

# Supplementary materials of “TransferBench: Benchmarking Ensemble-based Black-box Transfer Attacks”

## A Methods involved in the benchmark

For this benchmark we considered the works described in Table 4. DSA method has not been directly compared since features a near-zero performance on the targeted scenarios. All the attacks has been tested on the scenarios of the original papers, achieving the same performance. Original scenarios can be found in the `transferbench/config/scenarios` path of the benchmark.

Table 4: ASR and average-queries-per-success claimed in the original papers.

Attack	Venue	$m$	HeS	HoS+R	Targeted	$p$	$\varepsilon$	ASR [%]	$\bar{q}$
SubSpace [15]	NeurIPS 2019	3	✓	✗	✗	$\infty$	13/255	98.9%	462
SimbaODS [32]	NeurIPS 2020	4	✗	✗	✓	$\infty$	13/255	92.0%	985
GFCS [23]	ICLR 2022	4	✗	✗	✓	2	$\sqrt{0.001d}^1$	60.0%	20
BASES [8]	NeurIPS 2022	20	✗	✗	✓	$\infty$	16/255	99.7%	1.8
GAA [37]	PR 2024	4	✗	✗	✓	$\infty$	16/255	46.0%	3.9
DSA [29]	Usenix 2024	3	✓	✓	✗	$\infty$	16/255	96.9%	136
DSWEA [16]	PR 2025	10	✗	✗	✓	$\infty$	16/255	96.6%	2.7

<sup>1</sup>Images included in the experiments have  $d = 3 \cdot 299 \cdot 299$  pixels, from which  $\varepsilon \approx 16.37$

## B Instructions

The TransferBench codebase is accompanied by three main instructional resources:

- The primary [Readme.md](#) provides installation guidance and a quick-start tutorial for using the API with minimal setup.
- A companion [example notebook](#) offers in-depth, hands-on instructions, demonstrating how to use the framework with varying levels of customization.
- The [attacks\\_zoo/Readme.md](#) explains the implementation of the TransferAttack protocol within the `attacks_zoo` module.
- Instructions for setting up and using the `trbench` CLI command are detailed in the dedicated [benchmark\\_tools/Readme.md](#).

Further details and the complete codebase are available on the official GitHub repository: <https://github.com/pralab/transfer-bench>.

## C Licenses of external assets

The benchmark involved external assets for the models and query-free attacks.

**Robust models** The robust models Mim-Sw-L [36], Amini-Sw-L [2], Peng-RWRN-70 [27], Barto-WRN-94 [3] have been imported from RobustBench [12] released under MIT license, except for Pub-RN-50 [30], which has been taken from its original repository, released under Apache 2.0 license.

**Black-box attacks** Query-free black box attacks involved for comparison have been imported from TransferAttack [14] under the MIT license.

## D Additional plots

We include in this section further plots not displayed in the main paper. Figure 4 include the success vs average-queries-per-success curves for the ImageNet dataset, while the same curves relative to the

833 CIFAR-10 dataset are visualized in Figure 5. Figure 6 shows aggregated success rates of the various  
834 attacks for the CIFAR-10 dataset. The empty plots are due to the fact that when the attack reaches  
zero success-rate the average-queries-per-success is not defined and curves can not be displayed.

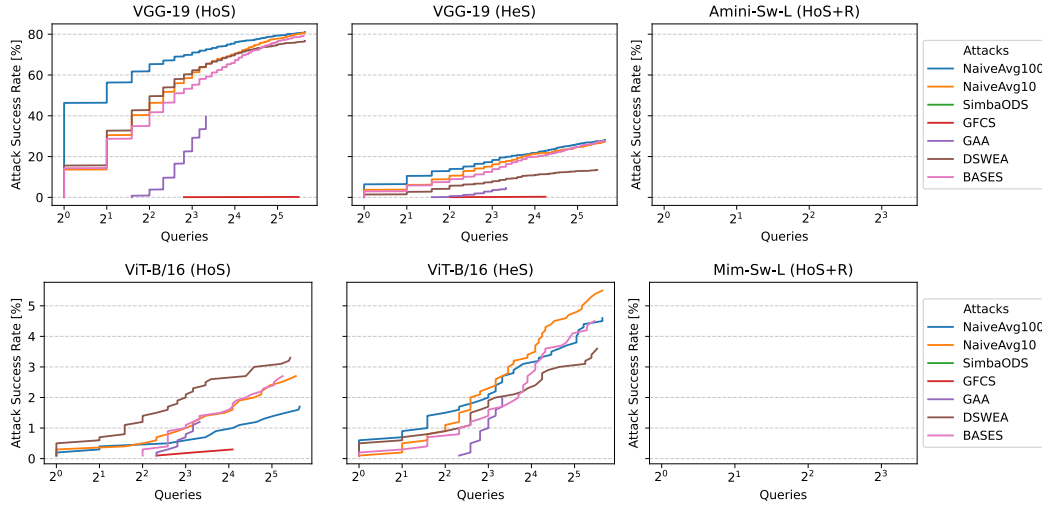


Figure 4: ASR-vs-Query curves on the ImageNet dataset.

835

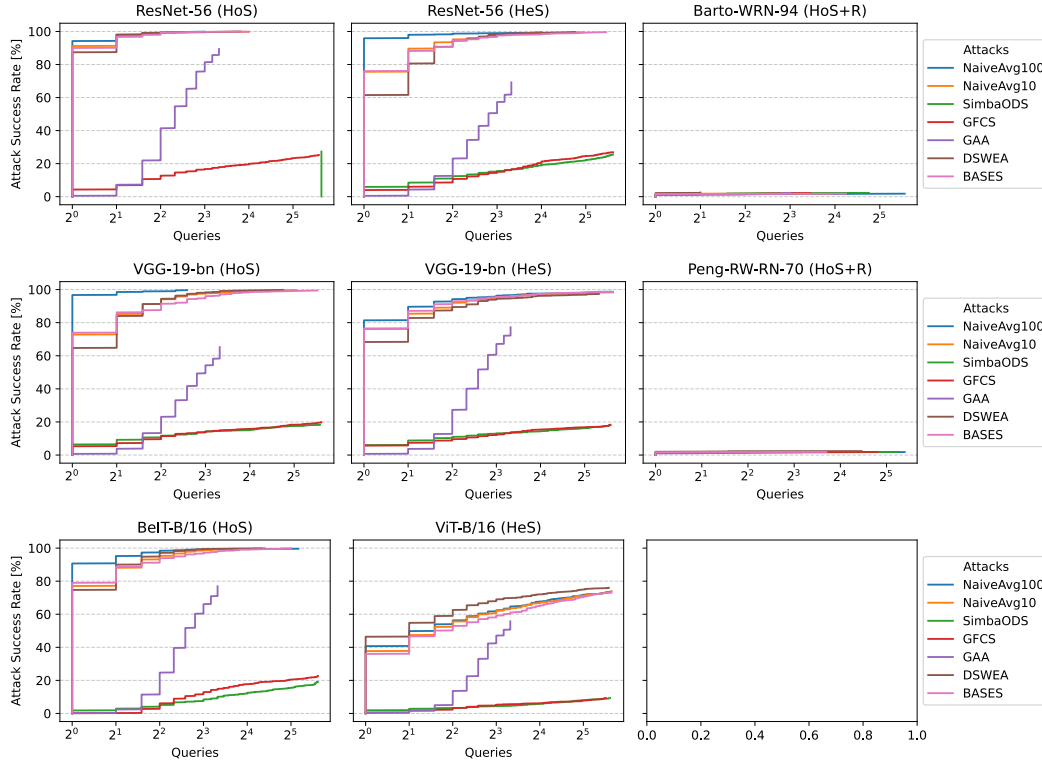


Figure 5: ASR-vs-Query curves on the ImageNet dataset for different victims.

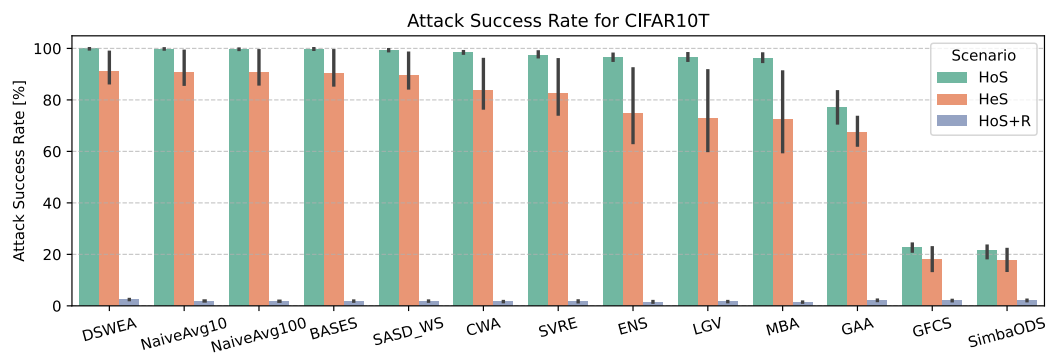


Figure 6: Aggregated attack success rate on the CIFAR10 dataset. Various attacks have a perfect success rate