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

This supplementary material provides more details about the principles of loss landscapes (Fig. 2 in the paper) and decision-based attacks. For comprehensive experiments, we also provide further evaluations on defensive models (Sec. 4.2 in the paper) and ablation studies on SOP and In-Shop datasets (Sec. 4.5 in the paper). Besides, we also plot a scatter map to visualize the relationship between the Attack Success Rate (ASR) metric and Recall@K drop for validating the rationality of the proposed attack goal experimentally. More details about attacking real-world visual search engine are also provided (Sec. 4.6 in the paper).

1. Loss Landscape

The visualization of loss landscape is implemented with the toolbox provided by [10]. The loss is designed as follows:

$$\text{loss}(i,j) = \mathcal{L}(\hat{x}, y),\ s.t.\ \hat{x} = x + i * \gamma + j * \eta \quad (1)$$

where coordinate (i, j) determines the perturbation added on input image. γ is a random direction sampled from Gaussian distribution while η is the sign of gradient and can be generated with:

$$\eta = \text{sign}(u) = \text{sign}(\frac{\partial(||f(\hat{x}) - f(x)||_2)}{\partial \hat{x}}). \quad (2)$$

Note that the gradient is directly derived from the target model for its loss landscape visualization. As shown in Fig 2, compared with same perturbations in the Gaussian noise direction, the loss gets to 0 faster in the adversarial direction, showing the model's vulnerability against adversarial examples. Besides, with the proposed relevance-based loss, the loss gets to 0 with smaller adversarial perturbations (see the red dotted line).

2. Decision-based Attack

Decision-based attacks is a kind of query-based attack that requires only the decision of whether the attack succeeds. They usually treat an irrelevant or target image as a start point and decrease the perturbation gradually to make the adversarial similar to the input image visually during optimization [1, 4, 2]. For example, OptAttack [3] starts the attack from an image that lies in the target class with a searched direction. Then it reduces the distance of the perturbed image towards the original input in input space with binary search. Though it can always succeed in subverting the outputs results in a great recall@K drop, it requires a tremendous number of queries to achieve small perturbations. Thus, attack success rate, which takes both recall@K drop and mean perturbations into consideration, is a relatively comprehensive metric. As shown in Fig. 11, though the resulted adversarial examples can subvert the top-k results (which can lead to a high Recall@K drop), the re-

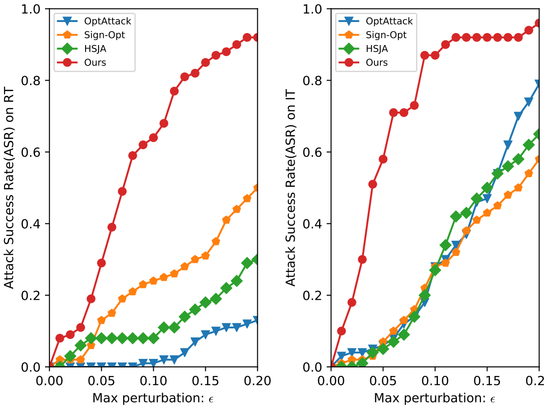

Figure 8. Attacks results on defensive models, including robust training (RT, left) and input transformations (IT, right).

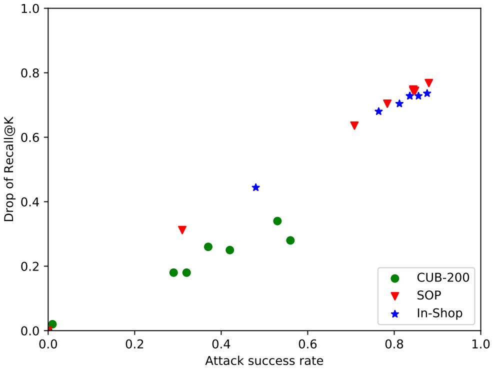

Figure 9. Scatter map of ASR and drop of Recall@K metric.

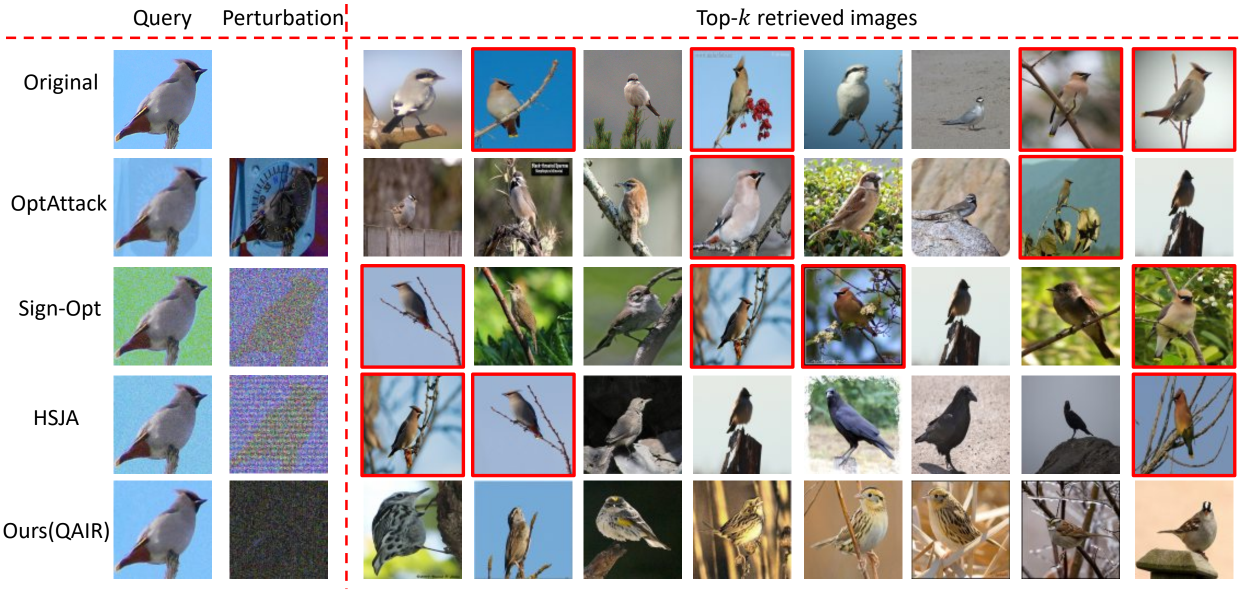

Figure 10. Query results before and after attacks. Images in the first column are the queries. Images in the second column are the adversarial perturbations added on original images. Darker perturbation images mean smaller disturbances needed, which in turn indicates that the attacks tend to be more effective. The red boxes represent the correctly matched images.

Attacks		Recall@K before attacks						Recall@K after attacks						AQ	ASR	DRR@1
		1	10	20	30	40	50	1	10	20	30	40	50			
BN-Inception [9]	Multi-Similarity [14]	0.853	0.959	0.965	0.973	0.976	0.979	0.008	0.044	0.132	0.256	0.312	0.352	35.19	0.92	**99.06%**
	Contrastive [7]	0.832	0.956	0.976	0.980	0.984	0.984	0.008	0.068	0.124	0.260	0.320	0.372	38.93	0.90	**99.04%**
	HardMining [13]	0.868	0.980	0.988	0.992	0.996	0.996	0.028	0.112	0.208	0.336	0.412	0.464	57.32	0.82	**96.77%**
	Lifted [12]	0.828	0.944	0.960	0.972	0.976	0.988	0.032	0.080	0.172	0.292	0.380	0.436	42.51	0.88	**96.14%**
DenseNet121 [8]	Multi-Similarity [14]	0.864	0.964	0.964	0.976	0.980	0.988	0.028	0.156	0.204	0.232	0.280	0.292	19.34	0.98	**96.76%**
	Contrastive [7]	0.868	0.948	0.960	0.964	0.976	0.976	0.016	0.112	0.148	0.184	0.220	0.232	17.85	0.97	**98.16%**
	HardMining [13]	0.852	0.968	0.980	0.988	0.988	0.988	0.036	0.148	0.200	0.252	0.292	0.320	17.16	0.97	**95.77%**
	Lifted [12]	0.828	0.952	0.964	0.976	0.984	0.984	0.044	0.152	0.228	0.276	0.320	0.340	30.49	0.92	**94.69%**

Table 5. Recall@K performances on In-Shop dataset before and after attacks.

Attacks		Recall@K before attacks				Recall@K after attacks				AQ	ASR	DRR@1
		1	10	100	1000	1	10	100	1000			
BN-Inception [9]	Multi-Similarity [14]	0.729	0.855	0.932	0.978	0.016	0.064	0.472	0.832	35.45	0.90	**97.81%**
	Contrastive [7]	0.701	0.839	0.920	0.975	0.008	0.028	0.440	0.792	27.57	0.94	**98.86%**
	HardMining [13]	0.723	0.861	0.937	0.980	0.016	0.032	0.428	0.824	31.73	0.94	**97.79%**
	Lifted [12]	0.703	0.839	0.923	0.974	0.008	0.068	0.472	0.832	36.37	0.91	**98.86%**
DenseNet121 [8]	Multi-Similarity [14]	0.720	0.824	0.904	0.964	0.024	0.140	0.312	0.612	20.97	0.96	**96.67%**
	Contrastive [7]	0.692	0.808	0.908	0.956	0.040	0.136	0.356	0.660	19.62	0.96	**94.22%**
	HardMining [13]	0.706	0.842	0.927	0.976	0.048	0.144	0.340	0.620	19.10	0.97	**93.20%**
	Lifted [12]	0.704	0.808	0.900	0.964	0.088	0.216	0.444	0.728	25.99	0.94	**87.50%**

Table 6. Recall@K performances on SOP dataset before and after attacks.

quired perturbations from OptAttack, Sign-Opt and HSJA are much more larger than ours. Besides, these perturbations are also beyond the perturbation budgets, leading to a low attack success rate.

3. Comparison on Defensive Models

We further validate the effectiveness of the proposed method against several defensive models on CUB-200 dataset, including the classical robust training (RT) [5] and input transformation (IT) [6]. The results in Fig. 8 show that compared to state-of-the-art methods, our attack can achieve a much higher attack success rate under the same perturbation budgets. This demonstrates the superiority of our method on attacking defensive models.

4. Ablation Study on More Datasets

Tab. 5 and Tab. 6 show more detailed experiments of attacking various deep metric learning models on In-Shop and SOP datasets, respectively. It can be found that the proposed query-based attack can achieve a high attack success rate on both datasets, demonstrating its effectiveness in different scenarios.

5. Attack Goal and Objective Function

Under the black-box setting, the attack success rate can only be calculated based on the observation of retrieved list. The rationality needs to be further explored. For this, we plot a scatter map of ASR and drop of Recall@K (obtained based on the true label), which is shown in Fig. 9 (under the same perturbations). It can be found that the ASR is in proportion to Recall@K drop, indicating the rationality of our attack goal experimentally. When comparing with state-of-the-art methods, the ASR metric takes both recall@K drop and maximum perturbations into consideration, making it a relatively comprehensive metric.

For the proposed bidirectional relevance-based loss, we provide some examples to make it more comprehensible. Suppose only the top three candidates are considered. Given an input image x, the target image retrieval system will output the top three similar images $\{a_1, a_2, a_3\}$ and others $\{b_4, b_5, b_6, ...\}$. After attacks, there are several kinds of situations:

- $\{a_1, b_4, b_5\}$ and $\{a_3, b_4, b_5\}$. In general, higher rank denotes higher relevance to the input image x. Thus, the loss of situation $\{a_1, b_4, b_5\}$ should be greater than $\{a_3, b_4, b_5\}$. Thus, the rank-sensitive relevance before attacks should be considered.
- $\{a_3, a_2, a_1\}$. The loss of situation $\{a_3, a_2, a_1\}$ should be smaller than $\{a_1, a_2, a_3\}$ since a_1 is the most relevant one to input image x. The lower it ranks, the more successful the attack is. Thus, how the candidates is ranked after attacks should also be considered.

6. Visualization Comparison

Fig. 10 shows the top 8 retrieved images of different input images (the first column). Images in the red boxes are from the same category with the input query. It can be found that after 10,000 queries, the perturbations generated by other methods are still much greater than ours (darker perturbation images indicate smaller perturbations), which only needs 200 queries. Besides, though all the adversarial examples can subvert the top-k retrieved results successfully, the retrieved results produced by other methods may contain images that share the same categories (red boxes) with the original ones. On the contrary, our method tries to push the adversarial example further from the original cluster in the feature space, which can relieve the inconsistency between attack goal and true labels.

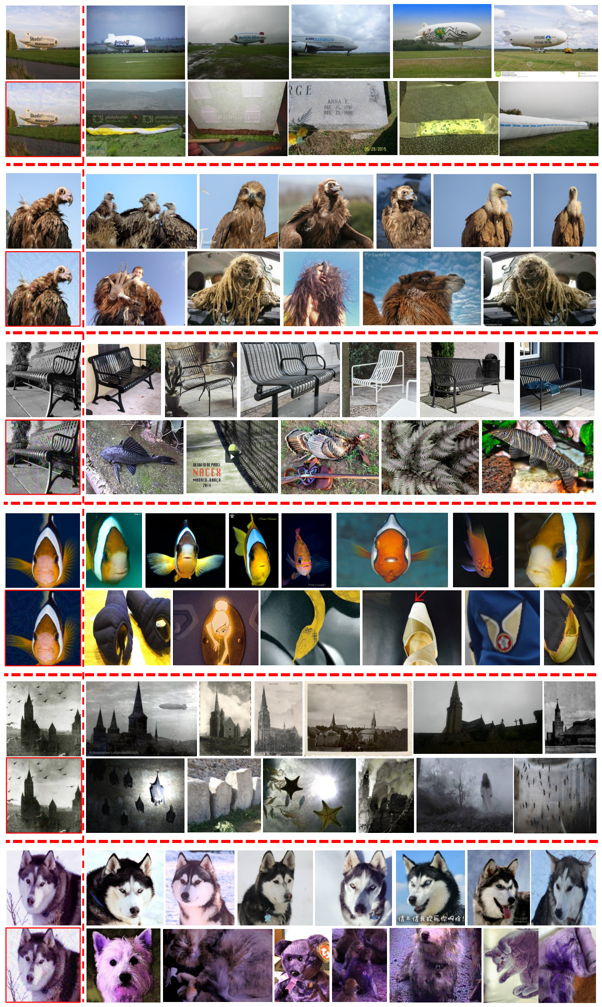

Figure 11. Query results on Bing Visual Search. Images in red boxes are adversarial examples generated with the proposed method. Images in the first column are queries while others are corresponding search results.

7. Attacks on Real-world Visual Search Engine

7.1. Implementation Details

Unlike most existing transfer attacks, query-attack that we study in this paper needs to query search engine constantly. Bing Visual Search is the only image retrieval API that can be automatically queried. Thus, we only provided attack results on Bing Visual Search.

Since Bing Visual Search is a frequently-used search engine and it has a huge amount of data in its gallery. Given an input image x, there are thousands of similar images with x. Thus, we need to take more candidates into consideration. For this, we set $\mathcal{K} = 100$ to ensure the adversarial examples far away enough from the original clusters in the feature space. Besides, the max query time and perturbation are set to 200 and 0.1. We only employ ResNet50 pretrained on ImageNet as our substitute model since it can make a good performance already.

7.2. Attack Results

As shown in Fig. 11, the generated adversarial examples can mislead the Bing Visual Search to output images actually irrelevant to the input image successfully with human-imperceptible perturbations. To quantitatively measure the performance, we randomly sample 1000 images from ImageNet for testing and the proposed method can achieve 98% attack success rate with only 33 queries on average. This demonstrates the practicability of our attack in real-world scenarios.

We have also attached a video recording the image retrieval results on Bing Visual Search before and after attacks. It should be noted that Bing Visual Search updates their engine frequently. Thus some generated adversarial examples may be ineffective after a few days.

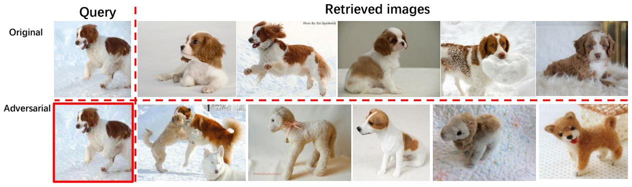

Figure 12. Failure attack examples. Image in red box is adversarial example generated with the proposed method. We can find that the retrieved images are still relevant to input image after attack even when its top-k images are subverted.

Methods	MQ	CUB-200		SOP		In-Shop	
		ASR	AQ	ASR	AQ	ASR	AQ
OptAttack [3]	10,000	0.04	9708	0.288	7931	**0.948**	3017
Sign-Opt [4]		0	8833	0.372	6746	0.492	5564
HSJA [2]		0	10,000	0.420	5888	0.472	5379
Ours	200	0.69	**93**	0.904	**35**	0.916	**35**
	500	0.72	180	0.918	64	**0.924**	58
	1000	**0.73**	315	**0.920**	109	0.924	96

Table 7. Attack performance under different max query limitations (MQ). Higher attack success rate (ASR), smaller average queries (AQ) indicate stronger attacks.

8. Limitations and Future Work

One limitation of the proposed method is that the attack may fail in practice even when the top-k images are subverted, especially when the number of truly relevant images in the gallery is large, as shown in Fig. 12. Apart from this, we also find the potential of the proposed QAIR may be limited due to the leverage of the substitute model. In OptAttack [3] or traditional RGF attack [11], they require lots of queries due to the randomness during optimization. Though we can improve the attack efficiency by leveraging the transfer-based priors as the guidance for optimization, the attack may fail due to the lack of adjustments of substitute model during attacks. Under this circumstance, more queries may not help much, as shown in Tab. 7. In future work, we aim to go further for a more advanced objective and interactive model stealing method towards stronger black-box attacks for developing robust image retrieval models.

We would like to thank Bing Visual Search for making their API public and available.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[6] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[7] 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.

[8] 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.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[10] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

[11] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[12] 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.

[13] 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.

[14] 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.