Simple Black-Box Adversarial Attack

ESHIKA KHANDELWAL 2020114018
AMEYA SHARMA 2020101059
ADITH JOHN RAJEEV 2020114010

PROBLEM STATEMENT

The problem we aim to address in our paper "Simple Black-box Adversarial Attack" is the generation of effective adversarial examples that can fool machine learning models without knowledge of their internal architecture or parameters.

SOLUTION

The paper proposes a gradient-free optimization approach called SimBA to generate adversarial examples. SimBA iteratively perturbs the input data by adding small noise until the model's output changes, and then uses a heuristic search to refine the perturbations.

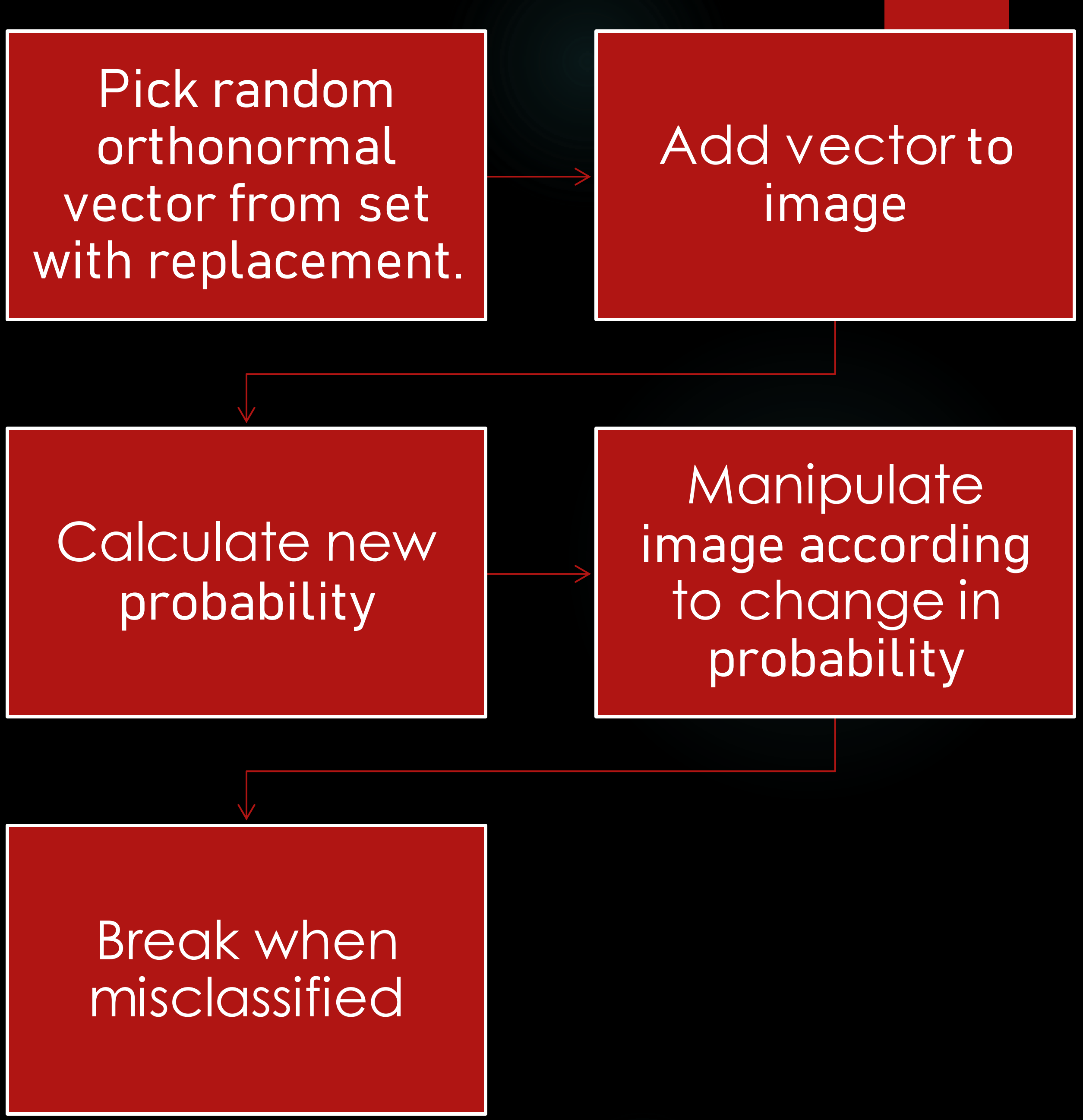
SimBA

- SimBA is a gradient-free optimization approach for generating adversarial examples that can fool machine learning models without knowledge of their internal architecture or parameters.
- SimBA works by iteratively perturbing the input data by adding small random noise until the model's output changes. Once a change in output is observed, SimBA uses a heuristic search to refine the perturbation and find the most effective one.
- Lowercase "e" refers to the perturbation or noise added to the input data by the SimBa attack algorithm.

The Approach

Algorithm 1 SimBA in Pseudocode

```
procedure SIMBA(x, y, Q, ε)
    δ = 0
    p = p_h(y | x)
    while p_y = max_{y'} p_{y'} do
        Pick randomly without replacement: q ∈ Q
        for α ∈ {ε, −ε} do
            p' = p_h(y | x + δ + αq)
            if p'_y < p_y then
                δ = δ + αq
                p = p'
                break
  return δ
```



Orthogonal Search Vectors (Q)

- Cartesian Basis
  - The standard basis Q = I
  - Increasing/Decreasing color of one pixel in every iteration
  - Corresponds to L0 attack
- Discrete Cosine Basis
  - Representing image in the frequency domain by breaking it down into a sum of cosine functions of varying frequencies and amplitudes. (DCT)
  - Adding noise in the frequency domain and then getting it back to the image space using IDCT.

Untargeted vs Targeted

Untargeted

- Aim to reduce the probability of original class
- Break when the highest probability is assigned to a different class

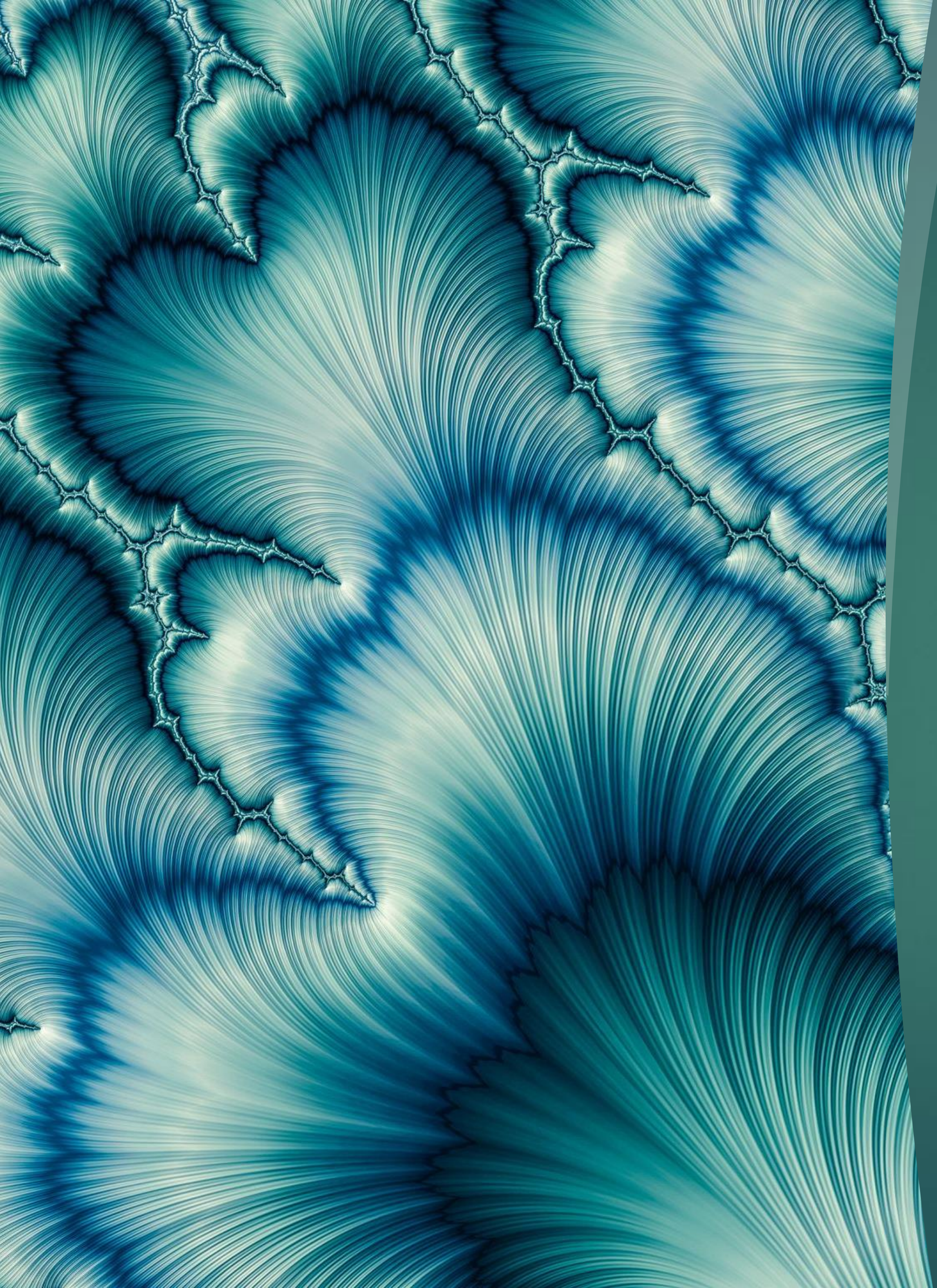
Targeted

- Aim to increase the prob of target class
  - Target class not chosen randomly.
  - A class that is not too close to original class.(10th from prediction)
- Break when highest probability is assigned to target class

SCOPE

(AS DISCUSSED WITH TA)

- TinyImageNet
- Attacks
  - Untargeted Attack (cartesian bias)
  - Targeted Attack
  - Discrete Cosine Bias (if time permits)



Implementation Details

- Finetuned the Resnet50 model to work on TinyImageNet database.
- Implemented proposed solution in paper, SimBa, to execute targeted and untargeted attack.
- Tested observations on 500 test images in untargeted and 100 images in targeted.

Dataset

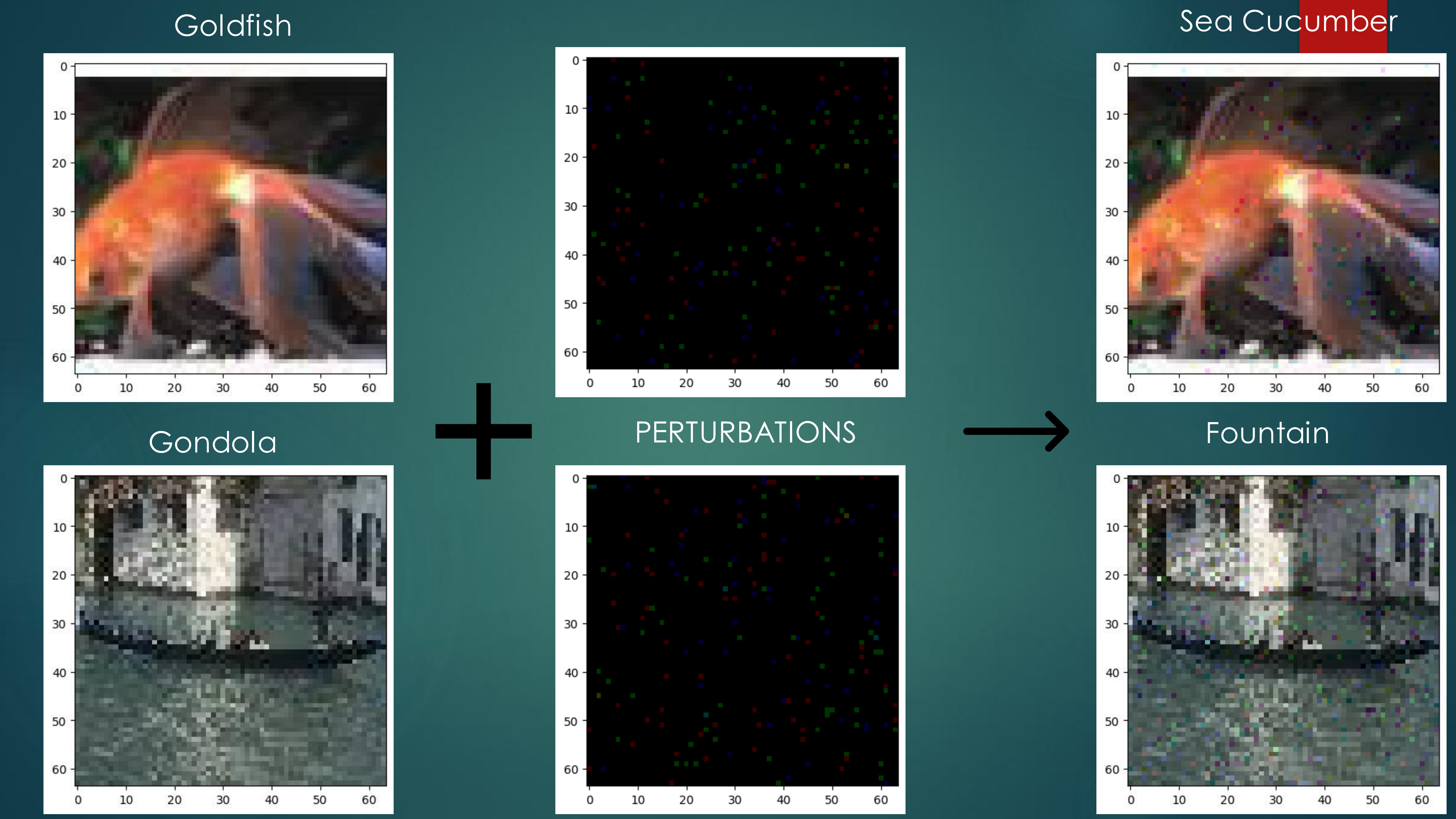
Finetuning Resnet50

```
100%|██████████| 2/2 [00:01<00:00,  1.03it/s]
100%|██████████| 1563/1563 [03:16<00:00,  7.94it/s]
100%|██████████| 157/157 [00:09<00:00, 17.01it/s]
Epoch: 1        Training Loss: 2.267500         Training Accuracy: 0.507110    Validation Loss: 1.307188        Validation Accuracy: 0.670800
100%|██████████| 1563/1563 [02:48<00:00,  9.29it/s]
100%|██████████| 157/157 [00:08<00:00, 19.04it/s]
Epoch: 2        Training Loss: 1.145882         Training Accuracy: 0.707690    Validation Loss: 1.193575        Validation Accuracy: 0.696100
100%|██████████| 1563/1563 [02:48<00:00,  9.26it/s]
100%|██████████| 157/157 [00:08<00:00, 18.93it/s]
Epoch: 3        Training Loss: 0.826008         Training Accuracy: 0.783250    Validation Loss: 1.172133        Validation Accuracy: 0.706400
```

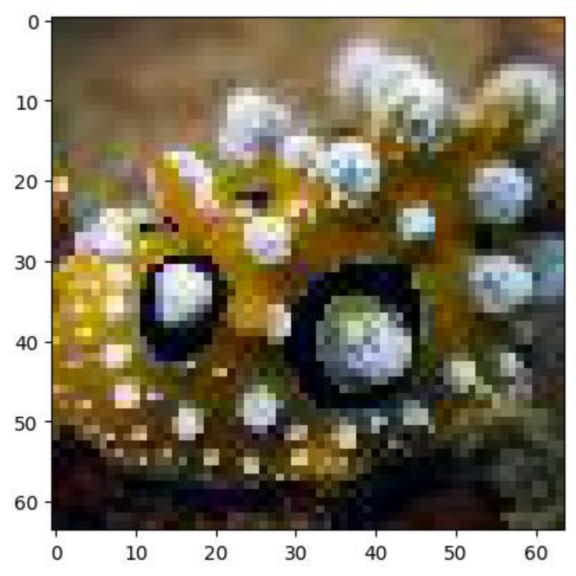
Attacks

- Untargeted attack (max 5000 iter, 500 correctly predicted images)
  - SimBA
  - SimBA-DCT
- Targeted attack (max 5000 iter, 100 correctly predicted images)
  - SimBA (can easily be extended to DCT with the same code)
- We run all attacks for e = [0.2, 0.4, 0.6, 0.8, 1.0]

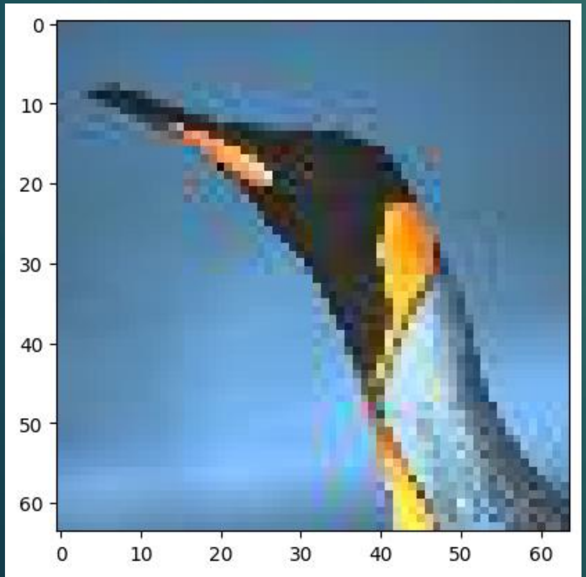
RESULTS UNTARGETED

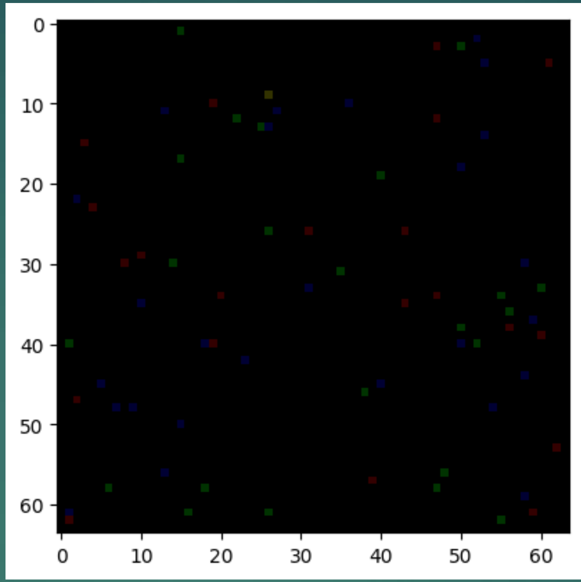

Goldfish
Gondola
PERTURBATIONS
Sea Cucumber
Fountain

Sea Slug

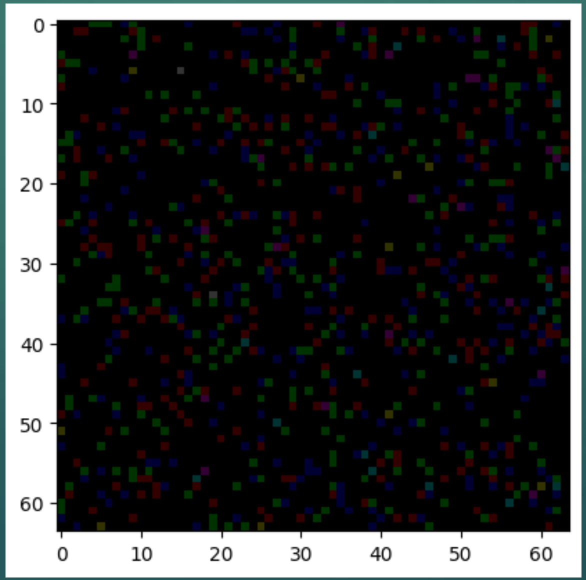


King Penguin

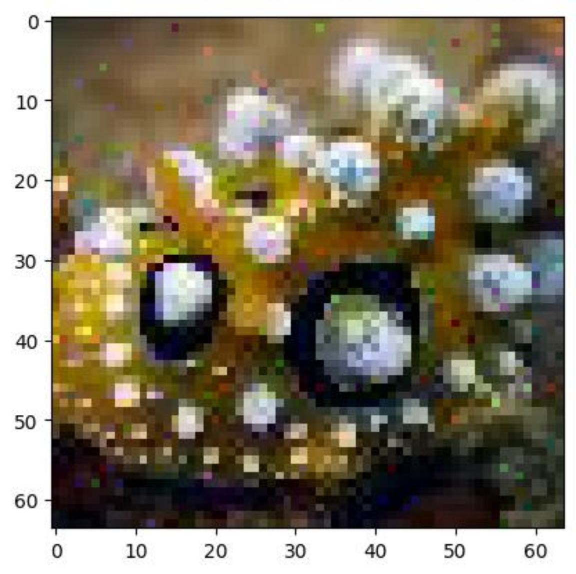




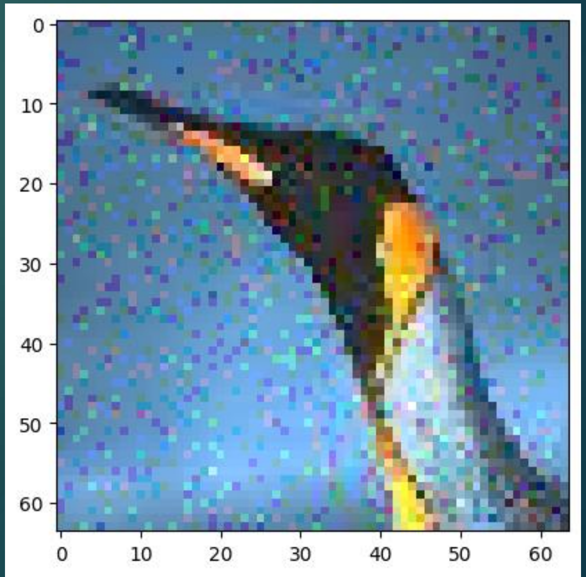
PERTURBATIONS



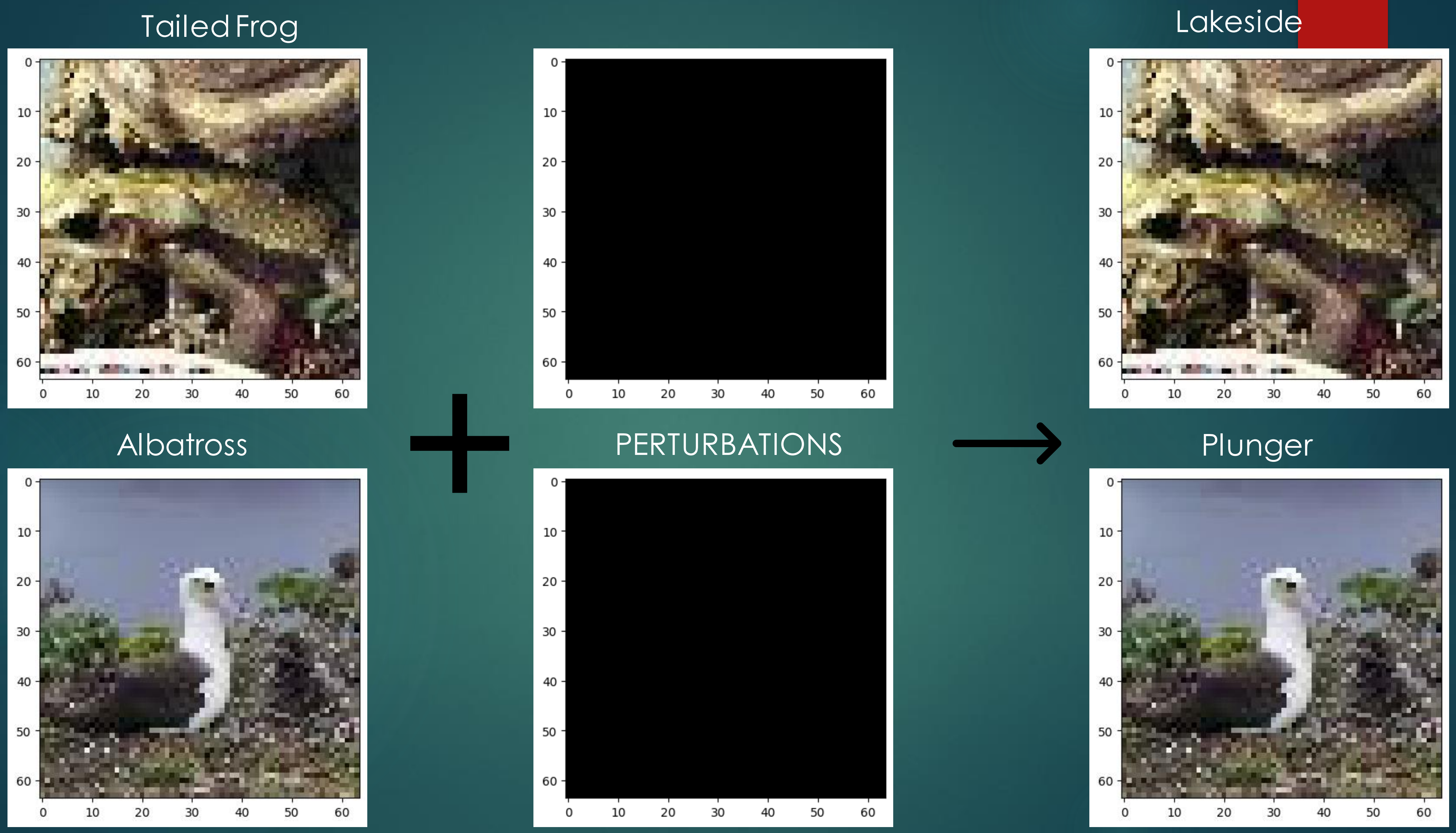
Pill Bottle

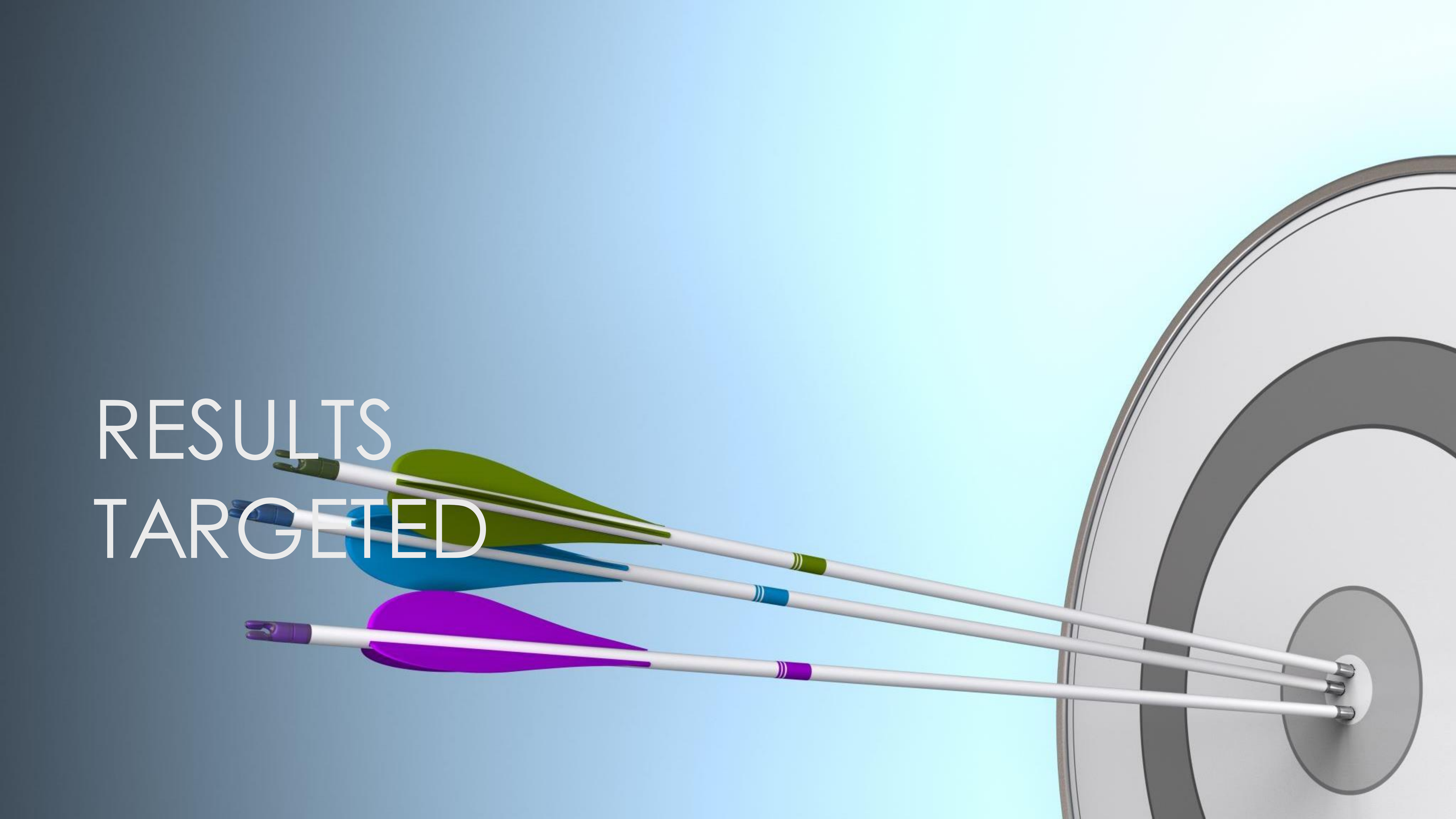


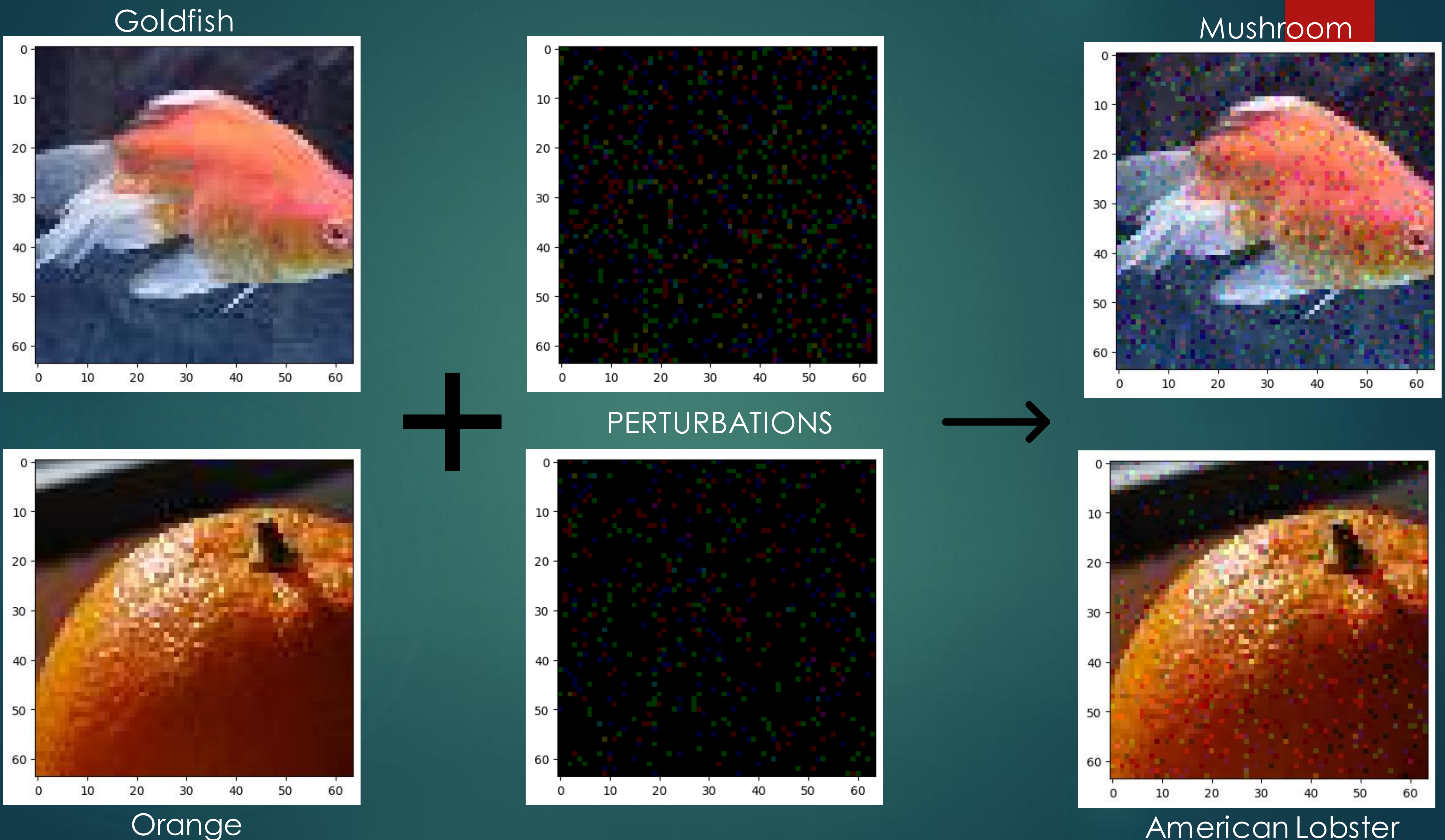
Flagpole



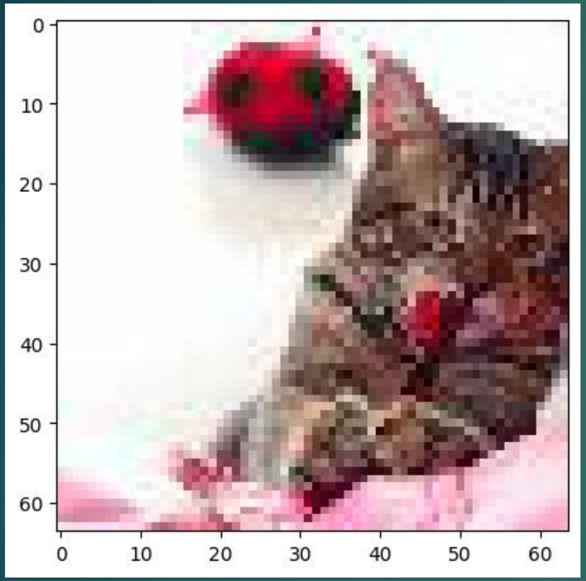
RESULTS UNTARGETED DCT

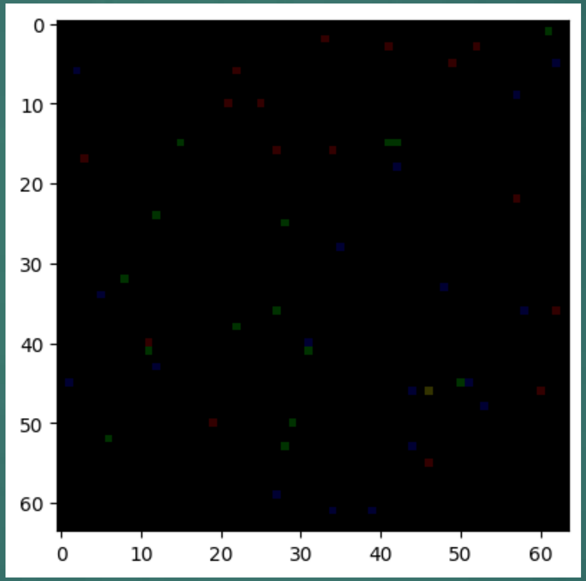

Tailed Frog
Lakeside
Albatross
PERTURBATIONS
Plunger


RESULTS
TARGETED


Goldfish
PERTURBATIONS
Mushroom
Orange
American Lobster

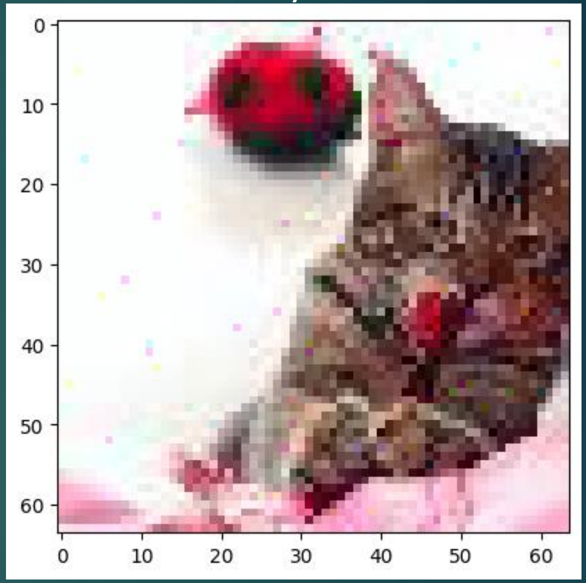
Tabby





PERTURBATIONS

Teddy Bear

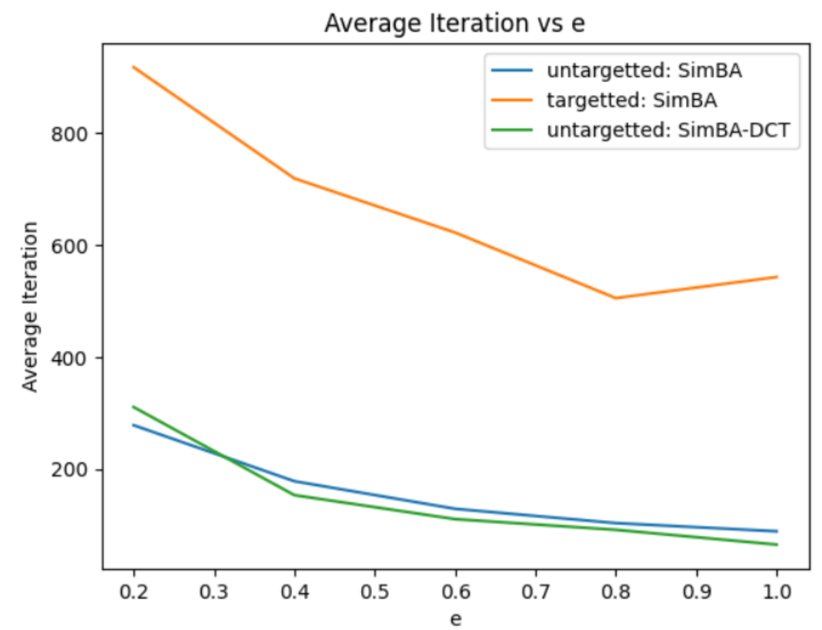


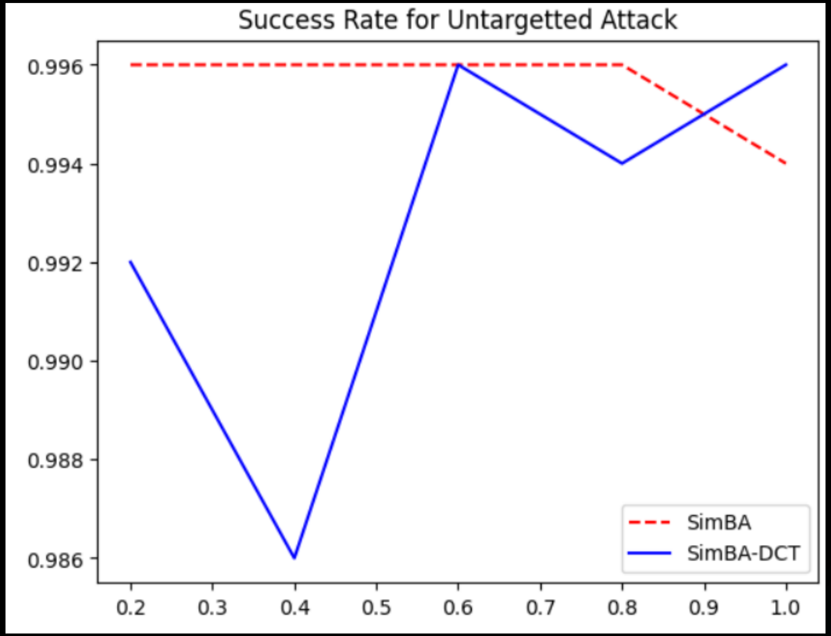
Paper: ImageNet

Attack	Average queries	Average L_2	Success rate
Untargeted			
Label-only			
Boundary attack	123,407	5.98	100%
Opt-attack	71,100	6.98	100%
LFBA	30,000	6.34	100%
Score-based			
QL-attack	28,174	8.27	85.4%
Bandits-TD	5,251	5.00	80.5%
SimBA	1,665	3.98	98.6%
SimBA-DCT	**1,283**	3.06	97.8%

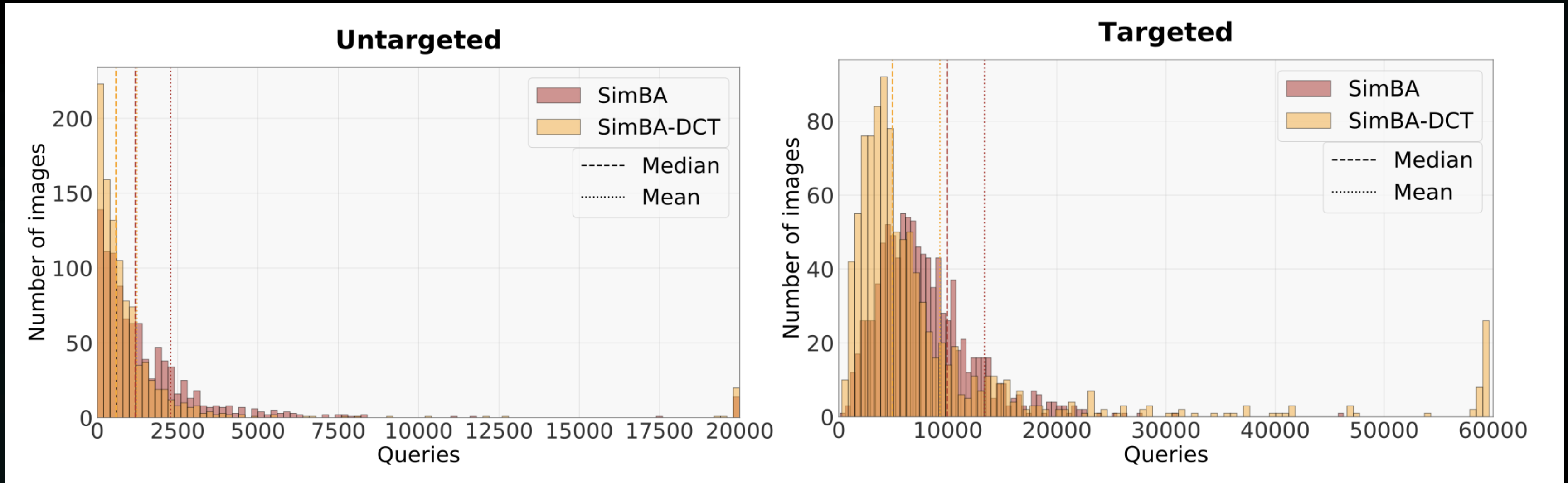
Attack	Average queries	Average L_2	Success rate
Targeted			
Score-based			
QL-attack	20,614	11.39	98.7%
AutoZOOM	13,525	26.74	100%
SimBA	**7,899**	9.53	100%
SimBA-DCT	8,824	7.04	96.5%

Our Implementation: TinyImageNet



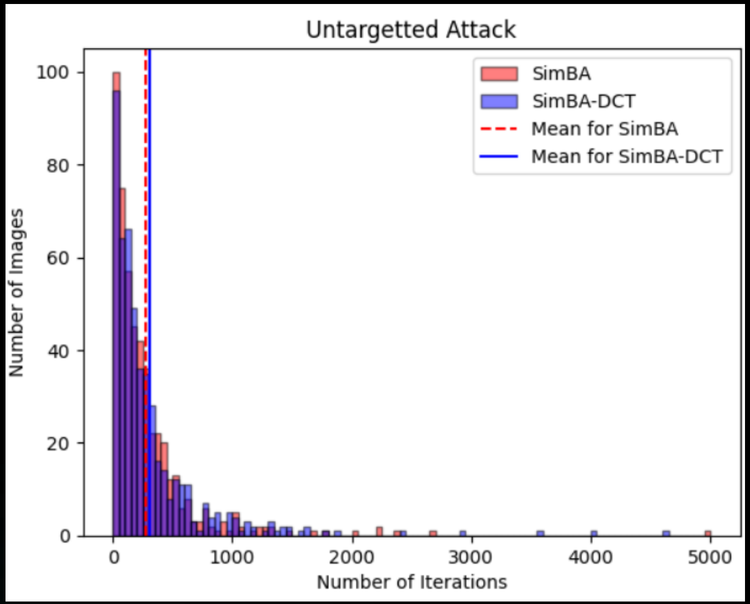


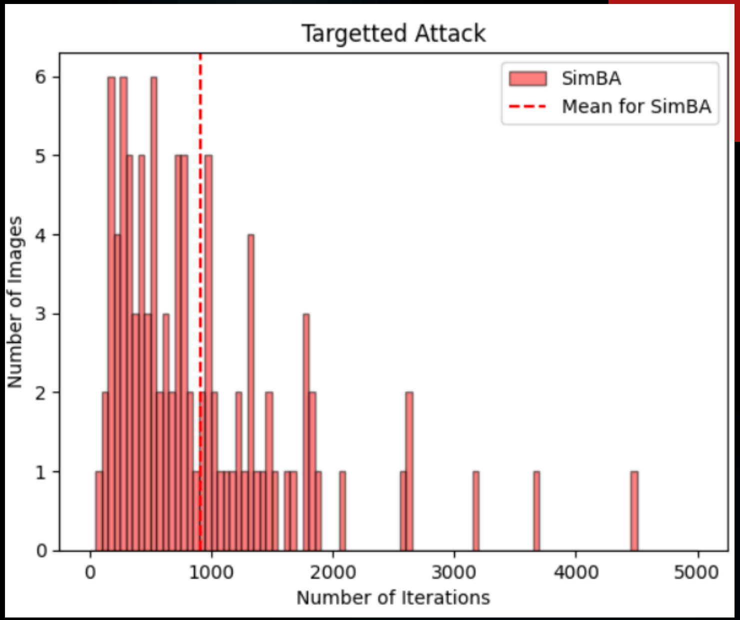
Paper:
- maximum iterations are more than ours
- more number of images

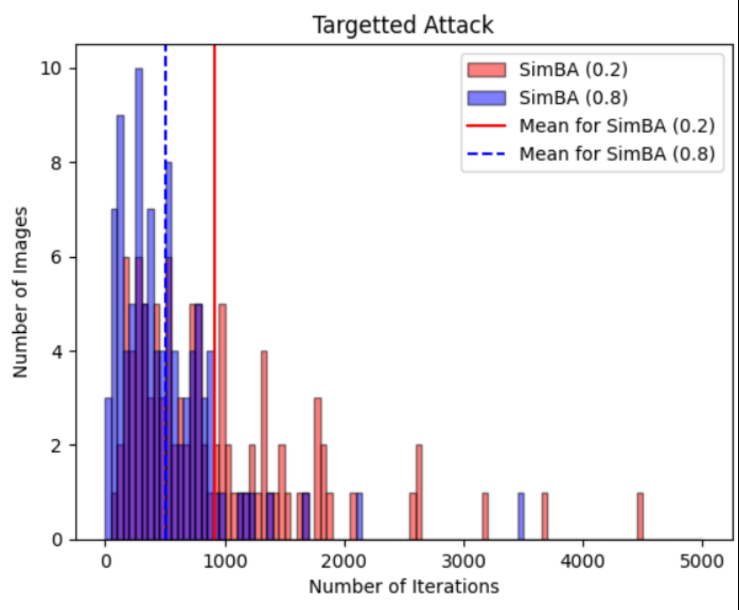


where e = 0.2

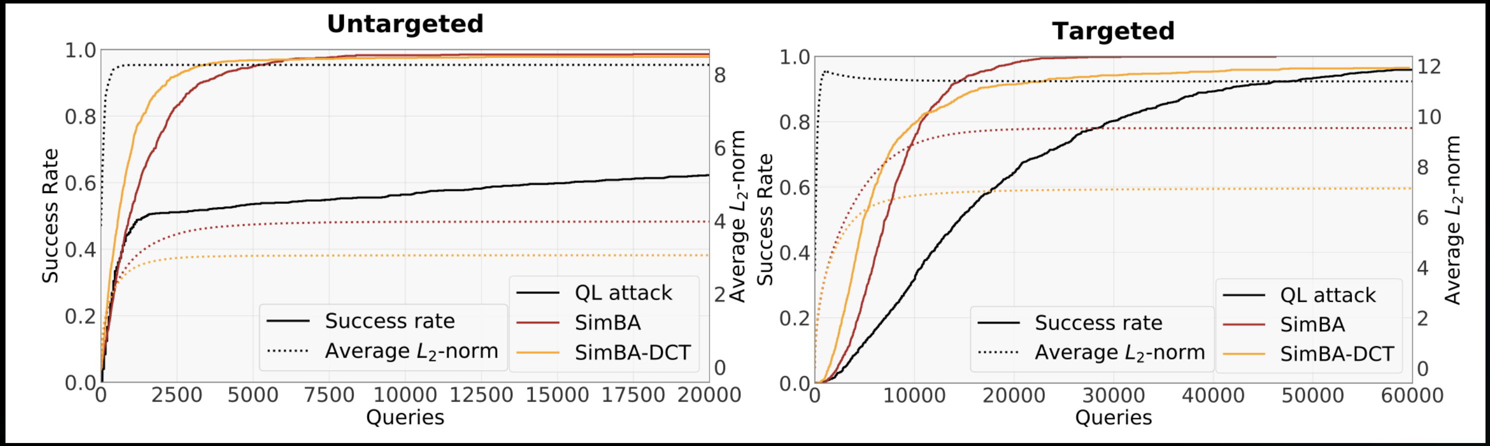
Ours :



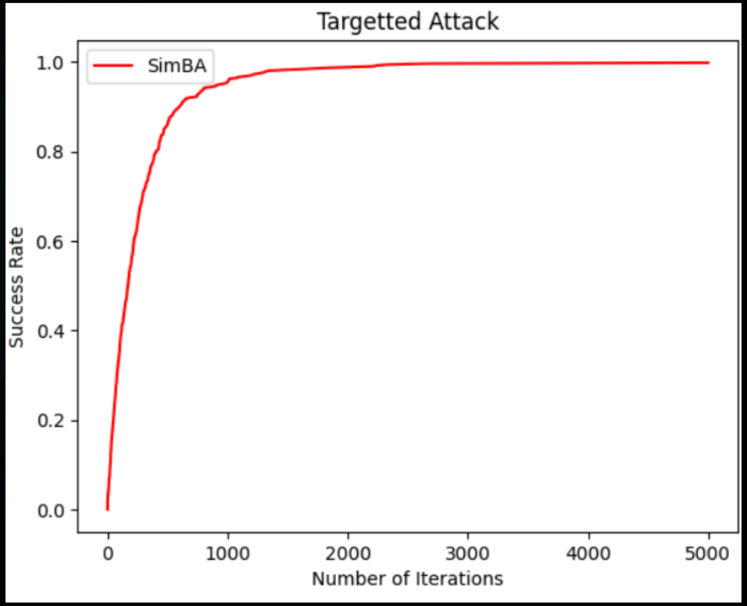


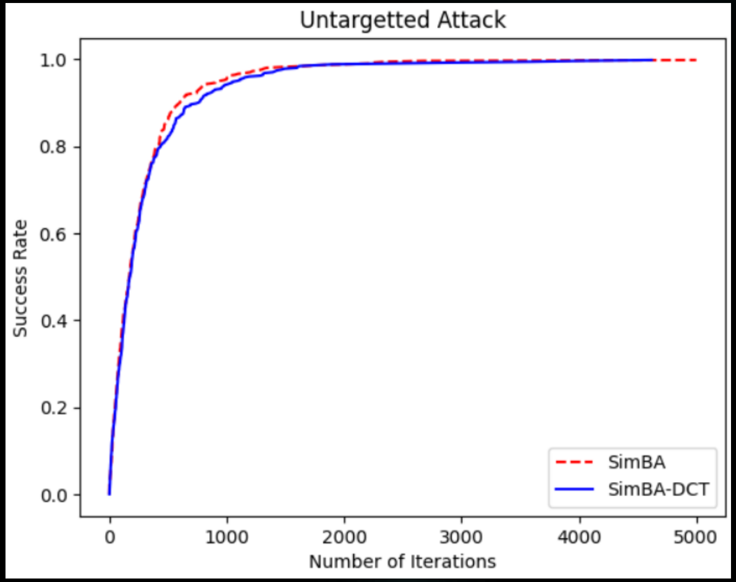


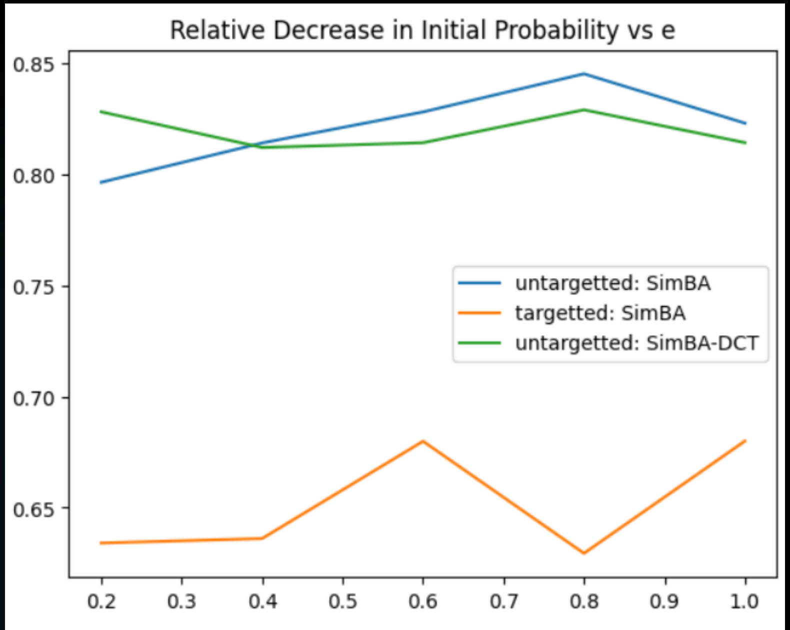
Paper:

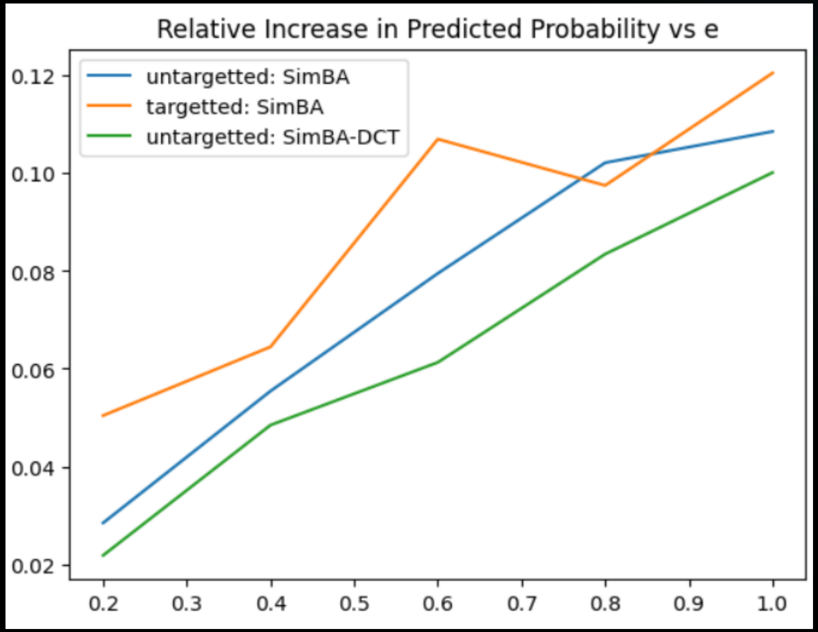

Untargeted
Targeted
Success Rate
Average L_2-norm
Queries
Success rate
Average L_2-norm
QL attack
SimBA
SimBA-DCT

Ours :


Targetted Attack
SimBA
Success Rate
Number of Iterations


Untargetted Attack
SimBA
SimBA-DCT
Success Rate
Number of Iterations


Relative Decrease in Initial Probability vs e
untargetted: SimBA
targetted: SimBA
untargetted: SimBA-DCT
0.85
0.80
0.75
0.70
0.65
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


Relative Increase in Predicted Probability vs e
untargetted: SimBA
targetted: SimBA
untargetted: SimBA-DCT
0.12
0.10
0.08
0.06
0.04
0.02
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Limitations

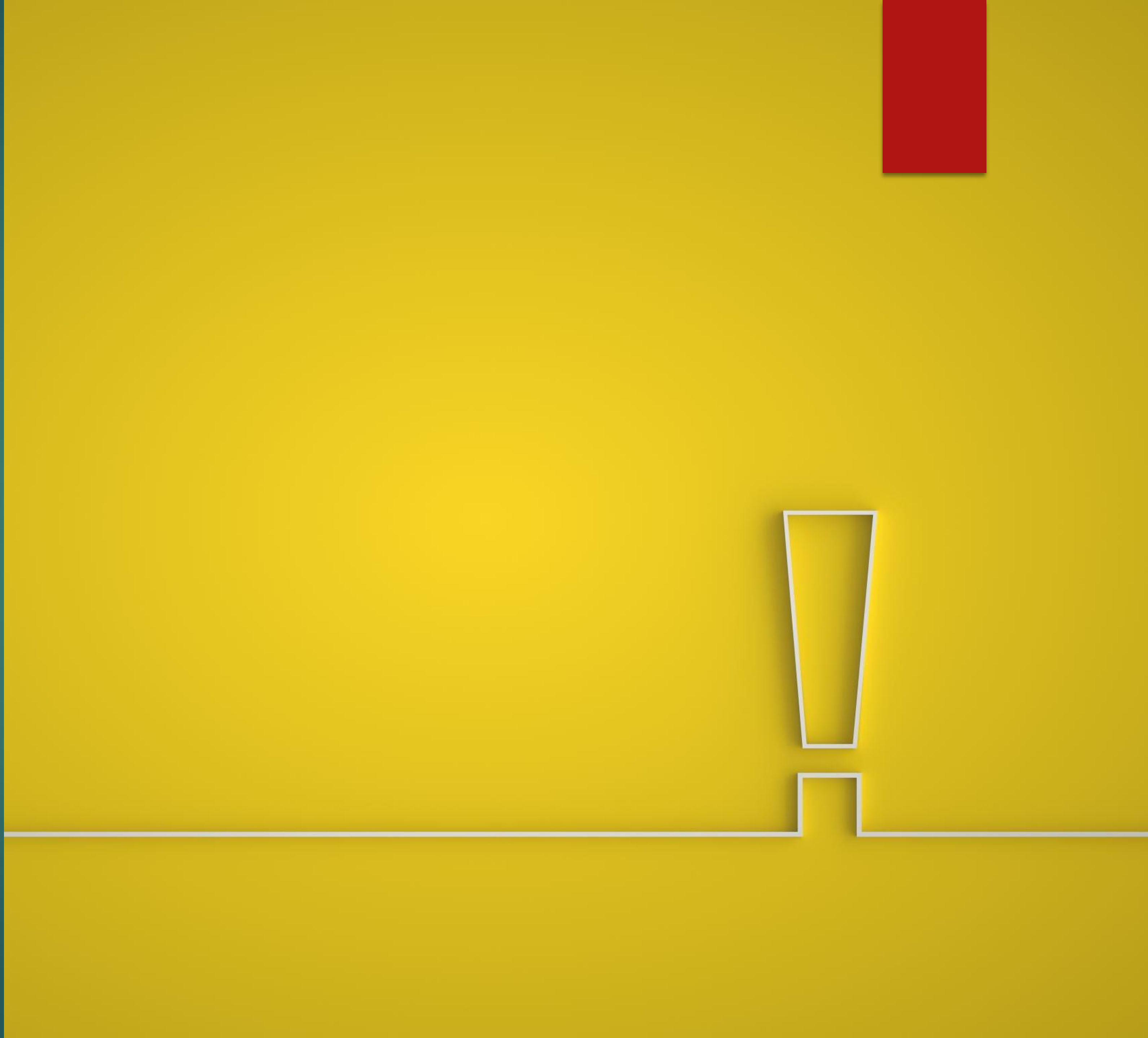
- Smaller images were harder to imperceptibly attack since not all pixels contribute to the confidence of a particular class.
- Finding effective perturbations for smaller images requires more fine-grained control over the perturbation, which can be challenging to achieve while maintaining imperceptibility.

Future Work

- General Basis to get Q
- Try on more images and higher iterations
- Try ImageNet instead of TinyImageNet

References

- Simple Black-box Adversarial Attacks; Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, Kilian Q. Weinberger
- Low Frequency Adversarial Perturbation; Chuan Guo, Jared S. Frank, Kilian Q. Weinberger
- https://huggingface.co/datasets/Maysee/tiny-imagenet



CONTRIBUTION

- ESHIKA – FINETUNE RESNET + SIMBA-DCT
- AMEYA – TARGETED ATTACK
- ADITH - UNTARGETED ATTACK (SIMBA)