HUANG: A Robust Diffusion Model-based Targeted Adversarial Attack Against Deep Hashing Retrieval

Chihan Huang , Xiaobo Shen*

Nanjing University of Science and Technology, Nanjing, China
huangchihan@njust.edu.cn, njust.shenxiaobo@gmail.com

Abstract

Deep hashing models have achieved great success in retrieval tasks due to their powerful representation and strong information compression capabilities. However, they inherit the vulnerability of deep neural networks to adversarial perturbations. Attackers can severely impact the retrieval capability of hashing models by adding subtle, carefully crafted adversarial perturbations to benign images, transforming them into adversarial images. Most existing adversarial attacks target image classification models, with few focusing on retrieval models. We propose HUANG, the first targeted adversarial attack algorithm to leverage a diffusion model for hashing retrieval in black-box scenarios. In our approach, adversarial denoising uses adversarial perturbations and residual image to guide the shift from benign to adversarial distribution. Extensive experiments demonstrate the superiority of HUANG across different datasets, achieving state-of-the-art performance in black-box targeted attacks. Additionally, the dynamic interplay between denoising and adding adversarial perturbations in adversarial denoising endows HUANG with exceptional robustness and transferability.

Introduction

With the exponential growth of information on the internet in recent years, the need for fast and accurate information retrieval has become increasingly important. In the Approximate Nearest Neighbor (ANN) algorithms widely used in search engines, hashing techniques convert high-dimensional media data into compact binary codes, enabling rapid computation of Hamming distances between hash codes to search for similar images (Andoni and Indyk 2006). Due to the powerful learning capabilities of Deep Neural Networks (DNNs), deep hashing that leverages DNNs for automatic feature extraction has made significant progress (Mnih et al. 2015). This approach substantially reduces retrieval time through efficient similarity computation of hash codes while achieving excellent retrieval accuracy, generally outperforming traditional hashing methods.

Although deep hashing methods have achieved remarkable retrieval performance (Cao et al. 2018), research has shown that DNNs are vulnerable to adversarial attacks, posing significant security risks to existing models (Nguyen, Yosinski, and Clune 2015). Attackers can introduce small, imperceptible perturbations to images, causing DNNs to make incorrect classifications or predictions. Unfortunately, deep hashing inherits this vulnerability from DNNs, resulting in the retrieval system returning irrelevant or specific other images. Most existing research on adversarial attack focuses on classification, with limited work addressing retrieval tasks. Retrieval models map query images to hash codes rather than categories (Wang et al. 2015), presenting a different attack surface for image retrieval. Consequently, the target labels cannot be directly used as learning objectives for adversarial samples, making adversarial attacks on retrieval more challenging.

Research on adversarial attacks can be divided into two categories: untargeted attacks and targeted attacks. In untargeted attacks, the generated adversarial samples cause the model's retrieval results to be semantically unrelated to the query image. Targeted attacks, on the other hand, aim to make the retrieval results correspond to a specified category, making them more challenging and destructive. Currently, major untargeted attack methods include HAG (Yang et al. 2020) and AACH (Li et al. 2021). Targeted attack methods include ProsGAN (Zhang et al. 2022), AdvHash (Hu et al. 2021), and TTA-GAN (Zhu et al. 2024b).

Nevertheless, in black-box scenarios, where targeted model details are inaccessible, existing adversarial attack methods for deep hashing generally suffer from poor robustness. Additionally, the generated images often have inferior visual quality and require longer inference time. The main contribution of the paper can be summarized as follows:

- We propose **H**ashing Rob**U**st **A**ttack with Diffusio**N** model for Tar**G**eted scenarios (HUANG), an adversarial attack algorithm for black-box hashing retrieval. To the best of our knowledge, it is the first time to use diffusion model for adversarial attacks on deep hashing.
- We propose adversarial denoising within the diffusion model. By incorporating the residual image, we enhance the robustness and transferability of the final generated adversarial image through the dynamic interplay between adding adversarial perturbations and denoising.
- Comprehensive experiments reveal that HUANG greatly outperforms prior models, setting a new benchmark for targeted black-box attacks. Evaluation against adversar-

*Corresponding author.

ial defense and other assessments validate the model's superior capabilities in terms of transferability and robustness.

Related Work

Deep Hashing

Conventional hashing methods seek to identify optimal subspaces within the data distribution to preserve the semantic features of the original samples for evaluating semantic similarity (Wang et al. 2016). Instead, deep hashing techniques directly learn semantic features from raw data, enabling end-to-end parameter optimization (Shen et al. 2022). Based on the use of label information, deep hashing models can be categorized into supervised (Xia et al. 2014; Li, Wang, and Kang 2016; Zhang et al. 2015) and unsupervised methods (Wang et al. 2015; Shen et al. 2018; Lin et al. 2016). Compared to unsupervised methods, supervised methods leverage semantic information from labels to train networks, achieving better performance. The pioneering supervised strategy CNNH (Xia et al. 2014) divided the training process into two stages, the first stage learns the hash codes and the second stage trains a CNN to generate hash codes. DPSH (Li, Wang, and Kang 2016), guided by paired labels, introduced a cross-entropy loss function to evaluate the quality of hash codes and added a regularization component to minimize quantization errors. HashNet (Cao et al. 2017) employed a convergence-driven sequential method to directly learn hash codes, producing precise binary codes from semantically similar data. CSQ (Yuan et al. 2020) enhanced efficiency and accuracy by leveraging central similarity.

Adversarial Attack

Adversarial attacks involve introducing imperceptible, intentional perturbations into input samples, leading to inaccurate model outputs (Szegedy et al. 2014). Depending on the attacker's familiarity with the model and dataset, these attacks can be categorized into white-box attacks and black-box attacks. In white-box attacks, the attacker has complete knowledge of the targeted model's details, including its structure, parameters, and gradients. Prominent white-box attack models include FGSM (Goodfellow, Shlens, and Szegedy 2015), PGD (Madry et al. 2019), DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016), and C&W (Carlini and Wagner 2017). However, in real-world scenarios, attackers rarely have access to such comprehensive information about the targeted model. This has led to the development of various black-box attack methods, such as ZOO (Chen et al. 2017) and DeepSearch (Zhang, Chowdhury, and Christakis 2020).

Despite the extensive research on adversarial attacks in image classification, adversarial attacks in image retrieval, particularly in deep hashing, have often been overlooked (Chen et al. 2021). In classification, adversarial perturbations are guided by image category labels to misclassify images into incorrect categories. However, in hashing retrieval, which focuses on similarity instead of explicit labels, directing the model learning process becomes much more complex. (Yang et al. 2020) first proposed an untargeted attack algorithm for deep hashing adversarial samples, known as HAG, which selectively modifies pixels to increase the attack success rate while maintaining the image visual quality. (Bai et al. 2020) introduced DHTA, a targeted attack algorithm for deep image hashing adversarial attack, which employs a hash code voting mechanism, wherein the hash codes of images within the same category are consolidated into a universal hash code through voting. This universal hash code is then input into the target function, and the final adversarial samples are obtained through backpropagation of gradients. (Hu et al. 2021) developed AdvHash, which uses a dot-product-based weighted gradient aggregation technique to create adversarial patches that can target multiple images under one label. (Fu et al. 2023) presented LGWAE, leveraging multi-label learning to produce hash codes as category features and using autoencoders for hashing attacks. (Zhu et al. 2024a) introduced a feature consistency loss aimed at universal adversarial perturbations, improving both their stability and aggressiveness.

Diffusion Model

In recent years, diffusion models have demonstrated superior performance over GANs in image generation (Song and Ermon 2019). Diffusion models are likelihood-based models widely used for high quality image generation. These models generate samples by progressively removing noise from latent variables, typically random noise (Ho, Jain, and Abbeel 2020). A diffusion model consists of a diffusion process and a reverse diffusion process, which is regarded as the image generation process. The diffusion process can be viewed as a fixed Markov chain, starting from the original data $\boldsymbol{x}_0$ and gradually adding noise over T diffusion steps to obtain $\boldsymbol{x}_T$. The reverse diffusion process involves gradually restoring $\boldsymbol{x}_T$ back to $\boldsymbol{x}_0$ through a Markov chain, which can be represented as follows:

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) := \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}\left(\boldsymbol{x}_t, t\right), \boldsymbol{\Sigma}\left(\boldsymbol{x}_t, t\right)\right) \quad (1)$$

where the mean $\boldsymbol{\mu}\left(\boldsymbol{x}_t, t\right)$ is fitted by a neural network with zero-mean parameters. The goal of training a diffusion model is to make the data distribution generated by the reverse diffusion process approach the original distribution, thereby generating more realistic and high quality images.

(Dhariwal and Nichol 2021) introduces a conditional diffusion model, which can be guided by a classifier during the reverse diffusion process to generate images of specified categories. Specifically, the pretrained diffusion model $p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ can be directed by adding conditions, derived from the gradient of the classifier $p\left(y \mid \boldsymbol{x}_t\right)$, to the sampling mean during the reverse diffusion process. This results in the generation of samples that satisfy the condition y on the original basis, and is referred to as the conditional diffusion model $p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, y\right)$. The derivation process of the conditional diffusion model is as follows:

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, y\right) = p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) \cdot e^{\log p(y|\boldsymbol{x}_{t-1}) - \log p(y|\boldsymbol{x}_t)} \quad (2)$$

$$\begin{aligned} &\log\left(p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) p_\phi\left(y \mid \boldsymbol{x}_t\right)\right) \\ &= \log p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) + \left(\boldsymbol{x}_{t-1} - \boldsymbol{\mu}\right) \boldsymbol{g} + c_1 \\ &\approx \log p(\boldsymbol{z}) + c_2 \end{aligned} \quad (3)$$

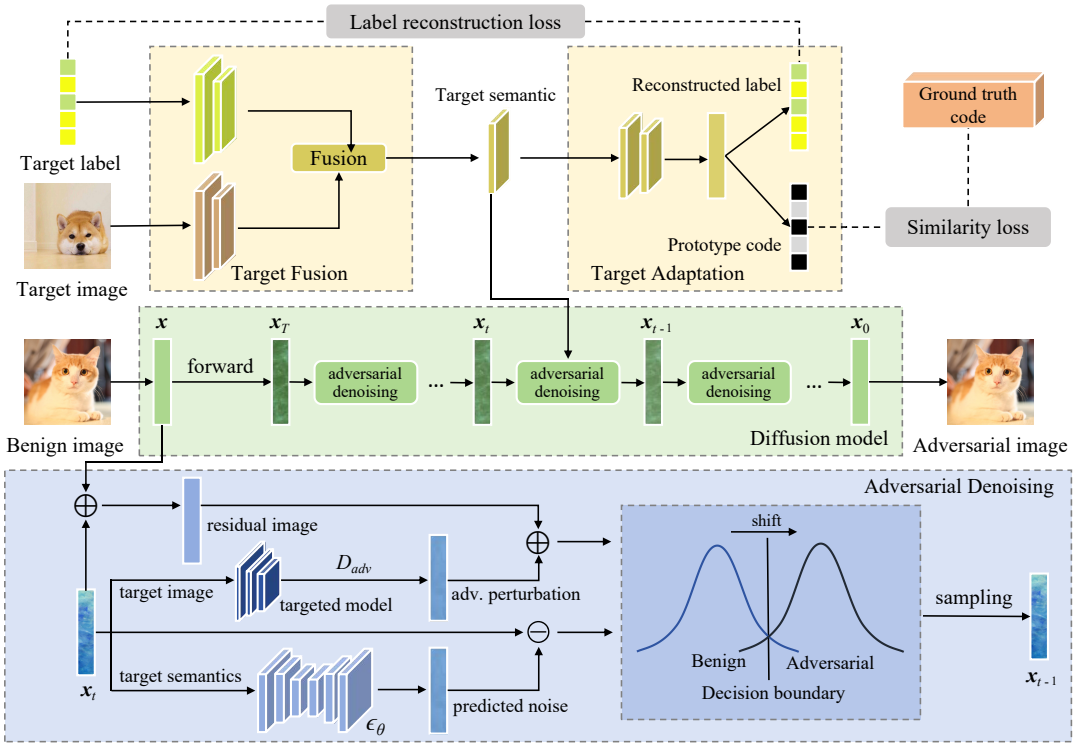


Figure 1: The overview structure of our model HUANG. The target label and image are fed into Target Fusion ($\mathcal{TF}$) to derive target semantics, used to reconstruct the label and produce a prototype code in Target Adaptation ($\mathcal{TA}$). In adversarial sample generation, we first input a benign image and add noise to obtain $\boldsymbol{x}_T$. During the reverse process, we use adversarial denoising to gradually restore the noisy $\boldsymbol{x}_T$ back to $\boldsymbol{x}_0$. In adversarial denoising, $\boldsymbol{x}_t$ and the predicted noise form a benign distribution, while the adversarial perturbation and residual image guide the shift of the image distribution from benign to adversarial. The subsequent $\boldsymbol{x}_{t-1}$ is sampled from this adversarial distribution. Finally, the output of the diffusion model is the adversarial image.

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{g}\boldsymbol{\Sigma}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is the mean of the pretrained diffusion model, $\boldsymbol{\Sigma}$ is the variance. c_1 and c_2 are constants, and $\boldsymbol{g}$ is the gradient computed by the classifier used to guide the diffusion model, which is $\boldsymbol{g} = \nabla_{\boldsymbol{x}_t} \log p\left(y \mid \boldsymbol{x}_t\right)$, ϕ is the parameter of the classification model.

Methodology

Preliminary

Let $\mathcal{X} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ denote a training set consisting of N images labeled with L classes, where $\boldsymbol{x}_i$ represents the i-th image, and $\boldsymbol{y}_i$ represents its corresponding label. Specifically, $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iL}] \in \{0,1\}^L$ denotes the multi-label vector associated with $\boldsymbol{x}_i$, where $y_{ij} = 1$ indicates that $\boldsymbol{x}_i$ belongs to class j. For a given image $\boldsymbol{x}_i$, its hash code can be obtained by the following expression:

$$\boldsymbol{h}_i = H\left(\boldsymbol{x}_i\right) = \operatorname{sign}\left(F\left(\boldsymbol{x}_i\right)\right) \quad \text{s.t.} \quad \boldsymbol{h}_i \in \{-1,1\}^K \tag{4}$$

where $F(\cdot)$ denotes a deep hashing model, and $F\left(\boldsymbol{x}_i\right)$ is the output of the last hash layer. The $\operatorname{sign}(\cdot)$ function is a sign function, binarizing the output to either -1 or 1, and K is the bit length of the hash code.

In targeted image hashing retrieval attack, for a given benign image $\boldsymbol{x}$ and a target label $\boldsymbol{y}_{tar}$, the goal is to generate an adversarial example, which should be visually imperceptible and capable of making the retrieval results semantically similar to $\boldsymbol{y}_{tar}$. We need to obtain the optimal $\Phi : (\boldsymbol{x}, \boldsymbol{y}_{tar}) \rightarrow \boldsymbol{x}_{adv}$ to generate effective and efficient adversarial image. In short words, the goal is to generate adversarial samples that make the hash retrieval system yield results closer to a target category than the original one.

Formulation

Overall framework We introduce HUANG, which is capable of generating effective adversarial examples with diffusion models under black-box scenarios to fool hashing retrieval systems. The overall framework of our method is shown in Figure 1. Structurally, HUANG comprises two stages: Semantic learning and Adversarial generation, where the former contains the target fusion module $\mathcal{TF}$ and target adaptation module $\mathcal{TA}$, and the latter contains the Diffusion model $\mathcal{DM}$.

Semantic learning The target label contains critical category information, and the target image holds significant target semantic information. These two modalities complement each other. Therefore in $\mathcal{TF}$, we first integrate their features to obtain the target semantic $\boldsymbol{f}^s$.

To enrich $\boldsymbol{f}^s$ with more semantic and category information from the targeted model, we design $\mathcal{TA}$ to generate a reconstructed label and prototype code:

$$\tilde{\boldsymbol{y}}, \boldsymbol{h}^p = \mathcal{TA}\left(\boldsymbol{f}^s \mid \Theta_{\mathcal{TA}}\right) \tag{5}$$

where $\tilde{\boldsymbol{y}}$ is the reconstructed label obtained from $\boldsymbol{f}^s$ through $\mathcal{TA}$, $\boldsymbol{h}^p$ is the prototype code, and $\Theta_{\mathcal{TA}}$ represents the parameters of $\mathcal{TA}$. We control this optimization process using reconstruction loss $\mathcal{L}_{rec}$ and similarity loss $\mathcal{L}_{sim}$.

The reconstruction loss $\mathcal{L}_{rec}$ ensures that $\boldsymbol{f}^s$ contains the target category information. It is expressed as follows:

$$\mathcal{L}_{rec} = \sum_i \left\|\tilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\right\|_2^2 \tag{6}$$

where $\|\cdot\|_2$ represents the ℓ_2 norm. The similarity loss $\mathcal{L}_{sim}$ controls the Hamming distance between the prototype code generated by $\boldsymbol{f}^s$ and the ground truth code generated by the targeted model. Similar to the triplet loss, the similarity loss minimizes the Hamming distance between the prototype code and the ground truth code, while maximizing that between the prototype code and codes from unrelated categories. It is expressed as follows:

$$\mathcal{L}_{sim} = \sum_i \sum_j \left(\boldsymbol{s}_{ij}\Delta_{ij} - \log\left(1 + e^{\Delta_{ij}}\right)\right) \tag{7}$$

where $\boldsymbol{s}_{ij}$ represents the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $\boldsymbol{s}_{ij} = 1$ if they share at least one category, and 0 otherwise, and $\Delta_{ij} = \frac{1}{2}\left(\boldsymbol{h}_i^p\right)^T \boldsymbol{h}_j$. Thus, the loss function for the semantic learning is $\mathcal{L}_{SL} = \mathcal{L}_{rec} + \lambda\mathcal{L}_{sim}$. Through the target adaptation process, the target semantic $\boldsymbol{f}^s$ offers high discriminative capability via the reconstructed label. By fitting prototype code to the ground truth code, it enhances the attack ability, leading to more destructive attacks. When semantic learning finishes, its parameters are fixed.

Adversarial denoising When using a pretrained diffusion model to generate images, the reverse process that transforms noise into an image is typically employed. In each step, the image $\boldsymbol{x}_{t-1}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$. Generally, the image obtained from the reverse process will have similar semantics to the benign image. However, by incorporating adversarial denoising into this process, the benign distribution is gradually shifted towards the adversarial distribution, resulting in an adversarial image. To achieve this, we modify the reverse process of $p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ in (1) to a conditional distribution:

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_{tar}\right) = C_1 p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) p_\theta\left(\boldsymbol{x}_{tar} \mid \boldsymbol{x}_{t-1}\right) \tag{8}$$

where $\boldsymbol{x}_{tar}$ is the target image, C_1 is a normalization factor. (Dhariwal and Nichol 2021) has demonstrated that a Gaussian distribution with a shifted mean can be used to approximate the distribution. We approximate the distribution using the following equation:

$$\begin{aligned} &p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) p_\theta\left(\boldsymbol{x}_{tar} \mid \boldsymbol{x}_{t-1}\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\Sigma}\nabla_{\boldsymbol{x}_t} \log p_\theta\left(\boldsymbol{x}_{tar} \mid \boldsymbol{x}_t\right), \boldsymbol{\Sigma}\right) \end{aligned} \tag{9}$$

where $p_\theta\left(\boldsymbol{x}_{tar} \mid \boldsymbol{x}_t\right) = C_2 \exp\left(s\mathcal{D}_{adv}\left(\boldsymbol{x}_{tar}, \boldsymbol{x}_t\right)\right)$ can be seen as the probability that $\boldsymbol{x}_{t-1}$ will be recovered to an adversarial image, and $\mathcal{D}_{adv}\left(\boldsymbol{x}_{tar}, \boldsymbol{x}_t\right) = -\frac{1}{K} H\left(\boldsymbol{x}_{tar}\right)^T H\left(\boldsymbol{x}_t\right) + 1$ measures the hamming distance. $p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ is the unconditional reverse process that we already possess, C_2 is a normalization factor, s is a scaling factor.

The same operation can be done on the target semantic $\boldsymbol{f}^s$: $p_\theta\left(\boldsymbol{f}^s \mid \boldsymbol{x}_t\right) = C_3 \exp\left(s\mathcal{D}_{sim}\left(\boldsymbol{f}^s, \boldsymbol{x}_t\right)\right)$, in which:

$$\mathcal{D}_{sim}\left(\boldsymbol{f}^s, \boldsymbol{x}_t\right) = \frac{\boldsymbol{f}^s \cdot \mathcal{TF}\left(\boldsymbol{x}_t, \boldsymbol{y}_{tar}\right)}{\|\boldsymbol{f}^s\| \|\mathcal{TF}\left(\boldsymbol{x}_t, \boldsymbol{y}_{tar}\right)\|} \tag{10}$$

where $\mathcal{TF}\left(\boldsymbol{x}_t, \boldsymbol{y}_{tar}\right)$ is $\mathcal{TF}$ output with input $\boldsymbol{x}_t$ and $\boldsymbol{y}_{tar}$. Denote $\boldsymbol{g} = \nabla_{\boldsymbol{x}_t}\mathcal{D}_{adv}\left(\boldsymbol{x}_{tar}, \boldsymbol{x}_t\right) + \mathcal{I}\left(\nabla_{\boldsymbol{x}_t}\mathcal{D}_{sim}\left(\boldsymbol{f}^s, \boldsymbol{x}_t\right)\right)$, where $\mathcal{I}$ is an interpolation funtcion, we modify the mean of the conditional distribution by $s\boldsymbol{\Sigma}\boldsymbol{g}$, and we can sample $\boldsymbol{x}_{t-1}$ from distribution $\mathcal{N}\left(\boldsymbol{\mu} + s\boldsymbol{\Sigma}\boldsymbol{g}, \boldsymbol{\Sigma}\right)$.

However, solely relying on adversarial perturbations to guide the distribution shift in the reverse process may lead to excessive deviations, resulting in generated images with significant visual differences from benign ones. To address this, we supervise the distribution shift using the residual image $\boldsymbol{r} = \boldsymbol{x}_t - \boldsymbol{x}_{ben}$. By subtracting this noise from the mean of the distribution, the mean shifts towards the benign distribution, thereby generating images that are more similar to benign images. Finally, $\boldsymbol{x}_{t-1}$ can be viewed as being sampled from the distribution:

$$\mathcal{N}(\boldsymbol{\mu} + s\boldsymbol{\Sigma}\boldsymbol{g} - \sqrt{\boldsymbol{\Sigma}}\boldsymbol{r}, \boldsymbol{\Sigma}) \tag{11}$$

The incorporation of adversarial perturbations and the denoising process create a dynamic interplay, significantly enhancing the robustness of the final adversarial image. After the reverse process we obtain the adversarial image. Algorithm 1 shows the process of adversarial denoising. In it, β_t controls the noise intensity added at each diffusion step, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$. They are all defined in the diffusion model (Ho, Jain, and Abbeel 2020).

Algorithm 1: Adversarial denoising

Input: Benign image $\boldsymbol{x}_{ben}$, target image $\boldsymbol{x}_{tar}$, pretrained diffusion model ϵ_θ, timestep T, scaling factor s, target semantic $\boldsymbol{f}^s$, interpolation function $\mathcal{I}$;

Output: Adversarial image $\boldsymbol{x}_{adv}$;

$z \sim \mathcal{N}(0, I)$;
$x_T = \sqrt{\bar{\alpha}_T}\boldsymbol{x}_{ben} + \sqrt{1 - \bar{\alpha}_T}z$;
for $t = T, T-1, \cdots, 1$ **do**
  $\boldsymbol{r} = \boldsymbol{x}_t - \boldsymbol{x}_{ben}$;
  $\boldsymbol{\Sigma} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$;
  $\boldsymbol{\mu} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\boldsymbol{x}_t, t\right)\right)$;
  $\boldsymbol{g} = \nabla_{\boldsymbol{x}_t}\mathcal{D}_{adv}\left(\boldsymbol{x}_t, \boldsymbol{x}_{tar}\right) + \mathcal{I}\left(\nabla_{\boldsymbol{x}_t}\mathcal{D}_{sim}\left(\boldsymbol{x}_t, \boldsymbol{f}^s\right)\right)$;
  $\boldsymbol{x}_{t-1} \leftarrow$ sample from $\mathcal{N}(\boldsymbol{\mu} + s\boldsymbol{\Sigma}\boldsymbol{g} - \sqrt{\boldsymbol{\Sigma}}\boldsymbol{r}, \boldsymbol{\Sigma})$;
end for
$\boldsymbol{x}_{adv} = \boldsymbol{x}_0$.
return $\boldsymbol{x}_{adv}$.

Experiments

Datasets

We conduct experiments on three datasets: FLICKR-25K (Huiskes and Lew 2008), NUS-WIDE (Chua et al. 2009),

Model	Method	FLICKR-25K				NUS-WIDE				MS-COCO			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
DPSH	Original	55.52	55.91	56.06	54.76	46.32	46.39	46.47	47.62	34.81	35.61	38.47	40.52
	DHTA	56.69	57.21	59.41	56.36	46.68	48.42	48.76	48.89	36.04	39.49	42.76	44.17
	ProS-GAN	26.93	58.17	60.26	57.62	46.81	48.87	49.13	49.25	38.14	42.62	43.59	45.92
	THA	59.14	59.01	60.88	62.76	49.01	49.13	49.41	49.15	37.80	40.96	43.01	44.85
	PTA	61.07	62.55	60.94	60.85	46.02	46.16	46.35	46.24	39.88	43.05	46.47	48.73
	SAAT	62.43	63.07	65.55	60.02	49.82	51.28	51.63	51.72	41.82	45.68	48.34	50.61
	HUANG	**71.64**	**78.15**	**73.66**	**71.32**	**61.74**	**63.49**	**66.50**	**68.08**	**52.31**	**55.90**	**57.85**	**61.41**
HashNet	Original	43.37	47.02	48.90	48.16	30.47	33.85	35.28	37.76	21.94	24.55	24.63	26.85
	DHTA	49.23	50.99	51.14	51.69	31.23	36.25	39.83	41.29	26.62	28.33	29.47	31.88
	ProS-GAN	50.16	51.10	52.82	53.13	35.29	37.06	40.95	43.48	28.42	30.84	33.36	34.80
	THA	47.01	47.61	48.21	48.58	36.62	38.39	42.32	44.91	30.65	31.33	33.91	35.26
	PTA	57.26	59.13	60.45	60.98	38.95	41.36	44.61	46.04	32.89	34.26	36.75	37.49
	SAAT	54.92	56.36	58.64	59.38	43.82	46.20	49.52	50.38	35.11	37.15	38.79	40.61
	HUANG	**64.18**	**68.67**	**64.67**	**66.63**	**52.44**	**56.11**	**58.69**	**61.53**	**43.82**	**45.91**	**47.16**	**49.77**
CSQ	Original	51.02	52.16	51.32	50.78	39.11	41.48	39.45	38.07	28.20	30.43	31.17	31.79
	DHTA	53.59	56.49	54.57	53.08	41.22	44.23	42.67	40.31	31.42	34.35	33.65	32.88
	ProS-GAN	56.74	57.99	58.74	60.39	43.01	45.19	43.92	41.15	34.89	36.71	35.61	34.21
	THA	56.79	60.19	59.40	57.88	44.65	47.77	46.86	44.54	35.95	37.71	35.08	32.51
	PTA	57.43	59.81	60.41	58.37	43.59	46.86	47.33	47.88	37.66	38.65	39.44	40.36
	SAAT	59.21	61.42	60.78	59.67	46.49	48.95	49.37	49.59	40.47	41.63	43.28	44.62
	HUANG	**70.15**	**73.56**	**72.94**	**71.65**	**57.24**	**59.87**	**60.77**	**62.43**	**47.83**	**49.36**	**51.89**	**53.67**

Table 1: The targeted attack performance comparison between HUANG and other advanced attack methods. The evaluation metric is t-MAP.

Figure 2: Visual comparison of benign images and adversarial images generated by HUANG on three datasets.

and MS-COCO (Lin et al. 2014).

The FLICKR-25K dataset comprises 25,000 images labeled across 38 categories, from which we have selected 1,700 images as our query set, 5,000 images for the training set, and the remaining images to form the database.

The NUS-WIDE dataset consists of 193,734 images labeled across 21 categories, with 2,100 images chosen as the query set, 10,000 for training, and the remaining images serving as the database.

The MS-COCO dataset contains 123,287 images labeled across 80 categories, with 5,000 images designated as the query set, 10,000 for training, and the remaining images serving as the database.

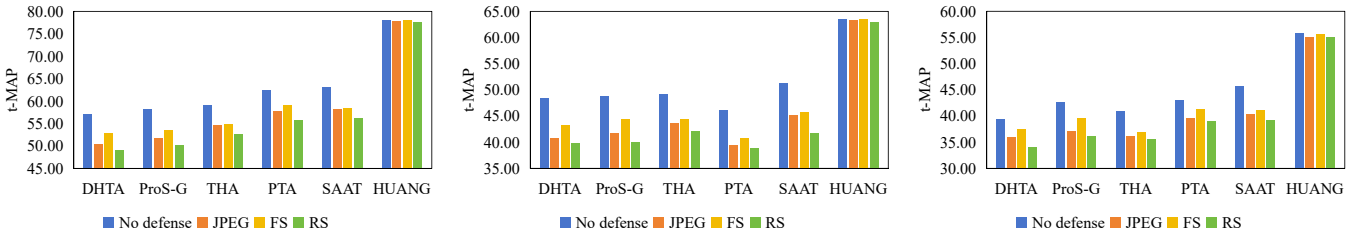
Experiment Setting

The experimental setup is anchored by a hardware environment running Windows 11 and powered by a GeForce RTX 4080 GPU. The software environment utilizes Python 3.8 and the PyTorch 2.0.0 + cu118. All images are resized to 224×224. The Adam optimizer is employed, with hyperparameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The perturbation threshold is established at $8/255$.

We select DPSH (Li, Wang, and Kang 2016), HashNet (Cao et al. 2017) and CSQ (Yuan et al. 2020) as the models to be attacked, with DPSH being the primary target. For targeted hash attacks, we utilize targeted mean average precision (t-MAP) as the evaluation metric, considering the target label as the reference label. A higher t-MAP value indicates greater precision in retrieving images related to the target label, thereby indicating a more effective attack.

Results

Targeted attack performance Table 1 demonstrates t-MAP of various targeted attack methods on the above datasets. Our method, HUANG, consistently outperforms competitors across different hash bit lengths and datasets against a range of targeted attack methods including DHTA (Bai et al. 2020), ProS-GAN (Zhang et al. 2022), THA (Wang et al. 2021), PTA (Zhao et al. 2023), and SAAT (Yuan et al. 2023). Using the primary targeted model DPSH



(a) FLICKR-25K (b) NUS-WIDE (c) MS-COCO

Figure 3: Comparison of t-MAP of HUANG and other SOTA algorithms after three kinds of defense methods.

Method	FLICKR-25K		NUS-WIDE		MS-COCO	
	t-MAP ↑	Perceptibility ↓	t-MAP ↑	Perceptibility ↓	t-MAP ↑	Perceptibility ↓
DHTA+FGSM	51.34	2.97	36.25	2.94	28.33	2.95
DHTA	57.21	**0.69**	48.42	**0.64**	39.49	**0.54**
ProS-GAN	58.17	2.38	48.87	1.75	42.62	2.11
HUANG	**78.15**	1.29	**63.49**	1.10	**55.90**	0.98

Table 2: t-MAP(%) and perceptibility ($\times 10^{-2}$) of different attack methods for the hashing model with 32-bits code length.

on the FLICKR-25K dataset, HUANG achieved an average t-MAP improvement of 10.93% over the previous SOTA model, SAAT. For the other two targeted models, HashNet and CSQ, our HUANG achieved more than a 10% performance boost over the previous best algorithms. From a dataset perspective, HUANG outperformed SAAT by an average of 10.48% on FLICKR-25K, 11.68% on NUS-WIDE, and 8.98% on MS-COCO. Notably, HUANG demonstrated even stronger performance on smaller datasets, attributed to its exceptional ability to capture fine-grained semantic information of target labels. From a targeted model perspective, HUANG surpassed SAAT by an average of 11.68% on DPSH, 19.67% on HashNet, and 10.49% on CSQ. This indicates HUANG possesses superior robustness and transferability, effectively leveraging the representational strengths of the attacked models.

Perceptibility & Efficiency In evaluating the quality of generated adversarial images, perceptibility is as crucial as attack performance. We calculate perceptibility using $\sqrt{\frac{1}{Z}\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2}$, where Z is the total number of pixels, normalized between [0,1]. A lower perceptibility indicates higher quality images. We assessed t-MAP, perceptibility, and generation time for 32-bit DPSH, as detailed in Table 2. HUANG consistently delivers superior targeted attack performance across all datasets while maintaining high visual quality. Although DHTA demonstrates the least perceptibility, its extensive iteration requirements reduce efficiency and lack practicality. Both DHTA+FGSM and ProS-GAN produce adversarial images quickly but compromise on attack performance and image quality, failing to adequately deceive the model. HUANG, while slightly more perceptible, significantly outperforms in attack performance. It has 15.96% higher t-MAP than DHTA on average. The visual quality of HUANG's adversarial images is also noteworthy, with the adversarial images appearing nearly identical to benign ones, as shown in Figure 2.

Robustness against adversarial defense methods To validate the effectiveness of HUANG in enhancing the robustness of adversarial images, we evaluated its performance against three adversarial defense methods: JPEG compression (JPEG) (Dziugaite, Ghahramani, and Roy 2016), feature squeezing (FS) (Xu, Evans, and Qi 2018), and randomized smoothing (RS) (Cohen, Rosenfeld, and Kolter 2019).

The robustness against the three defense methods is illustrated in Figure 3. It is evident that previous SOTA methods experience a significant decrease in t-MAP when subjected to the three defense methods, showing the weakest robustness against RS. In contrast, our HUANG maintains high robustness under all three defense methods, with only a slight decrease in t-MAP, indicating minimal impact from the adversarial defenses. The high robustness of HUANG is attributed to the dynamic interplay between denoising and the addition of adversarial perturbations in the adversarial denoising process, allowing the adversarial image to achieve high attack potency while preserving robustness.

Transferability

In adversarial attacks on image hashing retrieval, the concept of cross-hash bit transferability refers to the ability of adversarial images generated for one hash bit configuration to be applicable to other. We conducted experiments on cross-hash bit transferability, and the results are presented in Table 3. It is observable that the adversarial images possess strong robustness across different hash bits, achieving comparable t-MAP values across various configurations. Compared to the previous SOTA model SAAT, HUANG exhibits significantly enhanced cross-hash bit transferability, demonstrating the robustness of our approach.

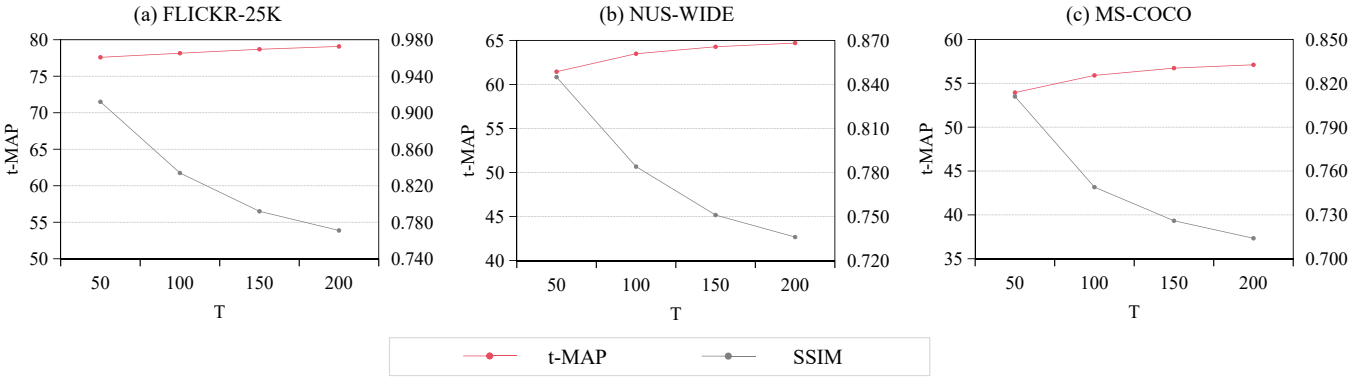


Figure 4: t-MAP and image quality of HUANG on three datasets when varying timestep T

Method	Code length	FLICKR-25K			NUS-WIDE			MS-COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SAAT	16 bits	62.43	58.91	56.48	49.82	46.19	42.83	41.82	39.13	36.76
	32 bits	59.28	63.07	57.86	45.64	51.28	46.44	38.62	45.68	46.80
	64 bits	57.10	58.12	60.02	41.72	45.88	51.72	35.40	44.45	50.61
HUANG	16 bits	71.64	69.86	69.13	61.74	58.73	57.66	52.31	50.94	48.16
	32 bits	69.09	78.15	70.25	59.99	63.49	65.48	51.15	55.90	55.17
	64 bits	68.46	69.25	71.32	59.15	64.79	68.08	50.43	56.67	61.41

Table 3: Transferability comparison result of HUANG and SAAT on DPSH.

Parameter Analysis

Experiments on time steps T use SSIM for image quality evaluation, where lower values indicate better quality. In Figure 4, as T increases, SSIM decreases, this is because larger T results in larger degree of noise, thus increasing denoising complexity. Regarding target attack performance, when T is smaller, the adversarial perturbations are insufficiently introduced, leading to suboptimal transferability of the adversarial examples. However, an excessively large T results in an overabundance of noise, which diminishes the effectiveness of the adversarial disturbance, making it harder for perturbations to manifest within the dynamic interplay.

Ablation Study

To validate the enhancement capabilities of each component module in HUANG for generating adversarial images, we conducted ablation experiments on the aforementioned three datasets. The results are illustrated in Table 4. In this context, TF, TA, AP, RI, and TS represent target fusion, target adaptation, adversarial perturbation, residual image, and target semantic, respectively. It is evident that removing any module from HUANG results in a decline in overall attack performance, underscoring the importance of each module. Specifically, take FLICKR-25K as an example, the t-MAP decreases by 16.41% and 12.30% upon removing AP and TS, respectively, indicating these two modules play a critical role in enhancing the attack potency. Similarly, the TF, TA, and RI modules are also crucial, as they effectively capture the semantic information of the target categories.

	FLICKR-25K	NUS-WIDE	MS-COCO
SAAT	63.07	51.28	45.68
w/o TF	72.55	60.14	52.97
w/o TA	75.40	62.85	54.16
w/o AP	61.74	50.06	41.52
w/o RI	69.42	57.88	47.28
w/o TS	65.85	54.72	45.77
HUANG	**78.15**	**63.49**	**55.90**

Table 4: Ablation results on different module components on different datasets, here w/o means without.

Conclusion

We propose HUANG, a hashing retrieval adversarial attack model for black-box targeted scenarios. HUANG operates in two stages: semantic learning and adversarial generation. Semantic learning fuses features from target label and image to extract target semantics, which guide the diffusion model. Adversarial generation stage introduces adversarial denoising, leveraging perturbations and predicted noise to shift image distributions, producing adversarial images. Supervised by multiple losses, HUANG achieves superior attack performance. Extensive experiments on three datasets demonstrate that HUANG significantly outperforms previous SOTA algorithms, setting a new benchmark in this field. Moreover, due to the dynamic interplay between denoising and adding adversarial perturbations in adversarial denoising, HUANG's robustness is greatly enhanced.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62472226, 62176126, the Natural Science Foundation of Jiangsu Province, China under Grant No. BK20230095.

References

Andoni, A.; and Indyk, P. 2006. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. In *Symposium on Foundations of Computer Science*, 459–468.

Bai, J.; et al. 2020. Targeted Attack for Deep Hashing Based Retrieval. In *Computer Vision - ECCV 2020*, 618–634.

Cao, Y.; Long, M.; Liu, B.; and Wang, J. 2018. Deep Cauchy Hashing for Hamming Space Retrieval. In *Computer Vision and Pattern Recognition*, 1229–1237.

Cao, Z.; Long, M.; Wang, J.; and Yu, P. S. 2017. HashNet: Deep Learning to Hash by Continuation. In *International Conference on Computer Vision*, 5609–5618.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *Symposium on Security and Privacy*, 39–57.

Chen, M.; Lu, J.; Wang, Y.; Qin, J.; and Wang, W. 2021. DAIR: A Query-Efficient Decision-based Attack on Image Retrieval Systems. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1064–1073.

Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; and Hsieh, C.-J. 2017. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. In *ACM Workshop on Artificial Intelligence and Security*, 15–26.

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

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning*, 1310–1320.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat GANs on image synthesis. In *Neural Information Processing Systems*, 8780–8794.

Dziugaite, G. K.; Ghahramani, Z.; and Roy, D. M. 2016. A study of the effect of JPG compression on adversarial images. arXiv:1608.00853.

Fu, S.; Cao, C.; Tao, F.; Zou, B.; Lin, X.; and Sun, J. 2023. LGWAE: Label-Guided Weighted Autoencoder Network for Flexible Targeted Attacks of Deep Hashing. In *International Joint Conference on Neural Networks*, 1–9.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, 6840–6851.

Hu, S.; Zhang, Y.; Liu, X.; Zhang, L. Y.; Li, M.; and Jin, H. 2021. AdvHash: Set-to-set Targeted Attack on Deep Hashing with One Single Adversarial Patch. In *ACM International Conference on Multimedia*, 2335–2343.

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

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

Li, W.-J.; Wang, S.; and Kang, W.-C. 2016. Feature learning based deep supervised hashing with pairwise labels. In *International Joint Conference on Artificial Intelligence*, 1711–1717.

Lin, K.; Lu, J.; Chen, C.-S.; and Zhou, J. 2016. Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks. In *Computer Vision and Pattern Recognition*, 1183–1192.

Lin, T.-Y.; et al. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV*, 740–755.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

Mnih, V.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518: 529–33.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Computer Vision and Pattern Recognition*, 2574–2582.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition*, 427–436.

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

Shen, X.; Dong, G.; Zheng, Y.; Lan, L.; Tsang, I. W.; and Sun, Q.-S. 2022. Deep Co-Image-Label Hashing for Multi-Label Image Retrieval. *IEEE Transactions on Multimedia*, 24: 1116–1126.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems*, 11918–11930.

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

Wang, D.; Cui, P.; Ou, M.; and Zhu, W. 2015. Learning Compact Hash Codes for Multimodal Representations Using Orthogonal Deep Structure. *IEEE Transactions on Multimedia*, 17(9): 1404–1416.

Wang, K.; He, R.; Wang, L.; Wang, W.; and Tan, T. 2016. Joint Feature Selection and Subspace Learning for Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10): 2010–2023.

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

Xia, R.; Pan, Y.; Lai, H.; Liu, C.; and Yan, S. 2014. Supervised hashing for image retrieval via image representation learning. In *AAAI Conference on Artificial Intelligence*, 2156–2162.

Xu, W.; Evans, D.; and Qi, Y. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Network and Distributed System Security Symposium*.

Yang, E.; Liu, T.; Deng, C.; and Tao, D. 2020. Adversarial Examples for Hamming Space Search. *IEEE Transactions on Cybernetics*, 50(4): 1473–1484.

Yuan, L.; et al. 2020. Central Similarity Quantization for Efficient Image and Video Retrieval. In *Computer Vision and Pattern Recognition*, 3080–3089.

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

Zhang, F.; Chowdhury, S. P.; and Christakis, M. 2020. DeepSearch: a simple and effective blackbox attack for deep neural networks. In *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 800–812.

Zhang, R.; Lin, L.; Zhang, R.; Zuo, W.; and Zhang, L. 2015. Bit-Scalable Deep Hashing With Regularized Similarity Learning for Image Retrieval and Person Re-Identification. *IEEE Transactions on Image Processing*, 24(12): 4766–4779.

Zhang, Z.; Wang, X.; Lu, G.; Shen, F.; and Zhu, L. 2022. Targeted Attack of Deep Hashing Via Prototype-Supervised Adversarial Networks. *IEEE Transactions on Multimedia*, 24: 3392–3404.

Zhao, W.; Song, J.; Yuan, S.; Gao, L.; Yang, Y.; and Shen, H. 2023. Precise Target-Oriented Attack against Deep Hashing-based Retrieval. In *ACM International Conference on Multimedia*, 6379–6389.

Zhu, F.; Zhang, W.; Wu, D.; Wang, L.; Li, B.; and Wang, W. 2024a. Exploring Targeted Universal Adversarial Attack for Deep Hashing. In *International Conference on Acoustics, Speech and Signal Processing*, 3335–3339.

Zhu, F.; Zhang, W.; Wu, D.; Wang, L.; Li, B.; and Wang, W. 2024b. Targeted Transferable Attack against Deep Hashing Retrieval. In *ACM International Conference on Multimedia in Asia*, 1–7.