Decision-based Black-box Attack Against Vision Transformers via Patch-wise Adversarial Removal

Yucheng Shi[1] Yahong Han[1] Yu-an Tan[2] Xiaohui Kuang[3]

[1]College of Intelligence and Computing, and Tianjin Key Lab of Machine Learning, Tianjin University
[2]School of Cyberspace Science and Technology, Beijing Institute of Technology
[3]National Key Laboratory of Science and Technology on Information System Security

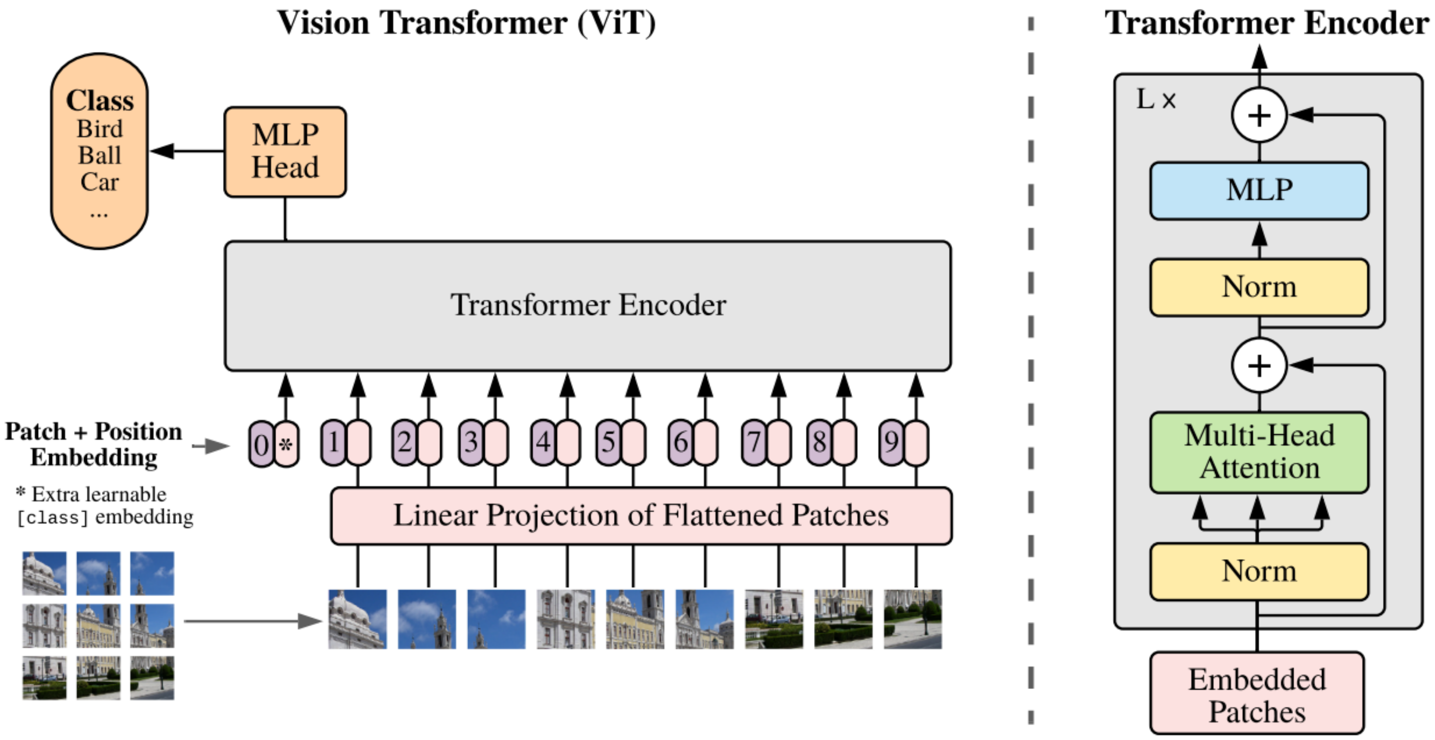
Code: [***https://github.com/shiyuchengTJU/PAR***](https://github.com/shiyuchengTJU/PAR)

Outline

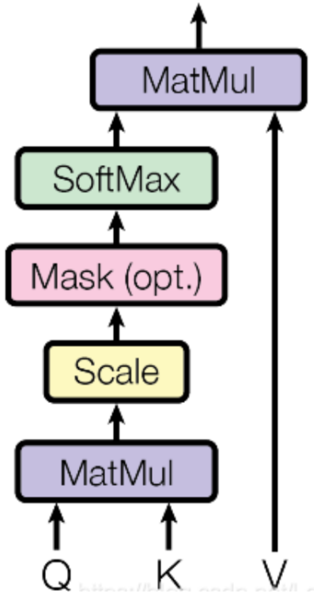
- Robustness of Vision Transformers
- Decision-based Attack Against ViTs
- Experiments

Robustness of Vision Transformers

Vision Transformers



$$Q = XW^Q$$
$$K = XW^K$$
$$V = XW^V$$



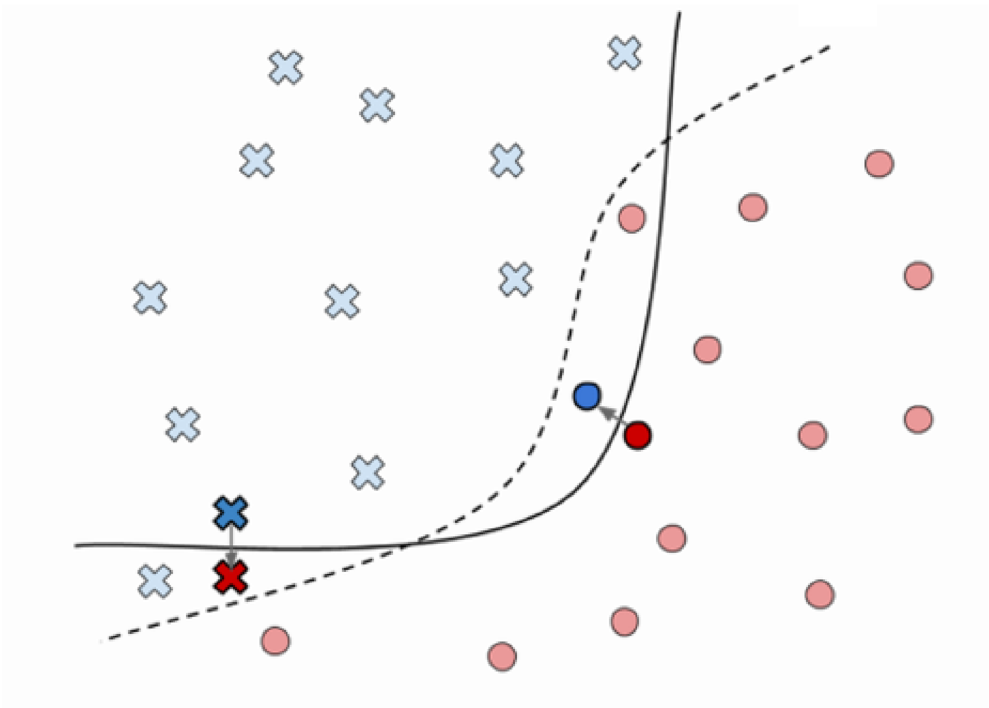
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR, 2020.

Robustness of Vision Transformers

Adversarial Example



$$\text{Find } x'$$

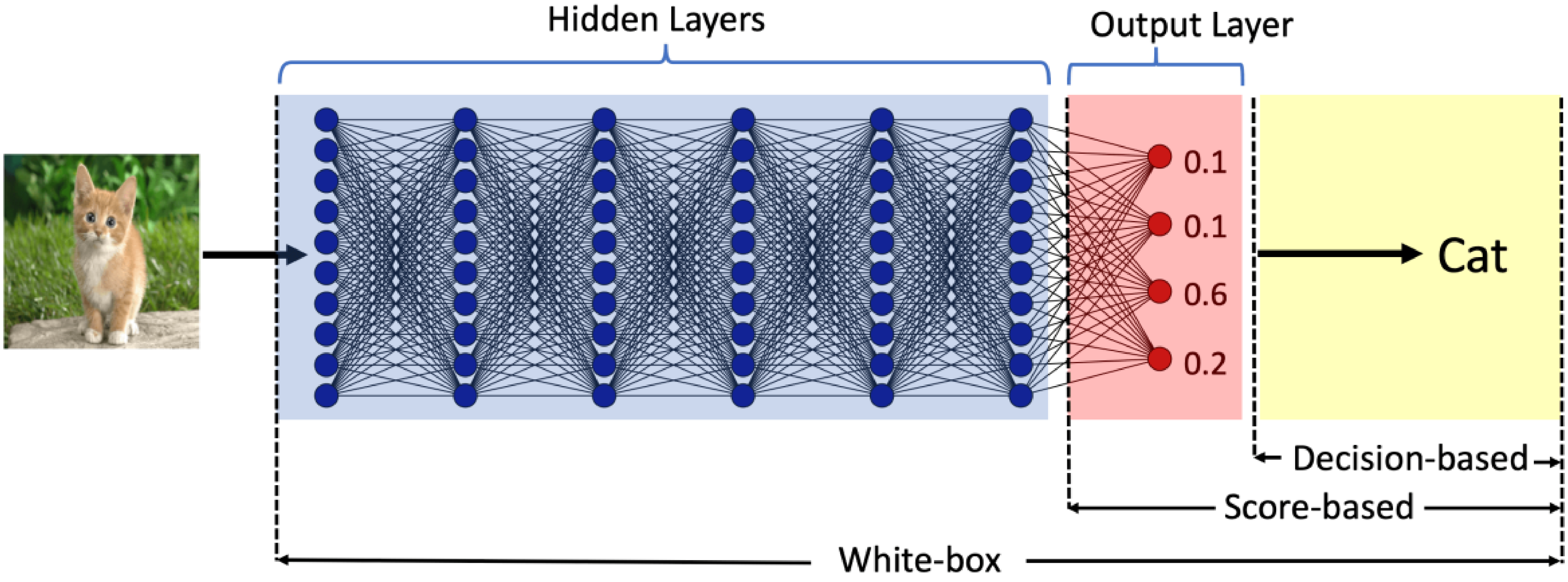
$$\text{s.t.} \quad f_1(x) \neq f_1(x')$$

$$\Delta(x, x') < \varepsilon$$

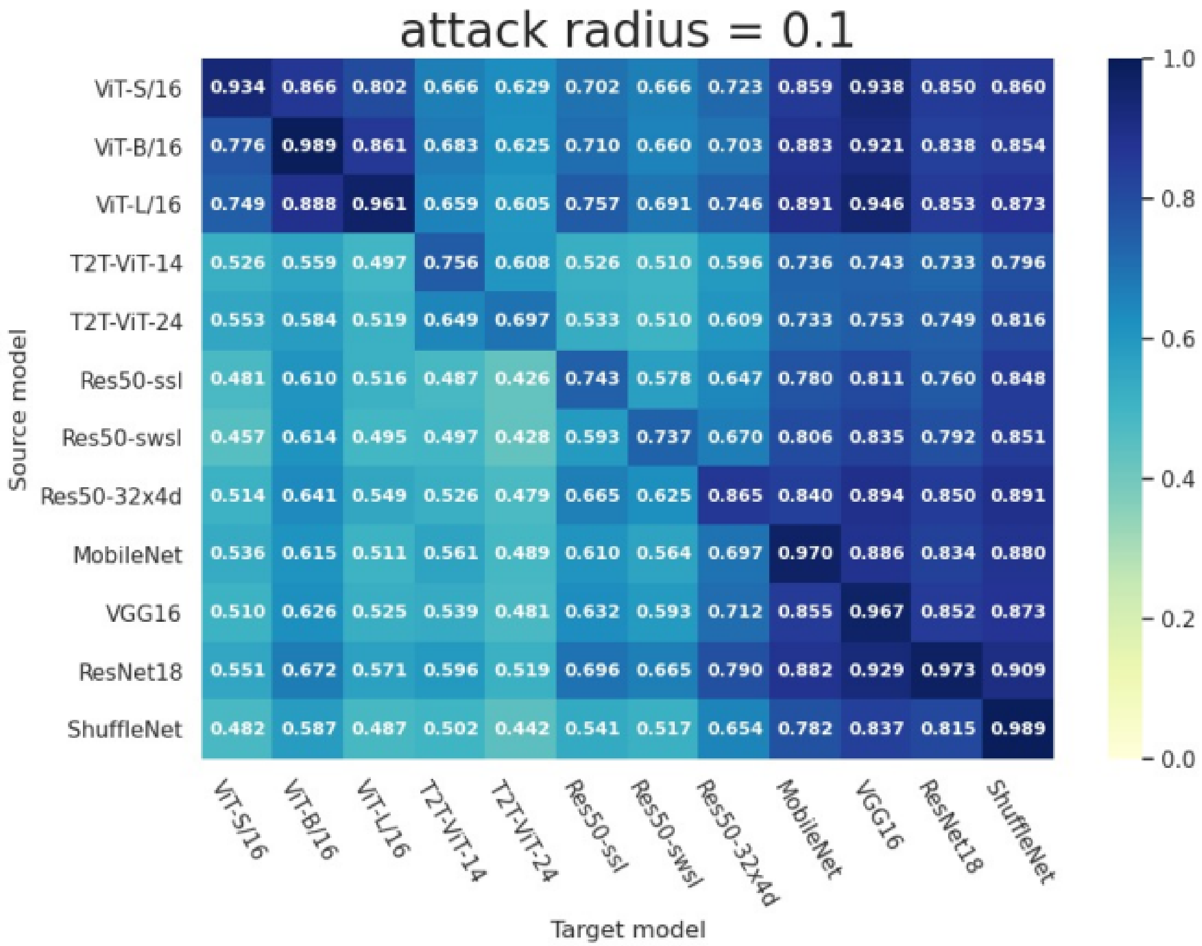
$$f_2(x) = f_2(x')$$

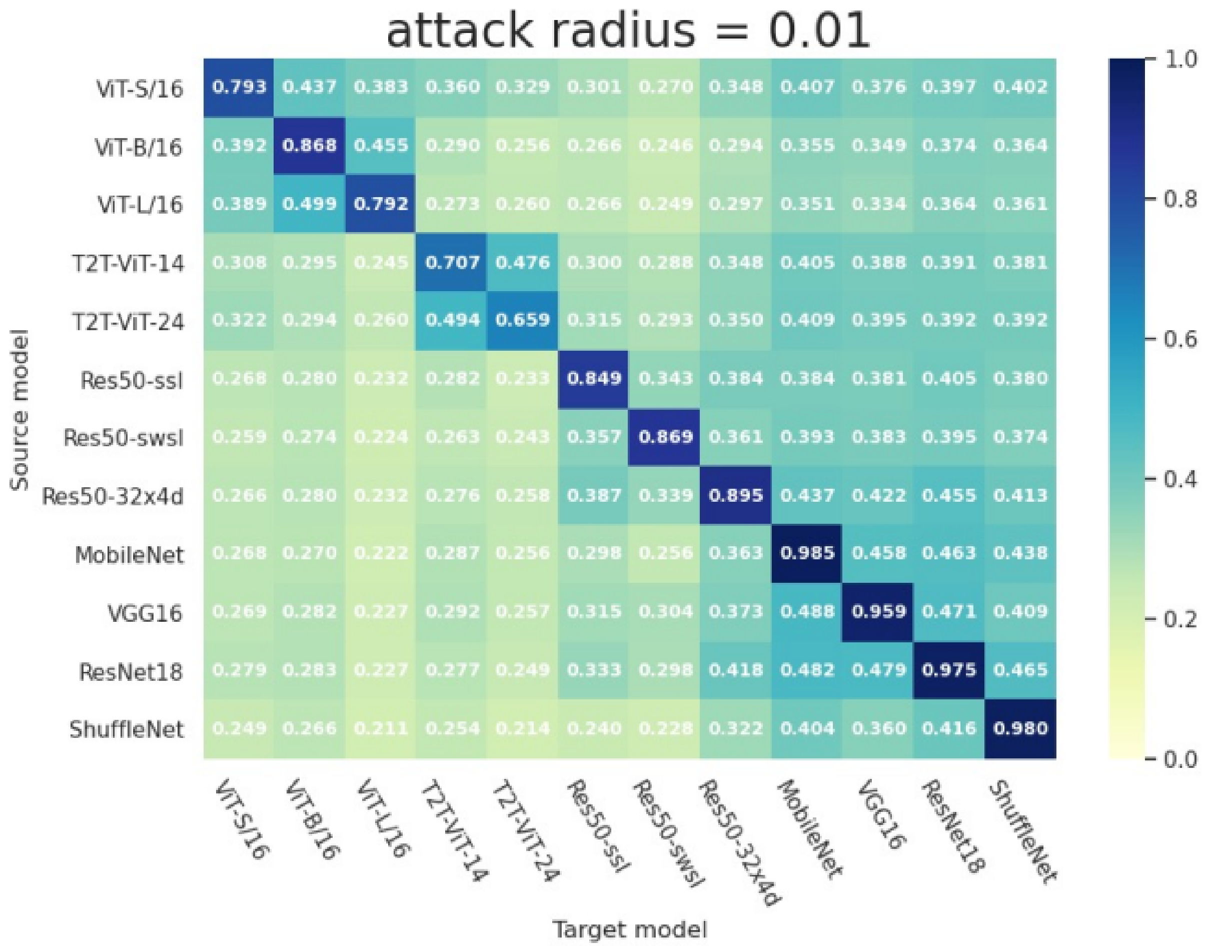
Robustness of Vision Transformers

- Black-box Attacks



Robustness of Vision Transformers



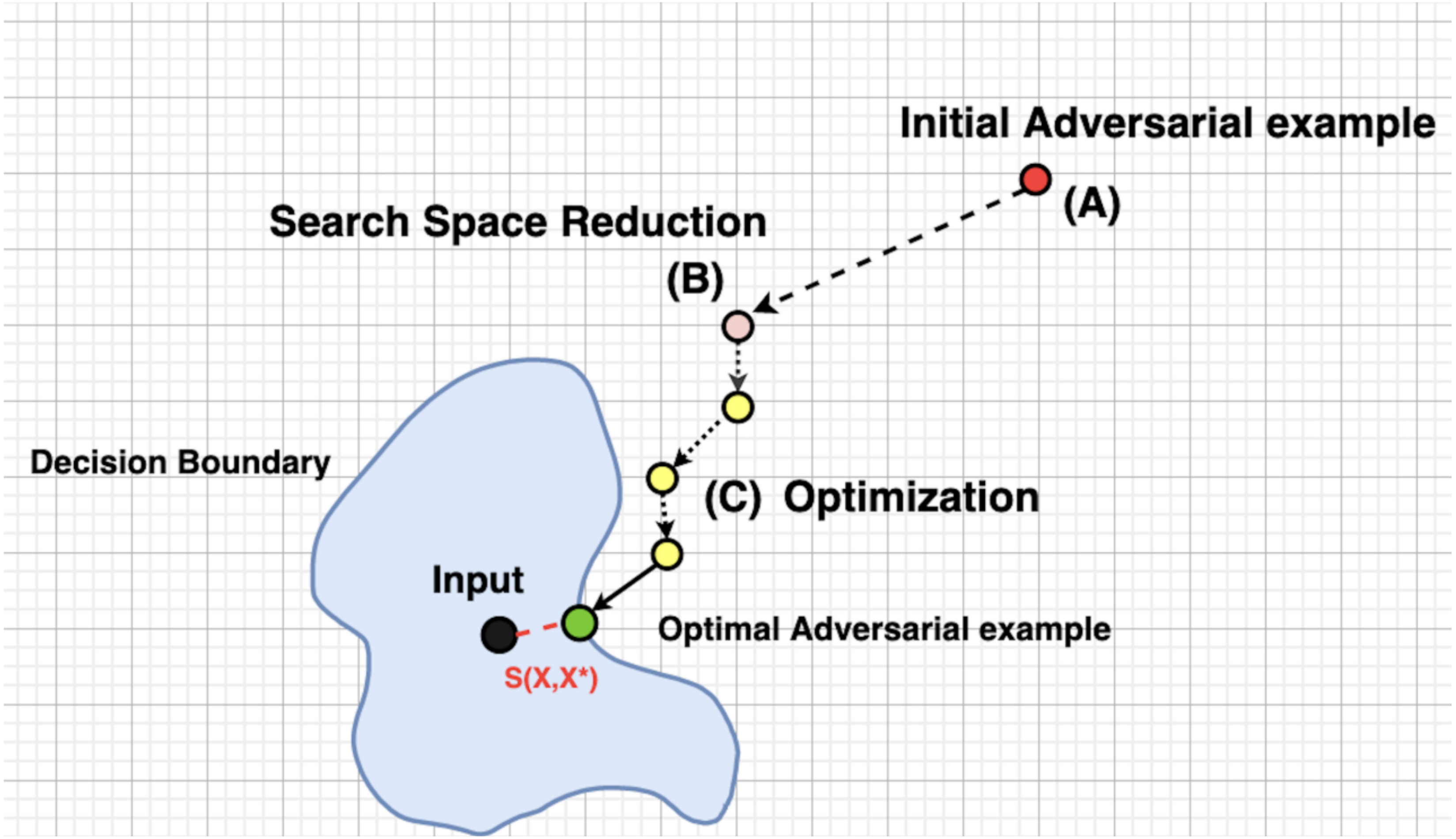


- Structural differences between ViT and CNN models lead to poor cross structure transferability
- Without prior knowledge of the target model structure, the transfer attack is prone to fail

- Shao R, Shi Z, and Yi J. On the adversarial robustness of vision transformers. arXiv preprint arXiv:2103.15670, 2021.

Robustness of Vision Transformers

- Decision-based adversarial attack



$$\min_{x' \in S_Q} \|x' - x\|_v, \quad s.t. \ \arg\max F(x') \neq y \ and \ |S_Q| \leq T$$

Robustness of Vision Transformers

Definition 1. *Let x' be an adversarial example of ViT model F on the original image x, i.e., $F(x') \neq F(x)$, and z be the current adversarial noise $z = x' - x$. Let $\tilde{z}$ be a new adversarial noise compressed from z in a rectangle patch with width of w, height of h, top left corner of sr, sc:*

$$\tilde{z}(sr, sc, h, w, \kappa)_{r,c} = \begin{cases} z_{r,c} \cdot \kappa, & if\ sr \leq r < sr + h\ and\ sc \leq c < sc + w, \\ z_{r,c}, & else, \end{cases}$$

where r and c refer to the row and column index of one pixel in noise z, respectively. $\kappa \in [0, 1]$ denotes the noise compression ratio. Define the noise sensitivity of a rectangle patch as the minimum noise compression ratio κ_{min} when F misclassifies $x + \tilde{z}$:

$$\begin{aligned} Sens(F, x, x', sr, sc, h, w) = \kappa_{min}, \quad s.t. \quad & F(x + \tilde{z}(sr, sc, h, w, \kappa_{min})) \neq F(x) \\ and\ \forall\ \kappa' < \kappa_{min}, \quad & F(x + \tilde{z}(sr, sc, h, w, \kappa')) = F(x). \end{aligned}$$

Sens: quantify the noise sensitivity of models between regions of an image.
Smaller ***Sens***: more noise can be removed without affecting misclassification.

Robustness of Vision Transformers

- CNN

Target	res-101		dense		vgg-19		senet	
Methods	Mid	Avg	Mid	Avg	Mid	Avg	Mid	Avg
Initial	58.60	54.71	54.38	52.77	34.80	34.67	49.52	53.96

- ViT

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

Target	vit_b32		r50_l32		ti_s16		r26_s32		vit_s16	
Methods	Mid	Avg	Mid	Avg	Mid	Avg	Mid	Avg	Mid	Avg
Initial	97.8	89.433	70.962	79.394	41.607	42.921	94.72	88.49	96.25	92.94

- Experimental results on ILSVRC-2012 , the overall high noise sensitivity of the ViT model results in a much larger initial adversarial noise required to achieve misclassification than CNN

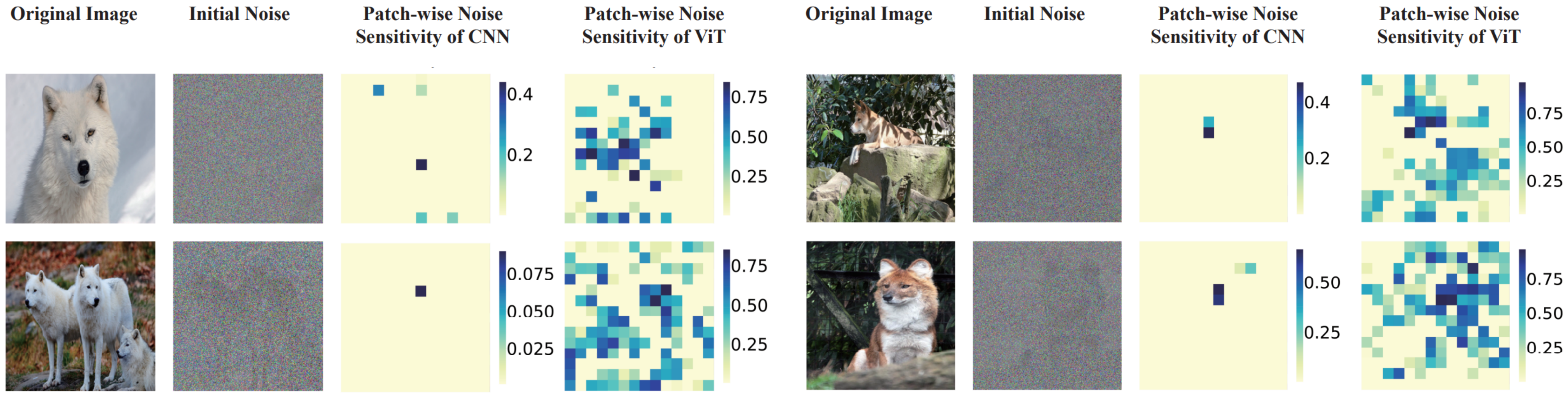
Robustness of Vision Transformers

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 $\iota = 1, 2$*

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

- Under decision-based attack, removing noise in regions with high ***Sens*** is more likely to be the cause of decision attack compression failure
- **Failures in noise compression are more likely to be caused by highly sensitive regions of the image.**

ViT 和 CNN的对抗鲁棒性对比

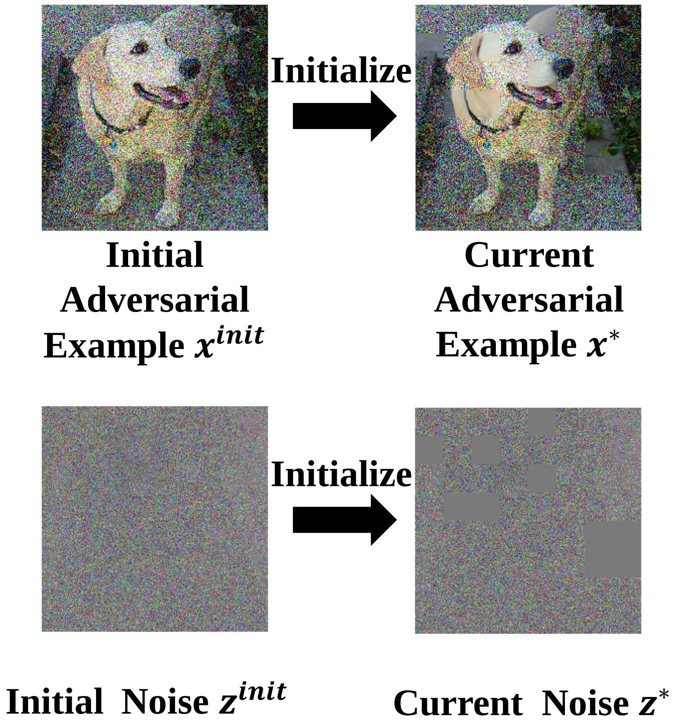


- CNNs: most regions are not sensitive and easy to compress
- ViTs: sensitivity of different regions varies greatly, therefore very difficult to compress the noise on the entire image as a whole.

Outline

- Robustness of Vision Transformers
- Decision-based Attack Against ViTs
- Experiments

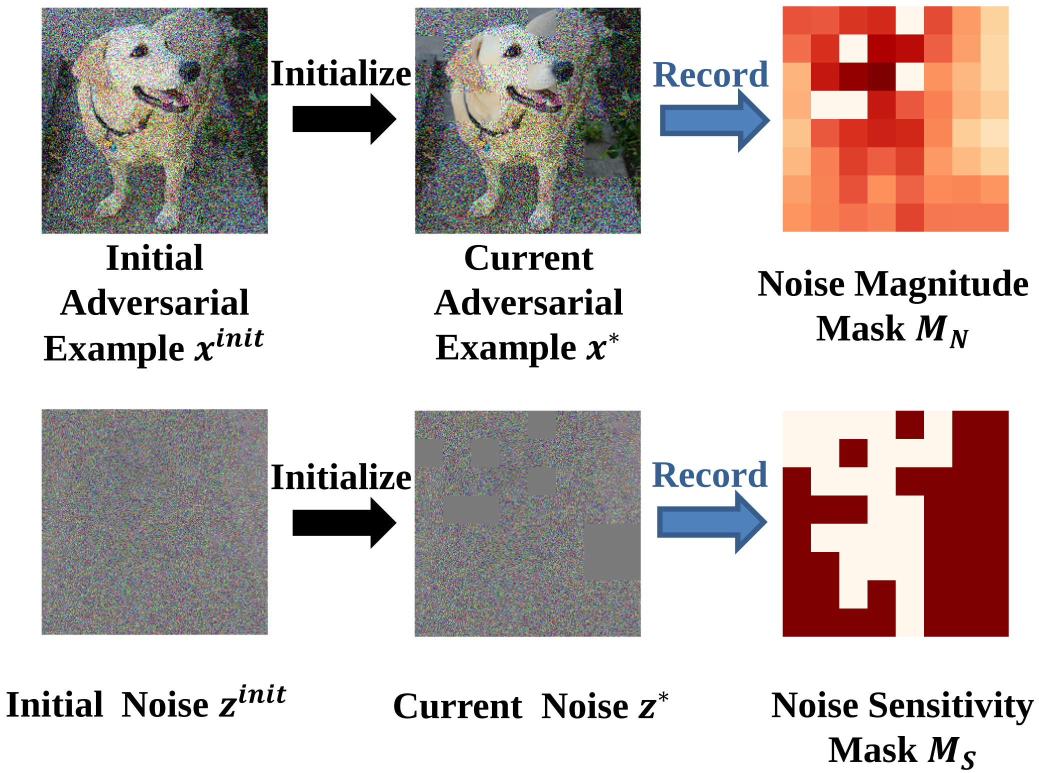
Patch-wise Adversarial Removal



- Adversarial noise initialization

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

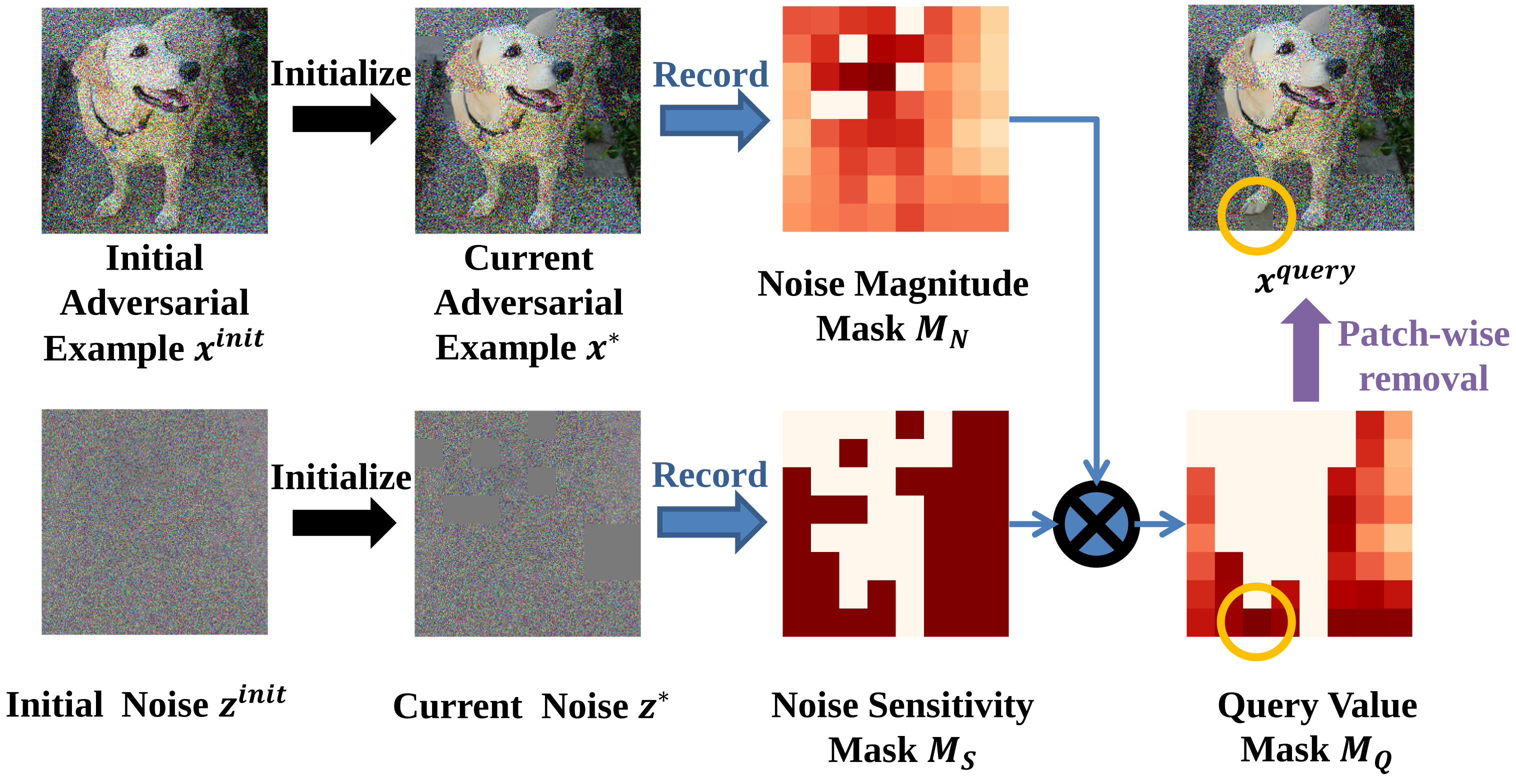
Patch-wise Adversarial Removal



- Noise sensitivity mask M_S: whether the noise is misclassified after removing the noise
- Noise magnitude mask M_N: records the noise amplitudes of different patches

$$M_N(row, col) = \sqrt{\sum_{i=row*PS_0+1}^{(row+1)*PS_0} \sum_{j=col*PS_0+1}^{(row+1)*PS_0} (x_{i,j}^{init} - x_{i,j})^2} \qquad M_S = J_{row,col}$$

Patch-wise Adversarial Removal

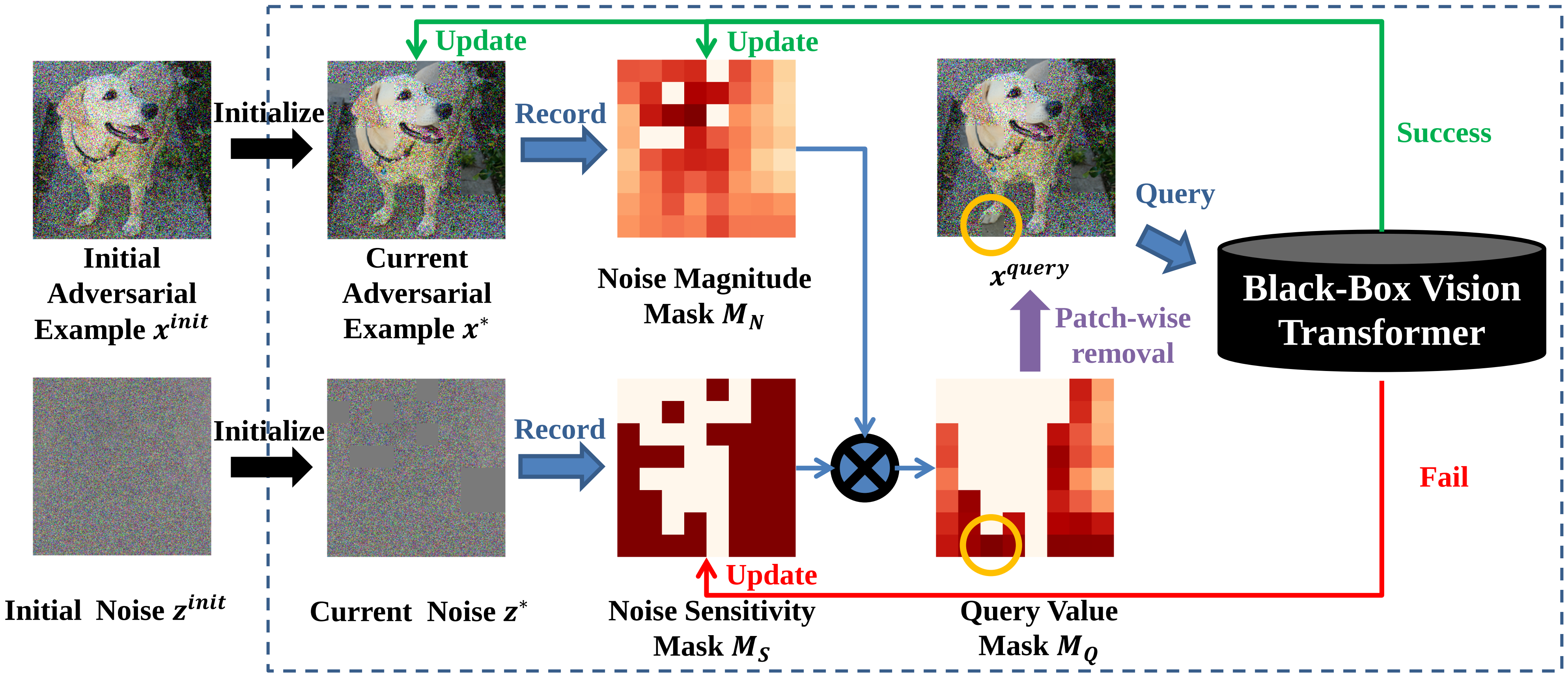


- Calculate the value of eliminating noise for a single patch
- Select the patch with low noise sensitivity and large noise magnitude

$$M_Q = M_N \odot M_S$$

$$row^*, col^* \leftarrow argmax(M_Q)$$
$$z^{query} \leftarrow x^* - x$$

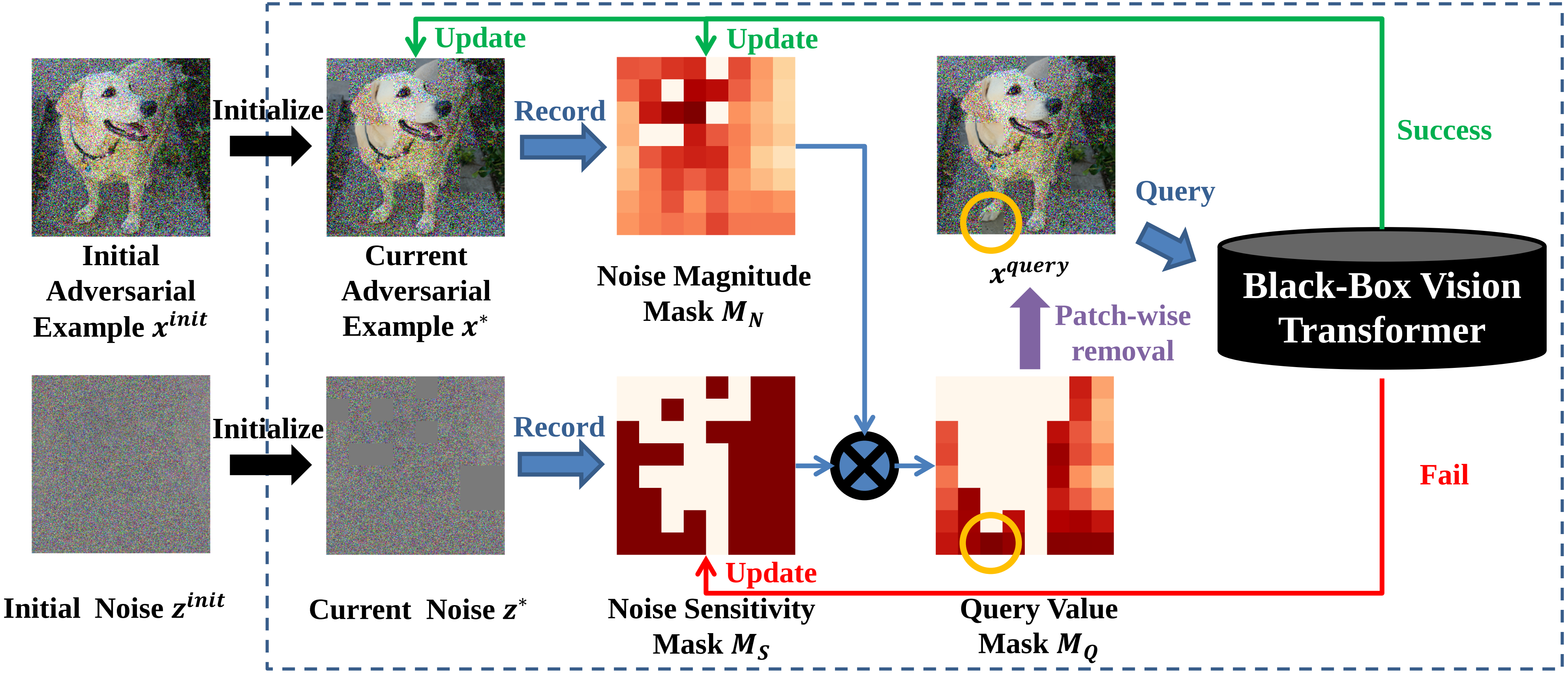
Patch-wise Adversarial Removal



- Query success: update the current adversarial example
- Query fail: update the noise sensitivity mask

$$z^{query}_{row^* * PS+1:(row^*+1)*PS,col^* * PS+1:(col^*+1)*PS} \leftarrow 0$$

Patch-wise Adversarial Removal



- Firstly eliminate noise in non-sensitive areas, and gradually optimizing sensitive areas
- Can be combined with other decision-based attack methods as an **efficient noise initialization means**

Outline

- Robustness of Vision Transformers
- Decision-based Attack Against ViTs
- Experiments

Experiments

- Results on ILSVRC-2012

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**

PAR: smaller noise magnitude than most decision-based attacks without using all the queries

Experiments

Results on ImageNet-21k

Target	r26_s32		ti_s16		vit_s16		ti_l16		r_ti_16	
Methods	median	average	median	average	median	average	median	average	median	average
Initial	41.161	43.24	21.376	26.847	40.52	45.828	23.591	43.866	8.075	14.297
PAR	5.706	9.189	2.771	3.992	4.326	7.516	5.016	10.18	1.554	2.592
HSJA	20.356	25.011	8.06	13.444	16.369	25.268	14.434	25.535	4.367	8.373
PAR+HSJA	4.752	7.781	2.388	3.719	3.644	6.688	4.517	9.093	1.522	2.51
BBA	5.849	9.069	1.643	3.125	3.692	6.422	5.423	10.875	1.263	2.315
PAR+BBA	3.899	6.953	0.982	2.21	2.098	4.547	3.456	7.816	0.921	1.759
Evo	8.195	12.047	4.133	6.253	5.223	9.82	7.847	15.358	3.093	3.924
PAR+Evo	4.091	7.122	2.055	3.284	2.427	5.236	4.041	8.576	1.487	2.223
Boundary	11.25	14.102	4.8	6.068	7.963	11.533	8.047	13.583	2.442	3.876
PAR+Boundary	4.762	8.073	2.145	3.34	3.535	5.888	4.604	8.795	1.296	2.307
SurFree	6.331	10.485	1.505	3.486	3.048	7.849	5.979	11.001	0.949	2.25
PAR+SurFree	4.078	6.989	1.224	2.589	2.183	4.603	4.015	7.959	1.008	1.912
CAB	4.214	8.034	1.966	3.978	2.364	10.554	3.646	12.058	1.121	2.084
PAR+CAB	**1.963**	**4.879**	**1.012**	**1.824**	**1.244**	**3.484**	**1.752**	**6.145**	**0.694**	**1.423**
Sign-OPT	30.581	36.062	19.56	22.152	29.566	38.994	20.952	38.496	6.392	12.083
PAR+Sign-OPT	4.525	8.067	2.602	3.73	3.578	6.679	4.91	9.387	1.353	2.548

More significant performance improvement combined with the existing decision-based attacks

Experiments

Results on Tiny-Imagenet

Target	res-18		inc-v3		inc-res		nasnet	
Methods	median	average	median	average	median	average	median	average
Initial	2.542	5.024	8.238	8.402	10.255	9.933	8.853	8.428
PAR	0.45	1.104	1.457	1.961	1.805	2.279	1.723	2.022
HSJA	0.959	2.762	3.479	4.576	5.053	5.603	4.226	5.237
PAR+HSJA	0.396	1.067	1.392	1.899	1.793	2.236	1.668	1.992
BBA	0.23	0.787	1.091	1.669	1.565	2.041	1.361	1.815
PAR+BBA	0.142	0.605	0.723	1.25	1.126	1.59	**0.948**	1.463
Evo	0.522	1.518	2.043	2.971	2.892	3.516	2.411	3.448
PAR+Evo	0.294	0.882	1.183	1.701	1.662	2.01	1.532	1.835
Boundary	0.577	1.194	1.552	2.091	2.38	2.807	1.967	2.388
PAR+Boundary	0.296	0.813	1.034	1.457	1.478	1.852	1.425	1.773
SurFree	0.143	0.653	0.627	1.233	1.126	1.772	0.963	1.639
PAR+SurFree	**0.14**	**0.599**	**0.629**	**1.171**	**1.087**	**1.479**	0.952	**1.453**
CAB	0.397	0.977	1.103	1.819	1.372	2.245	1.23	2.301
PAR+CAB	0.248	0.728	0.803	1.326	1.11	1.604	0.968	1.474
Sign-OPT	2.134	4.293	6.669	7.268	7.037	8.274	7.332	7.394
PAR+Sign-OPT	0.433	0.957	1.426	1.926	1.712	2.012	1.573	2.008

Effective when the target model is CNN

Experiments

	Initial Patch Size	112	112	112	112	56	56	56	28	28	14
	Minimum Patch Size	7	14	28	56	7	14	28	7	14	7
vgg-19	Mid Noise	**4.31**	5.07	5.55	6.21	4.34	4.88	5.54	4.60	5.09	4.79
	Avg Noise	**5.83**	7.11	8.17	8.84	5.92	7.20	8.33	6.11	7.43	6.47
	Avg Query Number	195.69	97.30	44.54	16.80	202.98	100.88	45.79	238.24	130.58	415.06
vit_s16	Mid Noise	8.76	9.32	9.54	10.35	**8.62**	9.17	9.67	9.01	9.88	9.17
	Avg Noise	17.24	19.08	19.96	20.52	**17.08**	18.84	19.69	17.43	19.16	17.90
	Avg Query Number	249.34	122.81	49.53	17.04	247.01	120.93	49.90	289.67	153.07	448.60

Ablation study different initial and final patch sizes

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

Query time and noise compression efficiency for decision-based attacks

Experiments

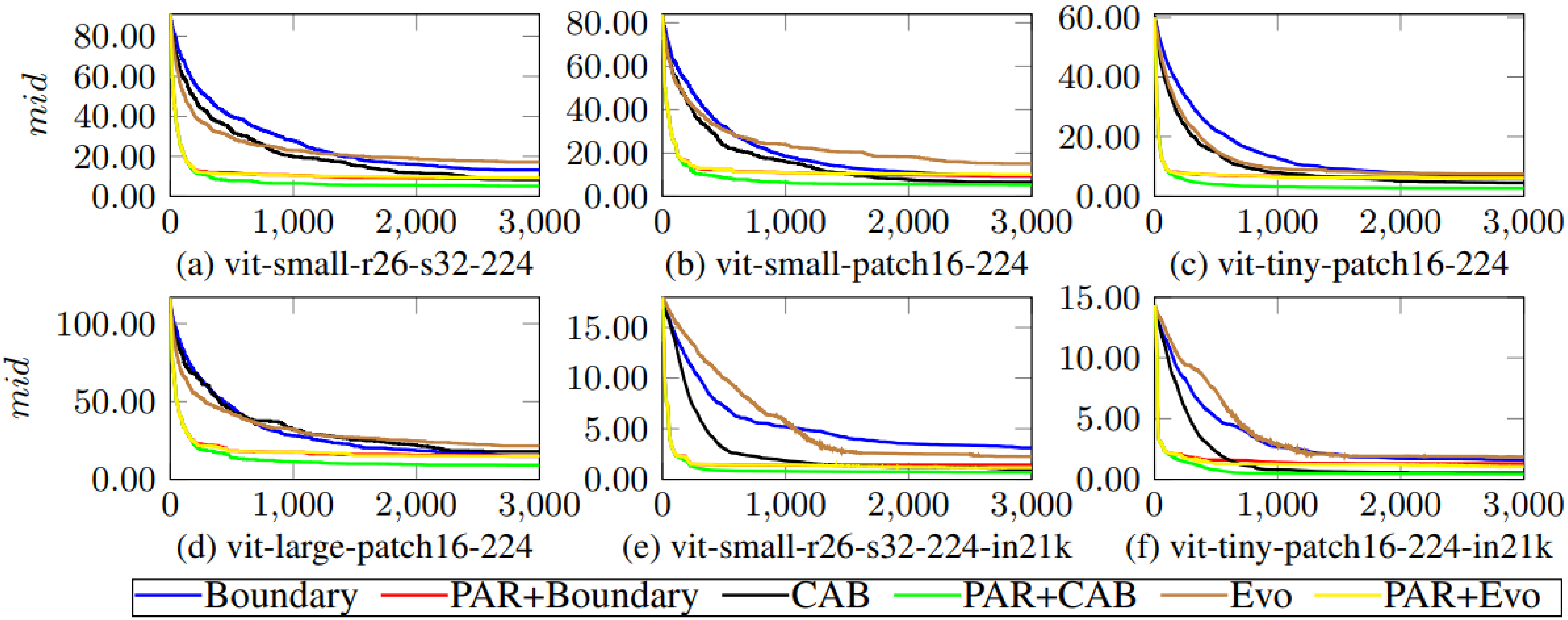
	Initial	PAR	HSJA	BBA	Evo	Boundary	SurFree
Mid	152.296	**39.821**	92.183	67.728	69.397	52.584	57.808
Avg	154.797	**40.792**	93.767	70.01	69.039	51.272	55.378

Results on targeted attack

	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

PAR compress noise under various patch size combinations

Experiments



Average noise magnitude decreases with the number of queries

Experiments