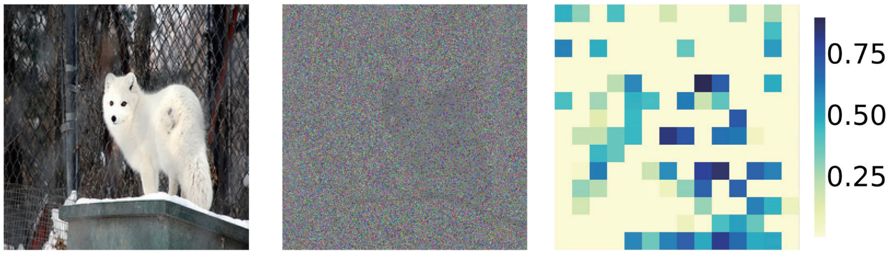
A Appendix

A.1 Visualization of patch-wise noise sensitivity



Original Image

Initial Noise $\ell_2 = 57.02$

Patch-wise Noise Sensitivity

Figure 1: Illustrations of original images (left), initial random noises (middle), and corresponding visualizations of patch-wise noise sensitivity (right).

Existing decision-based attack methods use random noises to initialize adversarial examples x^{init} [1, 2, 3, 4]. For example, a common practice is to add Gaussian noise with mean of 0 and a gradually increasing variance on the original image until the target model is misclassified:

$$x^{init} = Clip_{x,\tau}\{x + \xi^{Gau}\}, \quad \xi^{Gau} \sim \mathcal{N}(0, var^2 I), \tag{1}$$

where ξ^{Gau} refers to the random noise with the same dimension as the original image x and follows the Gaussian distribution with mean of 0 and variance of var. I is an identity matrix of the same dimension as x. The decision-based attack can only obtain the hard-label returned by the target model, and the attacker does not have any prior knowledge about the target model. Therefore, the noise z^{init} on the initial adversarial example is generally uniform at each pixel, as shown in the middle column of Fig. 1. After adding random noises to the original image until misclassification, the decision-based attacks use the initial adversarial example as the starting point of the noise compression process.

Fig. 2 compares the differences of patch-wise noise sensitivity between res-101 and r26-32. It can be seen that only removing the noises on a few patches on the res-101 will affect the misclassification, while the patch-wise noise sensitivity on r26-32 varies greatly. This reflects the reason why it is difficult to attack ViTs using existing decision-based attacks.

A.2 Proof of Proposition 1

Proposition 1. *Assume x' is an initial adversarial example generated by Boundary Attack against ViT F starting from original image x, $F(x) \neq F(x')$. For any $0 < r_1, r_2, h \leq Height$, $0 < c_1, c_2, w \leq Width$, if $Sens(F, x, x', r_1, c_1, h, w) < Sens(F, x, x', r_2, c_2, h, w)$, and the new noise added by one step by Boundary Attack is z', then $P(F(x' + z'_1) \neq F(x)|F(x' + z') = F(x)) < P(F(x' + z'_2) \neq F(x)|F(x' + z') = F(x))$, where for $\iota = 1, 2$*

$$z'_{\iota,r,c} = \begin{cases} 0, & if \quad r_\iota \leq r < r_\iota + h \quad and \quad c_\iota \leq c < c_\iota + w, \\ z'_{r,c}, & else, \end{cases} \tag{2}$$

Proof. According to the attack process of Boundary Attack:

$$x^*_{new} = x^* + \delta \cdot \frac{\eta}{\|\eta\|_2} + \varepsilon \cdot \frac{x - x^*}{\|x - x^*\|_2}, \quad \eta \sim \mathcal{N}(0, I), \tag{3}$$

New noise $z' \sim \mathcal{N}(\varepsilon \cdot \frac{x-x'}{\|x-x'\|_2}, \delta^2)$. Noise compression ratio after one-step Boundary Attack satisfies $\frac{z'}{x'-x} \sim \mathcal{N}(\frac{\varepsilon}{\|x-x'\|_2}, \frac{\delta^2}{(x-x')^2})$. Since the initial noise x' generated by Boundary Attack follows Gaussian distribution with mean of 0 and equal variance on each pixel, the expectation of the initial noise is equal. Therefore, the noise compression ratio after one-step Boundary Attack for each

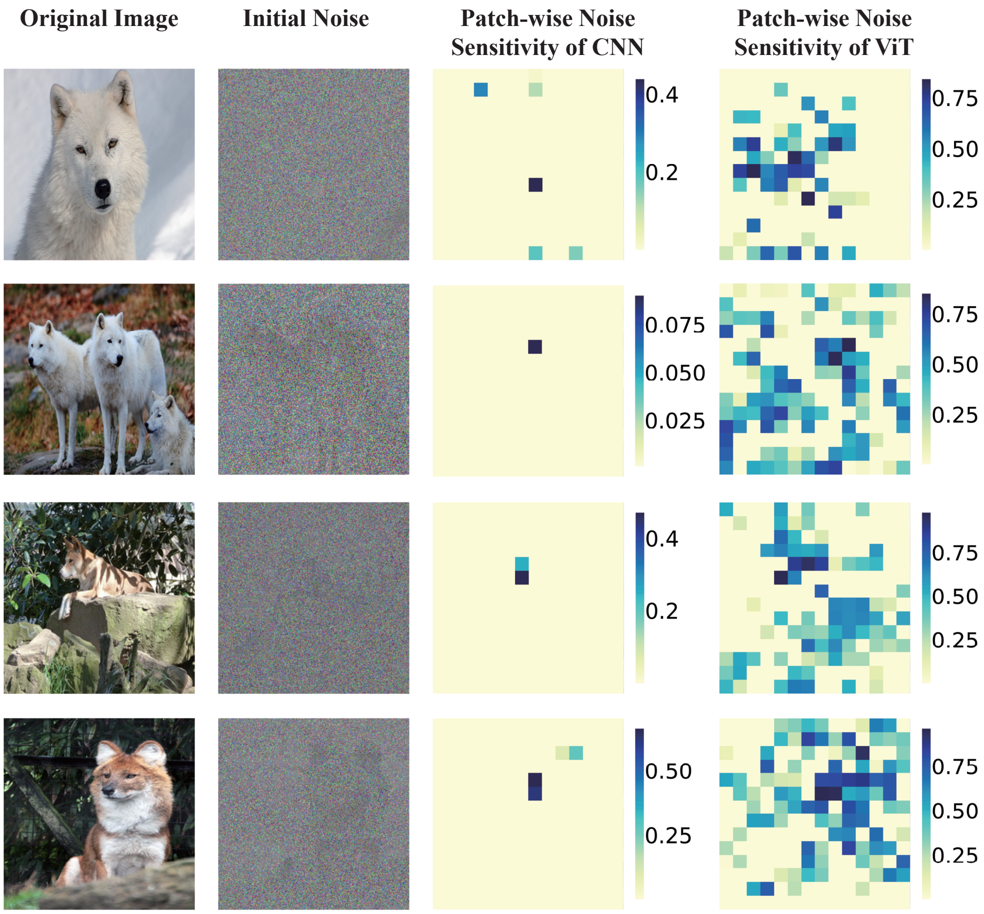


Figure 2: Comparison of patch-wise noise sensitivity between res-101 and r26-32 on ILSVRC-2012.

pixel is i.i.d. The possibility that the noise compression ratio on at least one pixel exceeds κ is the same for any pixel:

$$P(\exists \quad 0 < r^* \leq Height \quad and \quad 0 < c^* \leq Width, \frac{z'_{r^*,c^*}}{x'-x} > \kappa), \tag{4}$$

where $0 < r^* \leq Height, 0 < c^* \leq Width, 0 < \kappa \leq 1$. For any $0 < \kappa_1 \leq \kappa_2 \leq 1$:

$$P(\frac{z'_{r^*,c^*}}{x'-x} > \kappa_1) - P(\frac{z'_{r^*,c^*}}{x'-x} > \kappa_2) = P(\kappa_2 \geq \frac{z'_{r^*,c^*}}{x'-x} \geq \kappa_1) \geq 0, \tag{5}$$

$$P(\frac{z'_{r^*,c^*}}{x'-x} < \kappa_2) - P(\frac{z'_{r^*,c^*}}{x'-x} < \kappa_1) = P(\kappa_2 \geq \frac{z'_{r^*,c^*}}{x'-x} \geq \kappa_1) \geq 0, \tag{6}$$

The equality holds when $\kappa_1 = \kappa_2$. Since the probability that noise compression ratio on at least one pixel exceeds the noise sensitivity $Sens$ increases monotonically with respect to the noise sensitivity on the whole patch, and $Sens(F, x, x', r_1, c_1, h, w) < Sens(F, x, x', r_2, c_2, h, w)$, we have:

$$\begin{aligned} &P(F(x' + z'_2) \neq F(x) | F(x' + z') = F(x)) \\ =&P(\exists r_2 \leq r_2^* \leq r_2 + h \; and \; c_2 \leq c_2^* \leq c_2 + w, \frac{z'_{r_2^*,c_2^*}}{x'-x} < Sens(F, x, x', r_2, c_2, h, w)) \\ >&P(\exists r_1 \leq r_1^* \leq r_1 + h \; and \; c_1 \leq c_1^* \leq c_1 + w, \frac{z'_{r_1^*,c_1^*}}{x'-x} < Sens(F, x, x', r_1, c_1, h, w)) \\ =&P(F(x' + z'_1) \neq F(x) | F(x' + z') = F(x)). \end{aligned} \tag{7}$$

Table 1: Median and average ℓ_2 distance of adversarial perturbations on ILSVRC-2012 against 4 ViTs.

Target	ti_l16		r_ti_16		vit_s32		vit_b16	
Methods	Mid	Avg	Mid	Avg	Mid	Avg	Mid	Avg
Initial	122.666	121.669	49.142	47.79	79.332	74.452	104.872	95.847
PAR	25.372	58.037	5.353	6.5	11.82	16.149	17.518	32.103
HSJA	79.806	91.875	28.195	30.339	57.971	51.718	76.448	73.613
PAR+HSJA	24.363	56.813	5.194	6.316	11.451	15.842	15.599	31.158
BBA	26.871	58.071	4.767	7.091	8.887	12.957	16.682	30.617
PAR+BBA	19.215	53.288	2.932	4.465	5.309	11.292	11.737	26.72
Evo	35.033	65.997	7.042	10.81	11.805	17.721	28.219	40.623
PAR+Evo	20.887	55.168	4.201	5.578	9.166	13.339	13.358	28.76
Boundary	39.43	66.223	9.116	12.512	18.191	20.409	26.333	38.064
PAR+Boundary	21.075	55.263	4.62	5.971	10.452	14.368	13.842	29.304
SurFree	30.971	61.017	5.69	9.325	11.024	15.758	17.341	33.533
PAR+SurFree	18.868	53.815	3.899	5.229	8.454	12.885	12.18	27.57
CAB	57.069	77.707	4.071	10.841	13.122	22.509	26.268	48.165
PAR+CAB	15.209	52.193	**2.627**	**4.419**	**5.156**	10.598	**8.171**	25.306
Sign-OPT	34.884	38.06	114.027	113.639	40.168	41.231	71.778	65.801
PAR+Sign-OPT	**5.264**	**6.793**	23.801	53.313	5.18	**6.135**	10.696	**15.447**

Table 2: Median and average ℓ_2 distance of adversarial perturbations on ILSVRC-2012 against ViTs.

Target	vit_b32		r50_l32		ti_s16	
Methods	Mid	Avg	Mid	Avg	Mid	Avg
Initial	97.8	89.433	70.962	79.394	41.607	42.921
PAR	15.897	26.216	13.083	26.662	5.449	7.772
HSJA	65.213	64.582	46.57	56.298	24.181	28.403
PAR+HSJA	15.376	25.845	11.106	25.49	4.897	7.538
BBA	11.835	24.534	14.954	24.47	4.182	5.99
PAR+BBA	10.196	22.026	9.775	22.162	2.787	4.772
Evo	17.234	30.62	19.952	28.534	6.616	8.872
PAR+Evo	12.179	23.134	10.159	22.639	4.39	6.182
Boundary	21.407	31.815	21.173	31.358	8.296	10.757
PAR+Boundary	13.786	24.294	10.506	24.255	4.818	6.705
SurFree	14.838	27.774	16.263	26.861	4.386	7.701
PAR+SurFree	11.684	22.92	9.381	22.719	3.701	5.76
CAB	19.376	38.092	19.226	33.201	4.559	10.665
PAR+CAB	**8.949**	**21.314**	**7.894**	**22.077**	**2.158**	**4.594**
Sign-OPT	95.78	88.212	88.657	81.727	34.884	38.06
PAR+Sign-OPT	16.477	31.713	15.212	25.67	5.264	6.793

Therefore, $P(F(x'+z'_1) \neq F(x) | F(x'+z') = F(x)) < P(F(x'+z'_2) \neq F(x) | F(x'+z') = F(x))$. $\square$

Although the sensitivity evaluation of PAR slightly resembles that of ℓ_0 sparse attacks [7, 8], there are huge differences which make the comparison hardly possible. Firstly, the goal of PAR is to compress noise from initial adversarial examples while the goal of ℓ_0 attacks is to minimize the number of perturbed pixels. Secondly, ℓ_0 attacks usually need some additional information, e.g., random adversarial images for sparse decomposition in LSDAT [7], while PAR only needs hard label of the target model.

Boundary Attack's ignorance of the difference in noise sensitivity between patches results in two serious consequences. First of all, since the initial noise z^{init} and compression noise are uniform for each pixel, the magnitude of noise on each pixel after multiple steps of compression is also close. When the noise in the most sensitive region of the image is compressed, it is difficult for the updated adversarial example to maintain misclassification, and the subsequent query is likely to fail. To some extent, this explains why the noise compression efficiency of Boundary Attack gradually decreases as the query number grows [1].

Except for Boundary Attack, most of the existing decision-based attack methods are essentially local random search starting from a random noise. For example, SurFree [9] focuses on the geometric properties in the neighborhood of current adversarial example x^*. HSJA [3] estimates the decision boundary near x^*. BBA [4] and CAB [10] samples in the entire image space based on x^* with adaptive

Table 3: Median and average ℓ_2 distance of adversarial perturbations between four models on ImageNet-21k.

Target	vit_s32		vit_b16		vit_b_32		r50_s32	
Methods	median	average	median	average	median	average	median	average
Initial	42.939	47.117	28.839	44.511	34.515	44.885	56.912	41.267
PAR	4.968	7.814	5.637	10.397	5.614	9.699	3.191	9.306
HSJA	24.728	27.328	16.244	27.895	20.486	29.87	38.993	29.514
PAR+HSJA	4.573	7.487	4.476	9.684	5.185	9.159	2.218	7.788
BBA	4.008	6.063	4.012	10.119	4.086	8.264	8.666	17.211
PAR+BBA	2.162	4.482	3.202	8.071	2.877	6.431	2.218	7.322
Evo	5.311	7.617	5.562	12.347	6.107	11.24	14.355	13.757
PAR+Evo	3.361	5.728	3.965	8.777	4.335	8.134	2.218	8.006
Boundary	8.012	9.768	7.519	13.406	7.822	12.011	12.587	15.687
PAR+Boundary	4.265	6.324	4.42	9.176	4.737	8.372	2.218	8.55
SurFree	4.996	6.319	3.343	9.349	4.725	8.64	6.83	13.943
PAR+SurFree	2.951	5.387	3.412	8.187	3.608	7.291	**2.218**	8.067
CAB	4.749	8.815	2.4	9.127	4.749	11.391	8.275	13.307
PAR+CAB	**1.696**	**4.24**	**1.72**	**6.235**	**2.225**	**6.007**	**2.218**	**5.437**
Sign-OPT	27.239	36.776	23.278	38.681	24.656	37.362	47.589	36.398
PAR+Sign-OPT	4.335	7.057	5.251	10.238	4.728	8.353	2.684	8.793

Table 4: Noise compression comparison when minimum patch size $PS_{min} = 1$.

	Initial Patch Size	112	56	28	14	7
	Minimum Patch Size	1	1	1	1	1
vgg-19	Mid Noise	4.73	4.95	5.20	5.98	13.05
	Avg Noise	6.32	6.31	6.55	7.05	11.31
	Avg Query Number	810.22	811.86	835.30	882.28	945.43
vit_s16	Mid Noise	8.89	8.97	9.38	11.88	24.93
	Avg Noise	17.68	17.53	17.49	18.90	26.84
	Avg Query Number	825.60	831.32	855.66	909.22	969.57

distribution. Existing decision-based attack methods mainly focus on searching for adversarial examples with smaller noise magnitude in the neighborhood of current adversarial example, but ignore the noise in x^{init} with larger magnitude and easier to compress due to the difference in noise sensitivity.

A.3 More Experimental Results

To further verify the advantage of PAR over existing decision-based attacks on different ViTs and CNNs, we report the median and average adversarial perturbation of more target models on ILSVRC-2012 and ImageNet-21k in Table 1, Table 2, and Table 3. The first row of three tables represents target models with different structures. We compare the average (Avg) and median (Mid) noise magnitude generated by PAR and other 6 attacks on different target models. We also use PAR as the noise initialization for other decision-based attacks. The noise compressed by PAR is handed over to other decision-based attacks for further compression. It can be seen that when PAR is used to initialize adversarial noise, the average and median noise magnitude drops significantly compared with only using the original decision-based attack. This verifies the strong noise compression ability of PAR.

In Table 4 we add experimental results with a minimum patch size of 1. A minimum patch size of 1 means that PAR will try to remove noise on a single pixel. It can be seen from the results that using a too small minimum patch size will also lead to low compression efficiency. Because when $PS_{min} = 1$, a single query can only remove noise on a single pixel at most even if it succeeds. At the same time, the number of queries consumed by PAR will also increase sharply with a too small minimum patch size.

In Table 5, we compare the time consumption and noise compression efficiency of PAR and other decision-making attacks on the Imagenet. The target model is r-ti-16. The total number of queries is 1000 times. Among them, the first 50 times are used for generating Gaussian noise to find initial

Table 5: Comparison of time cost and compression efficiency.

Methods	Time Cost (s)	Used step	Time Per Query (s)	Noise Compression Per Query
PAR	2.22	60	0.037	**0.673**
Evo	28.28	950	0.030	0.035
PAR+Evo	27.22	950	**0.029**	0.045
Boundary	31.37	950	0.033	0.040
PAR+Boundary	34.72	950	0.037	0.044
CAB	36.09	950	0.038	0.044
PAR+CAB	70.15	950	0.074	0.047

adversarial examples. When PAR is not applied, the next 950 times are all used for decision-based attacks. When initialized with PAR, 60 queries are used for PAR, and then the remaining 890 queries are used for decision-based attack. The experimental results report the total time consumption, number of queries, query time per query and average compression noise per query. Since the main time-consuming of the query lies in the forward propagation process of the target model, the used time of a single query for each method is similar. But it can be seen that the noise compression efficiency of each decision attack method is improved after initializing with PAR. During the first 60 queries of PAR, the noise compression efficiency is significantly higher than other decision-based attacks, which demonstrates the effectiveness of PAR.

References

[1] J. R. Wieland Brendel and M. Bethge, “Decision-based adversarial attacks: Reliable attacks against black-box machine learning models,” in *ICLR*, 2018.

[2] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, “Efficient decision-based black-box adversarial attacks on face recognition,” in *CVPR*, 2019.

[3] J. Chen, M. I. Jordan, and M. J. Wainwright, “Hopskipjumpattack: A query-efficient decision-based attack,” in *2020 ieee symposium on security and privacy (sp)*. IEEE, 2020, pp. 1277–1294.

[4] T. Brunner, F. Diehl, M. T. Le, and A. Knoll, “Guessing smart: Biased sampling for efficient black-box adversarial attacks,” in *ICCV*, 2019, pp. 4958–4966.

[5] R. Wightman, “Pytorch image models,” https://github.com/rwightman/pytorch-image-models, 2019.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[7] A. Esmaeili and M. Edraki, “Lsdat: Low-rank and sparse decomposition for decision-based adversarial attack,” *arXiv*, 2021.

[8] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE TEC*, 2019.

[9] T. Maho, T. Furon, and E. Le Merrer, “Surfree: a fast surrogate-free black-box attack,” in *CVPR*, 2021, pp. 10 430–10 439.

[10] Y. Shi, Y. Han, and Q. Tian, “Polishing decision-based adversarial noise with a customized sampling,” in *CVPR*, 2020, pp. 1030–1038.