UCSF ci2 center for intelligent imaging

Addressing The False Negative Problem of Deep Learning MRI Reconstruction Models by Adversarial Attacks and Robust Training

Paper #28

MIDL Montréal 2020

Kaiyang Cheng, Francesco Calivà*, Rutwik Shah, Misung Han, Sharmila Majumdar, Valentina Pedoia*

Disclosure

I have no financial interests or relationships to disclose with regard to the subject matter of this presentation.

Funding source

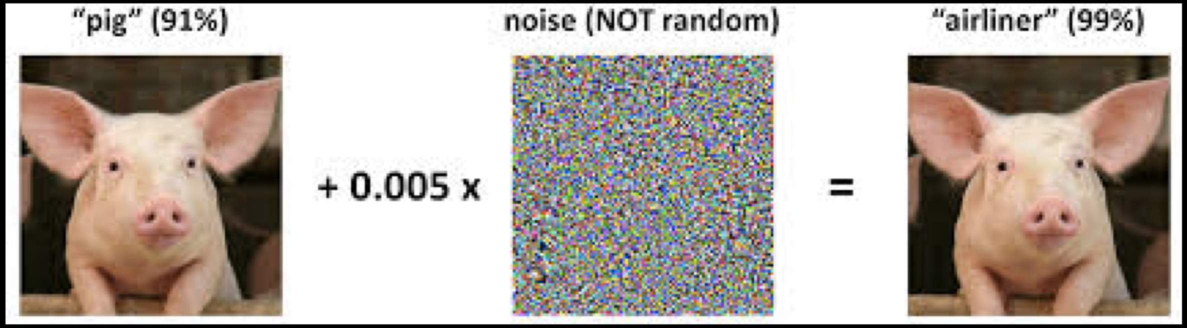
This project was supported by R00AR070902 (VP), R61AR073552 (SM/VP) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, (NIH-NIAMS).

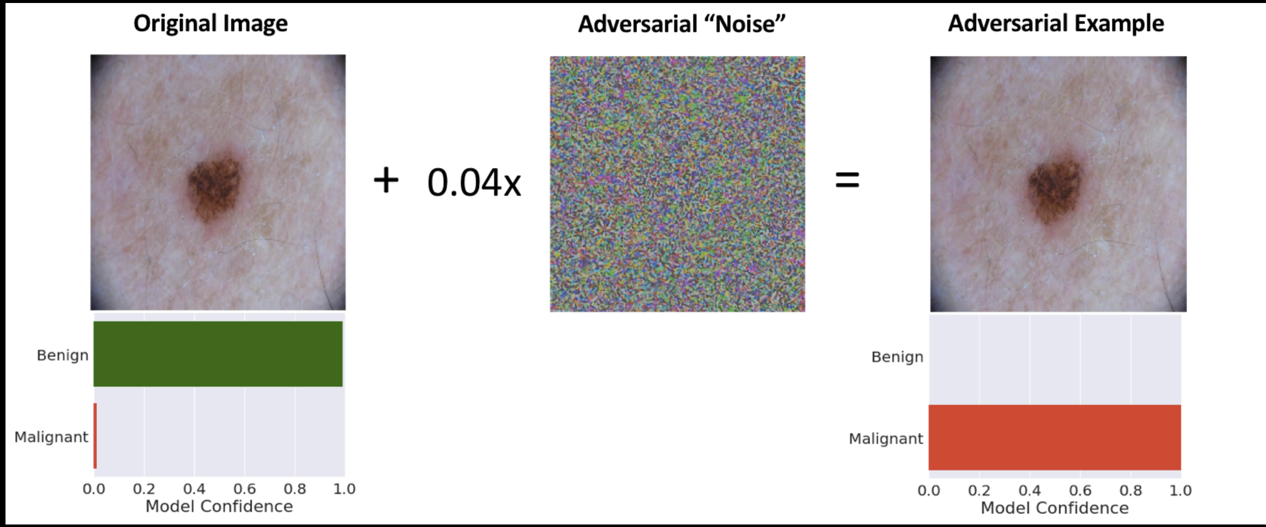
NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases

Outline

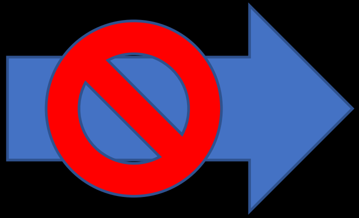
- *Motivation*
- *False negative problem in accelerated MRI reconstruction*
- *Adversarial examples*
- *FNAF attack*
- *Adversarial robustness training*
- *FNAF robust training*
- *Experimental results*
- *Conclusions*

Adversarial Examples in Medical Imaging Analysis





Adversarial Examples in Medical Imaging Analysis



IID Machine Learning vs Adversarial Machine Learning

$$\mathbb{E}_{(x,y)\sim D}[L(x,y,\theta)]$$

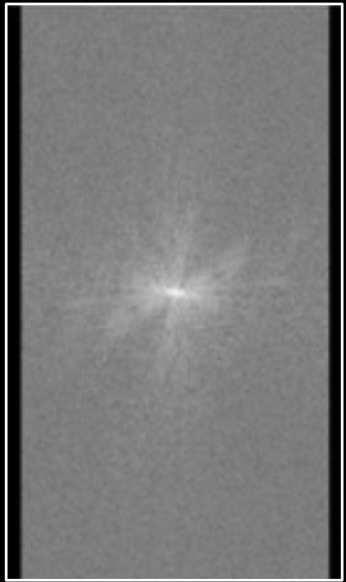
IID:
Average Case

$$\mathbb{E}_{(x,y)\sim D}[\max_{\delta\in S} L(\theta,x+\delta,y)]$$

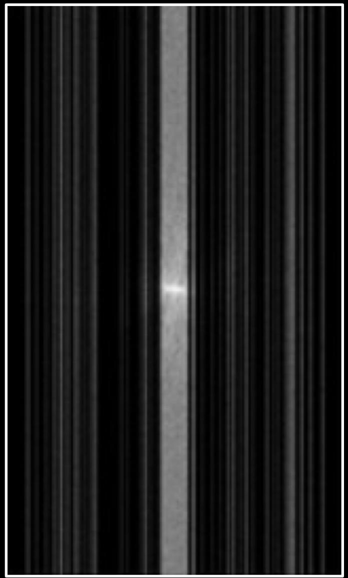
Adversarial:
Worst Case

Accelerated MRI Reconstruction

Fully-sampled k-space



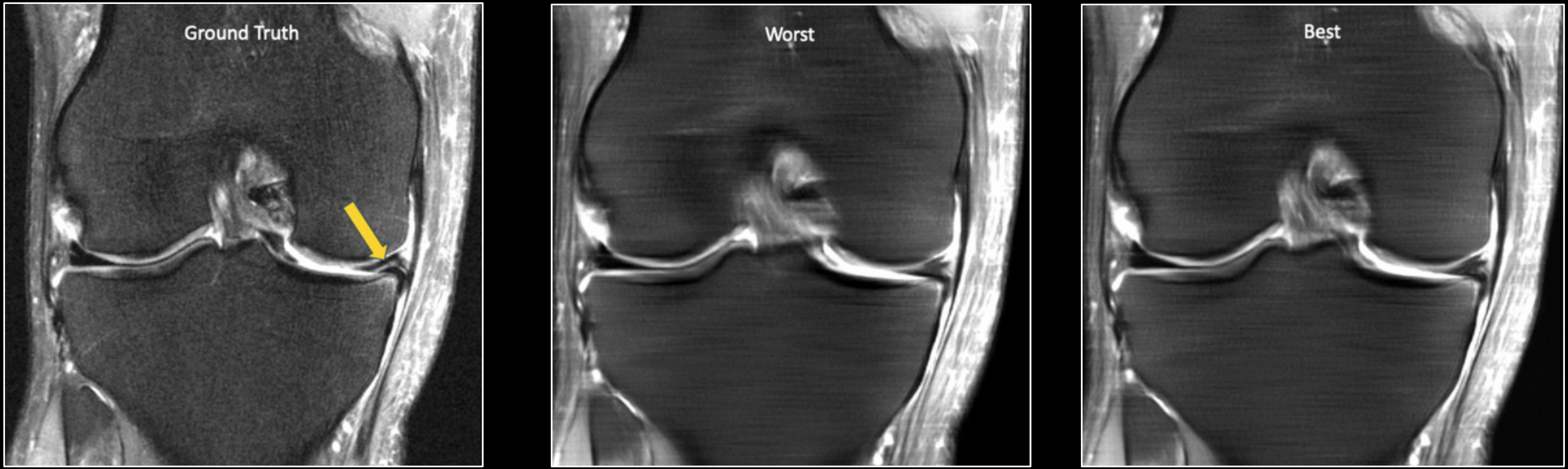
Under-sampled k-space



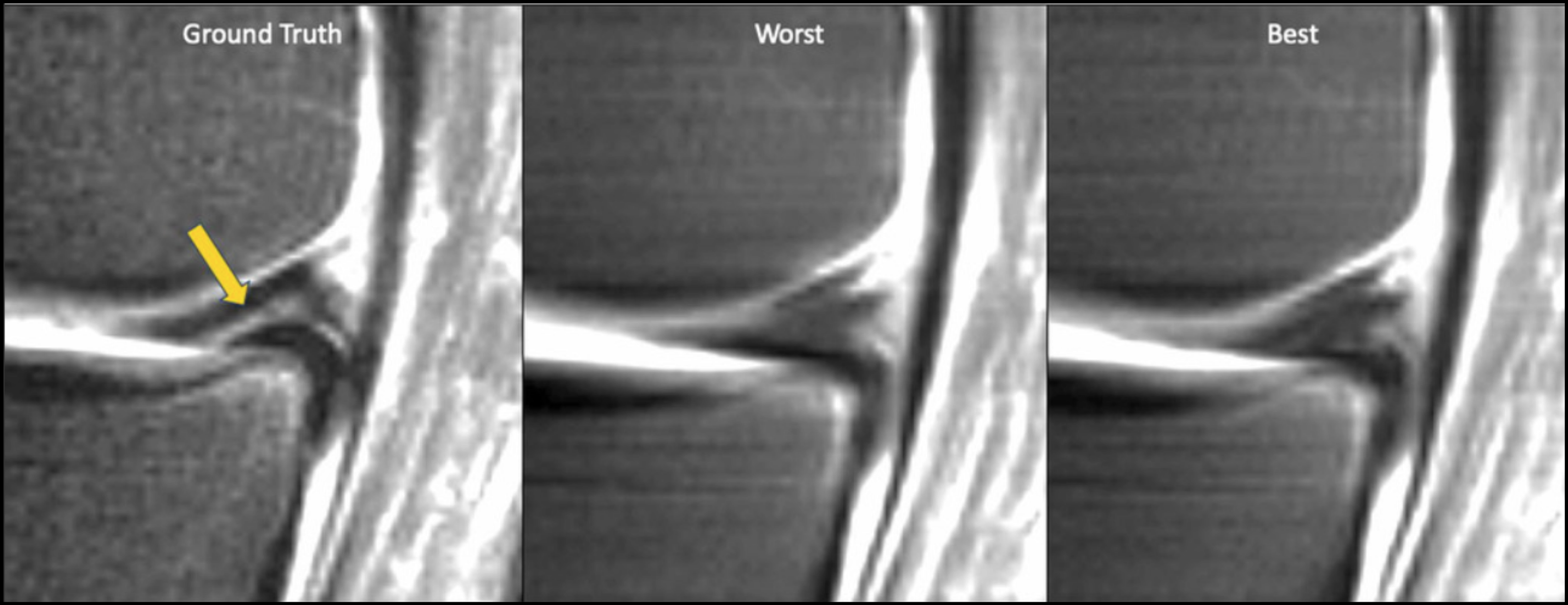
Methods

victorcheng21@Berkeley.edu

FastMRI results: loss of meniscal tear



The False Negative Phenomenon



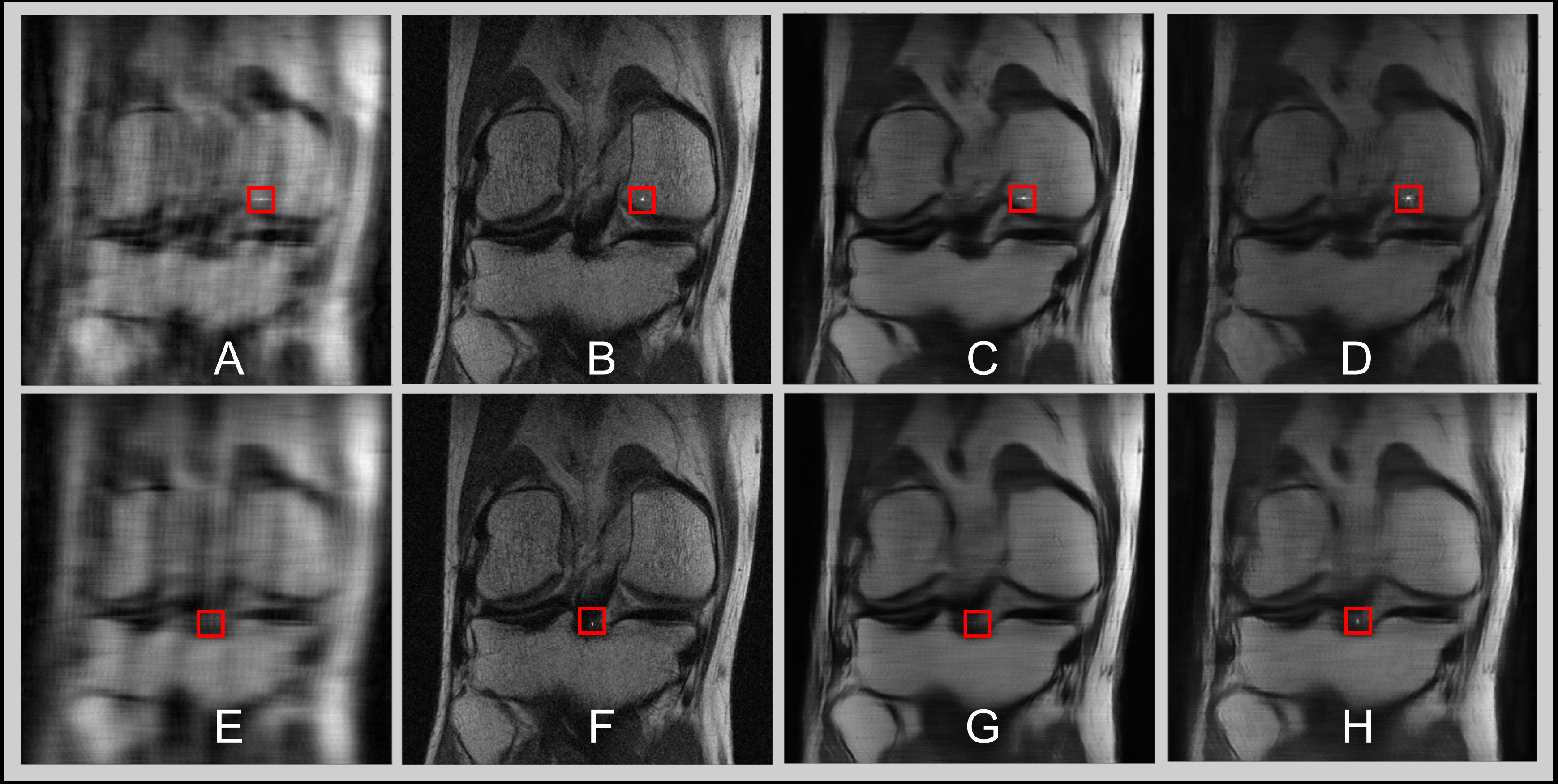
Two hypotheses for the false negative problem:

1) *The information of small abnormality features is completely lost through the under- sampling process*

2) *The information of small abnormality features is not completely lost. Instead, it is attenuated and laid in the tail-end of the task distribution, hence is rare*

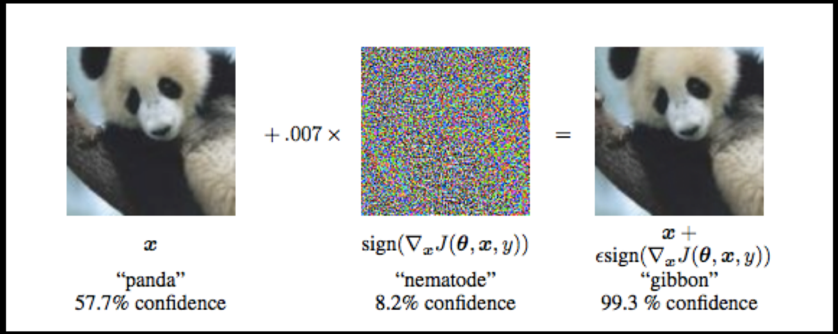
FNAF: false-negative adversarial feature

A perceptible small feature which is present in the ground truth MRI but has disappeared upon MRI reconstruction.

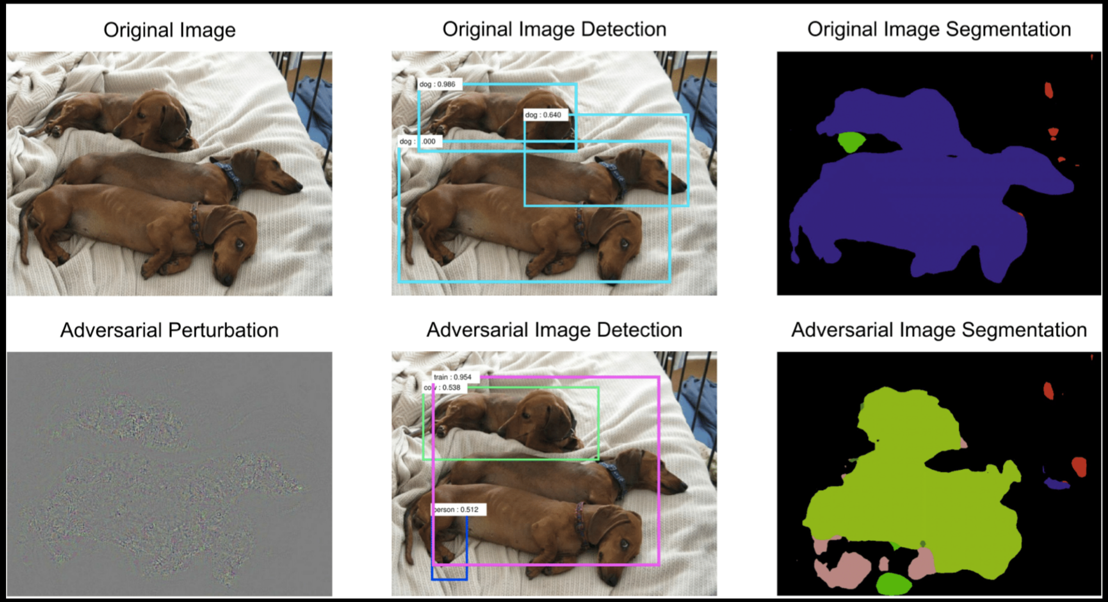


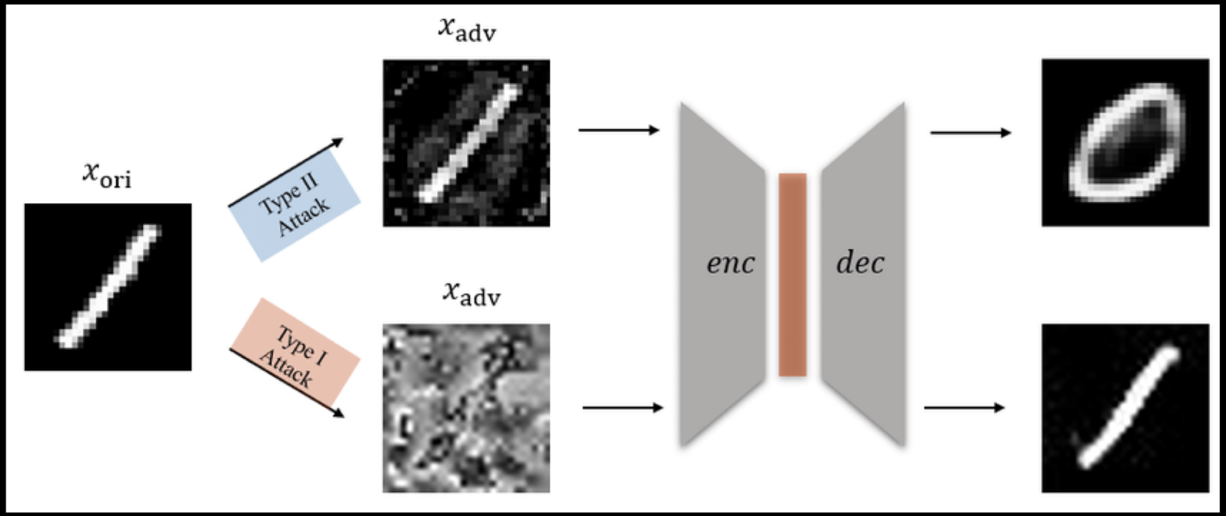
Adversarial Examples and Attacks

$$\max_{\delta \in S} L(\theta, x + \delta, y)$$



Adversarial Examples and Attacks





FNAF Attack

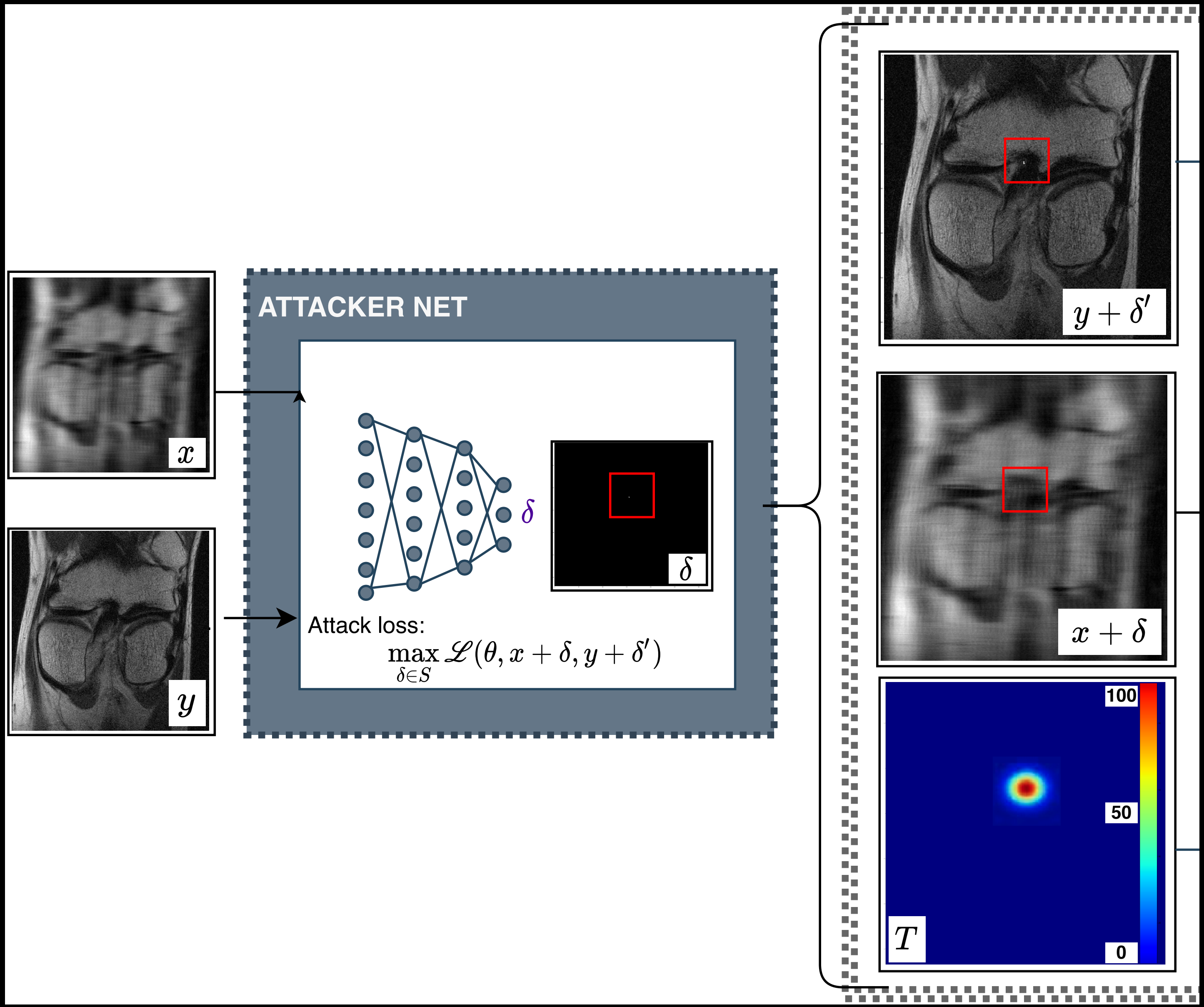
$$\max_{\delta \in S} L(\theta, x + \delta, y)]$$

$$\max_{\delta \in S} L(\theta, x + \delta, y + \delta')]$$

$$\delta = U(\delta')$$

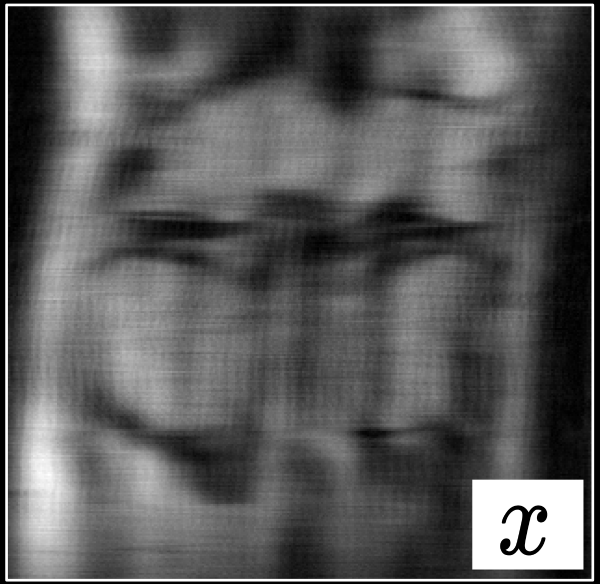
$$U(y) = \mathcal{F}^{-1}(M(\mathcal{F}(y)))$$

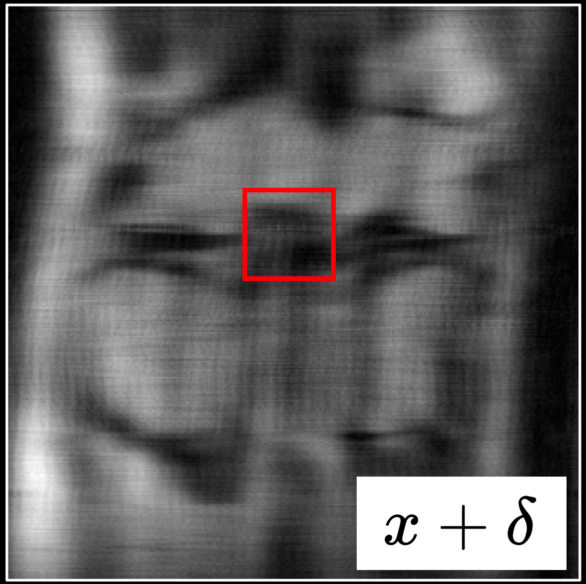
$$MSE(T(x), T(y))$$



Under-sampling information preservation

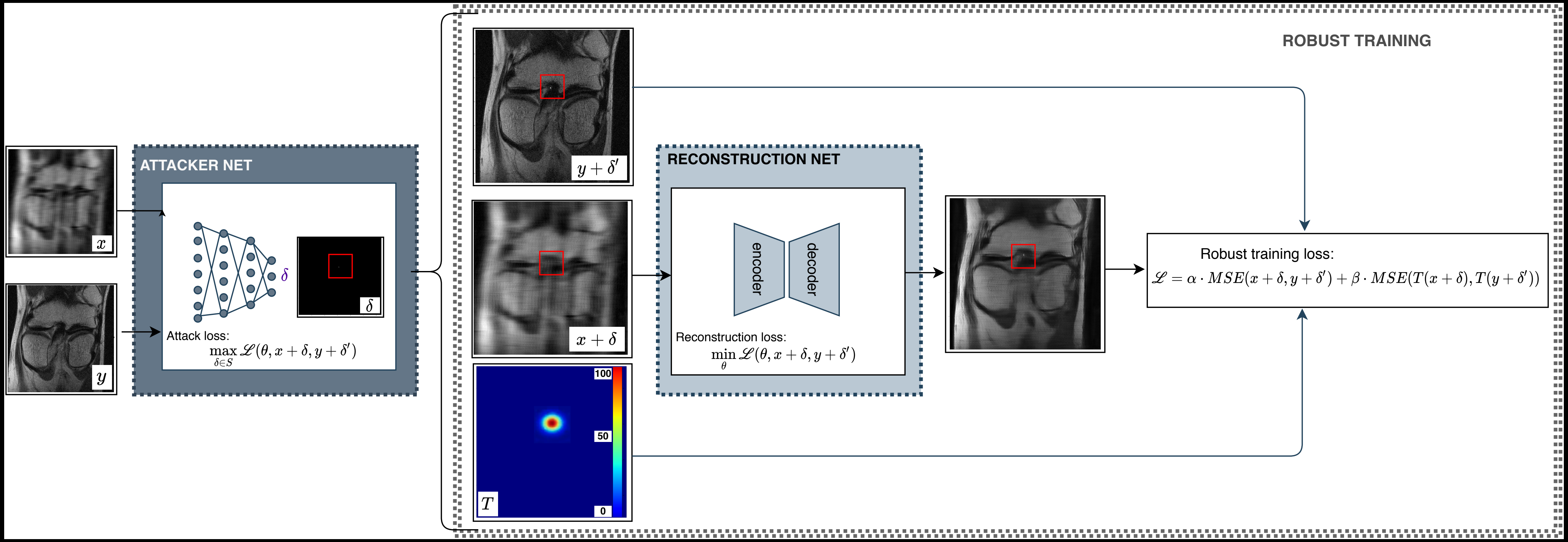
$$D(x + \delta, x) > \varepsilon$$





Adversarial robustness training

$$\mathbb{E}_{(x,y)\sim D}[\max_{\delta\in S} L(\theta, x+\delta, y)]$$



victorcheng21@Berkeley.edu

Experimental Results

Table 1: Standard validation set evaluation with SSIM and normalized mean-square error (NMSE)

$4\times$	SSIM	NMSE
U-Net	0.7213 ± 0.2621	0.03455 ± 0.05011
I-RIM	0.7501 ± 0.2546	0.03413 ± 0.05800
FNAF-robust U-Net	0.7197 ± 0.2613	0.03489 ± 0.05008

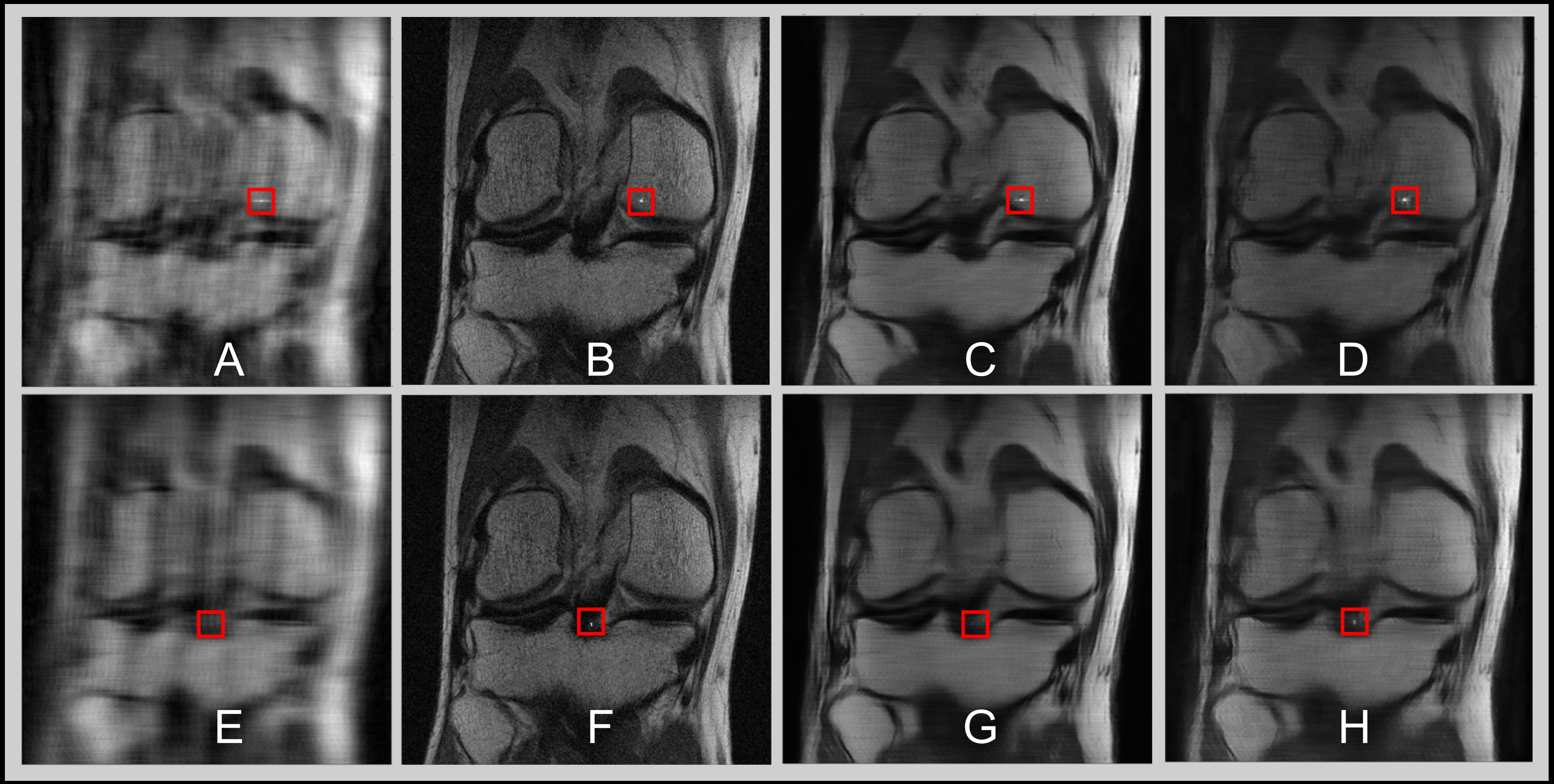
$8\times$	SSIM	NMSE
U-Net	0.6548 ± 0.2942	0.04935 ± 0.04962
I-RIM	0.6916 ± 0.2941	0.04438 ± 0.06830
FNAF-robust U-Net	0.6533 ± 0.2924	0.04962 ± 0.05670

Table 2: FNAF attack evaluations.

$4\times$	RS (Attack Rate %)	FD (Attack Rate %)	RS (MSE)	FD (MSE)
U-Net	84.44	72.17	0.001530	0.001386
I-RIM	44.49	34.60	0.001164	0.001080
FNAF-robust U-Net	12.71	10.48	0.000483	0.000466

$8\times$	RS (Attack Rate %)	FD (Attack Rate %)	RS (MSE)	FD (MSE)
U-Net	86.00	74.84	0.001592	0.001457
I-RIM	77.39	63.88	0.001470	0.001349
FNAF-robust U-Net	15.09	13.30	0.000534	0.000467

Qualitative Results



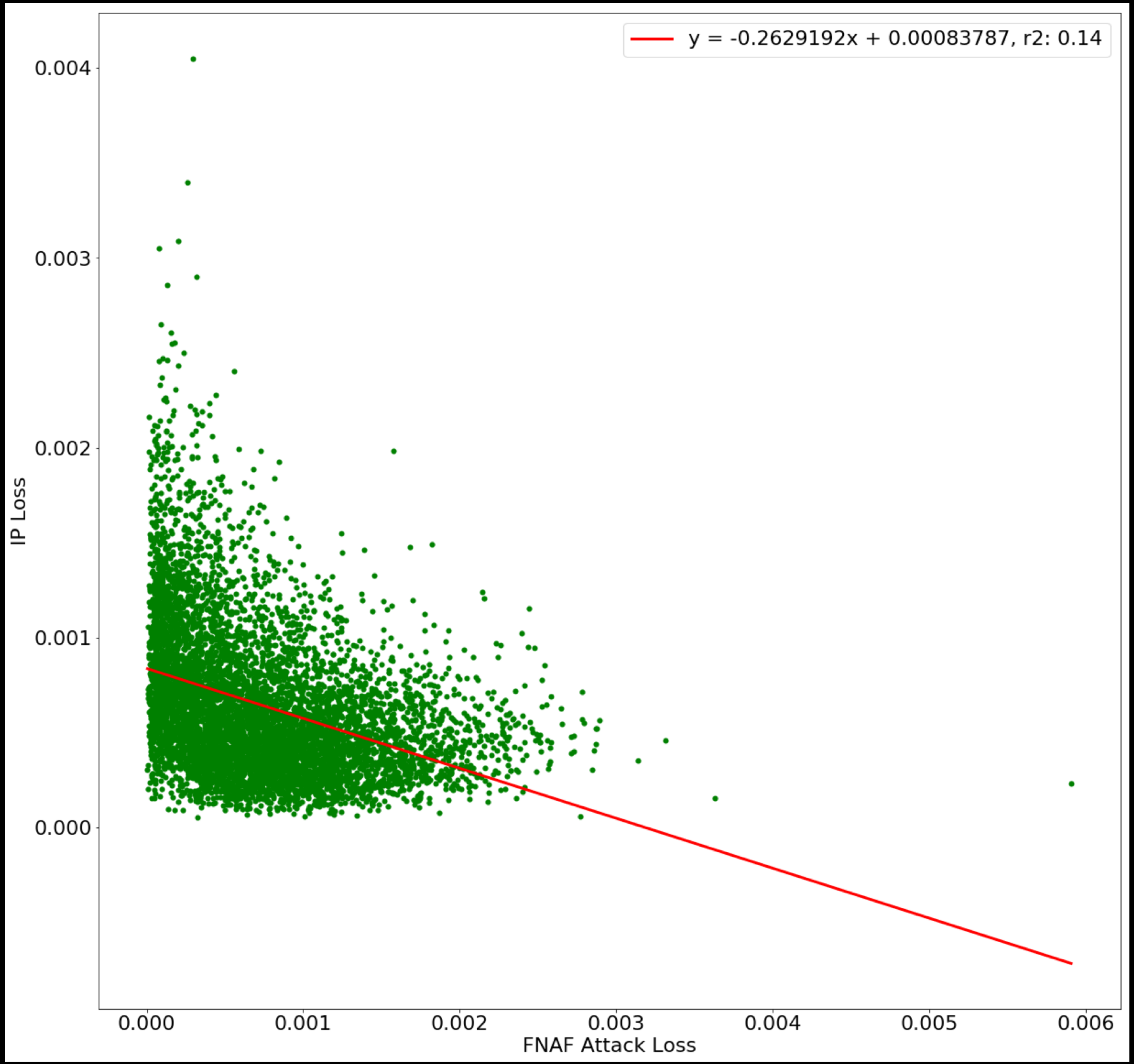
The top row (A-D) shows a "failed" FNAF attack. The bottom row (E-H) shows a "successful" FNAF attack. Column 1 contains the under-sampled zero-filled images. Column 2 contains the fully-sampled ground truth images. Column 3 contains U-Net reconstructed images. Column 4 contains FNAF-robust U-Net reconstructed images. (C-G-D-H) FNAF reconstruction: (C) adversarial loss of 0.000229. (G) adversarial loss of 0.00110. (D) adversarial loss of $9.73 \cdot 10{-}5$. (H) adversarial loss of 0.000449

Information Preservation (IP)

$$D(x+\delta, x) > \varepsilon$$

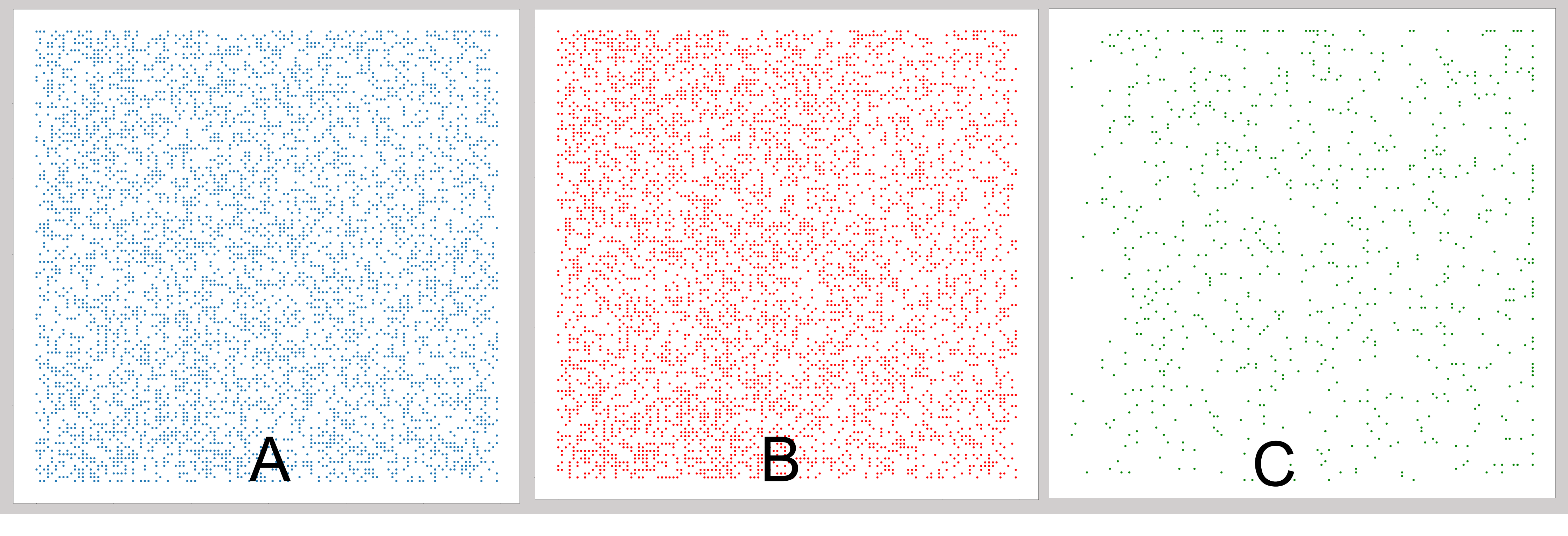
	Random	U-Net FNAF	I-RIM FNAF	Robust U-Net FNAF
Acceptance Rate (%)	99.82	99.72	99.76	99.34
IP Loss (MSE)	0.00064	0.00050	0.00051	0.00052

FNAF Attack Loss vs. IP Loss



victorcheng21@Berkeley.edu

FNAF Location Distribution and Transferability

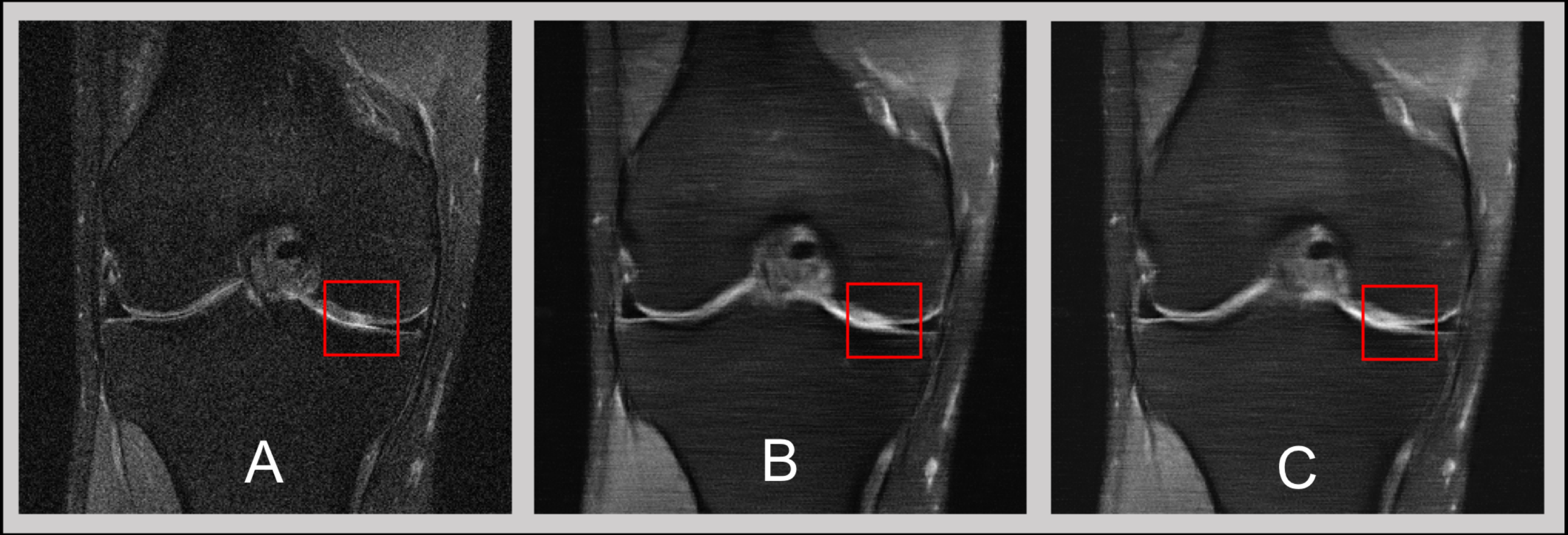


FNAF location distribution within the 120x120 center crop of the image of (A) U-Net, (B) I-RIM, (C) FNAF-robust U-Net

We take FNAF examples from U-Net and apply them to I-RIM, and observe a 89.48% attack rate.

Real-world Abnormalities reconstruction

	Cartilage Lesion Rate	Meniscus Lesion Rate
U-Net	1/8	8/9
FNAF-robust U-Net	3/8	9/9



(A) Ground truth: small cartilage lesion in femur. (B) U-Net: Area of cartilage lesion not defined and resembles increased signal intensity. (C) FNAF-robust U-Net: Cartilage lesion preserved but less clear.

Limitations

- *FNAF attack hit rate was defined heuristically*
- *Attack inner maximization optimization has no guarantee and can be expensive*
- *Adversarial training is only empirically robust*
- *Limited real world abnormalities evaluation*

Conclusions and Future directions

- *Two hypotheses*

1) *The information of small abnormality features is completely lost through the under- sampling process*

2) *The information of small abnormality features is not completely lost. Instead, it is attenuated and laid in the tail-end of the task distribution, hence is rare*

- *Address our limitations*
- *Robustness in other medical imaging tasks*

Acknowledgements

Valentina Pedoia's Lab
Francesco Calivà
Rutwik Shah

Sharmila Majumdar's Lab
Misung Han
Claudia Iriondo

Funding source

This project was supported by R00AR070902 (VP), R61AR073552 (SM/VP) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, (NIH-NIAMS).

National Institute of Arthritis and Musculoskeletal and Skin Diseases

UCSF center for intelligent imaging

victorcheng21@berkeley.edu

Paper #28