Efficient Multi-branch Black-box Semantic-aware Targeted Attack Against Deep Hashing Retrieval

Chihan Huang, Xiaobo Shen

Nanjing University of Sciences and Technology

huangchihan@njust.edu.cn, njust.shenxiaobo@gmail.com

IEEE International Conference on Acoustics, Speech, and Signal Processing
ICASSP 2025
6-11 April 2025

Abstract

Deep hashing excels in retrieval tasks but remains vulnerable to adversarial attacks, where small perturbations cause incorrect results. While many attack methods exist, targeted black-box attacks on deep hashing models are underexplored. We propose the Efficient Multi-branch Black-box Semantic-aware Targeted Attack (EmbSTar), which performs targeted black-box attacks. EmbSTar distills the target model into a knockoff model and introduces Target Fusion and Target Adaptation modules to enhance the semantic alignment between the adversarial and target images. This approach enables effective attacks with minimal queries. Experiments show that EmbSTar achieves state-of-the-art performance in targeted black-box attacks.

Introduction

The growth of multimedia data has increased the need for efficient retrieval. Deep hashing, a popular ANN method, leverages binary codes for fast similarity search but is vulnerable to adversarial attacks, which exploit subtle modifications to mislead models. Most research focuses on classification attacks, with fewer exploring hashing retrieval, particularly targeted black-box attacks.

Contributions

- We propose EmbSTar, an Efficient Multi-branch Black-box Semantic-aware Targeted Attack for deep hashing, addressing a rarely explored domain.
- EmbSTar requires no target model knowledge and achieves targeted results with minimal queries by effectively extracting semantic information from target labels.
- Experiments show EmbSTar outperforms previous methods, achieving state-of-the-art performance in targeted attacks.

Components

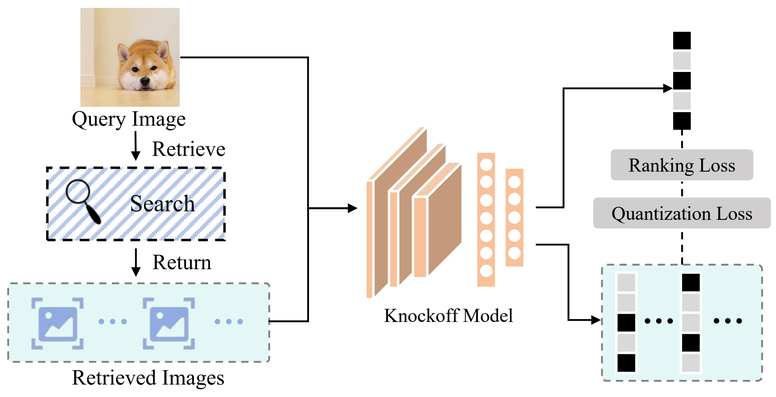

Fig. 1: The illustration of target model distilling. Query mages are transmitted to the target model, which then generates a sequence of search outcomes. The outcomes are used to train the knockoff model based on relevance ranking.

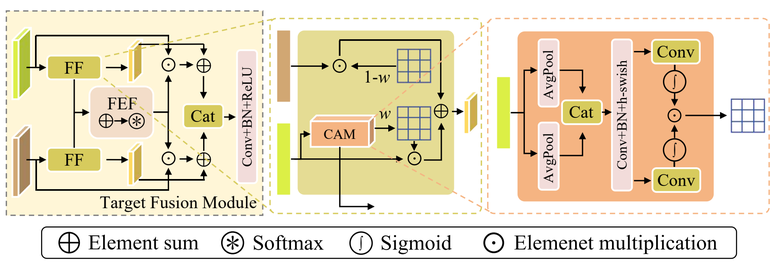

Fig. 2: The illustration of target fusion module.

Methodology

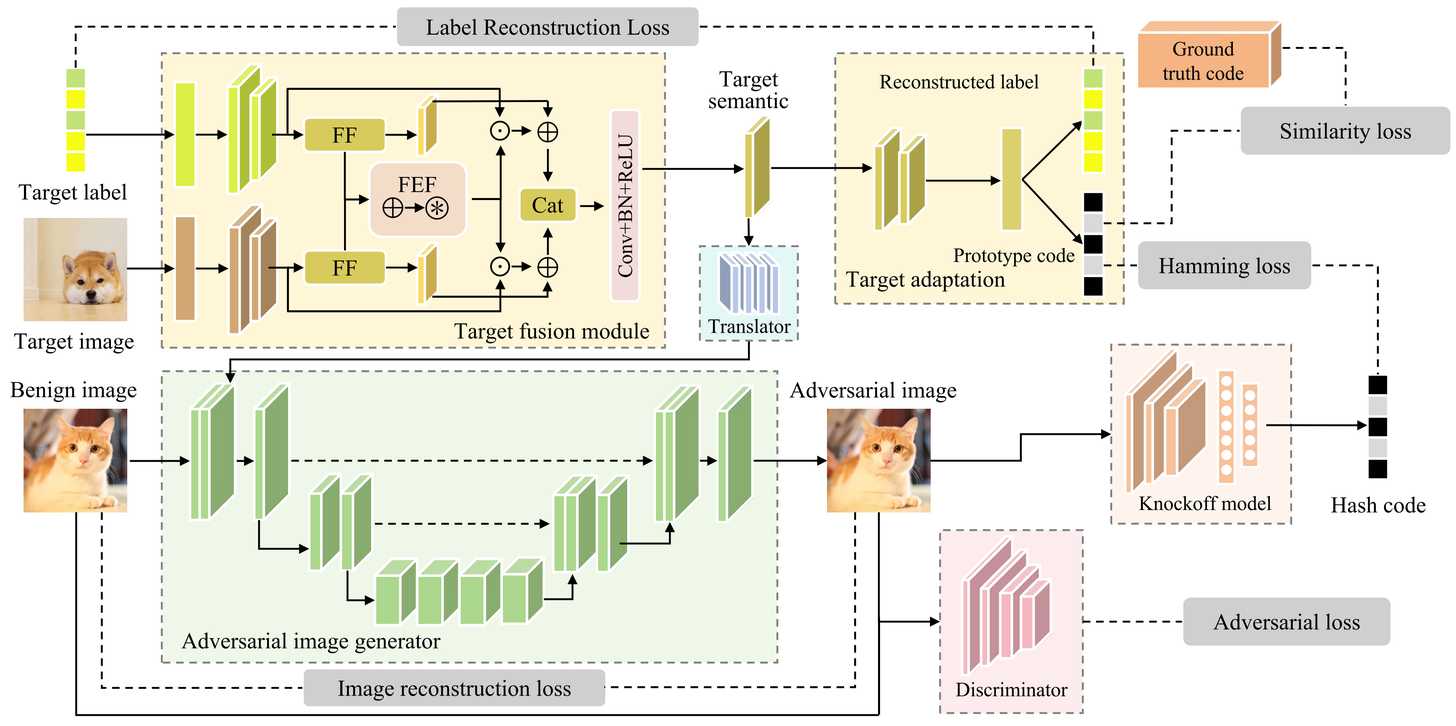

Fig. 3: The illustration of EmbSTar. Target label and image are fed into target fusion to derive target semantic, used to reconstruct the label and produce a prototype code in target adaptation. In adversarial generation, target semantic and benign image are input into generator, yielding an adversarial image. Discriminator ensures visual fidelity and category discrimination in adversarial images. Parameters of the knockoff model remain unchanged during this phase.

Quantitative results

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
	EmbSTar	**70.59**	**76.67**	**72.24**	**70.58**	**58.16**	**60.40**	**63.34**	**67.02**	**50.96**	**53.72**	**56.27**	**59.45**
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
	EmbSTar	**62.51**	**67.10**	**62.89**	**65.34**	**51.16**	**54.77**	**57.20**	**59.63**	**42.37**	**44.24**	**45.84**	**48.61**
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
	EmbSTar	**68.52**	**71.49**	**70.67**	**69.26**	**55.68**	**57.44**	**58.95**	**60.33**	**46.25**	**48.93**	**50.71**	**52.53**

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

Optimization

Stage 1: When knockoff distillation stage converges, its parameters $\Theta_{\mathcal{K}}$ are frozen.

$$\Theta_{\mathcal{K}} \Leftarrow \arg\min \left(\mathcal{L}_{tri} + \lambda_q \mathcal{L}_{quant}\right) \tag{1}$$

Stage 2: When prototype learning stage converges, its parameters $\Theta_{\mathcal{PL}}$ are also frozen.

$$\Theta_{\mathcal{PL}} \Leftarrow \arg\min \left(\mathcal{L}_{rec} + \lambda_s \mathcal{L}_{sim}\right) \tag{2}$$

Stage 3: In adversarial generation, we alternatively optimize the translator $\mathcal{T}$ + generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. When optimizing the former, we froze the latter's parameters, and vice versa.

$$\begin{aligned} \Theta_{\mathcal{T}}, \Theta_{\mathcal{G}} &\Leftarrow \arg\min \left(\mathcal{L}_{rec} + \lambda_h \mathcal{L}_{ham} + \lambda_a \mathcal{L}_{adv}\right) \\ \Theta_{\mathcal{D}} &\Leftarrow \arg\min \left(\mathcal{L}_D\right) \end{aligned} \tag{3}$$

Qualitative results

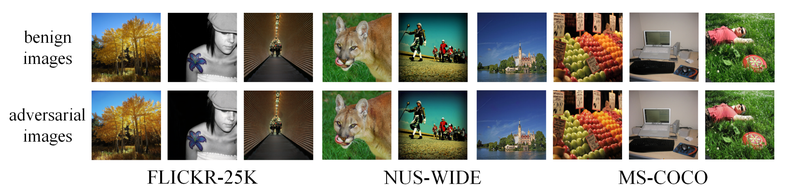

Fig. 4: Visual comparison of benign and adversarial images by EmbSTar.

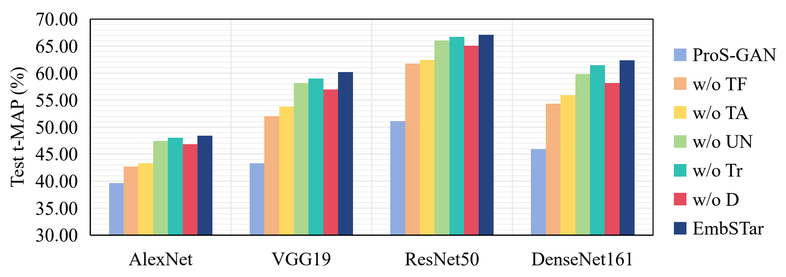

Fig. 5: Ablation results on different backbones, w/o means without. TF, TA, DB, SB, Tr, D are target fusion, target adaptation, detail branch, semantic branch, translator and discriminator respectively.