reply for pv33

Q1: The reviewer questions the absence of an efficiency comparison between the current attack and previous ones regarding query numbers, the rationale behind fixed query limits, and the lack of a queries vs. attack strength analysis for identifying the Pareto optimal curve.

A1: Thanks for your comments. We consider adding the attack intensity of different query budgets to our experimental scenario. An increase in query budgets does increase the intensity of attacks. The comparison in Table 2 was designed to demonstrate the efficiency of our method under a specific set of conditions. We compare the performance of all algorithms with a limited query budget. As far as we know in the baseline, the setting of this query budget is not very strict, and can effectively compare the differences between different algorithms.

We set different query budget that is 5,000 , 10,000, 15,000, 20,000, 25,000 instead of fixed budget. Notwithstanding the inherent disparities in our respective problem, we divided three sets of experiments for different problem(Global attack, Region-wise attack, Fixed attack). First, we compared with global attack methods, i.e. Square attack, Parsimonious attack and ZO-NGD attack, with ℓ_∞ constraints in Fig.1. Thus, we set the same ℓ_∞ constraint for all algorithms is 0.05. Second, we compared with region-wise attack, i.e. Patch-RS, with a patch in Fig.2. We set the patch size of the patch to be 80 in Patch-RS. The ℓ_0 constraint is 19200($80 \times 80 \times 3$). Finally, we compared with fixed-version of global attack algorithms in Fig. 3. For all algorithms, we set the same double constraint, ℓ_∞ is 0.1, ℓ_0 is 26820 for ImageNet dataset, Inceptionv3 model, and ℓ_∞ is 0.1, ℓ_0 is 15052 for ImageNet dataset, vision transformer model. That is, for double constraint, we perturb only 10% of the pixels.

Fig1 Global attack algorithms in different query budget

From the Fig. 1, all algorithms have ℓ_∞ constraint 0.05. We can see that the success rate of the attack will decrease significantly if the query is less than 15,000 for high-resolution image attack tasks. When the query budget reaches 20,000, the success rate becomes stable.

Fig2 Region-wise attack algorithms in different query budget

From the Fig. 2, all algorithms perturb up to 19200 of the pixels. Patch-RS is a heuristic attack method, it can be seen from the Fig. 2 ViT-B model that Patch-RS performs very poorly under strict query budget.

Fig3 Fixed attack algorithms in different query budget

From the Fig 3, global attack algorithms with fixed region perform poorly under different budgets. Due to time constraints and according to the attack performance of different regions in Table 18, we only provide the results of attacks in the central area.

Q2. Are there diminishing returns in attack success with higher resolution? While the proposed attack appears to be stronger and less perceptible than baselines of small resolution datasets (cifar10, mnist), the trend doesn't fully hold on ImageNet dataset (table 14 in appendix). Square attack [1] achieves equally high success rate and lower average queries.

A2: High-resolution images often contain more complex and varied features, which might affect the efficiency and success rate of adversarial attacks compared to smaller resolution datasets. As shown in Table 14 in the appendix, the attack performance trend observed on CIFAR10 and MNIST does not fully extend to the ImageNet dataset . *In the latest results provided, we adopt the variance reduction method to estimate the gradient in the black box more effectively, and the results are shown in the table.[To be update]*

The problem addressed within the realms of the Square attack substantially diverges from the context elucidated within our paper. In our paper, our deliberations pivot around the conceptualization of regionally sparse attacks, a paradigm wherein we endow the capability to autonomously designate target regions for assault while concurrently regulating the number of regions to be subjected to such perturbation. In stark contrast, the Square attack methodology is fundamentally engrossed in the task of imbuing perturbations across the entire expanse of an image, bereft of the nuanced capacity to pinpoint specific regions meriting subjection to attack.

From the table 14 in the appendix, if we perturb all the pixels like Square attack, we can achieve similar success rates and queries to Square attack. However, from Tab.16 and Tab. 18 in the appendix, we can seen that Square attack doesn't do well with double constraint scenarios.

Q3: Can authors clarify how the attack complexity behaves with experimental setup, i.e., input resolutions, size of neural networks, number of classes, etc?

A3: Our experimental scenario is set up as an unknown black box attack, and we get no information about the model other than output. This is a big challenge for gradient-based attack algorithms, and we want to generate sparse and imperceptible perturbations, and the strict requirements make the number of queries significantly higher.

We tested our method on three datasets: MNIST, CIFAR10, and ImageNet, each varying in image resolution and complexity. In order to preserve the inherent characteristics of the baseline algorithm, we designed three experiments to comprehensively evaluate the performance of the algorithm. Our

evaluation models include both simple CNN models and complex models(Inception-v3, ViT-B) on high-resolution images. The results showed varying degrees of attack success and efficiency across these different setups. We have systematically analyzed these three different attack modes.

In the follow-up work, we adopted the variance reduction technology to effectively improve the query efficiency. Although our work still has some shortcomings, it also provides the direction for automatic region selection attack algorithm.

Q4: Can authors provide additional intuition of why the proposed approach has higher ASR than similar black-box attacks, e.g., square attacks? Is it because of attack strength or subtle design choices, such as patch based perturbations.

A4: First, our approach utilizes automatic region selection, enabled by the ℓ_0^G constraint, which represents structured sparsity defined on a collection of groups G. This automatic detection of the regions that need to be perturbed is a key differentiator from heuristic-based approaches like square attacks. It allows for more targeted and effective perturbations, increasing the ASR.

Second, we reformulated the function f as an optimization problem with ℓ_0^G and ℓ_∞ constraint. This formulation, combined with the use of natural evolution strategies and search gradients, provides a robust framework for generating perturbations that are more aligned with the target, contributing to a higher ASR.

Finally, for non-overlapping groups G, our method provided a closed-form solution to the first-order Taylor approximation of the objective function. In cases where G is overlapping, an approximate solution is derived. This dual approach ensures effective perturbation generation across various scenarios, enhancing the overall ASR.

reply for jvb5

Q1: Some descriptions are confusing. The deduction in sec. 3.1 seems to be unnecessary, and the gradient estimation proposed in equ. (4) does not seem to differ from the standard gradient estimation approach.

A1: Thanks for your correction, we have fixed these errors in the text.

Thank you for your observations regarding the deduction in Section 3.1 and the gradient estimation method in Equation (4). We integrate this estimation technique with the Natural Evolutionary

Strategies (NES). This integration is key to efficiently navigating the complex search space inherent in our proposed method. We will shorten the content of this section.

Q2: For computational cost and convergence, it only says high, medium and low. Are there any quantitative results to demonstrate it?

A2: In Table 1, we opted for qualitative descriptors (high, medium, low) to provide an initial, high-level comparative overview of our approach against others. This was intended to give readers a quick reference point for understanding the relative computational demands and convergence rates. We acknowledge that this presentation is flawed, and we will revise the table in the main text so that readers can better understand the different attack patterns. For cost analysis, we provide quantified results for reference in Appendix B.3.

Q3: It is not clear how the algorithm performs region selection. In algorithm 1, how the delta^0 is initialised? What is the initial perturbation group set G?

A3: We begin by dividing the input image into several potential regions based on predefined criteria. Our grouping method is to divide groups by a fixed sliding window. It can generate different perturbations by adjusting the window size and step size. In the section 3.3, we explain the process of the algorithm in detail, where the 9 line calculates the **DIS** of all groups according to the mask I_G. Then, k groups with the smallest **DIS** are selected by algorithm 2. The selection process is iterative and adaptive, meaning it can change in subsequent iterations based on the feedback received from the model's responses to previous perturbations.

In the version provided in the paper, we set the initial value of δ to be a uniformly distributed random disturbance. The initial perturbation group set G is empty.

Q4: Using the estimated gradient to perform black-box adversarial attacks is not new. Please refer to the SPSA attack [1].

A4: Thank you for pointing out the relevance of the SPSA attack in the context of our work. Our work contributes to the field by extending the concept of gradient estimation to more specific and challenging scenarios of adversarial attacks. So black box gradient attack method is not our main work and innovation point. Our method specifically addresses the challenge of region-wise adversarial attacks, focusing on improving efficiency and imperceptibility in this narrower domain. We propose specific algorithmic modifications and enhancements that are tailored to the unique requirements of region-wise attacks. This includes adaptations to handle structured sparsity constraints and the ℓ_∞ norm in a black-box setting. Our approach integrates the gradient estimation with Natural Evolutionary

Strategies (NES), creating a novel synergy that enhances the efficacy of our attack method in terms of query efficiency and perturbation control.

Q5: In the experiment section. I am not sure if the comparison is fair, as different black-box attacks select different regions. Besides, the result shows that the proposed methods may actually require more queries as the median is significantly higher than other models.

A5: According to the original characteristics of the baseline, we divided the experiment into three sets(Global attack, Region-wise attack, Fixed attack).

- In the global attack, all the baseline algorithms have only ℓ_∞ constraints, so the sparsity and undetectability of perturbations are measured by ℓ_0 and ℓ_2 metrics, respectively.
- In the region-wise attack, all the baseline algorithms have only ℓ_0 constraints, so the undetectability of perturbations is measured by ℓ_∞ and ℓ_2 metrics.
- In the fixed attack, all the baseline algorithms have $\ell_{0+\infty}$ constraint, thus, it just need to compare the ASR and query efficiency.

In the follow-up work, we adopted the variance reduction technology to effectively improve the query efficiency. We provide the latest results in the article.

Q6: Also, the authors used a simple CNN for the CIFAR10 dataset. It would be more convincing to evaluate pre-trained models from PyTorch model zoo or other resources.

A6: We evaluated the CIFAR10 dataset on two models, resnet18 and mobilenet_v2. The results of the evaluation are shown in the table below. In this evaluation, we also adopted variance reduction technology, and its query efficiency and attack success rate have been effectively improved.

Table Performance on resnet18 and mobilenet_v2

Q7: As the authors conducted a convergence analysis of the proposed attack. I am wondering if this can be further developed towards an adversarial verification method. Also, the robustness verification of adversarial patches has been done in [2]. I am also interested in the performance of the proposed attack on such a certified defence.

[To update]

reply for apqw

Q1: Please put the imagenet result into mainpart of paper. add median number of queries for your ImageNet experiments in Table 14 as you did for CIFAR10 and MNIST in Tables 3, 4. Just the average number of queries is not sufficient in my opinion.

A1: Thank you for your comments. We have modified the structure of the article according to your suggestion, adding the Median item in Table 14.

Q2: Figure 5 is rather misleading because for the Square Attack in the second row we only see stripe initializaton without any sampled squares.

A2: Thank you for your suggestion. We have amend the figure in the Figure 5 of the main paper.

Q3: Could you elaborate on why fixed versions of existing attacks (e. g. Fixed-ZO-NGD) would be valuable baselines? The attacks were not designed that way and introducing this additional constraint seems to be an unclear step to me.

A3: In order to preserve the inherent characteristics of the baseline algorithm, we have made conservative changes to the existing algorithm (fixed areas), and in order to create as fair a comparison environment as possible, we also provide attack results for other areas in the appendix.

As far as we know, in recent years, there are fewer SOTA black-box regional attack algorithms that conform to the same scenario setting, so we change the SOTA global attack algorithm to fixed version. The good thing about this is, first of all, fixed versions of existing attacks provide a valuable baseline as they allow for a direct comparison under similar constraints. Secondly, it also effectively measures the attack performance of SOTA global attacks with the perturbation quantity limitation. As can be seen from Table 4, both the attack success rate and the query efficiency show a significant decline after the disturbance quantity limitation is added, thus highlighting the challenge of our paper. Besides, from the results of the five positions provided in appendix, it can be seen that most of the best performance is concentrated in the center area, but it is still far behind us.

By comparing against these adapted baselines, the unique contributions and advantages of our method become more apparent. This includes demonstrating how automatic region selection can outperform traditional methods that do not focus on region perturbations. While these attacks were not originally designed with fixed constraints, adapting them in this manner helps to create a more comprehensive and robust evaluation framework. This approach is not intended to undermine the

original design of these attacks but rather to provide a clearer context for evaluating the specific advancements our method offers.

Q4: Why would considering ℓ_∞ and ℓ_2 metrics simultaneously e. g. in Table 2 be significant?

A4: The goal of our method is to generate sparse and imperceptible perturbations. Therefore, we designed two constraints i.e. ℓ_∞ and ℓ_0^G , where ℓ_0^G is the group sparsity and ℓ_∞ is the constraint that the control perturbation is not perceptible. The ℓ_2 norm is an metric to measure the imperceptibility of perturbations outside the constraint.

According to the original characteristics of the baseline, we divided the experiment into three sets(Global attack, Region-wise attack, Fixed attack).

- In the global attack, all the baseline algorithms have only ℓ_∞ constraints, so the sparsity and undetectability of perturbations are measured by ℓ_0 and ℓ_2 metrics, respectively.
- In the region-wise attack, all the baseline algorithms have only ℓ_0 constraints, so the undetectability of perturbations is measured by ℓ_∞ and ℓ_2 metrics.
- In the fixed attack, all the baseline algorithms have $\ell_{0+\infty}$ constraint, thus, it just need to compare the ASR and query efficiency.

Q5: If we wanted to minimize them simultaneously with the baseline attacks that you consider, we could include it as another term in the loss that they are trying to optimize. Have you considered such modifications to obtain better baselines?

[waiting for experiment result]

reply for 9mc2

Q1: To minimize $F(x_0 + \delta, y)$ under some constraints, the authors solve the problem (5) derived from inequalities originating from the smoothness of F and the Lipschitz continuity of the gradient. However, it is important to note that problem (5) does not necessarily entail the minimization of $F(x_0 + \delta, y)$.

A1: Thank you for your comments. Directly minimizing $F(x_0 + \delta, y)$ is not possible, so we relax this problem by miniming its upper bound, which is derived from the restricted L-smoothness condition. As

a result, The problem (5) serves as an intermediary towards achieving the ultimate goal of minimizing $F(x_0+\delta, y)$ under the given constraints. It has better properties to study optimization problem. Problem 5 is actually a proximal operator, it can be written as $prox(\delta) = min\frac{L}{2}\|\delta - S_L(\delta^t)\|_2^2 + h(\delta)$, where $h(\delta) = 0$ if $\|\delta\|_0^G \leq k, l \leq \delta \leq u$, otherwise $h(\delta) = \infty$. Thus, the problem (5) is closely linked to the optimization process in adversarial attacks. From this point of view, it is necessary to minimize the problem (5). We'll make changes to minimize $F(x_0 + y)$ in the paper.

Q2: Is the assumption 1 reasonable for adversarial attacks? Particularly in the context of adversarial attacks involving complex neural networks.

A2: Thank you for your valuable comments. Assumption 1 involves Restricted Strong Convexity (RSC) and Restricted Strong Smoothness (RSS), which are properties typically assumed in high-dimensional statistical theory. They imply that the objective function behaves like a strongly convex and smooth function over a sparse domain, even if the function itself is non-convex. Thus, assumption 1 is much weaker than the general strongly convex and smooth condition, which is common and general in the study of ℓ_0 problems. This problem itself is NP-hard, and despite the flaws in our assumption, we can still gain some insights from it.

In adversarial settings, the objective often involves crafting inputs that cause the network to misclassify. This process can be viewed through the lens of optimizing a loss function that characterizes the discrepancy between the current output and the desired adversarial outcome. The assumption 1 is more relevant to optimization problems, particularly in specific subsets of the domain, and is useful if the standard convexity may not hold. Thus, assumption 1 is relatively plausible in the adversarial attacks.

Q3: Does theorem 2 provide meaningful performance bound of the algorithm 1 in practice?

A3: Thank you for your insight regarding the tightness of the bound and the role of ρ. First, Theorem 2's guarantees are based on two critical assumptions: RSC,RSS and the boundedness of the function $f(x_0+\delta, y)$. The theorem establishes a geometric convergence rate for Algorithm 1, where ρ is a critical factor. The concern that ρ might be greater than 1 and hence could cause the bound to diverge as $T \to \infty$ is valid in a theoretical sense. In practical scenarios, ρ is a factor derived from the assumptions and the structure of the algorithm. It is not an arbitrary variable but is grounded in the algorithm's design and the conditions under which it operates. We acknowledge that the condition of $\rho \leq 1$ is critical for the convergence guarantee. Thus, we provide further analysis at Remark 1, where if you set $k = \frac{L_{k+k^*}^2}{\alpha_{k+k^*}^2}k^* + \hat{k}$, then $\rho = 1 - \frac{\alpha^2}{L^2} \leq 0$. In other words, we can make $\rho \leq 1$ by taking the value of k.