Training-Free Adversarial Robustness in Computational MRI

Mahdi Saberi[1,2], Chi Zhang[1,2], Mehmet Akçakaya[1,2]

[1] Department of Electrical and Computer Engineering, University of Minnesota, MN, USA
[2] Center for Magnetic Resonance Research, University of Minnesota, MN, USA.

MRI Reconstruction

- In MRI, raw measurements are collected in frequency domain (know is k-space)
- Accelerated MRI techniques, acquire subsampled data:

$$\mathbf{y}_\Omega = \mathbf{E}_\Omega \mathbf{x} + \mathbf{n}$$

Acquired sub-sampled data (i.e. Sampling Pattern) — Measurements noise — Multi-coil encoding matrix

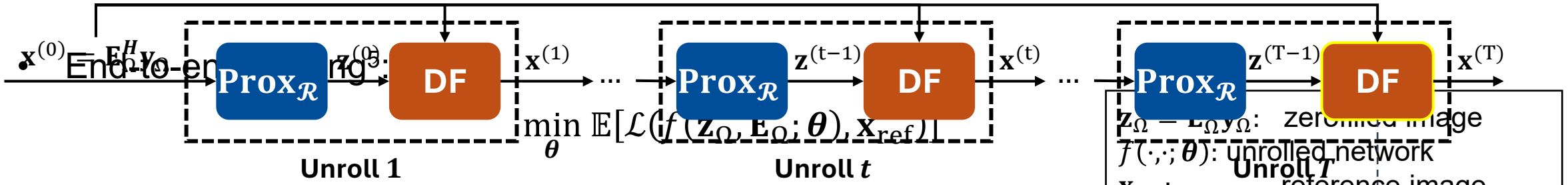
- Most modern reconstruction solves a regularized least squares problem

$$\arg\min_{\mathbf{x}} \|\mathbf{y}_\Omega - \mathbf{E}_\Omega \mathbf{x}\|_2^2 + \mathcal{R}(\mathbf{x})$$

Data Fidelity (DF) — Regularizer

conventionally solved iteratively, by alternating between DF and regularization[1]

- Physics-driven deep learning (PD-DL) employ algorithm unrolling[2,3]
- Unrolling variable splitting with quadratic penalty (VSQP)[1], as MoDL[4]



End-to-end training:

$$\min_{\boldsymbol{\theta}} \mathbb{E}[\mathcal{L}(f(\mathbf{z}_\Omega, \mathbf{E}_\Omega; \boldsymbol{\theta}), \mathbf{x}_{\text{ref}})]$$

$\propto$ Final DF step enforces measurement consistency $\mathbf{y}_\Omega$

[1] Fessler et al, IEEE Sig. Proc. Mag., 2020; [2] Monga et al, IEEE Sig. Proc. Mag., 2021; [3] Hammernik et al, IEEE Sig. Proc. Mag., 2023;
[4] Aggarwal et al, IEEE Trans.. Med. Imaging., 2019; [5] Hammernik et al, Magn. Reson. Med., 2019;

Adversarial Attacks in MRI

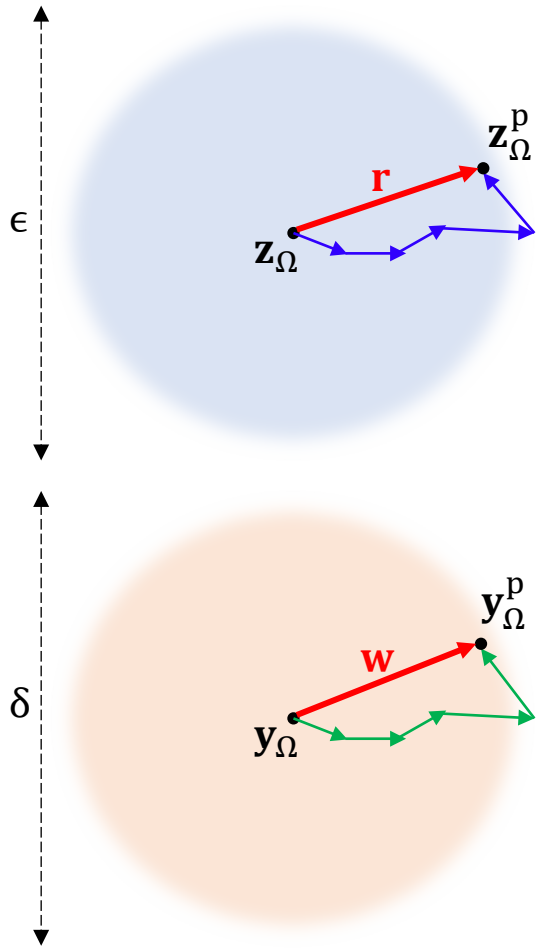
- PD-DL MRI reconstructions are vulnerable to adversarial attacks[1,2,3]
- Small perturbations → large reconstruction errors[1,2,3]
- Worst-case perturbations in two domains:
  - Image domain (ℓ_∞ attacks):

$$\arg \max_{\mathbf{r}:\, \|\mathbf{r}\|_\infty < \epsilon} \mathcal{L}\big(f(\mathbf{z}_\Omega + \mathbf{r}, \mathbf{E}_\Omega; \boldsymbol{\theta}), f(\mathbf{z}_\Omega, \mathbf{E}_\Omega; \boldsymbol{\theta})\big)$$

  - k-space domain (ℓ_2 attacks):

$$\arg \max_{\mathbf{w}:\, \|\mathbf{w}\|_2 < \delta} \mathcal{L}\left(f\left(\mathbf{E}_\Omega^{\mathbf{H}}(\mathbf{y}_\Omega + \mathbf{w}), \mathbf{E}_\Omega; \boldsymbol{\theta}\right), f\left(\mathbf{E}_\Omega^{\mathrm{H}}\mathbf{y}_\Omega, \mathbf{E}_\Omega; \boldsymbol{\theta}\right)\right)$$

- Typically solved by:
  - Fast gradient sign method (FGSM)[4]
  - Projected gradient descent (PGD)[5]

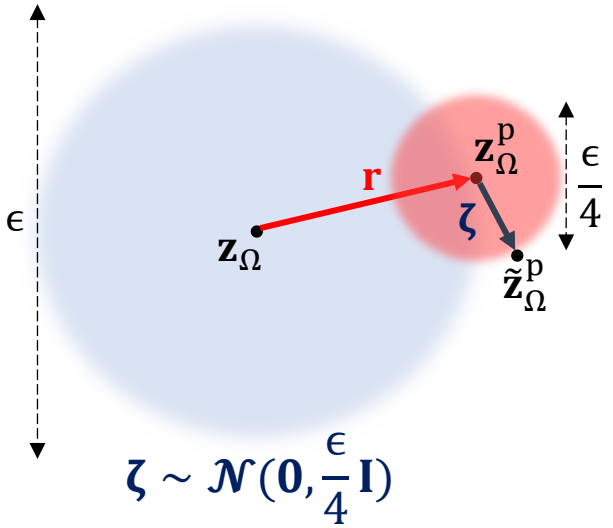


ICML International Conference On Machine Learning

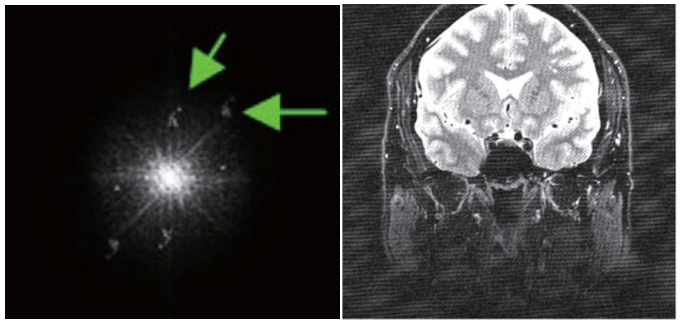
[1] Antun et al, PNAS., 2020; [2] Caliva et al, MELBA., 2021; [3] Zhang et al, Asilomar Conf.. Sig. Sys. and Comp., 2021;
[4] Goodfellow et al, ICLR , 2015; [5] Mkadry et al, ICLR., 2018;

Motivation – Why Adversary in MRI?

- **Worse-case perturbations – Non-zero probability,**
  - Practical[1] and theoretical[2] evidence
  - Sample from the perturbed input,
    - ➢ Another failed reconstruction[1]
  - Sample Gaussian noise from perturbed input,
    - ➢ Instability with non-zero probability[2]



- **Herringbone artifacts,**
  - Other MRI measurements perturbations: Body motion[3], Hardware issues[4]
  - Electromagnetic spikes:
    - ➢ Gradient power fluctuation
    - ➢ Inadequate room shielding
  - Impulse noise in k-space → Herringbone artifact in image domain[5,6]



ICML International Conference On Machine Learning

[1] Antun et al, PNAS., 2020; [2] Gottschling et al, SIAM., 2025; [3] Zaitsev et al, Journal of MRI., 2015;
[4] Kashani et al, American Journal of Neuroradiology, 2020; [5] Stadler et al, European Radiology, 2007; [5] Jin et al, MRM, 2017;

Defense Against Adversaries in MRI

- **Training-based defenses:**
  - Adversarial training (**AT**)[1,2]
  - Smooth unrolling (**SMUG**)[3]

- **Training-free defenses:**
  - Randomized smoothing (**RS**)[3,4]
  - **JPEG** compression[5]
  - Total variation (**TV**)[6] denoising

ICML International Conference On Machine Learning

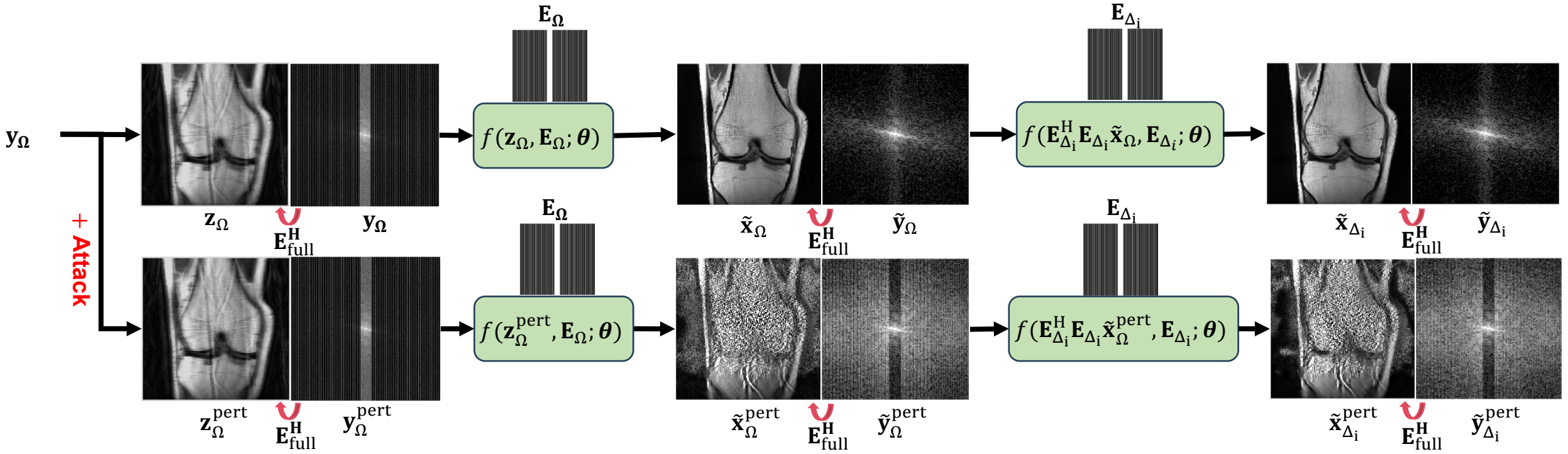
[1] Jia et al, NeurIPS Workshop., 2022;
[2] Raj et al, ICML, 2020;
[3] Liang et al, IEEE JSTSP, 2025;
[4] Cohen et al, PMLR, 2017;
[5] Cucu et al, ICSTCC, 2023;
[6] Sheikh et al, Journal of Imaging Informatics in Medicine, 2024;

Attack Propagation in Simulated k-space

- A well-trained model generalizes to undersampling patterns drawn from similar distribution[1]
- For Ω, Δ_i drawn with same acceleration rate, and number of ACS lines:

$$\tilde{\mathbf{x}}_{\Omega} = f(\mathbf{z}_{\Omega}, \mathbf{E}_{\Omega}; \boldsymbol{\theta})$$

$$\tilde{\mathbf{x}}_{\Delta_i} = f(\mathbf{E}_{\Delta_i}^{\mathrm{H}} \mathbf{E}_{\Delta_i} \tilde{\mathbf{x}}_{\Omega}, \mathbf{E}_{\Delta_i}; \boldsymbol{\theta})$$



[1] Knoll et al, IEEE Sig. Proc. Mag., 2020;

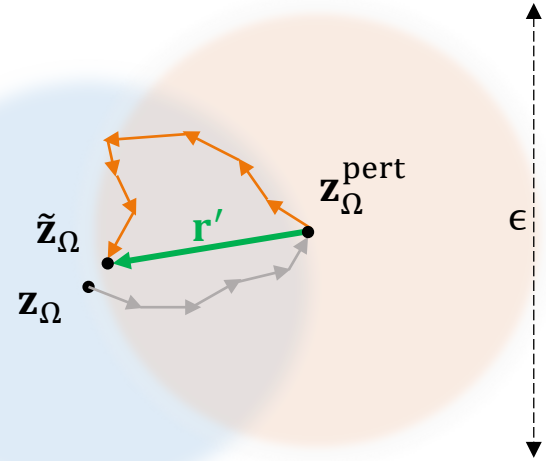
[2] Saberi et al, ICML, 2026;

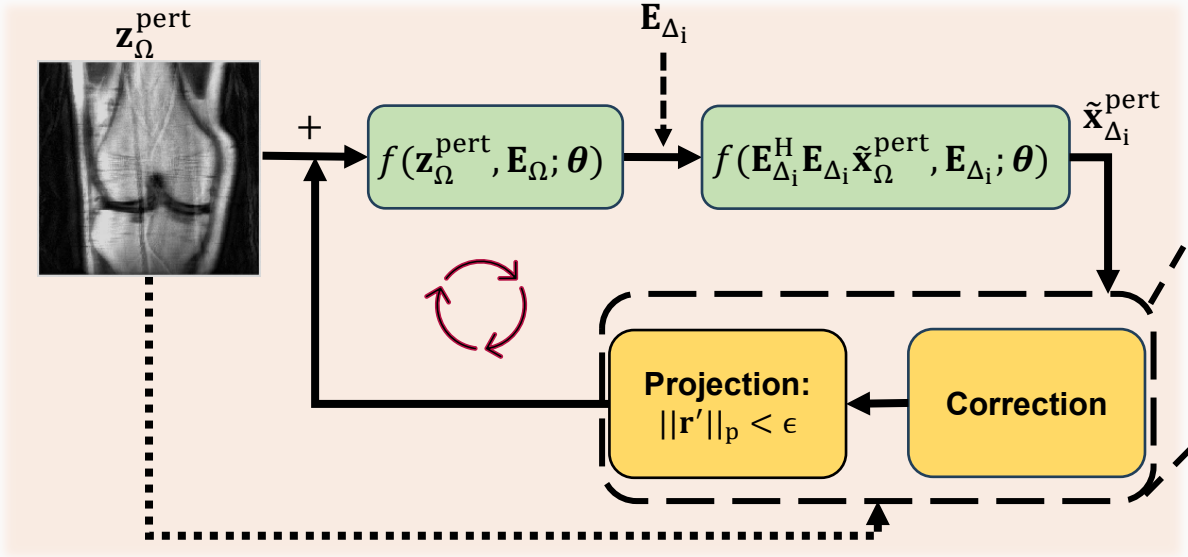
Attack Mitigation with Cyclic Consistency

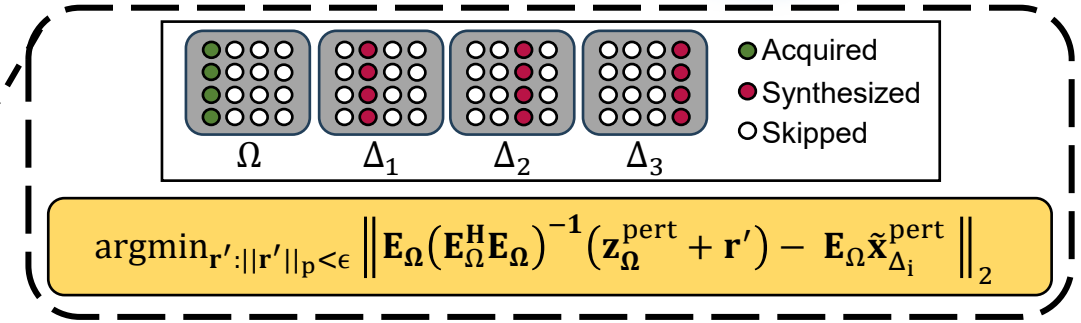
- Adversarial attack: Small perturbation within a ball around clean input
- **Goal:** reverse this process
- We devise a novel objective based on ***cyclic consistency***:

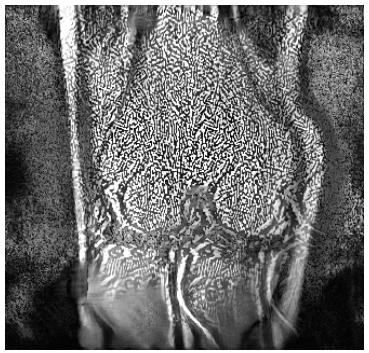
$$\arg \min_{\mathbf{r}':||\mathbf{r}'||_p<\epsilon} \mathbb{E}_{\Delta} \left[\left\| \left(\mathbf{E}_{\boldsymbol{\Omega}}^{\boldsymbol{H}}\right)^{\dagger}\left(\mathbf{z}_{\boldsymbol{\Omega}}^{\text{pert}} + \mathbf{r}'\right) - \mathbf{E}_{\Omega} f\left(\mathbf{E}_{\Delta_i}^{\mathbf{H}}\left(\mathbf{E}_{\Delta_i} f\left(\mathbf{z}_{\Omega}^{\text{pert}} + \mathbf{r}', \mathbf{E}_{\Omega}; \boldsymbol{\theta}\right) + \tilde{\mathbf{n}}\right), \mathbf{E}_{\Delta_i}; \boldsymbol{\theta}\right) \right\|_2 \right]$$

$\left(\mathbf{E}_{\boldsymbol{\Omega}}^{\boldsymbol{H}}\right)^{\dagger}\left(\mathbf{z}_{\boldsymbol{\Omega}}^{\text{pert}} + \mathbf{r}'\right)$: minimum ℓ_2 k-space solution of zerofilled image[1]





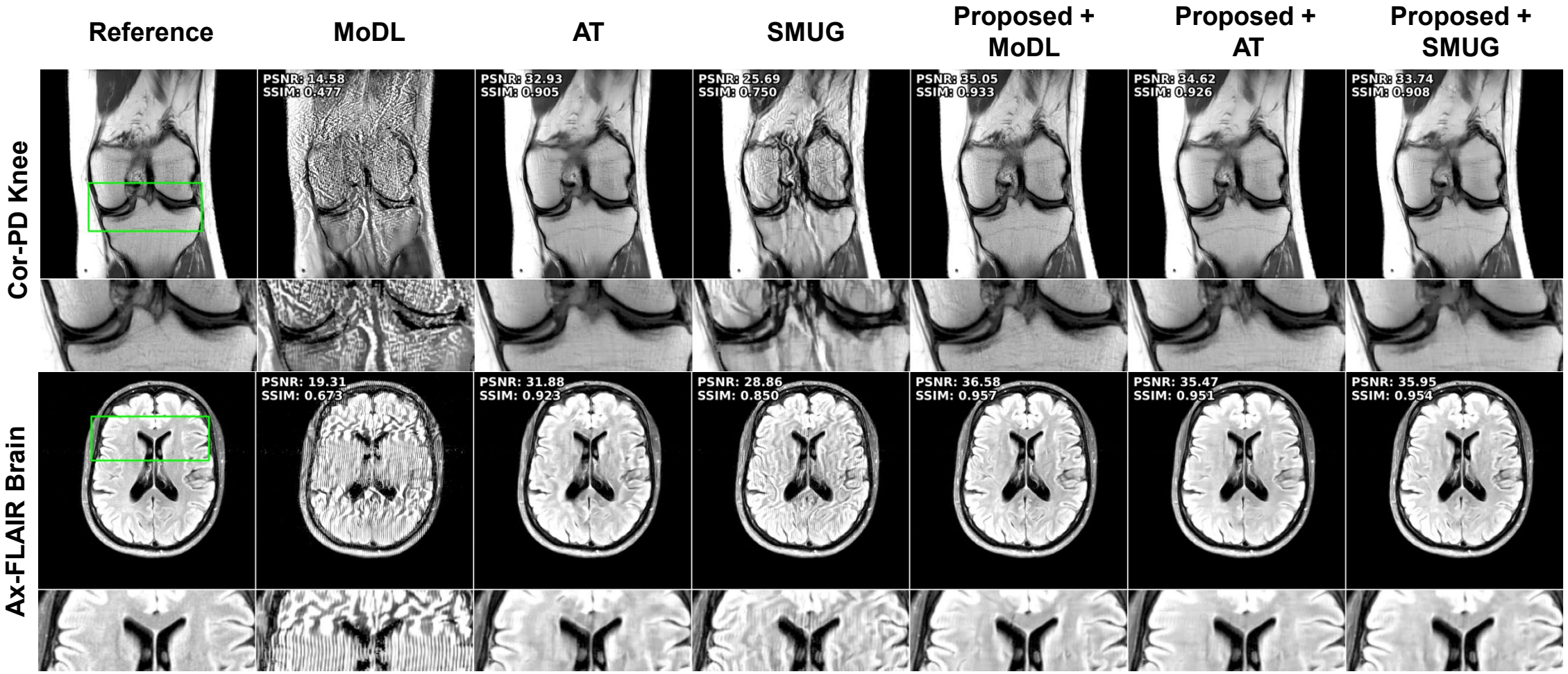




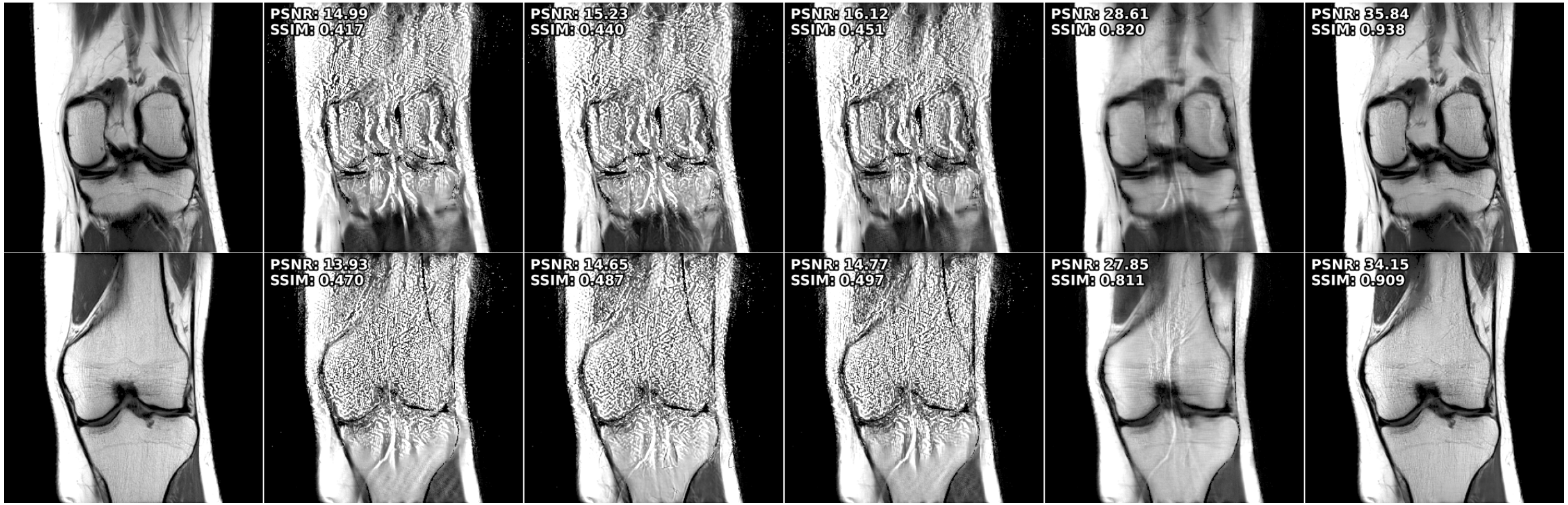
ICML International Conference On Machine Learning

[1] Zhang et al, Asilomar Conf.. Sig. Sys. and Comp., 2021;

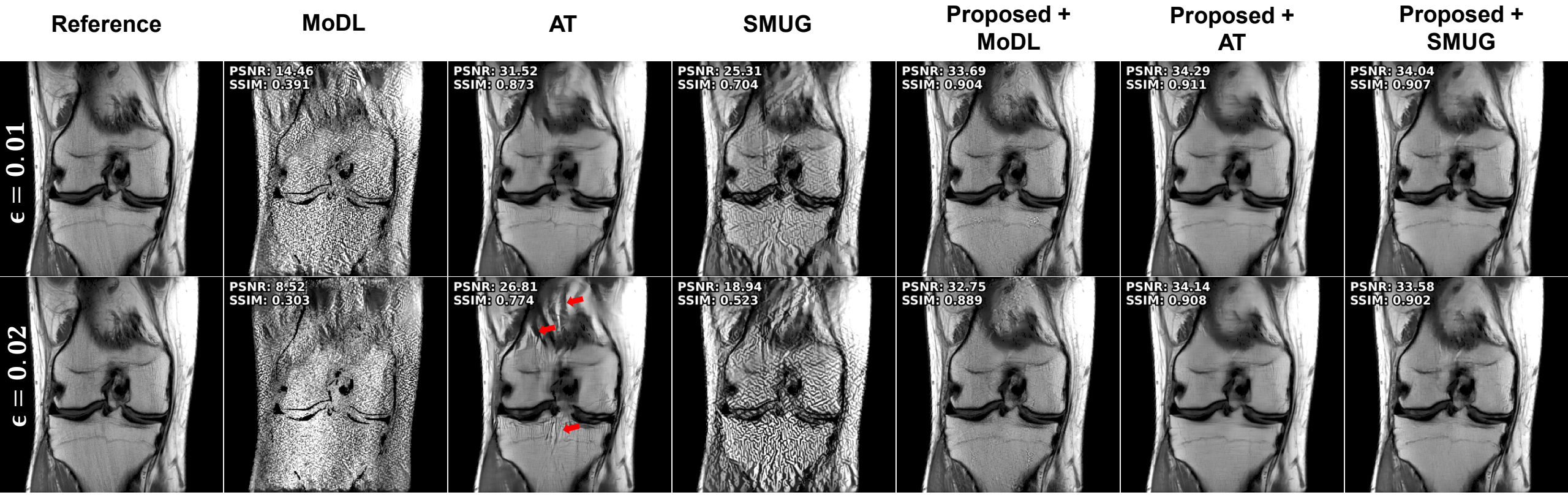
Ours vs. Training-Based Defenses



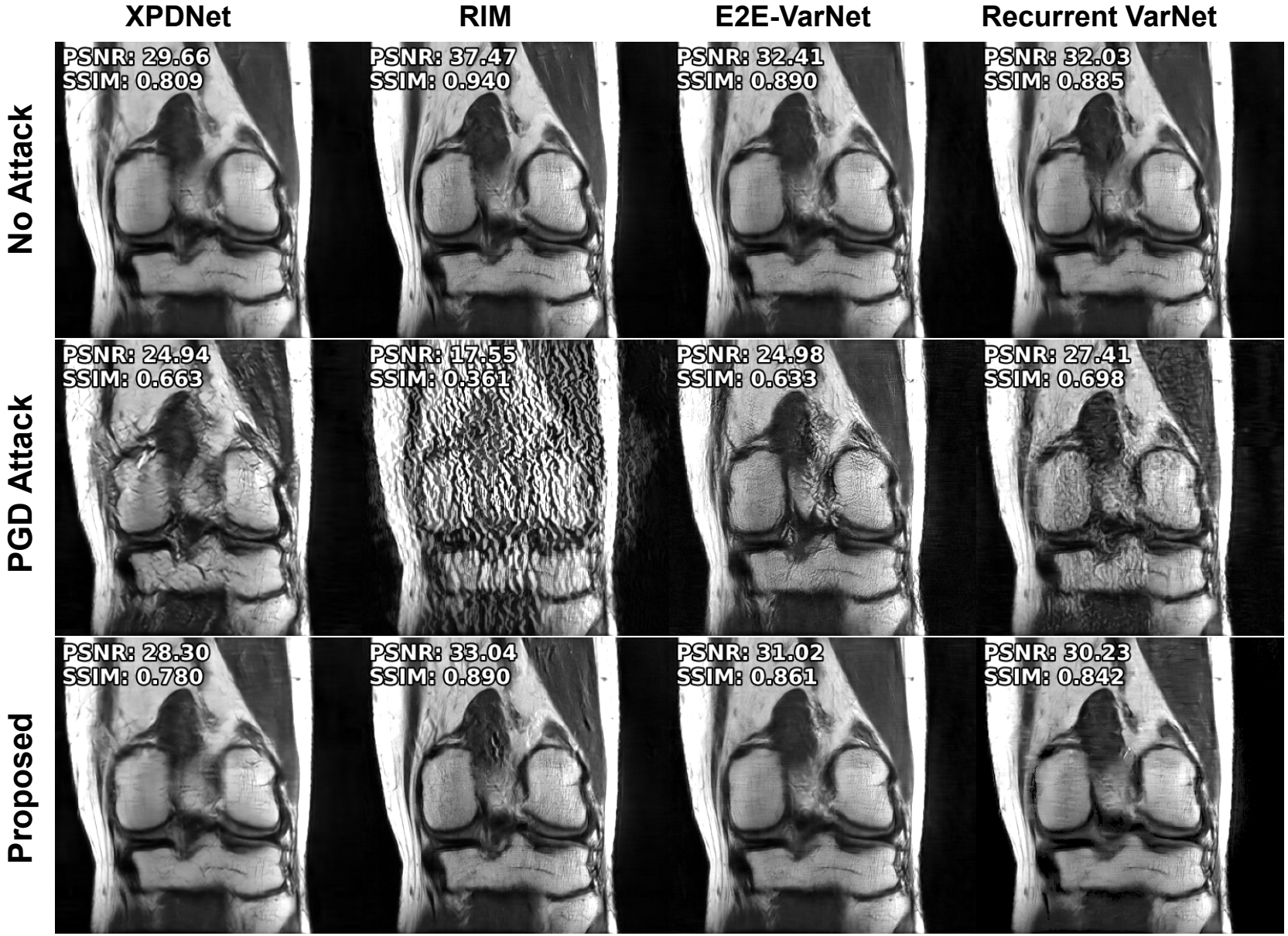
Ours vs. Training-Free Defenses



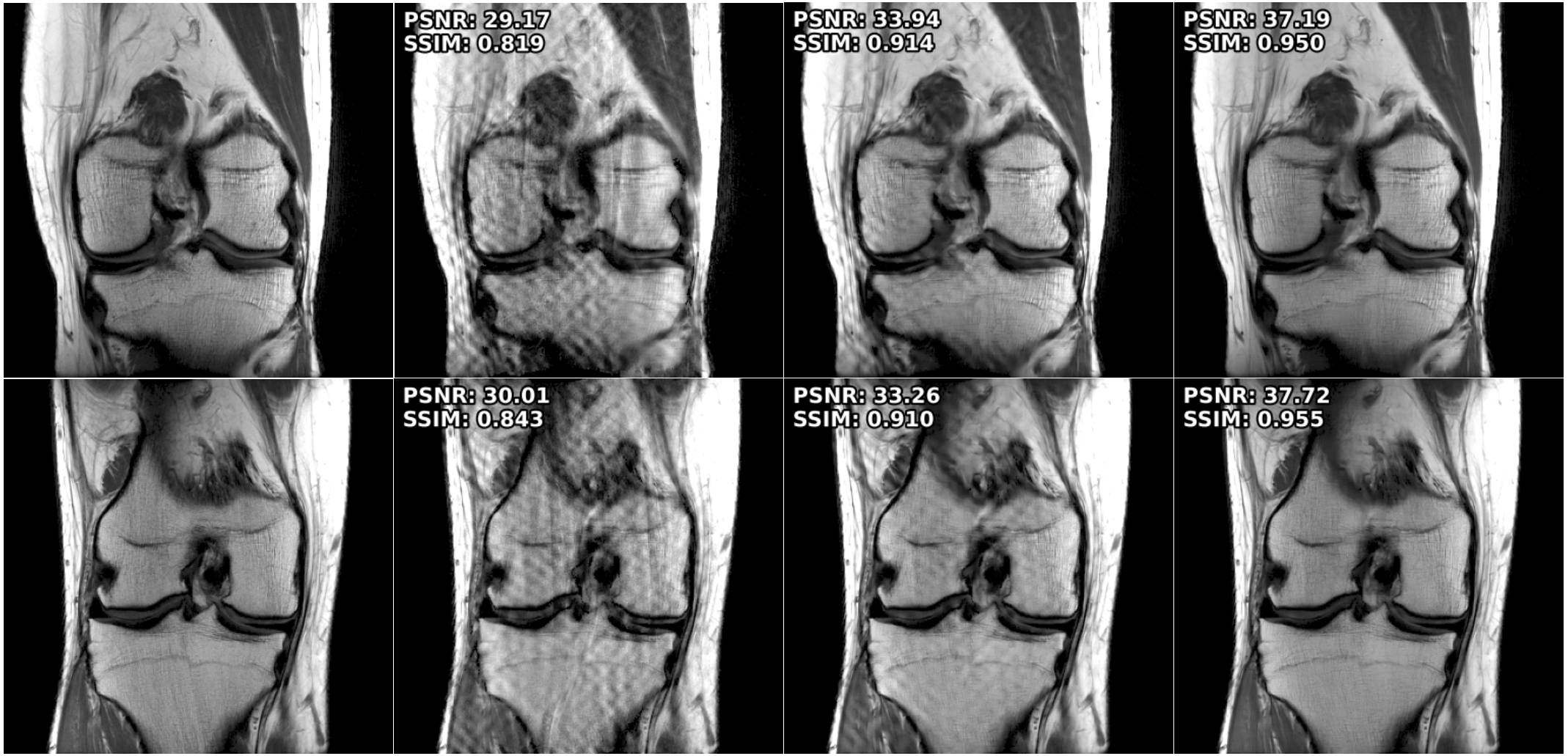
Results on Different Attack Strengths



Ours with Various Unrolled Networks



Practical Use Case– Herringbone Artifacts



Additional Findings

- **In this paper, we also show our method works for many cases:**
  - Adaptive attacks
  - Supervised attacks
  - ℓ_2 attacks in k-space
  - Non-uniform undersampling patterns
  - Blind setup with no predefined ϵ
  - Non-optimal reconstruction conditions
  - Other inverse problems (i.e., inpainting)

Conclusion

- **In this work, we proposed a novel mitigation strategy for computational MRI,**
  - Leverages cyclic measurement consistency
  - Optimizes the input within a small neighborhood of the attack
  - Requires no training or modification of network

- **Results demonstrate,**
  - Our method is robust across different,
    - Datasets
    - Inverse problems
    - Unrolled networks
    - Attack strengths, Adaptive attacks
  - Our method can,
    - Perform in a blind setup
    - Be combined with existing training-based methods

ICML International Conference On Machine Learning

Thank you!

IMAGINE
imagine.umn.edu

IMAGINE Lab,
University of Minnesota

ICML
International Conference
On Machine Learning

Questions?

Code & Paper

Supported by,

NIH R01EB032830
NIH P41EB027061