ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

SPARC: Superpixel-based Black-Box Adversarial Attack with Regional Confidence

Tram Ho, Ngoc-Thao Nguyen and Bac Le✉

Faculty of Information Technology, University of Science,
Vietnam National University, Ho Chi Minh City, Vietnam
Email: 23120421@student.hcmus.edu.vn,
nnthao@fit.hcmus.edu.vn, lhbac@fit.hcmus.edu.vn

Abstract. Deep learning models for critical vision tasks remain vulnerable to adversarial attacks. We present SPARC, the first superpixel-targeted black-box attack offering high interpretability and strong performance. Our method uses regional confidence maps to guide perturbations to the most important regions and controls their magnitude using L_2 and L_1 constraints, keeping changes small and spatially coherent. In targeted attacks, SPARC achieves a competitive success rate with subtle perturbations, while in untargeted attacks, it attains the highest success among superpixel-based and score-based methods with the fewest black-box queries. SPARC provides a practical balance of performance and interpretability, suitable for real-world black-box scenarios.

1 Introduction

Adversarial examples demonstrate that subtle, imperceptible perturbations can deceive deep neural networks, underscoring their sensitivity and the necessity of robustness [1]. Attacks are categorized by attacker knowledge: white-box attacks provide full access to the model and gradients, gray-box attacks expose the model but not input transformations, and black-box attacks limit access to model outputs and, at most, the test dataset, requiring external queries or surrogate models to craft adversarial examples [14]. This work focuses on black-box attacks, reflecting realistic scenarios with restricted model access.

Our method belongs to the class of black-box score-based attack and leverages superpixel segmentation with random-search strategies. Existing baseline methods include query-based attacks such as ParsiBA [9], SimBA [5], Square-attack [6], and BABIES-DCT [7], which achieve high success rates but typically require thousands of queries. A separate line of work investigates superpixel-based attacks, including M-SAI-FGM [2], Superpixel Attack [3], and Saliency [11], which improve interpretability but generally underperform pixel-wise methods in attack success. To address these limitations, we propose SPARC, a superpixel-guided black-box attack that reconciles interpretability and effectiveness. Our contributions are as follows:

- **Leading Score-based Black-Box attack (untargeted):** Delivers the highest success rate among superpixel-based and score-based methods while requiring the fewest black-box queries.
- **First superpixel-targeted attack:** Achieves competitive targeted success (65.7%) with minimal perturbation ($L_2 = 0.482$) and the lowest query

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

count (1,614), highlighting efficiency and subtlety, while slightly trading off success rate compared to pixel-level methods.

- **Confidence-guided controlled perturbation:** Allocates perturbations according to model confidence, yielding interpretable attacks and regulating $L2$ growth via $L1$ constraints for small updates.

2 Related Work

Black-box score-based attacks exploit model outputs to craft adversarial examples via iterative optimization. Random-search methods, including ParsiBA [9] (combinatorial updates), SimBA [5] (coordinate-wise perturbations), Square Attack [6] (random square updates), and evolutionary strategies like BABIES [7], operate at the pixel level, lacking interpretability and applying uniform perturbations despite varying pixel importance between foreground and background regions. Superpixel-based methods such as M-SAI-FGM [2], Superpixel Attack [3], and Saliency [11] reduce redundancy by perturbing superpixels, often guided by CAM [12], but rely on uniform perturbations per superpixel, limiting fine-grained control. In contrast, our method uses RISE [15] to estimate attention via random masking and distributes perturbations based on normalized confidence, enabling more targeted and effective attacks.

3 Methodology

3.1 Problem Definition

Let $f : x \mapsto p \in [0,1]^c$ be an image classifier, where p denotes the output scores or probabilities over c classes, with $f(x) = y = \arg\max(p)$. The goal is to find a perturbation δ such that either $f(x+\delta) \neq y$ (untargeted) or $f(x+\delta) = y^*$ (targeted), equivalent to reducing $p_f(y \mid x+\delta)$ or increasing $p_f(y^* \mid x+\delta)$. Typical score-based black-box attacks iteratively query the model for the output probability vector while applying gradually evolving perturbations. Attacks are evaluated by (i) query count, (ii) success rate, and (iii) perturbation size (e.g. $\|\delta\|_2$ or $\|\delta\|_\infty$), usually subject to a similarity constraint $\|\delta\|_p \leq \epsilon$.

We propose a superpixel-constrained, importance-weighted black-box attack that enforces constant perturbations within superpixels (exploiting local pixel correlation), prioritizes the most discriminative superpixels for the predicted class. This approach yields stronger, more imperceptible adversarial examples with faster convergence and greatly reduced query complexity compared to global pixel-wise perturbations.

3.2 Perturbation Formulation

We partition the image I into K superpixels $\{S_1, S_2, ..., S_K\}$, applying uniform perturbation δ_k within each superpixel S_k to ensure spatial coherence. The importance of each superpixel is measured by the confidence drop when occluded: $\phi(S_k) = f_{c^*}(I) - f_{c^*}(I \setminus S_k)$, where c^* is the original predicted class.

For uniform perturbations within superpixels, the norms satisfy:

$$\|\delta\|_1 = \sum_{k=1}^{K} n_k \|\delta_k\|_1, \qquad \|\delta\|_2 = \Big(\sum_{k=1}^{K} n_k \|\delta_k\|_2^2\Big)^{1/2},$$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

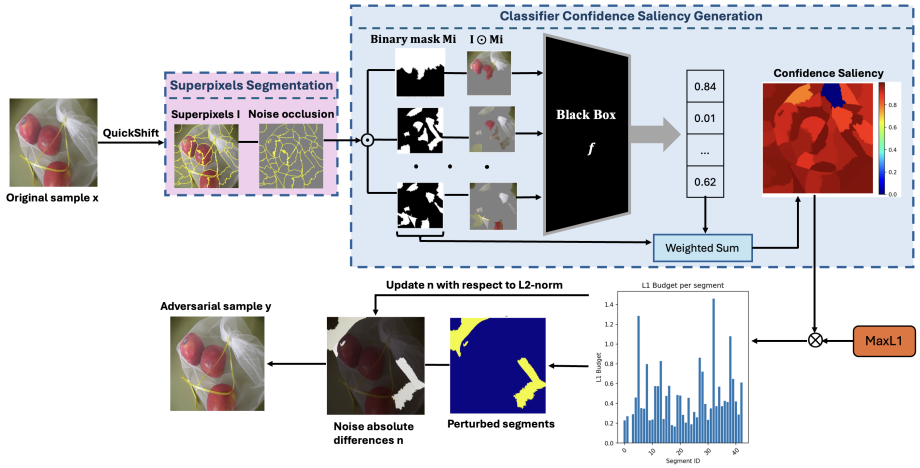

Fig. 1: Pipeline for our superpixel-based adversarial attack: superpixel segmentation, saliency estimation, and ℓ_1-guided perturbation respecting the ℓ_2-norm.

where n_k is the number of pixels in superpixel k. If $\|\delta_k\|_1 \approx \|\delta_k\|_2$ for small per-superpixel perturbation, the ratio grows as:

$$\frac{\|\delta\|_1}{\|\delta\|_2} \approx \frac{\sum_{k=1}^{K} n_k \epsilon_k}{\sqrt{\sum_{k=1}^{K} n_k \epsilon_k^2}} \sim \sqrt{K_{\text{eff}}},$$

where K_{eff} is the effective number of perturbed superpixels. This observation indicates that the ℓ_2-norm of the perturbation grows more slowly than the ℓ_1-norm as K_{eff} increases, and by attacking superpixels in descending order of importance $\phi(S_k)$, the cumulative attack strength $\Delta_m = \sum_{i=1}^{m} \phi(S_{(i)}) \cdot \|\delta_{(i)}\|$ increases rapidly, achieving misclassification after perturbing only $m^* \ll K$ superpixels. This reduces query complexity from $O(K)$ to $O(m^*)$ while maintaining $\|\delta\|_2 \leq \epsilon_{\max}$ through construction.

3.3 Superpixel-based Adversarial attack with Regional Confidence

Inspired by these observations, we propose a fully black-box, superpixel-guided attack that (i) requires no access to internal activations (unlike CAM-based approaches in [2]) and (ii) allocates the perturbation budget more effectively than prior superpixel attacks [3, 11].

Our pipeline (Fig.1) first segments the image with QuickShift [4] and initializes an occlusion noise vector over superpixels. We then compute occlusion-based confidence saliency maps by adapting RISE [15] and occlusion scoring to black-box classifiers to identify the most discriminative segments. Finally, we apply an L_1-based perturbation under an L_2-norm budget to the top-ranked segments, iteratively updating the noise toward the objective.

Superpixel Segmentation. We adopt QuickShift [4] to generate superpixels as modification templates, exploiting its efficient mode-seeking algorithm to produce coherent image regions. The medoid distance threshold τ controls the granularity of segmentation, balancing between under- and over-segmentation.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Classifier Confidence Saliency. We extend RISE [15] by using random binary masks $M \in \{0,1\}^K$ over K superpixels, and compute the expected confidence score when superpixel λ is preserved: $S_{I,f}(\lambda) = \mathbb{E}_M[f(I \odot M) \mid M(\lambda) = 1]$, where $\odot$ is the element-wise product that applies the mask to the image. Using Monte Carlo sampling with N masks $\{M_1, \ldots, M_N\}$, the importance map is estimated as:

$$S_{I,f}(\lambda) \approx \frac{1}{N\,\mathbb{E}[M]} \sum_{i=1}^{N} f(I \odot M_i)\, M_i(\lambda)$$

L-norm Budget. We adopt a two-phase, norm-constrained strategy. For untargeted attacks, the L_1 budget is allocated proportionally to superpixel importance, applying uniform perturbations $\delta_i = (B_1 \cdot w_k)/n_k$ within each superpixel S_k and processing segments in chunks until misclassification. If unsuccessful, the remaining L_2 budget is applied to the most salient segments (default 5%) with stronger perturbations $= \delta_i \cdot \delta_i$. For targeted attacks, a candidate direction pool is built from images of the target class, using differences with the original image and their negations; the pool is expanded when progress stalls and pruned to remove unused directions at the end of each iteration. Both approaches respect L_2 constraints, allocate perturbations efficiently, and avoid blind random search while remaining fully black-box.

4 Experiments

4.1 Experimental Setup

We evaluate SPARC on both untargeted and targeted black-box attacks using an Inception-V3 model [8] trained on ImageNet [10]. The evaluation uses 1,000 correctly classified validation images to ensure meaningful success rate measurement. First, we benchmark SPARC against superpixel-based attack baselines for the untargeted task: M-SAI-FGM [2], Superpixel [3], and Saliency [11]. Additionally, for pixel-wise and random-search black-box attack baselines, we benchmark against Parsimonious Attack [9], SimBA [5], EigenBA [13], Square Attack [6], and BABIES-DCT [7] for both untargeted and targeted scenarios.

The attack processes superpixels in chunks of size $\max(2, 0.05 \times \text{segments})$, with saliency map queries included in the total query count. SPARC operates in two phases: an initial L_1-constrained perturbation ($B_1 = 20$ for untargeted, $B_1 = 100$ for targeted), followed, if necessary, by an L_2-constrained refinement ($B_2 = 20$ for both), applied to the most discriminative superpixels. For targeted attacks, the search terminates upon exceeding budget constraints (L_1, L_2) or reaching query limits (20,000 queries, 100 iterations).

4.2 Comparison with Other Methods

Table 1 shows that SPARC attains a 99.6% untargeted success rate, 12.1 points higher than Saliency (87.5%) and 32.2 points higher than M-SAI-FGM (67.4%), demonstrating its leading performance among superpixel-based attacks.

Table 2 compares SPARC with state-of-the-art random-search black-box attacks; superpixel-based baselines are omitted as their success rates are substantially lower than pixel-wise methods, while included baselines cover ParsiBA [9],

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 1: Untargeted attack success rates for superpixel-based methods. Bold is best, underline is second-best.

	M-SAI-FGM [2]	Superpixel [3]	Saliency [11]	SPARC (Ours)
SR	0.674	0.872	<u>0.875</u>	**0.996**

SimBA [5], EigenBA [13], Square-attack[6], and BABIES-DCT[7].

Table 2: Attack comparison on ImageNet for untargeted and targeted scenarios. Bold is best, underline is second-best.

	Untargeted			Targeted		
	Avg QY	Avg L2	SR	Avg QY	Avg L2	SR
ParsiBA [9]	997	3.957	0.965	5075	8.422	0.634
SimBA [5]	1433	3.958	<u>0.990</u>	5762	8.424	0.744
EigenBA [13]	<u>383</u>	<u>3.622</u>	0.986	<u>2730</u>	7.926	<u>0.806</u>
Square-attack [6]	1100	5.0	0.929	16746	<u>3.0</u>	0.780
BABIES-DCT [7]	1907	5.0	0.879	9429	12.0	**0.992**
SPARC (Ours)	**72**	**2.622**	**0.996**	**1614**	**0.482**	0.657

SPARC achieves the lowest average queries in both untargeted (72) and targeted (1,614) attacks, outperforming baselines like EigenBA (383 / 2,730) and Square Attack (1,100 / 16,746), and reaches the highest untargeted success rate (0.996), slightly above SimBA (0.990). While its targeted success (0.657) is lower than BABIES-DCT (0.992), whose pixel-wise Fourier perturbations create visible artifacts, SPARC's superpixel-based optimization faces a harder loss landscape but still produces much smaller perturbations ($\ell_2 = 0.482$), enabling more controlled and visually subtle attacks. Overall, our method balances query efficiency, perturbation magnitude, and success more effectively.

Figure 2 shows SPARC adversarial examples with low, smooth L_2 perturbations along saliency-guided superpixels.

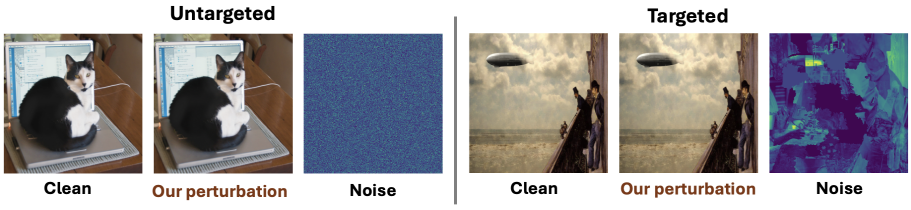

Fig. 2: SPARC adversarial examples: clean images, our perturbed outputs, and noise maps for untargeted (left) and targeted (right) attacks.

5 Conclusion

We introduced SPARC, a superpixel-based black-box attack with attention-guided dual-norm perturbations. It achieves strong untargeted and competitive targeted performance with low queries and small, interpretable noise. Future work will aim to boost targeted success and evaluate SPARC on medical imaging, where subtlety and trustworthy assessment are crucial.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[2] X. Dong, J. Han, D. Chen, J. Liu, H. Bian, Z. Ma, H. Li, X. Wang, W. Zhang, and N. Yu, Robust Superpixel-Guided Attentional Adversarial Attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12892-12901, 2020.

[3] I. Oe, K. Yamamura, H. Ishikura, R. Hamahira, and K. Fujisawa, Superpixel Attack: Enhancing Black-Box Adversarial Attack with Image-Driven Division Areas. In *AI 2023: Advances in Artificial Intelligence, 36th Australasian Joint Conference on Artificial Intelligence (AI 2023), Part I*, pages 141-152, Springer-Verlag, 2023.

[4] A. Vedaldi and S. Soatto, Quick Shift and Kernel Methods for Mode Seeking. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision – ECCV 2008*, pages 705-718, Springer, 2008.

[5] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, Simple black-box adversarial attacks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2484-2493, 2019.

[6] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*, pages 484-501, 2020.

[7] H. Tran, D. Lu, and G. Zhang, Exploiting the local parabolic landscapes of adversarial losses to accelerate black-box adversarial attack. In *European Conference on Computer Vision (ECCV)*, pages 317-334, 2022.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818-2826, 2016.

[9] S. Moon, G. An, and H. O. Song, Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4636-4645, 2019.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248-255, 2009.

[11] Z. Dai, S. Liu, K. Tang, and Q. Li, Saliency Attack: Towards Imperceptible Black-box Adversarial Attack. arXiv preprint, 2022.

[12] B. Zhou, A. Khosla, Á. Lapedriza, A. Oliva, and A. Torralba, Learning Deep Features for Discriminative Localization. *CoRR*, abs/1512.04150, 2015.

[13] L. Zhou, P. Cui, Y. Jiang, and S. Yang, Adversarial Eigen Attack on Black-Box Models. *CoRR*, abs/2009.00097, 2020.

[14] H. Hu, Z. Salcic, G. Dobbie, and X. Zhang, Membership Inference Attacks on Machine Learning: A Survey. *CoRR*, abs/2103.07853, 2021.

[15] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," *arXiv preprint arXiv:1806.07421*, 2018.