attack (SDHA) employed a dimension-wise Hamming distance surrogate function to improve the effectiveness of attack [47]. Recently, Yuan et al. [48] proposed a theoretically guaranteed measure of discriminant learning called SAAT that can obtain representative pillar codes. For targeted attacks, Bai et al. [49] proposed two targeted hashing attack methods called point-to-point (P2P) and deep hashing targeted attack (DHTA), which is to make the hash code of the adversarial example similar to the images of targeted images. Compared with them, DHTA is more effective. Because DHTA determines the category hash code from multiple examples in a given category, P2P randomly selects only one example. Wang et al. [32] proposed a neural network called PrototypeNet to generate hash codes for specified category labels and achieved the optimal targeted attack effect. Then Wang et al. [50] proposed a prototype supervised adversarial network (ProS-GAN), which formulates a generative architecture for efficient and effective targeted attacks.

The above attack methods mainly focus on adversarial examples of natural images, and there has been little research in remote sensing in recent years. To comprehensively evaluate the impact of adversarial examples on the remote-sensing image scene classification, Chen et al. [51] tested eight state-of-the-art classification DNNs on six remote-sensing image benchmarks, which include both optical and synthetic-aperture radar images of different spectral and spatial resolutions, and the experimental result shows that the fooling of attacks is over 98%. Studies have shown that DNNs on remote-sensing images are more vulnerable to adversarial examples, and adversarial examples on deep hashing networks for RSIR need to be studied [52].

At present, the hash attack method based on gradient descent is commonly used, but this method does not consider three problems: the particularity of hash network symbolization, the local optimal problem caused by gradient attack, and the hash code representation problem caused by using instance-level hash code and P2P attack.

C. Adversarial Defense on Deep Hashing Networks

Since the discovery of adversarial examples for deep neural networks, increasing efforts have been made to build systems that are robust against adversarial examples. Currently, the defenses against the adversarial examples are being developed in three main directions: 1) using modified training during learning or modified input during testing; 2) modifying networks; and 3) using external models as network add-ons when solving unseen examples.

The first direction does not directly deal with the learning models. Goodfellow et al. [43] adopted adversarial training to augment the training data of the classifier with adversarial examples. The second direction is modifying networks, for example, by adding more layers or subnetworks, changing the loss or activation function, and so on. Papernot et al. [33] used distillation techniques to train networks, which can greatly reduce the magnitude of gradients used for adversarial example creation. The last direction is using external models as network add-ons to detect and remove malicious perturbations before prediction. Carlini and Wagner [34] invested in ten recent proposals for adversarial example detection and showed that all of these defense methods are ineffective at dealing with newly constructed loss functions.

There have been studies on remote sensing that show that DNNs on remote-sensing images are more vulnerable to adversarial examples, so the defense strategy on remote-sensing images should be thoroughly investigated [52]. Wang et al. [32] proposed a training framework designed to train the classifier by introducing the examples generated during the image reconstruction process. To achieve standardized adversarial training, Yuan et al. [48] completed adversarial training by minimizing the distance between the hash code of the adversarial example and the body code.

In RSIR, the defenses of deep hashing networks are very few, but they are very important in practical applications. For example, in the military field, when remote-sensing images are used to judge the terrain, it is necessary to be alert to the judgment error caused by the adversarial example.

III. Adversarial Deep Hashing Learning for RSIR

In RSIR, there are many tasks about military and national defense security, so it is very important to ensure the reliability of the results. Although the image retrieval method of the deep hashing network performs well in terms of retrieval speed and accuracy, it still shows its vulnerability in the face of adversarial examples. By adding small perturbations to the image, the adversarial examples are imperceptible to the human eye but can still deceive the deep neural networks, and the wrong classification results will be generated. The existence of such adversarial examples will hinder the development and practical application of RSIR. As a result, determining how to repel the attack becomes a challenge. The intuitive idea is to directly participate in model training with adversarial examples as training examples so that the model can learn this deception and make correct judgments when faced with adversarial examples again. This section presents the proposed attack and defense methods.

A. Adversarial Attack on Deep Hashing Networks

In general, a deep hashing model $E(\cdot)$ consists of a deep model $H(\cdot)$ and a sign function, where $H(\cdot)$ consists of a feature extractor followed by fully connected layers. Given an image x, the hash code of this image can be calculated as

$$b_i = E(x_i) = \text{sign}(H(x_i)) \quad \text{s.t. } b_i \in \{1, -1\}^k. \tag{1}$$

The deep hashing model will return a list of images that are organized according to the hamming distances between the hash code of the query and all images in the database. To obtain the hashing method $E(\cdot)$, most of supervised hashing methods are trained on dataset $D = \{(x_i, y_i)\}_{i=1}^N$ that contains N examples collection labeled with C classes, where x_i indicates the retrieval image, and $y_i \in [0, 1]^C$ corresponds to a label vector. The cth component of indicator vector $y_i^c = 1$ means that the example x_i belongs to class c. Besides, the sign$(\cdot)$ function is approximated and replaced by the tanh$(\cdot)$ function during the training process in deep hashing to alleviate the gradient vanishing problem. The output of the image through deep hashing networks is

$$f_i = F(x_i) = \tanh(H(x_i)). \tag{2}$$

Given a benign query x with label y_t, the image of attacks in retrieval is to generate an adversarial example x', which would cause the targeted model to retrieve objects with different labels. So, we can make the output of adversarial example x' semantically irrelevant to the original label y_t. The objective can be achieved through maximization of the distance between the hash code of the adversarial example x_i' and the retrieval objects with the target label y_t

$$\max_{x'} d_H\left(F\left(x'\right), E(x)\right) \tag{3}$$

where x' is the adversarial example of x and $d_H(\cdot)$ is the hamming distance metric.

Since RSIR does not need a clear target when a remote-sensing image is attacked, the semantic-aware hash code can be obtained by definition 1, to realize an untargeted attack. The optimality of semantic-aware hash codes is verified in Theorem 1.

Definition 1: Given a set of points $\mathcal{A} \in \{-1, +1\}^K$, the semantic-aware hash code of the set is defined as follows:

$$h_a = \text{sgn}\left(\frac{1}{|\mathcal{A}|}\sum_{h\in\mathcal{A}} h\right) = \text{sgn}\left(\sum_{h\in\mathcal{A}} h\right). \tag{4}$$

Theorem 1: Given a set of points $\mathcal{A}$ in $\{-1, +1\}^K$, the problem of maximizing the Hamming distance between a point and a set can be translated to the point and h_a

$$\begin{aligned} h' &= \underset{h\in\mathcal{A}}{\text{argmax}} \sum_{h_i\in\mathcal{A}} d_H(h, h_i) \\ &= \underset{h\in\mathcal{A}}{\text{argmax}}\, d_H\left(h, \sum_{h_i\in\mathcal{A}} h_i\right) \\ &= \underset{h\in\mathcal{A}}{\text{argmax}}\, d_H(h, h_a). \end{aligned} \tag{5}$$

Due to the representative property of semantic-aware code for the set of retrieval images with the target label y_t, we can choose the semantic-aware code (definition 1) as the hash code to optimize

$$\max_{x'} d_H\left(F\left(x'\right), h_a\right) \quad \text{s.t. } \left\|x' - x\right\|_\infty \le \epsilon. \tag{6}$$

Given a pair of binary codes h_i and h_j, since $d_H(h_i, h_j) = (1/2)(K - h_i^T h_j)$, we can equivalently replace Hamming distance with an inner product in the objective function. The optimization objective is as follows:

$$\min_{x'} L_a = h_a^T F\left(x'\right) \quad \text{s.t. } \left\|x' - x\right\|_\infty \le \epsilon. \tag{7}$$

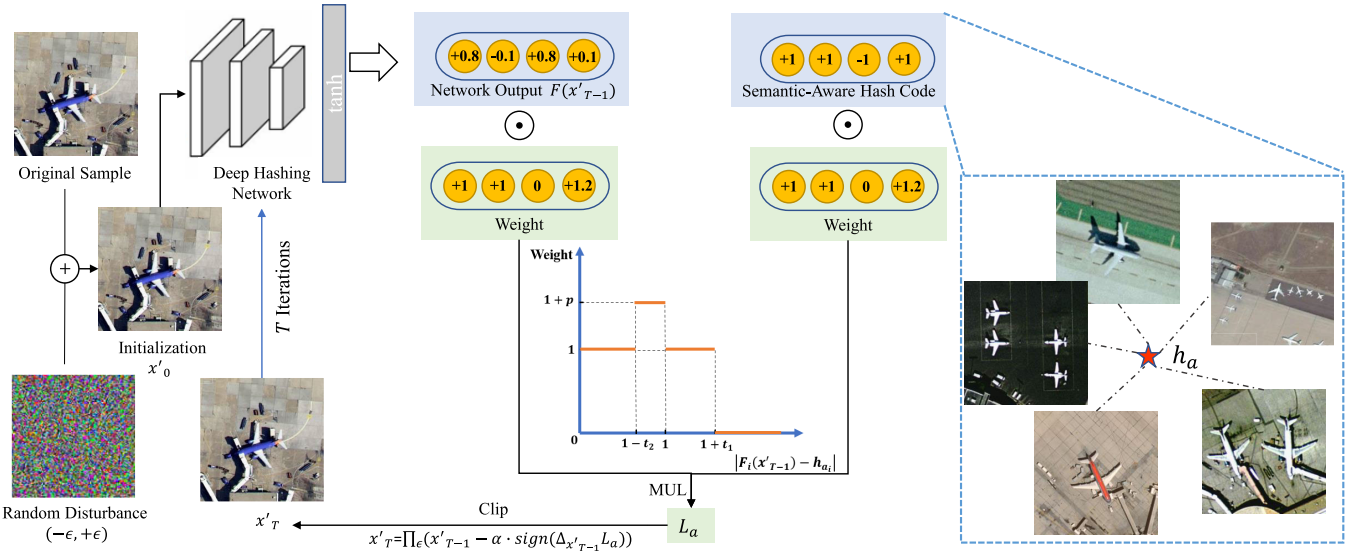

Fig. 1. Flowchart of the proposed RWC attack method on deep hashing networks. The deep hashing network has been pretrained with original examples. The adversarial example is generated by adding perturbation to the original example. In the use of it for image retrieval to get the wrong classification results.

However, similar to [30], since continuous values $F(x')$ will be transferred to binary codes via a sign function, if $F(x')$ and h_a already have different signs for some dimensions with a large margin, continuing to optimize in these dimensions will not change the final results; if $F(x')$ and h_a will have different signs for some dimension with a small margin, enlarging optimize in these dimensions will significantly enhance the final results. To solve the problem, we add the weight of the hash bits of the adversarial example. The objective can be rewritten as

$$\begin{aligned} &\min_{x'} \; L_a = h_a^{'T} F'(x') = (\omega \odot h_a^T) * (\omega \odot F(x')) \\ &\text{s.t. } \|x' - x\|_\infty \le \epsilon \end{aligned} \tag{8}$$

where h_a denotes the semantic-aware hash code. The operation $\odot$ represents the dot product and ω denotes the weight of hash bits for the adversarial example. We can obtain ω by

$$\omega_i = \begin{cases} 0, & \text{if } |F_i(x') - h_{a_i}| > 1 + t_1 \\ 1 + p, & \text{if } 1 > |F_i(x') - h_{a_i}| > 1 - t_2 \\ 1, & \text{otherwise} \end{cases} \tag{9}$$

where ω_i denotes the ith element of the weight vector and h_{a_i} denotes the ith element of the semantic-aware code h_a. t_1 and t_2 are thresholds to control the margin. When there is a large gap in the hash code between the adversarial example and the category, the disturbance can be stopped. Otherwise, when the hash codes of the adversarial example and the boundary are close, the disturbance weight is decreased by p. In detail, we optimize x' with gradient iterative of T iterations as follows:

$$\begin{aligned} x'_T &= \prod_\epsilon (x'_{T-1} - \alpha \cdot \text{sign}(\Delta_{x'_{T-1}} L_a)) \\ x'_0 &= (x - \epsilon, x + \epsilon) \end{aligned} \tag{10}$$

where α is the step size of each iteration, $\prod_\epsilon$ clips x' to ϵ-neighbor of x, and x'_0 is the random initialization. The adversarial examples after each iteration are truncated to ensure that the disturbance range is always within the limit range, as shown below

$$x' = \min\{255, x + \epsilon, \max(0, x - \epsilon, x')\}. \tag{11}$$

In the end, following [30], to avoid the gradient vanishing problem, we optimize the output of deep hashing networks as follows:

$$F(x) = \tanh(\beta H(x)) \tag{12}$$

where β is first set at 0.1 and then gradually enlarged until eventually becoming 1.

Based on the above description, the RWC method is proposed. The detailed steps of RWC are shown in Algorithm 1 and Fig. 1.

B. Adversarial Defense on Deep Hashing Networks

Much discussion will revolve around an optimization view of adversarial hashing robustness. Consider a standard hash retrieval task with an underlying data D. Assume a suitable loss function $L(\theta, x_i, x_j, s_{ij})$, where s_{ij} is the similarity information of the original example x_i and x_j and $\theta \in R^p$ is the set of model parameters. The goal is to find model parameters θ that minimize the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\theta, x_i, x_j, s_{ij})]$ to preserving the similarity relationship of original clean examples.

Although empirical risk minimization (ERM) can yield excellent performance on the original clean examples, it often does not yield models that are robust to adversarially crafted examples. To reliably train models that are robust to adversarial attacks, we modify the definition of the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\theta, x_i, x_j, s_{ij})]$ by incorporating the above adversarial examples. Instead of feeding examples from the distribution D directly into the loss L, we allow the adversarial examples to train first. Accordingly, we formulate the following object:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L\left(\theta, x_i, x_j + \delta, s_{ij}\right)\right] \tag{13}$$

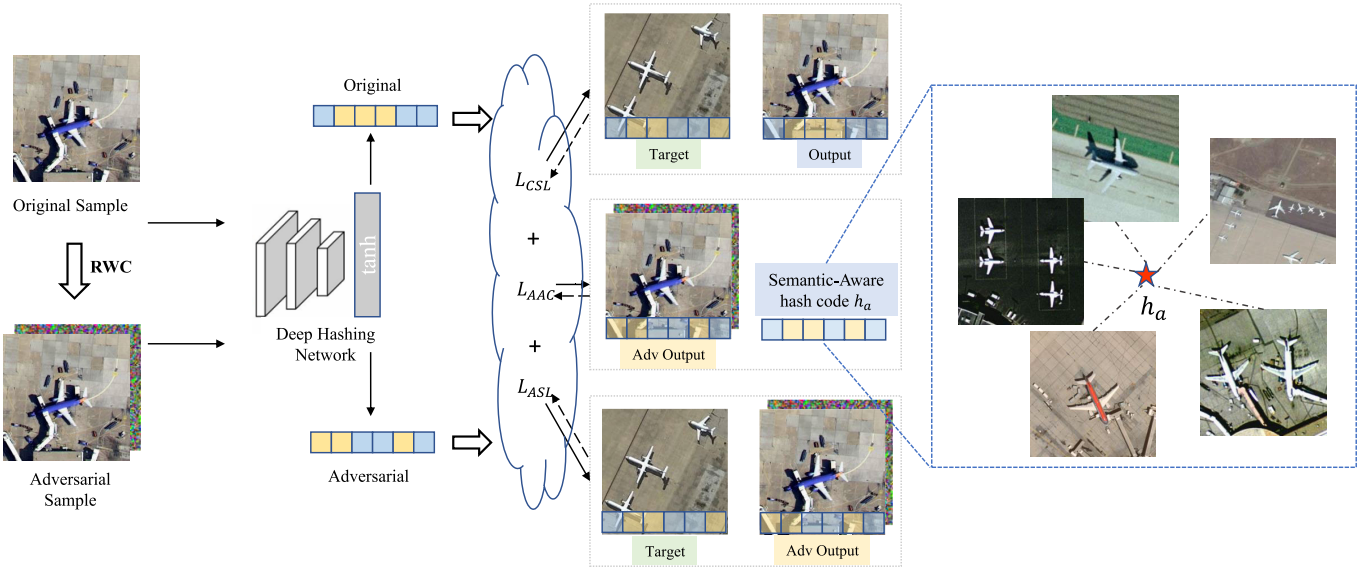

Fig. 2. Flowchart of the proposed ACN defense training on deep hashing networks. First, the original examples are processed by a deep hashing network and then the RWC attack method is used to generate adversarial examples. Then, the generated adversarial examples are used to pass through the network and the joint loss function together with the original examples. The parameters of the deep hashing network model are updated through backpropagation, to improve the network's robustness in the training process.

Algorithm 1 Semantic-Aware Attacks With Weights Using Multiple Random Initialization

Input: An original image (x, y), a deep hashing model $E(\cdot)$, the number of random initialization N, the number of iterations T, a sequence of values β_0, and the hyperparameter t_1, t_2, p.

Output: The adversarial example x'.

Compute the category-level point h_a of category y in the database image set via Eq. (4)

for $n = 0$ to N **do**

- Random initialization $x'_0 = (x - \epsilon, x + \epsilon)$
- **for** $t = 0$ to T **do**

  $\beta = \beta_T$

  The adversarial example x'_t passes through the hashing network to obtain $F(x'_t)$

  Compute weight matrix ω via Eq. (9)

  Compute x'_t via Eq. (10)

  Recompute x'_t via Eq. (11)

  end for
- **if** $d_H(F(x'), h_a) < d_H(F(x'_T), h_a)$

  $x' = x'_T$

  end if

end for

where x_i denotes the original clean example and δ denotes the crafted perturbation added to the original example x_j. The minimization problem seeks model parameters that preserve the similarity relationship between the original clean and adversarial examples. This is precisely the problem of training a robust hash network using adversarial training techniques. In practice, the models should be robust to adversarial crafted examples, while also preserving the similarity order in the original clean examples. We often train networks with a mixture of adversarial examples and clean images

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L\left(\theta, x_i, x_j, s_{ij}\right) + L\left(\theta, x_i, x_j + \delta, s_{ij}\right)\right]. \quad (14)$$

To enhance the robustness of deep hashing networks, except for the similarity relationship between the original clean examples and adversarial examples, and considering the consistency relationship between the adversarial examples and the semantic-aware code of the set with the same label proposed in definition 1, we propose an external loss function called the adversarial semantic-aware consistency (AAC) constraint, the whole loss function L_{ADV} should be

$$\begin{aligned} L_{\text{ADV}} &= L_{\text{CSL}} + L_{\text{ASL}} + L_{\text{AAC}} \\ &= L\left(\theta, x_i, x_j, s_{ij}\right) + L\left(\theta, x_i, x_j + \delta, s_{ij}\right) \\ &\quad + L\left(\theta, h_a, x_j + \delta\right) \end{aligned} \quad (15)$$

where L_{CSL} denotes the clean similarity loss, L_{ASL} denotes the adversarial loss, and $L_{\text{AAC}} = L(\theta, h_a, x_j + \delta) = \gamma D(h_a, F(x_j + \delta))$ denotes the distance constraint between the network output of an adversarial example and the semantic-aware code of the set. γ is the hyperparameter. The overall adversarial training defense on deep hashing networks with AAC constraint terms process is in Fig. 2.

C. Experiment Details

The dataset division is consistent with previous work [14]. In three datasets, 10% of the images are sampled as a query set, and the rest of the samples are used as a database. 20%, 60%, or 80% of the images for each scene category are randomly chosen from the database as a training set. In addition, we resize all images into $224 \times 224 \times 3$. Implementation details for different stages are as follows.

1) In Attack Stage: We adopt Alexnet pretrained on ImageNet as the backbone network, which is simple but mainstream to extract features in datasets and then replace the last fully connected layer of SoftMax classifier with the hashing layer. We set the training epochs at 100 and the batch size at 32. When generating the semantic-aware hash code, we choose examples from images in the database with the query label to form the hash code set. α and T of our attack method are set to $1/255$ and 100, the perturbation ϵ is $8/255$, and the number of random initializations N is 5. The hyperparameters t_1, t_2, and p of weight vector are set 0.5, 0.05, and 0.2. Following [30], the parameter β is set as 0.1 during the first 50 iterations and is updated every ten iterations according to $[0.2, 0.3, 0.5, 0.7, 1.0]$ during the last 50 iterations.

2) In Defense Stage: α and T of our attack method are set $2/255$ and 7. The perturbations ϵ and N are also $8/255$ and 5, and the semantic-aware hash codes are generated from mini-batch training examples. We set a batch size of 64 to generate category center hash codes with more images of the same category in a small batch of examples during adversarial training. The other setting is the same as the attack stage. Compared with HashNet and FAH, the loss value is smaller and more sensitive to datasets with different data volumes on DPSH. The hyperparameter γ on UCM, AID, and NWPU-RESISC45 with DPSH is set to 12, 13, and 19, respectively. The hyperparameter γ is set to 1 on all datasets with HashNet and FAH.

IV. Experimental Results

Section IV-A is mainly about the introduction of remote-sensing image scene datasets used in experiments. The adopted evaluation index is used for RSIR. Section III-C consists of two parts, including experimental implementation details such as the proportional division of the training set and test set, the related experimental configuration, and so on. Section IV-B demonstrates the excellence of our proposed attack method RWC through experiments. Finally, Section IV-C demonstrates the excellence of our proposed defense method ACN through experiments.

A. Datasets and Evaluation Metrics

To evaluate the performance of our proposed method for the attack and defense on deep hash neural networks for the CBRSIR tasks, we conduct extensive experiments on three benchmarks: the UC Merced Land Use Dataset (UCM), the AID dataset, and the NWPU-RESISC45 dataset.

1) *UCM Dataset [53]:* The images are manually extracted from the USGS National Map Urban Area Imagery collection for various urban areas around the country. This dataset consists of 2100 overhead scene images, including 21 land-use classes. Each class contains 100 aerial images measuring 256 × 256 pixels, with a spatial resolution of 0.3 m/pixel in red–green–blue color space.
2) *AID Dataset [54]:* The images are extracted from Google Earth, and AID is a multisource. Compared with UCM, it has more challenges. This dataset is composed of 30 scene categories, and each category contains 200–420 image scenes with a size of 600 × 600 pixels. It has 10 000 images. The pixel resolution of this public-domain imagery is 0.5 to 8 m.
3) *NWPU-RESISC45 Dataset [55]:* The images are extracted from Google Earth, like an AID dataset. The dataset has 31 500 within 45 scene categories. This category contains 700 image scenes with a size of 256 × 256 pixels. The pixel resolution of this public-domain imagery is 0.2–30 m.

In the article, the performance of RSIS attack methods and defense methods is evaluated using the widely adopted mean average precision (MAP). More specifically, the MAP score can be computed from

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{R_i}\sum_{j=1}^{n}\mathrm{Precision}\left(r_i^j\right)\delta\left(r_i^j\right) \tag{16}$$

where $q_i \in Q$ denotes the inquiry image, $|Q|$ denotes the volume of the inquiry image dataset, and R_i is the number of ground-truth neighbors of the query image q_i in the database image dataset, n is the number of all the entities in the database image dataset, $\mathrm{Precision}(r_i^j)$ denotes the precision of the top j retrieval entities, and $\delta(r_i^j) = 1$ in the jth retrieval entity is a ground-truth neighbor, and otherwise, $\delta(r_i^j) = 0$.

Note that, for a retrieval method, the satisfactory output is that more similar images can be ranked in the top positions. Therefore, we use the top 1000 retrieved images to count the MAP values in the following experiments.

B. Experimental Results of Deep Hashing Networks With Adversarial Attack

The overall attack performance of different methods is shown in Table I, where None and Gaussian Noise are to query with original clean examples and noisy examples by adding random noise from the uniform distribution $U(-\epsilon, +\epsilon)$ to original clean examples. HAG is the method proposed in [30]. SAAT is the method proposed in [48]. RWC is the attack method we proposed. It can be observed that deep hashing networks can yield good performance querying with the original clean examples for all three datasets, and the MAP values of Gaussian noise are only slightly less than the original clean examples, which shows that adding random noise cannot bias the outputs of the deep hashing network. However, when confronted with crafted adversarial examples (i.e., HAG, SAAT, and RWC), all the deep hashing networks fail to yield satisfactory performance, and the MAP drops to less than 1.5%, which fully demonstrates the vulnerability of deep hashing networks for RSIR. Even FAH, which currently obtains the best performance for RSIR, cannot defend against crafted adversarial examples. In addition, the proposed RWC attack method achieves the best attack performance. Especially on UCM, the MAP of HashNet is approximately 0, confronted with adversarial examples produced by RWC, which means that all the images retrieved by HashNet are irrelevant.

The example of a query result in HashNet under 64-bit code length with original clean images and adversarial samples is shown in Figs. 3–5. As is shown, the results querying with

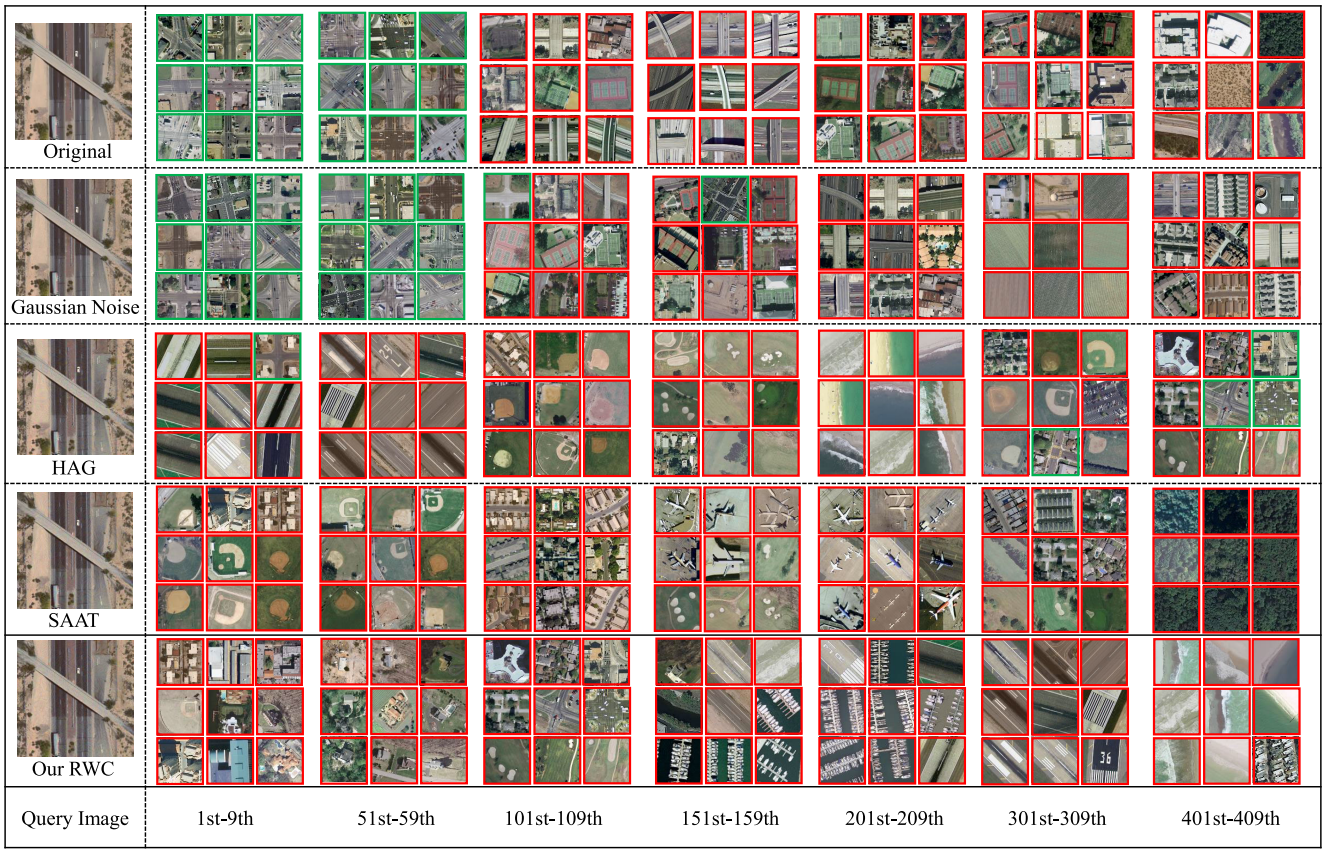

Fig. 3. Example of adversarial samples retrieval of attack methods on the UCM dataset with HashNet under 64-bit code length.

TABLE I
MAP (%) OF DIFFERENT DEEP HASHING METHODS ON THE THREE DATASETS

Deep Hashing Networks	Attack Methods	UCM		AID		NWPU-RESISC45	
		64 bits	96 bits	64 bits	96 bits	64 bits	96 bits
DPSH [25]	None	96.54	96.82	90.20	92.76	84.29	85.59
	Gaussian Noise	95.93	96.65	89.89	92.85	83.51	85.03
	HAG [30]	0.398	0.122	0.578	0.980	0.932	1.461
	SAAT [48]	0.311	0.012	0.854	1.025	1.083	1.302
	Our RWC	**0.003**	**0.003**	**0.213**	**0.274**	**0.341**	**0.511**
HashNet [42]	None	95.43	96.86	91.08	92.80	88.93	88.90
	Gaussian Noise	93.71	95.99	90.36	92.02	87.90	88.64
	HAG [30]	0.072	0.196	0.325	0.467	0.494	0.639
	SAAT [48]	0.004	0.007	0.292	0.753	0.370	0.581
	Our RWC	**0.000**	**0.000**	**0.171**	**0.211**	**0.069**	**0.100**
FAH [14]	None	97.88	97.90	92.19	92.86	88.94	88.60
	Gaussian Noise	96.89	97.23	91.57	92.24	88.80	88.08
	HAG [30]	0.187	0.254	0.474	0.387	0.294	0.171
	SAAT [48]	0.006	0.000	0.495	0.669	0.330	0.108
	Our RWC	**0.003**	**0.000**	**0.034**	**0.092**	**0.123**	**0.050**

original clean images and Gaussian noise adversarial samples are the most semantically relevant images, and the results querying with HAG and RWC adversarial samples are the most semantically irrelevant images, especially on AID and NWPU-RESISC45. Because there are only 100 images in the UCM retrieval category selected, the images in the lower position are all retrieval errors.

Perceptibility is also an important criterion for evaluating the quality of adversarial examples. Following [30], given an original clean example x, the perceptibility of an adversarial example x' can be calculated by $((1/n)\|x' - x\|_2^2)^{1/2}$, where n is the number of image pixels and the pixel values of x and x' are normalized to $[0, 1]$. The higher the perceptibility, the worse the visual quality of adversarial examples. The results of HashNet with 64 bits on three datasets are shown in Fig. 6, indicating that these disturbances are very small and undetectable. The disturbances shown in Figs. 3–5 also verify this view. Original and

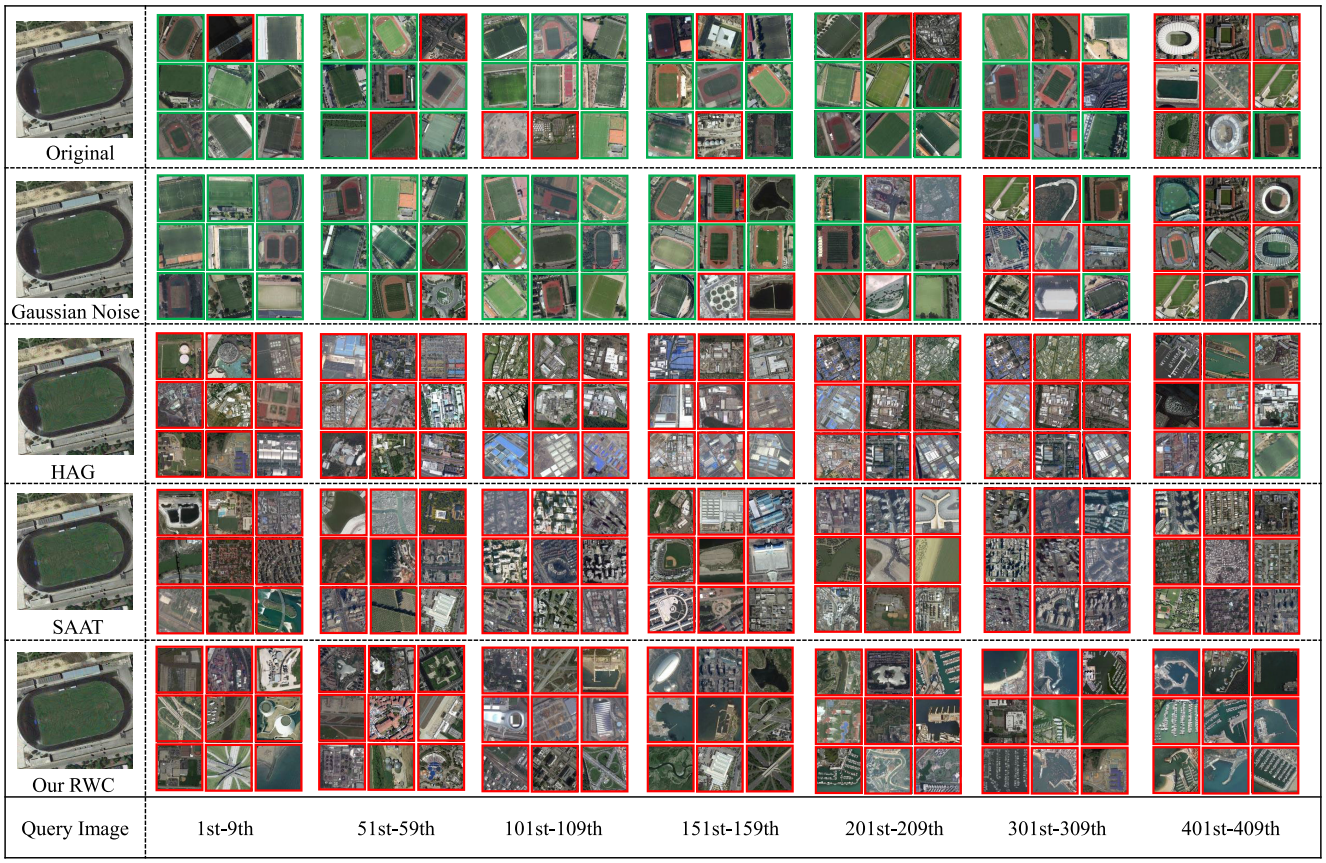

Fig. 4. Example of adversarial samples retrieval of attack methods on the AID dataset with HashNet under 64-bit code length.

RWC adversarial samples are indistinguishable to the naked eye.

Moreover, to visually show the comparison effect of attacks, Fig. 7 shows the topN curves, which is the accuracy of the first N retrieved images, of HAG and RWC attacks on HashNet. As shown in the figure, the MAP of the HashNet attacked by the RWC is always lower than that of HAG. The MAP is basically zero when confronted with adversarial examples produced by RWC. This fully demonstrates the advanced nature of the RWC attack method.

As described in Section III-A, the proposed RWC is composed of three parts: random initializations, weight vectors, and semantic-aware hash codes. When the three parts are used separately, we present the results to demonstrate the effectiveness of each part.

The attack effect of the three parts in HashNet with 64 bits is shown in Fig. 8, where None is the gradient attack method introduced in Section III. Rand is the random initialization. To better display the random initialization effect, the number of random initializations N is set to 5. Weight is the vector of added weight. Semantic is the objective function of the semantic-aware hash code. It is clear that rand, weight, and semantic have all improved their roles. Among them, Rand achieves a better attack effect by avoiding falling into the local optimal value through multiple starts. Weight achieves a better attack effect by assigning different attack weights to hash codes on different hash bits. Semantic selects a more category-based image hash code as the target attack function through semantic-aware, to obtain a better attack effect. The superposition of these three components also determines the advanced nature of RWC. However, the point pair setting method that plays a decisive role is semantic.

The parameter analysis for the number of random initializations N is shown in Table II. Different numbers of random initialization were added on the basis of the original gradient attack to conduct experiments. The experiments were all carried out on HashNet with 64 bits. $N = 0$ indicates that the adversarial examples were initialized directly to the original examples. It can be seen from the table that a better attack effect is achieved in the case of a random initial iteration. When the number of iterations is 0, the result has no significant change with the increase in the number of random initializations, which indicates that only when the attack method is used together with random initialization, the result be improved. In the case of 100 iterations near convergence, the random initialization achieves a better attack effect, which indicates that the global optimum in the disturbance space can be found with greater probability when approaching convergence through multiple random initializations, and multiple local optimals can be found through multiple random initializations. Then choose the best result as the global optimum. In addition, except that the result after 100 iterations of the UCM dataset is 0, AID and NWPU show a trend toward better attack effects as the number of iterations increases. To make reasonable use of resources, $N = 7$ has no obvious impact on the results compared with $N = 5$, so $N = 5$ is selected.

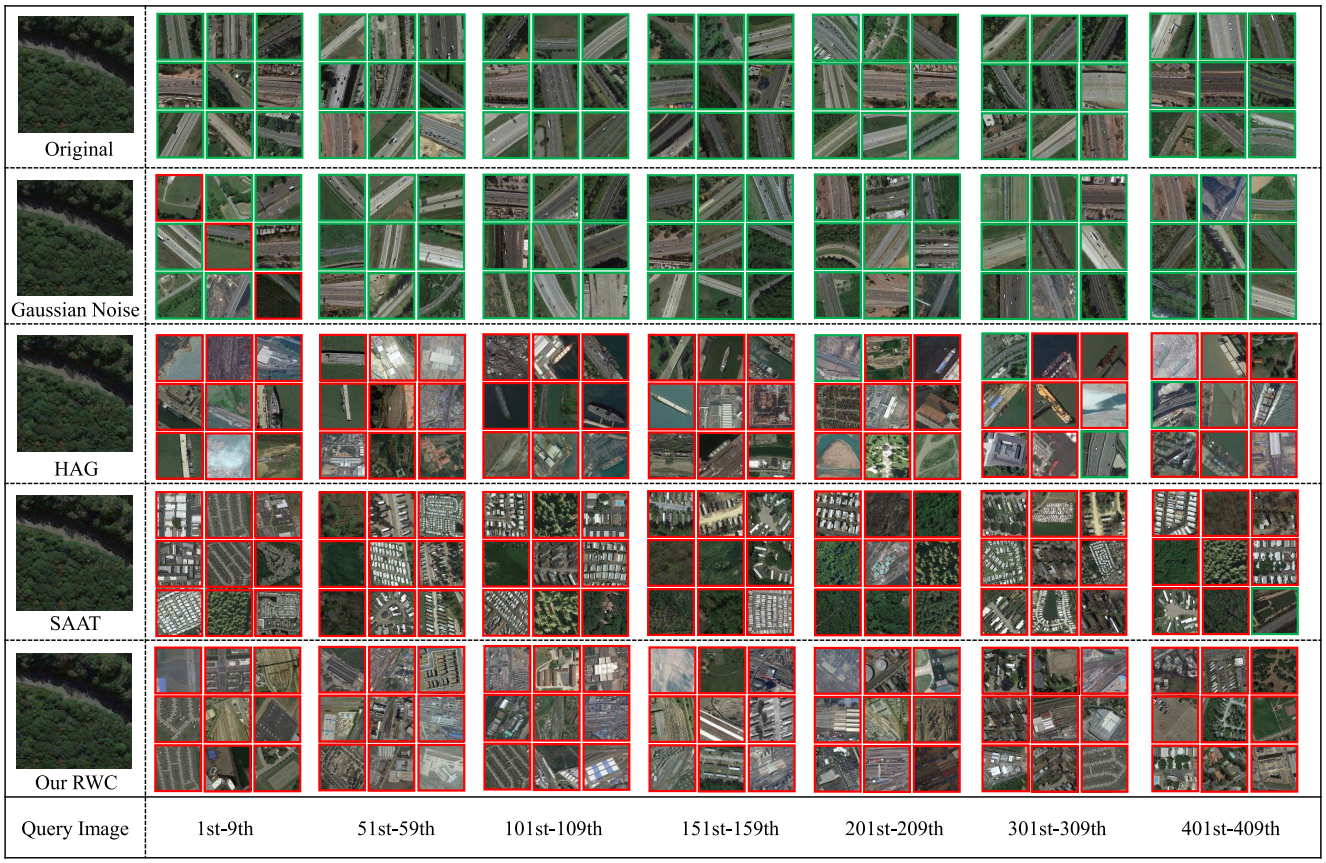

Fig. 5. Example of adversarial samples retrieval of attack methods on the NWPU-RESISC45 dataset with HashNet under 64-bit code length.

TABLE II
MAP (%) UNDER DIFFERENT NUMBER OF RANDOM INITIALIZATIONS

Attack	UCM					AID					NWPU-RESISC45				
Iteration	N=0	N=1	N=3	N=5	N=7	N=0	N=1	N=3	N=5	N=7	N=0	N=1	N=3	N=5	N=7
0	95.43	95.42	95.18	95.18	95.18	91.07	91.08	91.08	91.08	91.08	87.63	87.41	87.39	87.41	87.40
100	0.000	0.000	0.000	0.000	0.000	0.184	0.181	0.177	0.171	0.169	0.108	0.074	0.073	0.069	0.068

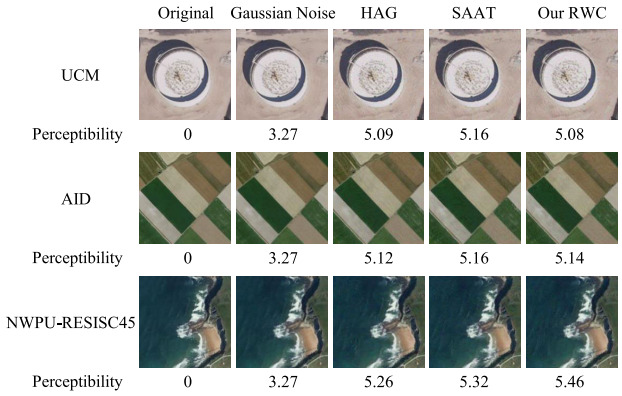

Fig. 6. Examples of attack images and perceptibility score ($\times 10^{-4}$) on three datasets with HashNet under 64-bit code length.

Finally, the hyperparameter p in (9) is analyzed experimentally. The results on HashNet under 64-bit code length are shown in Table III. To eliminate other distractions, the attack method includes only the weight part. The experimental results

TABLE III
MAP (%) UNDER PARAMETER ANALYSIS OF p

Dataset	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5
UCM	0.108	**0.078**	0.136	0.159	0.118
AID	0.367	**0.308**	0.336	0.346	0.346
NWPU-RESISC45	0.471	**0.347**	0.463	0.436	0.436

show that the optimal behavior is achieved at $p = 0.2$ on all three datasets.

C. Experimental Results of Deep Hashing Networks With Adversarial Defense

The results of the defense experiment are as follows. The general defense performance of different methods is shown in Table IV, where None means the original deep hashing network in the face of attacks. BLT [32] is to add a specified loss term on the basis of the original training loss term to improve the network's robustness. Its related parameters are

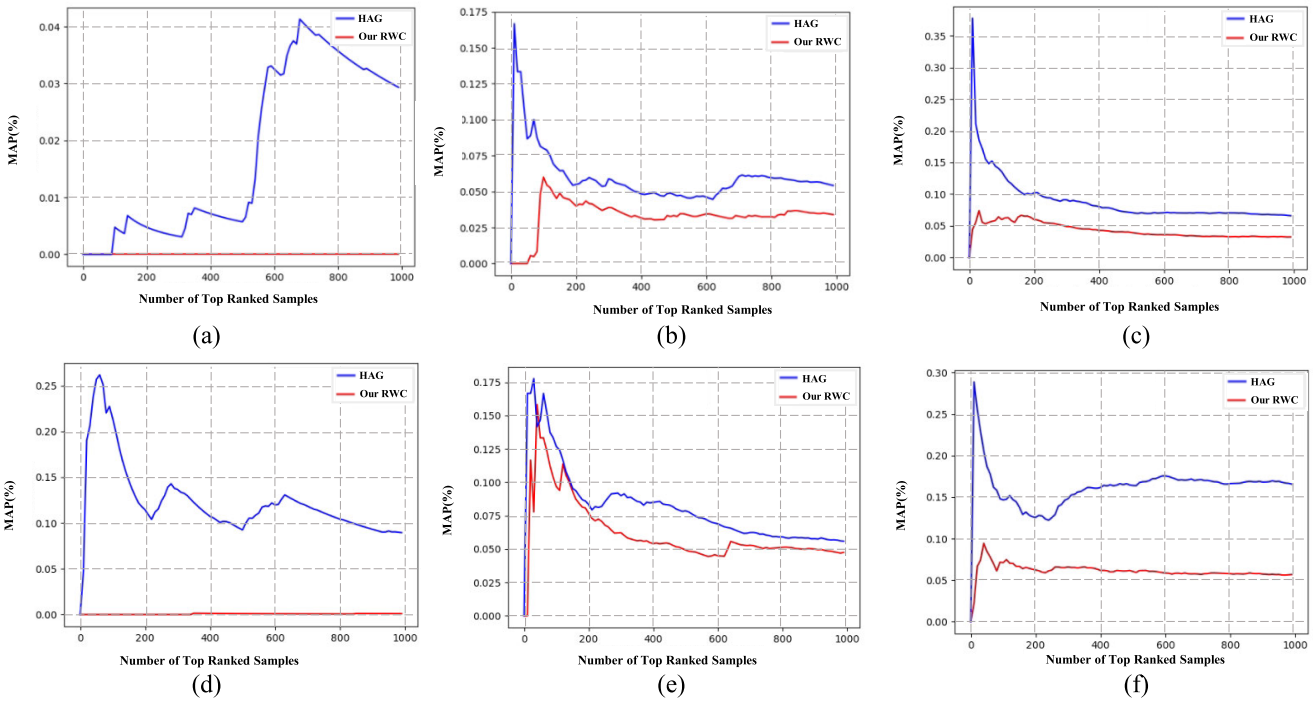

Fig. 7. TopN curves of HAG and RWC attack HashNet method. (a) UCM dataset with 64 bits. (b) AID dataset with 64 bits. (c) NWPU-RESISC45 dataset with 64 bits. (d) UCM dataset with 96 bits. (e) AID dataset with 96 bits. (f) NWPU-RESISC45 dataset with 96 bits.

TABLE IV
MAP (%) OF ADVERSARIAL TRAINING DEFENSE WITH ACN LOSS TERMS

Deep Hashing Networks	Defense Methods	UCM		AID		NWPU-RESISC45	
		64 bits	96 bits	64 bits	96 bits	64 bits	96 bits
DPSH [25]	None	2.46	3.14	0.85	1.02	0.57	0.71
	BLT [32]	9.68	13.34	4.09	4.39	2.23	2.35
	SAAT [48]	30.63	35.88	11.79	17.38	8.12	9.40
	Our ACN	**40.12**	**45.92**	**18.48**	**20.90**	**10.61**	**9.65**
HashNet [42]	None	1.66	1.57	0.52	0.40	0.39	0.31
	BLT [32]	8.80	6.05	2.84	2.85	2.52	2.46
	SAAT [48]	21.75	21.33	12.06	10.18	8.04	7.78
	Our ACN	**41.44**	**40.75**	**26.72**	**26.67**	**19.00**	**19.01**
FAH [14]	None	2.65	1.47	0.33	0.45	0.14	0.21
	BLT [32]	11.15	14.74	6.52	5.39	4.07	3.73
	SAAT [48]	13.66	10.79	8.57	6.93	6.78	4.33
	Our ACN	**38.95**	**39.90**	**22.30**	**18.73**	**13.54**	**14.36**

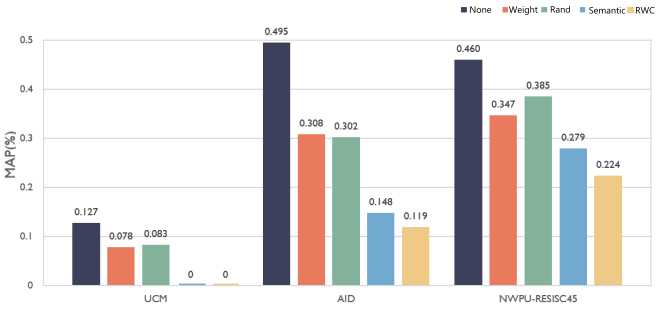

Fig. 8. MAP (%) of the different components in the RWC attack method.

consistent with the original text. SAAT [48] is to take the mainstay code as a label to guide the adversarial attack. Our ACN is a deep hashing network adversarial training defense method with ACNs proposed by Section III-B. As can be seen from Table IV, the result of the network in the face of adversarial examples is significantly improved after the adversarial defense. However, the results of the defense method BLT and SAAT in the three deep hashing network methods are all worse than the ACN, especially in HashNet and FAH methods, because HashNet contains the weighted loss term of similar image pairs, and FAH contains the semantic constraint term. Therefore, the method of giving a single loss term will have a poor effect when the form of the loss term is quite different. On the basis of the robustness analysis of the original hashing network, ACN makes full use of the similarity relationship between adversarial examples and original examples and the consistency relationship between adversarial examples and original class hash codes, to achieve a better defense effect.

TABLE V
ABLATION EXPERIMENT OF LOSS ITEMS FOR THE PROPOSED DEFENSE METHOD

L_{CSL}	L_{ASL}	L_{AAC}	UCM			AID			NWPU-RESISC45		
			Natural	Adversarial	OA GAP	Natural	Adversarial	OA GAP	Natural	Adversarial	OA GAP
✓			95.43	1.66	93.77	91.08	0.52	90.56	88.93	0.39	88.54
✓	✓		93.39	39.94	53.45	79.58	26.19	53.39	74.02	18.24	55.78
✓	✓	✓	91.48	41.44	50.04	79.16	26.72	52.44	73.86	19.00	54.86

TABLE VI
MAP (%) OF DIFFERENT ATTACK METHODS AFTER ADVERSARIAL TRAINING BY OUR ACN

Methods	UCM					AID					NWPU-RESISC45				
	Original	Gaussian Noise	HAG	SAAT	Our RWC	Original	Gaussian Noise	HAG	SAAT	Our RWC	Original	Gaussian Noise	HAG	SAAT	Our RWC
ST	95.43	93.71	0.07	0.00	**0.00**	91.08	90.36	0.33	0.29	**0.17**	88.93	87.90	0.49	0.37	**0.07**
Our ACN	91.48	91.59	37.26	37.01	**34.34**	79.16	79.05	24.99	22.56	**22.10**	73.86	73.72	17.39	**15.28**	15.49

To demonstrate the effect of the proposed defense loss items, an ablation study is performed. The results are shown in Table V, where natural represents the retrieval accuracy of the network on the original clean image. Adversarial represents the retrieval accuracy of the network in the face of the adversarial example attack, and OA Gap is the difference between the retrieval accuracy of the network on the original examples and adversarial examples. L_{CSL} is the clean example similarity constraint loss term of the original hash network, $L_{\mathrm{CSL}} + L_{\mathrm{ASL}}$ is the addition of the proposed adversarial example and clean example on the basis of the original loss term, $L_{\mathrm{CSL}} + L_{\mathrm{ASL}} + L_{\mathrm{AAC}}$ is the addition of the similarity constraint loss term between the adversarial example and clean example and the adversarial semantic-aware consistency loss term. The results show that both $L_{\mathrm{CSL}} + L_{\mathrm{ASL}}$ and $L_{\mathrm{CSL}} + L_{\mathrm{ASL}} + L_{\mathrm{AAC}}$ improve the robustness of the network. $L_{\mathrm{CSL}} + L_{\mathrm{ASL}} + L_{\mathrm{AAC}}$ further improves the robustness of the network on the basis of $L_{\mathrm{CSL}} + L_{\mathrm{ASL}}$, but the retrieval accuracy on the original clean examples will be slightly decreased. This involves a tradeoff between network generalization and robustness. The experiment proves that adding adversarial examples on the training stage can effectively improve the robustness of the network by maintaining a similar relationship between adversarial examples and original clean examples and adding the consistency relationship between the hash codes of adversarial examples and original category examples can further improve the network defense effect.

To further reveal the advanced nature of ACN, Table VI shows the retrieval accuracy of the network in the face of different attack methods, where ST represents the original training method and ACN is the defense method proposed in this article. To fully demonstrate the effectiveness of the attack method and the defense method, all attack methods adopt the same parameters as in the attack stage. After the adversarial training, the network robustness is greatly improved, and it is effective against all attack methods, which proves the advanced nature of the proposed defense method and its ability to resist the current network attack methods with good performance.

V. CONCLUSION

In this article, we propose adversarial semantic-aware attack and defense methods to mitigate the malicious attack on deep hashing networks for RSIR. To begin, we propose a gradient-based attack, which is a semantic-aware attack with weights via multiple random initializations, in light of the shortcomings of existing hash gradient attack methods. The experimental results show that the deep hashing network with excellent performance in the RSIR field is vulnerable to an adversarial example attack and the advanced nature of the proposed attack method. The adversarial semantic-aware defense method is then investigated on deep hashing networks. Experiments show that adversarial training is a simple but effective method to improve network defense performance, which improves the robustness of the RSIR system in the face of adversarial example attacks. Ablation experiments and visualization demonstrate the effectiveness of the proposed modules. Furthermore, it is necessary to continue studying the adversarial examples in future work. We intend to optimize the deep hashing network and optimize the loss function considering class semantic information, to improve the robustness of the model.

REFERENCES

[1] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
[2] Y. Wang, D. Yu, S. Ji, Q. Cheng, and M. Luo, "The joint spatial and radiometric transformer for remote sensing image retrieval," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 227–231, Aug. 2020.
[3] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Vancouver, BC, Canada: Curran Associates, 2008, pp. 1–8.
[4] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. ICML*, 2011, pp. 1–8.
[5] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 892–904, Feb. 2016.
[6] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.
[7] P. Li and P. Ren, "Partial randomness hashing for large-scale remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 464–468, Mar. 2017.

[8] Y. Li, Y. Zhang, C. Tao, and H. Zhu, "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion," *Remote Sens.*, vol. 8, no. 9, p. 709, Aug. 2016.

[9] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2018.

[10] L. Liu, Y. Wang, J. Peng, and A. Plaza, "DFLLR: Deep feature learning with latent relationship embedding for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608114.

[11] Z. Yuan et al., "MCRN: A multi-source cross-modal retrieval network for remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, Dec. 2022, Art. no. 103071.

[12] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[13] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520303778

[14] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Deep hash learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3420–3443, Apr. 2021.

[15] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[16] C. Liu, J. Ma, X. Tang, X. Zhang, and L. Jiao, "Adversarial hash-code learning for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4324–4327.

[17] J. T. Hoe, K. W. Ng, T. Zhang, C. S. Chan, Y.-Z. Song, and T. Xiang, "One loss for all: Deep hashing with a single cosine similarity based learning objective," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24286–24298.

[18] X. Tang, C. Liu, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Large-scale remote sensing image retrieval based on semi-supervised adversarial hashing," *Remote Sens.*, vol. 11, no. 17, p. 2055, Sep. 2019.

[19] P. Li et al., "Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7331–7345, Oct. 2020.

[20] X. Tang et al., "Meta-hashing for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615419.

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[22] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4480–4488.

[23] S. Sankaranarayanan, A. Jain, R. Chellappa, and S. N. Lim, "Regularizing deep networks using efficient layerwise adversarial training," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[24] R. Venkatesan, S.-M. Koon, M. H. Jakubowski, and P. Moulin, "Robust image hashing," in *Proc. Int. Conf. Image Process.*, 2000, pp. 664–666.

[25] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," 2015, *arXiv:1511.03855*.

[26] Y. Li, E. X. Fang, H. Xu, and T. Zhao, "Inductive bias of gradient descent based adversarial training on separable data," 2019, *arXiv:1906.02931*.

[27] Y. Cao et al., "DML-GANR: Deep metric learning with generative adversarial network regularization for high spatial resolution remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8888–8904, Dec. 2020.

[28] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[29] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610426.

[30] E. Yang, T. Liu, C. Deng, and D. Tao, "Adversarial examples for Hamming space search," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1473–1484, Apr. 2020.

[31] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," 2018, *arXiv:1803.06373*.

[32] X. Wang, Z. Zhang, G. Lu, and Y. Xu, "Targeted attack and defense for deep hashing," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2298–2302.

[33] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[34] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[35] X. Liu et al., "Adversarial training for large neural language models," 2020, *arXiv:2004.08994*.

[36] T. Zhang and Z. Zhu, "Interpreting adversarially trained convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7502–7511.

[37] X. Chen, C. Xie, M. Tan, L. Zhang, C.-J. Hsieh, and B. Gong, "Robust and accurate object detection via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16617–16626.

[38] Y. Fang, P. Li, J. Zhang, and P. Ren, "Cohesion intensive hash code book coconstruction for efficiently localizing sketch depicted scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629016.

[39] S. K. Sudha and S. Aji, "A review on recent advances in remote sensing image retrieval techniques," *J. Indian Soc. Remote Sens.*, vol. 47, no. 12, pp. 2129–2139, Dec. 2019.

[40] L. Han, P. Li, A. Plaza, and P. Ren, "Hashing for localization (HfL): A baseline for fast localizing objects in a large-scale scene," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609916.

[41] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1–7.

[42] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5609–5618.

[43] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[44] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*.

[45] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[46] Y. Xu and P. Ghamisi, "Universal adversarial examples in remote sensing: Methodology and benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619815.

[47] J. Lu, M. Chen, Y. Sun, W. Wang, Y. Wang, and X. Yang, "A smart adversarial attack on deep hashing based image retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 227–235.

[48] X. Yuan, Z. Zhang, X. Wang, and L. Wu, "Semantic-aware adversarial training for reliable deep hashing retrieval," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4681–4694, 2023.

[49] J. Bai et al., "Targeted attack for deep hashing based retrieval," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 618–634.

[50] X. Wang, Z. Zhang, B. Wu, F. Shen, and G. Lu, "Prototype-supervised adversarial network for targeted attack of deep hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16352–16361.

[51] L. Chen, Z. Xu, Q. Li, J. Peng, S. Wang, and H. Li, "An empirical study of adversarial examples on remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7419–7433, Sep. 2021.

[52] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, Feb. 2021.

[53] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2010, pp. 270–279.

[54] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[55] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

Yansheng Li (Senior Member, IEEE) received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He is currently a Full Professor with the School of Remote Sensing and Information Engineering, Wuhan University (WHU), Wuhan. He has authored more than 100 peer-reviewed journal articles and conference papers. His research interests include knowledge graph, deep learning, and their applications in remote-sensing big data mining.

Dr. Li was awarded the Young Surveying and Mapping Science and Technology Innovation Talent Award of the Chinese Society for Geodesy, Photogrammetry and Cartography in 2022. He received the recognition of the Best Reviewers of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) in 2021 and the Best Reviewers of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2022. He is an Associate Editor of IEEE TGRS, a Junior Editorial Member of *The Photogrammetric Record*, and a Lead Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Mengze Hao received the B.S. degree in data science and big data technology from China Agricultural University, Beijing, China, in 2022. She is currently pursuing the M.S. degree in pattern recognition and intelligent systems with Wuhan University, Wuhan, China.

Her research interests include remote-sensing image change detection, attacks, and defense.

Rongjie Liu received the B.Eng. degree in photogrammetry and remote sensing and the M.S. degree in environmental engineering from Wuhan University, Wuhan, China, in 2020 and 2022, respectively.

His research interests include remote-sensing image retrieval, attacks, and defense.

Zhichao Zhang received the B.S. degree in surveying engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University (WHU), Wuhan, China, in 2005 and 2010, respectively.

He is currently a Lecturer with the School of Remote Sensing and Information Engineering, WHU. His research interests include knowledge graph, deep learning, and their applications in cultural heritage conservation and remote-sensing big data mining.

Hu Zhu (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from the Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2013, respectively.

Now, he is a Professor with the School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include pattern recognition.

Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

From 2014 to 2015, he was a Senior Visiting Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. From 2015 to 2018, he was a Senior Scientist at Environmental Systems Research Institute Inc. (Esri), Redlands, CA, USA. He is currently the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. He holds 23 Chinese Patents and 26 copyright-registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote-sensing image interpretation, integration of light detection and ranging (LiDAR) point clouds and images, and 3-D city reconstruction.

Dr. Zhang was a Key Member of ISPRS Workgroup II/I from 2016 to 2020. He is the PI Winner of the Second-Class National Science and Technology Progress Award in 2017 and the PI Winner of the Outstanding-Class Science and Technology Progress Award in Surveying and Mapping (Chinese Society of Surveying, Mapping and Geoinformation, China) in 2015. In recent years, he has also served as the session chair for above 20 international workshops or conferences. He has been frequently serving as a referee for over 20 international journals. He is the Coeditor-in-Chief of *The Photogrammetric Record*.