RisingAttacK

Adversarial Perturbations Are Formed by Iteratively Learning Linear Combinations of the Right Singular Vectors of the Adversarial Jacobian

https://github.com/ivmcl/ordered-topk-attack

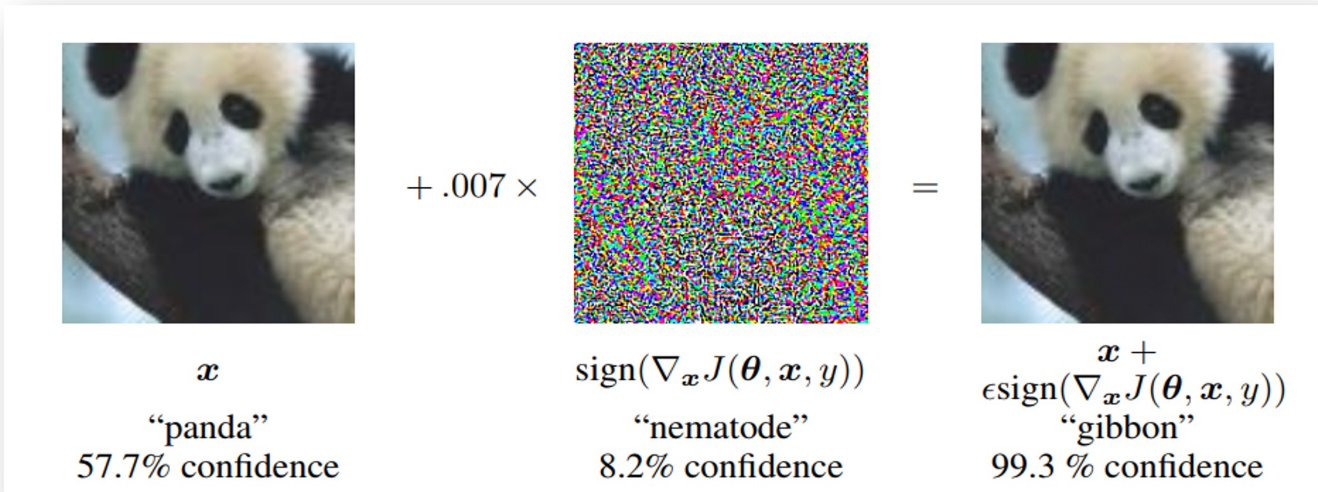
Thomas Paniagua, Chinmay Savadikar, Tianfu (Matt) Wu

NC STATE UNIVERSITY
Electrical and Computer Engineering

ICML
International Conference On Machine Learning
2025

Adversarial Attacks

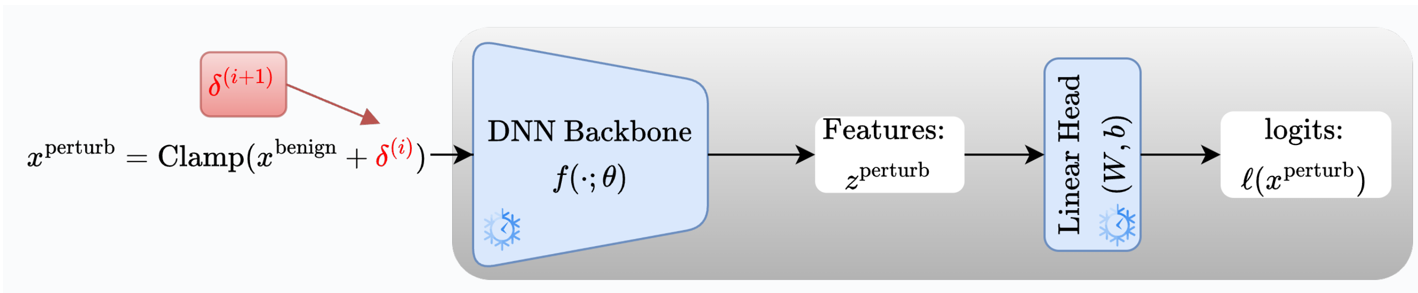
- Adversarial vulnerability of Deep Neural Networks (DNNs) has been intriguingly well-known.



- Adversarial attacks cause a catastrophic reduction in deep learning capability, especially via white-box targeted attacks.

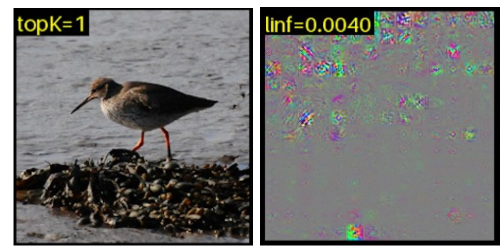
- C. Szegedy et al, *Intriguing properties of neural networks*, ICLR2013
- I. Goodfellow, J. Shlens and C. Szegedy, *Explaining and harnessing adversarial examples*, ICLR2015

Targeted Attacks: From Top-1 to Ordered Top-K

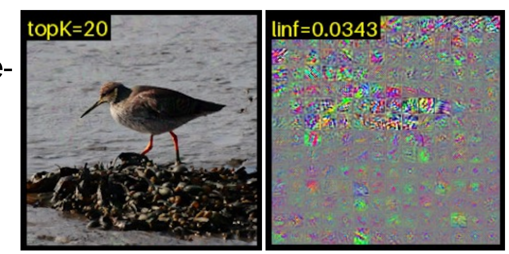


ViT-B
Benign Top-K predictions: redshank, ruddy turnstone, red-backed sandpiper, dowitcher, ...

Top-1 attack target: mask



Ordered Top-20 attack targets: mask, analog-clock, slide-rule, Siberian-husky, harmonica, African-chameleon, dowitcher, yena, wing, pillow, garter-snake, Great-Pyrenees, puffer, banana, West-Highland-white-terrier, whippet, brown-bear, snowplow, tarantula, space-heater

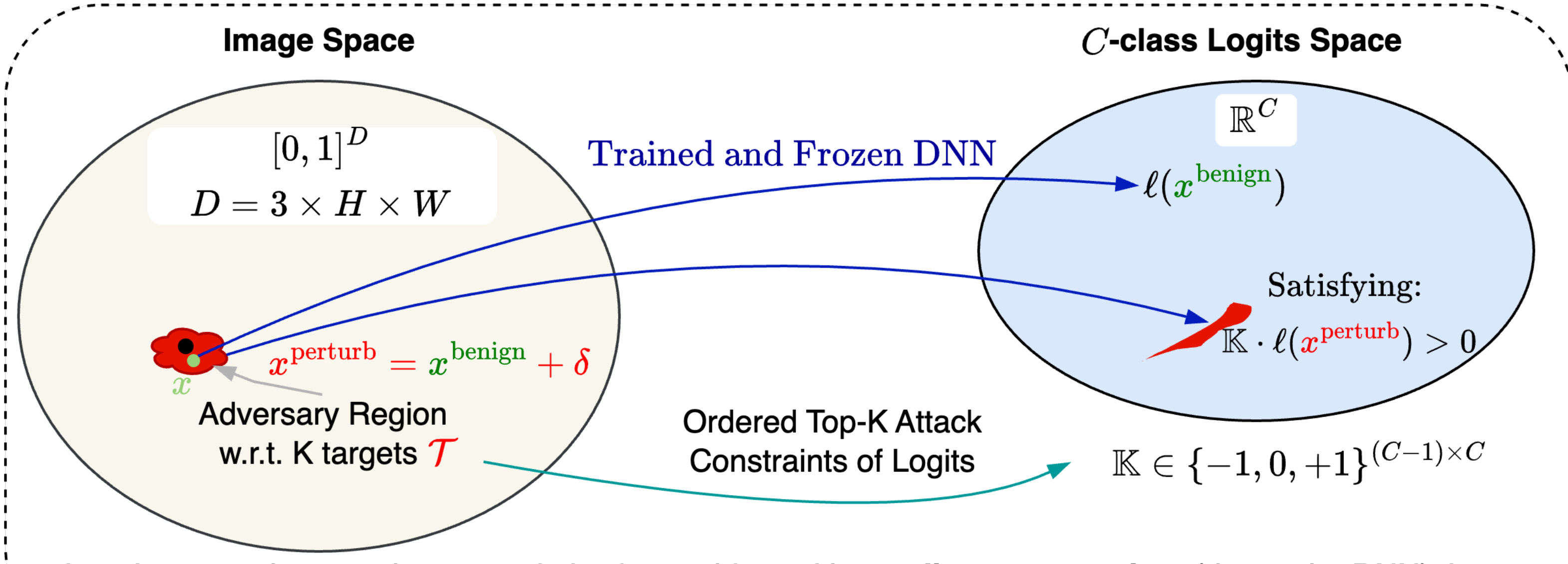


- Z. Zhang and T. Wu, *Learning Ordered Top-k Adversarial Attacks via Adversarial Distillation*, In CVPRW'20
- T. Paniagua, R. Grainger and T. Wu, *QuadAttacK: a quadratic programing approach to learning ordered top-K adversarial attacks*, In NeurIPS'23

Why Do Ordered Top-K Attacks Matter?

- Reveal deeper vulnerability of DNNs
- Safety-critical systems (e.g., face unlock, medical triage, content moderation) reason over entire ranked lists.
  - An attacker dictating all top predictions (e.g., using semantic coherent attack targets) obtains finer control and evades simple “Top-1 changed” detectors.
- Security evaluations now recommend K > 1.
  - E.g., NIST SP 800-226 (March 2025)

Learning Ordered Top-K Attacks



Leaning δ can be posed as an optimization problem with **non-linear constraints** (due to the DNN) in the image space, which is **intractable** in the vanilla form.

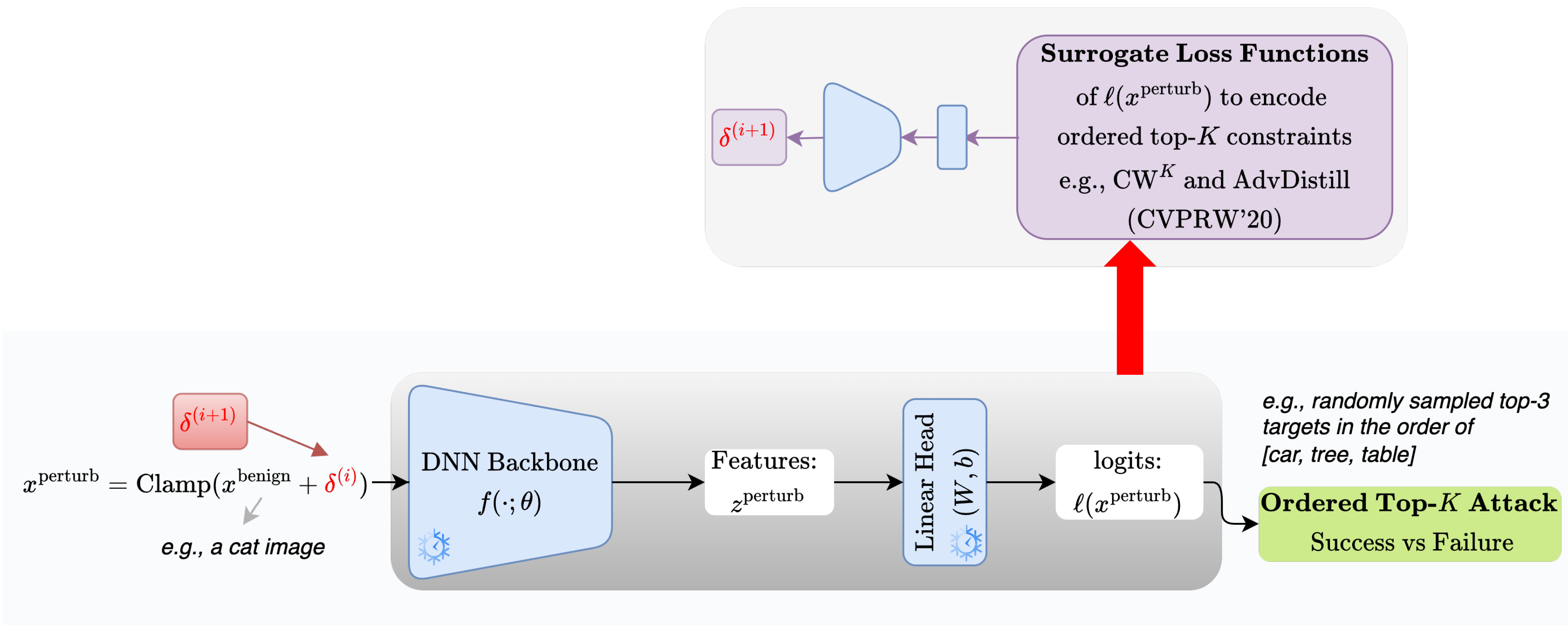
$$\underset{x \in [0,1]^D}{\text{minimize}} \quad ||x - x^{\text{benign}}||_p,$$
$$\text{subject to} \quad \mathbb{K} \cdot \ell(x) > 0,$$

Contributions of Our Proposed RisingAttacK

- Novel Theoretical Insights:
  - It introduces explicit derivations connecting adversarial perturbations to singular vectors of the adversarial Jacobian, providing new theoretical clarity.
- Methodological Innovation:
  - It is the first method to directly optimize ordered top-K adversarial attacks in image space via SQP, significantly improving alignment between optimized solutions and visually coherent perturbations.
- Empirical Advances:
  - It provides comprehensive evaluation across multiple architectures and attack levels, consistently outperforming the previous state-of-the-art, QuadAttacK using a proposed holistic metric, Figure of Merits (FoM) covering both success rates and perturbation magnitudes.

Learning Ordered Top-K Attacks – The Prior Art

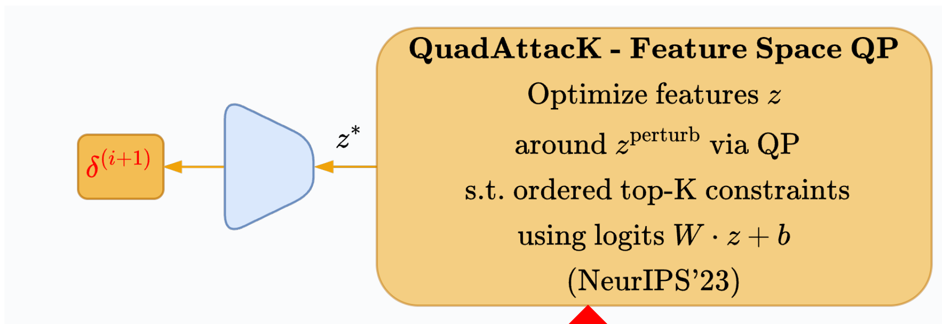
- From *constrained* optimization to *unconstrained* ones via designing surrogate loss functions

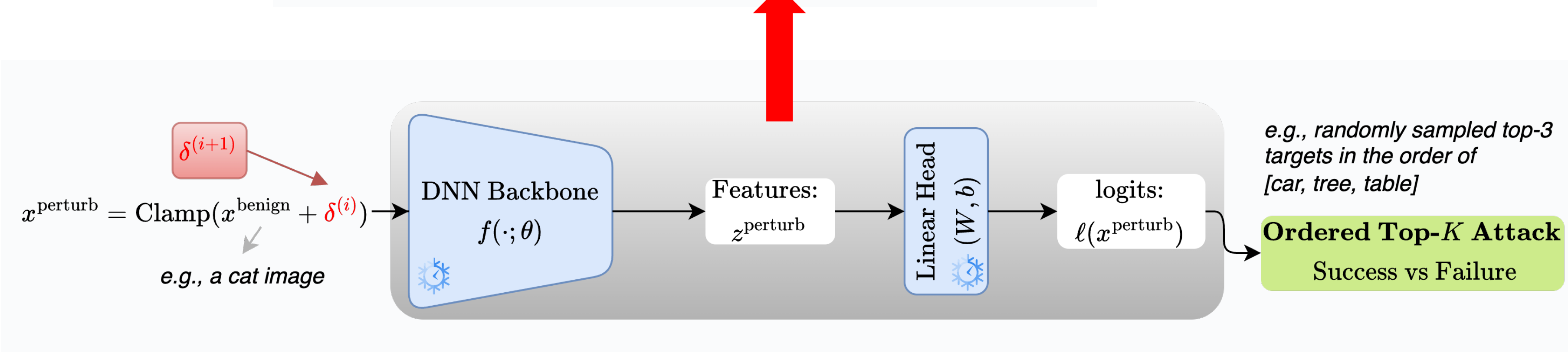


- Z. Zhang and T. Wu, *Learning Ordered Top-k Adversarial Attacks via Adversarial Distillation*, In CVPRW'20

Learning Ordered Top-K Attacks – The Prior Art

- From latent feature perturbation (via Quadratic Programing, QP) to image perturbation (via one-step back-propagation)

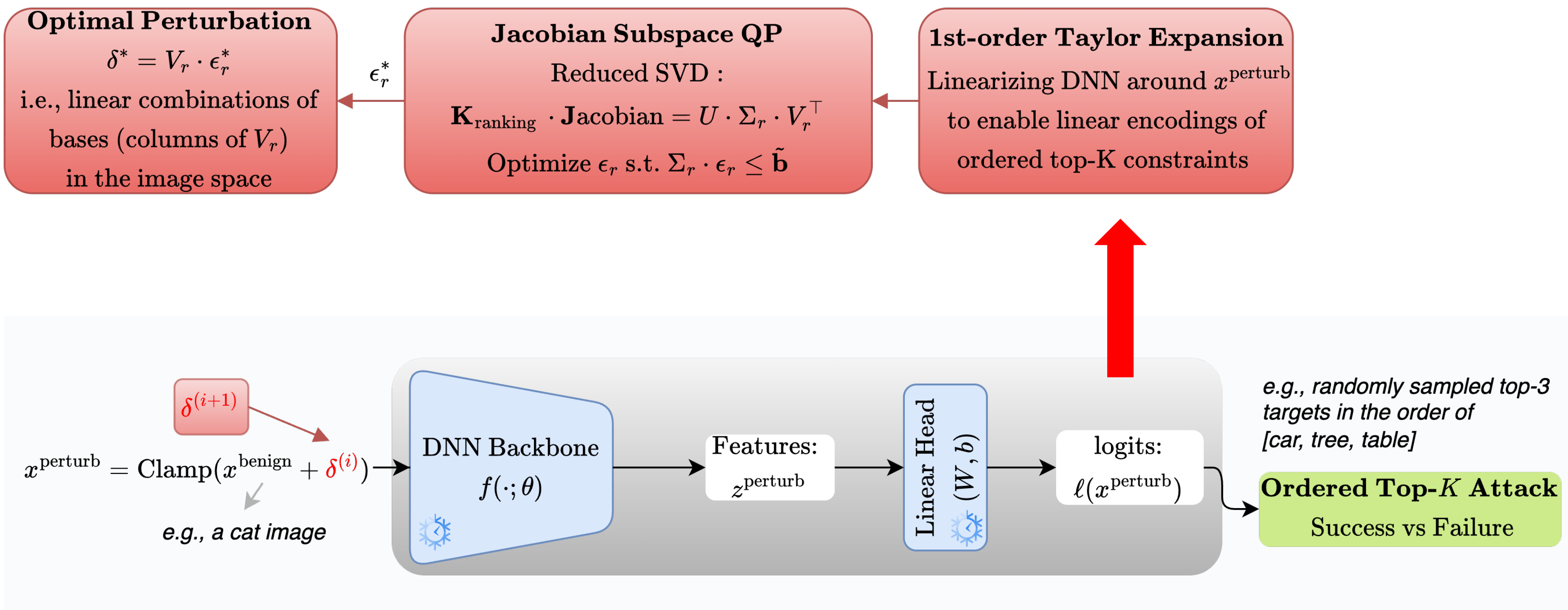




- T. Paniagua, R. Grainger and T. Wu, *QuadAttacK: a quadratic programing approach to learning ordered top-K adversarial attacks*, In NeurIPS'23

Learning Ordered Top-K Attacks – Our RisingAttacK

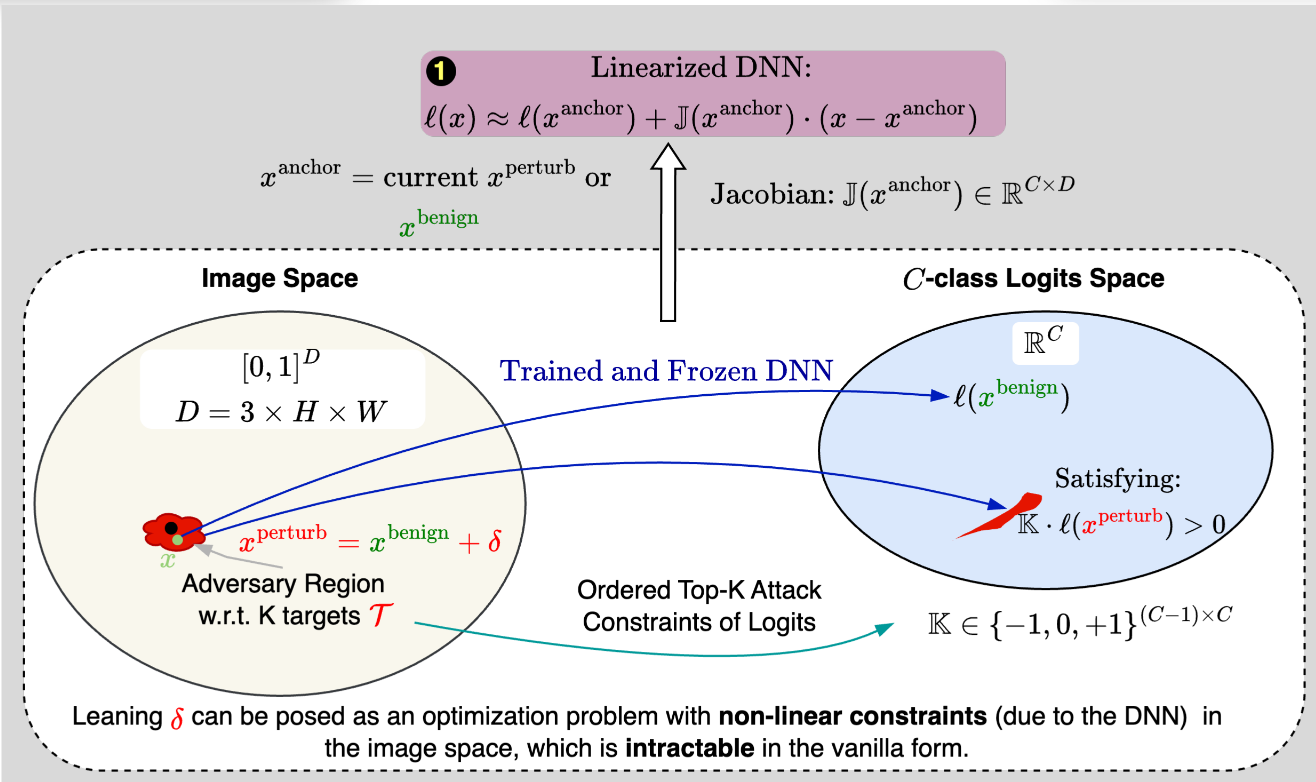
- Directly optimize adversarial perturbation in the image space via Sequential QP (SQP)



RisingAttack

$$\underset{x\in[0,1]^D}{\text{minimize}} \quad ||x - x^{\text{benign}}||_p$$
$$\text{subject to} \quad \mathbb{K}\cdot\ell(x) > 0,$$

$$\underset{x\in[0,1]^D}{\text{minimize}} \quad ||x - x^{\text{anchor}}||_2^2,$$
$$\text{s.t.} \quad \mathbb{K}\cdot\left(\ell(x^{\text{anchor}}) + \mathbb{J}(x^{\text{anchor}})\cdot(x - x^{\text{anchor}})\right) > 0$$

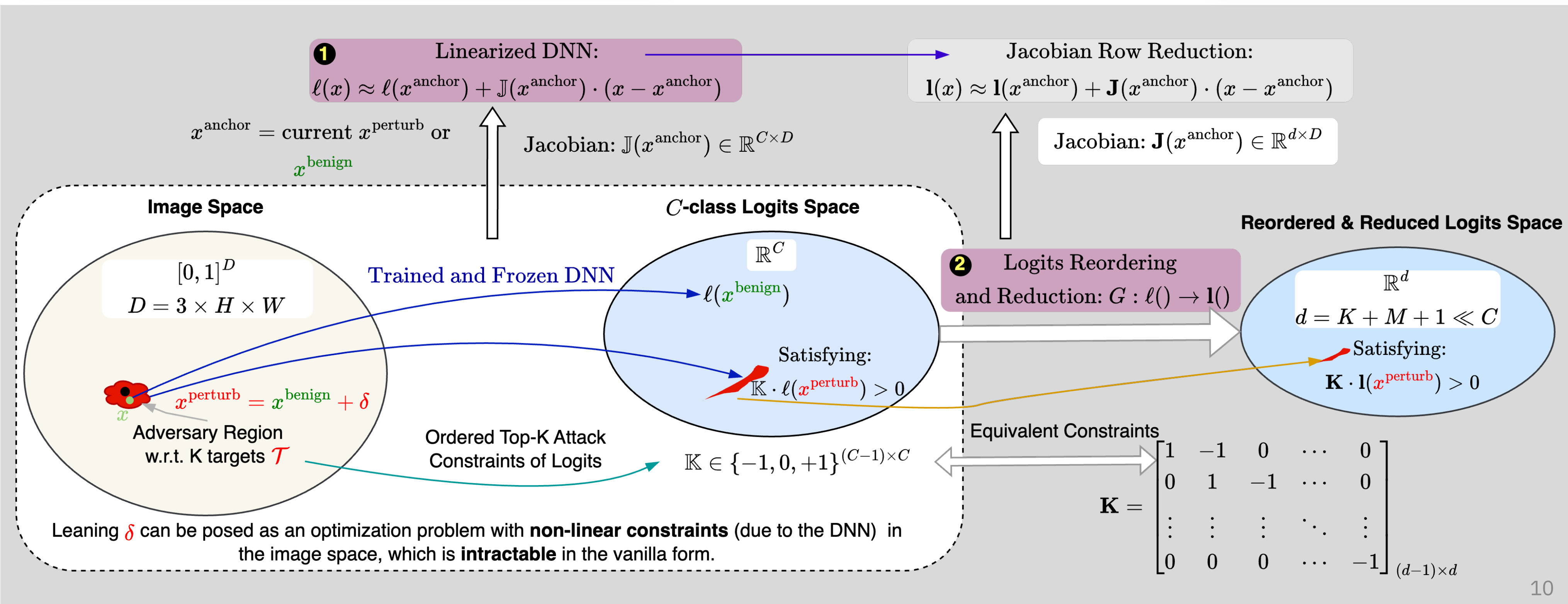


RisingAttacK

$$\underset{x\in[0,1]^D}{\text{minimize}} \quad ||x - x^{\text{anchor}}||_2^2,$$
$$\text{s.t.} \quad \mathbb{K}\cdot\left(\ell(x^{\text{anchor}}) + \mathbb{J}(x^{\text{anchor}})\cdot(x - x^{\text{anchor}})\right) > 0$$

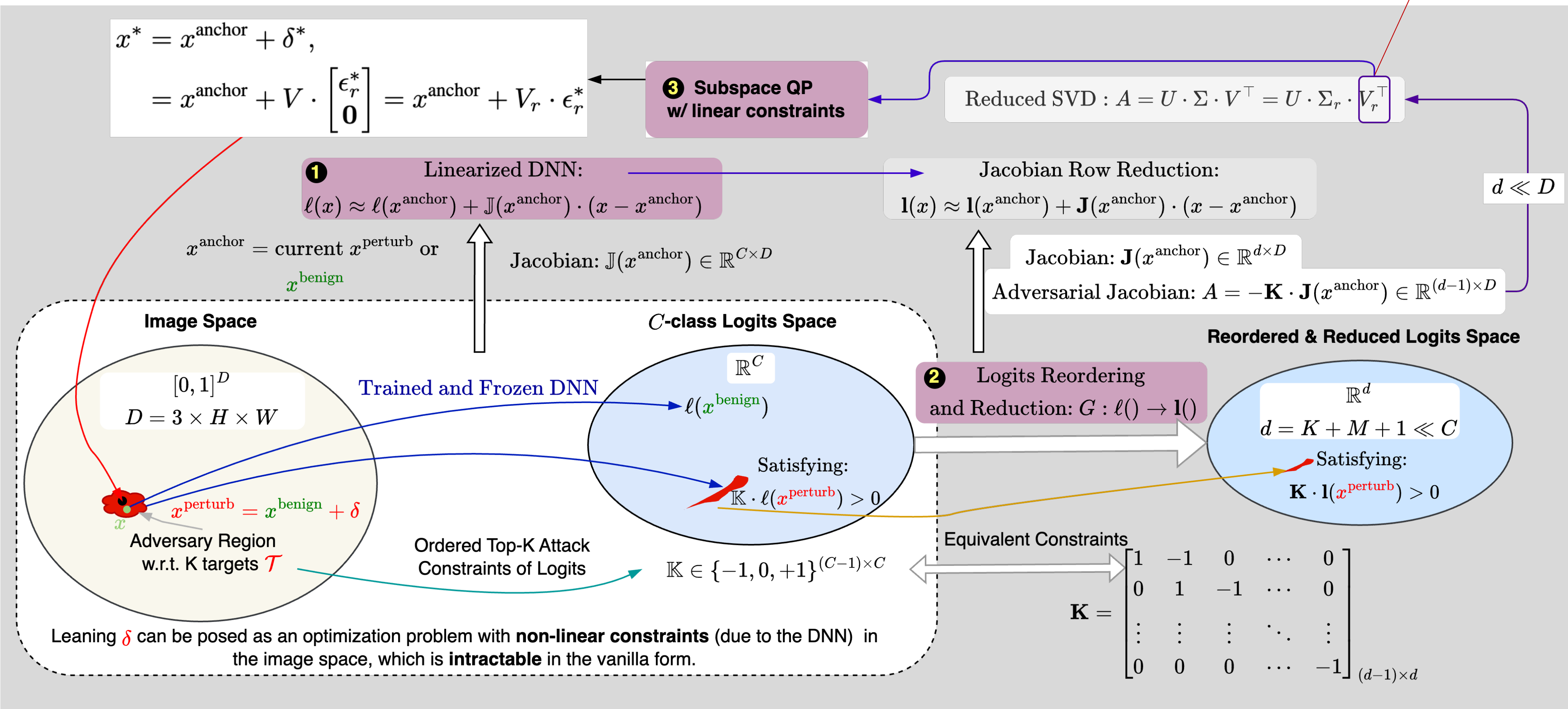
$$A = -\mathbf{K}\cdot\mathbf{J}(x^{\text{anchor}}) \quad \text{Adversarial Jacobian}$$
$$\mathbf{b} = \mathbf{K}\cdot\left(\mathbf{l}(x^{\text{anchor}}) - \mathbf{J}(x^{\text{anchor}})\cdot x^{\text{anchor}}\right) + \mathbf{m}$$

$$\underset{x\in[0,1]^D}{\text{minimize}} \quad ||x - x^{\text{anchor}}||_2^2,$$
$$\text{subject to} \quad A\cdot x \le \mathbf{b},$$



RisingAttack

Its columns span a subspace in which adversarial perturbations are most effective towards satisfying ordered top-K constraints.



Experiments

- Four models trained on ImageNet-1k
  - ResNet50, DenseNet121, ViT-B, DEiT-B
- 1000 images from ImageNet-1k val
  - which can be correctly classified by all the four models
- K=1,5,10,15,20,25,30
- For each K, 5 random seeds are used
- Metrics
  - Attack Success Rate (ASR), l_p $(p = 1,2,\infty)$ energy
  - Figure of Merits (FoM)
    - a holistic comparison between Method 1 and 2

$$\mathrm{FoM} = \frac{\mathrm{ASR}^1}{\mathrm{ASR}^2} \cdot \frac{1}{3} \cdot \sum_{p \in \{1,2,\infty\}} \frac{\ell_p^2}{\ell_p^1}$$

Experiments

(a) ResNet-50 (He et al., 2016)

Top-K	Method	Mean				Time (s/img) ↓	FoM↑
		ASR ↑	ℓ_1 ↓	ℓ_2 ↓	ℓ_∞ ↓		
Top-30	QuadAttackK_{60}	0.2076	11.8070	3654.9139	0.1349	**3.3947**	6.4793
	RisingAttackK_{60}	**0.6642**	**7.0271**	**2081.8960**	**0.0511**	17.0013	
	QuadAttackK_{30}	Failed				**1.6539**	inf
	RisingAttackK_{30}	**0.0022**	6.2378	1844.3013	0.0470	8.5619	
Top-25	QuadAttackK_{60}	0.6018	11.6214	3599.8101	0.1301	**3.4167**	3.6439
	RisingAttackK_{60}	**0.8420**	**5.2960**	**1561.6462**	**0.0393**	14.0839	
	QuadAttackK_{30}	0.0018	10.4263	3259.2773	0.0991	**1.7058**	48.9628
	RisingAttackK_{30}	**0.0392**	**5.1218**	**1511.3347**	**0.0388**	7.0999	
Top-20	QuadAttackK_{60}	**0.8344**	10.0891	3133.6199	0.1079	**3.4039**	3.1100
	RisingAttackK_{60}	0.8306	**3.7474**	**1101.1521**	**0.0281**	6.7267	
	QuadAttackK_{30}	**0.0978**	9.0948	2850.0433	0.0858	**1.7264**	1.9481
	RisingAttackK_{30}	0.0666	**3.4854**	**1022.5585**	**0.0269**	3.7216	
Top-15	QuadAttackK_{60}	0.9440	8.3368	2600.7510	0.0822	**3.4839**	3.2229
	RisingAttackK_{60}	**0.9868**	**3.0150**	**878.9222**	**0.0233**	5.1634	
	QuadAttackK_{30}	0.4922	7.8296	2451.8036	0.0717	**1.7382**	3.3674
	RisingAttackK_{30}	**0.5856**	**2.9944**	**873.3877**	**0.0234**	2.8794	
Top-10	QuadAttackK_{60}	0.9866	6.5228	2044.5753	0.0576	3.7396	3.3482
	RisingAttackK_{60}	**0.9936**	**2.0825**	**602.1784**	**0.0167**	**3.3991**	
	QuadAttackK_{30}	**0.8460**	6.3547	1994.8023	0.0544	**1.7593**	2.9244
	RisingAttackK_{30}	0.8064	**2.1748**	**630.0922**	**0.0175**	1.7965	
Top-5	QuadAttackK_{60}	**0.9968**	4.0029	1261.2314	0.0309	4.5257	3.3373
	RisingAttackK_{60}	0.9558	**1.1534**	**330.1495**	**0.0098**	**1.8225**	
	QuadAttackK_{30}	**0.9590**	3.9539	1246.4929	0.0300	2.1458	2.6681
	RisingAttackK_{30}	0.9504	**1.4693**	**420.0254**	**0.0124**	**0.9517**	
Top-1	QuadAttackK_{60}	**0.9996**	1.4443	467.1178	0.0083	5.3373	2.1564
	RisingAttackK_{60}	0.9992	**0.6144**	**165.8517**	**0.0064**	**0.6114**	
	QuadAttackK_{30}	0.9772	1.4244	461.1199	0.0080	2.6411	1.4638
	RisingAttackK_{30}	**0.9986**	**0.9155**	**251.6174**	0.0088	**0.3201**	

(b) DenseNet-121 (Huang et al., 2017)

Top-K	Method	Mean				Time (s/img) ↓	FoM↑
		ASR ↑	ℓ_1 ↓	ℓ_2 ↓	ℓ_∞ ↓		
Top-30	QuadAttackK_{60}	Failed				**4.5409**	inf
	RisingAttackK_{60}	**0.4074**	14.7263	4393.8482	0.1051	20.3156	
	QuadAttackK_{30}	Failed				**2.3266**	0
	RisingAttackK_{30}	Failed				10.2335	
Top-25	QuadAttackK_{60}	0.1734	13.1825	4053.5759	0.1531	**4.1657**	8.5496
	RisingAttackK_{60}	**0.9370**	**9.9898**	**2945.3574**	**0.0747**	16.8643	
	QuadAttackK_{30}	Failed				**2.2016**	inf
	RisingAttackK_{30}	**0.1094**	9.9203	2921.6770	0.0756	8.5279	
Top-20	QuadAttackK_{60}	0.8340	11.6266	3583.3589	0.1268	**4.0066**	2.6290
	RisingAttackK_{60}	**0.9812**	**5.9921**	**1744.8239**	**0.0468**	8.2901	
	QuadAttackK_{30}	0.0330	9.8564	3072.6790	0.0923	**2.0206**	23.7723
	RisingAttackK_{30}	**0.4500**	**6.1377**	**1786.5613**	**0.0485**	4.6070	
Top-15	QuadAttackK_{60}	0.9866	9.2713	2884.2755	0.0887	**3.8963**	2.3310
	RisingAttackK_{60}	**1.0000**	**4.3657**	**1252.9889**	**0.0359**	6.2878	
	QuadAttackK_{30}	0.5088	8.6281	2697.9823	0.0771	**1.8919**	3.5524
	RisingAttackK_{30}	**0.9362**	**4.7380**	**1362.7350**	**0.0387**	3.5501	
Top-10	QuadAttackK_{60}	0.9986	6.7558	2123.4894	0.0545	**3.8256**	2.5458
	RisingAttackK_{60}	**1.0000**	**2.6903**	**759.1986**	**0.0235**	4.2223	
	QuadAttackK_{30}	0.9392	6.6701	2098.0095	0.0531	**1.8918**	2.4272
	RisingAttackK_{30}	**0.9880**	**2.9210**	**827.1606**	**0.0253**	2.2937	
Top-5	QuadAttackK_{60}	**0.9998**	3.9671	1258.1706	0.0264	3.8644	3.0870
	RisingAttackK_{60}	0.9994	**1.2169**	**331.6714**	**0.0119**	**2.2643**	
	QuadAttackK_{30}	0.9924	3.9526	1253.5745	0.0262	1.8502	2.2794
	RisingAttackK_{30}	**0.9982**	**1.6603**	**457.9204**	**0.0156**	**1.2082**	
Top-1	QuadAttackK_{60}	**1.0000**	1.5191	503.0047	0.0070	3.0413	1.9466
	RisingAttackK_{60}	1.0000	**0.7001**	**177.1356**	0.0085	**0.8046**	
	QuadAttackK_{30}	0.9960	1.5144	501.4779	0.0070	1.5519	1.2739
	RisingAttackK_{30}	**1.0000**	**1.0708**	**280.0300**	0.0116	**0.4255**	

Experiments

(c) ViT-B (Dosovitskiy et al., 2020)

Top-K	Method	Mean				Time (s/img) ↓	FoM↑
		ASR ↑	ℓ_1 ↓	ℓ_2 ↓	ℓ_∞ ↓		
Top-30	QuadAttacK$_{60}$	0.3272	**9.6708**	2938.2587	0.1032	**5.2135**	3.1589
	RisingAttacK$_{60}$	**0.9534**	9.7262	**2721.2876**	**0.0876**	43.3954	
	QuadAttacK$_{30}$	Failed				**2.7870**	inf
	RisingAttacK$_{30}$	**0.5568**	11.4132	3206.4565	0.1029	21.7179	
Top-25	QuadAttacK$_{60}$	0.6872	9.4331	2860.6667	0.1002	**5.2723**	2.6425
	RisingAttacK$_{60}$	**0.9944**	**5.5706**	**1520.5703**	**0.0526**	36.0486	
	QuadAttacK$_{30}$	Failed				**2.7354**	inf
	RisingAttacK$_{30}$	**0.7536**	7.7050	2126.3211	0.0721	18.0775	
Top-20	QuadAttacK$_{60}$	0.7828	7.9108	2393.0875	0.0815	**5.0069**	2.8308
	RisingAttacK$_{60}$	**0.9864**	**3.7609**	**1007.6887**	**0.0360**	15.8230	
	QuadAttacK$_{30}$	0.0004	6.3770	1992.7502	0.0533	**2.6210**	1610.8632
	RisingAttacK$_{30}$	**0.4956**	**4.9482**	**1343.2135**	**0.0473**	7.9615	
Top-15	QuadAttacK$_{60}$	0.8404	6.2661	1893.5173	0.0620	**4.7622**	2.7231
	RisingAttacK$_{60}$	**0.9988**	**2.8751**	**753.1852**	**0.0284**	11.9841	
	QuadAttacK$_{30}$	0.0056	4.7982	1495.8188	0.0385	**2.4245**	164.8583
	RisingAttacK$_{30}$	**0.7510**	**3.8944**	**1038.7394**	**0.0379**	6.0305	
Top-10	QuadAttacK$_{60}$	0.9130	4.5246	1374.2282	0.0410	**4.6368**	2.5247
	RisingAttacK$_{60}$	**0.9936**	**1.9915**	**508.8791**	**0.0206**	8.2583	
	QuadAttacK$_{30}$	0.0252	3.4999	1094.6987	0.0261	**2.3034**	36.7947
	RisingAttacK$_{30}$	**0.7112**	**2.6247**	**684.1576**	0.0267	4.1602	
Top-5	QuadAttacK$_{60}$	**0.9980**	3.6439	1128.3054	0.0288	**4.3981**	1.7630
	RisingAttacK$_{60}$	0.5712	**1.1650**	**292.6494**	**0.0128**	4.4038	
	QuadAttacK$_{30}$	0.5024	3.2930	1029.8490	0.0242	**2.1108**	2.3688
	RisingAttacK$_{30}$	**0.5980**	**1.6101**	**406.4644**	**0.0174**	2.2197	
Top-1	QuadAttacK$_{60}$	**0.9998**	1.5736	509.7575	0.0081	2.6007	3.2121
	RisingAttacK$_{60}$	0.9388	**0.4365**	**96.0745**	**0.0060**	**1.2715**	
	QuadAttacK$_{30}$	**0.9958**	1.5681	508.0591	**0.0081**	1.3040	2.1102
	RisingAttacK$_{30}$	0.9362	**0.6578**	**149.1661**	0.0086	**0.6417**	

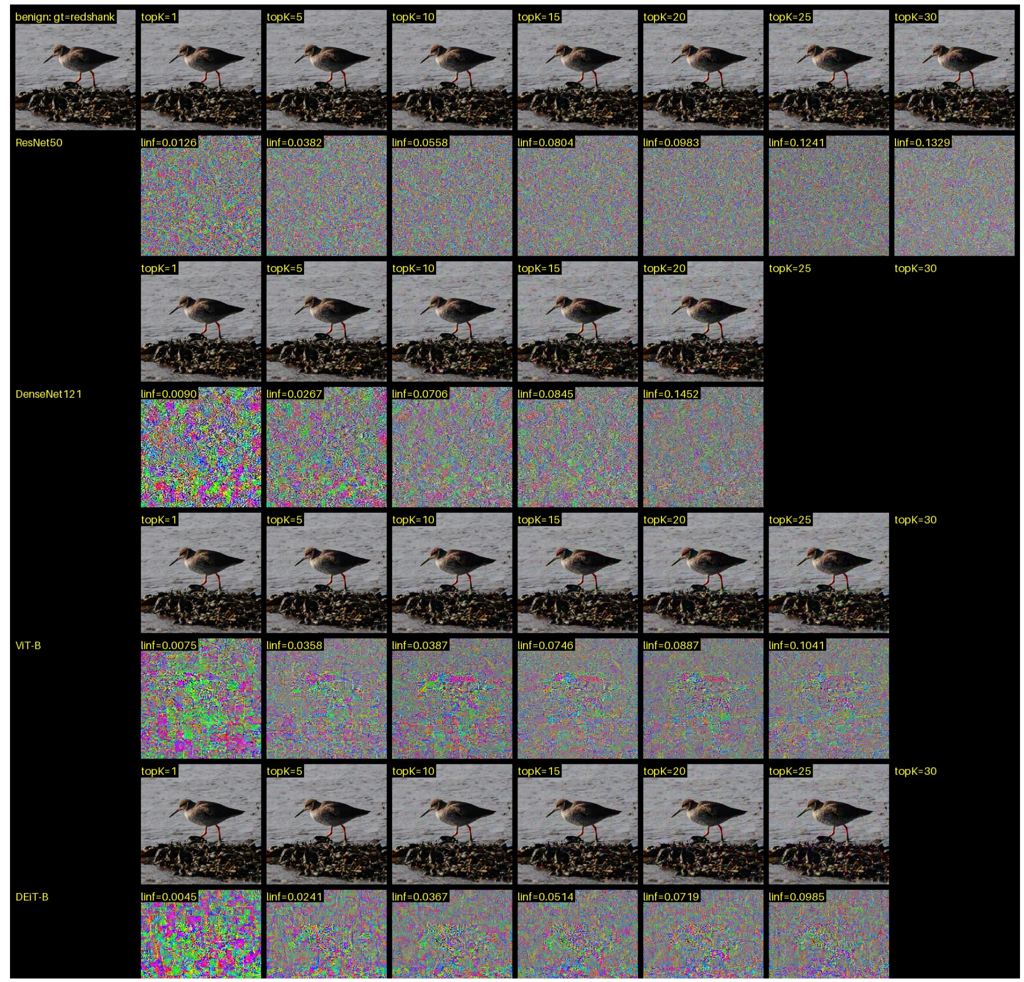
(d) DEiT-B (Touvron et al., 2021)

Top-K	Method	Mean				Time (s/img) ↓	FoM↑
		ASR ↑	ℓ_1 ↓	ℓ_2 ↓	ℓ_∞ ↓		
Top-30	QuadAttacK$_{60}$	0.0640	**9.3734**	2860.9240	0.0997	**4.1792**	8.8333
	RisingAttacK$_{60}$	**0.5150**	9.4432	**2697.9176**	**0.0804**	43.3521	
	QuadAttacK$_{30}$	Failed				**2.3032**	inf
	RisingAttacK$_{30}$	**0.0600**	11.0771	3165.6910	0.0957	21.6930	
Top-25	QuadAttacK$_{60}$	0.8644	9.3780	2849.8222	0.0960	**4.0966**	2.1975
	RisingAttacK$_{60}$	**0.9854**	**5.1921**	**1434.6160**	**0.0482**	36.1084	
	QuadAttacK$_{30}$	Failed				**2.2173**	inf
	RisingAttacK$_{30}$	**0.6748**	6.3220	1763.4334	0.0581	18.1108	
Top-20	QuadAttacK$_{60}$	0.9612	7.6974	2343.5441	0.0735	**4.1868**	2.8466
	RisingAttacK$_{60}$	**0.9956**	**2.9174**	**781.2607**	**0.0282**	15.8331	
	QuadAttacK$_{30}$	0.0032	6.2491	1950.0525	0.0524	**2.1503**	325.7059
	RisingAttacK$_{30}$	**0.6348**	**3.8373**	**1045.3953**	**0.0366**	7.9624	
Top-15	QuadAttacK$_{60}$	0.9750	6.0671	1852.4958	0.0544	**3.9525**	2.8819
	RisingAttacK$_{60}$	**1.0000**	**2.2015**	**573.3811**	**0.0223**	11.9810	
	QuadAttacK$_{30}$	0.0338	4.9874	1558.1460	0.0386	**2.0234**	42.8983
	RisingAttacK$_{30}$	**0.9278**	**3.1490**	**838.2295**	**0.0310**	6.0263	
Top-10	QuadAttacK$_{60}$	0.9762	4.3693	1346.6326	0.0353	**3.8755**	2.9455
	RisingAttacK$_{60}$	**0.9996**	**1.5076**	**379.6582**	**0.0162**	8.2610	
	QuadAttacK$_{30}$	0.1298	3.5782	1123.3760	0.0256	**1.9552**	11.4300
	RisingAttacK$_{30}$	**0.9200**	**2.1465**	**556.2741**	**0.0222**	4.1613	
Top-5	QuadAttacK$_{60}$	0.9984	3.3975	1064.7252	0.0243	**3.4381**	3.1378
	RisingAttacK$_{60}$	**0.9992**	**1.0575**	**254.5953**	**0.0121**	4.4027	
	QuadAttacK$_{30}$	0.7794	3.2526	1024.0607	0.0225	**1.7718**	2.6286
	RisingAttacK$_{30}$	**0.8800**	**1.3450**	**334.9398**	**0.0149**	2.2165	
Top-1	QuadAttacK$_{60}$	**1.0000**	1.3910	459.6084	0.0063	2.9955	3.9437
	RisingAttacK$_{60}$	0.9794	**0.3340**	**68.5738**	**0.0052**	**1.2708**	
	QuadAttacK$_{30}$	**0.9994**	1.3899	459.3060	0.0063	1.4404	2.4502
	RisingAttacK$_{30}$	0.9772	**0.5249**	**114.2980**	0.0073	**0.6426**	

Experiments

RisingAttacK

QuadAttacK



RisingThanK You!

This research is partly supported by NSF IIS-1909644, ARO Grant W911NF1810295, ARO Grant W911NF2210010, NSF CMMI-2024688 and NSF IUSE-2013451. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ARO, NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon.

RisingAttacK https://github.com/ivmcl/ordered-topk-attack

Thomas Paniagua, Chinmay Savadikar, Tianfu (Matt) Wu



NC STATE UNIVERSITY Electrical and Computer Engineering

ICML International Conference On Machine Learning 2025