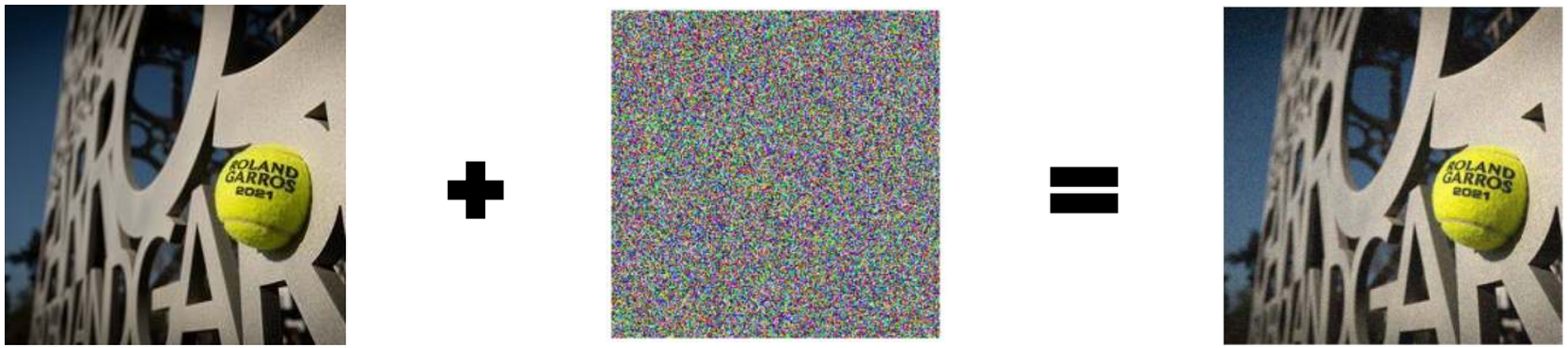
SurFree: a fast surrogate-free black-box attack

Thibault Maho, Teddy Furon, Erwan Le Merrer

CVPR 2021

Introduction

- Objective: Forge an adversarial:

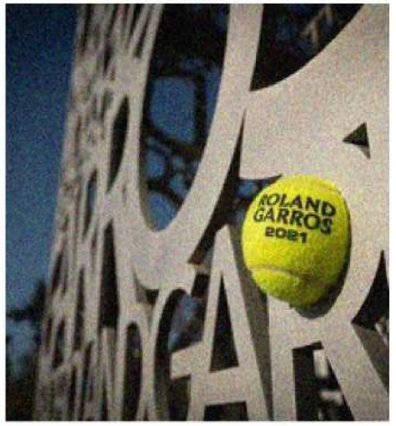


I_o, y_o Perturbation I_a, $y_a \neq y_o$

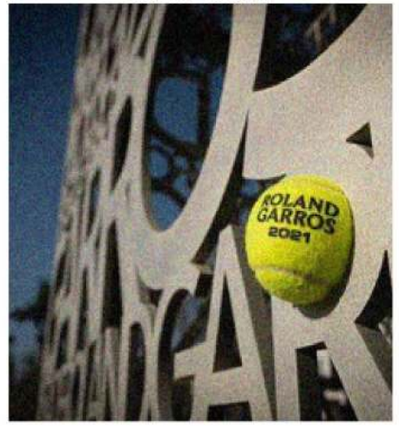
- Ideal Adversarial for a model *M*:

$$I_a^* = \arg \min_{M(I_a) \neq y_o} \|I_a - I_o\|_2$$

Distorsion : the lower ...



$L_2 = 17.8$

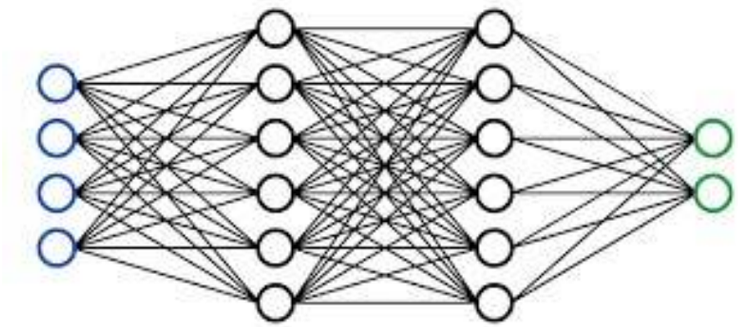


$L_2 = 12.4$

$L_2 = 7.6$

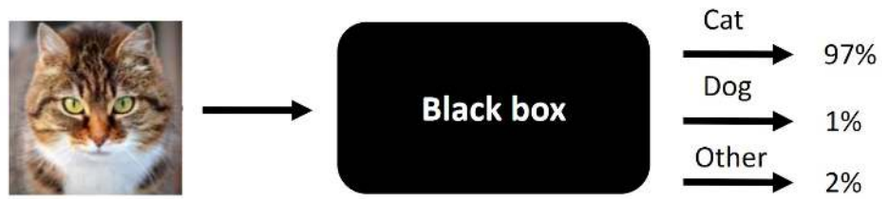
Introduction

White Box Attack



Total access to the model:
gradient, loss, ...

Black Box Attack



Limited access to the model:

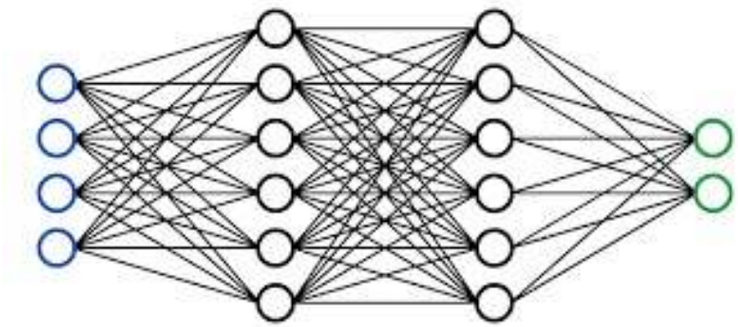
Score

OR

Top-1 label

Introduction

White Box Attack



Total access to the model:
gradient, loss, ...

Black Box Attack

Limited access to the model:

Score

OR

Top-1 label

Problem

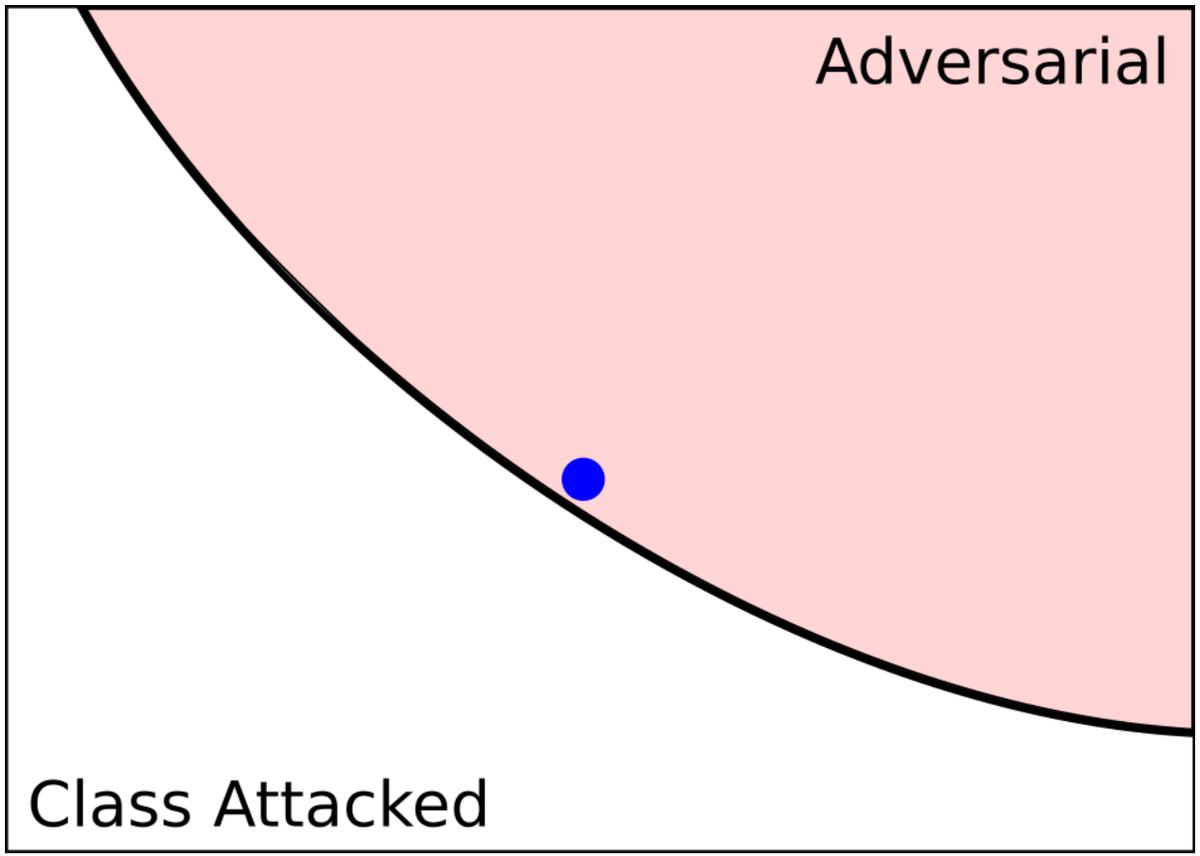
- SOTA decision-based black-box attacks use surrogates (copies):
  - Model Surrogates → expensive
  - Loss Surrogates
  - Gradient Surrogates

Problem

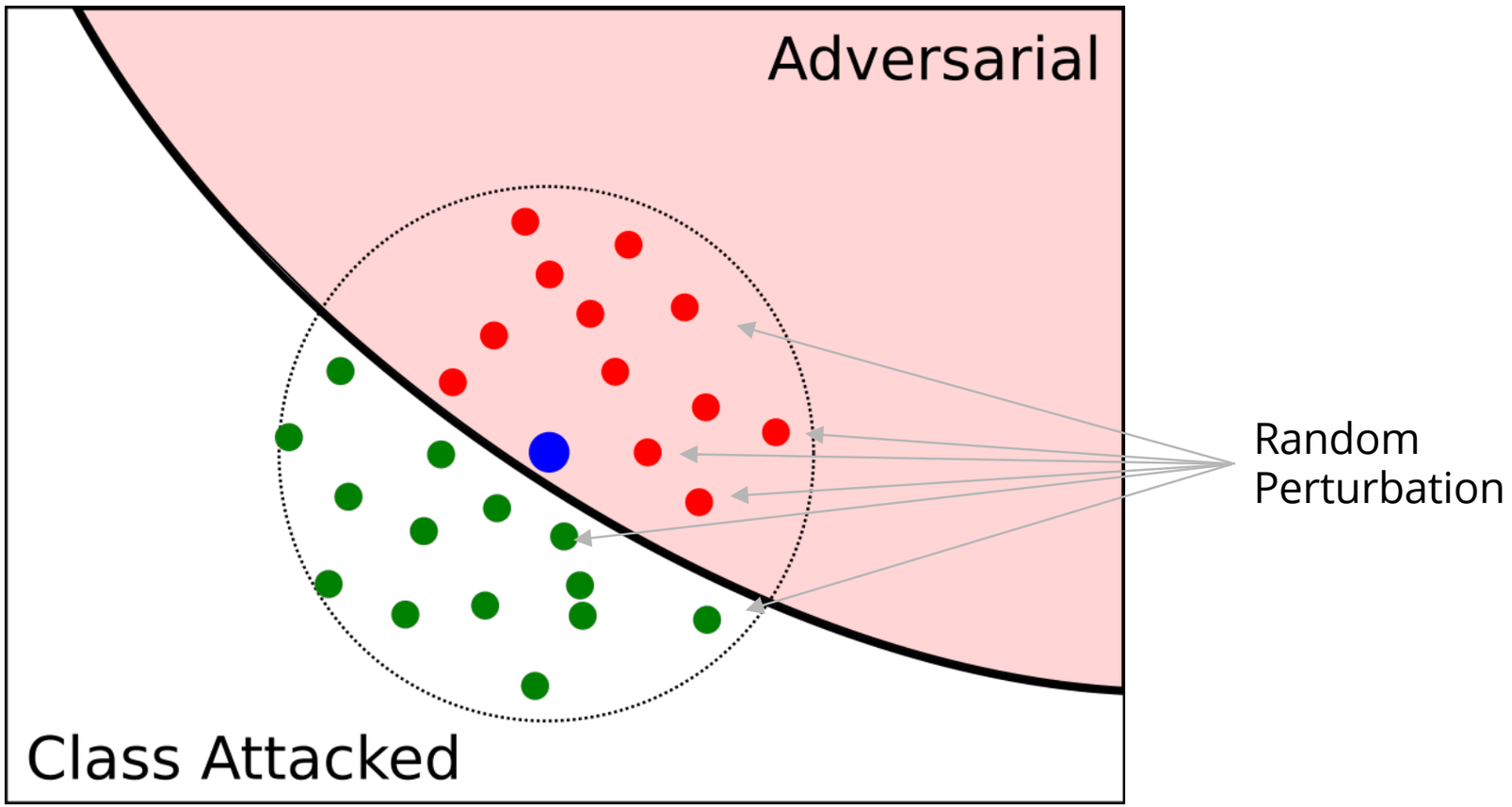
- SOTA decision-based black-box attacks use surrogates (copies):
  - Model Surrogates → expensive
  - Loss Surrogates
  - Gradient Surrogates

- Best Gradient Surrogates:
  - HopSkipJump
  - GeoDA
  - QEBA
- Main Difficulty: Attack with the lowest number of queries

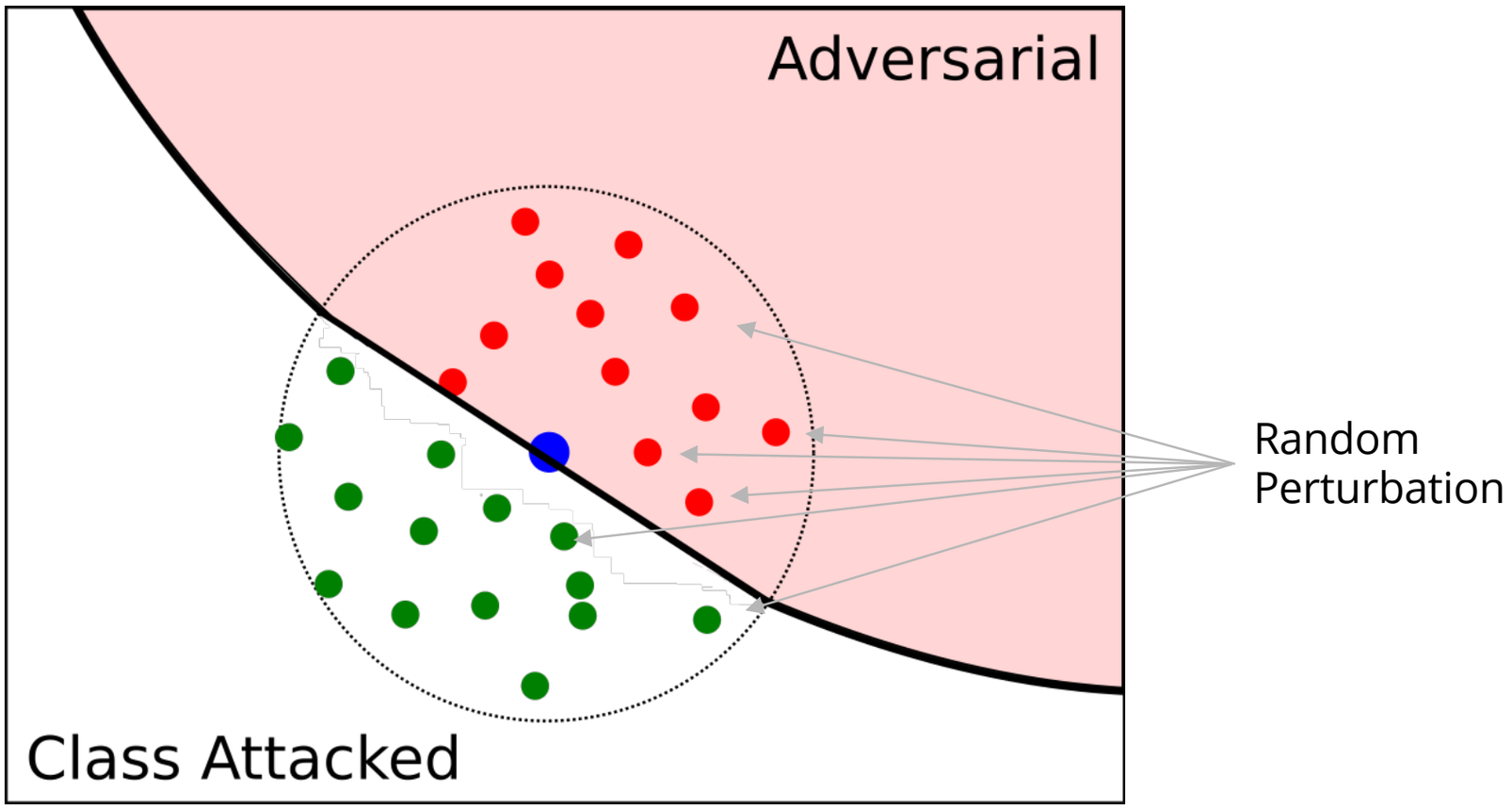
Problem - gradient surrogates



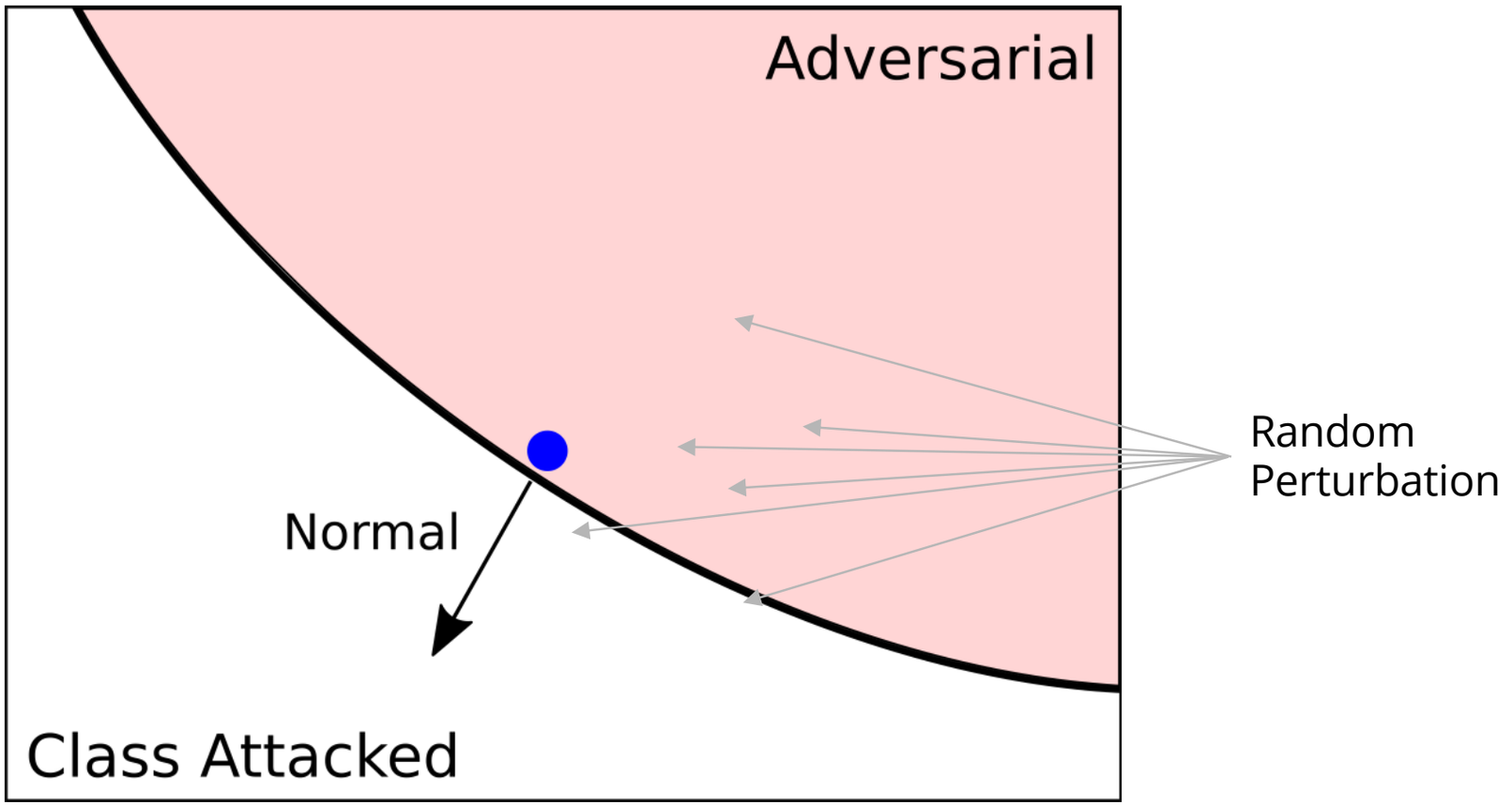
Problem - gradient surrogates



Problem - gradient surrogates



Problem - gradient surrogates



Problem

- “Useless” queries ?

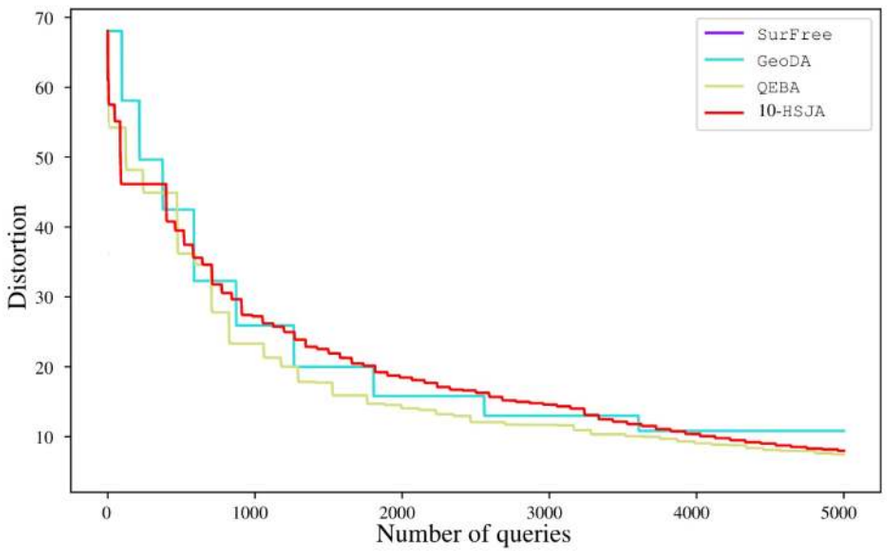


Figure: Monitoring distortion vs. number of queries for a single image

[1] Ali Rahmat et al, “Geoda: a geometric framework for black-box adversarial attacks”, CVPR 2020
[2] H. Li et al, “QEBA: Query-Efficient Boundary-based blackbox Attack”, CVPR 2020
[3] J. Chen et al, “HopSkipJumpAttack: A query-efficient decision-based attack”, IEEE Symp. on Security and Privacy 2020

Problem

- “Useless” queries ?

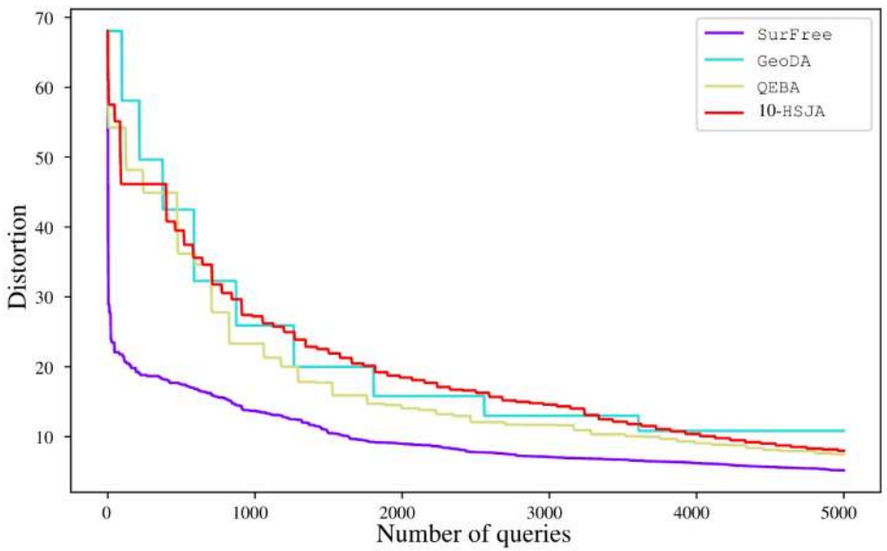
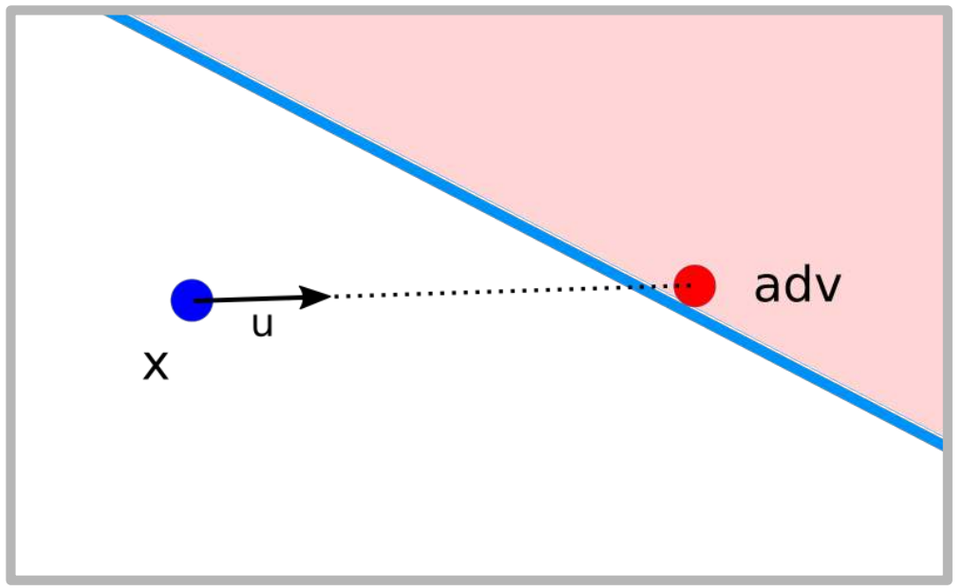


Figure 1: Monitoring distortion vs number of queries for a single image

We will build no surrogate to save queries = more trials

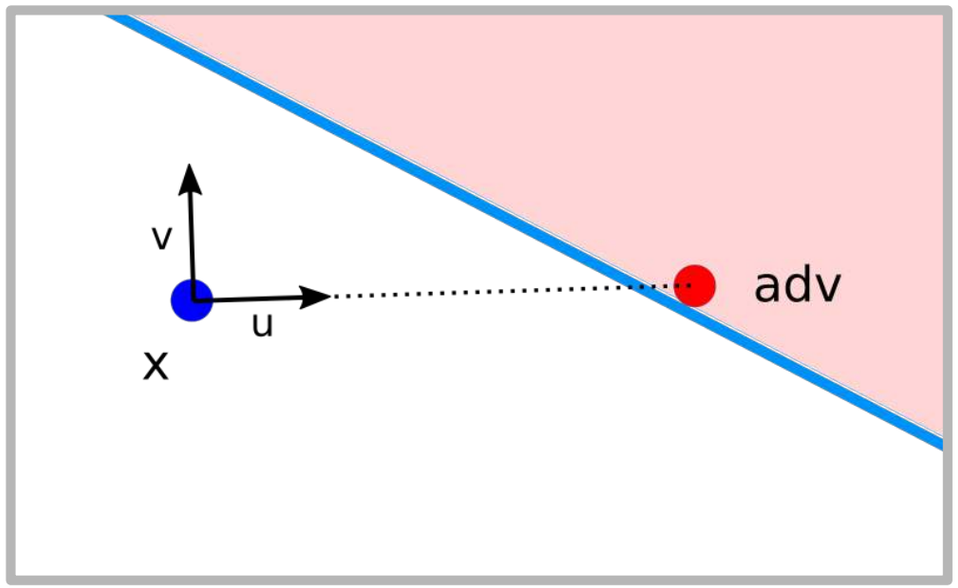
[1] Ali Rahmat et al, “Geoda: a geometric framework for black-box adversarial attacks”, CVPR 2020
[2] [illegible] “QEBA: Query-Efficient Boundary-based Black-box Attack”, CVPR 2020
[3] J. Chen et al, “HopSkipJumpAttack: A query-efficient decision-based attack”, IEEE Symp. on Security and Privacy 2020

Approach



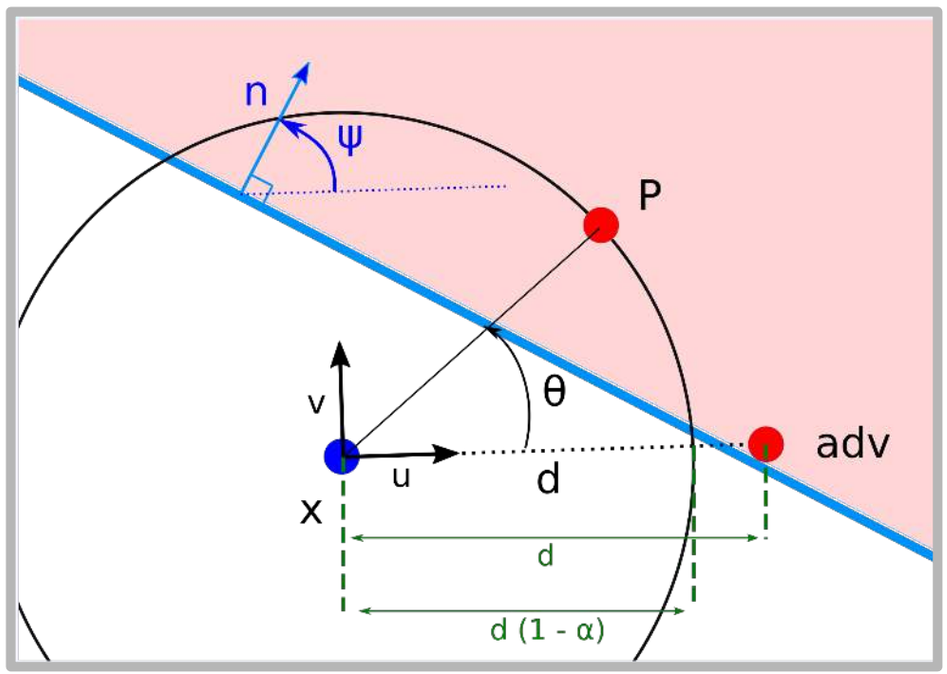
- Image attacked x
- Adversarial on the boundary (obtained by line search between x and a very noisy x)
- Direction u given by u and x

Approach



- Pick a random direction v orthogonal to u
- This iteration looks for a closer adversarial in (x, u, v)

Approach



- We search for an adversarial in polar coordinate. For a given point P, we have the following coordinates:

  .

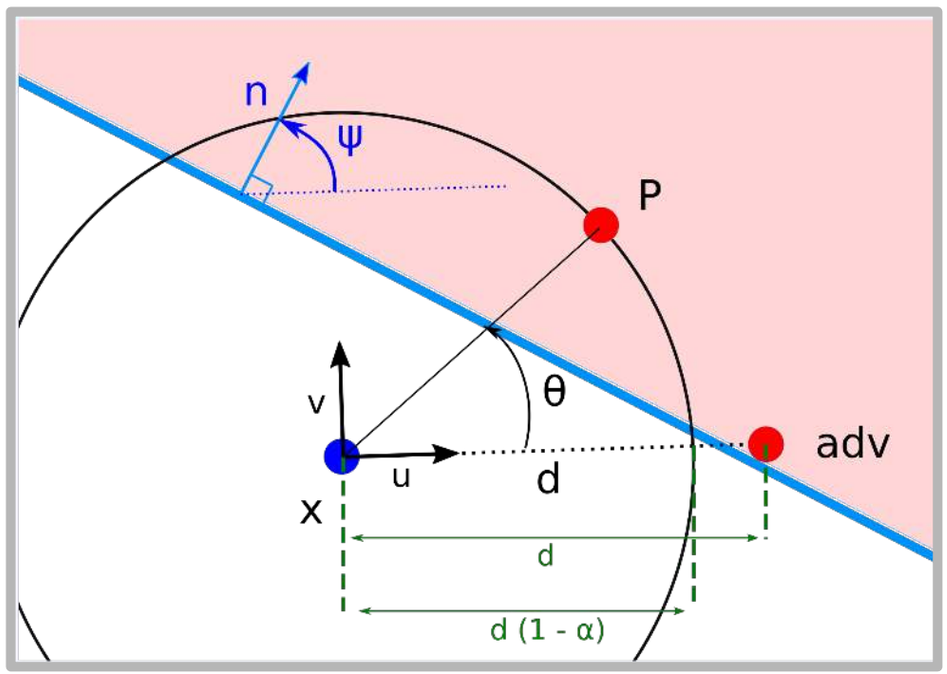
  - α controlled the gain of our adversarial

$$\mathbf{z}(\alpha, \theta) = d(1-\alpha)\left(\cos(\theta)\mathbf{u} + \sin(\theta)\mathbf{v}\right) + \mathbf{x}$$

- *n* is the normal of the hyperplan, unknown. In polar coordinates:

$$\mathbf{n} := \cos(\psi)\mathbf{u} + \sin(\psi)\mathbf{v}$$

Approach



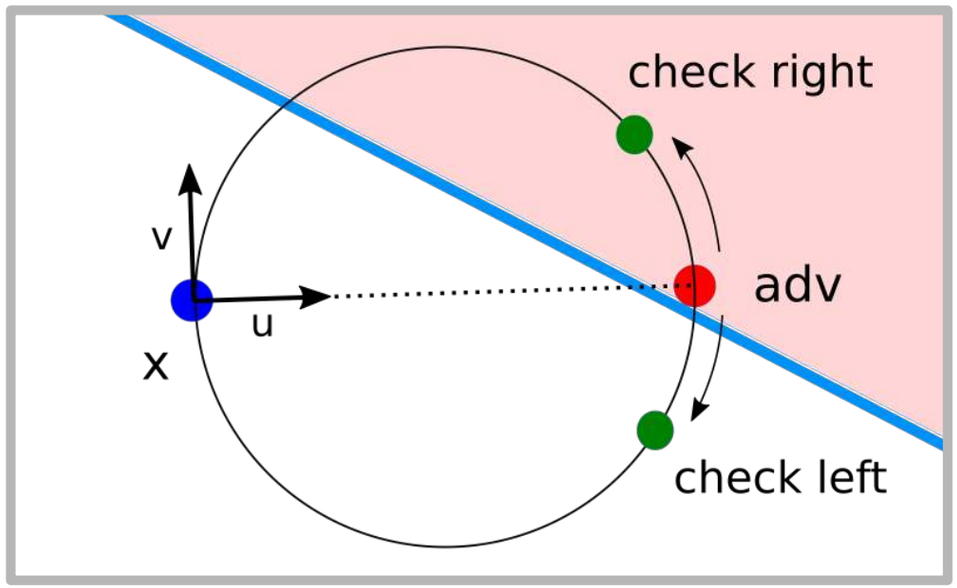
It gives, the point P is adversarial if:

$$g_\alpha(\theta) := \left| \frac{1 - (1-\alpha)\cos(\theta)}{(1-\alpha)\sin(\theta)} \right| \leq \tan(\psi)\mathrm{sign}(\theta)$$

By minimizing the left term, we have:

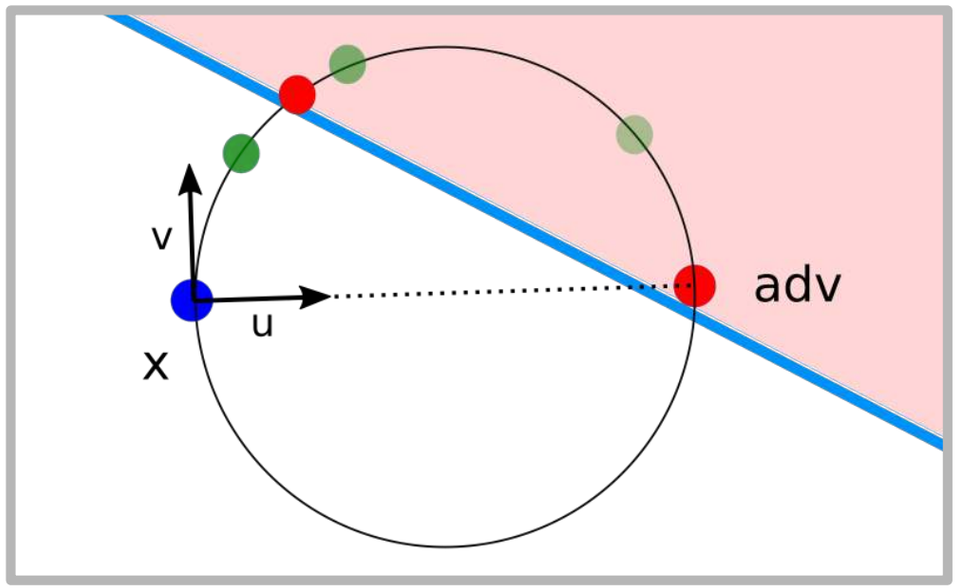
$$\theta = \pm \arccos(1 - \alpha)$$

Approach



- Thanks to the duality of theta and alpha, we have this circle.
- Find the direction by probing a small step to the left and to the right

Approach



- Line Search over the circle to find the intersection with the boundary

Basic Results

- Results compared between random directions and GeoDA and QEBA

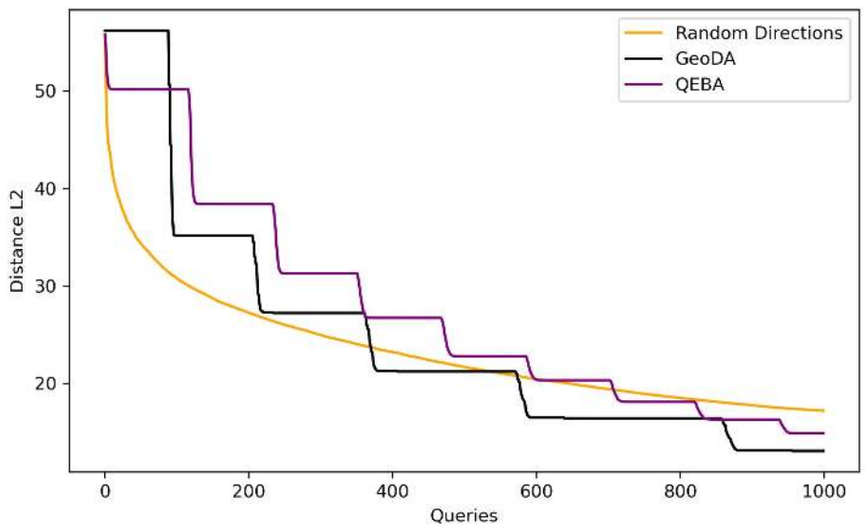
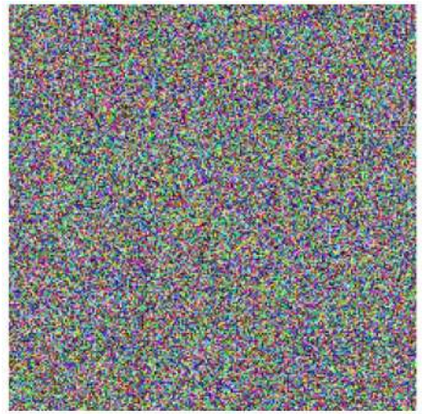


Figure: Basic Approach Benchmark. The amount of queries k (x-axis) w.r.t. mean distortion d(k) (y-axis).

Speed It Up

- Random Directions on Pixels

Image



Random
Direction

- 150 528 parameters
- At 1.000 queries:
  $L_2 = 17.20$

Speed It Up

- Frequency Domain / DCT

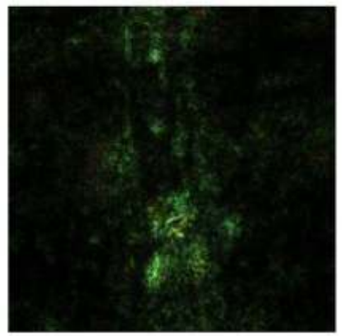
Speed It Up

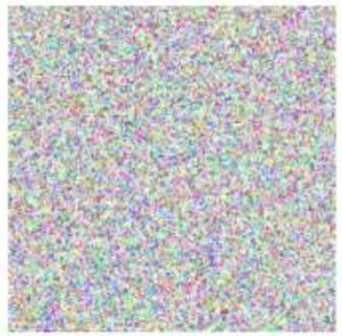
- Frequency Domain / DCT

- Parameters: -75%
- Almost the same image

Speed It Up







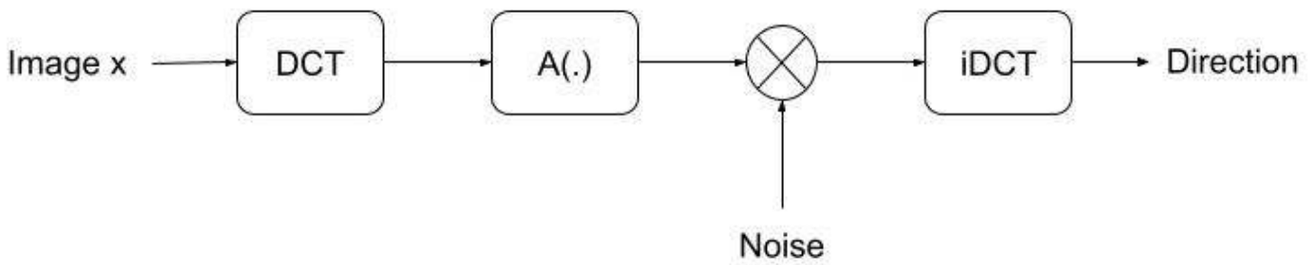
Speed It Up

Perturbation less perceptible if adapted to the image

Speed It Up

Perturbation less perceptible if adapted to the image

→ shape the power distribution of the perturbation as the one of the image



Speed It Up

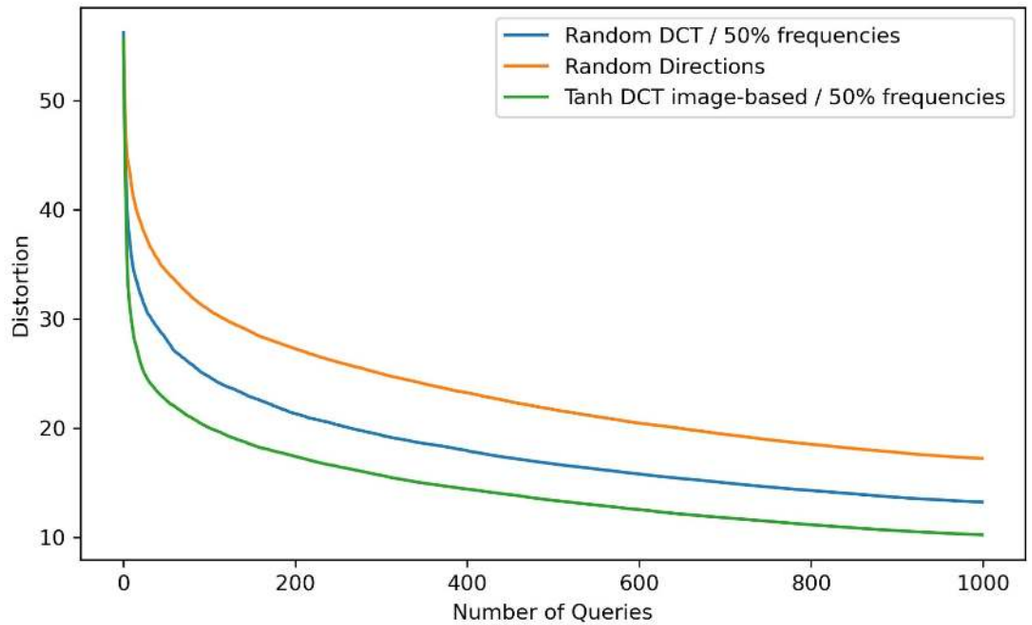


Figure: Dimension Reduction on directions. The amount of queries k (x-axis) w.r.t. mean distortion d(k) (y-axis).

Results

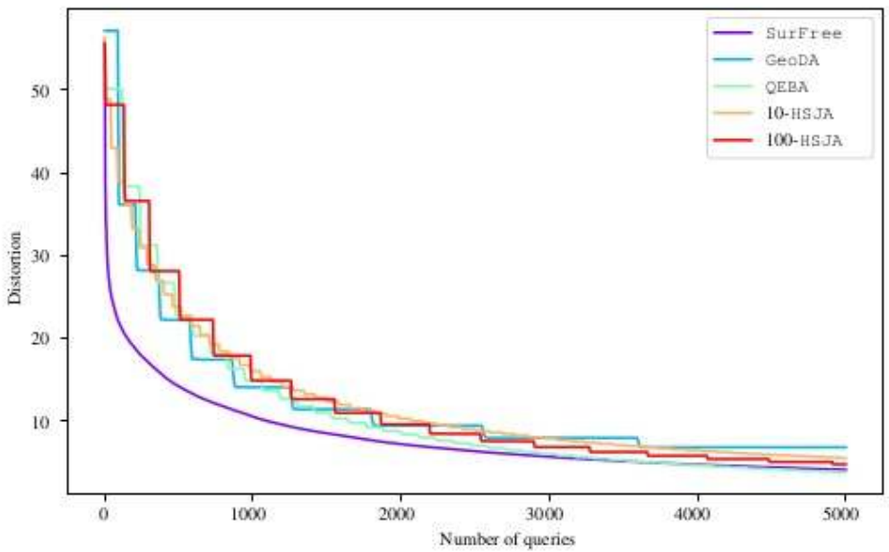


Figure: Benchmark on ImageNet. The amount of queries k (x-axis) w.r.t. mean distortion d(k) (y-axis).

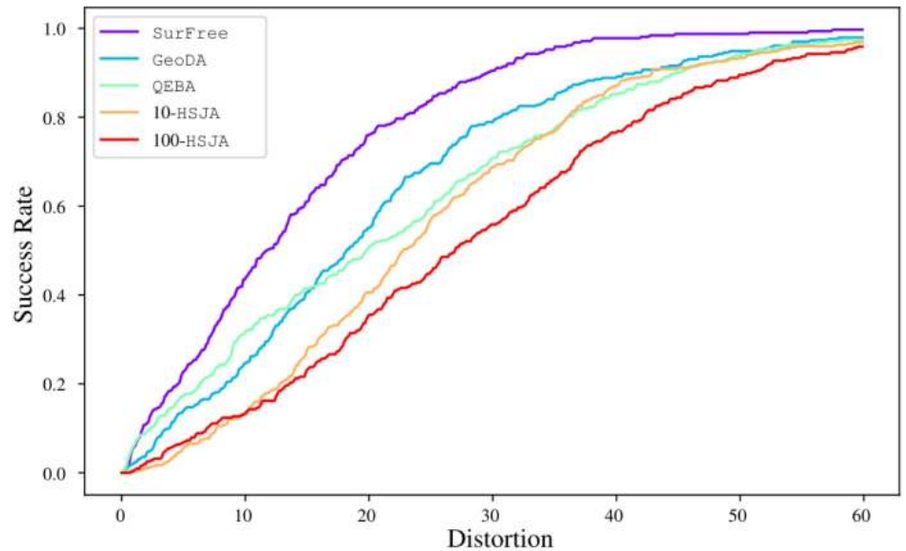


Figure: Global performances: accuracy vs. Euclidean diction in pixel domain

A picture's worth a thousand words

Original	SurFree	Geoda [1]	QEBA [2]
			
0	2.6	18.9	60.6
Chickadee	Amer. Dipper	Brambling	Stingray

Figure: Comparison of visual quality after 100 queries. Euclidean distortion in pixel domain

Thanks for your
Attention

Speed It Up

- Final perturbation is dependent of the x
- Example at 400 queries:

$L_2 = 17.8$

$L_2 = 12.4$

$L_2 = 7.6$