Unrestricted Targeted Deep Hashing Attack via Contrastive Latent Diffusion

Fan Yang[1] , **Chuan Ma**[2] , **Yuhui Zheng**[3] , **Xiaobo Shen**[1,4*] and **Joey Tianyi Zhou**[5,6]
[1]Nanjing University of Science and Technology, China
[2]Chongqing University, China
[3]Qinghai Normal University, China
[4]State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, China
[5]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[6]Institute of High Performance Computing (IHPC), A*STAR, Singapore
fyang@njust.edu.cn, chuan.ma@cqu.edu.cn, zhengyh@vip.126.com, njust.shenxiaobo@gmail.com, joey.tianyi.zhou@gmail.com

Abstract

Deep hashing is widely used for large-scale image retrieval but remains vulnerable to adversarial examples, raising practical security concerns. Existing targeted adversarial attacks on deep hashing typically rely on ℓ_p-norm constrained perturbations, which struggle to balance attack effectiveness and imperceptibility, often requiring perceptible noise and limiting their practicality in real-world retrieval scenarios. We propose UTDHA, the first unrestricted targeted attack for deep hashing models using contrastive-guided latent diffusion. UTDHA generates adversarial examples with a latent diffusion model and performs optimization in the latent space rather than the pixel space, enabling semantic manipulation while preserving image naturalness. Through contrastive guidance, the attack pulls adversarial examples toward the target label while pushing them away from non-target labels. Meanwhile, UTDHA enforces structural and perceptual consistency, producing adversarial examples that are both imperceptible and visually natural. Extensive experiments on three benchmarks demonstrate that UTDHA outperforms existing targeted adversarial attack baselines for deep hashing models in both attack effectiveness and imperceptibility.

1 Introduction

With the rapid expansion of multimedia data, there is an urgent demand for efficient large-scale image retrieval techniques. Hashing [Wang *et al.*, 2016] reduces the storage and computational cost by mapping high-dimensional data to compact binary codes, making it highly effective for large-scale retrieval. Leveraging the powerful feature representation and nonlinear modeling capabilities of deep learning, deep hashing [Liu *et al.*, 2016; Yuan *et al.*, 2020] employs deep neural networks (DNNs) to learn nonlinear hashing functions, thereby achieving superior performance compared to shallow hashing methods.

(a) Original (b) ProS-GAN (c) SAAT (d) UTDHA

Figure 1: The adversarial perturbation patterns generated by different attacks. The targeted mean average precision (t-MAP) in the four cases are (a) 42.60%, (b) 73.03%, (c) 81.80%, (d) 85.55%. The higher t-MAP indicates more successful targeted attacks.

Recent studies have shown that DNNs are vulnerable to adversarial examples [Goodfellow *et al.*, 2015], where inputs modified with imperceptible perturbations can mislead the model into making incorrect predictions. Deep hashing models inevitably inherit the adversarial vulnerability of DNNs [Bai *et al.*, 2020] that raises serious security concerns and degrades retrieval performance. Adversarial attacks on deep hashing models are commonly divided into untargeted attacks [Yang *et al.*, 2020] and targeted attacks [Bai *et al.*, 2020; Wang *et al.*, 2021; Meng *et al.*, 2024]. Untargeted attacks aim to make the retrieved results semantically irrelevant to the query, whereas targeted attacks manipulate the model to retrieve results from a specific target class. Compared to untargeted attacks, targeted attacks pose greater threats. For example, a deliberately perturbed image of a cat may cause the retrieval system to return violent results. Therefore, investigating targeted adversarial attacks on deep hashing models is essential to enhance their robustness and security.

Existing targeted adversarial attacks on deep hashing models typically impose ℓ_p-norm constraints in pixel space to ensure imperceptibility. However, such constraints often cause adversarial examples to deviate from the natural distribution [Hosseini and Poovendran, 2018], resulting in a trade-off between attack performance and visual imperceptibility. As illustrated in Figure 1, the attacks in (b) and (c) achieve good attack performance, but inevitably introduce

*Corresponding author.

visible perturbations, compromising imperceptibility. Consequently, unrestricted adversarial attacks [Yuan *et al.*, 2022; Xue *et al.*, 2023] have emerged, relaxing pixel-level constraints and instead modifying semantic attributes such as color and texture to synthesize visually natural adversarial examples. However, prior work has focused mainly on classification tasks, leaving unrestricted adversarial attacks on deep hashing retrieval largely unexplored.

Diffusion models [Ho *et al.*, 2020] have emerged as the new state-of-the-art family of deep generative models. Diffusion models progressively destruct data by injecting noise, then learn to reverse this process for sample generation. Latent Diffusion Models (LDMs) [Rombach *et al.*, 2022] perform denoising in a learned latent space, enabling more efficient computation and higher-quality image generation. Leveraging the superiority of LDMs, we propose the first **U**nrestricted **T**argeted **D**eep **H**ashing **A**ttack (UTDHA). We adopt DDIM inversion [Mokady *et al.*, 2023] to project the original sample into the latent space of a pretrained LDM, where adversarial optimization is performed. A contrastive mechanism is introduced to exploit the structured semantic similarity encoded in hash codes and guide the optimization. Figure 2 provides an overview of the proposed UTDHA. The main contributions of this work are as follows:

- We propose the first unrestricted targeted attack for deep hashing models, dubbed UTDHA, which leverages LDMs to synthesize imperceptible adversarial examples. The proposed UTDHA is optimized in the latent space of LDMs instead of pixel space to generate visually natural adversarial examples.
- A contrastive guidance mechanism is introduced to exploit the semantic structure encoded in hash codes, guiding adversarial examples toward target semantics for effective targeted attacks.
- Extensive experiments demonstrate that the proposed method achieves superior attack performance and generates visually imperceptible adversarial examples compared to existing targeted adversarial attacks on deep hashing retrieval.

2 Related Work

2.1 Deep Hashing-based Retrieval

Deep hashing methods leverage DNNs to learn hash functions that transform data into compact binary codes, improving retrieval efficiency and reducing storage costs. Existing deep hashing methods are generally divided into unsupervised [Tu *et al.*, 2020] and supervised methods [Li *et al.*, 2017; Jiang and Li, 2018], depending on whether label supervision is employed during training. Unsupervised deep hashing usually relies on semantic cues inferred from the data itself or extracted using pretrained models [Peng *et al.*, 2026]. Supervised deep hashing leverages label information to explicitly encode semantic similarity in the Hamming space [Wang *et al.*, 2026a; Wang *et al.*, 2026b]. For example, the first supervised deep hashing method CNNH [Xia *et al.*, 2014] employs a two-stage framework where semantic labels guide the learning of hash codes after feature extraction. DPSH [Li *et al.*, 2015] introduces an end-to-end framework that jointly learns features and hash codes from pairwise labels. CSQ [Yuan *et al.*, 2020] learns hash codes by aligning them with predefined semantic centers using center-based similarity and quantization losses. In this work, we investigate unrestricted targeted attacks on supervised deep hashing.

2.2 Adversarial Attacks on Deep Hashing

The vulnerability of DNNs has been revealed [Szegedy *et al.*, 2014], and adversarial attacks have been extensively investigated to expose the vulnerability of classification models and improve their robustness. Compared to adversarial attacks on classification models, adversarial attacks on deep hashing-based retrieval are rarely explored. Existing adversarial attack methods on deep hashing models can be broadly categorized into untargeted and targeted attacks. HAG [Yang *et al.*, 2020] is the first method for untargeted adversarial attacks on deep hashing, aiming to generate adversarial examples that mislead retrieval systems toward irrelevant samples. [Bai *et al.*, 2020] introduces DHTA that constructs an anchor code to represent the target class through different heuristic strategies and use it to guide optimization of adversarial examples. ProS-GAN [Wang *et al.*, 2021] first uses an auxiliary network to produce a prototype code for the target class, which then guides the generator to craft adversarial examples. SAAT [Yuan *et al.*, 2023] generates a mainstay code representing the target class through mainstay feature learning, which is then used to guide the synthesis of adversarial examples. HUANG [Huang and Shen, 2025] is a diffusion-based targeted attack method that constructs a prototype code using an auxiliary network, and generates adversarial examples by progressively adding perturbations in the pixel space. In contrast, our method performs attack optimization directly in the latent space of diffusion models and synthesizes adversarial examples without explicitly adding pixel-level perturbations.

2.3 Unrestricted Attacks

Conventional ℓ_p-norm-based adversarial attacks add perturbations in the pixel space, often pushing adversarial examples away from the natural image distribution and making them perceptible [Hosseini and Poovendran, 2018]. In contrast, unrestricted attacks relax pixel-level constraints and achieve imperceptibility by introducing semantic modifications to the original samples. Early unrestricted attack methods manipulate a single semantic attribute of the sample, such as texture and color. For instance, SemanticAdv [Qiu *et al.*, 2020] modifies texture attributes via semantic image editing to generate realistic adversarial examples. Natural Color Fool [Yuan *et al.*, 2022] attacks by altering color semantics using a learned color distribution transformation. Recently, diffusion models [Ho *et al.*, 2020] have been leveraged for their powerful generative capabilities to produce natural adversarial examples. AdvDiffuser [Chen *et al.*, 2023a] applies PGD to perturb diffusion outputs, producing high-quality unrestricted adversarial samples. ACA [Chen *et al.*, 2023b] generates visually realistic examples by editing images on a natural image manifold using Stable Diffusion [Rombach *et al.*, 2022]. However, these methods primarily target classification tasks,

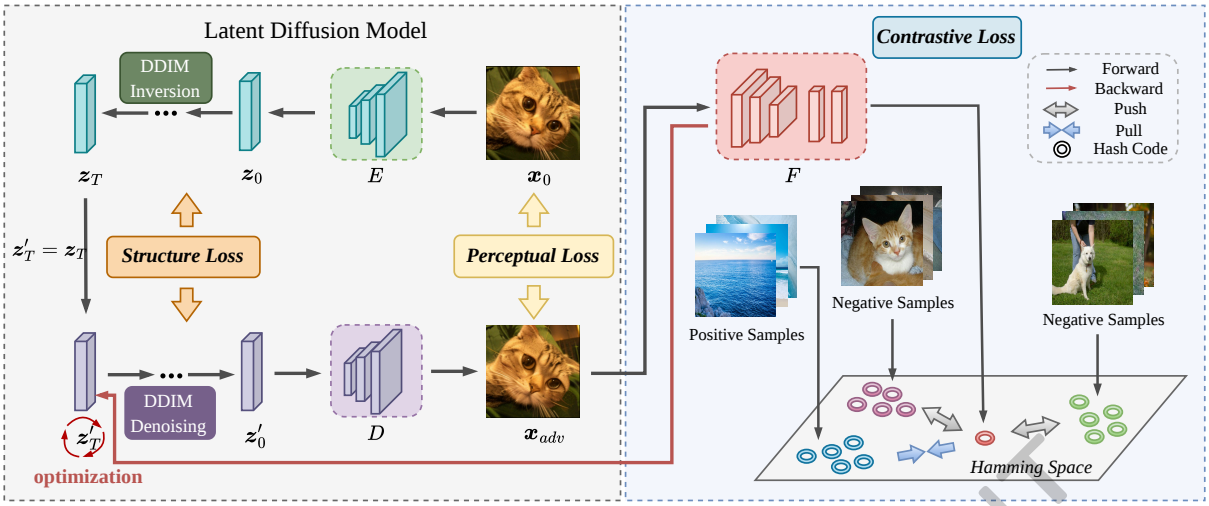


Figure 2: Overview of the proposed UTDHA. The original sample $\boldsymbol{x}_0$ is encoded into an initial latent representation $\boldsymbol{z}_0$ by the VAE encoder $E(\cdot)$. DDIM inversion then transforms $\boldsymbol{z}_0$ into an optimizable latent variable $\boldsymbol{z}'_T$. This latent variable is subsequently optimized via contrastive guidance that leverages hash code semantics to align it with the target label. Additionally, structure and perceptual preservation losses are employed to maintain visual consistency. Finally, the adversarial example $\boldsymbol{x}_{adv}$ is synthesized through denoising and decoding by the VAE decoder $D(\cdot)$ to attack the target hashing model $F(\cdot)$.

and unrestricted attacks on deep hashing retrieval remain largely unexplored. Deep hashing learns semantic similarity and outputs discrete binary codes rather than categories, thus classification-oriented attack methods are difficult to directly transfer.

3 Problem Definition

Let $\mathcal{O} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ denote a dataset of N instances labeled with P classes, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i \in \{0,1\}^P$ indicate the i-th original sample and its label vector, respectively. Given $\boldsymbol{x}_i$, its K-bit hash code $\boldsymbol{b}_i$ can be generated by a deep hashing model $F(\cdot)$ following:

$$\boldsymbol{b}_i = F(\boldsymbol{x}_i) = \text{sign}(f_\theta(\boldsymbol{x}_i)), \quad \text{s.t.} \quad \boldsymbol{b}_i \in \{-1,+1\}^K, \tag{1}$$

where $f_\theta(\cdot)$ is a DNN that approximates $F(\cdot)$, and the sign function $\text{sign}(\cdot)$ is applied to the output of $f_\theta(\boldsymbol{x}_i)$ to generate the hash code $\boldsymbol{b}_i$.

In this work, we explore unrestricted targeted attacks on deep hashing models. Specifically, given a benign query sample $\boldsymbol{x}_0$ and a target label $\boldsymbol{y}_t$, an unrestricted targeted attack aims to craft a visually realistic adversarial example $\boldsymbol{x}_{adv}$ that misleads the target hashing model to retrieve results belonging to $\boldsymbol{y}_t$, without ℓ_p-norm constrained perturbations.

4 Methodology

This section details the proposed UTDHA, the first unrestricted targeted attack method for deep hashing models, as shown in Figure 2.

4.1 Unrestricted Attack Framework

Existing targeted attacks mainly operate under ℓ_p-norm constraints in pixel space, and often lead to perceptible perturbations. In contrast, the proposed UTDHA performs adversarial optimization directly in the latent space of a pretrained LDM, producing imperceptible adversarial examples without pixel-space norm constraints.

Specifically, given a benign sample $\boldsymbol{x}_0$, we first obtain an initial latent representation $\boldsymbol{z}_0$ by the VAE [Kingma and Welling, 2014] encoder: $\boldsymbol{z}_0 = E(\boldsymbol{x}_0)$. We then apply DDIM inversion [Mokady *et al.*, 2023] for T steps, transforming $\boldsymbol{z}_0$ into $\boldsymbol{z}_T$. At each timestep t, $\boldsymbol{z}_t$ is inverted to $\boldsymbol{z}_{t+1}$ as follows:

$$\begin{aligned}\boldsymbol{z}_{t+1} = \text{Inversion}(\boldsymbol{z}_t) = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\boldsymbol{z}_t \\ + \sqrt{\alpha_{t+1}}\left(\sqrt{\frac{1}{\alpha_{t+1}}-1} - \sqrt{\frac{1}{\alpha_t}-1}\right)\cdot \epsilon_\theta(\boldsymbol{z}_t, t),\end{aligned} \tag{2}$$

where α_t is a hyper-parameter controlling the level of noise and $\epsilon_\theta(\boldsymbol{z}_t, t)$ is the U-Net model [Song *et al.*, 2021] that predicts the noise added at step t. The latent variable $\boldsymbol{z}'_T$ is initialized from $\boldsymbol{z}_T$ as $\boldsymbol{z}'_T = \boldsymbol{z}_T$. Although directly denoising $\boldsymbol{z}'_T$ produces visually natural examples, they retain the original semantics and fail to achieve targeted attacks. Therefore, we perform adversarial latent optimization on $\boldsymbol{z}'_T$ to guide its semantics toward the target label. After adversarial latent optimization, we denoise the perturbed latent variable $\boldsymbol{z}'_T$ to obtain $\boldsymbol{z}'_0$ using:

$$\begin{aligned}\boldsymbol{z}_{t-1} = \text{Denoise}(\boldsymbol{z}_t) = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\boldsymbol{z}_t \\ + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}}-1} - \sqrt{\frac{1}{\alpha_t}-1}\right)\cdot \epsilon_\theta(\boldsymbol{z}_t, t).\end{aligned} \tag{3}$$

Finally, the adversarial example is produced by decoding $\boldsymbol{z}'_0$ through the VAE decoder: $\boldsymbol{x}_{adv} = D(\boldsymbol{z}'_0)$.

4.2 Adversarial Latent Optimization

By optimizing the latent variable $\boldsymbol{z}'_T$ that encodes high-level semantics, we efficiently synthesize natural adversarial exam-

ples. To achieve effective targeted attacks, we employ contrastive guidance throughout the latent optimization process. Meanwhile, we incorporate structure and perceptual losses to ensure visual consistency and imperceptibility.

Contrastive Guidance

We introduce a contrastive guidance mechanism to align the semantics of $\boldsymbol{z}'_T$ with the target class y_t. It exploits semantic similarity in the Hamming space, enabling unrestricted targeted attacks for deep hashing models. Specifically, we minimize the distance between the hash code of the adversarial example $\boldsymbol{x}_{adv}$ and those of semantic-relevant samples $\boldsymbol{x}_i^{(t)}$ with the target label $\boldsymbol{y}_t$, while maximizing the distance from unrelated samples $\boldsymbol{x}_j^{(n)}$ belonging to other categories. The objective is defined as:

$$\min \sum_{i\in\mathcal{P}} d\big(F(\boldsymbol{x}_{adv}), F(\boldsymbol{x}_i^{(t)})\big) - \sum_{j\in\mathcal{N}} d\big(F(\boldsymbol{x}_{adv}), F(\boldsymbol{x}_j^{(n)})\big), \tag{4}$$

where $\boldsymbol{x}_{adv} = D(\boldsymbol{z}'_0) = D(\underbrace{\text{Denoise} \circ \cdots \circ \text{Denoise}}_{T}(\boldsymbol{z}'_T))$, $d(\cdot,\cdot)$ denotes the Hamming distance metric and $D(\cdot)$ represents the VAE decoder. Here, $\mathcal{P}$ and $\mathcal{N}$ denote the positive and negative sets, consisting of samples semantically related and unrelated to the adversarial example $\boldsymbol{x}_{adv}$, respectively. However, directly optimizing this objective may result in suboptimal attack performance, particularly under imbalanced similarity distributions between positive and negative pairs. Inspired by [Sun *et al.*, 2020], we formulate a weighted contrastive loss that dynamically emphasizes more informative sample pairs, and define the following contrastive loss:

$$\mathcal{L}_c = \log\Big(\sum_{i\in\mathcal{P}} e^{(\boldsymbol{\Theta}_{ti}-m)\cdot\boldsymbol{\Theta}_{ti})}\Big) + \log\Big(\sum_{j\in\mathcal{N}} e^{(\boldsymbol{\Theta}_{tj}+m)\cdot\boldsymbol{\Theta}_{tj})}\Big), \tag{5}$$

where $\boldsymbol{b}_t$ and $\boldsymbol{b}_i$ denote the hash codes of the adversarial sample and the i-th positive sample, respectively, $\boldsymbol{\Theta}_{ti} = \frac{1}{K}\boldsymbol{b}_t^\top \boldsymbol{b}_i$ indicates the similarity between the adversarial sample and the i-th positive sample. The weight factor m is used to balance the numbers of $\mathcal{P}$ and $\mathcal{N}$.

Structural Consistency Preservation

Although contrastive guidance effectively optimizes the latent variable, it may excessively alter high-level semantics, leading to noticeable structural and perceptual deviations from the original sample. LDMs commonly leverage the self-attention mechanism [Vaswani *et al.*, 2017] in U-Net during denoising to preserve structural consistency, as it preserves global structure [Chen *et al.*, 2024]. Inspired by this, we introduce self-attention guidance into latent optimization to maintain structural consistency. Specifically, we first extract the self-attention maps from the benign latent $\boldsymbol{z}_t$ without perturbation. During adversarial latent optimization, the self-attention map of each adversarial latent $\boldsymbol{z}'_t$ is encouraged to closely align with the corresponding map of $\boldsymbol{z}_t$ using the following loss:

$$\mathcal{L}_s = \sum_{t=1}^{T} \|S(\boldsymbol{z}'_t) - S(\boldsymbol{z}_t)\|_2^2, \tag{6}$$

where $S(\cdot)$ computes the self-attention map of the latent.

Perceptual Consistency Preservation

To further preserve semantic similarity between adversarial and original samples, we incorporate a perceptual loss derived from the Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018]. LPIPS measures perceptual differences in a deep feature space rather than pixel space, effectively capturing crucial texture and structure for human vision. Concretely, we decode the final denoised latent $\boldsymbol{z}'_0$ to obtain the adversarial example $\boldsymbol{x}_{adv}$. Then $\boldsymbol{x}_{adv}$ and $\boldsymbol{x}_\mathbf{0}$ are fed into a pretrained deep network $\phi(\cdot)$ to obtain deep features. The perceptual loss is then defined by measuring the distance between the extracted deep features as follows:

$$\mathcal{L}_p = \sum_{q\in\mathcal{Q}} \frac{1}{U_q W_q C_q} \|\phi_q(\boldsymbol{x}_{adv}) - \phi_q(\boldsymbol{x}_0)\|_2^2, \tag{7}$$

where $\mathcal{Q}$ is the set of layers of $\phi(\cdot)$, and $\phi_q(\cdot)$ denotes the feature extractor at layer q with spatial dimensions $U_q \times W_q$ and C_q channels.

Final Objective Function

The final objective function includes adversarial contrastive loss, structure, and perceptual losses to ensure both attack effectiveness and visual consistency. The final objective function is defined as:

$$\min \mathcal{L} = \mathcal{L}_c + \alpha\mathcal{L}_s + \beta\mathcal{L}_p. \tag{8}$$

where the hyper-parameters α and β weight the relative importance of each loss.

5 Experiments

5.1 Experimental Setup

Datasets

We conduct experiments on three commonly used datasets including FLICKR-25K [Huiskes and Lew, 2008], NUS-WIDE [Chua *et al.*, 2009], and MS-COCO [Lin *et al.*, 2014].

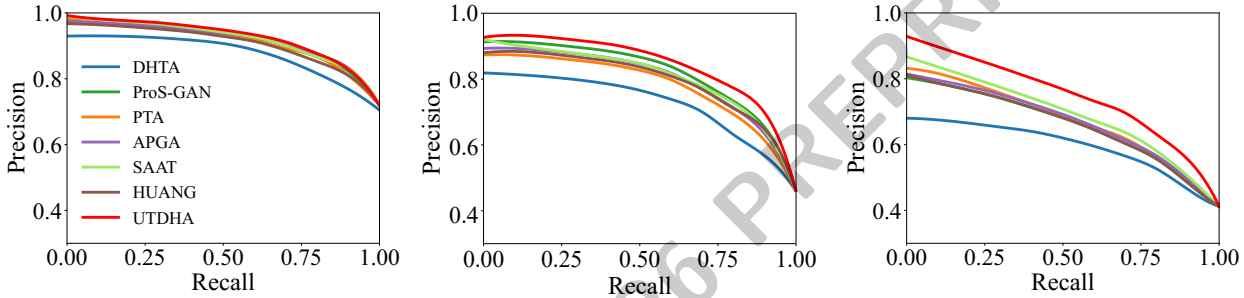
- FLICKR-25K contains 25,000 images labeled with 38 unique labels. We randomly select 1,700 images as the query set and the rest as the retrieval set, where 5,000 randomly selected images are used for training.
- NUS-WIDE consists of 269,648 images in 81 categories. We only select 195,834 images annotated with the 21 most frequent labels. Among them, 2,100 images are randomly sampled as queries, 10,500 as the training set, and the remaining images are used as the retrieval set.
- MS-COCO comprises 123,287 images labeled across 80 categories, from which 5,000 images are randomly sampled as the query set. The rest of the images are used as the retrieval set and 10,000 images randomly sampled from the retrieval set are used as the training set.

Metrics

Following [Bai *et al.*, 2020], we adopt targeted mean average precision (t-MAP) and precision-recall (PR) curves as evaluation metrics to evaluate the effectiveness of targeted attacks. The t-MAP is calculated based on the top 5,000

Method	Venue	FLICKR-25K				NUS-WIDE				MS-COCO			
		16 bit	32 bit	48 bit	64 bit	16 bit	32 bit	48 bit	64 bit	16 bit	32 bit	48 bit	64 bit
Original	—	69.35	69.32	69.19	68.80	57.66	58.63	58.78	59.43	43.23	43.11	43.54	42.60
Noise	—	69.12	68.81	69.05	68.68	57.39	58.41	58.24	59.22	43.02	42.89	43.27	42.53
DHTA	ECCV '20	94.40	94.51	94.83	94.54	84.31	84.57	82.92	77.41	64.30	63.97	64.26	63.78
ProS-GAN	CVPR '21	96.88	97.24	97.27	96.69	85.43	86.11	86.48	89.27	76.16	75.91	76.17	76.03
PTA	MM '23	97.15	98.29	97.68	97.14	85.79	85.75	84.44	85.73	78.62	78.44	79.12	79.95
SAAT	TIFS '23	96.98	97.57	98.02	97.43	86.94	88.16	87.27	88.54	78.56	78.70	80.77	81.80
APGA	TIFS '25	96.14	97.83	96.74	96.97	85.92	86.99	86.21	87.60	77.60	76.85	77.45	77.31
HUANG	AAAI '25	96.12	96.61	96.77	96.42	85.75	85.84	86.30	86.93	76.22	76.47	77.34	76.29
UTDHA	Ours	**98.17**	**98.67**	**98.34**	**97.88**	**90.52**	**89.61**	**91.83**	**92.32**	**81.06**	**84.12**	**85.09**	**85.55**

Table 1: t-MAP (%) of all the targeted attack methods with respect to different hash bits on three benchmark datasets. The bold indicates best.



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

Figure 3: Precision-Recall Curves of all the targeted attack methods on three benchmark datasets under 64 bits code length.

retrieved images, and a higher t-MAP indicates more effective targeted attack performance. In addition, we consider two reference-based metrics, i.e., LPIPS [Zhang *et al.*, 2018], FID [Heusel *et al.*, 2017], and one no-reference metric, i.e., BRISQUE [Mittal *et al.*, 2011] to evaluate the visual quality of adversarial examples. LPIPS and FID indicate visual similarity in terms of semantics and distribution, respectively, and BRISQUE measures perceptual quality by detecting deviations from natural scene statistics. Lower LPIPS, FID, and BRISQUE scores indicate better visual quality of adversarial examples.

Baselines

As there have been no unrestricted targeted attack baselines for deep hashing models, we compare the proposed UTDHA with state-of-the-art targeted adversarial attacks on deep hashing models, including DHTA [Bai *et al.*, 2020], ProS-GAN [Wang *et al.*, 2021], PTA [Zhao *et al.*, 2023], SAAT [Yuan *et al.*, 2023], APGA [Tu *et al.*, 2025] and HUANG [Huang and Shen, 2025]. These attack methods are conducted on three benchmark hashing models, including DPSH [Li *et al.*, 2015], HashNet [Cao *et al.*, 2017], and CSQ [Yuan *et al.*, 2020], where DPSH is the primary target deep hashing model and ResNet50 is the default backbone.

Implementation Details

For the proposed UTDHA, we utilize the pretrained Stable Diffusion [Rombach *et al.*, 2022] with a DDIM sampler [Song *et al.*, 2021]. The number of DDIM sampling steps T is set to 30. We optimize the latent variable using the AdamW optimizer [Loshchilov and Hutter, 2019], with a learning rate of $1e^{-4}$ and a total of 50 iterations. The hyper-parameters α and β are set to 10 and 5, respectively. The weight factor m is set to 3. For targeted attacks, we randomly select a label that differs from the original class as the target label. Following [Wang *et al.*, 2021], the perturbation budget for all ℓ_p-norm-based baselines is set to $8/255$. All experiments are conducted using PyTorch and run on an NVIDIA A6000 GPU.

5.2 Results

Attack Performance

Table 1 reports the t-MAPs of different targeted attack methods on the three benchmark datasets. Table 1 also reports the t-MAPs of querying with original and noisy samples as simple baselines, where noisy samples are obtained by adding noise sampled from a uniform distribution to benign samples. Figure 3 illustrates the PR curves of different attack methods

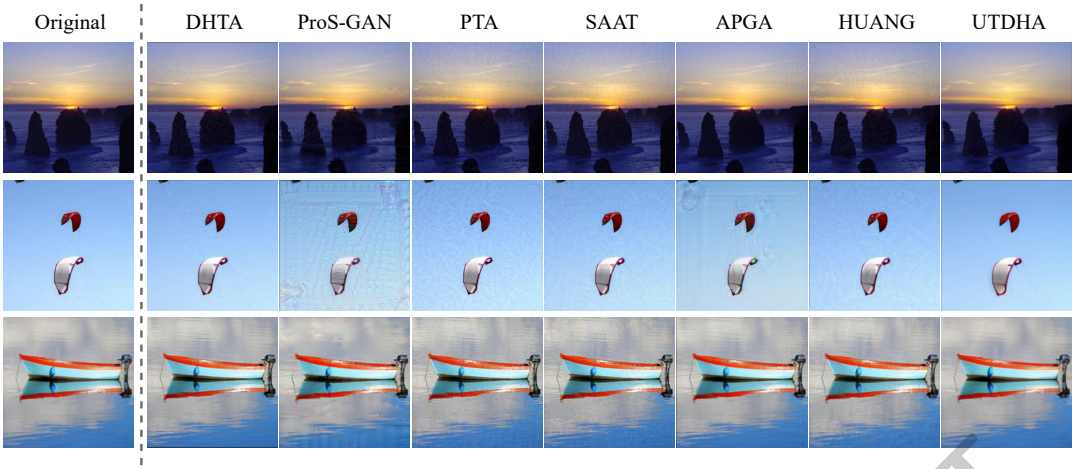


Figure 4: Qualitative visual quality comparisons of adversarial examples generated by different targeted attack methods on NUS-WIDE under 64 bits code length.

Method	t-MAP↑	LPIPS↓	FID↓	BRISQUE↓
DHTA	77.41	**0.044**	**17.53**	23.29
ProS-GAN	89.27	0.097	38.87	24.18
PTA	85.73	0.134	43.65	48.66
SAAT	88.54	0.113	36.04	31.07
APGA	87.60	0.081	35.98	23.84
HUANG	86.93	0.106	39.02	24.49
UTDHA	**92.32**	0.065	18.61	**22.96**

Table 2: Quantitative visual quality comparisons of adversarial examples generated by different targeted attack methods on NUS-WIDE under 64 bits code length. The bold indicates best.

on the three benchmark datasets. We can find the following interesting observations:

- The proposed method consistently achieves the highest t-MAPs among all 12 cases. For example, the proposed method outperforms the best baseline, i.e., SAAT, on average by 0.61%, 2.67%, and 3.99% on FLICKR-25K, NUS-WIDE, and MS-COCO, respectively. In addition, our method significantly outperforms HUANG, a diffusion-based baseline, improving t-MAP by an average of 4.68%. The PR curves of the proposed method are generally above those of the baselines. These results obviously demonstrate the superiority of the proposed method in targeted attacks for deep hashing models.
- Among the baselines, SAAT achieves the best average attack performance, as it may better capture the semantics of the target classes. HUANG, APGA, ProS-GAN, and PTA perform slightly worse, while DHTA shows the weakest results, possibly due to limited semantic alignment with target classes.
- The t-MAPs of querying noisy samples are nearly the same as t-MAPs of querying with original samples across three benchmark datasets, clearly indicating that simple random noise cannot bias the retrieval results of deep hashing models.

Visual Quality

This section conducts quantitative and qualitative comparisons of the visual quality of adversarial examples generated by different attack methods. Table 2 reports the quantitative results of adversarial examples on NUS-WIDE with 64-bit hash codes, evaluated using three visual quality metrics, i.e., LPIPS, FID and BRISQUE. As can be clearly observed, our method is the best in terms of t-MAP and BRISQUE, while ranking second on LPIPS and FID, showing effective attack performance while maintaining high visual quality. Although DHTA achieves better image quality metrics, this is likely due to their limited perturbation magnitude, which prioritizes imperceptibility at the cost of reduced targeted attack performance. Compared to ProS-GAN, our method achieves lower LPIPS, FID, and BRISQUE scores, indicating superior visual quality. This may benefit from the powerful generative capabilities of Stable Diffusion, as well as the preservation of structural and perceptual consistency.

Due to the limitations of quantitative metrics in fully capturing visual consistency, we performed a qualitative analysis, as illustrated in Figure 4. Compared to previous methods that add ℓ_p-norm perturbations to original samples, our approach modifies the semantic content, enabling the synthesis of more natural and imperceptible adversarial examples.

Universal Attack

This section empirically evaluates the universality of different targeted attack methods by applying them to more hashing models. We apply the attack methods to attack two representative hashing methods, i.e., HashNet and CSQ, and report t-MAPs on NUS-WIDE with 64 bits in Table 3. As can be observed, the proposed method achieves the highest t-MAPs in 5 out of 6 cases, and generally outperforms the existing attack baselines. This study empirically verifies the superiority of universality of the proposed method.

Method	FLICKR-25K		NUS-WIDE		MS-COCO	
	HashNet	CSQ	HashNet	CSQ	HashNet	CSQ
Original	69.45	68.49	58.38	58.54	43.38	49.07
DHTA	93.08	92.43	87.00	83.37	69.30	78.41
ProS-GAN	97.02	97.71	86.31	86.75	71.76	84.51
PTA	91.49	92.23	83.34	84.28	73.53	86.14
SAAT	97.75	98.46	87.38	86.46	75.09	**89.28**
APGA	97.30	97.88	86.27	86.93	72.41	87.74
HUANG	96.81	96.25	85.14	85.22	71.48	84.29
UTDHA	**98.63**	**98.59**	**89.98**	**90.09**	**78.28**	89.16

Table 3: t-MAP (%) of different targeted adversarial attacks on HashNet and CSQ hashing methods under 64 bits code length. The bold indicates best.

Method	VGG11	VGG16	AlexNet	ResNet18	ResNet50*
DHTA	54.07	54.53	54.64	54.95	77.41*
ProS-GAN	66.28	67.30	68.79	69.47	89.27*
PTA	68.83	69.06	67.85	67.12	85.73*
SAAT	59.41	59.72	59.92	60.39	88.54*
APGA	67.96	68.36	68.10	68.44	87.60*
HUANG	60.54	60.01	60.42	61.68	86.93*
UTDHA	**70.48**	**71.06**	**71.63**	**71.75**	**92.32***

Table 4: Transfer t-MAPs of different targeted attack methods on NUS-WIDE under 64 bits code length. The white-box attacks are represented with *. The bold indicates best.

5.3 Further Analysis

Transferability

This section empirically evaluates the black-box transferability of different targeted attack methods, where adversarial examples are first generated by one model and then applied to attack another model. Specifically, we generate adversarial examples using various targeted attack methods to attack DPSH with a ResNet50 backbone, and apply these generated adversarial examples to attack DPSH with AlexNet, ResNet18, VGG11, and VGG16 backbones. Table 4 reports the t-MAPs of different methods on NUS-WIDE with 64 bits. As can be observed, the proposed UTDHA has higher transfer t-MAPs than the baselines on the other four backbone networks, verifying its robustness in black-box transferability.

Ablation Study

This section conducts ablation studies and empirically evaluates the effectiveness of each component in the proposed UTDHA. We compare the proposed UTDHA with three variants, each omitting one of the three losses, and report the attack performance and visual quality of these methods on NUS-WIDE in Table 5. From Table 5, we observe that the removal of $\mathcal{L}_c$, $\mathcal{L}_s$, and $\mathcal{L}_p$ results in decreases of 21.25%, 0.35%, and 0.13% in t-MAP, respectively. It demonstrates that contrastive loss $\mathcal{L}_c$ in UTDHA is crucial for better targeted attack performance. Removing $\mathcal{L}_c$ leads to lower LPIPS and FID scores, indicating that adversarial examples become overly similar to the original samples due to the lack of contrastive guidance, thus losing their adversarial properties. In addition, $\mathcal{L}_s$ and $\mathcal{L}_p$ slightly enhance attack performance and significantly improve visual quality.

	t-MAP↑	LPIPS↓	FID↓	BRISQUE↓
w/o $\mathcal{L}_c$	71.07	**0.041**	**6.81**	24.97
w/o $\mathcal{L}_s$	91.97	0.069	27.47	24.28
w/o $\mathcal{L}_p$	92.19	0.094	33.28	23.04
UTDHA	**92.32**	0.065	18.61	**22.96**

Table 5: The performance of the proposed method and its three variants on NUS-WIDE under 64 bits code length. The bold indicates best.

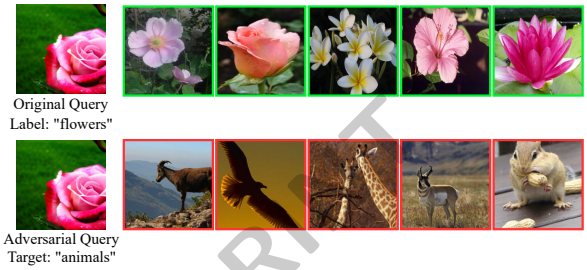


Figure 5: Top-5 retrieved results by querying a normal image and its adversarial example generated by the proposed method on NUS-WIDE under 64 bits code length.

Case Study

This section presents a case study and illustrates the Top-5 retrieved results of a randomly selected image on NUS-WIDE and its adversarial example generated by the proposed method with the target label of animals. As can be clearly observed, the attack is successful as the retrieved results by querying the adversarial image all belong to animal category that is consistent with the given target label. In addition, the generated adversarial example is natural and visually similar to the original example. This experiment qualitatively demonstrates the superiority of the proposed method in unrestricted targeted attack on deep hashing models.

6 Conclusion

In this paper, we observe that existing targeted adversarial attacks on deep hashing typically apply ℓ_p-norm constrained perturbations, which inevitably introduce visible noise to ensure attack effectiveness. To overcome this limitation, we propose UTDHA, the first unrestricted targeted attack on deep hashing retrieval. Instead of perturbing pixels, our method operates in the latent space of a latent diffusion model. This allows for semantic-level modifications, generating more natural adversarial examples. We leverage contrastive guidance to achieve effective targeted attack performance. To preserve the visual consistency of adversarial examples, we use the structure and perceptual losses. Empirical studies confirm the effectiveness of UTDHA.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62472226, 62572083, also supported by the Japan Science and Technology Agency (JST) and the Agency for Science, Technology and Research (A*STAR) under the Japan-Singapore Joint Call (Project No. R24I6IR133) and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2024-033). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore and Infocomm Media Development Authority.

References

[Bai *et al.*, 2020] Jiawang Bai, Bin Chen, Yiming Li, Dongxian Wu, Weiwei Guo, Shu-Tao Xia, and En-Hui Yang. Targeted attack for deep hashing based retrieval. In *Proceedings of European Conference on Computer Vision*, pages 618–634, 2020.

[Cao *et al.*, 2017] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Hashnet: Deep learning to hash by continuation. In *Proceedings of International Conference on Computer Vision*, pages 5609–5618, 2017.

[Chen *et al.*, 2023a] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of International Conference on Computer Vision*, pages 4539–4549, 2023.

[Chen *et al.*, 2023b] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

[Chen *et al.*, 2024] Jianqi Chen, Hao Wei Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhen Xia Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:961–977, 2024.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2009.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of International Conference on Learning Representations*, 2015.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6629–6640, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[Hosseini and Poovendran, 2018] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018.

[Huang and Shen, 2025] Chihan Huang and Xiaobo Shen. Huang: A robust diffusion model-based targeted adversarial attack against deep hashing retrieval. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3626–3634, 2025.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Multimedia Information Retrieval*, pages 39–43, 2008.

[Jiang and Li, 2018] Qing-Yuan Jiang and Wu-Jun Li. Asymmetric deep supervised hashing. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3342–3349, 2018.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of International Conference on Learning Representations*, 2014.

[Li *et al.*, 2015] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1711–1717, 2015.

[Li *et al.*, 2017] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. Deep supervised discrete hashing. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2479–2488, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision*, pages 740–755, 2014.

[Liu *et al.*, 2016] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2064–2072, 2016.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations*, 2019.

[Meng *et al.*, 2024] Fanlei Meng, Xiangru Chen, and Yuan Cao. Targeted universal adversarial attack on deep hash networks. In *Proceedings of ACM International Conference on Multimedia Information Retrieval*, pages 165–174, 2024.

[Mittal *et al.*, 2011] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Blind/referenceless image spatial quality evaluator. In *Conference Record of the Forty Fifth*

Asilomar Conference on Signals, Systems and Computers, pages 723–727, 2011.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[Peng *et al.*, 2026] Jinjia Peng, Junyu Liu, Xutao Zuo, Zeze Tao, and Huibing Wang. Bi-level inter-modality modulation for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 21:2652–2667, 2026.

[Qiu *et al.*, 2020] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. In *Proceedings of European Conference on Computer Vision*, pages 19–37, 2020.

[Rombach *et al.*, 2022] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of International Conference on Learning Representations*, 2021.

[Sun *et al.*, 2020] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6397–6406, 2020.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of International Conference on Learning Representations*, 2014.

[Tu *et al.*, 2020] Rong-Cheng Tu, Xianling Mao, and Wei Wei. Mls3rduh: Deep unsupervised hashing via manifold based local semantic similarity structure reconstructing. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 3466–3472, 2020.

[Tu *et al.*, 2025] Rongxin Tu, Xiangui Kang, Chee Wei Tan, Chi-Hung Chi, and Kwok-Yan Lam. All points guided adversarial generator for targeted attack against deep hashing retrieval. *IEEE Transactions on Information Forensics and Security*, 20:1695–1709, 2025.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[Wang *et al.*, 2016] Jingdong Wang, Ting Zhang, Jingkuan Song, N. Sebe, and Heng Tao Shen. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):769–790, 2016.

[Wang *et al.*, 2021] Xunguang Wang, Zheng Zhang, Baoyuan Wu, Fumin Shen, and Guangming Lu. Prototype-supervised adversarial network for targeted attack of deep hashing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 16352–16361, 2021.

[Wang *et al.*, 2026a] Huibing Wang, Yawei Chen, Mingze Yao, Qian Liu, Guangqi Jiang, and Xianping Fu. Tensorized parameter-free multi-view spectral clustering based on fair representation learning. *IEEE Transactions on Multimedia*, pages 1–13, 2026.

[Wang *et al.*, 2026b] Huibing Wang, Luyan Cui, Mingze Yao, and Xianping Fu. Tensorized fine-grained incomplete multiview clustering via intrinsic structure recovery. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1–14, 2026.

[Xia *et al.*, 2014] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 2156–2162, 2014.

[Xue *et al.*, 2023] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, pages 2894–2921, 2023.

[Yang *et al.*, 2020] Erkun Yang, Tongliang Liu, Cheng Deng, and Dacheng Tao. Adversarial examples for hamming space search. *IEEE Transactions on Cybernetics*, 50(4):1473–1484, 2020.

[Yuan *et al.*, 2020] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis E. H. Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3080–3089, 2020.

[Yuan *et al.*, 2022] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. In *Proceedings of Advances in Neural Information Processing Systems*, 2022.

[Yuan *et al.*, 2023] Xu Yuan, Zheng Zhang, Xunguang Wang, and Lin Wu. Semantic-aware adversarial training for reliable deep hashing retrieval. *IEEE Transactions on Information Forensics and Security*, 18:4681–4694, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[Zhao *et al.*, 2023] Wenshuo Zhao, Jingkuan Song, Shengming Yuan, Lianli Gao, Yang Yang, and Heng Tao Shen. Precise target-oriented attack against deep hashing-based retrieval. In *Proceedings of ACM International Conference on Multimedia*, pages 6379–6389, 2023.