Defending Against Universal Attacks Through Selective Feature Regeneration

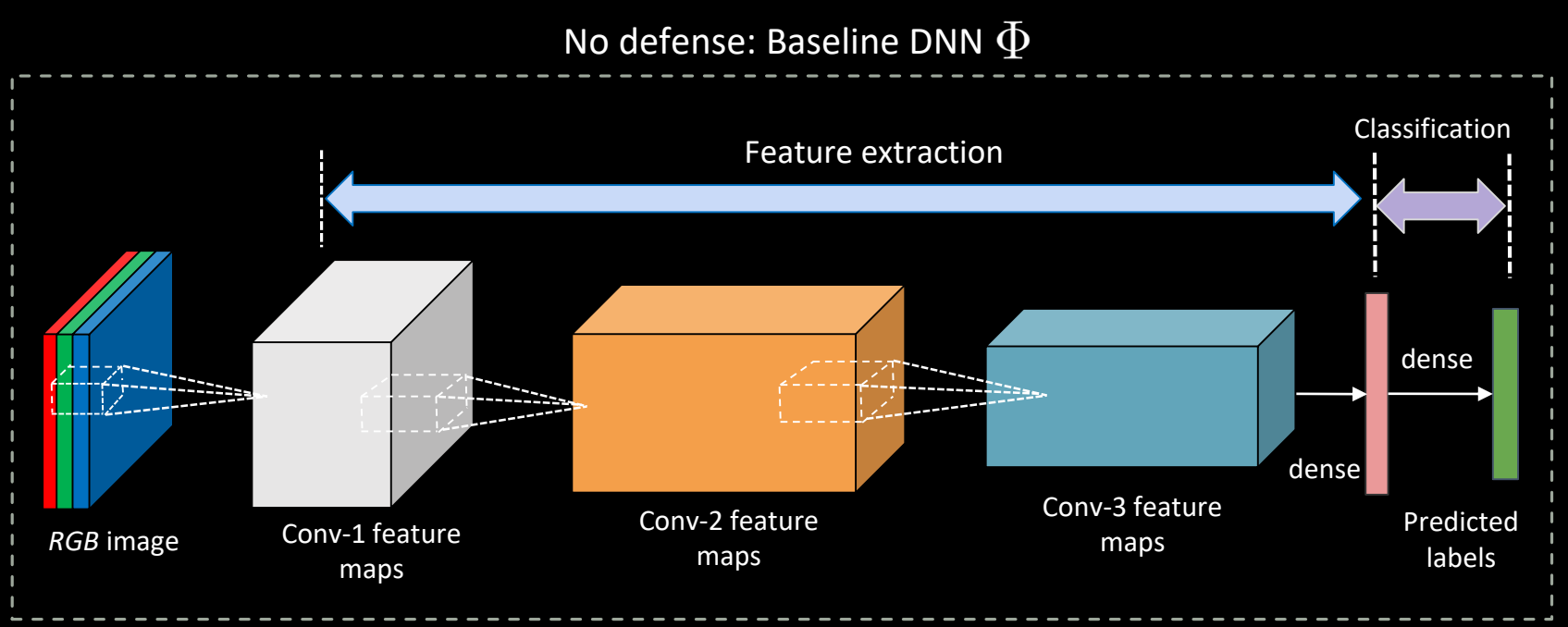
Tejas Borkar[1], Felix Heide[2,3], Lina Karam[1,4]

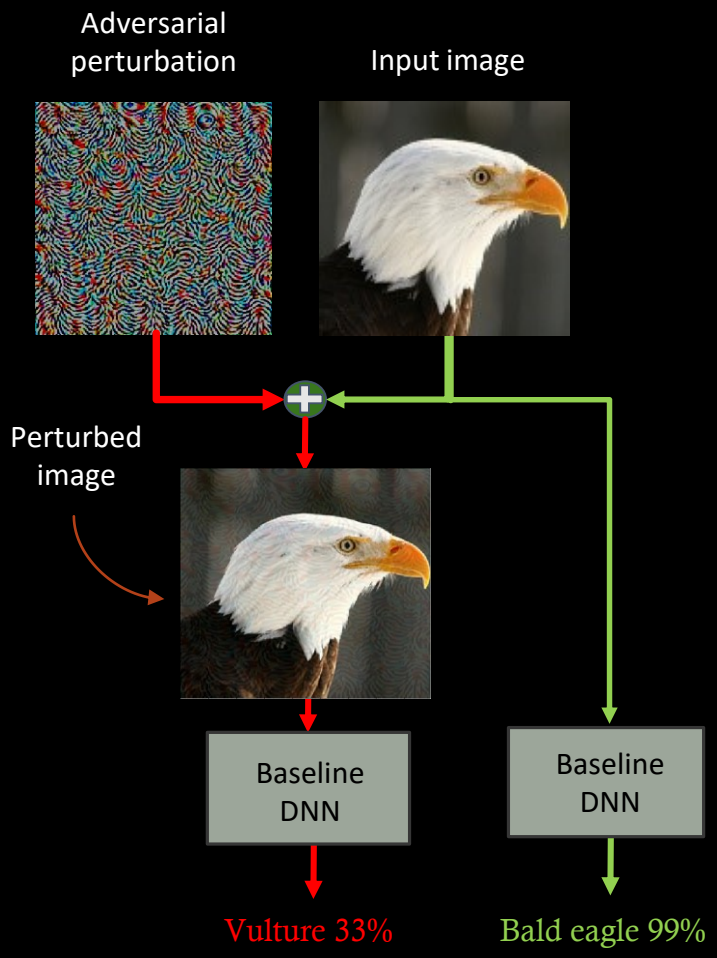
[1]Arizona State University [2]Princeton University [3]Algolux [4]Lebanese American University

CVPR 2020

Universal Adversarial Attacks

- Image agnostic and transferable across networks

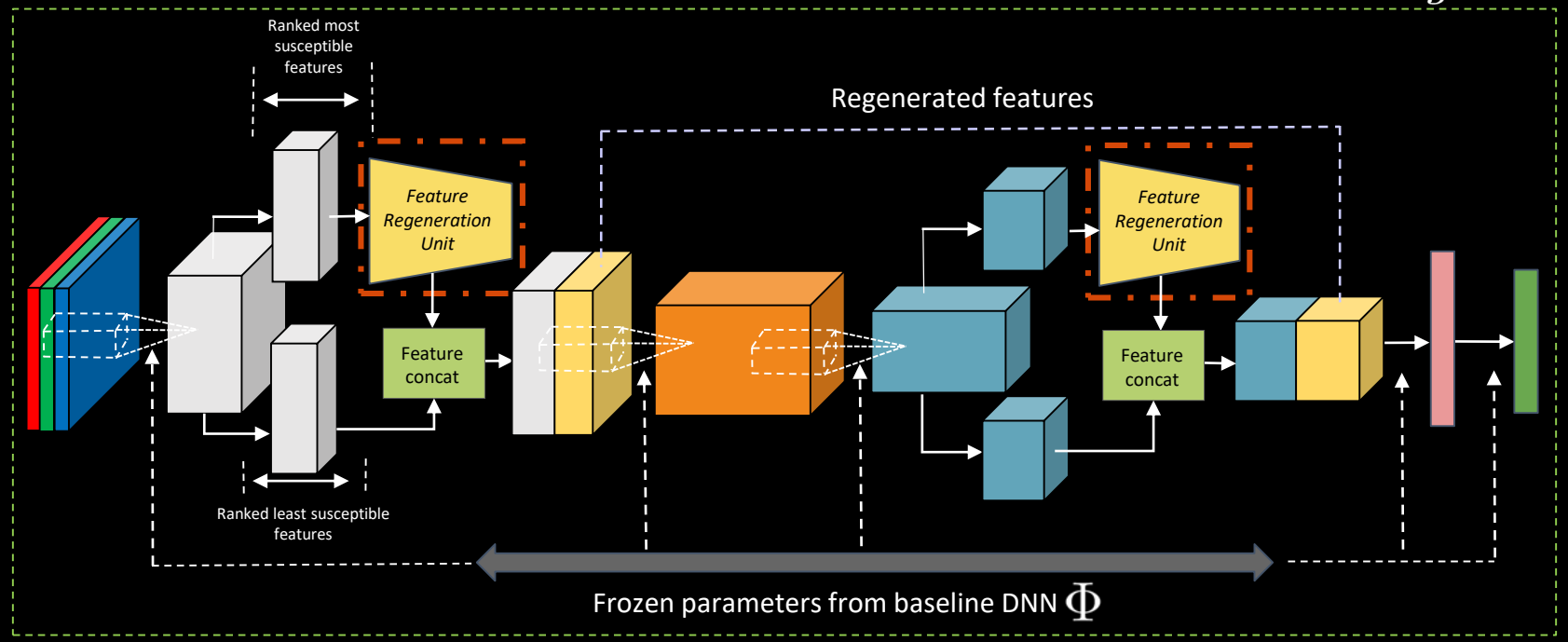


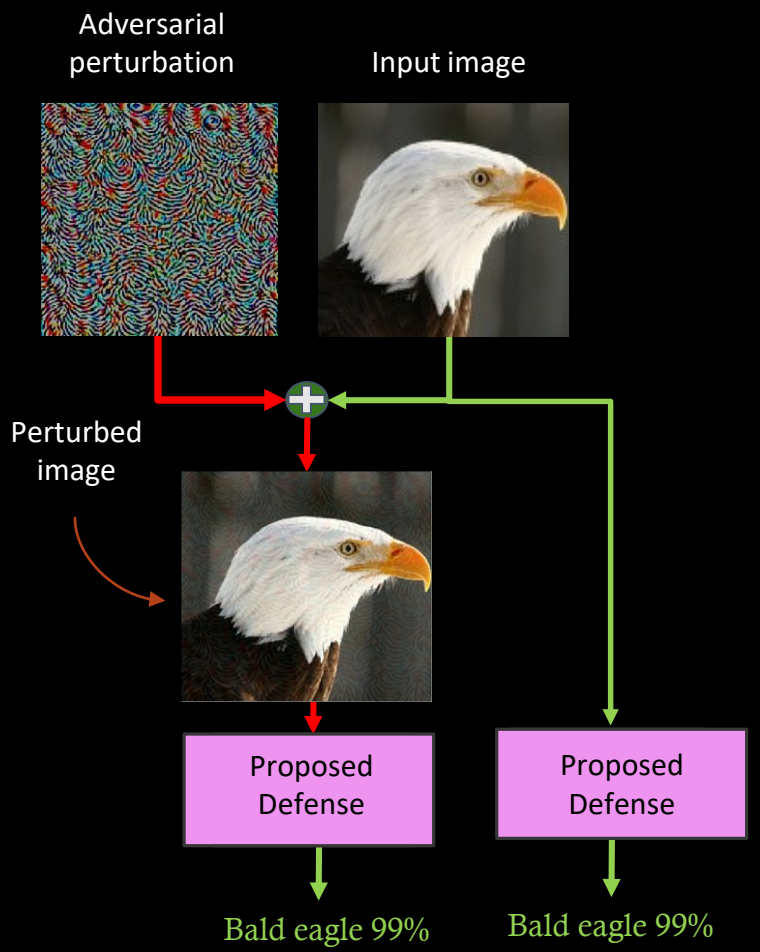


Defending against Universal Adversarial Attacks

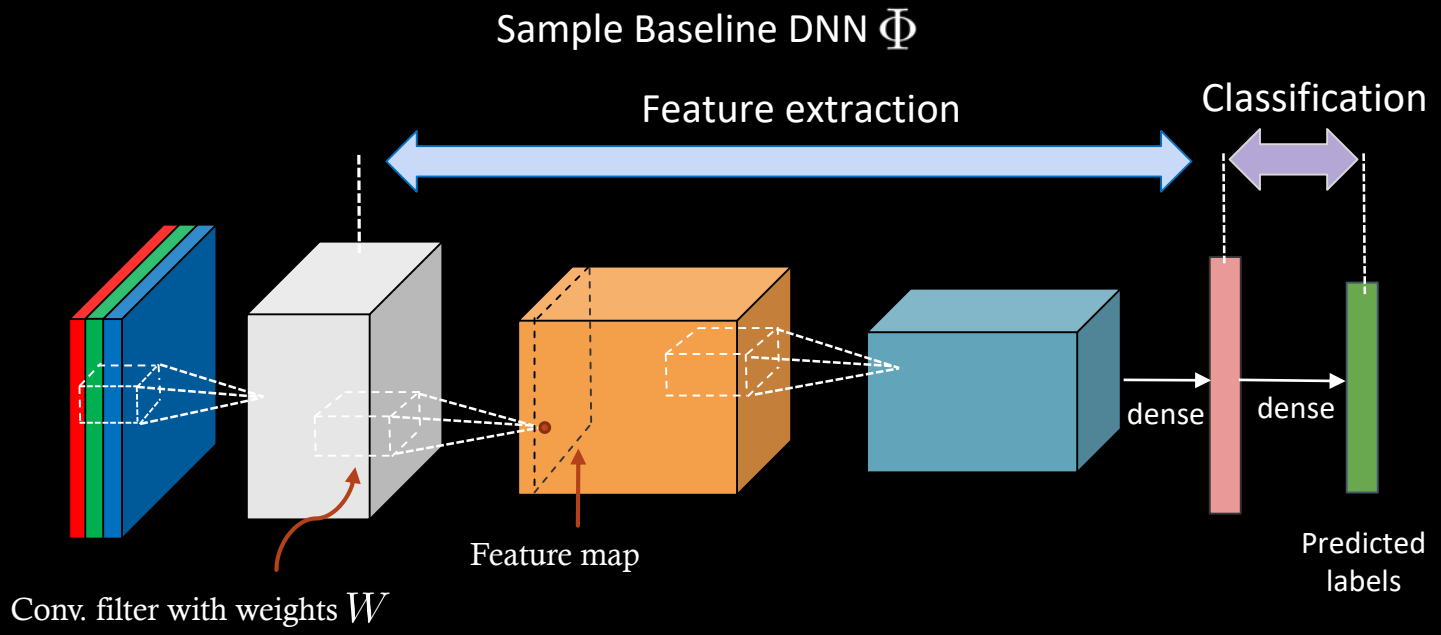
- Selective feature regeneration effectively restores robustness

Proposed defense: Baseline DNN with resilient feature regeneration Φ_{reg}





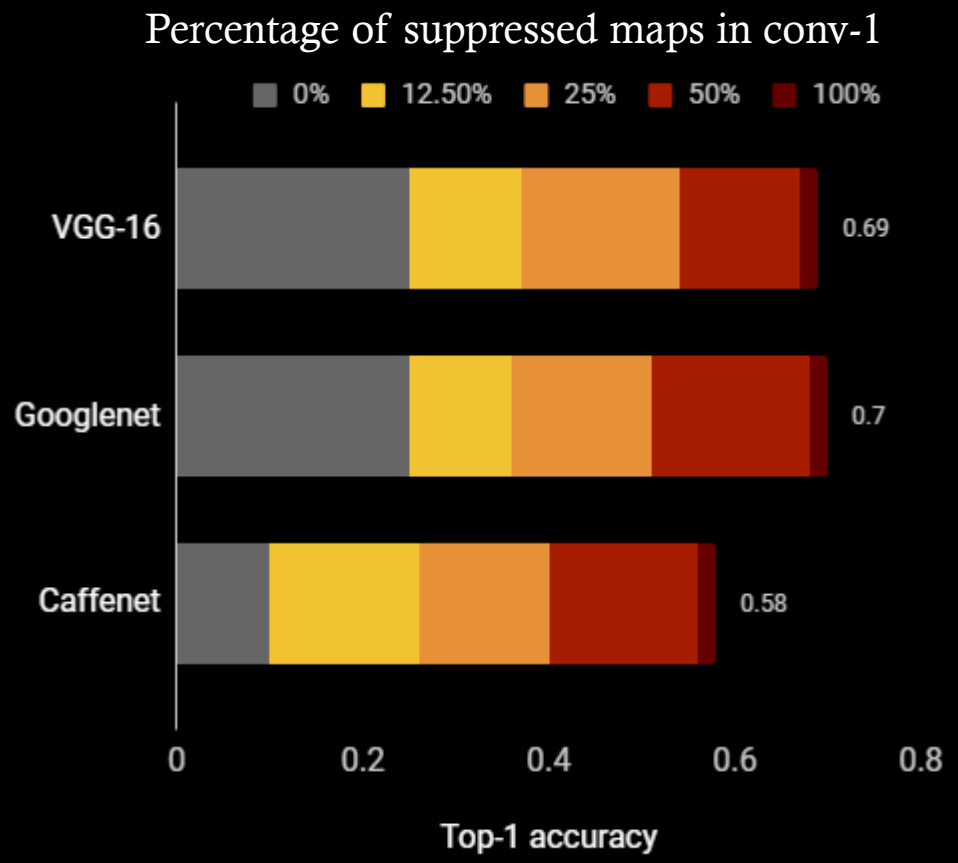
Ranking CNN Filters Based on Noise Susceptibility



We show:

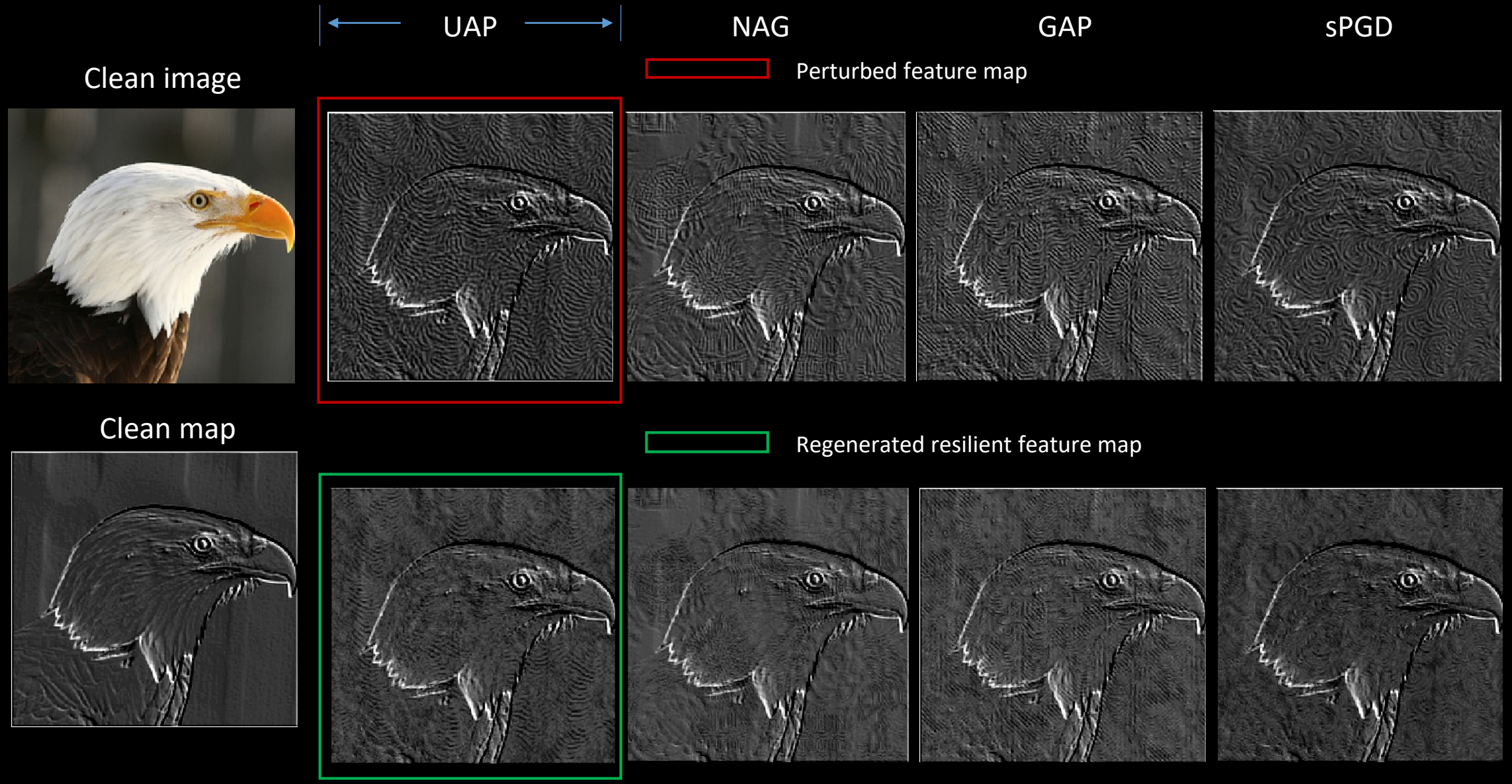
- Max perturbation level induced in feature map $\propto$ ℓ_1-norm of the filter weight ($\|W\|_1$)

Suppressing perturbations in ranked filters' output maps



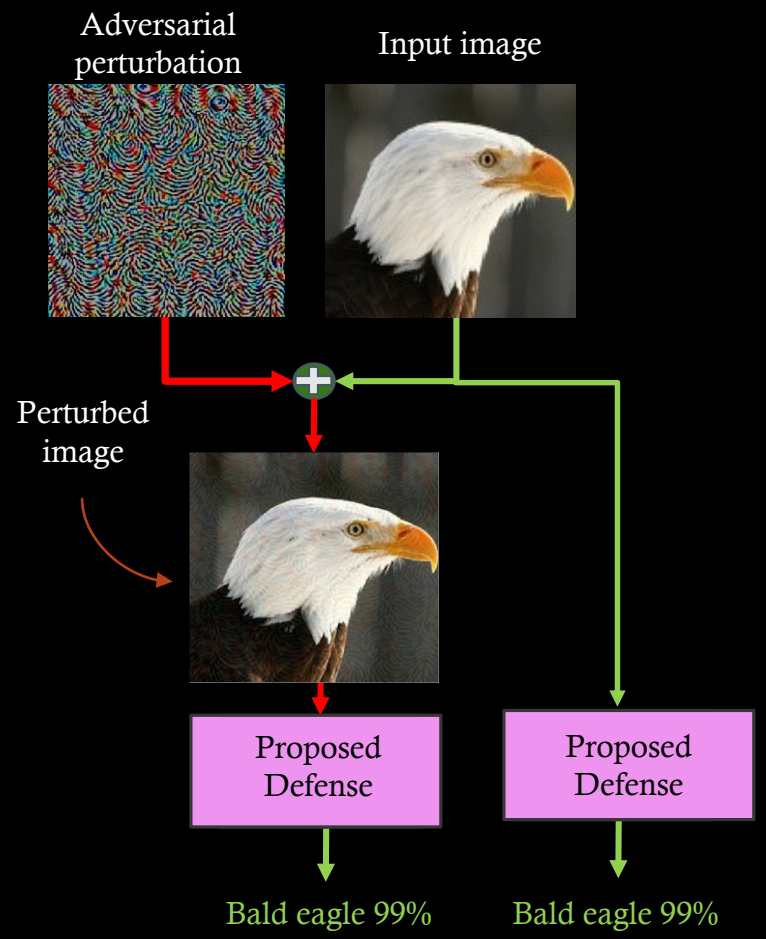
Robustness to Unseen Universal Adversarial Attacks

- Defense trained on only UAP noise samples

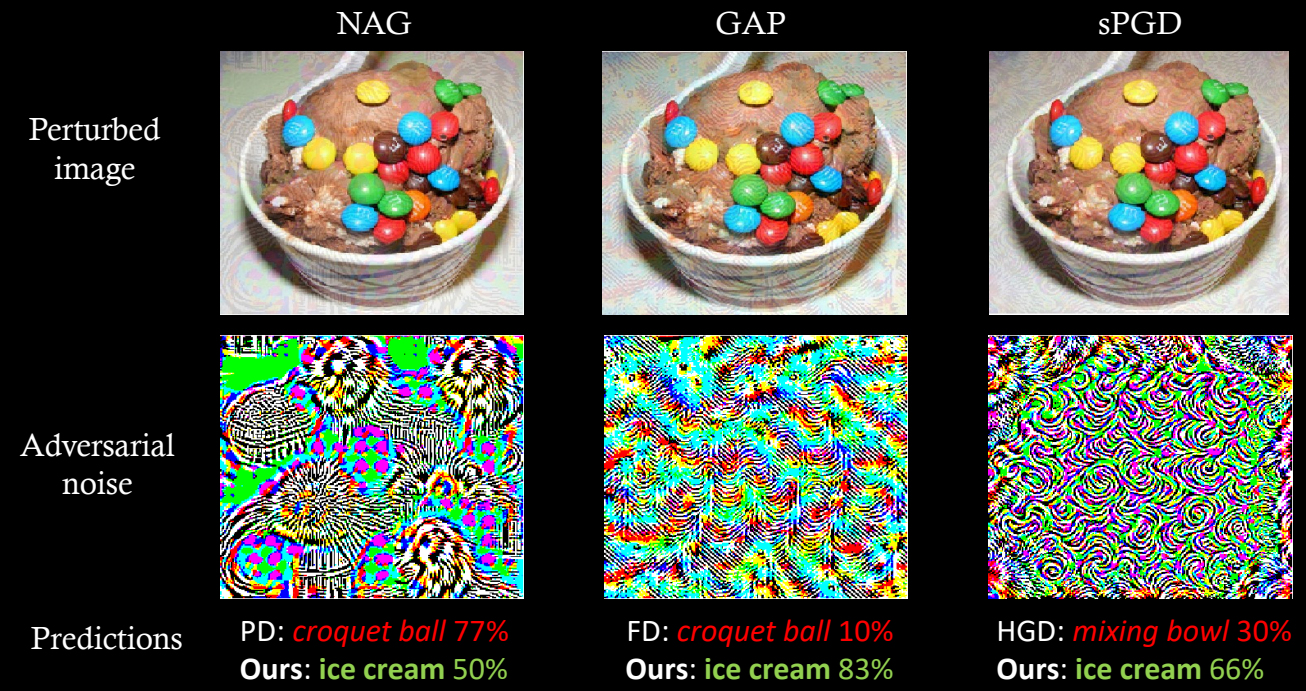


Defending Against Universal Attacks Through Selective Feature Regeneration

Robustness to image-agnostic noise:



Robustness to unseen universal attacks:



Summary:

- Novel ℓ_1 - norm measure identifies and ranks adversarially susceptible feature maps
- Selective regeneration of only the most vulnerable feature maps restores robustness

Code: https://github.com/tsborkar/Selective-feature-regeneration