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

Chihan Huang [1] Xiaobo Shen [1]

[1]Nanjing University of Science and Technology

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.

Motivation

- In our article published at ICASSP 2025 (EmbSTar), the adversarial generation process exhibits instability and struggles to converge.
- Adversarial attacks can be conceptualized as adding noise to benign images, while diffusion models are known for their strong capability in handling noise. Therefore, we aim to modify the sampling distribution during the denoising process of the diffusion model to progressively shift the image distribution towards the adversarial distribution.

Methodology

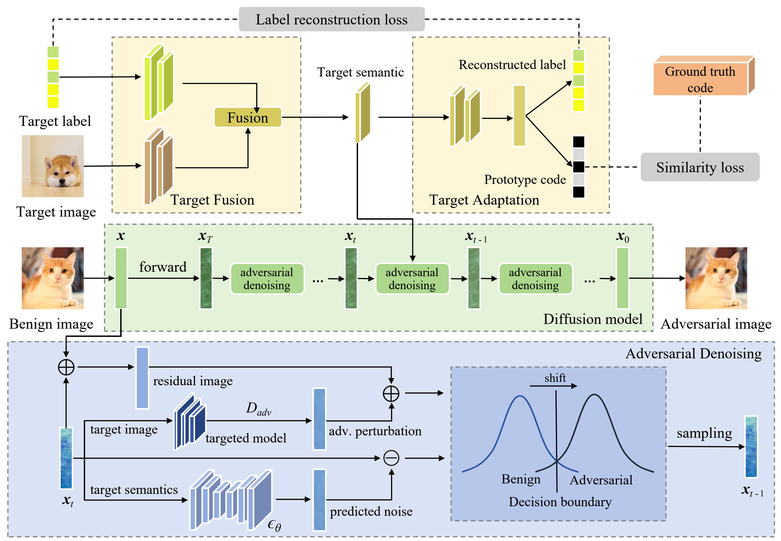

Figure 1. HUANG's structure: Target Fusion extracts target semantics, and Target Adaptation reconstructs the label and generates a prototype code. Adversarial generation adds noise to a benign image, followed by adversarial denoising to shift the image distribution from benign to adversarial, producing the final adversarial image.

Theory derivation

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 x_{t-1} is sampled from $\mathcal{N}(\mu_\theta, \Sigma_\theta)$. 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 $p_\theta(x_{t-1} \mid x_t)$ to a conditional distribution:

$$p_\theta(x_{t-1} \mid x_t, x_{tar}) = C_1 p_\theta(x_{t-1} \mid x_t)\, p_\theta(x_{tar} \mid x_{t-1})$$

Dhariwal 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:

$$p_\theta(x_{t-1} \mid x_t)\, p_\theta(x_{tar} \mid x_{t-1}) = \mathcal{N}\left(\mu + \Sigma \nabla_{x_t} \log p_\theta(x_{tar} \mid x_t), \Sigma\right)$$

So if we know $p_\theta(x_{tar} \mid x_t)$, we can direct the sampling distribution from benign to adversarial gradually.

Adversarial perturbation

We use an empirical measurement to describe $p_\theta(x_{tar} \mid x_t)$:

$$p_\theta(x_{tar} \mid x_t) = C_2 \exp\left(s\mathcal{D}_{adv}(x_{tar}, x_t)\right)$$
$$\mathcal{D}_{adv}(x_{tar}, x_t) = -\frac{1}{K} H(x_{tar})^T H(x_t) + 1$$

where $p_\theta(x_{tar} \mid x_t)$ can be seen as the probability that x_{t-1} will be recovered to an adversarial image, and $\mathcal{D}_{adv}(x_{tar}, x_t)$ measures the hamming distance between x_t and the target image.

The same operation can be done on the target semantic f^s:

$$p_\theta(f^s \mid x_t) = C_3 \exp\left(s\mathcal{D}_{sim}(f^s, x_t)\right)$$
$$\mathcal{D}_{sim}(f^s, x_t) = \frac{f^s \cdot \mathcal{TF}(x_t, y_{tar})}{\|f^s\| \, \|\mathcal{TF}(x_t, y_{tar})\|}$$

Residual image

Relying solely on adversarial perturbations may cause excessive deviations, producing visually distinct images. To mitigate this, we supervise the shift using the residual image $r = x_t - x_{ben}$, aligning the mean closer to the benign distribution for more visually similar outputs. Denote $g = \nabla_{x_t}\mathcal{D}_{adv}(x_{tar}, x_t) + \nabla_{x_t}\mathcal{D}_{sim}(f^s, x_t)$, x_{t-1} is then sampled from this adjusted distribution.

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

Looking for internship & graduate advisor

My name is Huang Chihan, a junior student at NJUST. I have received multiple awards and honors at my university and have published *one paper at AAAI* and *three at CCF-B conferences* as the *first author*. I am currently seeking **internship**, with a future research focus on *AI security*, particularly *LLM security*. I am also actively searching for a **graduate advisor**. If any professors from top universities in China are interested, please feel free to contact me! My personal homepage is shown bottom-left, where you can also find my WeChat QR code.

Quantitative results

Table 1 shows that HUANG outperforms existing methods across datasets and hash bit lengths, achieving over 10% higher t-MAP on DPSH, HashNet, and CSQ compared to previous SOTA. HUANG's strong performance, especially on smaller datasets, highlights its ability to capture fine-grained semantics and enhance robustness and transferability.

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.

Qualitative results

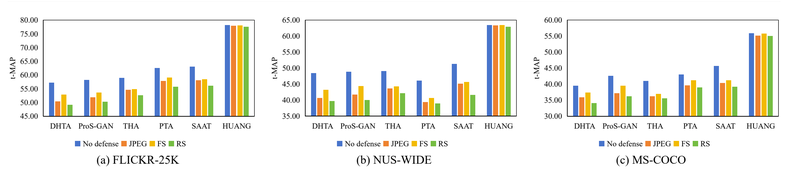

Figure 2. We evaluated HUANG's robustness against three defense methods: JPEG compression, feature squeezing, and randomized smoothing. HUANG outperforms prior methods, maintaining high t-MAP with minimal impact from defenses, thanks to its dynamic interplay of adding adversarial perturbations and denoising.

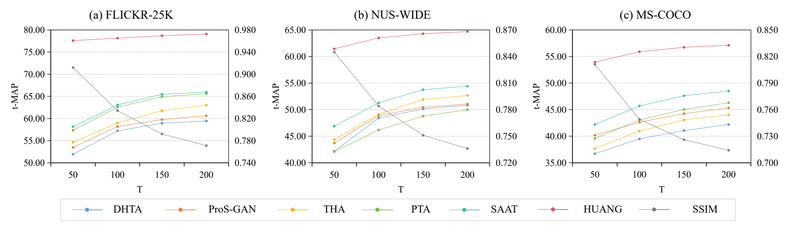

Figure 3. Experiments using SSIM show that as T increases, image quality decreases due to higher noise and denoising complexity. Smaller T weakens adversarial transferability, while excessively large T reduces adversarial effectiveness.