Attention-Guided Black-box Adversarial Attacks with Large-Scale Multiobjective Evolutionary Optimization

Anonymous Authors[1]

Abstract

Recent black-box adversarial attacks may struggle to balance their attack ability and visual quality of the generated adversarial examples (AEs) in tackling high-resolution images. In this paper, We propose an attention-guided black-box adversarial attack based on the large-scale multiobjective evolutionary optimization, termed as LMOA. By considering the spatial semantic information of images, we firstly take advantage of the attention map to determine the perturbed pixels. Then, a large-scale multiobjective evolutionary algorithm is employed to traverse the reduced pixels in the salient region. Extensive experimental results have verified the effectiveness of the proposed LMOA on the ImageNet dataset.

1. Introduction

In the past decade, a series of studies have shown that DNNs are vulnerable to adversarial examples (AEs) by imposing some designed perturbations to original images (Szegedy et al., 2013; Goodfellow et al., 2014b; Carlini & Wagner, 2017). These perturbations are imperceptible to human beings but can easily fool DNNs, which raises invisible threats to the vision-based automatic decision (Kurakin et al., 2016; Yin et al., 2020). Consequently, the robustness of DNNs encounters great challenges, and the issue of AEs has received considerable attention (Zhang & Li, 2019).

Szegedy *et al.* (2013)first pointed out the vulnerability of DNNs and proposed the definition of adversarial attacks. They also demonstrated that the AEs for one network could fool another, even DNNs were trained on different datasets. Then, a considerable amount of researches on adversarial attacks has been studied. These attacks are designed to fool the target DNN by adding a small perturbation $\boldsymbol{X}$ to the original image $\boldsymbol{I} : \mathcal{C}(\boldsymbol{I} + \boldsymbol{X}) \neq \mathcal{C}(\boldsymbol{I})$, where $\mathcal{C}(\cdot)$ is a m-class classifier that receives n-dimensional input and gives m-dimensional output. AEs can be easily generated by using internal information of the target DNN, *e.g.,* the gradient of the loss function of the original image (Goodfellow et al., 2014a; Kurakin et al., 2016; Madry et al., 2018; Dong et al., 2018). These attacks, called white-box attacks, are essentially an exploration of the robustness of DNNs. Other than the white-box attacks, researchers have shown an increased interest in black-box attacks. To be specific, the attacker can only obtain the output of the target DNN without accessing its structures and parameters. Since the structural prior knowledge of DNNs is usually unavailable, the works on black-box attacks are more practical than that of white-box ones in real cases. Therefore, numerous attempts have been made to realize black-box attacks(Bhambri et al., 2019), such as the gradient estimation-based (Tu et al., 2019), local search-based (Chen et al., 2019), or transferability of AEs-based (Dong et al., 2019) attacks.

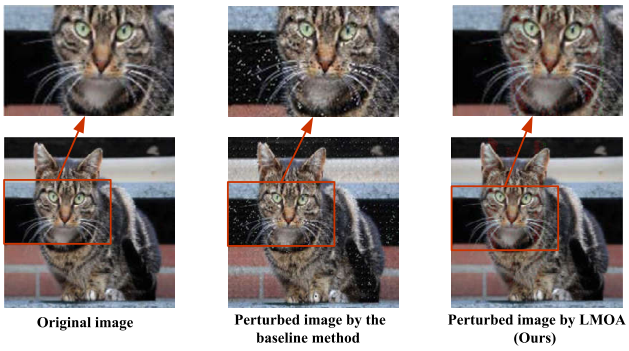


Figure 1. The original image and its corresponding AEs generated by the baseline (Suzuki et al., 2019) and LMOA, respectively.

To date, several works suggest that there is more than one objective should be taken into consideration in attacking, *e.g.,* minimizing the confidence probability of the true label and the perturbation intensity of the changed image simultaneously (Liu et al., 2020; Suzuki et al., 2019). They expected that the candidate AEs would mislead the target DNN as possible while exhibiting similar visual features with the original image(see Appendix for more details). Unfortu-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2021 workshop on A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning. Do not distribute.

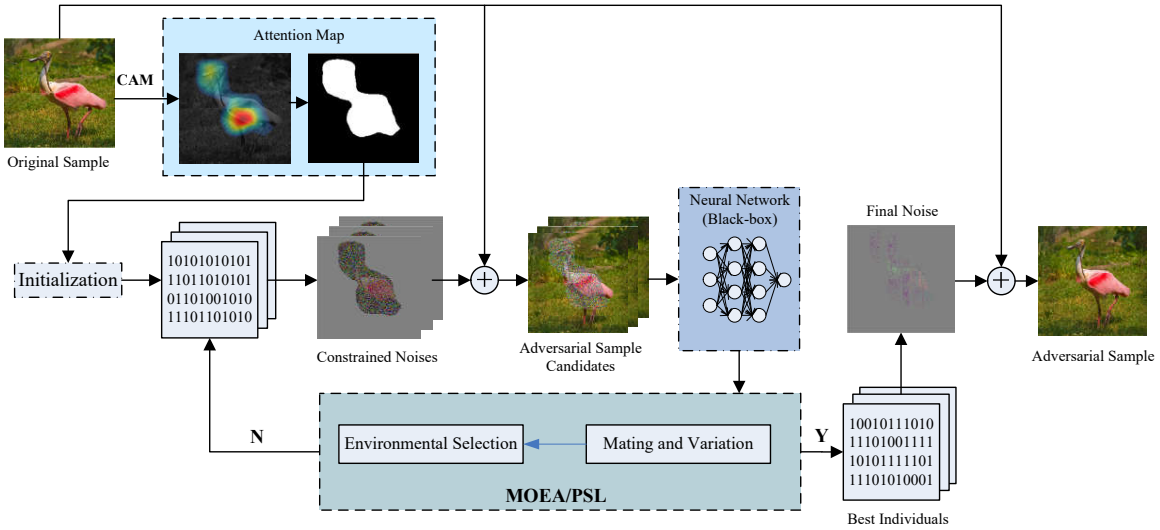


Figure 2. The general framework of the proposed LMOA.

nately, these two objectives are somewhat conflicting with each other. On the one hand, the great perturbations significantly influence the classification result of DNNs, but the generated AEs could be easily detected by the human vision, as they lose the majority of features of the original image. On the other hand, a slight change that is hardly observed by both human beings and computers is not enough to mislead DNNs. Most of the previous methods cannot balance these two objectives and thereby probably miss the optimal trade-off perturbation. The issue stimulates the efforts on evolutionary algorithm-based, especially multiobjective evolutionary algorithm-based black-box adversarial attacks. Moreover, since the traditional evolutionary algorithm are not competive on handling the optimization problems with large-scale decision variables (Yang et al., 2008), the existing works following this line may not achieve satisfactory performance on the high-resolution images(Su et al., 2019).

To overcome the above drawbacks(more details about the motivation in the Appendix), we propose an attention-guided black-box adversarial attack, where a large-scale multiobjective evolutionary algorithm is employed to traverse the salient region of an image. Since the proposed method involves Large-scale Multiobjective Optimization and Attentional mechanism, it is named by LMOA. The main contributions can be summarized as follows:

- **Using the attention mechanism to screen the attacked pixels.** We firstly use the class activation mapping (CAM) and a proxy model to obtain the attention map of the target image. The map strictly limits the attacked pixels so that the perturbations are only allowed to emerge within the salient region. On the one hand, attacking the salient pixels might be more efficient than the entire image(Dong et al., 2020), as these pixels can better reflect the spatial semantic information. On the other hand, screening the pixels is able to reduce the dimensionality of decision variables in the case of high-resolution images, which is beneficial to the black-box optimization.

- **Performing the black-box adversarial attack with a large-scale multiobjective evolutionary algorithm.** We secondly formulate the black-box attack into a large-scale multiobjective optimization problem, in which both attack ability and visual quality of the generated AEs are viewed as two objectives. Then, an optimizer tailored for large-scale optimization is employed. By doing so, a set of Pareto optimal solutions that achieves the balance between two objectives will be obtained, and the final generated perturbations can easily fool the target DNNs while being imperceptible by the human vision.

- **Attacking high-resolution images with high success rate and acceptable visual quality.** Extensive experimental results have been investigated on the ImageNet dataset. The results show that the proposed LMOA can achieve almost 100% success rate of attacks. Compared with the baseline method, LMOA is more competitive to contribute high-resolution AEs with better visual imperceptibility (see Fig. 1).

2. Proposed method

The general framework of the proposed LMOA is exhibited in Fig. 2. LMOA mainly consists of two steps: 1) Screening the perturbed pixels with the attention mechanism; 2) Generating the optimal perturbations with a large-scale MOEA.

2.1. Screening perturbed pixels with the attention mechanism

By considering the spatial semantic information, LMOA firstly employs the class activation mapping (CAM) (Zhou et al., 2016) to obtain the attention map of the target image. CAM can visualize predicted class scores on any given image, highlighting the region of the object detected by the target DNN. In other words, the obtained attention map reflects the pixels of the interest of DNNs in the classification. Obviously, attacking these pixels that contain the spatial semantic information can fool the target DNN with a relatively higher probability. However, it is tricky to know the gradient information of black-box DNNs, which brings a barrier for using CAM. Therefore, we suggest using a proxy model to obtain an approximated attention map of the input image.

To make the generated adversarial perturbations more effective, only the salient pixels derived from the attention map will be screened as the attacked pixels. To be specific, the attention map of the proxy model is binarized to represent the candidate pixels for attacking. By doing this, the dimension of the decision variables to be optimized is reduced, as the pixels for perturbation is changed. The reduction of the search space also facilitates the convergence of the subsequent MOEA.

2.2. Generating perturbations with large-scale MOEA

The black-box attack is firstly formulated as a multiobjective optimization problem below.

$$\begin{array}{ll} \min & f_1 = P(\mathcal{C}(\boldsymbol{I}+\boldsymbol{X}) = \mathcal{C}(\boldsymbol{I})) \\ \min & f_2 = \|\boldsymbol{X}\|_0 \\ \min & f_3 = \|\boldsymbol{X}\|_2 \\ \text{s.t.} & 0 \le u_i + x_i \le 255 \end{array} \tag{1}$$

where $P(\cdot)$ denotes the *confidence probability* of the classification result; $\boldsymbol{I}$ and $\boldsymbol{X}$ represent the *original sample* and *adversarial perturbation*, respectively; u_i is the value of pixel at the (l, w, c) position of $\boldsymbol{I}$, while x_i is the value of perturbation.

As shown in Eq. (1), the proposed multiobjective optimization based black-box attack involves three objective functions. The first one f_1 represents the probability that the target classifier $\mathcal{C}(\cdot)$ classifies the generated adversarial example $\boldsymbol{I}+\boldsymbol{X}$ into the correct class $\mathcal{C}(\boldsymbol{I})$. The remaining two functions are both distance metrics, each of which is employed to evaluate the similarity between $\boldsymbol{I}+\boldsymbol{X}$ and $\boldsymbol{I}$. Furthermore, minimizing l_0 distance (l_0 norm) is to restrict the number of pixels to be attacked, while minimizing l_2 distance aims to reduce the change of each pixel. Other than three objective functions, the constraint imposed to $\boldsymbol{X}$ defines the range of perturbation on each pixel based on the intrinsic property of images. Using such a constraint can effectively reduce the search space and facilitate the convergence of MOEA. Finally, we note that most of the decision variables in $\boldsymbol{X}$ are fixed as zero, and the dimension of perturbations to be optimized is relatively lower than that of $\boldsymbol{I}$. The reason is that the candidate attacked pixels are significantly reduced according to the technique introduced in the last subsection.

Note that, the objective f_2 is compatible with f_3 in some cases, *e.g.*, a solution with a small objective value on f_2 may also has an acceptable performance on f_3. Moreover, the value of f_2 reflects the sparsity of a solution, that is, minimizing f_2 is to find the most sparse adversarial attack. We thereby reformulate the black-box attack below and resort to one of MOEAs tailored for large-scale sparse multiobjective optimization problems (LSMOPs)..

$$\begin{array}{ll} \min & f_1 = P(\mathcal{C}(\boldsymbol{I}+\boldsymbol{X}) = \mathcal{C}(\boldsymbol{I})) \\ \min & f_2 = \|\boldsymbol{X}\|_2 \\ \text{s.t.} & 0 \le u_i + x_i \le 255 \end{array} \tag{2}$$

LSMOPs are characterized as the problems, where the value of most decision variables of their Pareto optimal solutions is zero, and the remaining large-scale variables are one or the other real numbers. Since recent methods can tackle this type of problems (Tian et al., 2020b), we employ the MOEA based on Pareto-optimal subspace learning (MOEA/PSL), which is recently proposed in (Tian et al., 2020a), to solve the problem depicted in Eq. (2).

3. Experiments and analysis

In this section, we frst introduce the experimental setup, in cluding the benchmark dataset and parameter setting. Then we compare the proposed algorithm with the work proposed in (Suzuki et al., 2019) in terms of attack ability and visual quality on benchmark dataset(see Appendix for more details). In the Appendix, the ablation study is investigated to verify the effectiveness of LMOA.

3.1. Experimental setup

The benchmark dataset contains 1000 high-resolution images that are randomly selected from ImageNet-1000 (Deng et al., 2009), which consists of 1000 categories in total. Two DNNs, including the pretrained ResNet-101 (He et al., 2016) and Inception-v3 (Szegedy et al., 2016), are selected as the target models for each compared algorithm. Before the optimization, the resolution of each image is resized according to the input layer of each model, *i.e.,* $224 \times 224 \times 3$ for ResNet-101 and $299 \times 299 \times 3$ for Inception-v3, respectively. All the experiments are carried out on a PC with Intel Core i7-6700K 4.0GHz CPU, 48GB RAM, Windows 10, and Matlab R2018b with PlatEMO (Tian et al., 2017).

Table 1. Classification results and corresponding confidences of the original and AEs.

Image No.	Recognition results and confidence							
	ResNet-101				Inception-v3			
	$\mathcal{C}(\boldsymbol{I})$		$\mathcal{C}(\boldsymbol{I}+\boldsymbol{X})$		$\mathcal{C}(\boldsymbol{I})$		$\mathcal{C}(\boldsymbol{I}+\boldsymbol{X})$	
I_1	Vulture:	62.05%	Kite:	52.37%	Kite:	15.30%	Kite:	70.98%
I_2	Hotdog:	95.69%	Cucumber:	47.46%	Banana:	10.03%	Banana:	59.93%
I_3	Street sign:	50.28%	Shopping cart:	57.47%	Street sign:	15.19%	Shopping cart:	88.76%
I_4	Wolf spider:	92.69%	Tarantula:	51.47%	Wolf spider:	93.70%	Barn spider:	43.05%
I_5	Knot:	97.67%	Swab:	53.19%	Knot:	92.69%	Swab:	42.84%
I_6	Fig:	81.77%	Jackfruit:	22.90%	Fig:	99.06%	Mushroom:	44.82%
I_7	Hip:	77.13%	pomegranate:	42.94%	Hip:	33.66%	Lenmon:	33.41%

Target model	Attack method	Classification accuracy (%)	Avg. l_2 norm
ResNet-101	N/A	77.10	N/A
	The baseline	9.30	73949.58
	LMOA	0.00	967.51
Inception-v3	N/A	78.40	N/A
	The baseline	15.60	152008.99
	LMOA	0.30	1186.34

Table 2. Comparison on the classification accuracy,and average l_2 norm of AEs between LMOA and the baseline (Suzuki et al., 2019).

3.2. Results and analysis

Table 1 shows the classification results and confidence probabilities of seven randomly selected original images and their corresponding AEs obtained with two DNNs. From the table, two remarks can be concluded as follows. Firstly, for the images that are correctly classified by the two DNNs, the proposed LMOA finally generates AEs that fool the models with high confidence probabilities (at least 22.9%, most of the results over 40%). Secondly, for the images that are misclassified by DNNs (I_1 and I_2 against Inception-v3), the proposed algorithm improves the confidence probabilities of the incorrect label (from 15.3% to 70.98% on I_1, 10.03% to 59.93% on I_2).

Table 2 compares LMOA and another MOEA-based black-box attack method (Suzuki et al., 2019), which adopts block-division method and formulates a MOP solved by MOEA/D (Zhang & Li, 2007). From the table, we can observe that 77.1% of the benchmark images can be correctly classified by ResNet-101, while 21.6% of them are misclassified by Inception-v3. After performing the attack with the baseline [35], only 9.3% and 15.6% of the images are correctly classified by the two DNNs, respectively. By contrast, LMOA has successfully attacked most of the images and fooled both of the two models. More concretely, LMOA has achieved a 100% success rate on ResNet-101. We also notice that LMOA performs better in terms of average l_2 norm between the original image and the generated AE. Fig. 3 also visualized attack results obtained by the two methods. From the figure, we show that the baseline method adds some visible noises to each entire image, which can be easily captured by the human vision. For LMOA, the perturbations are much more difficult to be perceived compared with the baseline method. The advantage is attributed to the usage of the attention mechanism and large-scale MOEA.

(a) AEs generated by the baseline method

(b) AEs generated by the proposed LMOA

Figure 3. The visual comparison of the AEs generated by the baseline method and the proposed LMOA.

4. Conclusions

In this paper, we propose a novel attention-guided black-box adversarial attack, where the adversarial perturbations are only added to several pixels in the salient region. Besides, MOEA/PSL is used to search for the optimal perturbation, and the algorithm is used to solve MOPs, where the Pareto optimal solutions are sparse. Experimental results show that the proposed LMOA can perform effective black-box attacks against high-resolution images (with nearly 100% success rate and high visual quality). Moreover, comparing with the baseline method, the proposed LMOA is more suitable to addressing high-resolution images from ImageNet dataset.

References

Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.-J., and Srivastava, M. B. Genattack: Practical black-box attacks with gradient-free optimization. In *Genetic and Evolutionary Computation Conference*, pp. 1111–1119, 2019.

Bhambri, S., Muku, S., Tulasi, A., and Balaji Buduru, A. A survey of black-box adversarial attacks on computer vision models, 2019.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Chen, J., Su, M., Shen, S., Xiong, H., and Zheng, H. Pobaga: Perturbation optimized black-box adversarial attacks via genetic algorithm. *Computers & Security*, pp. 89–106, 2019.

Chen, P., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

Deb, K. *Multi-objective optimisation using evolutionary algorithms: an introduction*. Springer, 2011.

Deb, K. and Goyal, M. A combined genetic adaptive search (geneas) for engineering design. *Computer Science and Informatics*, 26(4):30–45, 1996.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197, 2002.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Dong, X., Han, J., Chen, D., Liu, J., Bian, H., Ma, Z., Li, H., Wang, X., Zhang, W., and Yu, N. Robust superpixel-guided attentional adversarial attack. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12895–12904, 2020.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Dong, Y., Pang, T., Su, H., and Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014a.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples, 2014b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale, 2016.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, X., Hu, T., Ding, K., Bai, Y., Niu, W., and Lu, J. A black-box attack on neural networks based on swarm evolutionary algorithm. In *Australasian Conference on Information Security and Privacy*, pp. 268–284, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, pp. 1–28, 2018.

Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Suzuki, T., Takeshita, S., and Ono, S. Adversarial example generation using evolutionary multi-objective optimization. In *IEEE Congress on Evolutionary Computation*, pp. 2136–2144, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks, 2013. arXiv preprint arXiv:1312.6199.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tian, Y., Cheng, R., Zhang, X., and Jin, Y. PlatEMO: A MATLAB platform for evolutionary multi-objective optimization [educational forum]. *IEEE Computational Intelligence Magazine*, 12(4):73–87, 2017.

Tian, Y., Lu, C., Zhang, X., Tan, K. C., and Jin, Y. Solving large-scale sparse multiobjective optimization problems with sparse optimal solutions via unsupervised neural networks. *IEEE Transactions on Cybernetics*, pp. 1–14, 2020a.

Tian, Y., Zhang, X., Wang, C., and Jin, Y. An evolutionary algorithm for large-scale sparse multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 24(2):380–393, 2020b.

Tu, C., Ting, P., Chen, P., Liu, S., Zhang, H., Yi, J., Hsieh, C., and Cheng, S. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *AAAI Conference on Artificial Intelligence*, pp. 742–749, 2019.

Yang, Z., Tang, K., and Yao, X. Large scale evolutionary optimization using cooperative coevolution. *Information Sciences*, 178:2985–2999, 2008.

Yin, Z., Wang, H., Wang, J., Tang, J., and Wang, W. Defense against adversarial attacks by low−level image transformations. *International Journal of Intelligent Systems*, 2020. https://doi.org/10.1002/int.22258.

Zhang, J. and Li, C. Adversarial examples: Opportunities and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2578–2593, 2019.

Zhang, Q. and Li, H. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation*, 11(6):712–731, 2007.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

Zitzler, M., Laumanns, M., and Thiele, L. SPEA2: improving the strength Pareto evolutionary algorithm for multiobjective optimization. In *Proceedings of the fifth Conference on Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems*, pp. 95–100, 2001.

Appendix

A. Related work

In this section, we summarize recent approaches for generating AEs, in both the white-box and black-box cases. Besides, we emphasize existing evolutionary algorithm based attacks. Please refer to the cited works for further details.

A.1. White-box adversarial attacks

In white-box attacks, the attacker is allowed to obtain the structure and parameters of the target DNNs. Among the studies, the gradient-based approach is the most commonly used white-box attack. Goodfellow *et al*.(Goodfellow et al., 2014b) proposed a simple and fast attack, named fast gradient sign method (FGSM), where AEs are generated based on the classification loss gradient of the input image. In their following works, an iterative version of FGSM (IFGSM) is proposed to obtain better attack performance (Kurakin et al., 2016), whereas a momentum term was suggested in (Dong et al., 2018) to improve the transferability of adversarial samples. Other than the above works, a great deal of work has also been presented, aiming at obtaining AEs that are more robust to unknown and more defensive DNNs (Dong et al., 2019; 2020). However, the structural prior information of DNNs is usually unknown in realistic cases. Therefore, many researchers focus on the research toward black-box attacks.

A.2. Black-box adversarial attacks

Black-box attacks assume that the attacker only knows the outputs of the target DNN, *e.g.,* category labels or confidence scores. They are categorized into three main classes (Bhambri et al., 2019): gradient estimation-based (Chen et al., 2017), transferability-based (Dong et al., 2019), and local search-based methods (Alzantot et al., 2019). The gradient estimation-based methods perform adversarial attacks by estimating the gradient of the target model. Although they can effectively generate AEs, these methods highly rely on the accuracy of the estimated gradient. As for the transferability-based methods, the attacker usually utilizes the transferability of the target DNN, and trains an alternative/surrogate model for adversarial attacks. It is clear that this type of attack relaxes the necessity of the inner information of DNNs, but heavily depends on the transferability, which is sensitive to the mismatch between the alternative and target DNNs. The third class is the local search-based methods, which follow a query-feedback pattern. In detail, a candidate AE begins as a random input and changes by the disturbance noise with an acceptable level. The advantage of local search-based methods is that they can accomplish adversarial attacks without considering the gradient information and alternative model. However, their performance is also limited by the dimension and direction of the search.

A.3. Evolutionary algorithm based attacks

Evolutionary algorithm (EA) is a computational model that simulates the natural selection and genetic mechanisms of Darwinian evolutionary biology. The main features of EA are characterized as derivative-free and global search, enabling EA becomes a more suitable method for solving black-box adversarial attacks. Recently, Su *et al*.(2019) proposed an one-pixel attack based on differential evolution (DE), where only a limited number of pixels in the image are changed to construct AEs. Due to the usage of DE, this method does not require the gradient of the target DNN and can be integrated with a non-differentiated objective function. Note that, a few pixel changes cannot result in a significant decline of the classification confidence with large-scale pixels, such that the success rate of adversarial attacks is unfavorable in the case of high-resolution images. Liu *et al*. (2020) presented a EA-based method that focuses on attacking the entire image, rather than limiting the number of changed pixels. The optimization objective is formulated as a weight function, where the minimization of confidence of the original label is considered as the main objective, while the similarity between solutions and the original image is used as a penalty.

The above EA-based approaches usually formulate a single-objective constrained/unconstrained optimization model. In essence, the process of AEs generation requires considering more than one objective, such as the classification probability and the visual quality of the candidate AEs. Therefore, the task of black-box adversarial attacks can be reformulated as a multiobjective optimization problem (MOPs). Mathematically, a MOP is stated as (Deb et al., 2002)

$$\begin{aligned} \min \quad & \mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_m(\mathbf{x}))^{\mathrm{T}} \\ \text{s.t.} \quad & \mathbf{x} \in \Omega \end{aligned} \tag{3}$$

where Ω is the *decision space* and $\mathbf{F}$ consists of m objective functions, respectively. $\{\mathbf{F}(\mathbf{x}) | \mathbf{x} \in \Omega\}$ is the *attainable objective set*. Considering two decision vectors $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$, the vector $\mathbf{x}_1$ is said to *dominate* $\mathbf{x}_2$ (denoted by $\mathbf{x}_1 \prec \mathbf{x}_2$) if and only if $\forall i : f_i(\mathbf{x}_1) \leq f_i(\mathbf{x}_2) \wedge \exists j : f_j(\mathbf{x}_1) < f_j(\mathbf{x}_2)$, where $i, j \in \{1, \cdots, m\}$. Moreover, a solution $\mathbf{x}^*$ is *Pareto optimal*, if there is no other solution $\mathbf{x} \in \Omega$ such that $\mathbf{x} \prec \mathbf{x}^*$. The set of all the Pareto optimal solutions is defined as Pareto set (PS) and its image is called Pareto front (PF).

By contrast with single-objective EAs, multiobjective evolutionary algorithms (MOEAs) are characterized as population-based metaheuristics for handling MOP, each of which has more than one Pareto optimal solution. In this community, research efforts are devoted to discovering the tradeoff solutions, which are as close as possible to the Pareto optimal solutions and cover the Pareto fron-

t as uniformly as possible. As for MOEA-based attacks, Suzuki *et al*. (Suzuki et al., 2019) first demonstrated the feasibility of using MOEAs for black-box attacks in three scenarios. While this method can effective attack against low-dimensional images (*e.g.*, CIFAR-10), its performance will degenerate when the dimension of the optimization problem rises. Consequently, it is necessary to apply dimension reduction techniques or MOEAs tailored for large-scale tasks to attack high-resolution images (e.g., ImageNet).

B. Motivation

Based on the above descriptions, there are still some limitations in conventional studies on EA-based black-box attacks. Therefore, the motivations of this paper are summarized in the following two respects.

1) Screening the perturbed pixels can reduce the complexity of optimization while generating more effective AEs. Previous adversarial attacks usually focus on attacking the entire image, regardless of its spatial semantic information. It is also worth noting that the foreground region usually contains more texture information than the background one of an image. In other words, changing the pixels in the foreground region of an image is more effective than attacking the entire image (Dong et al., 2020). The fact motivated us to use an attention mechanism to restrict the attack within the salient pixels in the foreground region. More importantly, from the perspective of optimization, this implementation can greatly reduce the search space and the complexity of the optimization.

2) Previous EA-based black-box attacks encounter difficulty in tackling high-resolution images. Recent evidence suggests that the AEs generation can be formulated as a MOP. However, current EA-based approaches usually model the black-box attack as a constrained single-objective optimization problem or an unconstrained problem with weights or penalties. The studies cannot balance the solutions between two optimization objectives, and their optimization result is sensitive to the weights. Although several researchers have investigated the black-box attacks based on MOEAs, the performance of attacks will significantly degenerate as the dimension of the target image increases (see Fig. 1). As for a high-resolution image, the dimension of decision variables to be optimized can reach hundreds of thousands or more, such that the complexity of the optimization greatly increases. In this case, using traditional MOEAs cannot obtain an acceptable convergence performance. The issue inspires us to use one of the MOEAs tailored for large-scale MOPs.

C. Details of the MOEA/PSL

In general, Algorithm 1 presents the framework of

Algorithm 1 *MOEA/PSL*

Input: N (population size)
Output: P (final population)
$P \leftarrow$ *Initilization*(N);
Initialize the parameters ρ and K;
while termination critertion not fullfilled **do**
  $P' \leftarrow$ *MatingSelection*(P);
  $O \leftarrow$ *Variation*(P, P', ρ, K);
  $P \leftarrow$ *EnvironmentalSelection*$(P \cup O)$;
  $[\rho, K] \leftarrow$ *ParameterAdaption*(P, ρ) ;
end while
return: P

MOEA/PSL, which is very similar to the nondominated sorting genetic algorithm (NSGA-II) (Deb et al., 2002). In the beginning, the population P is initialized. It is worth noting that the candidate adversarial example is required to have a high similarity with the original image. Bearing this in mind, the population is randomly initialized at first, and then it is multiplied with a small number α, *e.g.,* $\alpha = 0.5$. Besides, two parameters ρ and K that will be detailed in the following contents are also initialized. In the main loop, N parents are selected according to their nondominated front numbers and crowding distances, through the binary tournament selection. Then, the offspring set O is generated via the population P and parent set P'. Then, N solutions with better nondominated front numbers and crowding distances survive from the combination of O and P'. Finally, the parameters ρ and K are adapted based on the new population P.

As for the representation, the length of each solution $\boldsymbol{x}$ is denoted by D, and each bit corresponds to one pixel. For instance, given a high-solution image with a size of $(224, 224, 3)$, D thereby equals to $224 \times 224 \times 3 = 150,528$ without using attention mechanism. To be specific, each solution $\boldsymbol{x}$ in P is represented by a binary vector $\boldsymbol{xb}$ and a real vector $\boldsymbol{xr}$, and the i th dimension ($1 \leq i \leq D$) of $\boldsymbol{x}$ is obtained by

$$x_i = xb_i \times xr_i \tag{4}$$

where xb_i indicates whether the perturbation on the pixel u_i is zero, and xr_i indicates the amount of perturbation on u_i. During the search, $\boldsymbol{xr}$ records the best perturbation on each pixel, while $\boldsymbol{xb}$ records the best positions that should be perturbed and controls the sparsity of adversarial perturbations. Before the generation of offspring solutions, $\boldsymbol{xb}$s and $\boldsymbol{xr}$s of all the nondominated solutions are collected and used to train a denoising autoencoder (DAE) and a restricted Boltzmann machine (RBM), respectively. The number of hidden layers is controlled by the parameter K ($K \ll D$) and iteratively updated by *ParameterAdaption*, which is detailed in (Tian et al., 2020a). By doing so, the dimensions of $\boldsymbol{xb}$ and $\boldsymbol{xr}$ can be reduced via two models, and two reduced vectors

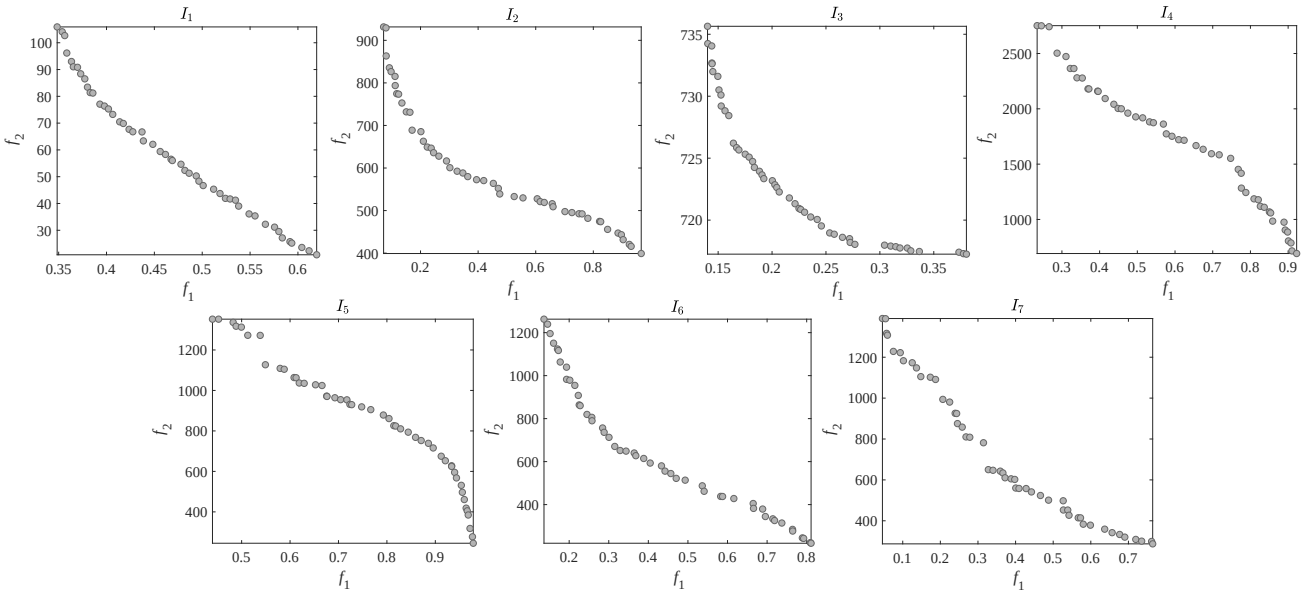


Figure 4. The Pareto optimal solutions obtained by LMOA with seven randomly selected images.

are therefore generated. Moreover, the reduced vectors can also be recovered to the original vectors. In *Variation*, two offspring solutions will be generated by simulated binary crossover (SBX) (Deb, 2011) and polynomial mutation (PM) (Deb & Goyal, 1996), when the parameter ρ is smaller than a random number. Otherwise, two parents will be fed into RBM and DAE to generate a pair of reduced vectors. Then, two new reduced vectors are accordingly obtained through SBX and PM, and they finally recovered as two offspring solutions. The update of the parameter ρ can also refer to (Tian et al., 2020a). To summarize, the aim of using DAE and RBM is to learn the Pareto-optimal subspace, which enables MOEA/PSL to evolve in a low-dimensional search space.

D. More details of experiments

D.1. The search results with LMOA

Fig. 4 shows the Pareto optimal solutions obtained by LMOA with seven images that randomly selected from the benchmark dataset. From the observations, we can conclude that the proposed LMOA can obtain the trade-off solutions between the two objectives including the confidence probability and l_2 norm. That is, the solution with a high confidence of the true label has a low l_2 norm, and vice versa. Note that, we select the perturbed image that successfully fools the target DNN while achieves the minimum l_1 norm as the final AE for each image.

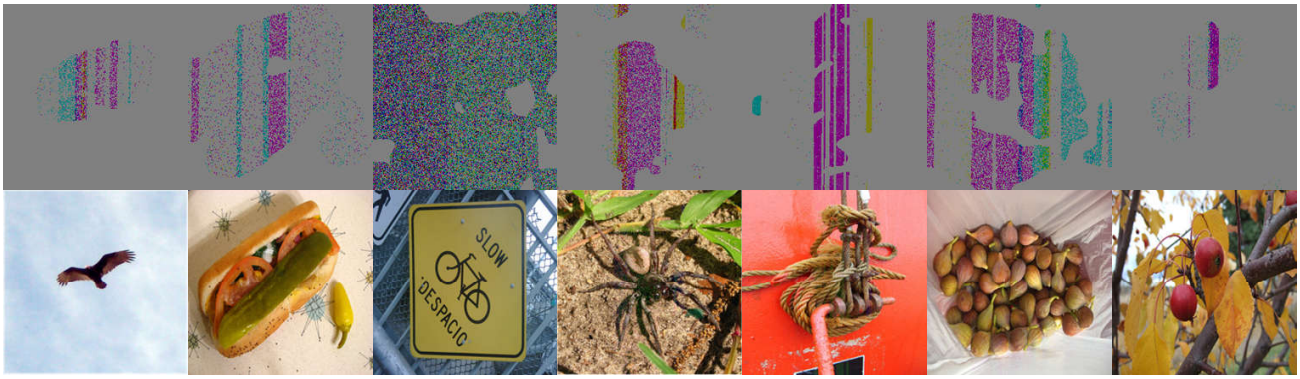
Fig. 5 further exhibits these AEs with the proposed LMOA and two DNNs. We note that the value of perturbations includes both positive and negative values. For better presentation, the unmodified pixels in each image are exhibited in gray pixels. Besides, the brighter pixels indicate increases in their intensity, while darker pixels represent changes in the opposite direction of their intensity. As can be seen in Fig. 5, the AEs generated by LMOA have similar visual perception compared with the original images. The reason is that the overall intensity of the perturbations is relatively low and the changes only emerge in part of the salient regions. Furthermore, even for the image that contains a simple background region with a sparse texture (I_1), the generated adversarial perturbation is still not easily perceptible. Therefore, the above evidence supports the analysis given in Section III, and it also verifies the effectiveness of the proposed method.

D.2. Comparison with the baseline method

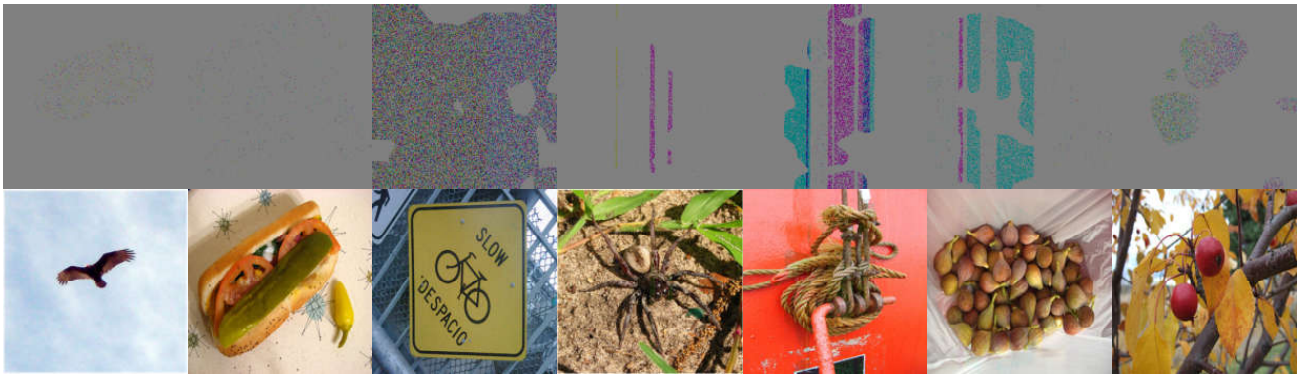
Fig. 6 further compares the adversarial attacks performed by the two methods. Similarly, the top row shows the generated AE and l_2 norm value on each channel obtained with the baseline method. The bottom row depicts the corresponding results of the proposed LMOA. From the figure, two observations are summarized as follows. Firstly, in terms of the region of perturbations, the baseline method attacks the entire image, resulting in easily perceptible perturbations. In contrast, this paper only perturbs the pixels in the salient region, which results in more sparse and smooth attacks. Secondly, the perturbation intensity of the bottom image is significantly smaller than that of the top one. This is also supported by the l_2 norm values of three channels.

I_1 I_2 I_3 I_4 I_5 I_6 I_7

(a) Original samples

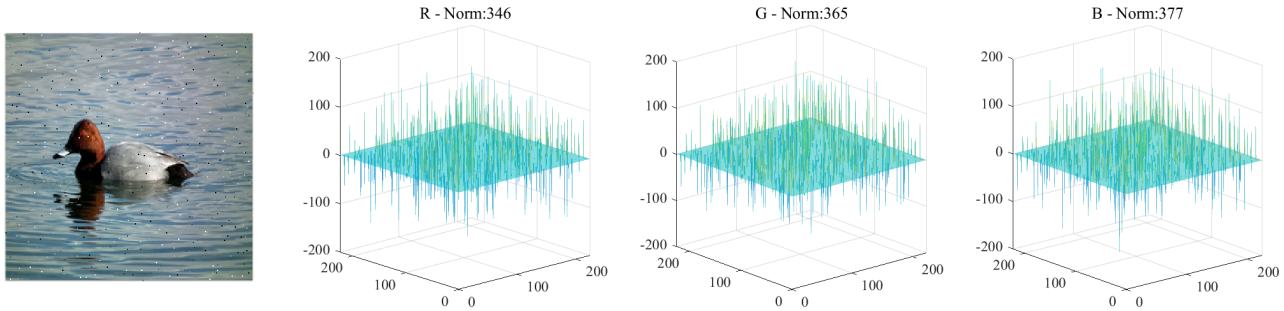


(b) Perturbation patterns and AEs generated by LMOA on ResNet-101

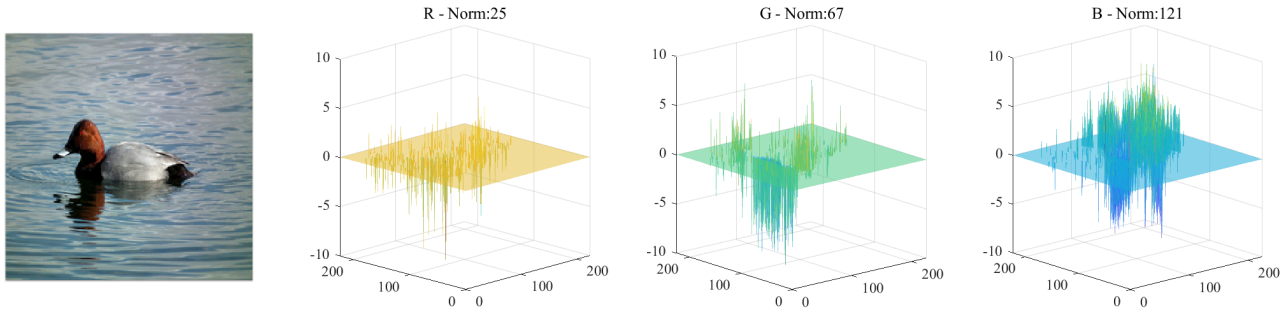


(c) Perturbation patterns and AEs generated by LMOA on Inception-v3

Figure 5. Illustration of AEs obtained by the proposed LMOA with different target DNNs.



(a) AE generated by the baseline method



(b) AE generated by LMOA

Figure 6. The perturbations of the baseline method and the proposed LMOA.

The fact indicates that the adversarial attack of LMOA is more effective than the baseline method on high-resolution images.

D.3. Ablation Study

In this part, we investigate the effectiveness of three components in LMOA through an ablation study.

D.3.1. Effectiveness of using attention guided search

The existing study has shown that attacking the salient region is more effective than attacking the background region (Dong et al., 2020). To verify the point, they performed the experiments, where the perturbations are imposed on the background and foreground regions separately. It was found that images that changed the foreground region can still fool the classifier, while another one failed. The above experiment result implies that adding perturbations to the salient region can get better attack performance. Moreover, we investigate the effectiveness of the attention mechanism when searching for the optimal adversarial attacks with MOEAs in this part. In other words, we attempt to explain the advantage of using the attention mechanism for black-box adversarial attacks with a new viewpoint.

We randomly pick an image from the benchmark dataset and perform the search by MOEA/PSL with the following two manners. The one is attacking the image without the guide of the attention mechanism, *i.e.,* the search space of the perturbation is the entire image. The other one aims at searching for the optimal perturbation guided by the attention map, which means that the region of perturbation is limited by the attention map.

Fig. 7 compares the results of two search paradigms. The original image, the AE obtained with 5 generations, and the AE obtained with 200 generations are shown from left to right in order. It can be seen that the AE obtained with the attention mechanism guided search contains significantly fewer noise than another one, regardless of the number of generations. The fact shows that restricting the perturbation region can facilitate the convergence of MOEA/PSL, which provides a more effective and invisible adversarial attack on the selected image. We also note that the reduction of the search space can greatly alleviate the computational cost of adversarial attacks with large-scale MOEAs.

D.3.2. Effectiveness of using Large-scale MOEA

This paper employs one of the MOEAs that are specially designed for tackling large-scale MOPs, due to the high

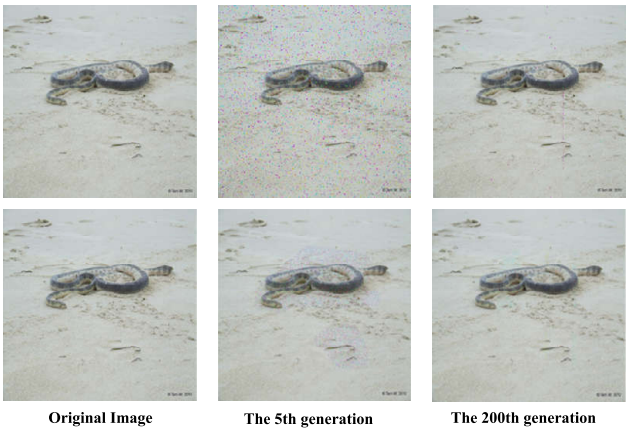
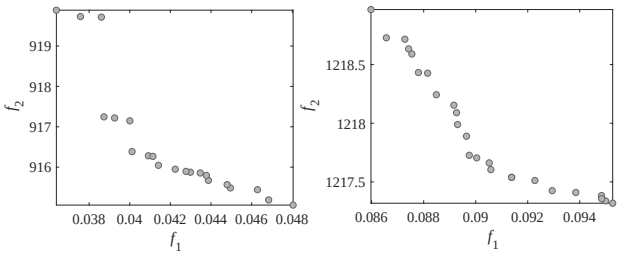
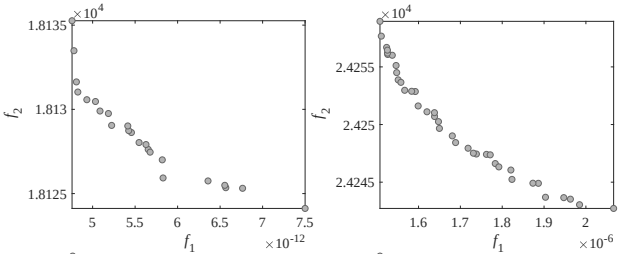


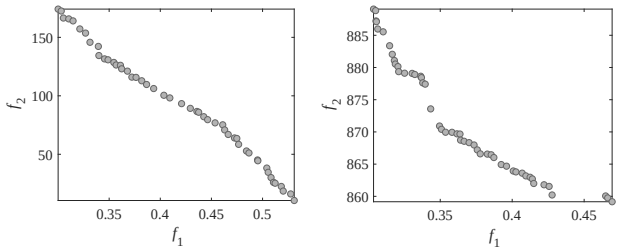
Figure 7. Illustration of the evolving processes of MOEA/PSL without or with the attention mechanism.



(a) The Pareto optimal solutions obtained by NSGA-II



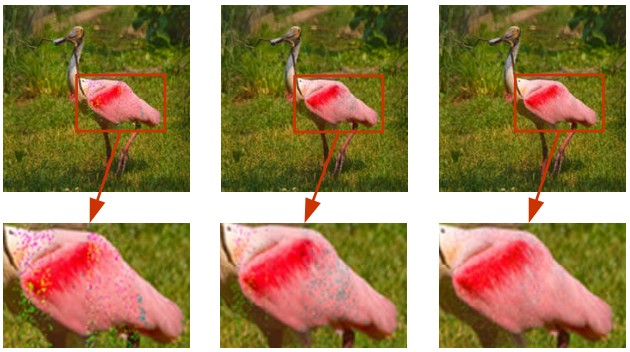
(b) The Pareto optimal solutions obtained by SPEA2



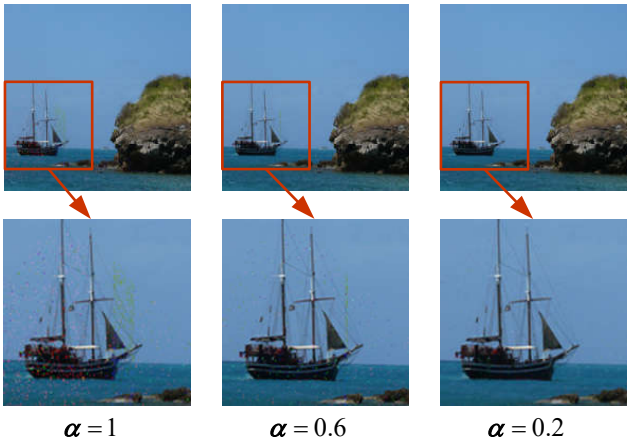
(c) The Pareto optimal solutions obtained by MOEA/PSL

Figure 8. The Pareto optimal solutions obtained by three MOEAs on ResNet-101 and Inception-v3 with two randomly selected images.

dimension of the optimization problem formulated for the high-resolution images. To verify the necessity of this technique, we further investigate the performance of the other two popular MOEAs, including NSGA-II (Deb et al., 2002), and SPEA2 (Zitzler et al., 2001), on the proposed MOP.



(a) AEs generated by LMOA with ResNet-101



(b) AEs generated by LMOA with Inception-v3

Figure 9. Illustration of the sensitivity of the parameter α involved in the improved population initialization.

Fig. 8 compares the Pareto optimal solutions obtained with three MOEAs on two DNNs, respectively. We note that using each of the three MOEAs can successfully attack the images in the experiment. Although both NSGA-II and SPEA2 are able to generate AEs with low confidence probabilities with respect to the true label, they fail to find out the AEs with acceptable l_2 norm values, especially SPEA2. Compared with the above two MOEAs, MOEA/PSL can search for the perturbations with lower intensities on each image, while accomplishing the adversarial attack. This result may be explained by the fact that traditional MOEAs may struggle to the large-scale MOPs, and they are not very suitable for attacking high-resolution images under the black-box settings. The above observations also confirm the effectiveness of using MOEAs tailored for large-scale

MOPs in this paper.

D.3.3. Effectiveness of the improved population initialization

As described in Section III, the initialized population is required to be multiplied with a small number α as the candidate AEs should maintain a high similarity with the original image. As known, a suitable initialization may help to result in better convergence of MOEAs. In this part, we will show that the improved population initialization can greatly influence the search results of LMOA.

Fig. 9 shows a comparison among the generated AEs with different values of α. It is clear that the AEs without the improved population initialization is much easier to be perceived. By contrast, the AEs with $\alpha = 0.6$ and $\alpha = 0.2$ achieves better visual masking. The observations indicate that selecting a proper α can generate an imperceptible AE while ensuring the success rate. Note that, the value of α can affect the success rate of the attack. An extremely small α is unfavorable to generate AEs that misclassify the target network. In this paper, we use $\alpha = 0.2$ for every image.