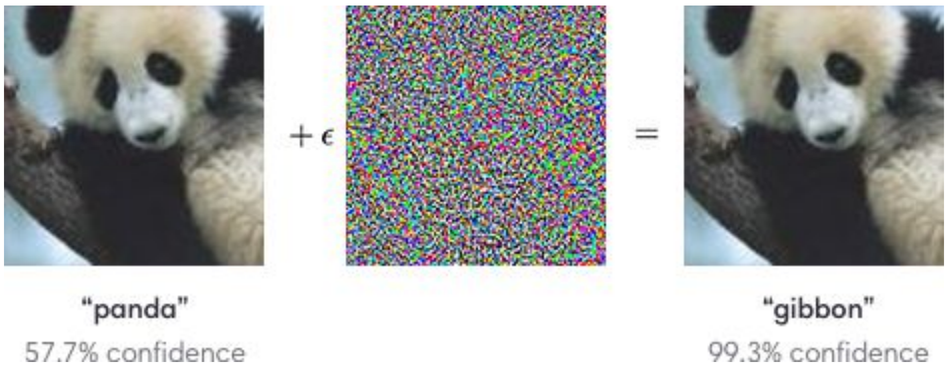
On the Defense Against Adversarial Examples Beyond the Visible Spectrum

Anthony Ortiz[1], Olac Fuentes[1], Dalton Rosario[2], Christopher Kiekintveld[1]

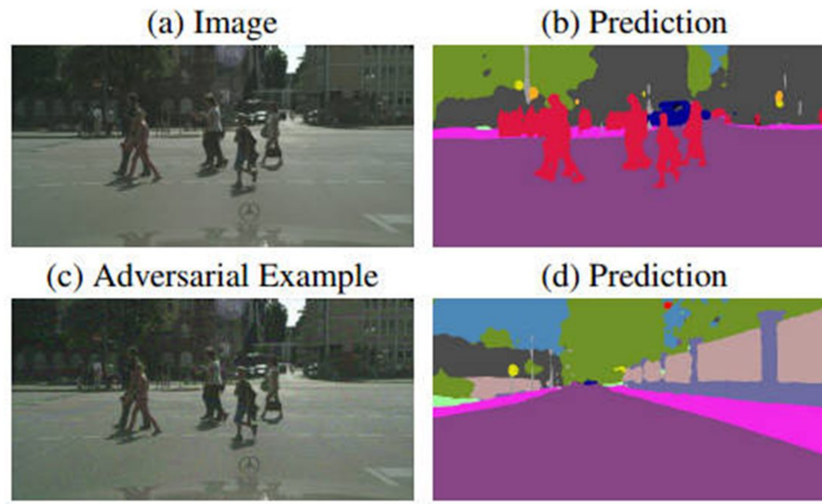
[1]Department of Computer Science, UTEP

[2]US Army Research Laboratory

Adversarial Examples on Natural Images



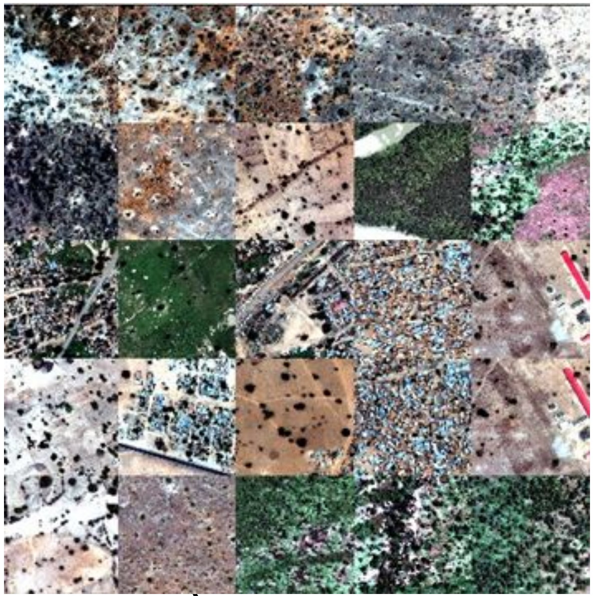
Goodfellow et al., 2015



Fisher et al., 2017

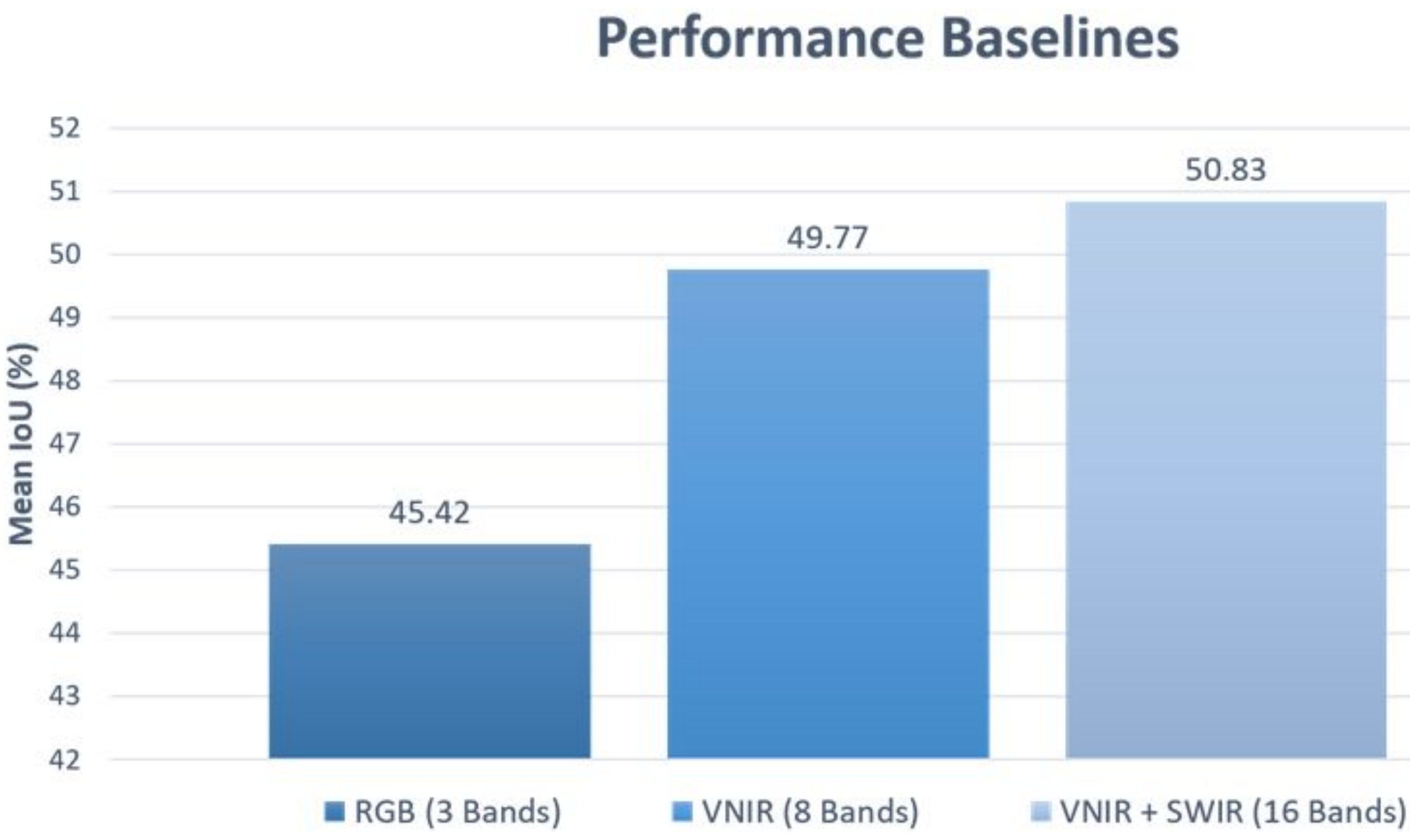
Adversarial Examples Beyond the Visible Spectrum

Experimantal Setup



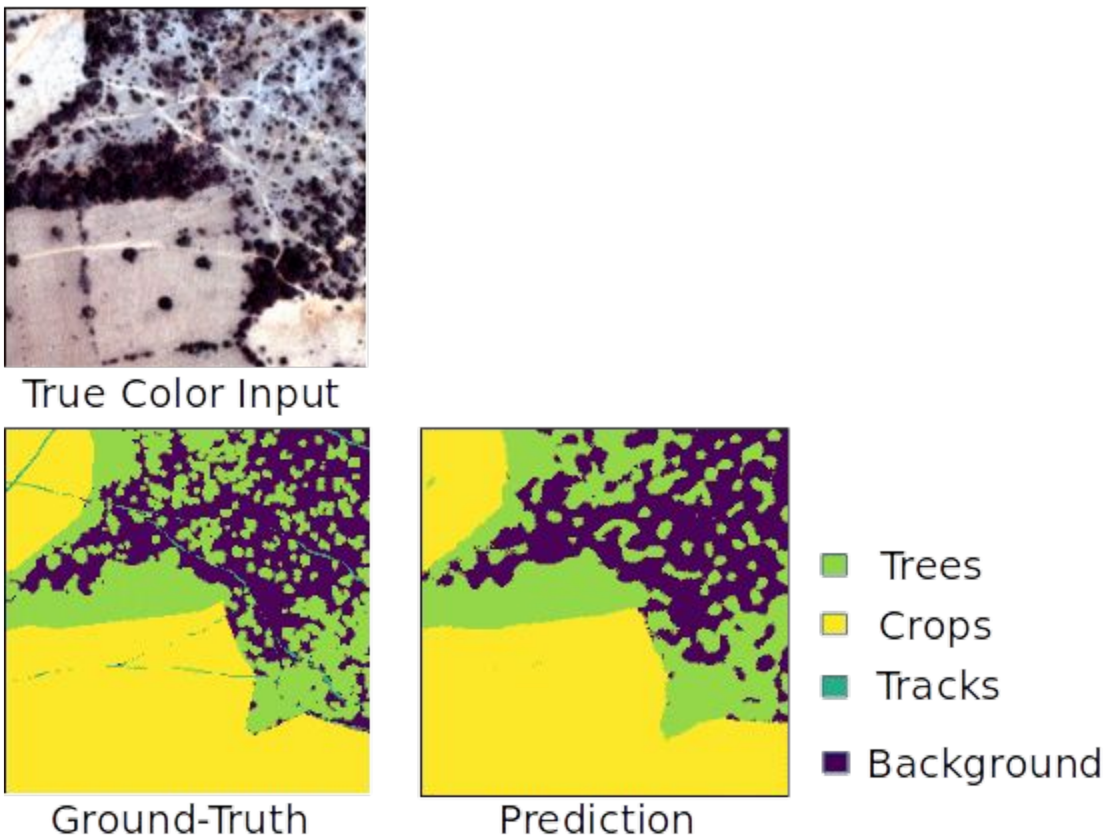
- DSTL Dataset:
- 1 km x 1 km Satellite Image
- Spatial resolution: 31 cm
- 3 channels RGB
- 8 Channels VNIR
- 8 Channels SWIR
- 10 Classes (Buildings, roads, track, trees, crops)
- DigitalGlobe's WorldView Satellite System
- Task: Semantic Segmentation
- Evaluation Metric: Mean IoU
- Architecture:
- Fully Convolutional Networks (FCN-8) with VGG-19 as backbone

Performance Evaluation DSTL Dataset

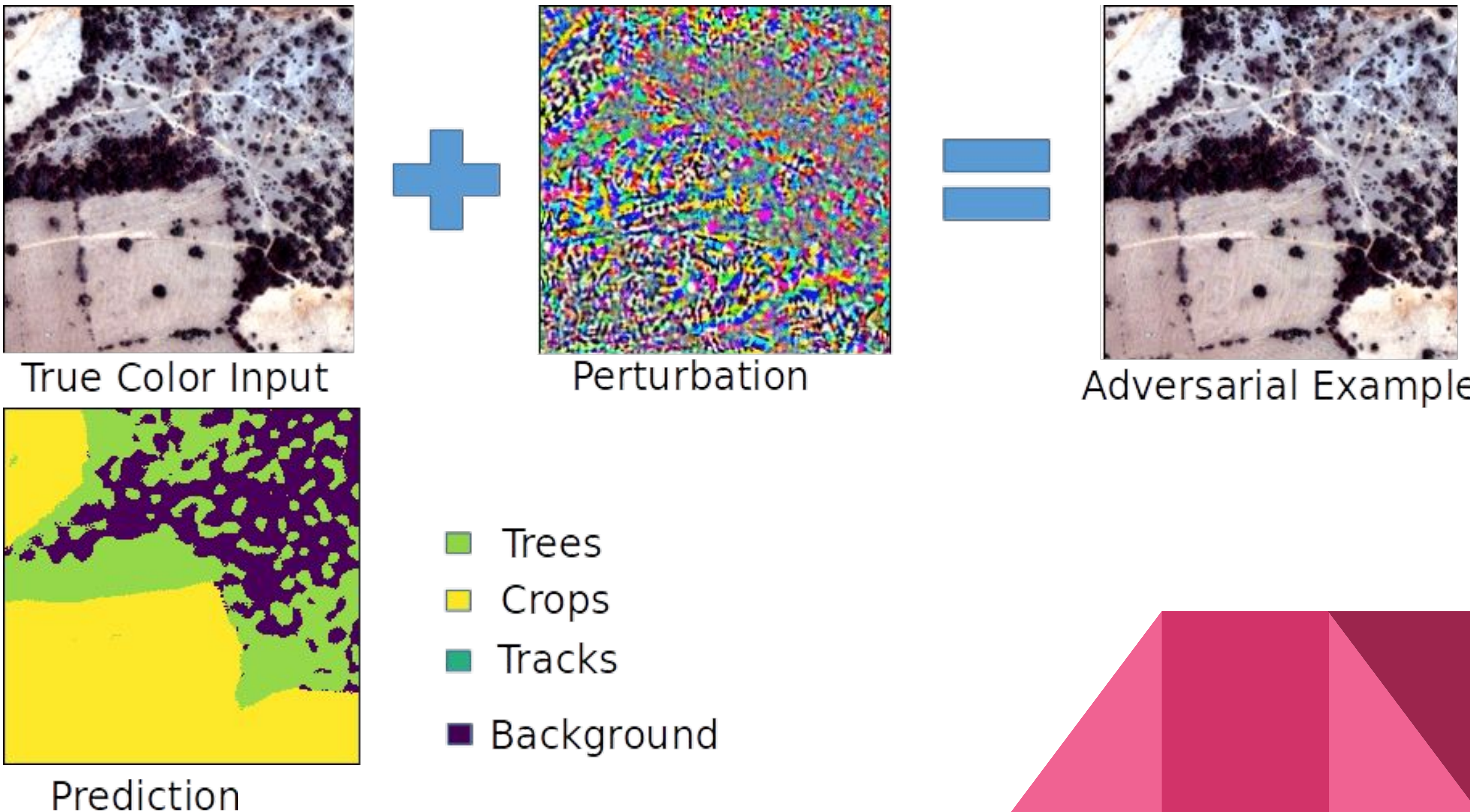


Adversarial Examples Beyond Visible Spectrum

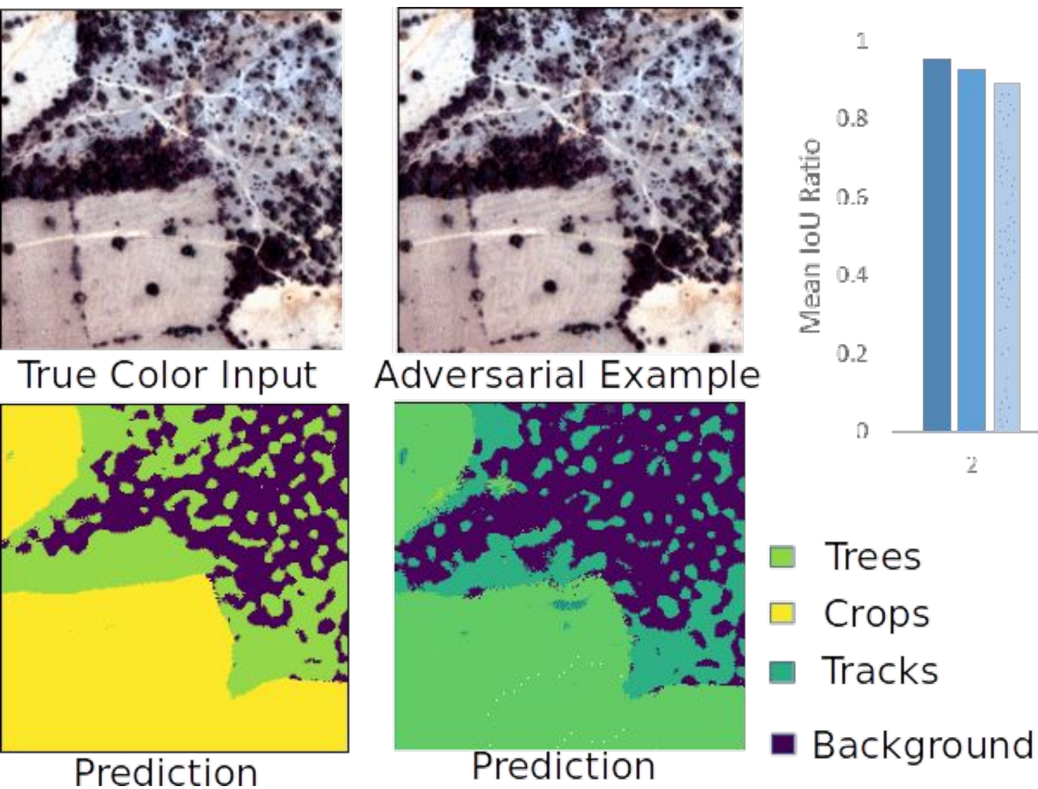
Adversarial Examples Beyond Visible Spectrum

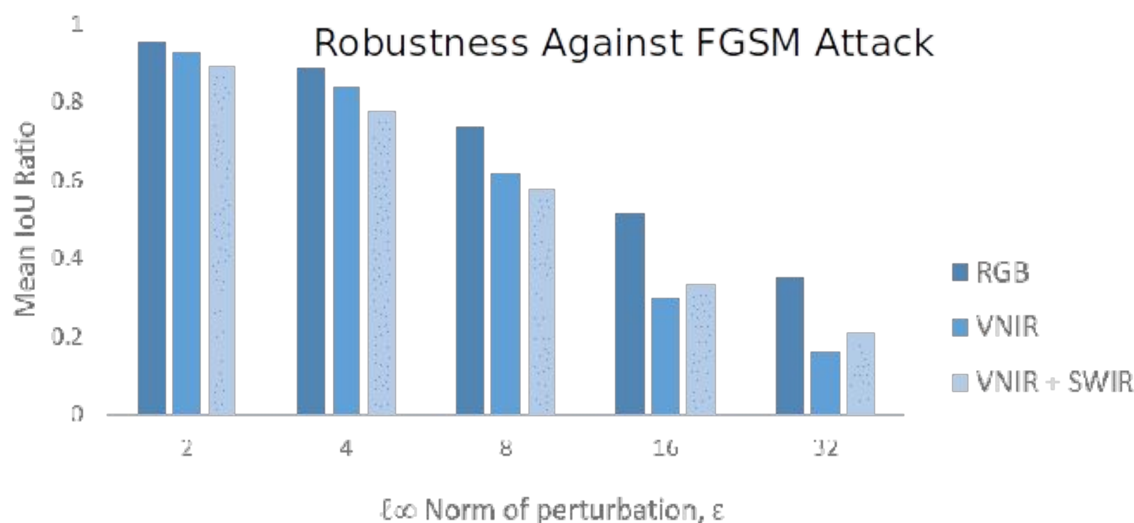


Adversarial Examples Beyond Visible Spectrum

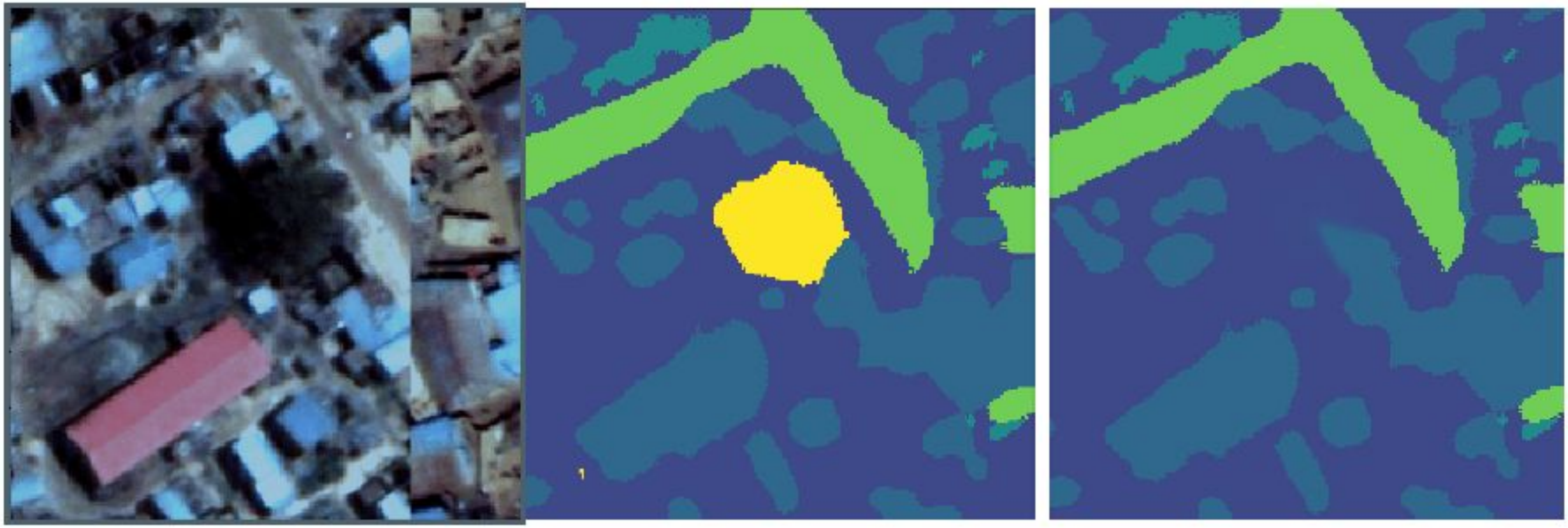


Adversarial Examples Beyond Visible Spectrum





Dynamic Adversarial Perturbation Attack

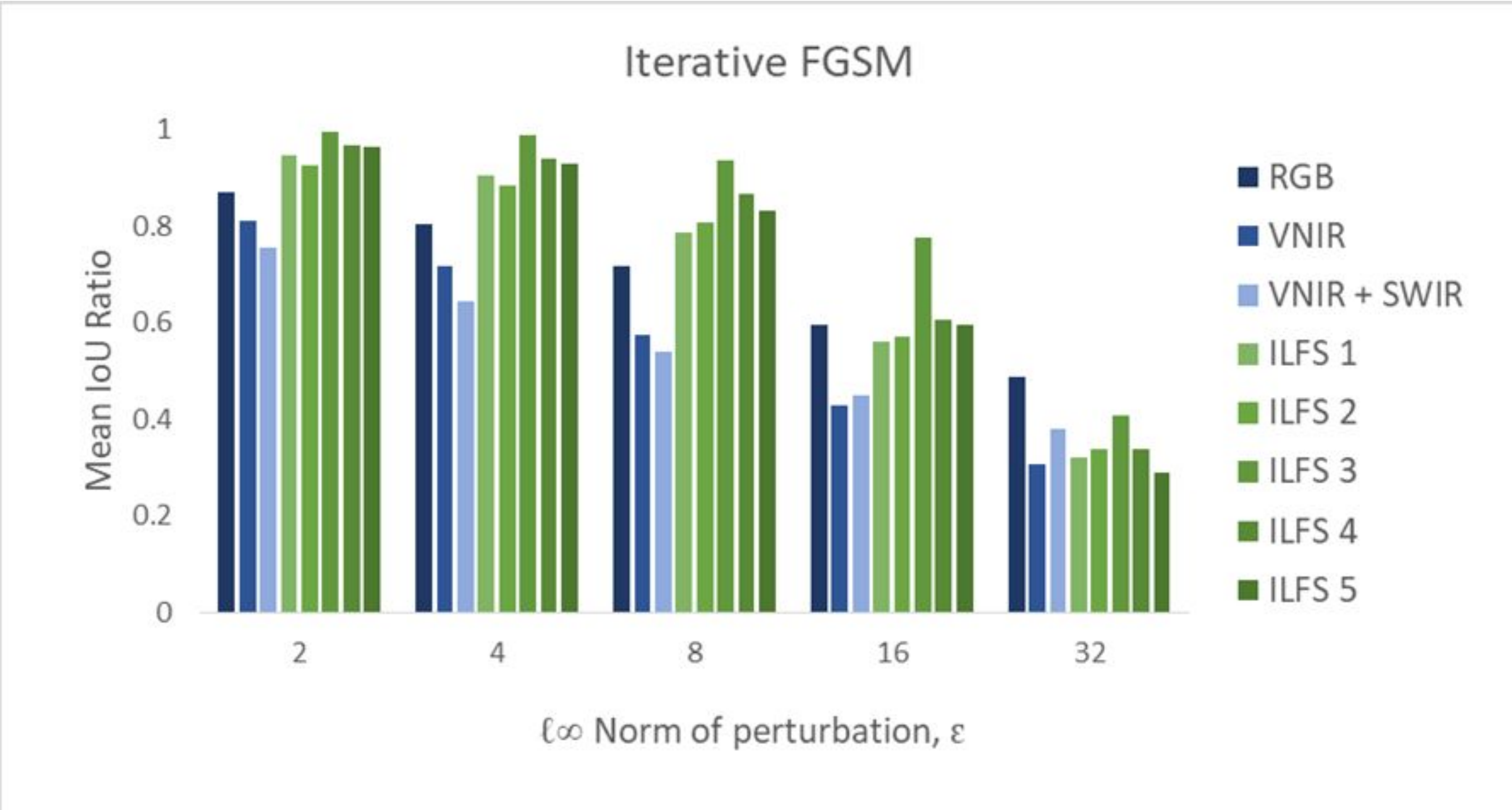


True Color Input

Prediction Clean

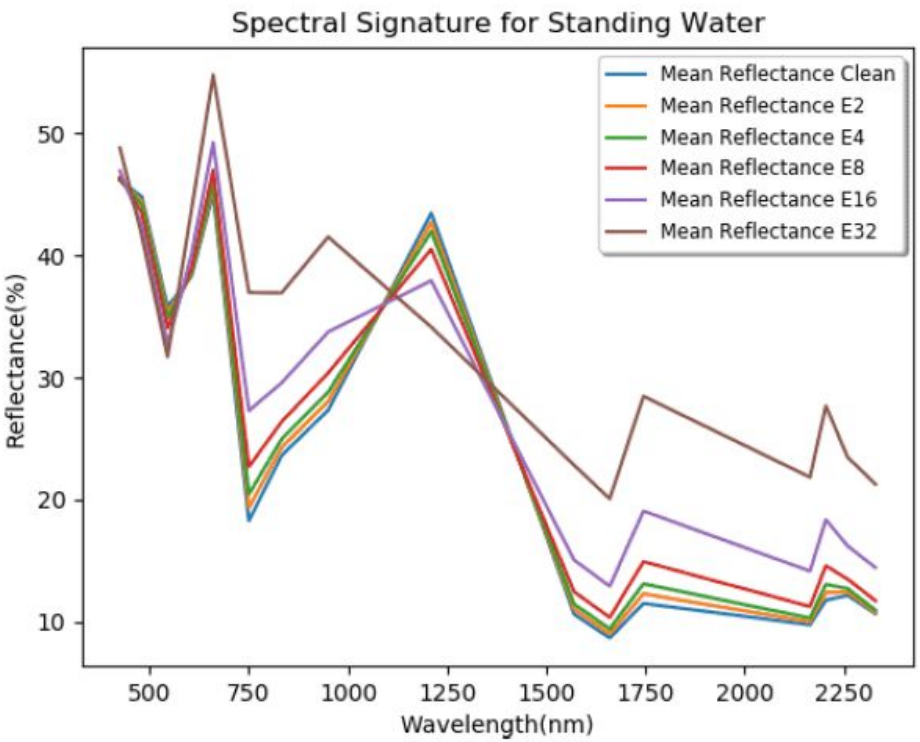
Prediction Adversarial

ILFS as a Defense Against Adversarial Examples

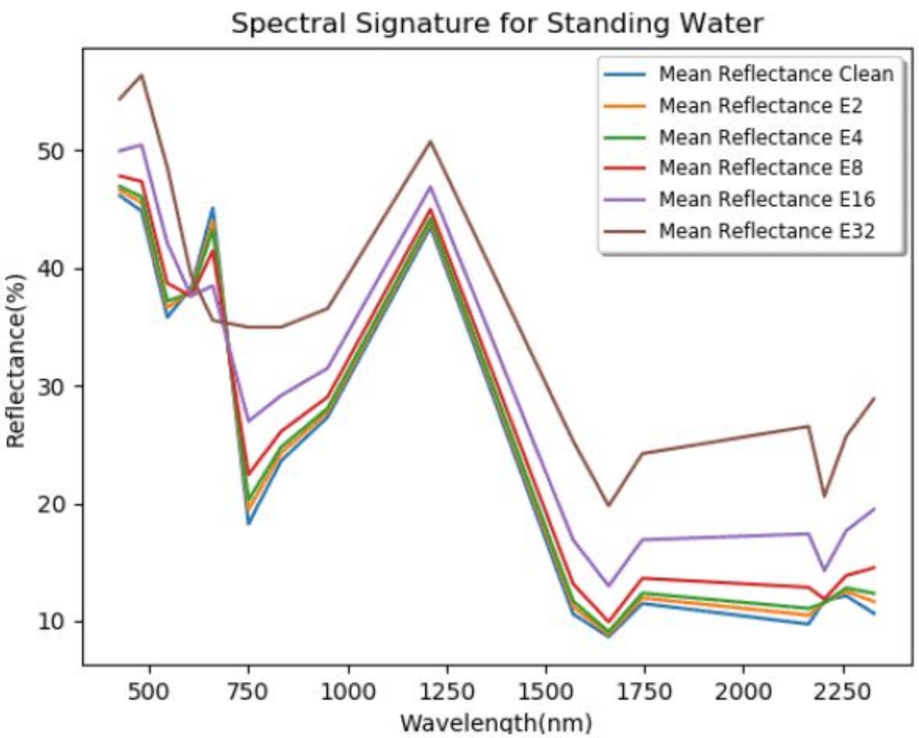


Detecting Adversarial Examples

Spectral Signature Adversarial Examples



FGSM

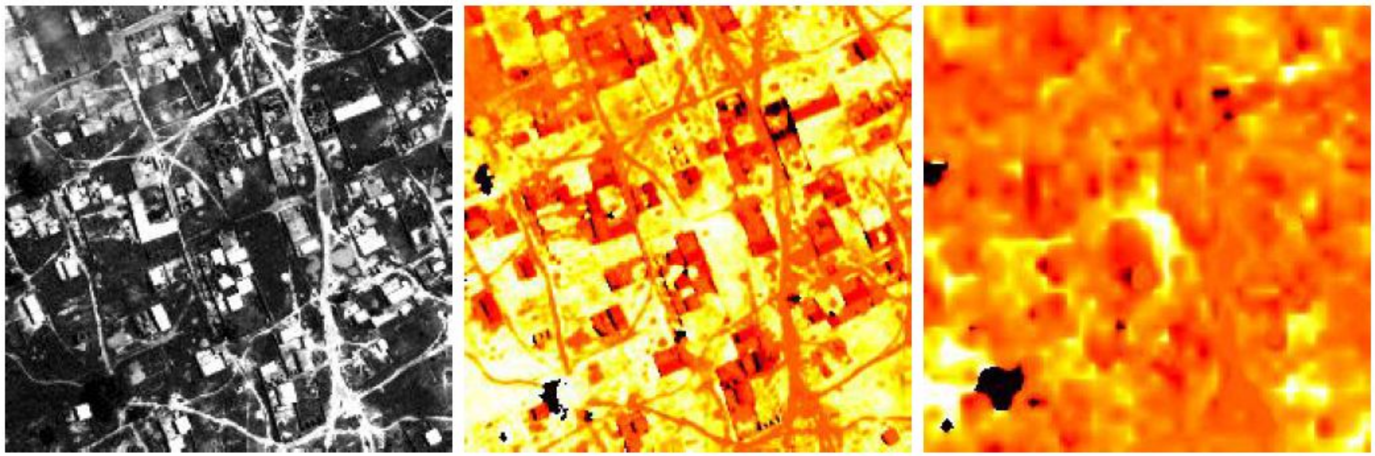


Iterative FGSM

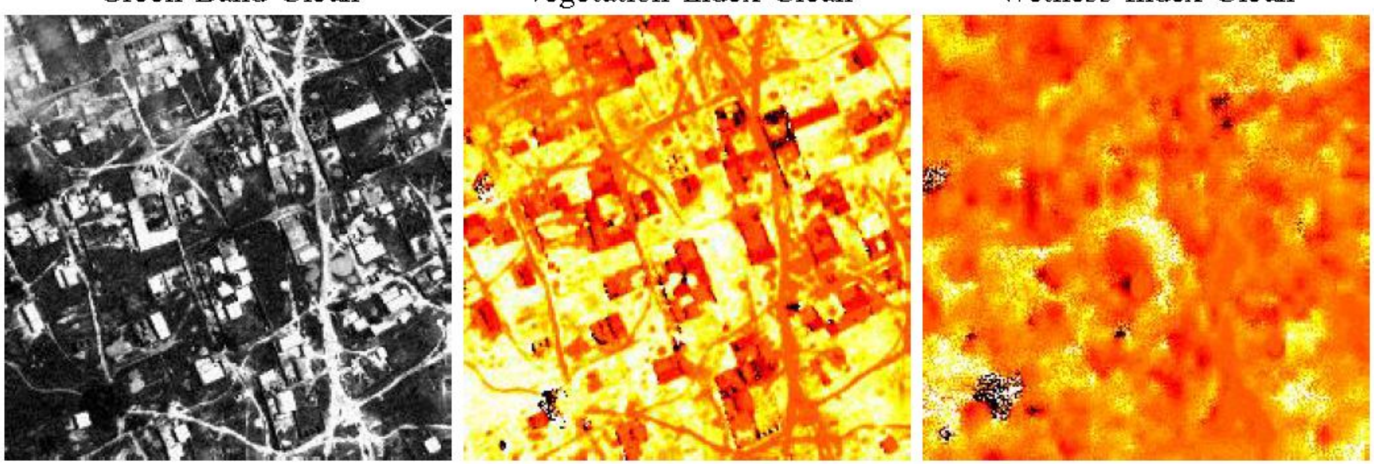
Wetness Index

$$wetness = \frac{b_{swir2} - b_{swir4}}{b_{swir2} + b_{swir4}}$$

Band swir2:1550-1590nm
Band swir4: 1710-1750nm

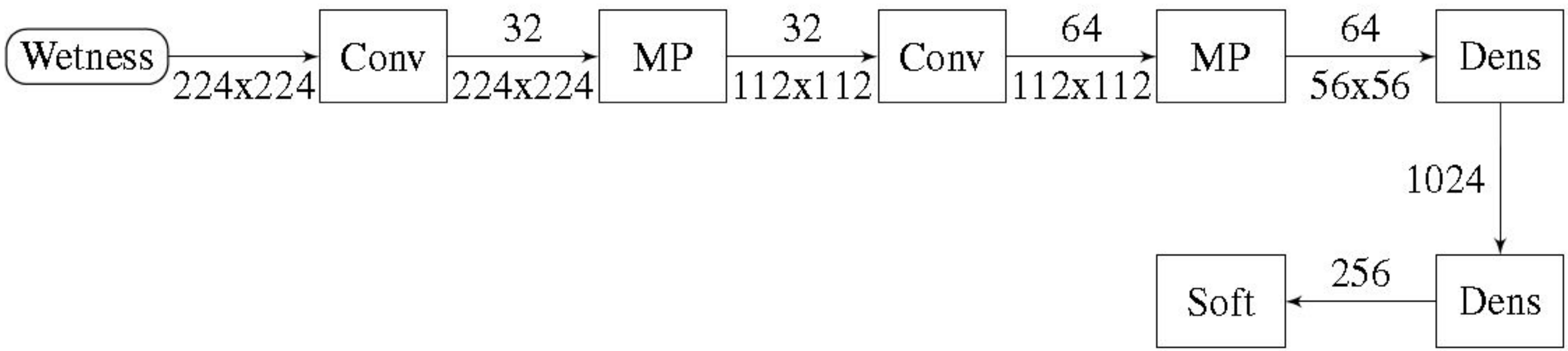


Green Band Clean | Vegetation Index Clean | Wetness Index Clean



Green Band Adversarial | Vegetation Index Adversarial | Wetness Index Adversarial

Detector Network Architecture

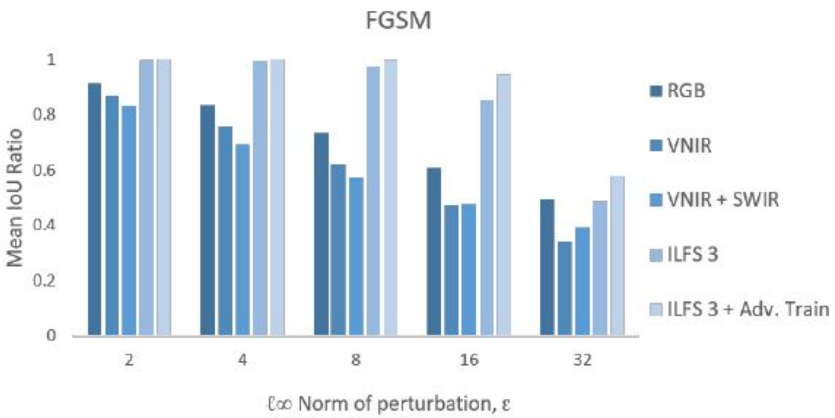


Detection Results

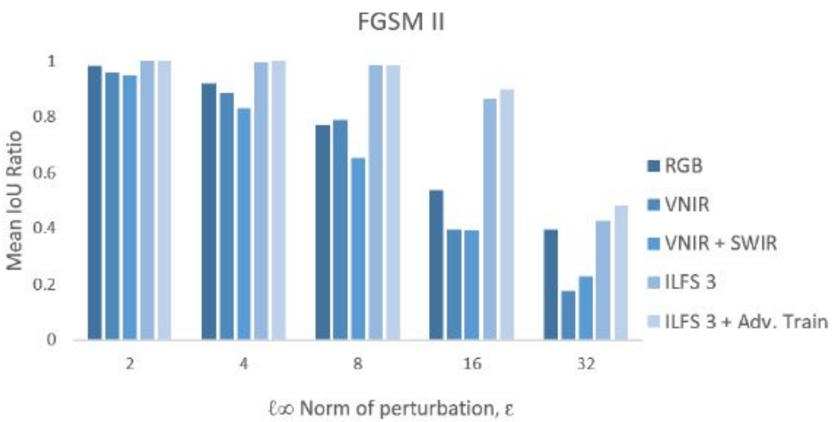
Wetness-based Detector Network Accuracy					
Attack	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = 16$	$\varepsilon = 32$
FGSM	0.84	0.99	1.00	1.00	1.00
FGSM ITER	0.94	0.99	1.00	1.00	1.00
FGSM ll	0.83	0.99	1.00	1.00	1.00
FGSM ll ITER	0.95	0.99	1.00	1.00	1.00

Adversarial Training Helps

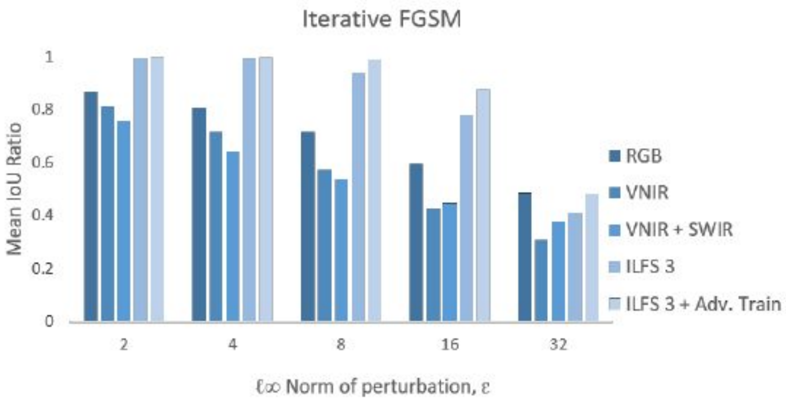
Adversarial Training Helps



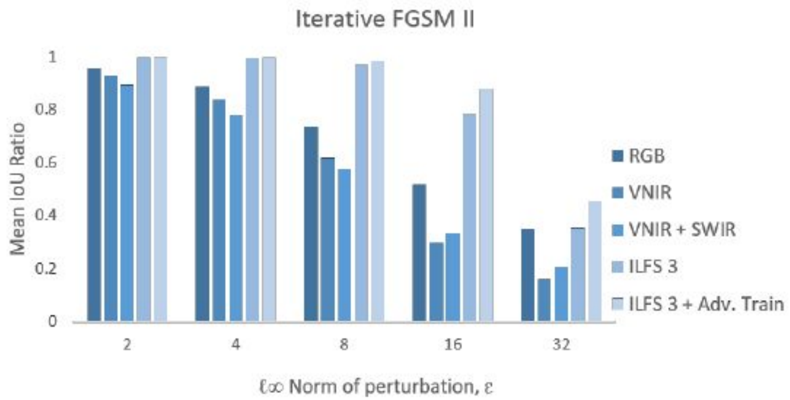
FGSM Attack



FGSM ll Attack



Iterative FGSM Attack



Iterative FGSM ll Attack

Conclusions

- Multispectral and Hyperspectral Images are vulnerable to adversarial examples.
- With the right prior, adversarial examples can successfully be detected.
- Adversarial Training improve models robustness beyond RGB and generalize across attacks.

Thank you