Hewlett Packard Enterprise

Learn how AI hackers detect fragility and how to thwart them with AI model resilience

Soumyendu Sarkar, Senior Distinguished Technologist and Senior Director - AI, HPE Labs
Ashwin Ramesh Babu, Expert AI Research Scientist, HPE Labs
Sajad Mousavi, Expert AI Research Scientist, HPE Labs

March 20, 2024 9AM PT/12PM ET/6PM CT

Agenda

- Introduction
- Robustness Evaluation for Image Classifiers
- Robustness Refinement
- Visual Explanation
- Signals and ECG Classification
- Video Classification
- Benchmark and Summary

Introduction

Vulnerability of ML models and Measuring Trustworthiness

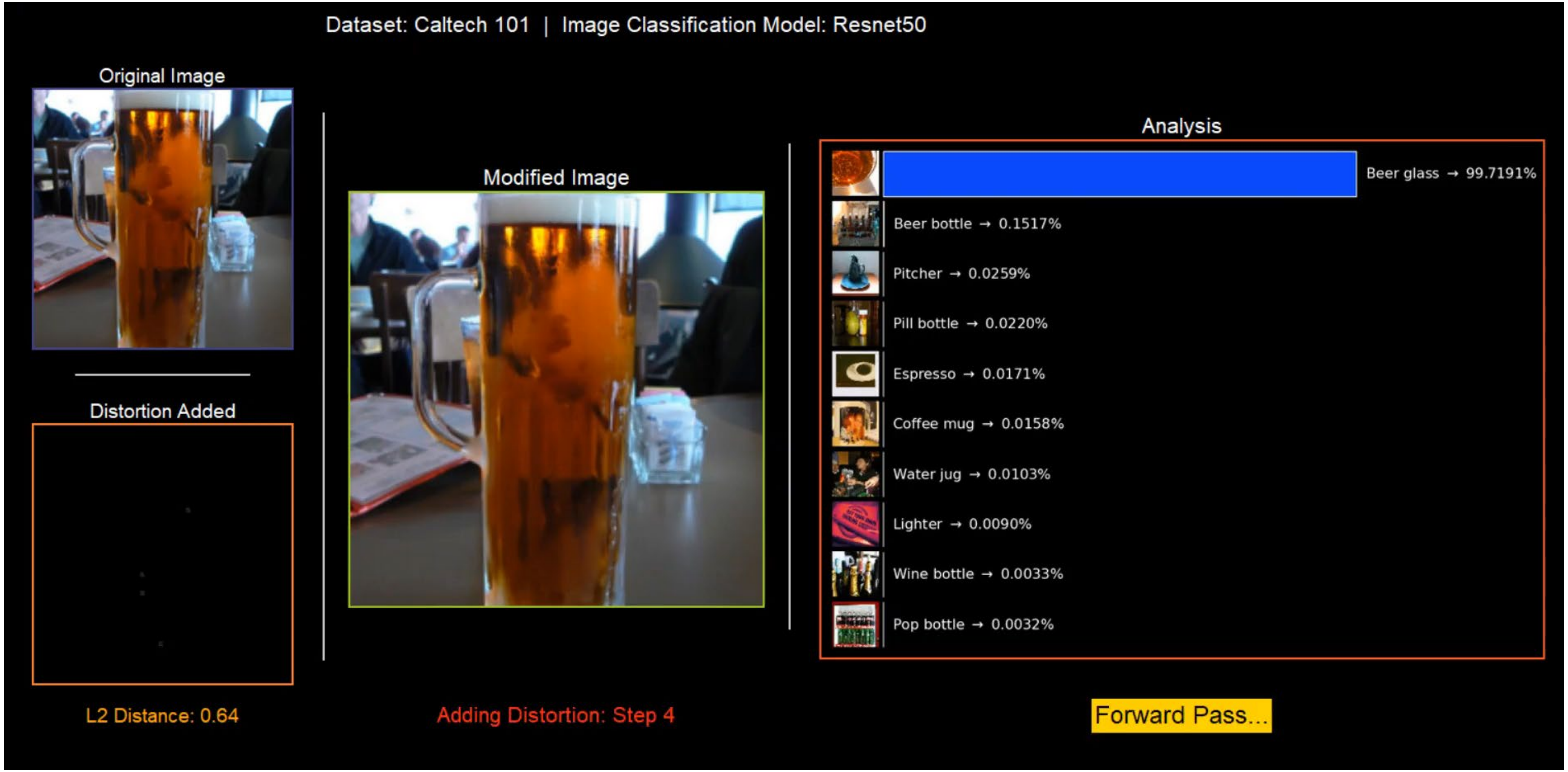
In Machine learning models a small perturbation of data may cause a model to misclassify

Increased Regulations for Machine Learning usage will require algorithmic audit and measurement of Trustworthiness

Test data sets used are limited to static testing and are unable to catch real world distortions or adversarial attacks

As ML models are complex, we need smart ML agents to analyze ML models

RLAB: Measuring robustness of image classification (CNN) Model



ML to test robustness of image classification model

Why do we need it ?

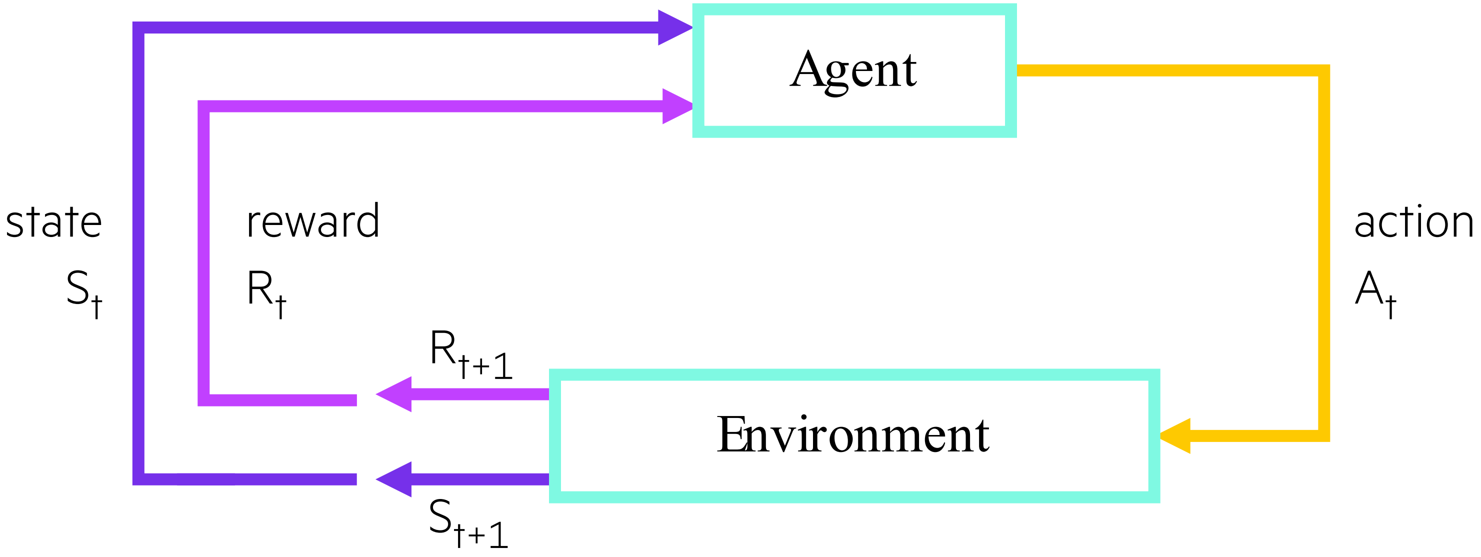
- Robustness of Machine Learning models is key for Trustworthiness
- Robustness of ML models is key for ensuring consistent classification accuracy with variations in input data and is an important element of trustworthiness.
- Quantitative metric for robustness and can be used for algorithmic audit for trustworthiness

What do we need it to do ?

- We model perturbations that occur naturally at deployment because of distortions from camera or from adversarial attacks.
- To tackle the complexity of ML models, Hewlett Packard Labs developed Reinforcement Learning based smart agents to evaluate robustness of ML models, by finding the minimum distortion needed for misclassification.
- *RLAB* – Reinforcement Learning based Adversarial Black-box attack, is HPE Lab's Platform for measuring robustness and other aspects of Trustworthy AI
- For robustness with image classification models, RLAB supports several types of naturally occurring distortions like Gaussian noise, Gaussian blur, and dead pixels.
- This technique can help retrain the ML models to enhance robustness against outliers.

Robustness Evaluation for Image Classifiers

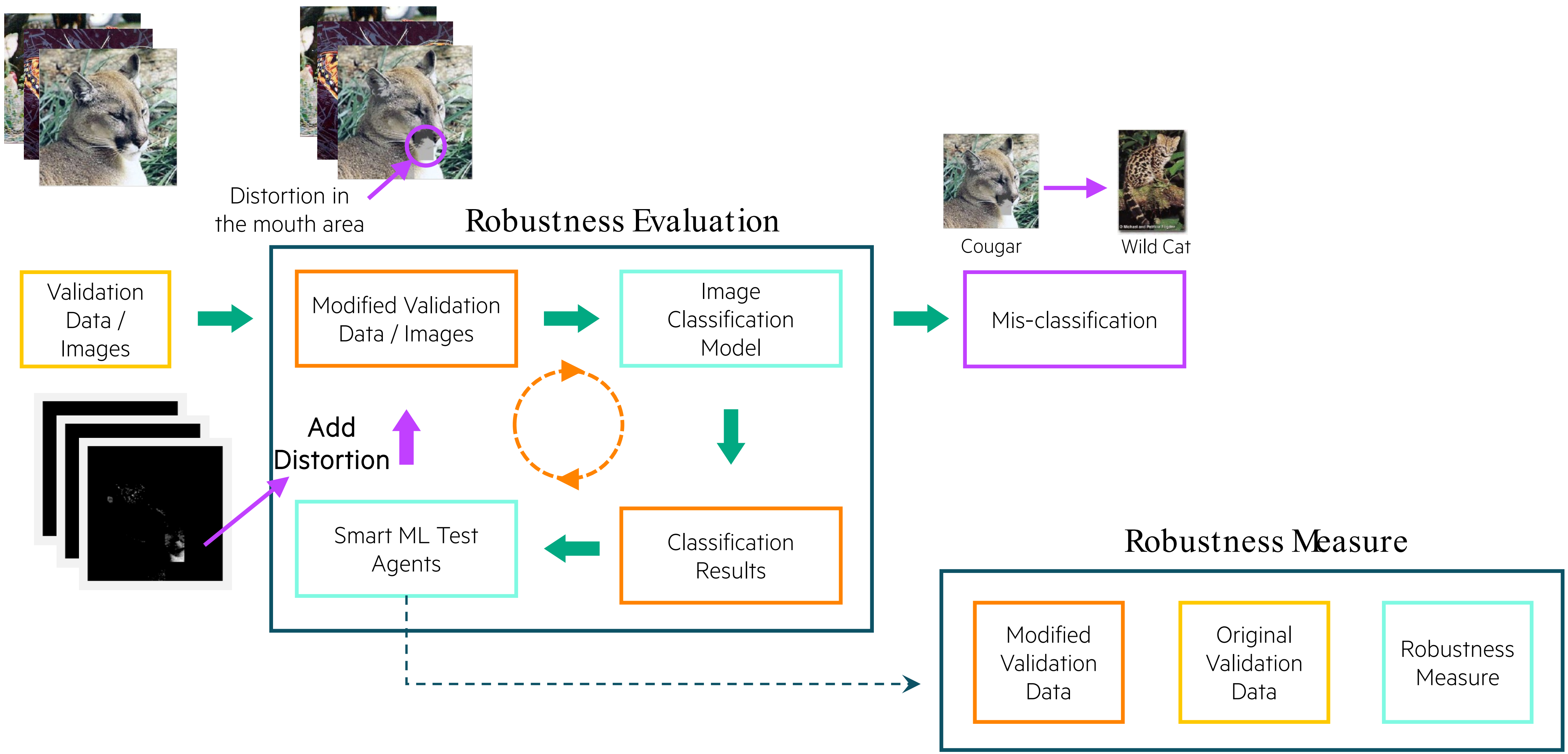
Reinforcement Learning



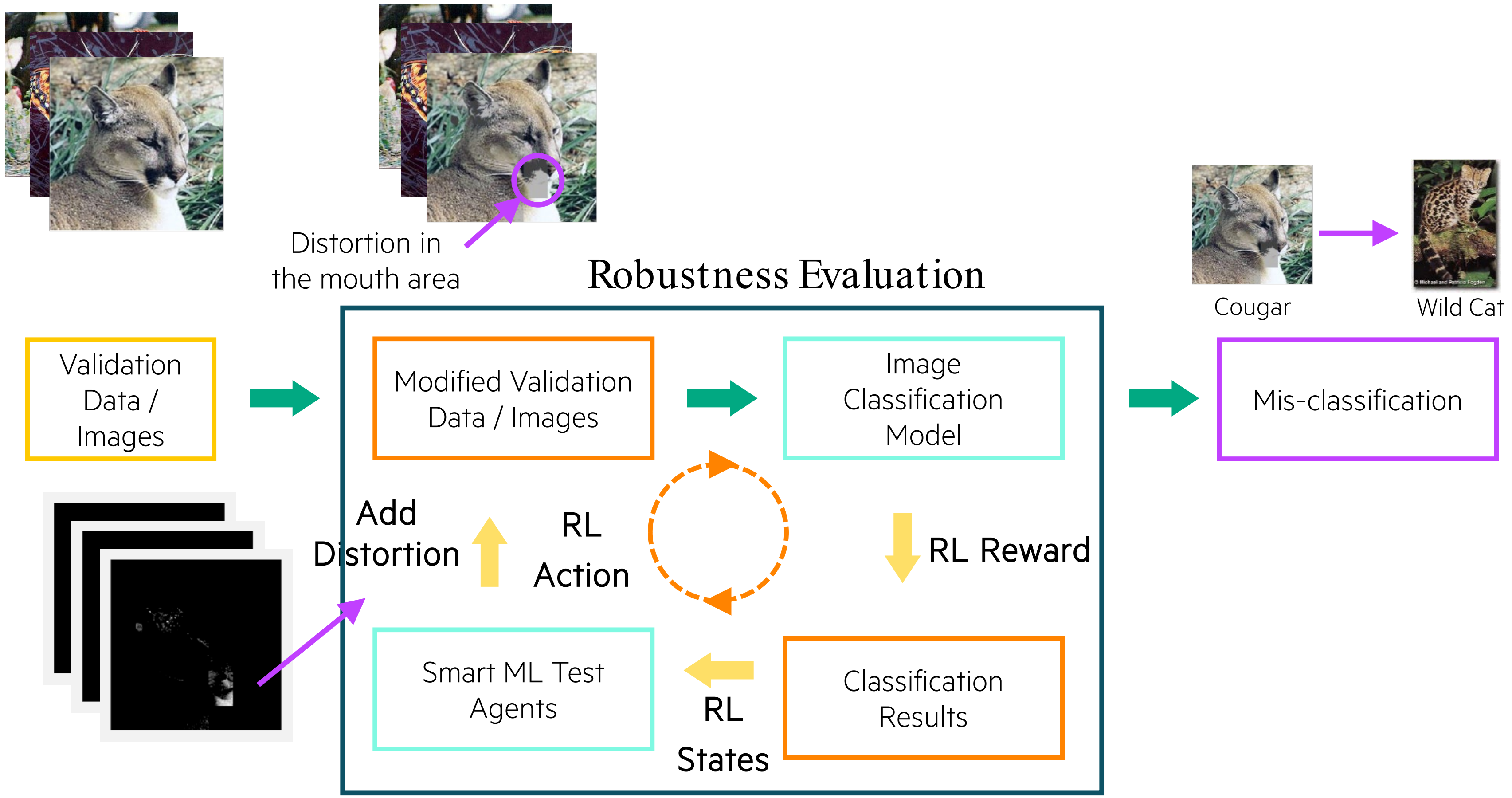
At each step, the agent:

- Executes action
- Observe new state
- Receive reward

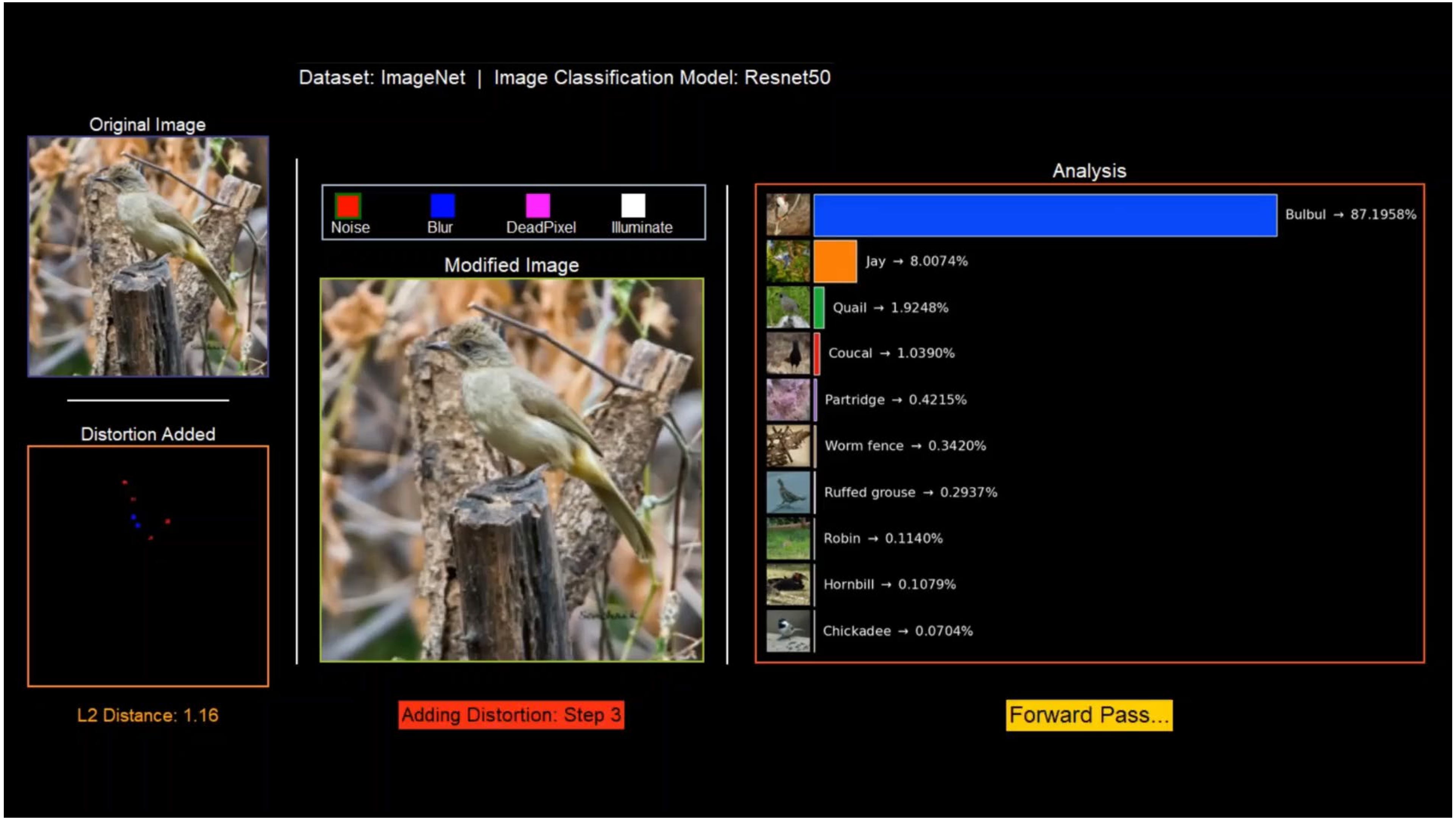
Measuring robustness of Black Box image classification (CNN) model



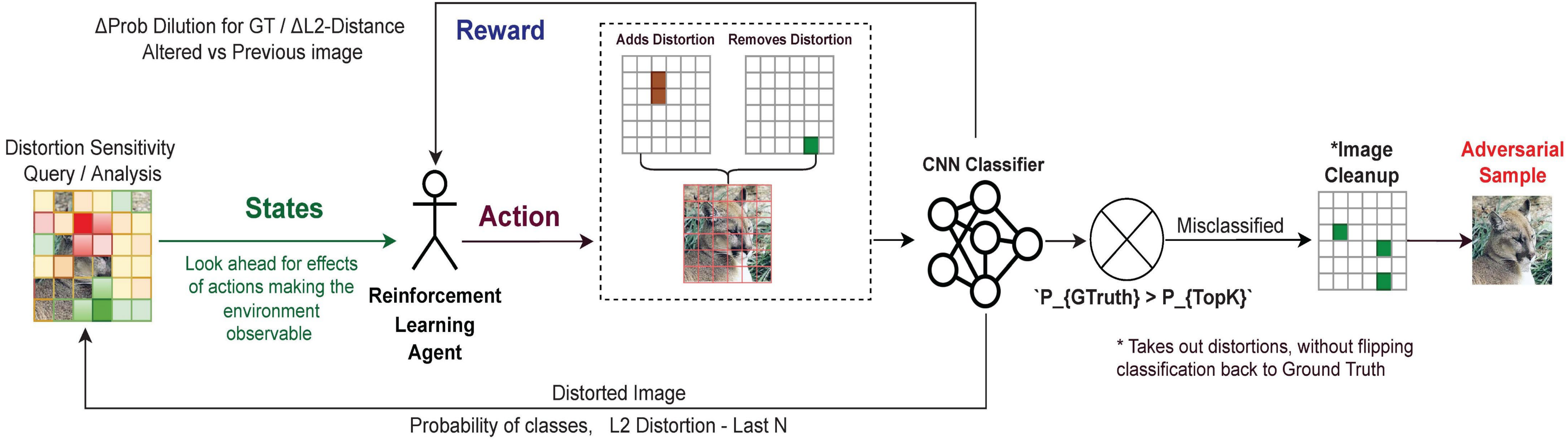
Measuring robustness : Reinforcement Learning Agent



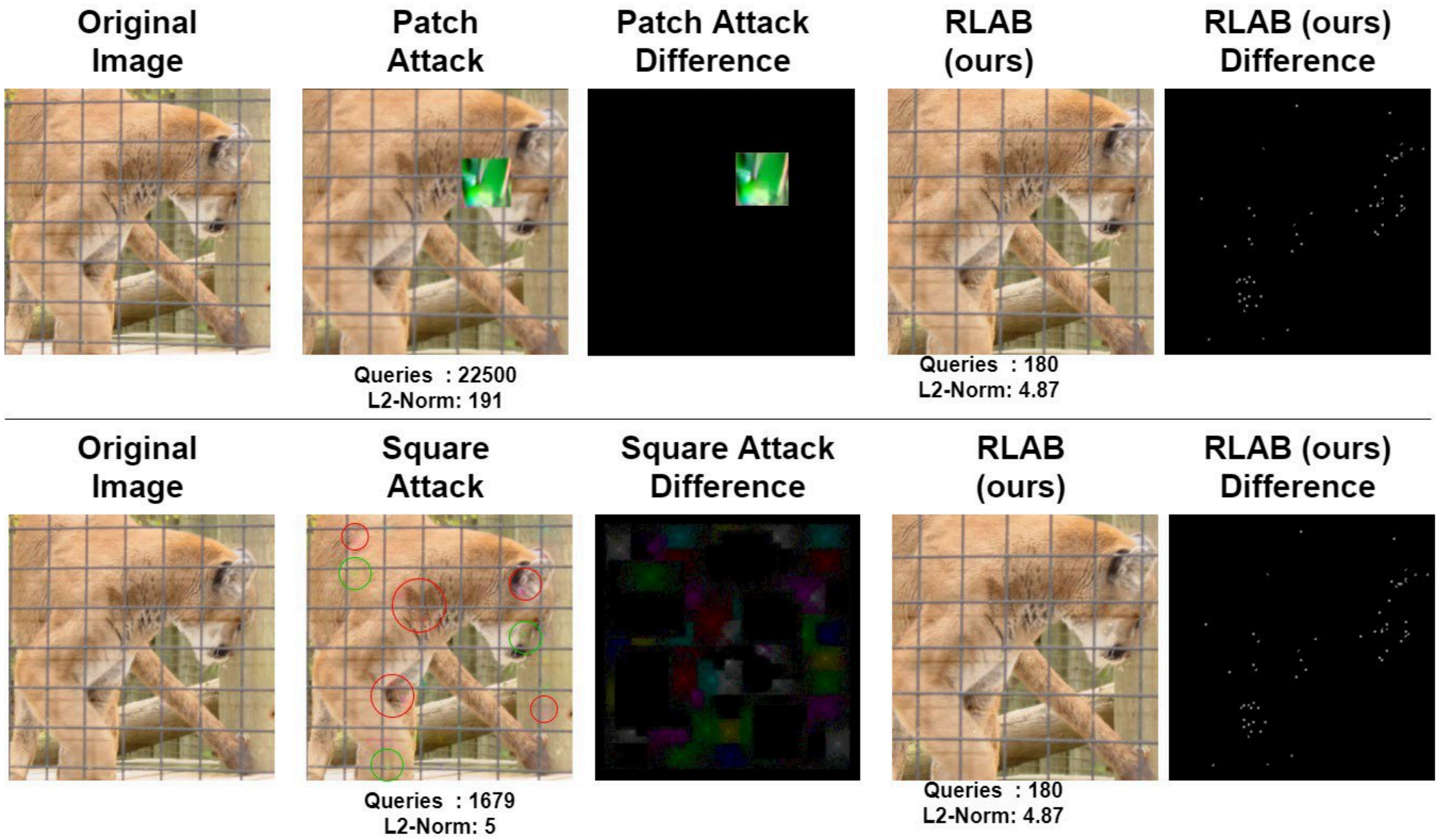
Measuring Robustness with Multiple Custom Distortions



Reinforcement Learning for Adversarial Black Box Attack (RLAB)



Comparison with SOTA: Natural Distortions relevant to Deployment



- Most state-of-the-art competitive solutions use unnatural modifications.
- In contrast, our proposed method preserves the true nature of the image with barely perceptible Gaussian noise.
- Patch Attack's distortion measured in L2-norm is significantly higher.
- The popular “Square Attack” has unnatural color blobs.

Comparison of RLAB to SOTA for Adversarial Attack

Table1: Comparing L_2 and average queries of the proposed method with competitors on the ResNet-50 model trained on Imagenet dataset.

Attack	AVG.Q	L_2	ASR
Q-Fool [26]	5000	7.52	-
NES (2018) [14]	1632	-	82.7
$Bandits_{TD}$(2018) [27]	5251	5	80.5
HopSkipJumpAttack [28]	1000	11.76	-
Subspace(2019) [29]	1078	-	94.4
P-RGF$_D$ (2019) [30]	270.5	-	99.3
LeBA (2020) [16]	178.7	-	99.9
Square (2020) [5]	401	5	99.8
SimBA-DCT (2021) [15]	1665	3.98	98.6
querynet (2021) [19]	-	5	-
AdvFlow (2021) [20]	746	-	96.7
EigenBA (2022) [17]	518	**3.6**	98
Pixle (2022) [18]	341	-	98
CG-Attack (2022)[21]	210	-	97.3
Patch Attack (2022) [16]	983	-	-
RLAB (ours)	**169**	4.01	**100%**

Table 2: Performance comparison of RLAB with State-of-the- art methods with Inception-V3, and VGG-16 on ImageNet dataset.

Method	Inception-v3		VGG-16	
	ASR %	AVG.Q	ASR %	AVG.Q
NES (2018) [14]	88.2	1726.2	84.8	1119
Bandits$_{TD}$ (2018) [27]	97.7	836.1	91.1	275.9
Subspace (2019) [29]	96.6	1035.8	96.2	1086
P-RGF$_D$ (2019) [30]	99	637.4	99.8	393.1
TIMI (2019) [32]	49	-	51.3	-
LeBA (2020) [16]	99.4	243.8	99.9	145.5
Sqr. Attack (2020) [5]	99.4	351.9	**100**	142.3
SimBA (2021) [15]	99.9	423.3	-	-
querynet (2021) [19]	-	518	-	-
AdvFlow (2021) [20]	99.3	694	95.5	1022
EigenBA (2022) [17]	95.7	968	-	-
Pixle (2022) [18]	-	-	99	519
CG-Attack (2022) [21]	**100**	139	99.4	**77**
Patch Attack [16]	-	-	-	-
RLAB(ours)	**100**	**132**	**100**	98

Table 3: Evaluation of the proposed method with competitors on ResNet-50 model trained on CIFAR-10 dataset

Attack	Avg. queries	S. Rate
SimBA-DCT [15]	353	100
AdvFlow [20]	841.4	100
MetaAttack [33]	363.2	100
AdvFlow [20]	598	97.2
CG-Attack [21]	81.6	100
EigenBA [17]	99	99.0
RLAB (ours)	**60**	**100**

Tab;e 4: Comparison between Dynamic policy driven patch selection and baseline for 'N'. **Dataset:** Ima- genet, **Model:** ResNet-50

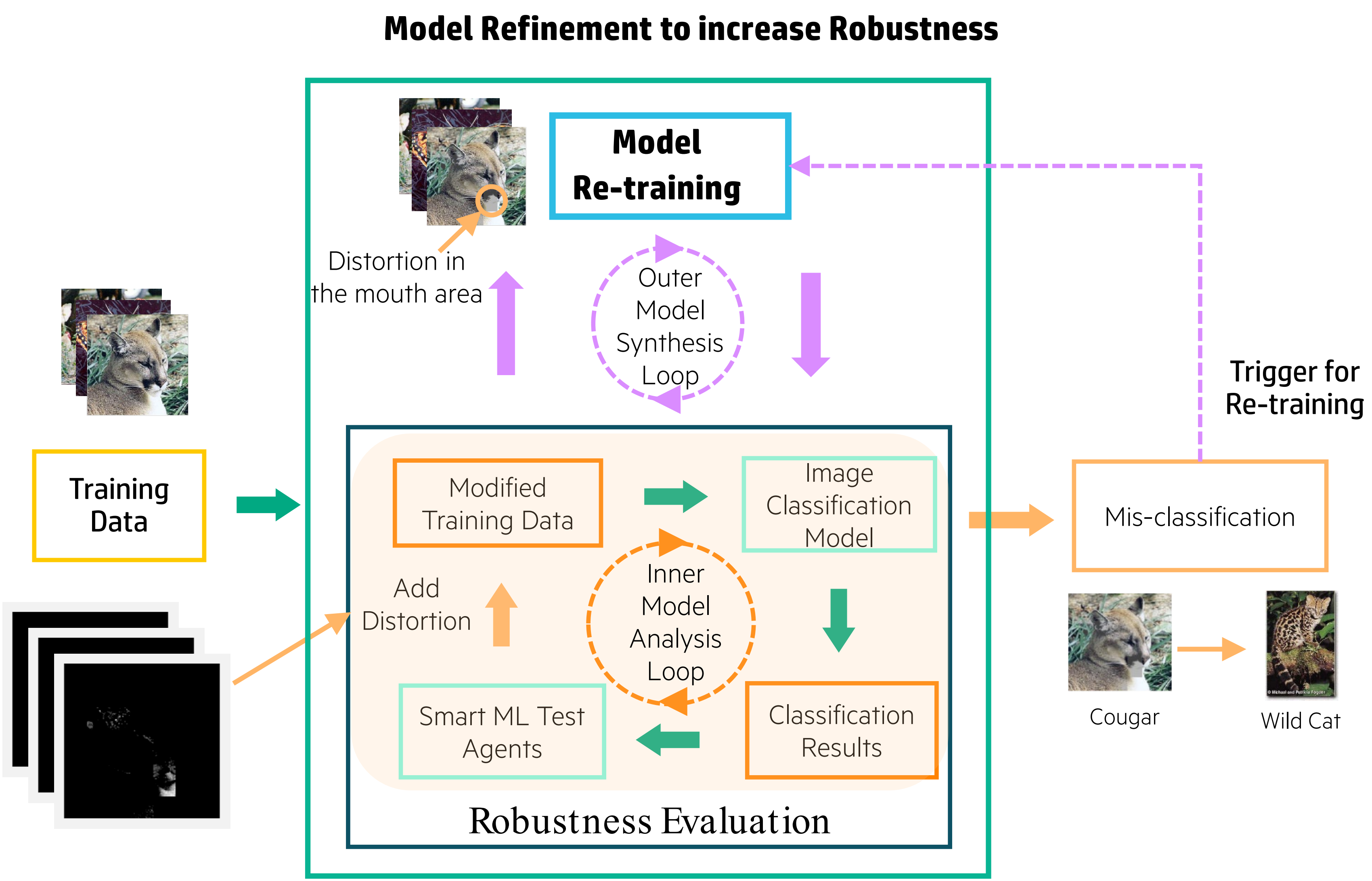
Approach	Average queries	Average L_2
Dynamic	169	4.03
Baseline	210	5.62

Table 5: Ablation study on different patch sizes **Dataset:** Imagenet, **Model:** ResNet-50

Patch Size	AVG. Q	Average L_2	ASR %
2x2	179	4.03	100
4x4	197	11.29	100
8x8	188	17.52	100
16x16	133	32.16	100
32x32	114	63.45	100

Robustness Enhancement

retrain and add robustness to cnn model



Results: Effectiveness of Re-training for Robustness with SOTA

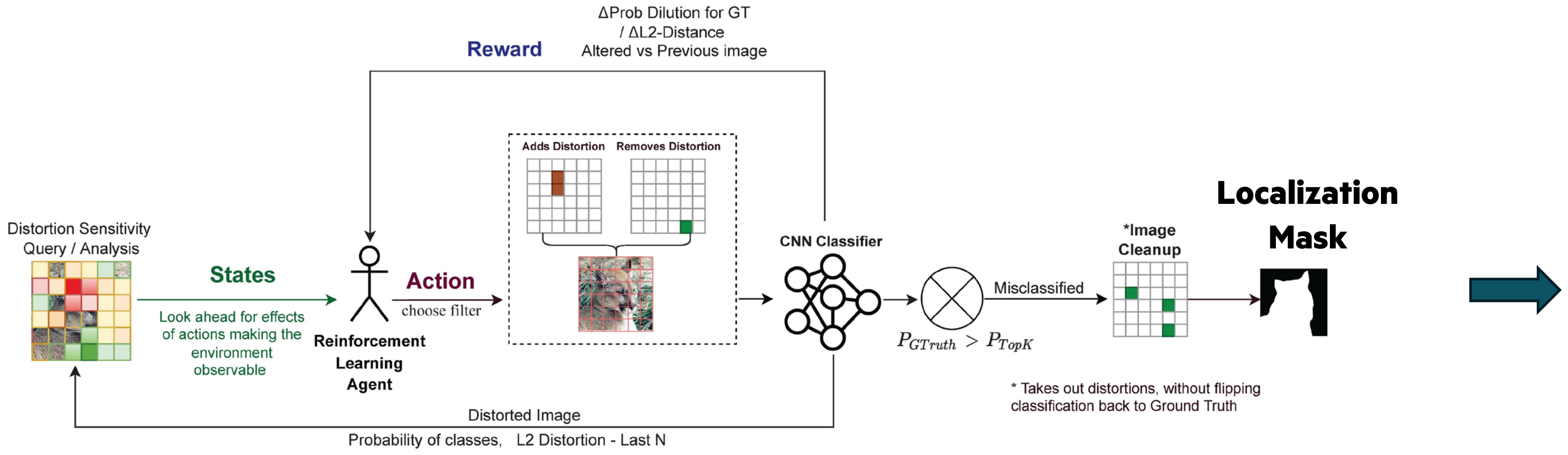
		Classification Error (%) with Re-training		
Dataset	Evaluated Against ↓	SimBA Adv Training	Square Adv Training	**RLAB Adv Training**
CIFAR-10	SimBA	-	99.80	**7.81**
CIFAR-10	Square	55.83	-	**51.61**
CIFAR-10	RLAB	88.60	97.80	-
Caltech-101	SimBA	-	2.15	**1.37**
Caltech-101	Square	32.77	-	**28.75**
Caltech-101	RLAB	75.00	75.04	-

Robustness comparison of our approach with Square and SimBA attack on ResNet-50 model with different datasets. Each attack was evaluated with the same 1000 samples generated from the test set.

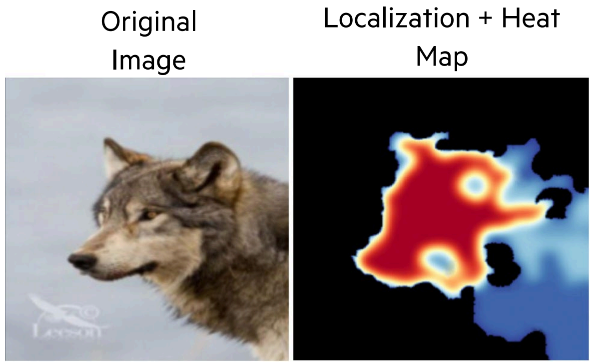
Visual Explanation

Visual Explanation: Localization Mask with Heat Map

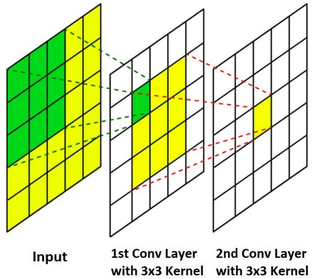
Important to understand the reasoning behind a model's predictions and to ensure that decisions are based on relevant features.



Reinforcement Learning based Architecture



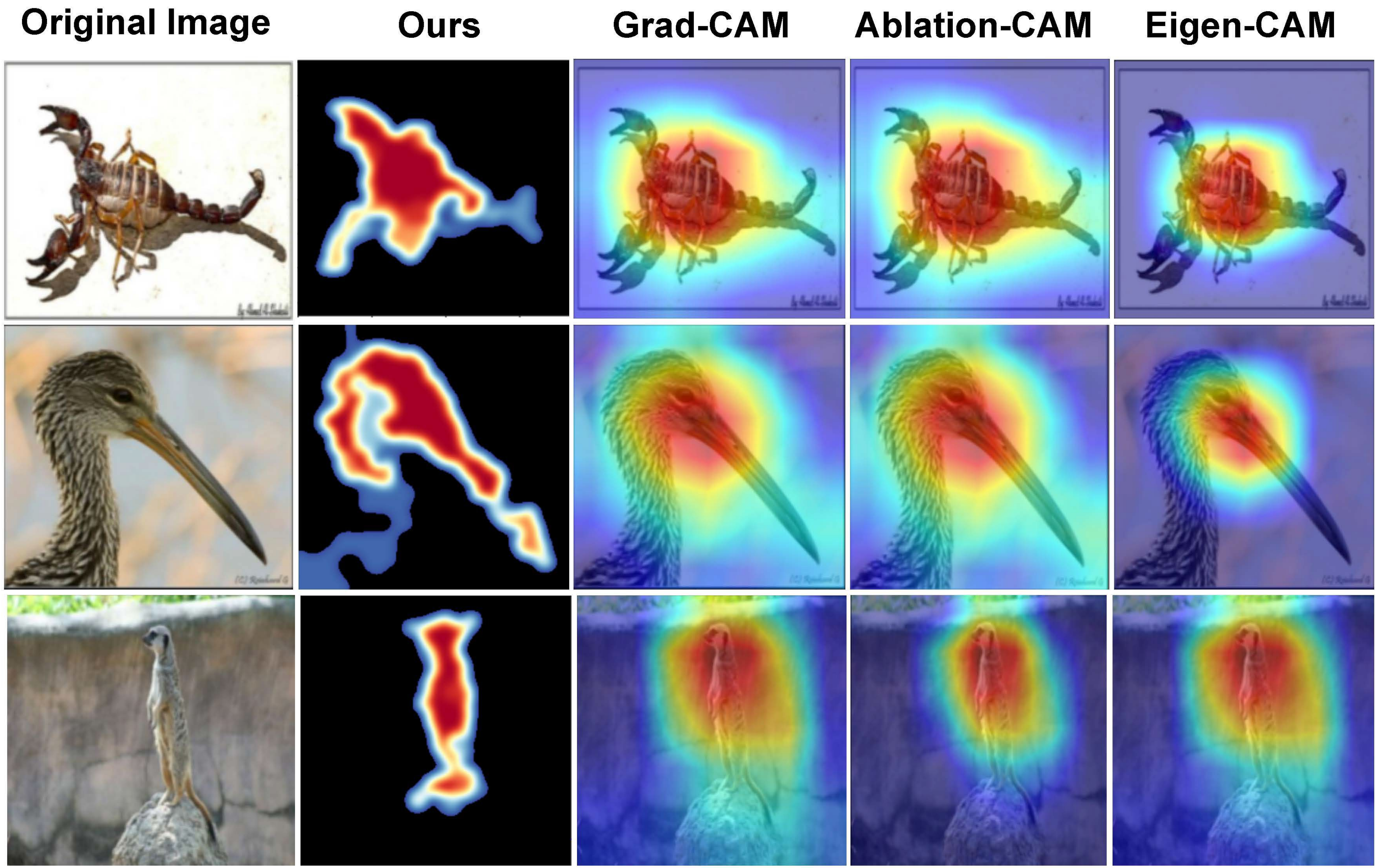
Heat Map



Receptive Field of Convolutional Neural Networks

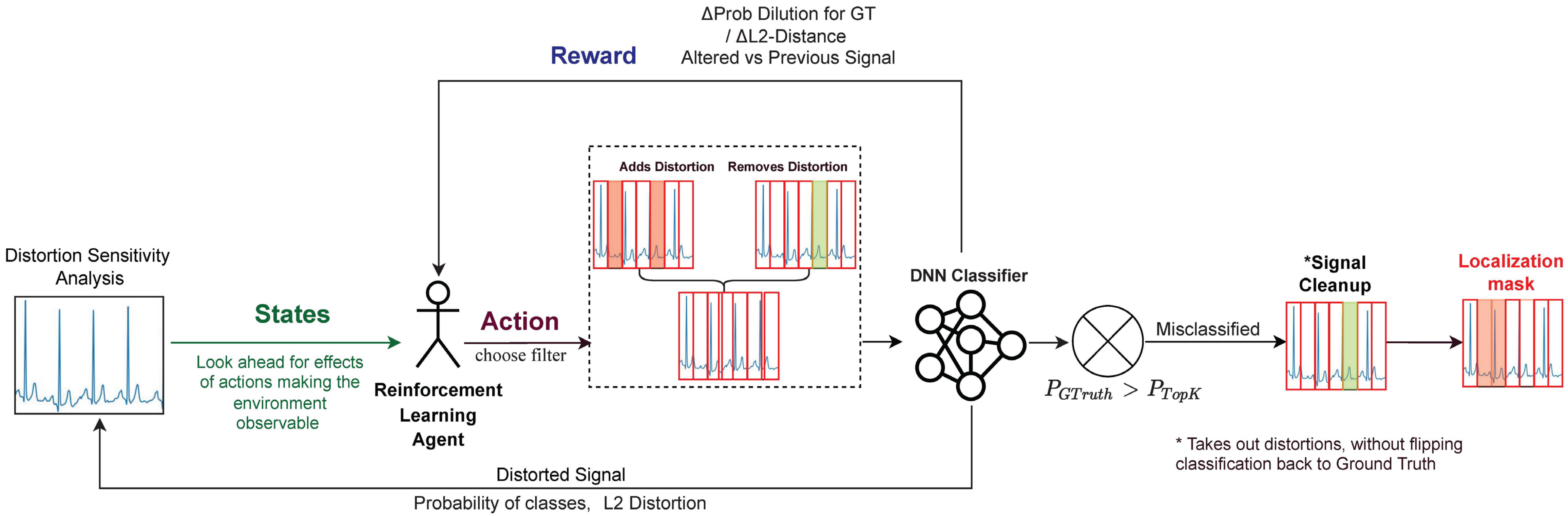
Gaussian Image Pyramid

Visual Explanation with RLAB



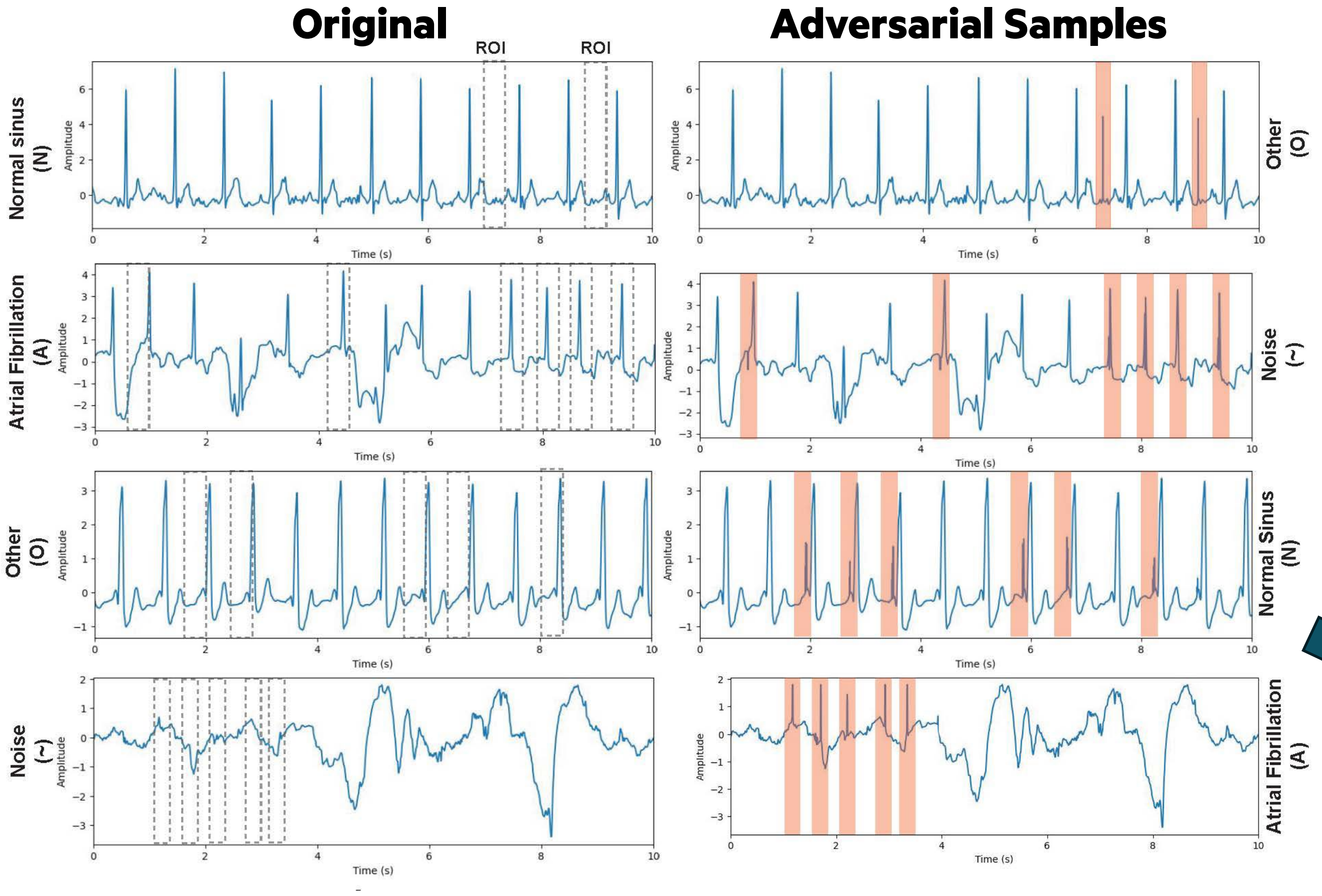
Signals and ECG Classification

ECG Arrhythmia Detection Models: Explainability and Robustness



* Takes out distortions, without flipping classification back to Ground Truth

ECG Arrhythmia Detection Models: Robustness



Robustness goes UP after adversarial training

Noise Type	Seg-length (ms)	AVG Q	AVG L_2	ASR
Motion Artifact	10	14.54	3.40	100%
	16.67	16.96	3.70	100%
	33.33	30.86	4.81	100%
	50	17.90	5.05	100%
	(10, 16.67)	20.88	3.20	100%
	(16.67, 33.33)	28.02	3.46	100%
Detached Device	3.33	4.58	8.37	100%
	6.66	2.36	11.27	100%
	10	3.12	12.94	100%
	(3.33, 6.66)	2.26	10.71	100%
	(6.66, 10)	2.48	11.68	100%

Avg L_2 and Q represent the average L2 and queries over all samples, respectively.
* ms stands for milliseconds.

Average L2 and Queries as a **measure of Robustness** of the ResNet model trained on PhysioNet Challenge 2017 dataset, **WITHOUT the Adversarial ECG signals.**
We see lower metrics.

Trained with Adversarial samples

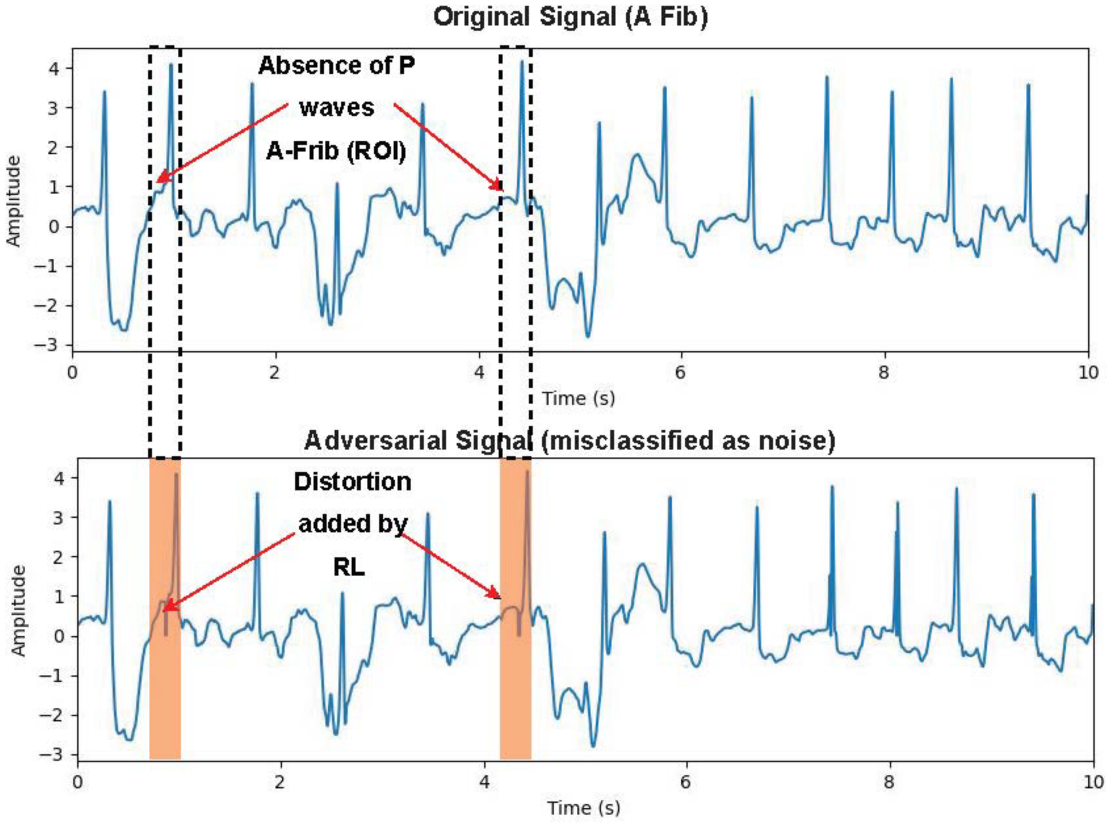
Noise Type	Seg-length (ms)	AVG Q	AVG L_2	ASR
Motion Artifact	10	85.10↑	3.90	100%
	16.67	61.34↑	4.41	100%
	33.33	33.86↑	5.32	100%
	50	27.70↑	5.54	100%
	(10, 16.67)	40.30↑	4.09	100%
	(16.67, 33.33)	42.74↑	4.78	100%
Detached Device	3.33	5.22↑	14.28	100%
	6.66	4.28↑	18.24	100%
	10	3.14↑	18.67	100%
	(3.33, 6.66)	4.78↑	16.36	100%
	(6.66, 10)	3.30↑	18.16	100%

Avg L_2 and Q represent the average L2 and queries over all samples, respectively.
* ms stands for milliseconds.

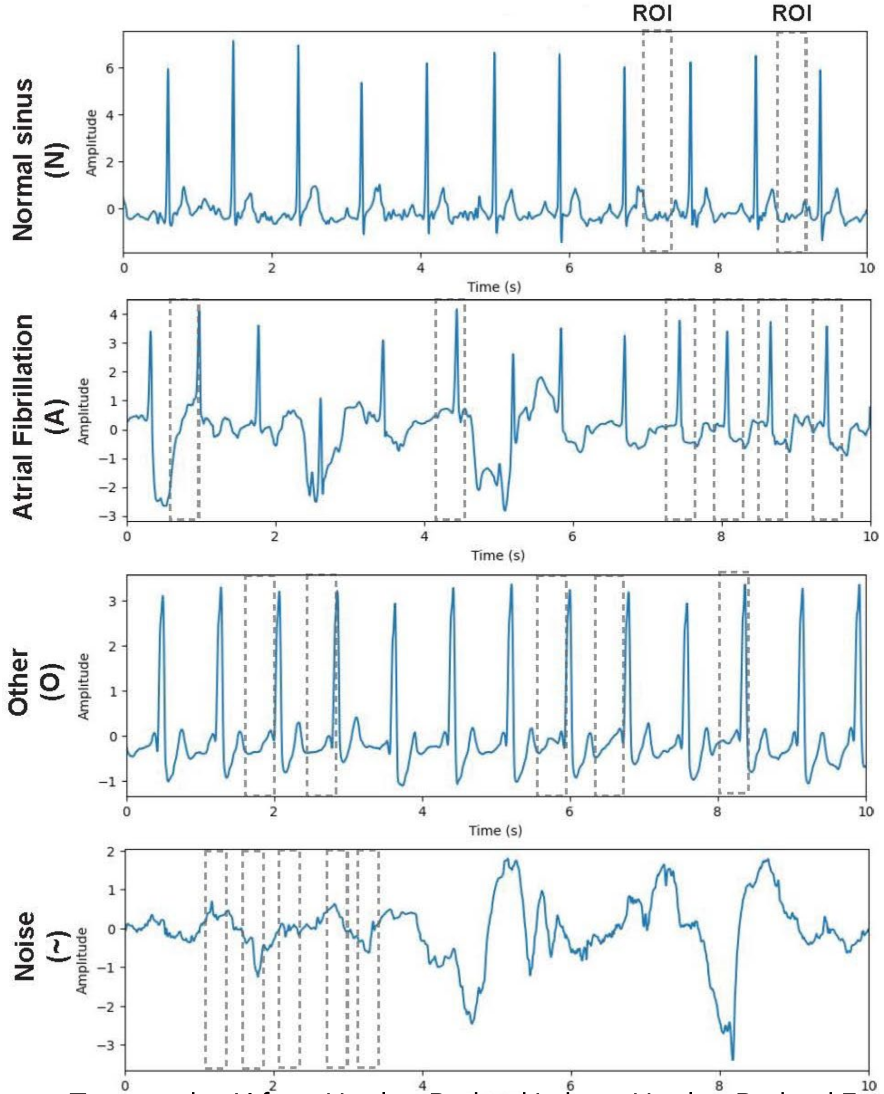
Average L2 and Queries as a **measure of Robustness** of the ResNet model trained on PhysioNet Challenge 2017 dataset, **WITH the Adversarial ECG signals.**
We see higher metrics indicating improved Robustness.

ECG Arrhythmia Detection Models: Explainability

Visual Explanation

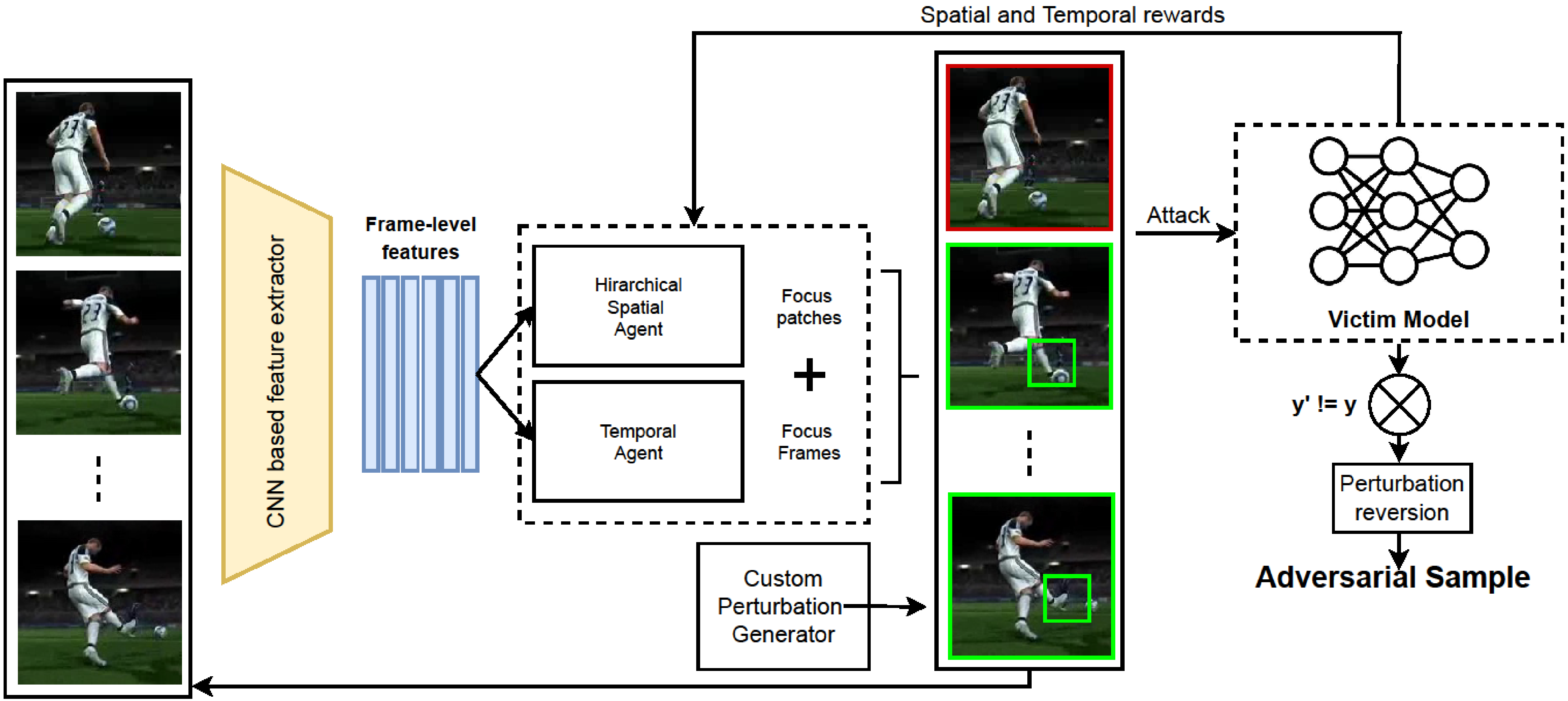


Marking the ROI

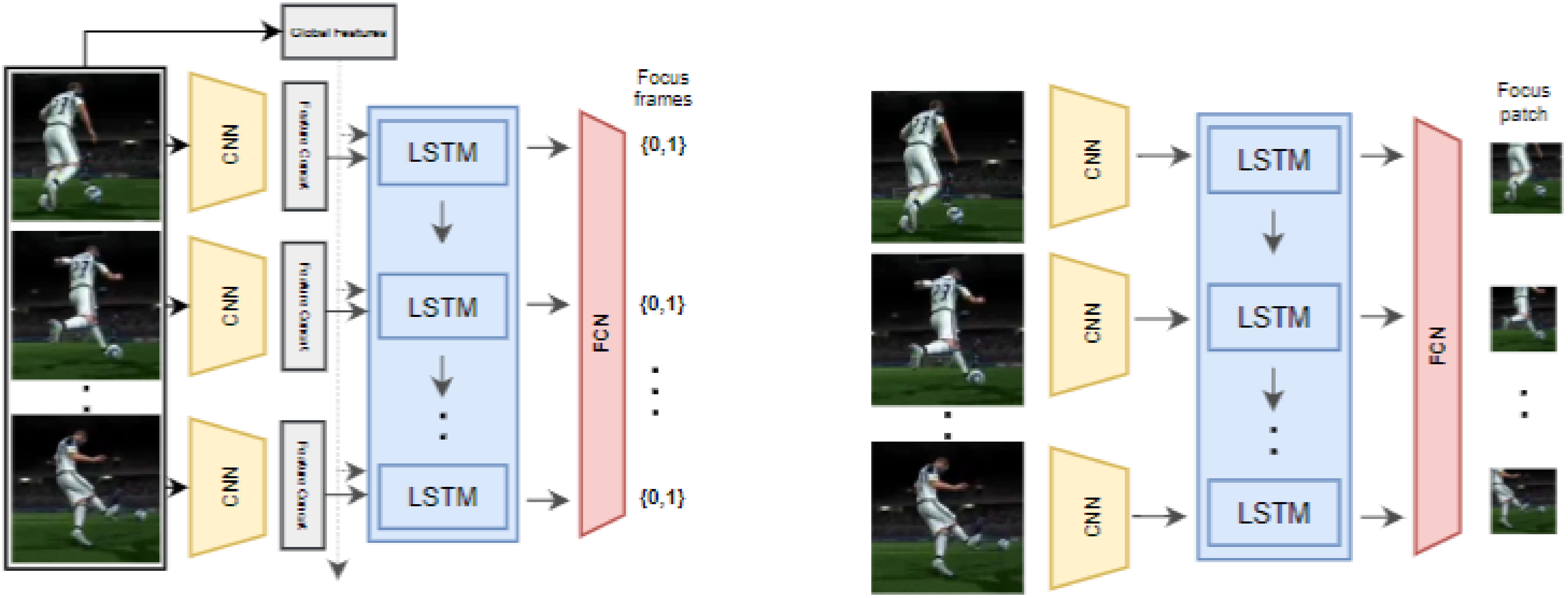


Video Classification

Video Classification: Adversarial Attack for Robustness Evaluation with RL



Video Classification: Adversarial Attack for Robustness Evaluation with RL



(a) Temporal agent (b) Spatial agent

Video Classification: Adversarial Attack for Robustness Evaluation with RL

Original video

(OURS)
Adversarial video

(SOTA)
Adversarial video

Adversarial Attack on Videos for Robustness Evaluation with RL

EFFECTIVENESS METRICS

THREAT MODELS

	Attack Methods	HMDB-51			UCF-101		
		MAP ↓	QN ↓	SR ↑	MAP ↓	QN ↓	SR ↑
TSM [8]	Heuristic attack	5.043	10385	58	5.956	10657	40
	Motion-sampler attack	7.229	3911	90	7.237	5187	83
	GEO-TRAP attack	5.919	3164	92	5.865	3782	88
	RLSB attack	5.323	5950	82	4.823	4898	87
	AstFocus attack	3.411	1529	**100**	3.355	1138	**96**
	Ours (GB)	**0.835**	921	83	**0.735**	863	85
	Ours (DP)	1.824	**600**	90	2.491	**306**	94
TSN [9]	Heuristic attack	5.395	10146	58	5.265	9135	51
	Motion-sampler attack	7.275	3667	88	6.895	4744	78
	GEO-TRAP attack	5.192	3392	88	5.472	3782	75
	RLSB attack	5.312	4217	92	5.238	3504	93
	AstFocus attack	3.520	2198	96	3.265	2015	**99**
	Ours (GB)	**0.677**	8433	96	**0.676**	3243	91
	Ours (DP)	2.373	**238**	**98**	2.417	**373**	96
C3D [10]	Heuristic attack	4.838	10534	42	6.295	14160	30
	Motion-sampler attack	7.035	6491	68	6.153	8132	62
	GEO-TRAP attack	5.666	5082	84	5.877	7045	75
	RLSB attack	4.688	7279	62	5.326	6568	68
	AstFocus attack	3.835	3628	92	4.015	4224	90
	Ours (GB)	**0.835**	8710	95	**2.258**	**1783**	90
	Ours (DP)	1.824	378		6.351	2602	81

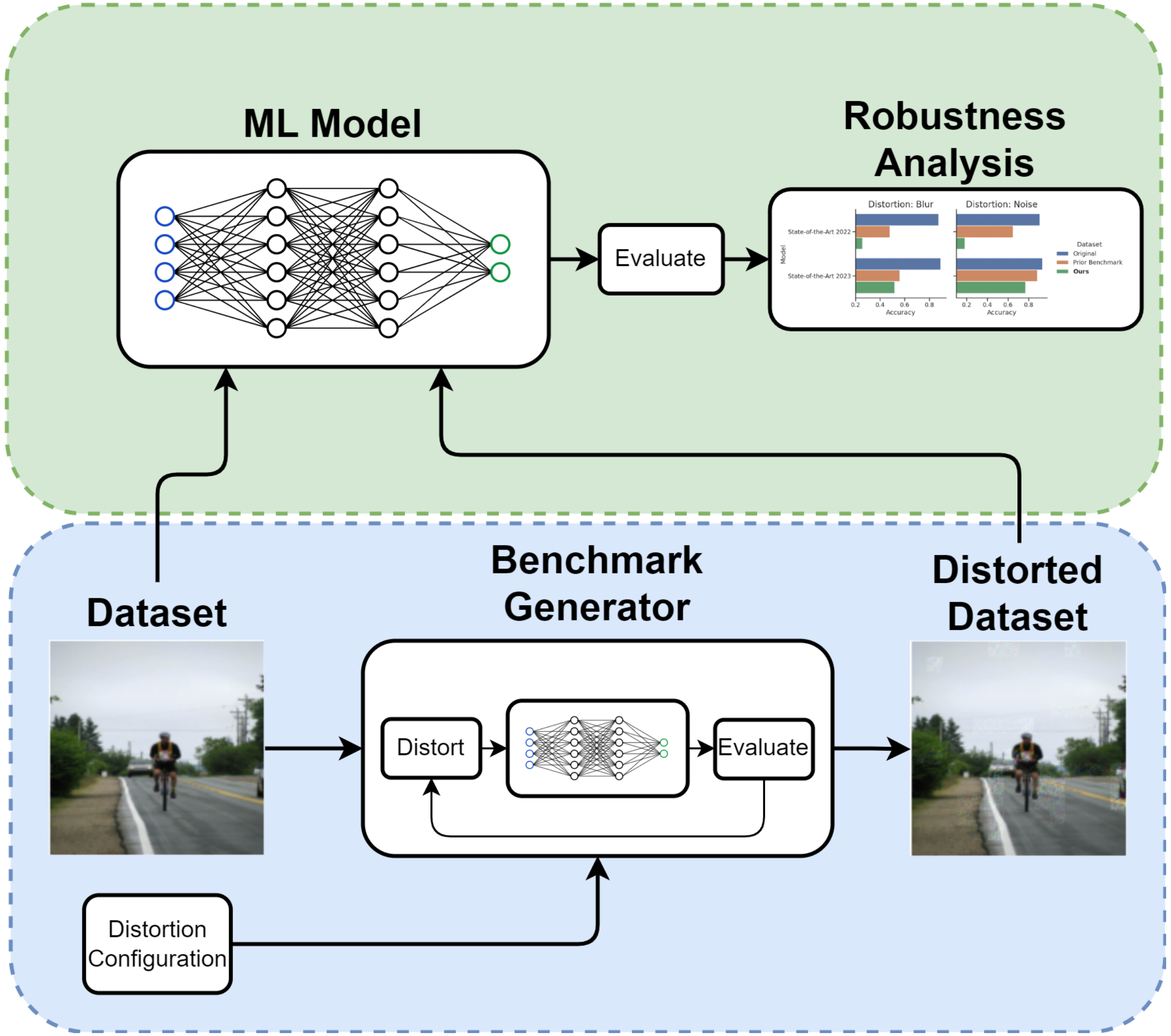
Benchmark and Summary

Generating Benchmark Dataset for Image Classification with RobustBench

ROBUSTBENCH

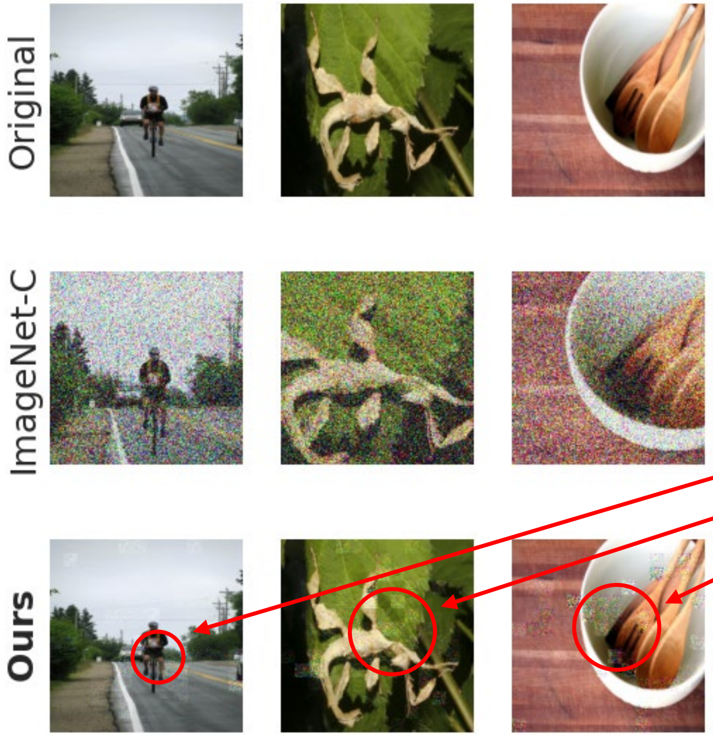
A standardized benchmark for adversarial robustness

- Highly reputable resource that tracks the state-of-the-art in robustness methods
- Robustness methods continuously evaluated against select challenging benchmarks
- Our generated benchmarks are experimentally demonstrated to be more difficult than those comprising RobustBench for state-of-the-art robustness methods

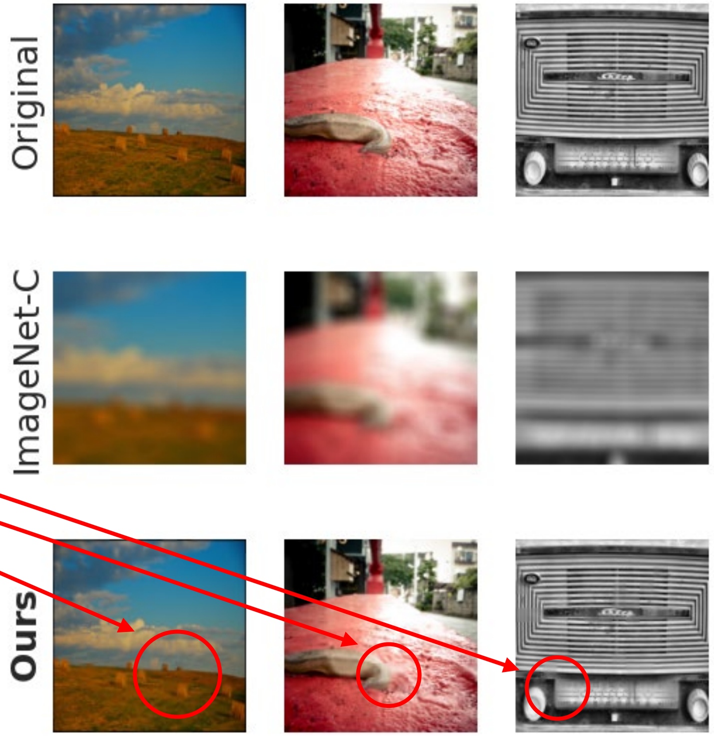


Superior Distortion Generation for Benchmark Dataset for Image Classification

- Higher fidelity & clarity of original images
- Our distortions are better for measuring robustness
- Our distortions enable effective auditing of model failures



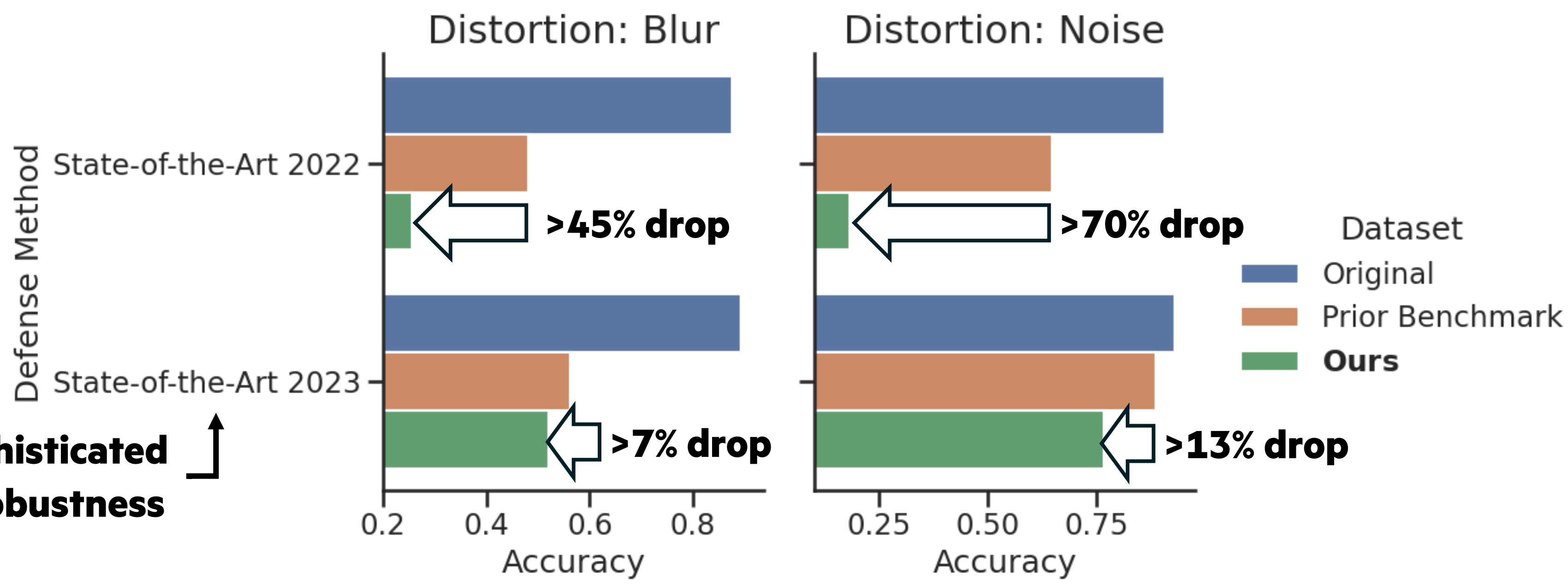
Distortions help identify points of failure



Over 3x less distortion needed!

Superior Distortion Generation for Benchmark Dataset for Image Classification

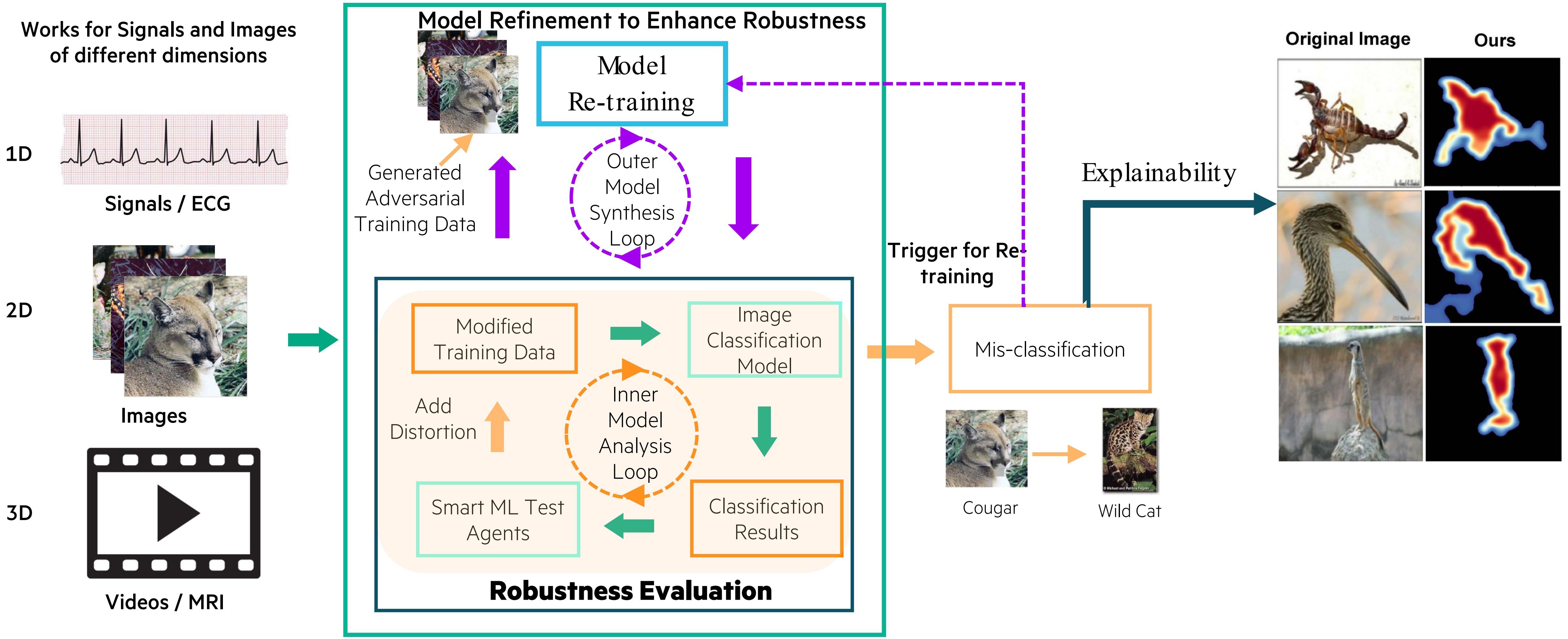
- Our benchmark generator more effectively identifies robustness issues than prior benchmarks
- Effective for multiple types of distortions and levels of distortion
- Experiments show the state-of-the-art is still susceptible to real-world distortions



Even the most sophisticated defenses exhibit robustness limitations

Reinforcement Learning for Robustness Evaluation, Enhancement, and Explanation

Important to understand the reasoning behind a model's predictions and to ensure that decisions are based on relevant features.



Publications

- Robustness With Query-Efficient Adversarial Attack Using Reinforcement Learning:
  - CVPR 2023 workshop proceedings
  - https://tinyurl.com/y2yvckp5
- RL-cam: Visual explanations for convolutional networks using reinforcement learning:
  - CVPR 2023 workshop proceedings
  - https://tinyurl.com/2skw93pw
- Benchmark Generation Framework With Customizable Distortions for Image Classifier Robustness:
  - WACV 2024 proceedings
  - https://tinyurl.com/38hwsst8
- RTDK-BO: High Dimensional Bayesian Optimization with Reinforced Transformer Deep kernels:
  - IEEE CASE 2023 proceedings
  - https://ieeexplore.ieee.org/abstract/document/10260520

Tune in for our work on Trust for LLMs

- ❑ Evaluating LLMs for Trust
- ❑ Refining LLMs for Trust

Thank you