Defense Mechanism Against Adversarial Attacks Using Density-based Representation of Images

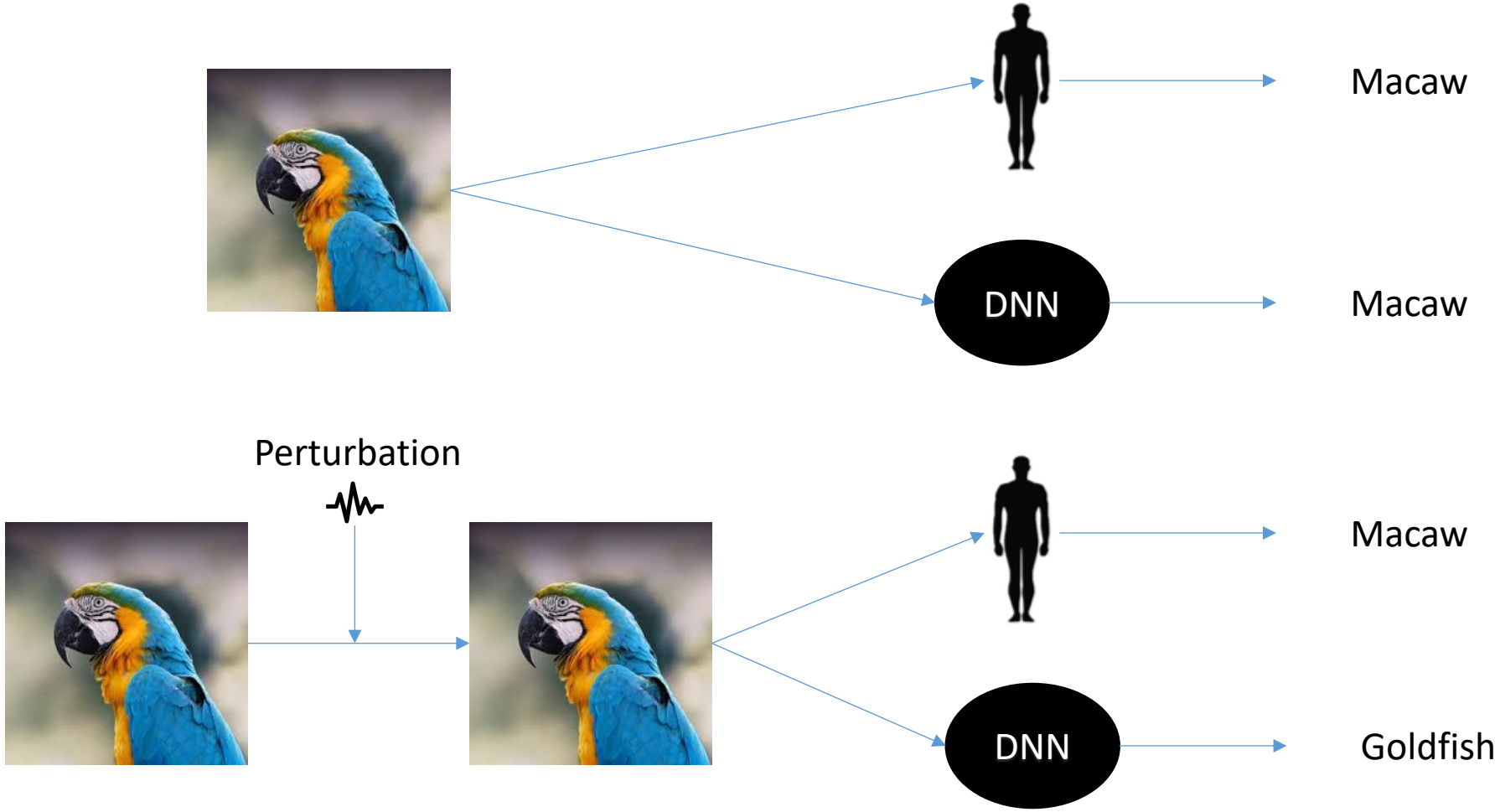
Yen-Ting Huang, Wen-Hung Liao, and Chen-Wei Huang

Dept. of Computer Science, National Chengchi University, TAIWAN

Pervasive Artificial Intelligence Research (PAIR) Labs, TAIWAN

ICPR 2020

What is Adversarial Attack?



Change of Decision Boundaries by Input Recharacterization

$$\mathcal{F}(\mathcal{C}(\mathcal{X}+\epsilon);\delta)=\mathcal{F}(\mathcal{X};\theta)$$

$$\mathcal{F}(\mathcal{R}(\mathcal{C}(\mathcal{X}+\epsilon));\theta)=\mathcal{F}(\mathcal{X};\theta)$$

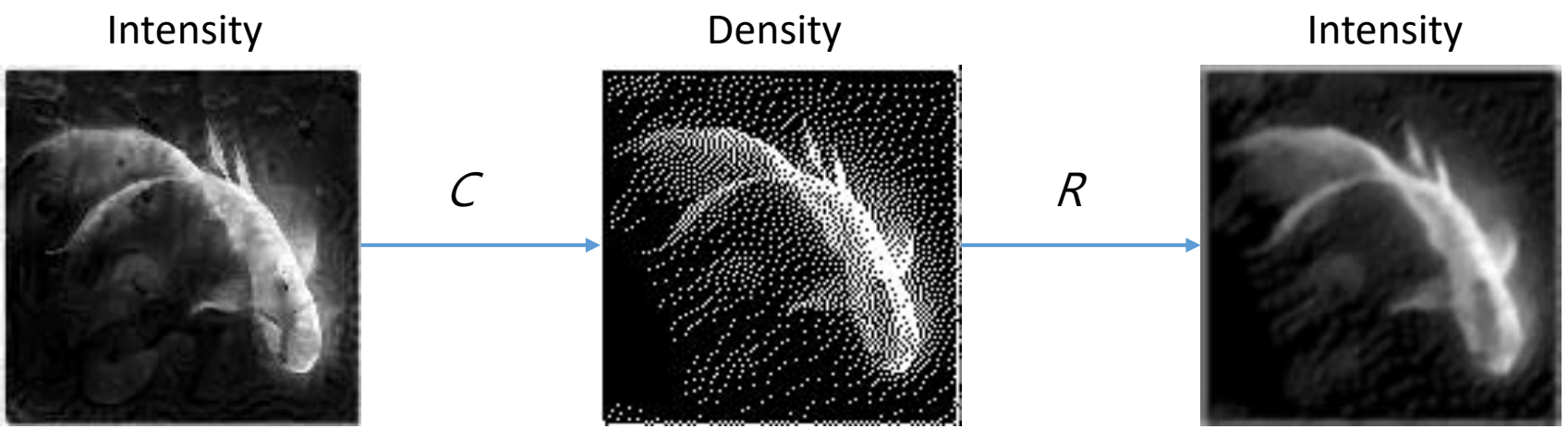
$$\mathcal{F}(\mathcal{R}(\mathcal{C}(\mathcal{X}+\epsilon));\hat{\theta})=\mathcal{F}(\mathcal{X};\theta)$$

$\mathcal{C}$: Forward Conversion
$\mathcal{R}$: Backward Reconstruction

ICPR 2020

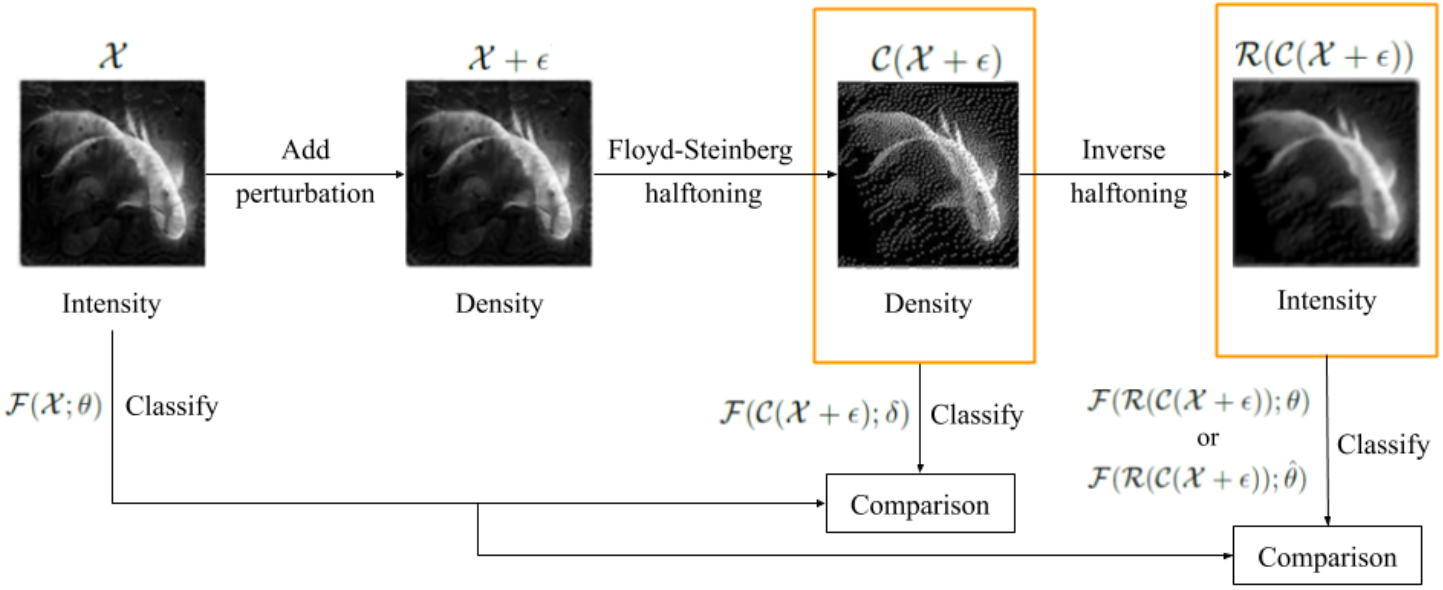
Proposed Defense Mechanism (1/2)

- Objective: perform effective defense on adversarial examples without incurring excessive computing costs.
- How?
  - Domain transformation with halftoning



Proposed Defense Mechanism (2/2)

- Three hypotheses need to be explored:
  - The transferability of adversarial examples between intensity-based and density- based domain
  - The attackability under the density-based representation
  - The feasibility of invalidating attacks with two-stage input recharacterization



Experimental Results - Transferability of Adversarial Examples

TABLE I: Performance of different input transform schemes

Attack	Accuracy Defense	Cropping and Rescaling	TVM	Grayscale	Halftone	Hybrid (intensity)	Hybrid (density)
Baseline	Top-1	56.98	59.13	62.0	61.1	66.01	60.06
	Top-5	77.23	78.56	76.5	80.4	85.14	82.31
FGSM	Top-1	43.65	36.46	12.0	57.78	59.93	59.40
	Top-5	69.96	69.07	31.4	80.34	81.13	80.97
I-FGSM	Top-1	45.10	43.15	10.1	52.01	34.93	52.51
	Top-5	72.52	70.21	17.4	78.35	69.31	78.77
PGD	Top-1	45.68	39.13	10.1	57.23	48.69	58.03
	Top-5	73.26	67.29	17.4	80.91	77.46	81.56

ICPR 2020

Experimental Results - Attackability under the Density-based Representation (1)

- Launching Attacks in the Halftone Domain
  1) Global Adversarial Perturbations: PGD Attack

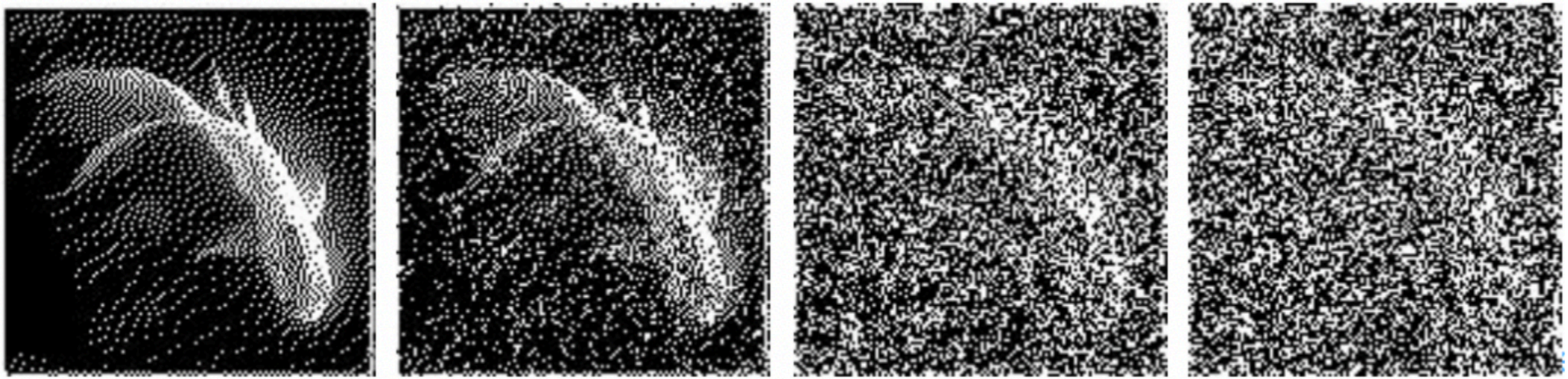
Original PGD Attack Original PGD Attack

Polluted samples artifact becomes easily detectable by human observer

ICPR 2020

Experimental Results - Attackability under the Density-based Representation (2)

- Launching Attacks in the Halftone Domain
  2) Local Adversarial Perturbations: JSMA Attack



Original | 10% modified | 20% modified | 30% modified

Polluted samples artifact becomes easily detectable by human observer

Experimental Results - Feasibility of Invalidating Attacks with Two-stage Input Recharacterization

TABLE II: One-way vs. two-stage transformation for defending adversarial attacks

Attack	Accuracy Defense	Grayscale (Original)	Grayscale (Inverse)	Hybrid (Original)	Hybrid (Inverse)
Baseline	Top-1	62.0	12.0	66.01	26.32
	Top-5	76.5	27.9	85.14	46.64
FGSM	Top-1	12.0	9.8	59.93	23.11
	Top-5	31.4	24.1	81.13	42.26
I-FGSM	Top-1	10.1	8.30	34.93	20.63
	Top-5	17.4	22.05	69.31	40.23
PGD	Top-1	10.1	9.33	48.69	21.57
	Top-5	17.4	23.41	77.46	41.50

ICPR 2020

Short Summary

- Answer to the hypotheses
  - (O) Transferability : Exhibits resistance against perturbations added to RGB images
  - (O) Attackability : Adversarial attacks (e.g., PGD, I-FGSM or JSMA) in density-based representation easily detected
  - (X) Feasibility : Two-way image recharacterization would result in excessive loss of texture.

Conclusion

- A lightweight procedure known as input recharacterization to counter adversarial attacks has been proposed in this research.
- We have generalized the input transform scheme for adversarial defense into input recharacterization and investigated its efficacy under different settings
- We demonstrated that input transform based method can exhibit resistance to adversarial examples only through model retraining

ICPR 2020

ICPR 2020