

# QAIR: Practical Query-efficient Black-Box Attacks for Image Retrieval

Xiaodan Li<sup>1\*</sup>, Jinfeng Li<sup>1</sup>, Yuefeng Chen<sup>1</sup>, Shaokai Ye<sup>2</sup>, Yuan He<sup>1</sup>, Shuhui Wang<sup>3</sup>, Hang Su<sup>4</sup>, Hui Xue<sup>1</sup>

<sup>1</sup>Alibaba Group <sup>2</sup>EPFL <sup>3</sup>Inst. of Comput. Tech., CAS, China

<sup>4</sup>Institute for AI, THBI Lab, Tsinghua University, Beijing, 100084, China

{fiona.lxd, jinfenglilj, yuefeng.chenyf, heyuan.hy, hui.xueh}@alibaba-inc.com

shaokai.yeah@gmail.com, wangshuhui@ict.ac.cn, suhangss@mail.tsi nghua.edu.cn

## Abstract

We study the query-based attack against image retrieval to evaluate its robustness against adversarial examples under the black-box setting, where the adversary only has query access to the top-k ranked unlabeled images from the database. Compared with query attacks in image classification, which produce adversaries according to the returned labels or confidence score, the challenge becomes even more prominent due to the difficulty in quantifying the attack effectiveness on the partial retrieved list. In this paper, we make the first attempt in Query-based Attack against Image Retrieval (QAIR), to completely subvert the top-k retrieval results. Specifically, a new relevance-based loss is designed to quantify the attack effects by measuring the set similarity on the top-k retrieval results before and after attacks and guide the gradient optimization. To further boost the attack efficiency, a recursive model stealing method is proposed to acquire transferable priors on the target model and generate the prior-guided gradients. Comprehensive experiments show that the proposed attack achieves a high attack success rate with few queries against the image retrieval systems under the black-box setting. The attack evaluations on the real-world visual search engine show that it successfully deceives a commercial system such as Bing Visual Search with 98% attack success rate by only 33 queries on average.

## 1. Introduction

Despite of its impressive performance in many tasks such as image classification [17], object detection [7] and image retrieval (IR) [37], deep neural network (DNN) has been shown to be vulnerable to adversarial examples that can trigger the misbehavior with human-imperceptible perturbations [14, 10, 13]. Such vulnerability has raised great concerns about the robustness and real-world deployment of DNNs for image retrieval [24, 39] and object detection [7],

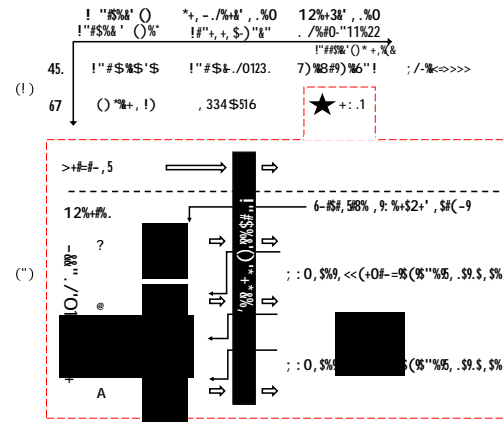


Figure 1. (a) (Left)Taxonomy of adversarial attacks. Different from existing attacks for IR, our attack is applicable to real-world scenarios since it only needs query access to the target model. (Right) The output of image classification(Cls) is a label or confidence score while it is a list of unlabeled images in IR. (b) Demonstration of the query-based black-box attack on IR. Given a target model, the attackers use queries to update and generate adversarial perturbations.

et al.. For example, in digital right management, the original graphic designs are protected by checking if there exists a same design in the top-k similar ones retrieved from the whole graphic design database. By adding adversarial perturbations on protected designs, attackers can deceive the target IR system into retrieving some irrelevant images for evading the censorship of professional monitors. Therefore, it is crucial to develop a practical robustness evaluation technology to explore the vulnerability of IR systems against adversarial attacks, and then facilitate the development of the corresponding countermeasures.

A general idea of adversarial attack is to generate adversarial examples with human-imperceptible perturbations along the gradient direction by *maximizing* a certain loss function, e.g., a classification loss [14, 11]. However, as shown in Fig. 1, the IR system produces a list of images for a query input. This makes it hard to define the objective

\*Corresponding author.

function for indicating the attack effectiveness only with the retrieved list. Under this circumstance, the gradients can hardly be estimated for deriving effective attacks. Though there exist some decision-based methods [3, 5] in attacking classification models which only rely on the final decision to indicate whether the attack succeeds, they usually require a significant attack cost by a tremendous number of queries to cross the decision boundary via greedy search [10].

Furthermore, in adversarial attacks, the gradients for guiding the attack process are usually calculated based on the knowledge of the target model, *e.g.*, the model structure and parameters. Existing studies on adversarial attacks against IR systems mainly focus on the white-box attacks in which attackers are assumed to have complete knowledge about the target model [41, 24, 36, 33], so the gradients can be directly acquired. However, the underlying white-box assumption does not hold in reality. Some studies try to use an approximate gradient instead for crafting adversarial examples [27, 36]. The approximate gradient could be either the gradient of a surrogate model (*a.k.a.* transfer-based attacks) or numerically estimated by methods (*a.k.a.* query-based attacks) such as the zero-order optimization [6]. The transfer-based methods attack the target model by leveraging adversarial examples generated against a white-box substitute model [36, 24], requiring training data that are usually protected. Besides, their attack success rate is still unsatisfactory due to the lack of adaptation procedure when the generated adversarial examples fail to attack the target model [10]. The query-based attacks produce the gradient with methods such as finite difference [6, 2], random gradient estimation [27]. However, they are not efficient enough due to the lack of knowledge about the target model.

To address the aforementioned challenges, we propose the first attempt on practical Query-efficient Attack against Image Retrieval (QAIR) under the black-box setting. First of all, we formulate the problem of black-box attacks on IR systems, and propose a new relevance-based loss to quantify the attack effects on target models with probabilistic interpretation. In this way, the structural output of IR systems can help to guide the gradient estimation during attacks. Besides, considering the fact that retrieved images are ranked based on similarities with the input image, which can generate plenty of labeled triplets, a recursive model stealing method is constructed on the ranking list to acquire transfer-based priors and generate the prior-guided gradients. Extensive experiments show that the proposed method can achieve a high attack success rate against IR systems with a remarkable Recall@K drop. We also evaluate our attack efficacy on the real visual search system<sup>1</sup>, which demonstrates its practicability in real-world scenarios.

Our main contributions can be summarized as follows:

- We formulate the problem of black-box attacks against image retrieval systems, and propose a new relevance-based loss to quantify the attack effects.
- We develop a recursive model stealing method to acquire transfer-based priors of target model for boosting the query-attack efficiency.
- We demonstrate the efficacy of our attack through extensive experiments on simulated environments and real-world commercial systems.

## 2. Related Work

In this section, we briefly introduce image retrieval and review existing adversarial attacks.

### 2.1. Image Retrieval

Image retrieval is a popular topic in computer vision and has been widely used in commercial systems such as Google Image Searching<sup>2</sup>, Bing Visual Search<sup>1</sup>, *etc.* [24]. A deep metric learning based image retrieval system usually consists of a metric learning model (*a.k.a.* image retrieval model) and a database (known as gallery) [37]. Given a query image, the metric learning model will extract and compare its feature with images in the gallery, then retrieve related ones based on their similarities with the query.

The metric learning model can be different in terms of training strategies. For example, contrastive loss [16] is proposed to make representations of samples from positive pairs to be closer while those from negative pairs to be far apart. Some researchers claim that pair-wise metric learning often generates a large amount of pair-wise samples, which are highly redundant. Training with random sampling may significantly degrade the model capability and also slow the convergence. Thus, hard mining strategy [32] and lifted structure loss [28] are proposed. Recently, multi-similarity loss [37] is proposed to establish a general pair weighting framework to formulate deep metric learning into a unified view of pair weighting and has achieved a state-of-the-art performance.

### 2.2. Adversarial Attack

Adversarial examples are maliciously crafted by adding human-imperceptible perturbations that trigger DNNs to misbehave [14]. The attacks for generating adversarial examples can be summarized into white-box [14, 26], transfer-based [11] and query-based attacks [3, 5] in terms of the information that attackers rely on [11]. The gradient calculation also differs a lot among these kinds of attacks.

**White-box.** Under the white-box setting, attackers have full access to the target model and they can directly acquire the true gradient of the loss *w.r.t.* the input. For instance,

<sup>1</sup><https://www.bing.com/visualsearch>

<sup>2</sup><https://images.google.com/>

Opposite Direction Feature Attack (ODFA) [40] generates adversarial examples by querying the target model's parameters and pushing away the feature of adversarial query in the opposite direction of their initial counterparts. To generate image-agnostic universal adversarial perturbations (UAP), Li *et al.* [24] try to optimize the traditional triplet loss inversely against metric learning on feature embeddings. However, the underlying white-box assumption usually does not hold in real-world scenarios.

**Transfer-based.** Transfer-based attacks do not rely on model information but need information about the training data to train a fully observable substitute model [29, 15]. For instance, DeepMisRanking (DeepMisR) [1] deceives the target models based on the transferability of adversarial examples generated against the substitute model by a white-box attack. But the training data may be unavailable in real applications. Though some work [42, 22] propose to steal model in a data-free manner, *e.g.*, producing inputs by generative models, this issue needs to be further investigated in image retrieval tasks. Besides, the performance of transfer-based attacks is limited due to the lack of adjustment when the gradient of the surrogate model points to a non-adversarial region of the target model [10].

**Query-based.** The query-based attack is more practical since the adversary in reality usually only has query access to the output of the target model. This kind of attack has been widely studied in the task of image classification and can be primarily divided into score-based attacks and decision-based attacks [3, 11].

Under the score-based setting, attackers have access to the confidence score of the prediction, which can be used to guide the attack process [6, 27]. Most score-based attacks usually estimate the gradient by zero-order optimization methods through query access to the output of the target model [10]. Specifically, a perturbation is firstly initialized and added to the input image. The output score will guide the algorithm to find out the optimization direction of the next step. For instance, Zero-Order Optimization Based Black-box Attack (Zoo) [6] estimates the gradient at each coordinate by using the symmetric difference quotient. To improve the query efficiency, a random gradient-free (RGF) method [27] is proposed to get an approximated gradient by sampling random vectors independently from a distribution.

Different from score-based ones, attacks under decision-based setting are more challenging since only the final decision is provided for indicating whether the attacks succeed. Existing decision-based attacks include Boundary Attack (BA) [3], HopSkipJumpAttack (HSJA) [5], *etc.* They usually treat an irrelevant or target image as the start point and decrease the perturbation gradually to make the adversarial examples visually similar to input image [3, 5, 9]. However, most of these attacks are proposed for image classification tasks, and to the best of our knowledge, there still

exists no query-based attacks for image retrieval.

### 3. Methodology

In this section, we first formulate the problem of attacking image retrieval models under the black-box setting and then elaborate our proposed attack. The whole attack pipeline is shown in Alg. 1, given an input image  $x$ , we first conduct a white-box attack on a substitute model  $s$  which is acquired with a recursive model stealing method beforehand (shown in Fig. 3), to provide the transfer-based priors for the following query-based attack. Then, we quantify the attack effects with a delicately designed relevance-based loss, and do gradient estimation following the basic idea of the score-based methods, aiming to provide the proper direction for the attack. Finally, we repeat the aforementioned steps till the generated adversarial image  $\hat{x}$  can deceive the target model successfully.

#### 3.1. Problem Formulation

As shown in Fig. 1, given a query image  $x$ , the image retrieval system with metric learning model  $f$  and gallery  $G$  returns a list of images

$$RLiSt^n(x, f) = \{x_1, x_2, \dots, x_i, \dots, x_n | x_i \in G\}, \quad (1)$$

ordered by their similarities to  $x$ , where  $n$  is the number of returned images and  $f$  projects  $x$  to the feature space as  $f(x)$ . In other words,  $D_f(x, x_i) = D_f(x, x_j)$ , s.t.  $i < j$  where  $D_f(x, x_i) = \|f(x) - f(x_i)\|_2$  is the metric that measures the feature distance between two images.

In this paper, the adversary aims to fool the target model into outputting a list of images whose top- $k$  has no overlap with original outputs under the assumption that the target model behaves well, *i.e.*, the returned images are well organized according to the similarities to the input image. Then, the attack goal can be formalized as

$$RLiSt^K(x, f) \cap RLiSt^K(x + \delta, f) = \emptyset \quad \text{s.t.} \quad \|\delta\|_p \leq \epsilon, \quad (2)$$

where  $K$  is the number of top-ranked images to be considered and  $\epsilon$  is the perturbation budget.  $p$  determines the kind of tensor norm (by default) to measure the perturbation.

The above goal can be solved by borrowing the idea of decision-based attacks proposed in the image classification task [5], in which only the final decision (*i.e.*, the predicted top label) instead of class probabilities is available to attackers. However, as shown in Fig. 2(left), the loss landscape is discontinuous, it hence requires combinatorial optimization or exhaustive search algorithms with a tremendous number of queries to perform a successful attack [8].

#### 3.2. Objective Function

To solve the above problems, we delicately design an objective function to quantify the attack effects on the retrieval model to guide the generation of adversarial images.

---

**Algorithm 1** The query-based attack for image retrieval

**Require:** Target model  $f$ ; input image  $x$ ; stolen model  $s$ ; number of iterations for momentum  $N_i$ ; max number of queries  $T$ ; max perturbation  $\epsilon$ ; step size  $\alpha$ ; learning rate  $\eta$ ; number of considered images  $K$ ;

```

1: Initialize  $\hat{x} = x, L^{\text{prev}} = 1.0, y = \text{RLiSt}^K(x, f)$ 
2: for  $t = 1$  to  $T/2$  do
3:   {Calculate basis  $u$  with stolen models} Eq. 11
4:   Initialize  $\hat{x}^t = \hat{x}, u = 0$ 
5:   for  $i = 1$  to  $N_i$  do
6:      $u = \eta \cdot u + \hat{x}^t (L_w(\hat{x}^t, y))$ 
7:      $\hat{x}^t = \text{CLIP}_x(\hat{x}^t + \alpha \cdot \text{sign}(u))$ 
8:   {Query attack with the resulted basis  $u$ } Eq. 9
9:    $\hat{g} = \frac{L(\hat{x} + u, y) - L(\hat{x}, y)}{\|u\|}$ 
10:   $\hat{x} = \text{CLIP}_x(\hat{x} + \alpha \cdot \text{sign}(\hat{g}))$ 
11:  if  $L(\hat{x}, y) = L^{\text{prev}}$  then
12:     $\epsilon = 2 \cdot \epsilon$ 
13:   $L^{\text{prev}} = L(\hat{x}, y)$ 
14: return adversarial sample  $\hat{x}$ 

```

---

Concretely, denote by  $P(\hat{x}, y)$  the probability that the adversarial image  $\hat{x}$  generated from the input  $x$  fails to trigger the target model  $f$  to misbehave, and denote by  $y$  the true label of  $x$ , *i.e.*,  $y = \text{RLiSt}^K(x, f)$ . Then, the objective is

$$\min L(\hat{x}, y) = P(\hat{x}, y), \text{ s.t. } \|\hat{x} - x\|_p \leq \epsilon. \quad (3)$$

To make  $P(\hat{x}, y)$  computable, density estimation methods such as kernel density estimators [30] can be applied. Since the computation cost is directly related to the number of samples, we need to sample as few samples as possible but approximate the distribution of  $x$  as accurately as possible. We leverage the nearest neighbor density estimation method to approximate  $P(\hat{x}, y)$  based on the top- $K$  nearest neighbors of  $x$  obtained by querying the target model. Then,  $P(\hat{x}, y)$  can be approximately rewritten as

$$P(\hat{x}, y) = \prod_{i=1}^K P(x_i) P(\hat{x}, y | x_i) = \prod_{i=1}^K p_i \cdot i_i, \quad (4)$$

where  $p_i = P(x_i)$  denotes the prior sampling probability and  $i_i = P(\hat{x}, y | x_i)$  denotes the conditioned attack failure probability.

In the typical image retrieval system,  $x_i$  is related to  $x$  with a specific similarity score. However, this score cannot be obtained under the black-box setting. The simplest strategy to tackle this problem is to treat each  $x_i$  equally, *i.e.*,  $i_i = [1, K], i_i = 1/K$  (denoted as the **Count-based Loss**). However, this is not the optimal strategy since it cannot reflect the attack effect in a fine-grained manner. Recall that  $x_i$  is ranked according to its similarity to  $x$ , thus we can use the ranking information to approximate their relevance.

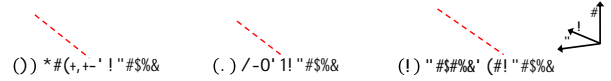


Figure 2. Loss (z-axis) landscape visualization of the target retrieval model. Compared with perturbations with Gaussian (x-axis), the loss gets to 0 faster with adversarial perturbations (y-axis), showing the model’s vulnerability against adversarial examples. We can find the hard-label problem is relaxed from left to middle. When relevance is considered, the loss gets to 0 with smaller perturbations.

Specifically, we refer to the Normalized Discounted Cumulative Gain metric (NDCG) used in classical ranking problem [21] and define  $r_i$  as the relevance between  $x_i$  and  $x$ . The probability  $P(x_i)$  is defined as:

$$P(x_i) = p_i = \frac{2^{r_i} - 1}{\sum_{i=1}^K (2^{r_i} - 1)}. \quad (5)$$

With  $r_i = K - i$ , the probability of the  $i$ -th result to be sampled is a decaying exponential.

$i_i = P(\hat{x}, y | x_i)$  indicates the attack failure probability of  $\hat{x}$  given  $x_i$ . It can be obtained from the retrieved results. If  $x_i = \text{RLiSt}^K(\hat{x}, f)$ , both  $x$  and  $\hat{x}$  are similar to  $x_i$  and thus  $\hat{x}$  should be similar to  $x$ , which also means the attack fails. Considering the aforementioned rank-sensitive relevance and supposing that  $x_i$  ranks at the  $j$ -th position in  $\text{RLiSt}^K(\hat{x}, f)$ ,  $i_i$  can be denoted as

$$i_i = \begin{cases} j, & x_i = \text{RLiSt}^K(\hat{x}, f) \text{ and } x_i = \hat{x}_j \\ 0, & x_i \neq \text{RLiSt}^K(\hat{x}, f) \end{cases}. \quad (6)$$

Then, the **Relevance-based** objective function  $L$  is rewritten as

$$L(\hat{x}, y) = \prod_{i=1}^K p_i \cdot i_i, \text{ s.t. } \|\hat{x} - x\|_p \leq \epsilon. \quad (7)$$

In this way, the attack effects can be evaluated only based on  $\text{RLiSt}^K(x, f)$  and  $\text{RLiSt}^K(\hat{x}, f)$  according to Eq. 7. As shown in Fig. 2, compared to the count-based loss, the attack with relevance-based loss requires a smaller perturbation to reduce the loss to 0.

### 3.3. Recursive Model Stealing

In adversarial attacks, the gradients for guiding the attack process are usually calculated based on the knowledge of the target model, which is unavailable under the black-box setting. Thus, some studies try to use surrogate models to obtain prior-guided gradients and improve the attack efficiency [4, 10, 15]. However, the training data of the target model required for training a surrogate model is usually unavailable. To tackle this problem, we propose to steal the gallery data of the IR system recursively via query access.



Specifically, as shown in Fig. 3, queried by a random image  $x$ , the image retrieval system returns a set of retrieved images  $\text{RLi st}^n(x, f)$ , from which we select  $N_c$  images evenly for greater diversity. These images again form a new image set as new queries to find more data. The above procedure will be repeated for  $C$  times to guarantee the diversity of collected images. To better obtain the priors for attacks, the surrogate model should be trained to have a similar ranking capacity as the target IR model. Hence, we query the target model with the collected  $M$  images to get final triplets as the ground-truth for training the surrogate model  $s$ , of which the objective function is defined as

$$\sum_{j>i} [D_s(x, x_i) - D_s(x, x_j) + \lambda]_+. \quad (8)$$

We set  $n = 1,000$ ,  $N_c = 10$ ,  $C = 3$  and  $\lambda = 0.05$  for all experiments. Thus, we only need 1,111 (summed by  $1+10+100+1,000$ ) queries to steal a model. Besides, the stolen model is dependent exclusively on the target model. In addition to the stolen model, the stealing cost is also shared by all the test samples. For example, the average query cost for each one is only  $1 + 1,111/1,000$  if the number of test samples is 1,000.

Our model stealing method is featured with the advantage that it requires no data beforehand. This is quite different from model distillation algorithms [24] which is usually performed based on the same training data with the target model. It also differs from generative model based methods [42], in which the collected data are usually out-of-distribution from the galleries of the target model. Besides, the diversity of the generated samples may be limited due to the problem of mode collapse. In contrast, by querying target models constantly, we can steal data from the gallery in a recursive manner and guarantee the performance.

### 3.4. Attack Optimization with Priors

Since the decision-based problem in Eq. 2 is relaxed with the proposed relevance-based loss, most of the query-based attacks proposed in image classification tasks can be extended to the retrieval tasks. We therefore adopt RGF-attack [27] as our base framework and define its loss by the proposed relevance-based loss for the extension to retrieval systems. The attack process can be summarized into two parts, *i.e.*, gradient estimation and perturbation optimization. Denote  $u_i$  as the  $i$ -th sampled basis vector which is sampled for  $q$  times and  $\hat{g}$  as the final estimated gradient. Then, gradient estimation and perturbation optimization are accomplished as follows:

$$\hat{g} = \frac{1}{q} \sum_{i=1}^q \hat{g}_i, \quad \hat{g}_i = \frac{L(x + u_i, y) - L(x, y)}{\|u_i\|}, \quad (9)$$

$$\hat{x} = \text{CLIP}_x(x + \eta \cdot \text{sign}(\hat{g})),$$

where  $\eta$  is the parameter to control the sampling variance and  $\eta$  is the learning rate. The  $\text{CLIP}_x$  operation aims to

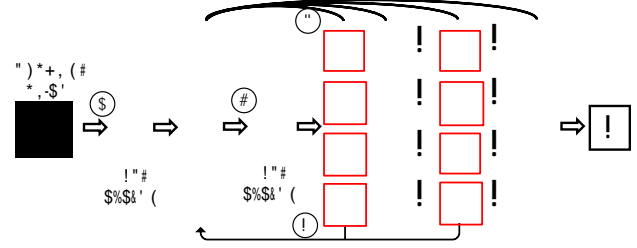


Figure 3. The pipeline of our model stealing. First, an arbitrary image is put into the target image retrieval (IR) system (1). The retrieved images are then evenly selected to construct a new query set, which will be put into the IR system in the next iteration (2, 3) for more triplets. Finally, The stolen triplets (4) will be used to train the substitute model  $s$ .

make the perturbation bounded in the budget [34]. Besides, the generated adversarial example is converted to integer before it is fed into the target model to ensure its validity.

In RGF, the basis initialization is achieved by sampling random vectors independently from a distribution such as Gaussian. This can be improved with transferable priors [15]. For this, we follow the state-of-the-art work Learnable Black-Box Attack [38], which utilizes the surrogate model  $s$  to obtain the transfer-based priors and guide the basis selection. Specifically, Momentum Iterative Method (MIM) [12] is firstly adopted to conduct the white-box attack based on the stolen model. The derived momentum item  $u$  is then used as the basis for the query-based attack. The loss  $L_w$  used for white-box attack is

$$L_w(\bar{x}, y) = \|\bar{s}(\bar{x}) - \sum_{i=1}^K w_i \cdot s(x_i)\|_2^2. \quad (10)$$

And the optimization procedure for momentum is:

$$u = \eta \cdot u + \bar{x} (L_w(\bar{x}, y)), \quad (11)$$

$$\bar{x} = \text{CLIP}_x(\bar{x} + \eta \cdot \text{sign}(u)),$$

where  $u$  is initialized with 0 and  $\eta = 0.9$ . The above procedure will be repeated for  $N_i$  times.

Note that our QAIR is different from previous transfer-based attacks against image retrieval that utilize substitute models for crafting adversarial examples directly. Instead, QAIR employs the stolen model for obtaining transfer-based priors and generating prior-guided gradients for query attack. In this way, adversarial examples can be further rectified with query response until the attack succeeds.

## 4. Experiments

In this section, we evaluate the proposed attack on various image retrieval models. More details can be found in our supplementary material.

### 4.1. Experimental Settings

**Datasets.** We evaluate our attack on three public datasets. Caltech-UCSD Birds-200-2011 (CUB-200) [35]:



Figure 4. Visualization of the attack procedures (left) and attack results (right). Images in red boxes are generated adversarial examples, which can fool the target model to return irrelevant images with imperceptible perturbations in the corresponding rows. Scores on the left are the corresponding loss to their searched results on their right. As more samples in the original sets disappear, the loss decays to 0.

Attacks	CUB-200								SOP						In-Shop									
	1	2	4	8	16	32	AQ	ASR	1	10	100	1000	AQ	ASR	1	10	20	30	40	50	AQ	ASR		
Original	0.61	0.73	0.87	0.91	0.98	0.99	0	0	0.724	0.816	0.904	0.960	0	0	0.642	0.868	0.910	0.926	0.938	0.945	0	0		
Comparsion with the State-of-the-art Methods																								
OptAttack [8]	0.08	0.15	0.30	0.49	0.63	0.89	9708	0.04	0.012	0.032	0.744	0.928	7931	0.288	0.004	0.020	0.564	0.680	0.764	0.828	3017	0.948		
Sign-Opt [9]	0.11	0.21	0.35	0.57	0.70	0.88	8833	0.00	0.008	0.024	0.696	0.916	6746	0.372	0.014	0.032	0.464	0.560	0.660	0.728	5564	0.492		
HSJA [5]	0.13	0.23	0.41	0.47	0.60	0.92	10000	0.00	0	0	0.632	0.880	5888	0.420	0.004	0.024	0.316	0.470	0.564	0.624	5379	0.472		
QAIR	0.16	0.23	0.32	0.45	0.56	0.76	93	0.69	0.016	0.064	0.472	0.832	35	0.904	0.008	0.044	0.132	0.256	0.312	0.352	35	0.916		
Component Analysis																								
QAIR <sub>C</sub>	0.59	0.76	0.83	0.94	0.96	0.97	199	0.01	0.176	0.372	0.724	0.916	113	0.480	0.296	0.556	0.716	0.764	0.800	0.832	147	0.310		
QAIR <sub>C-1</sub>	0.36	0.51	0.60	0.74	0.80	0.92	152	0.32	0.060	0.112	0.572	0.868	60	0.812	0.072	0.164	0.272	0.396	0.476	0.520	65	0.784		
QAIR <sub>C-S</sub>	0.31	0.46	0.52	0.58	0.72	0.85	142	0.37	0.056	0.088	0.532	0.848	51	0.836	0.052	0.124	0.204	0.320	0.404	0.432	50	0.844		
QAIR <sub>R-S</sub>	0.16	0.23	0.32	0.45	0.56	0.76	93	0.69	0.016	0.064	0.472	0.832	35	0.904	0.008	0.044	0.132	0.256	0.312	0.352	35	0.916		

Table 1. Comparison with state-of-the-art methods on CUB-200, SOP and In-Shop before (Original) and after attack (others). Smaller Recall@K, smaller average number of queries (AQ) over both successful and failed attacks as well as higher Attack Success Rate (ASR) mean stronger attack.

It has 200 classes of birds with 11788 images. The first 100 classes are split out for training and the rest for testing. It is a small but hard dataset for attack since it only has 100 classes in testing data. **Stanford Online Products (SOP)** [28]: It is a large scale dataset in image retrieval with 23k classes of 120k online product images from eBay.com. It is split into 11,318 classes of 59,551 images for training and 11,316 classes of 60,502 images for testing. **In-Shop Clothes (In-Shop)** [25]: This dataset contains 54,642 images of 11,735 clothing items from Forever21. It provides 3,997 and 3,985 classes for training (25,882 images) and testing (28,760 images).

**Evaluation metrics.** We use the commonly used metric Recall@K [28] in image retrieval for evaluation. Greater drop of Recall@K indicates stronger attack. Besides, the commonly used attack success rate (ASR) metric in adver-

sarial attack community is also employed. We treat the attack as successful when Eq. 2 satisfies, thus ASR can be evaluated as the percentage of successful attacks. Note that ASR is designed for evaluating attacks against image retrieval under the black-box setting. This is different from Recall@K where true labels are required.

**Implementation details.** We adopt the state-of-the-art image retrieval models<sup>3</sup> [37] as targets. They are implemented with BN-Inception Network [20] as most image retrieval works do for fairness and trained by their Multi-Similarity Loss. The image retrieval results are listed in the Tab. 1 (The row with “Original”). For model stealing, ResNet50 [17] is adopted as the default backbone and trained with random horizontal flip and resized crop only since the pre-processing of the target model is unavailable

<sup>3</sup>[https://github.com/bnu-wangxun/Deep\\_Metric/](https://github.com/bnu-wangxun/Deep_Metric/)

Metric Learning Models		Recall@K before our attacks						Recall@K after our attacks						AQ	ASR	DRR@1
		1	2	4	8	16	32	1	2	4	8	16	32			
BN-Inception [20]	Multi-Similarity [37]	0.61	0.73	0.87	0.91	0.98	0.99	0.16	0.23	0.32	0.45	0.56	0.76	93.40	0.69	73.77%
	Contrastive [16]	0.57	0.66	0.81	0.89	0.92	0.96	0.16	0.28	0.45	0.55	0.64	0.78	89.99	0.68	71.93%
	HardMining [32]	0.62	0.75	0.81	0.88	0.94	0.97	0.24	0.29	0.39	0.48	0.62	0.75	96.34	0.64	61.29%
	Lifted [28]	0.62	0.73	0.84	0.92	0.94	0.97	0.14	0.24	0.28	0.40	0.52	0.80	90.14	0.71	77.42%
DenseNet121 [19]	Multi-Similarity [37]	0.66	0.81	0.89	0.94	0.96	0.99	0.08	0.15	0.24	0.37	0.53	0.67	84.46	0.72	87.88%
	Contrastive [16]	0.66	0.80	0.88	0.91	0.95	0.98	0.10	0.15	0.23	0.29	0.42	0.60	83.80	0.70	84.85%
	HardMining [32]	0.66	0.76	0.85	0.92	0.97	0.99	0.12	0.17	0.27	0.33	0.47	0.65	161.92	0.27	81.82%
	Lifted [28]	0.66	0.79	0.87	0.92	0.95	0.98	0.07	0.15	0.26	0.41	0.51	0.65	84.36	0.68	89.39%

Table 2. Recall@K performances on the CUB-200 dataset before and after our attacks. It can be found that the proposed attack is effective on different image retrieval architectures trained with different metric learning methods. DRR@1 is the drop rate on Recall@1. The higher it is, the more vulnerable the image retrieval model is.

Attacks	CUB-200					
	1	2	4	8	16	32
Original	0.61	0.73	0.87	0.91	0.98	0.99
FGSM [14]	0.33	0.45	0.56	0.66	0.76	0.85
T BIM [23]	0.28	0.44	0.60	0.77	0.77	0.85
MIM [12]	0.20	0.28	0.39	0.52	0.61	0.75
Q Ours (QAIR)	<b>0.16</b>	<b>0.23</b>	<b>0.32</b>	<b>0.45</b>	<b>0.56</b>	0.76

Table 3. Comparison with transfer-based attacks (T). Q means query-based attack.

to attackers. We evaluate on randomly sampled 250 images in the test sets on SOP and In-Shop (100 for CUB-200). The perturbation budget is set to 0.05 under  $\ell_2$ -norm and the maximal number of query T is set to 200. The parameters in Eq. 9 are set as follows:  $q = 1$ ,  $\alpha = 0.1$ ,  $\beta = 0.01$  [27]. For each dataset, we set  $K = 16$  in Eq. 2 by default.

## 4.2. Comparison with State-of-the-art Methods

Since the adversarial attack against image retrieval systems under black-box setting is a kind of decision-based attack, we compare our QAIR with several state-of-the-art decision-based attacks including Optimization-based attack (OptAttack) [8], Sign-Opt [9] and HopSkipJumpAttack (HSJA) [5]. For these attacks, the maximum number of queries is set to 10,000 to find adversarial examples with small perturbations. As shown in Tab. 1, our method can achieve comparable attack effects and at the same time, require much fewer queries. This proves the practical value of our method. We found that though decision-based methods can completely subvert the top K results in most cases, the required maximum perturbation after 10,000 queries is usually much higher than  $\epsilon$ , resulting in a low ASR. For a comprehensive study, we evaluate the ASR under different max perturbation limitations further. As shown in Fig. 5 (left), our attack can always get a higher ASR than other methods, showing the effectiveness of the proposed approach. The visualization comparison of generated adversarial examples and comparison on defensive models can be found in our supplementary material.

Model	CUB-200							AQ	ASR
	1	2	4	8	16	32			
Original	0.61	0.73	0.87	0.91	0.98	0.99	0	0	
S <sub>r18</sub>	0.18	0.24	0.35	0.53	0.62	0.78	92.14	0.70	
S <sub>r50</sub>	0.16	0.23	0.32	0.45	0.56	0.76	93.40	0.69	
S <sub>r101</sub>	0.24	0.29	0.35	0.44	0.60	0.79	99.92	0.65	
S <sub>v16</sub>	0.28	0.37	0.44	0.52	0.65	0.82	121.84	0.54	
S <sub>d121</sub>	<b>0.14</b>	<b>0.23</b>	0.34	<b>0.42</b>	<b>0.55</b>	0.77	87.64	<b>0.73</b>	
S <sub>d169</sub>	0.18	0.24	<b>0.30</b>	0.45	0.56	<b>0.74</b>	<b>86.48</b>	0.71	

Table 4. Recall@K performance after our attack in terms of stolen models with different architectures.

Figure 5. Comparisons under different perturbation budgets on CUB-200 dataset (left) and Recall@K in terms of different numbers of queries used to steal the target model (right).

## 4.3. Comparison on Transfer and Query Attacks

We also compare the proposed query-based attack with transfer-based attacks. The evaluation results are listed in Tab. 3, from which we can see that the proposed attack outperforms transfer attacks developed based on different white-box attacks such as Fast Gradient Sign Method (FGSM) [14] and Basic Iterative Method (BIM) [23], as well as the Momentum Iterative Method (MIM) [12]. This is reasonable since the query-based attack can adjust the optimization direction with retrieval results, while transfer-based attack heavily relies on the transferability of generated adversarial examples.





## References

- [1] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip HS Torr. Metric attack and defense for person re-identification. *arXiv preprint arXiv:1901.10650*, 2019.
- [2] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *European Conference on Computer Vision*, pages 158–174. Springer, 2018.
- [3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.
- [4] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4958–4966, 2019.
- [5] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1277–1294. IEEE, 2020.
- [6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.
- [7] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018.
- [8] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.
- [9] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*, 2019.
- [10] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems*, pages 10934–10944, 2019.
- [11] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2020.
- [12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.
- [13] Yan Feng, Bin Chen, Tao Dai, and Shu-Tao Xia. Adversarial attack on deep product quantization network for image retrieval. *arXiv preprint arXiv:2002.11374v1*, 2020.
- [14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [15] Yiwen Guo, Ziang Yan, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In *Advances in Neural Information Processing Systems*, pages 3825–3834, 2019.
- [16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [18] Houdong Hu, Yan Wang, Linjun Yang, Pavel Komlev, Li Huang, Xi Chen, Jiawei Huang, Ye Wu, Meenaz Merchant, and Arun Sacheti. Web-scale responsive visual search at bing. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 359–367, 2018.
- [19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [21] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA, 2017.
- [22] Sanjay Kariyappa, Atul Prakash, and Moinuddin Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. *arXiv preprint arXiv:2005.03161*, 2020.
- [23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- [24] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4899–4908, 2019.
- [25] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [27] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

- [28] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.
- [29] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.
- [30] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.
- [31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [33] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1589–1604, 2020.
- [34] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2019.
- [35] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [36] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2020.
- [37] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.
- [38] Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning black-box attackers with transferable priors and query feedback. 2020.
- [39] Guoping Zhao, Mingyu Zhang, Jiajun Liu, and Ji-Rong Wen. Unsupervised adversarial attacks on deep feature-based retrieval with gan. *arXiv preprint arXiv:1907.05793*, 2019.
- [40] Zhedong Zheng, Liang Zheng, Zhilan Hu, and Yi Yang. Open set adversarial examples. *arXiv preprint arXiv:1809.02681*, 2018.
- [41] Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. Adversarial ranking attack and defense. *arXiv preprint arXiv:2002.11293*, 2020.
- [42] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2020.