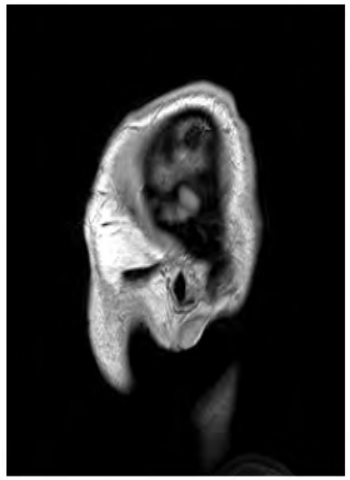
ADVERSARIAL MACHINE LEARNING

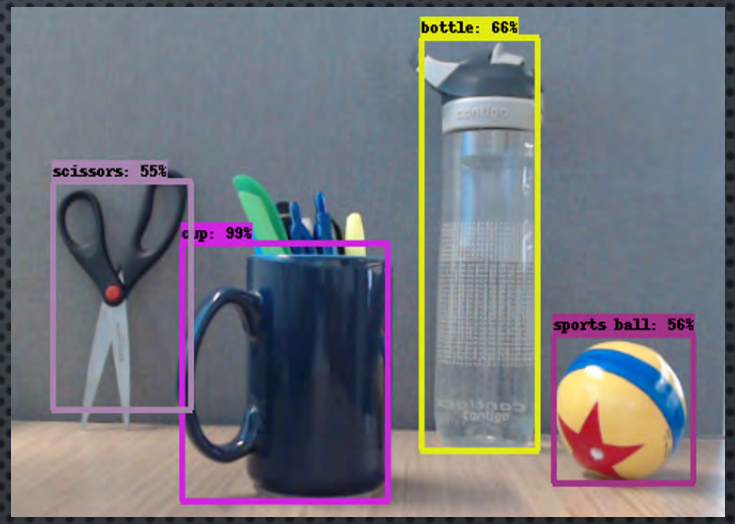
FINAL PRESENTATION

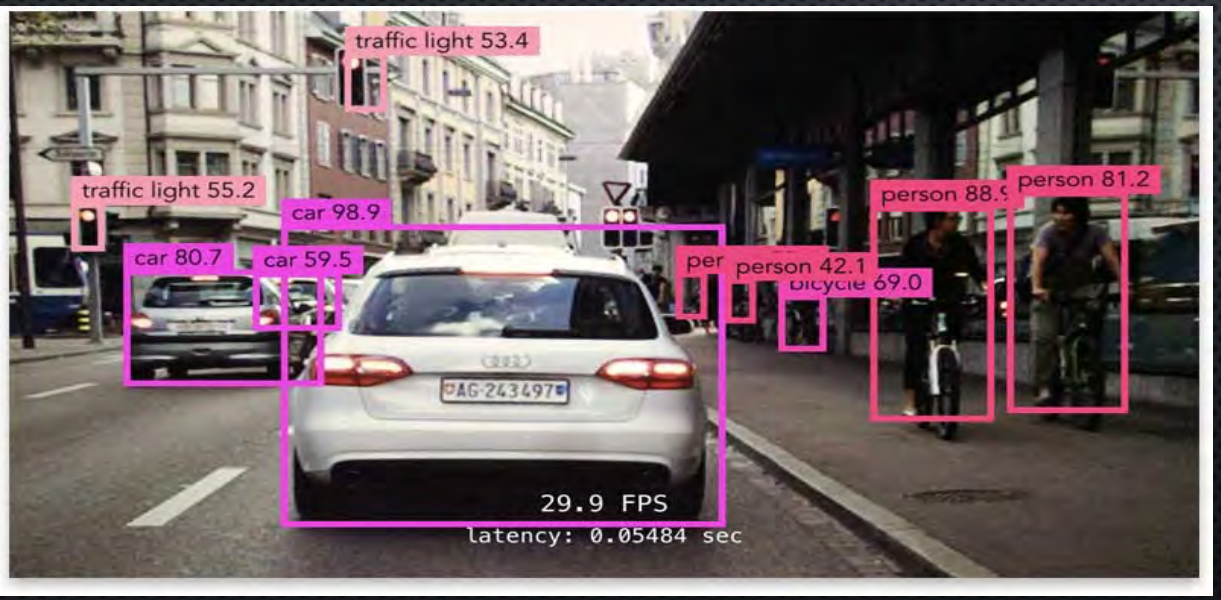
Object Detection and Classification

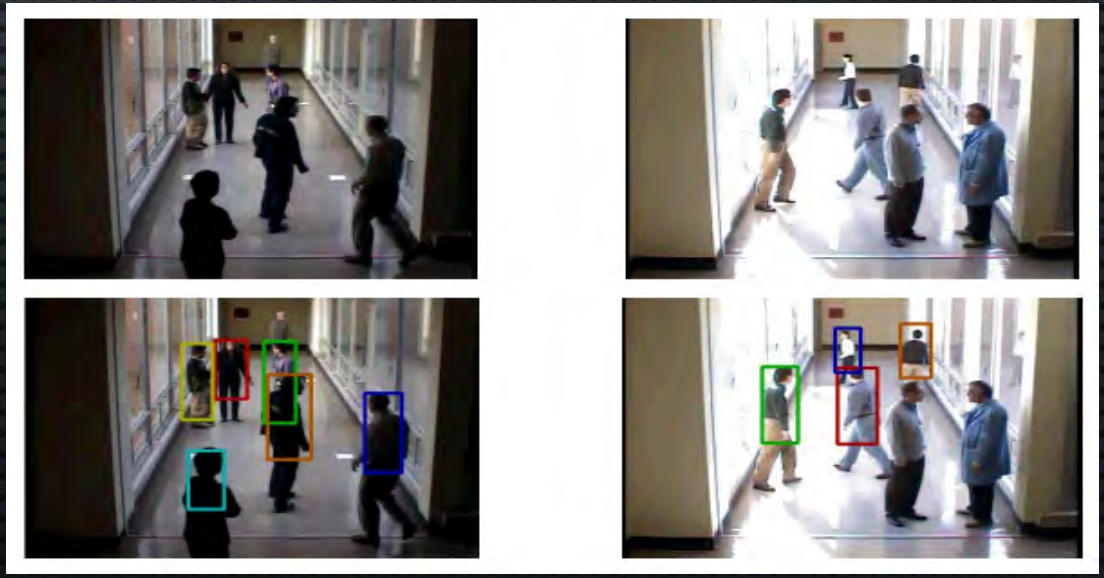
"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."



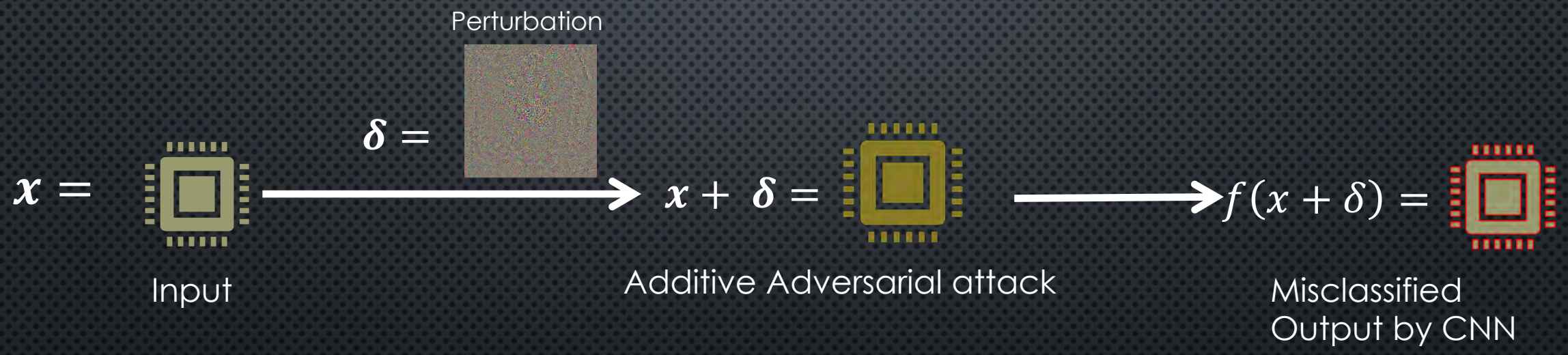




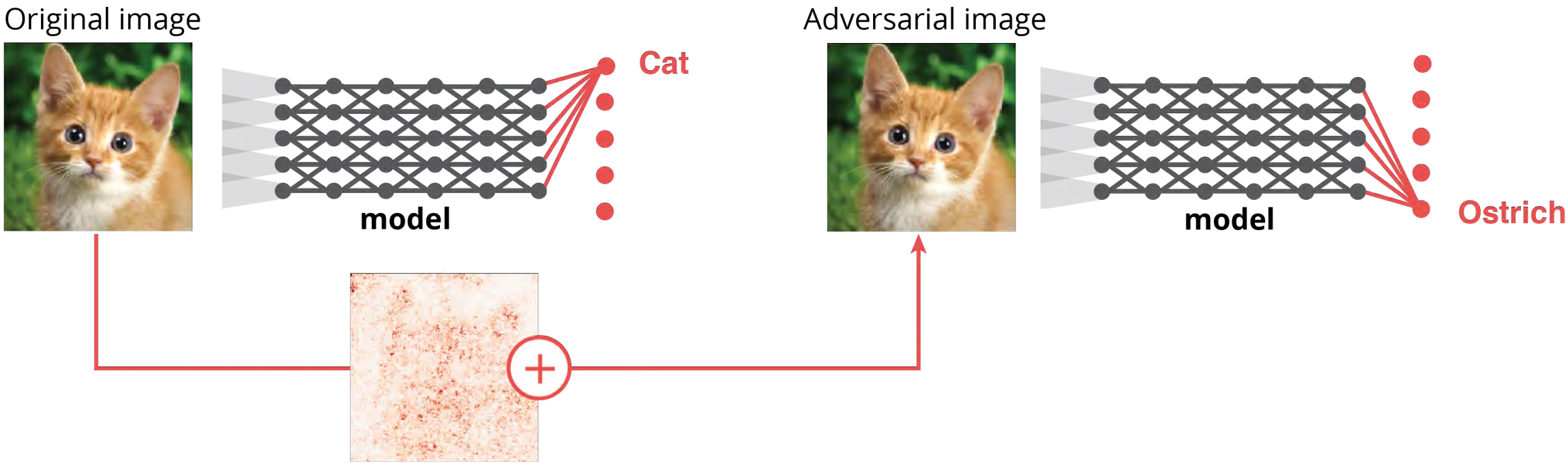


ADVERSARIAL MACHINE LEARNING

FOOLING ML MODELS USING ADVERSARIAL MACHINE LEARNING



Apply Perturbation to Input Image

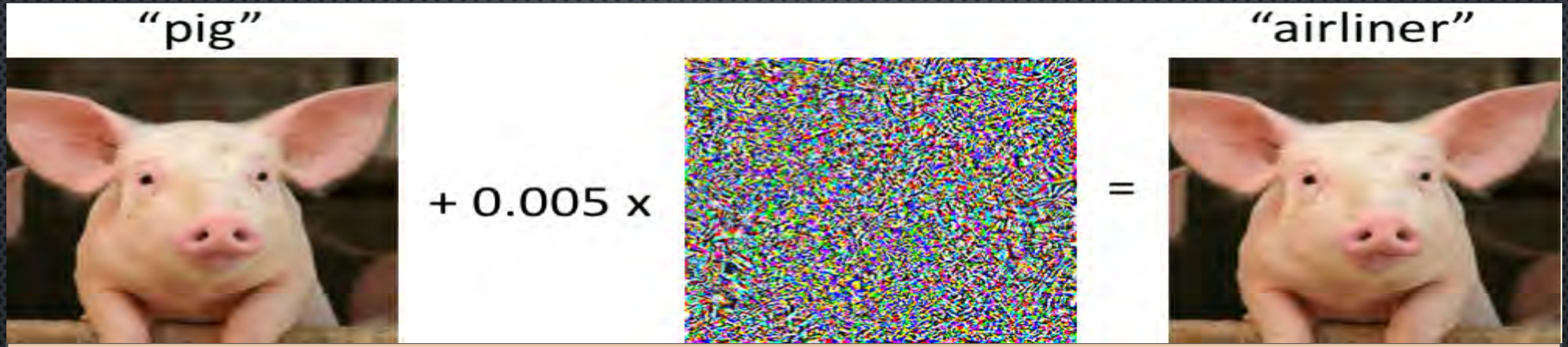


(small) adversarial perturbation created by **attack**

Computing Perturbations

- Fast Gradient Sign Method (FGSM)
- One Pixel Attack
- Universal Adversarial Perturbations

Fast Gradient Sign Method (FGSM)



x

60% Confidence

ϵ

$\mathbf{sign}(\nabla_x J(\theta, x, y))$

"nematode"

9% Confidence

$x + \epsilon\, sign(\nabla_x J(\theta, x, y))$

99.4% Confidence

One Pixel Attack

Airplane (Dog)

Automobile (Dog)

Automobile (Airplane)

Cat (Dog)

Dog (Ship)

Deer (Dog)

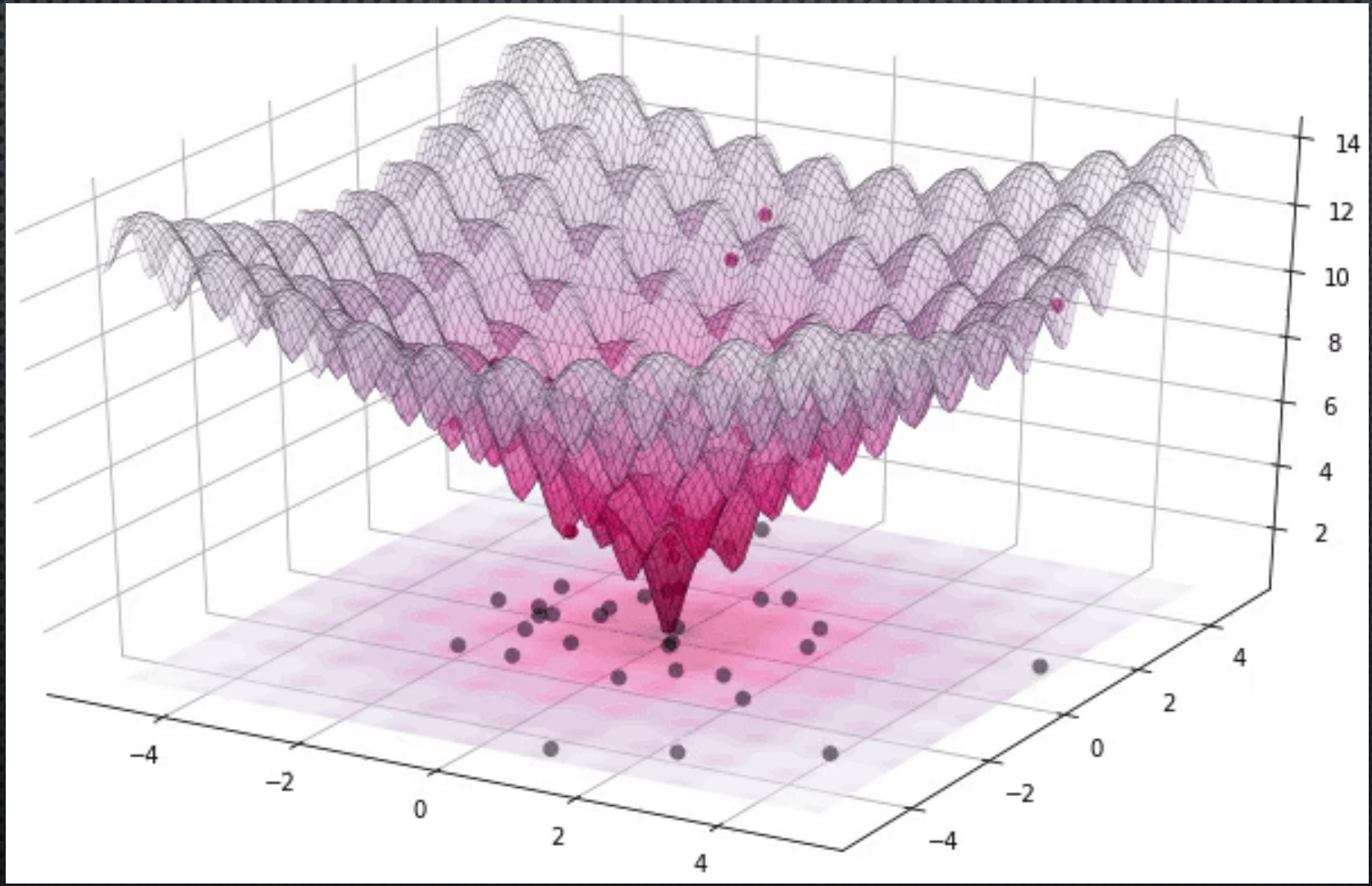
Frog (Dog)

Frog (Truck)

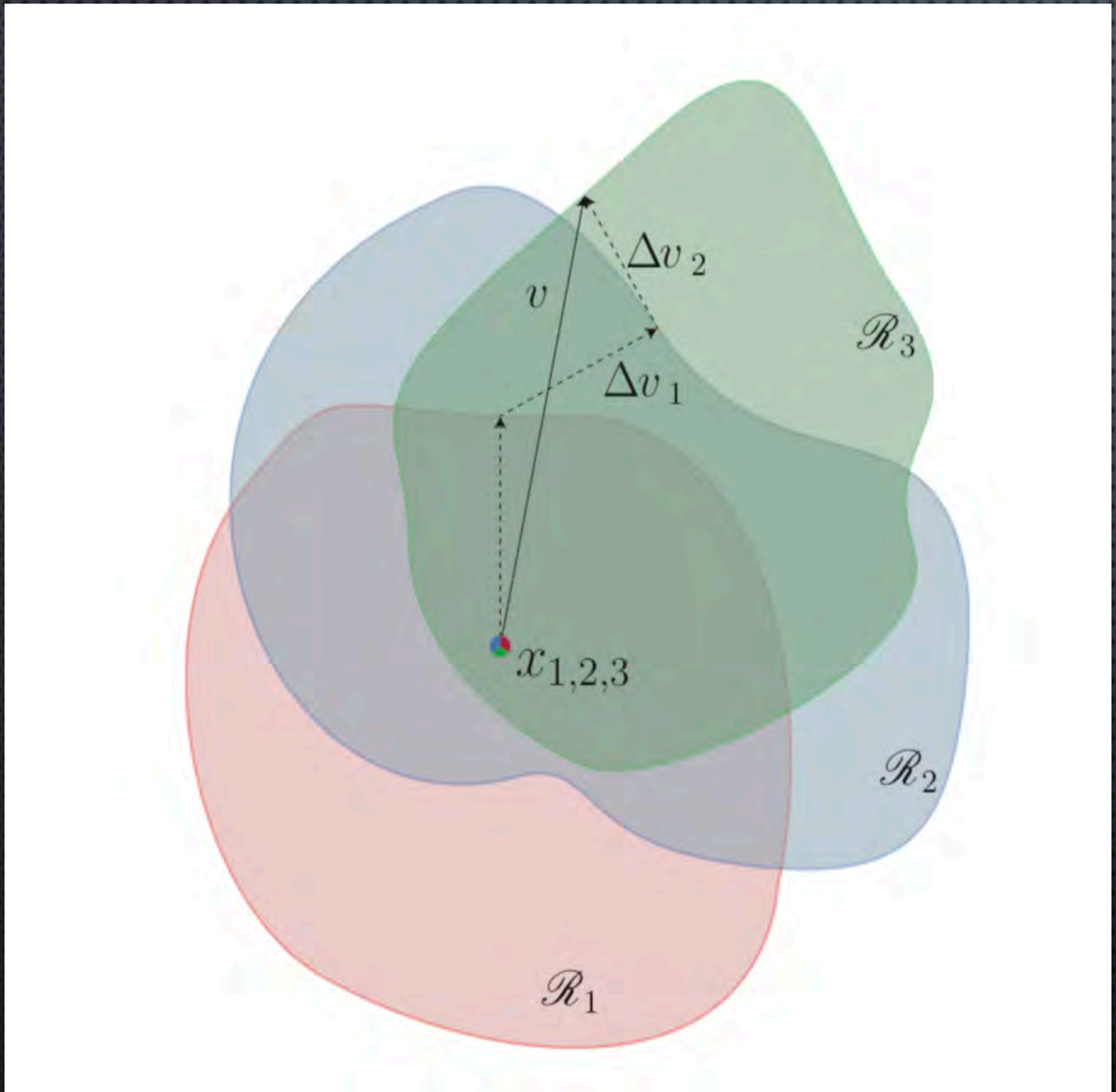
Dog (Cat)

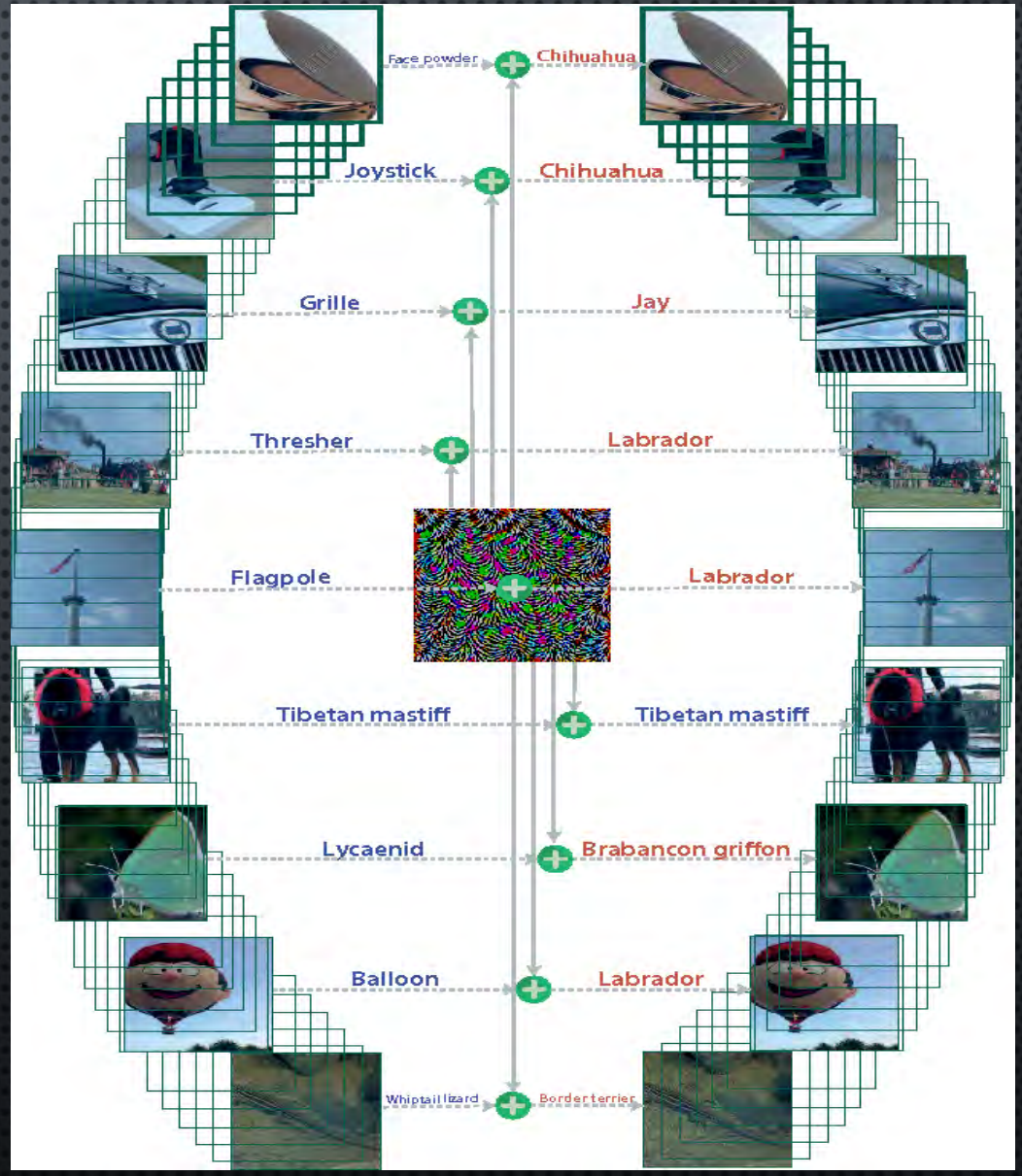
Bird (Airplane)

Differential Evolution



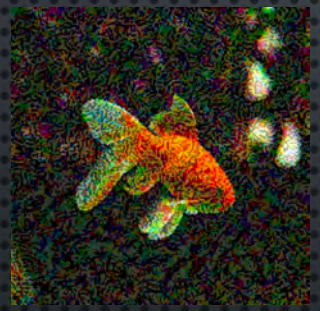
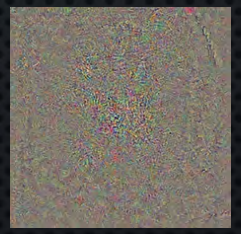
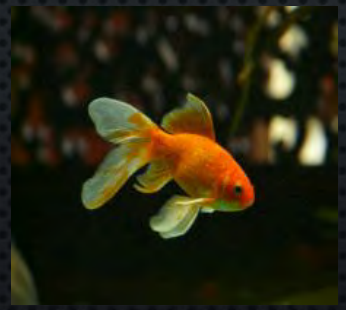
Universal Adversarial Perturbations



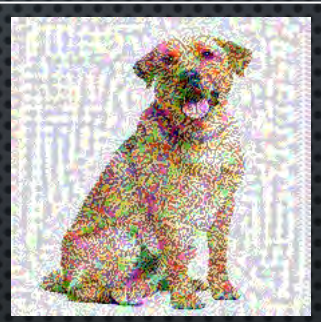
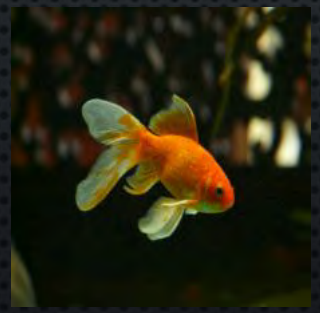


ADVERSARIAL ATTACK RESULTS

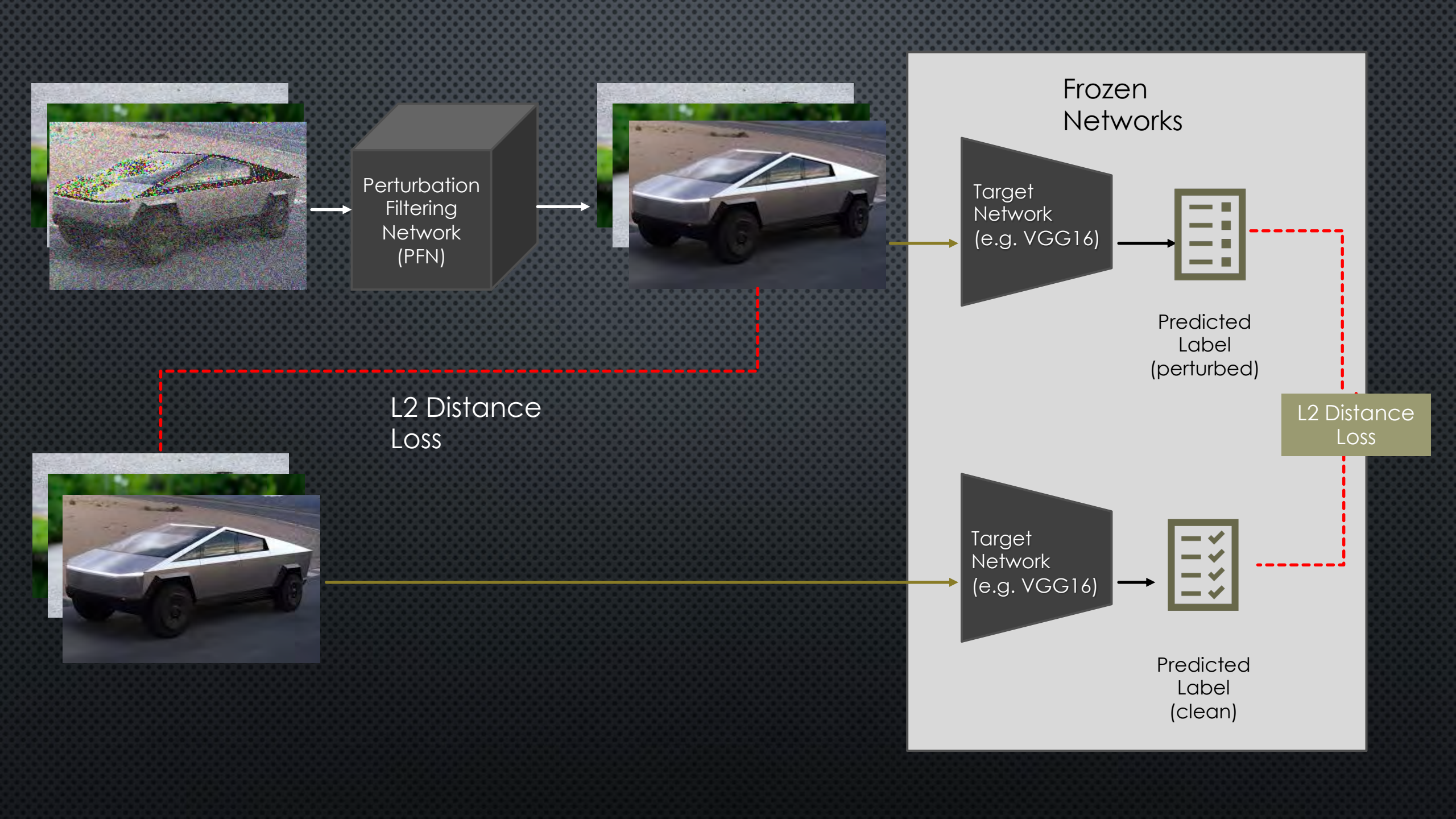
Model 1: **Resnet18**

Original Image:	Goldfish	Bee	Orange	Labrador	Airliner
Attack Type: **FGSM**	 Bubble	Tick	Loupe	Bull Mastiff	Space Shuttle
Attack Type: **Universal Perturbation** 	 Mud, puppy	Tick	Lemon	Bay retriever	Space Shuttle

Model 2: **Resnet101**

Original Image:	Goldfish	Bee	Orange	Labrador	Airliner
Attack Type: **FGSM**	Screw	Sea Urchin	Orange	 Airedale	Missile
Attack Type: **Universal Perturbation**	 Hook	Bucket	Claw	Tennis Ball	Hook

DEFENSE AGAINST ADVERSARIAL ATTACKS


Perturbation Filtering Network (PFN)
Frozen Networks
Target Network (e.g. VGG16)
Predicted Label (perturbed)
L2 Distance Loss
L2 Distance Loss
Target Network (e.g. VGG16)
Predicted Label (clean)

DEFENSE ON **IMAGENET** DATASET

Perturbed

Spaniel (a dog)
17% Confidence

Coffeepot 29%

Rug 12%

We reduce the fooling rate from <u>92% to 33%</u>

Filtered
(Defense)

Sports Car
41% Confidence

Airliner 57%

Tabby 20%