35TH USENIX SECURITY SYMPOSIUM



Low-Cost Hard-Label Adversarial Attack with Theoretical Foundations

Jun Liu[1,2], Leo Yu Zhang[3], Fengpeng Li[1], Isao Echizen[2,4], Jiantao Zhou[1]

[1]University of Macau, [2] National Institute of Informatics, [3]Griffith University, [4]University of Tokyo









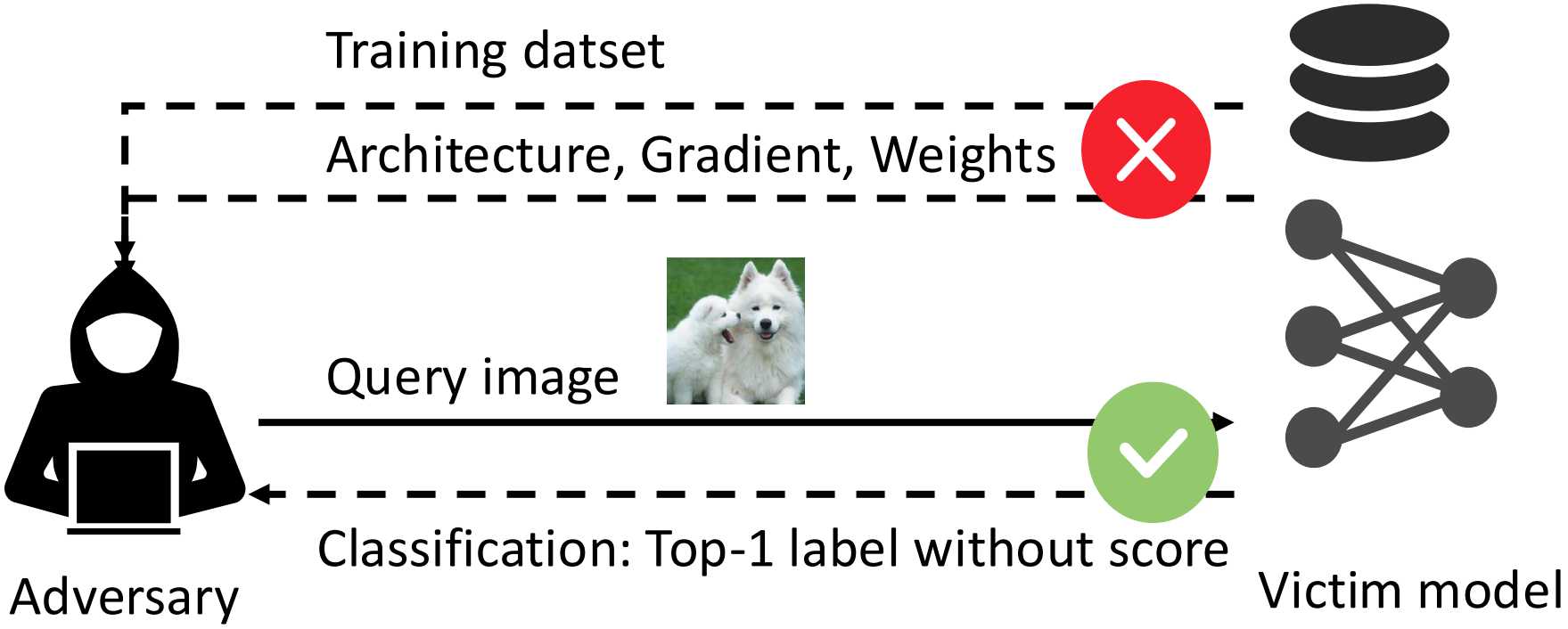
Outline

- **Background & Motivation**
- **Our Method (DPAttack)**
- **Representative Experimental Results**
- **Takeaway**

Our full paper: https://arxiv.org/abs/2601.14300

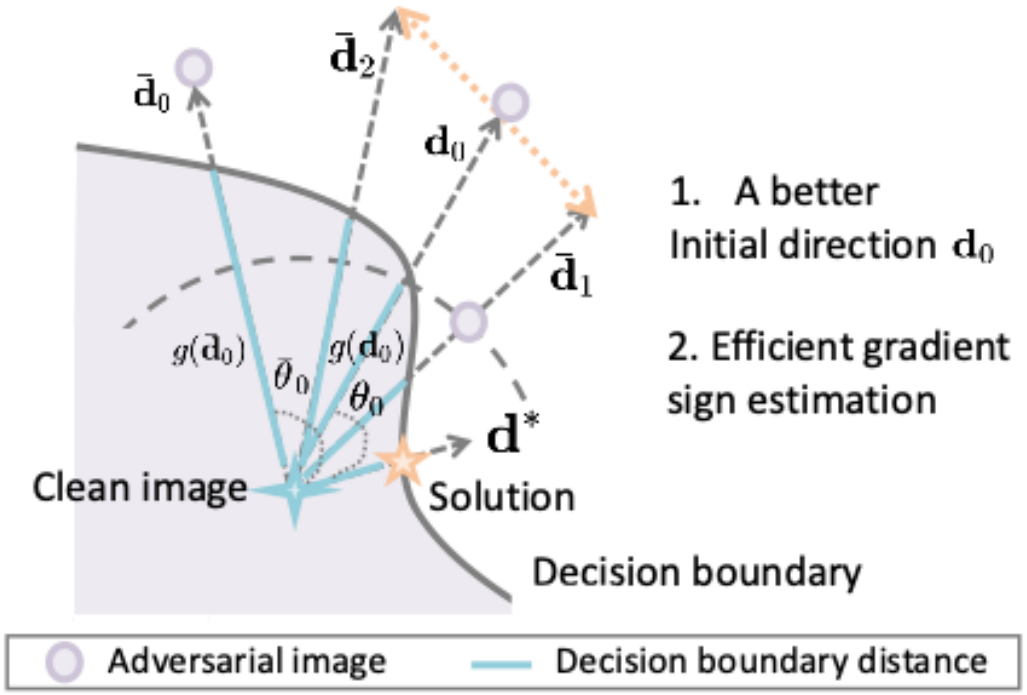
Background

Threat Model: Hard-label black-box attack setting:



Motivation

Traditional hard-label attack workflow:
Stage 1: Perturbation direction initialization.
Stage 2: Direction refinement and perturbation reduction.



1. A better Initial direction $\mathbf{d}_0$

2. Efficient gradient sign estimation

Our motivation:
Stage 1: Better initialization (w/o surrogate model).
Stage 2: More efficient gradient sign estimation.

Note: We theoretically model the state-of-the-art L_∞ sign-flip direction search as estimating the gradient sign

Theorem 2 (Gradient Approximation Guarantee for HRayS). *Assume that the gradient magnitude is uniformly bounded, i.e.,* $\|\nabla\mathcal{L}(\mathbf{x},y)\|_\infty \le G_{\max}$. *Let* η_l *denote the maximum deviation between the directional derivative influence of the subset* $B_{ks}^{(l)}$ *selected by HRayS at round* l *and that of the optimal subset* $\arg\max_{B_k^{(l)}} I(B_k^{(l)})$. *Let* $\hat{\mathbf{d}}^*$ *be the final direction produced by HRayS. Then, the alignment between* $\hat{\mathbf{d}}^*$ *and the true gradient sign satisfies*

$$\operatorname{sgn}(\nabla\mathcal{L}(\mathbf{x},y))^\top\hat{\mathbf{d}}^* \ge \frac{\|\nabla\mathcal{L}(\mathbf{x},y)\|_1}{G_{\max}}\cdot(1-\varpi), \quad (9)$$

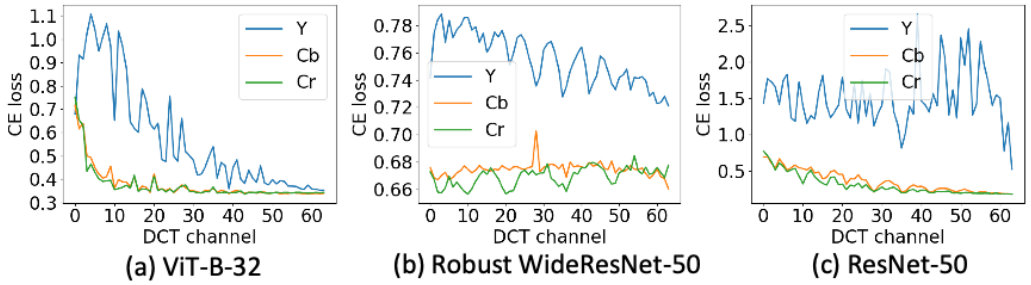
where $\varpi = \sum_{l=1}^{\lceil\log_2 d\rceil}\eta_l\|\nabla\mathcal{L}(\mathbf{x},y)\|_1^{-1}$ *denotes the cumulative relative error induced by suboptimal subset selections across all hierarchical rounds.*

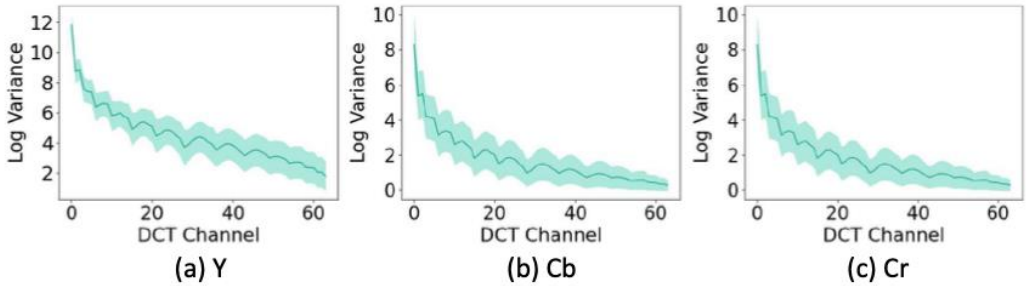
Motivation

- **Problem formulation:**
  Find a perturbation direction **d** that can minimize the adversarial perturbation g(**d**).

$$g(\mathbf{d}) = \min\{r > 0 \mid \mathcal{F}(\mathbf{x} + r \cdot \mathbf{d} \neq y)\}.$$

- **Our Observation**: Block-DCT model sensitivity curve is positively corelated with the frequency variance curve of clean images.





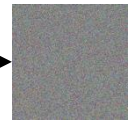
Channel	ViT-B-32	WRS-50	ResNet-50
Y	$0.8074^{7.69E-16}$	$0.7024^{9.87E-11}$	$0.0327^{7.97E-01}$
Cb	$0.9274^{3.61E-28}$	$-0.1714^{1.76E-01}$	$0.9137^{6.18E-26}$
Cr	$0.8844^{3.44E-22}$	$-0.4087^{7.99E-04}$	$0.9618^{1.40E-36}$

[1] $\bar{r}$: the correlation coefficient; p-value: the statistical significance; WRS-50: Robust WideResNet-50.

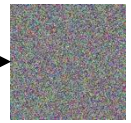
Method: Stage 1. A Better Initial Direction



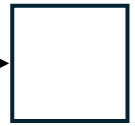
Traditional Initilizations:



Gaussian noise



Uniform noise

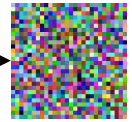


All white /black



Image from other classes





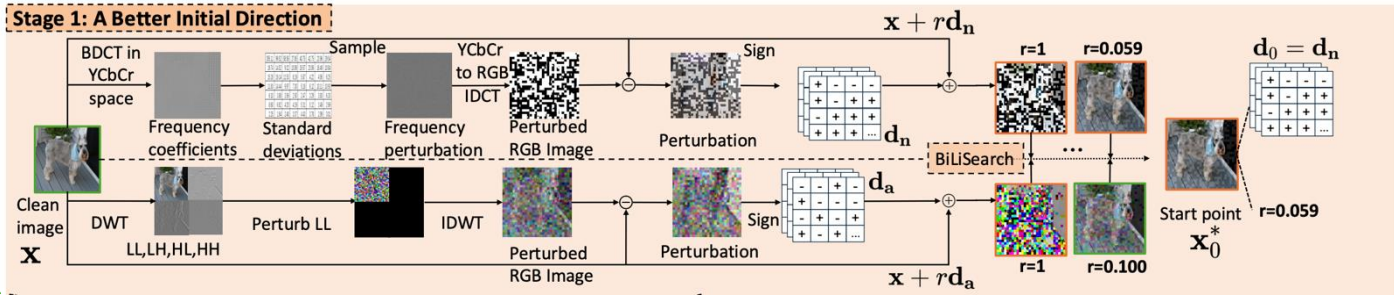
Ours frequency-based initialization

Binary search → Start point → Stage 2: Perturbation Direction Search

Our start point:
- Smaller decision boundary distance
- Higher curvature

Table 2: Comparison of curvature (κ) and perturbation distances (ℓ_∞, ℓ_2) across initialization methods. "Image" denotes a random sample from a different class.

Metrics	Curvature κ↑		ℓ_∞↓		ℓ_2↓	
Methods	Avg.	Med.	Avg.	Med.	Avg.	Med.
Gaussian	58,736	49,404	0.363	0.341	49.4	46.7
Uniform	58,273	48,932	0.315	0.318	49.7	48.6
Image	57,873	47,068	0.362	0.364	55.2	54.2
RayS	58,167	45,454	0.518	0.533	147.8	150.1
ADBA	39,408	33,592	0.354	0.339	114.9	121.2
Our $\mathbf{d_r}$	58,308	47,349	0.091	0.083	35.1	32.1
Our $\mathbf{d_b}$	60,934	50,192	0.108	0.072	39.3	27.2
Our $\mathbf{d_n}$	**68,502**	**54,743**	**0.078**	**0.069**	**28.9**	**26.1**

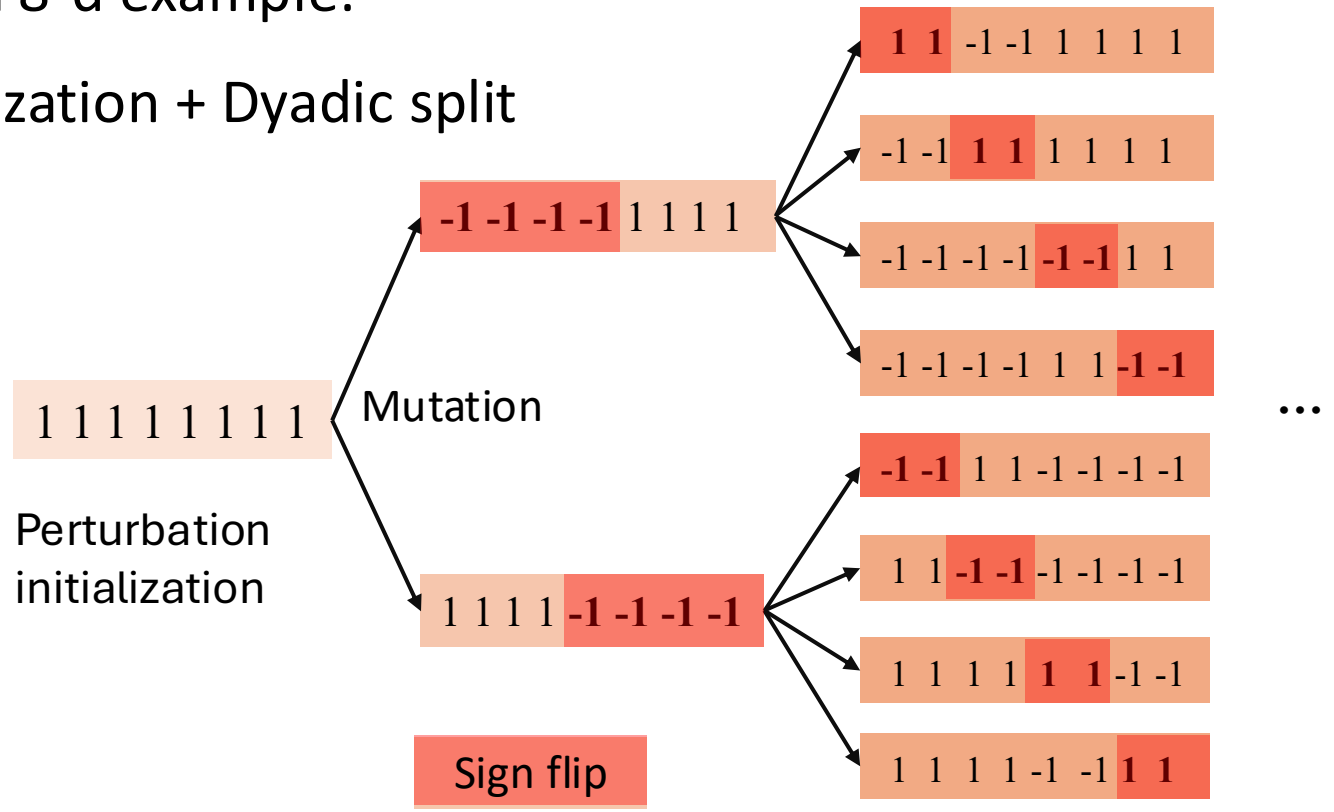


Method: Stage 2. More efficient gradient sign estimation

- Perturbation direction mutation of current L_∞untargeted attack SOTA ADBA[1]:

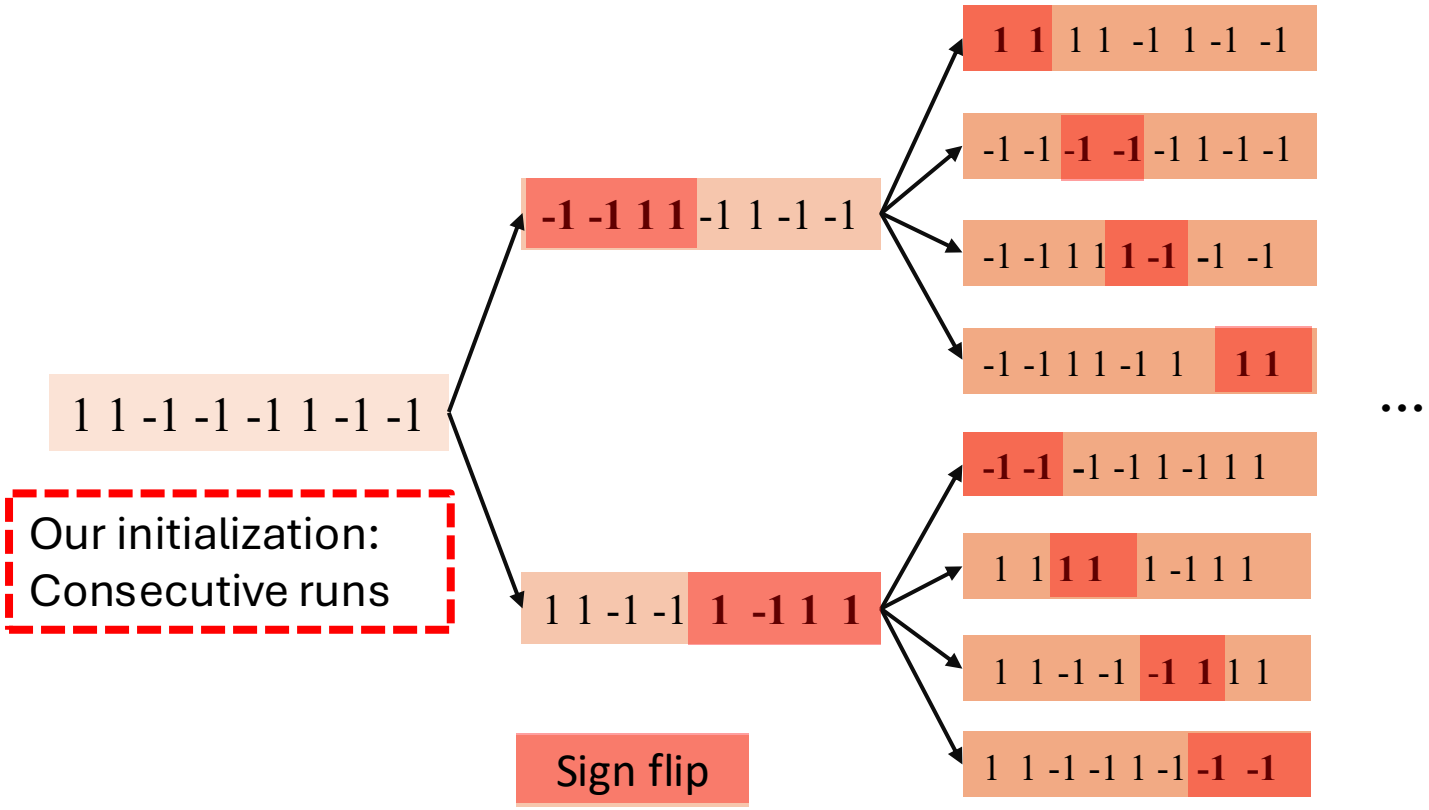
  A flattened 8-d example:

  All 1 initialization + Dyadic split



[1] F. Wang, et al. ADBA: Approximation Decision Boundary Approach for Black-Box Adversarial Attacks. AAAI 2025.

Method: Stage 2. More efficient gradient sign estimation
Our initialization + ADBA dyadic split



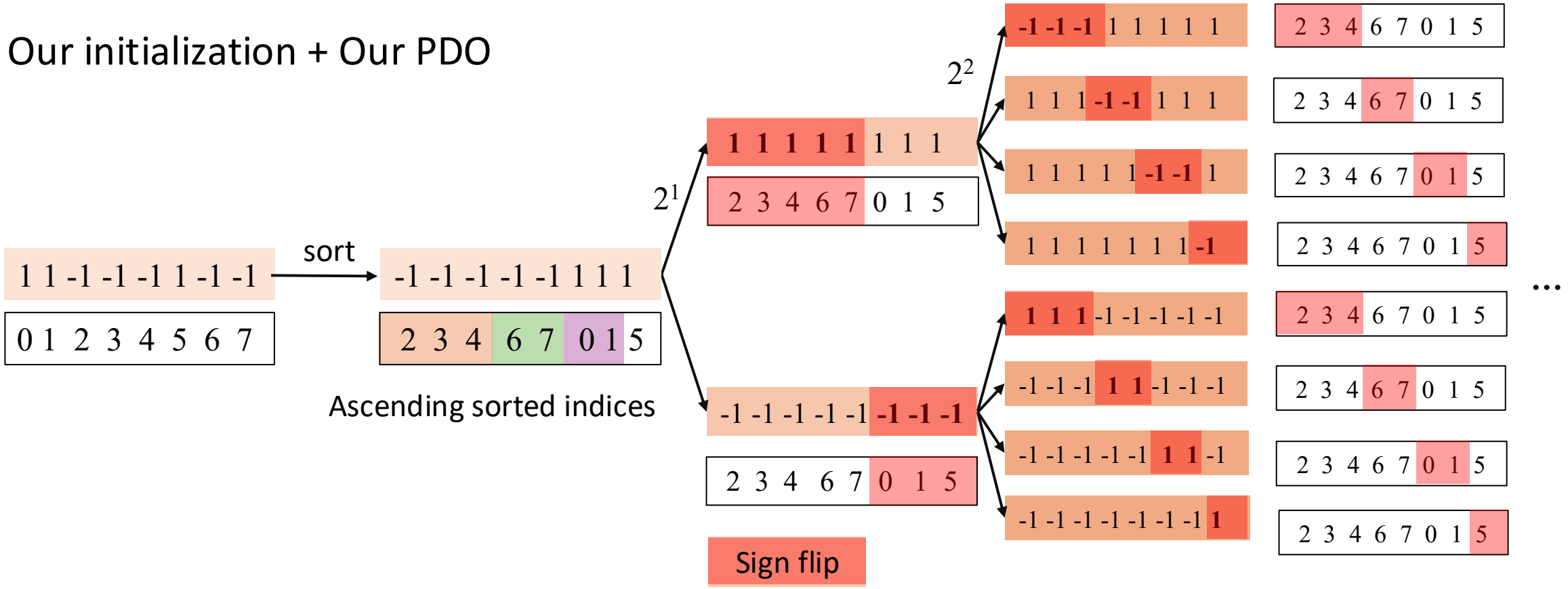
Weakness:

- Fragments the initial consecutive runs
- conflicting gradient signals within a partition offset each other, resulting in query inefficiency

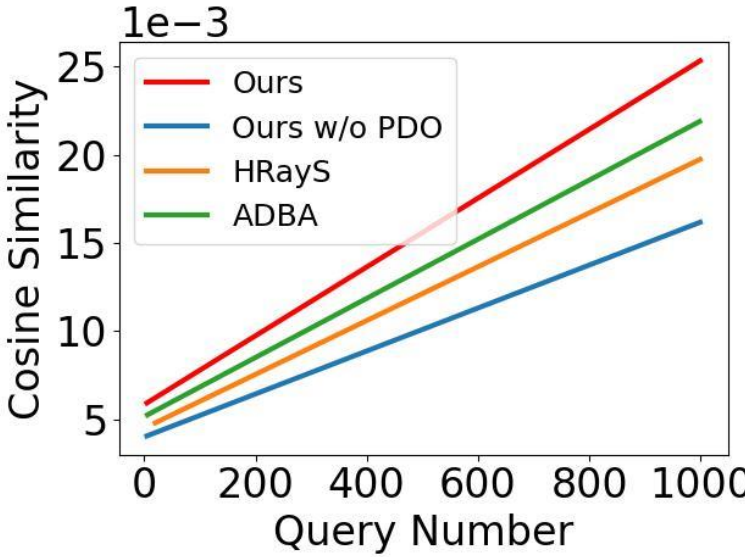
[1] F. Wang, et al. ADBA:Approximation Decision Boundary Approach for Black-Box Adversarial Attacks. AAAI 25

Method: Stage 2. More efficient gradient sign estimation
Our Pattern-driven Direction Optimization (PDO)

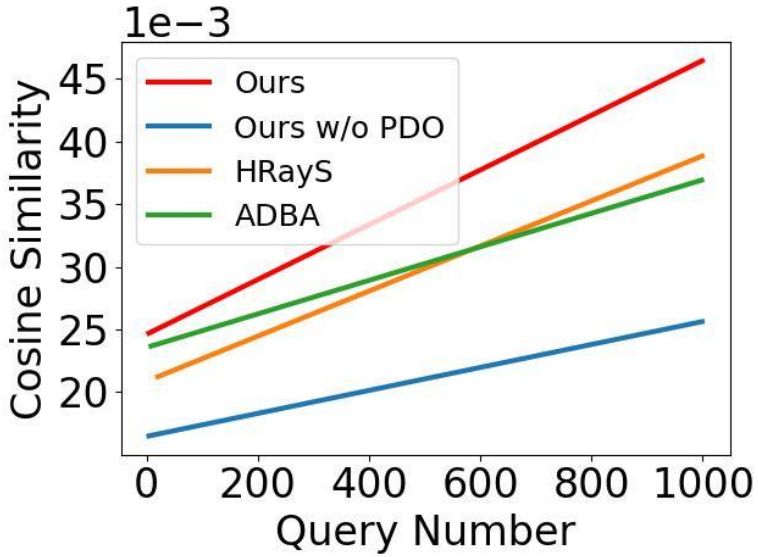
- ✓ PDO treats each sign-consistent run as an atomic search unit.
- ✓ Keep the superior alignment with the gradient sign of the initial runs.



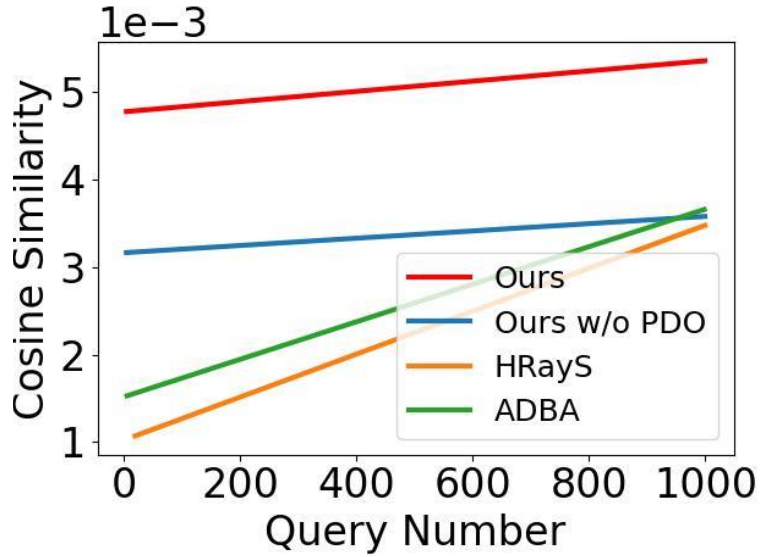
Method: Stage 2. More efficient gradient sign estimation PDO vs Dyadic split of ADBA



(a) ViT-B-32



(b) Robust WideResNet-50



(c) ResNet-50

- The "**Ours**" curve consistently upper-bounds the baseline "**Ours w/o PDO**" across architectures, demonstrating that our PDO generally yields **a better gradient descent direction** than dyadic split of ADBA.

Experimental results: Ablation

Table 9: The ablation study on the initialization methods and PDO module. Abbreviation: WideResNet-50 (WRS-50); all one vector (**1**); dyadic search used in ADBA (ADBAS), see Step 2 of ADBA in Sec. 2.3. Max.Q: 500.

Model	Search	Gauss.	Unif.	Img.	**1**	$\mathbf{d_b}$	$\mathbf{d_r}$	$\mathbf{d_n}$	DDM
ViT	ADBAS	52.5	62.5	76.5	75.5	63.5	75.5	75.5	77.0
	PDO	70.5	72.0	76.0	75.5	82.5	81.5	**85.5**	**85.5**
WRS-50	ADBAS	16.5	18.0	36.5	45.0	27.5	35.0	28.5	37.0
	PDO	42.0	43.5	43.0	45.0	44.0	48.0	47.0	**50.5**

1. Our **initialization** significantly outperforms traditional initializations.
2. **PDO** consistently enhances ASR across virtually all initialization methods.

Experimental results on Cifar-10 and ImageNet

Table 11: The untargeted attack performance on CIFAR-10 (VGG-16-BN) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
5	0.0	0.0	0.5	0.0	1.5	0.0	**4.0**	0.0	-	-	5	-	5	-	5	-	-	-	5	-	5	-	5	-
10	0.0	0.0	2.5	0.0	1.5	0.0	**14.5**	2.5	-	-	7	-	10	-	6	9	-	-	7	-	10	-	6	6
20	2.5	2.5	2.5	0.0	3.5	2.0	14.5	**18.5**	14	14	7	-	19	17	6	14	14	14	7	-	20	17	6	14
50	24.5	24.5	14.0	0.0	8.0	13.0	**30.0**	**30.0**	44	44	33	-	39	35	25	24	48	48	40	-	47	37	24	18
80	32.0	33.0	25.0	4.0	11.0	24.5	**45.5**	41.0	45	47	49	71	49	47	39	35	49	50	41	71	49	49	37	29
100	32.5	37.5	22.5	5.5	18.0	31.5	49.5	**50.5**	45	47	44	73	69	57	43	45	49	50	41		68	55	40	39
150	41.5	40.5	16.0	8.5	26.5	51.0	62.0	**64.0**	82	81	64	91	93	83	61	61	89	89	79	72	98	85	52	56
200	44.0	41.5	27.0	9.0	37.0	66.0	**76.0**	73.0	81	84	115	106	121	104	81	74	89	89	82	107	122	104	66	65
250	45.0	45.0	27.0	15.5	45.5	75.0	80.0	**80.5**	92	85	99	147	143	118	88	88	89	90	80	141	150	117	69	74
300	45.5	45.0	28.0	21.0	57.0	79.0	82.5	**85.5**	89	85	106	185	173	125	94	99	90	90	80	214	179	120	72	82
400	42.0	42.0	29.0	32.5	71.5	86.0	87.5	**90.5**	87	84	120	259	212	144	109	112	89	89	82	298	209	127	79	87
500	44.0	43.0	29.5	40.0	76.5	91.0	**92.5**	91.5	83	86	115	293	227	161	127	116	89	90	81	299	217	132	83	87
1k	47.0	49.5	35.0	64.0	91.0	98.0	**98.5**	**98.5**	137	133	239	491	305	205	163	161	140	139	133	487	265	147	96	92
2k	48.5	48.0	37.0	81.5	98.0	99.5	**100.0**	99.0	201	130	320	657	387	220	178	166	140	139	132	591	288	147	107	94
3k	52.0	47.5	40.0	92.5	99.5	**100.0**	**100.0**	**100.0**	180	136	346	873	393	229	171	188	139	139	134	698	290	147	104	98

Table 3: The untargeted attack performance on ImageNet (ResNet-50) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
5	0.0	0.0	3.5	0.0	2.0	0.0	**20.5**	0.0	0.0	-	5	-	5	-	5	-	-	-	5	-	5	-	5	-
10	0.0	0.0	9.5	0.0	2.0	0.0	**37.0**	-	-	-	6		10	-	6	-	-	-	6	-	10	-	5	-
20	6.5	6.5	9.5	0.0	2.0	3.5	37.0	**39.5**	14	14	6	-	20	17	6	19	14	14	6	-	20	17	5	19
50	6.0	6.0	17.0	0.0	16.5	22.5	51.5	**55.0**	13	13	20	-	42	32	19	23	13	13	6	-	40	31	13	20
80	14.0	14.0	17.5	6.5	21.0	38.0	57.0	**64.0**	40	40	21	75	52	46	24	29	59	59	6	75	56	42	15	20
100	15.0	16.0	16.0	7.5	25.0	49.5	62.0	**66.0**	40	37	23	77	60	56	29	31	59	38	6	75	57	59	15	20
200	17.5	17.5	22.0	10.5	51.0	71.0	**75.5**	75.0	75	64	72	87	109	87	49	46	99	99	48	75	104	80	26	22
300	20.0	18.5	21.0	14.0	54.0	79.5	**83.5**	**83.5**	77	71	54	126	136	104	67	66	99	99	6	82	121	84	34	22
400	16.5	16.0	23.5	16.0	69.0	85.0	88.5	**89.0**	61	60	72	158	160	117	83	85	99	99	7	144	143	85	38	24
500	20.5	18.0	25.5	19.0	70.5	87.5	90.0	**92.0**	92	81	85	184	198	128	88	97	99	99	88	146	174	88	38	29
1k	20.5	18.5	22.0	32.5	89.5	93.5	95.5	**97.5**	164	110	184	414	312	163	127	131	150	149	8	404	226	97	44	36
3k	23.0	19.5	23.0	59.5	97.5	99.0	99.5	**100.0**	352	209	196	1,117	431	236	199	180	150	149	139	965	252	100	48	39
5k	21.0	19.0	22.5	67.5	**100.0**	**100.0**	**100.0**	**100.0**	210	140	319	1,441	522	249	222	180	149	149	138	978	263	102	49	39

✓ Stronger attack capability

(ASR)

✓ Higher attack efficiency

(Avg.Q & Med.Q)

Compared methods:

ADBA: AAAI 2025

BouR&BouT: IEEE S&P 2024

HSJA: IEEE S&P 2020

RayS: SIGKDD 2020

Experimental results on APIs and CLIP

- Real-world APIs

Table 5: The untargeted attack performance on real-world APIs under different perturbation constraints. Abbreviations: ASR: A., Med.Q: M.Q, Avg.Q: A.Q.

ℓ_∞	APIs	Imagga			Google			Tencent			Baidu		
	Methods	A.	M.Q	A.Q	A.	M.Q	A.Q	A.	M.Q	A.Q	A.	M.Q	A.Q
0.05	ADBA	83	61	105	53	26	28	33	25	26	65	25	26
	Ours$_{opt}$	90	**6**	38	91	22	23	67	**10**	**15**	**88**	**9**	**13**
	Ours$_{dyn}$	**100**	23	**20**	**100**	**21**	**19**	**75**	23	20	**90**	21	20
0.1	ADBA	99	41	70	68	24	26	56	33	35	80	24	27
	Ours$_{opt}$	**100**	**4**	**8**	94	7	14	**100**	**14**	**14**	98	**7**	**14**
	Ours$_{dyn}$	**100**	16	16	**100**	**15**	**16**	97	17	16	**99**	16	16

- CLIP model

Table 4: The untargeted attack performance on ObjectNet (ViT-H-14-CLIPA-336) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	Ours$_{opt}$	Ours$_{dyn}$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	Ours$_{opt}$	Ours$_{dyn}$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	Ours$_{opt}$	Ours$_{dyn}$
10	0.0	0.0	10.0	0.0	0.0	0.0	**21.5**	0.0	-	-	6	-	-	-	5		-	-	5	-	-	-	5	
50	5.5	5.5	15.0	0.0	5.0	7.5	37.0	**39.5**	13	13	20	-	41	41	16	22	12	12	6	-	39	42	10	20
100	20.5	18.0	22.5	9.5	13.0	20.5	44.5	**50.0**	40	42	35	77	72	61	25	33	61	58	6	76	81	58	15	20
500	23.5	23.5	23.0	22.0	38.0	49.5	65.5	**67.5**	95	91	72	207	178	156	98	92	102	98	88	190	157	118	34	28
1k	25.0	23.5	25.5	32.5	51.0	61.0	77.0	**78.5**	136	129	137	363	326	257	194	186	151	149	6	313	220	162	55	49
3k	26.5	22.5	26.5	58.5	73.0	82.5	**88.5**	**88.5**	242	141	202	1,105	770	636	371	364	152	149	137	850	430	290	86	65
6k	27.0	25.0	29.5	73.0	82.0	88.0	**92.5**	**92.5**	637	130	197	1,692	1,171	872	536	523	152	149	8	1,356	565	313	98	82
15k	33.0	24.5	27.5	83.0	91.5	95.5	**98.5**	98.0	1,123	412	137	2,437	2,135	1,527	1,039	1,022	154	148	7	1,650	853	362	124	92
30k	37.5	22.5	28.5	89.5	96.5	99.0	99.0	**99.5**	3,880	107	203	3,498	3,024	2,244	1,366	1,399	155	148	138	1,952	940	509	171	99
40k	40.0	24.5	30.0	89.5	95.5	99.5	99.5	**100.0**	6,552	114	225	4,460	1,211	2,426	1,367	1,555	153	149	8	2,125	651	518	141	103

Experimental Results on Defensive Methods

1、Performance on **adversarially trained models**

Table 6: The untargeted attack performance on adversarially trained WideResNet. Max.Q: 1,000.

Dataset	Metrics	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
CIFAR-10	ASR	4.5	3.5	3.5	1.5	20.5	20.0	**22.5**	21.0
	Avg.Q	140	141	58	361	290	211	234	205
	Med.Q	140	140	8	398	222	101	143	83
ImageNet	ASR	10.5	10.5	11.0	3.5	44.5	45.0	**50.5**	48.5
	Avg.Q	112	111	84	302	189	143	143	191
	Med.Q	149	146	6	223	91	58	73	90

2、Performance on **stateful defensive model**, called Blacklight[2].

Table 7: The untargeted attack performance on Blacklight against ViT-B-32. Max.Q: 100. Outside/inside parentheses: DPAttack with/without Gaussian-noise injection.

Metrics	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
ASR↑	0	0	5	0	0	0	41 (22)	**43** (24.5)
Avg.Q↓	-	-	5	-	-	-	**39** (11)	40 (17)
Med.Q↓	-	-	5	-	-	-	31 (10)	**26** (17)
Det$R$↓	100	100	95	100	100	100	**0** (78)	**0** (76.5)
1stDQ↑	9	9	8	11	5	2	- (15.7)	- (20.3)

- The highest ASR and comparable query efficiency.
- For real-time defense:
  - Ours variant: **a detection rate of 0**[1]**.**
  - Variant: adding random Gaussian noise to the perturbation direction.
- Other methods are mostly detected.

1. For Ours$_{opt}$, DetR may occasionally be 1% due to the randomness of Gaussian noise during reproduction.

[2] Li H, et al. Blacklight: Scalable defense for neural networks against query-based black-box attacks. USENIX Security. 2022: 2117-2134.

Takeaway

1. Hard-label attacks are still dangerously effective under strict query budgets.
2. Initialization matters as much as search.
3. Our DPAttack brings theoretical grounding and practical effectiveness.
4. DPAttack also performs well on dense prediction task and in the L_2 setting.

Thank you very much for your attention!

Contact me

Jun Liu
Postdoc researcher
Echizen Lab, National Institute of Informatics (Japan)
csjunliu@nii.ac.jp
https://csjunjun.github.io/