Evaluating Adversarial Robustness of Low dose CT Recovery

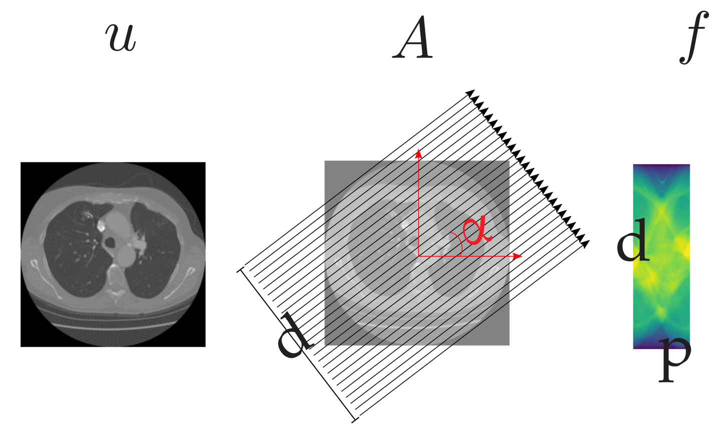
Kanchana Vaishnavi Gandikota, Paramanand Chandramouli, Hannah Droege, Michael Möller
University of Siegen

Background

Computed Tomography (CT):

✓ Diagnosis of various health conditions

✗ Radiation induced health risks.



Low-dose CT : Target is radiated with low-power radiation and/or using fewer projection angles.

✗ Noisy and severely ill-posed reconstruction.

CT Reconstruction

Reconstruction of a CT image u from a measured sinogram f

$$f = Au + n,$$

- Filtered back projection
- Algebraic reconstruction techniques
- Variational methods
$$\hat{u} = \arg\min_u \frac{1}{2}\|Au - f\|^2 + R(u)$$
- Neural Networks $\hat{u} = \mathcal{N}_\theta(f)$

Code

https://github.com/KVGandikota/robustness-low-dose-ct/

References

[1] S. G. Armato III et al. The lung image database consortium (lidc) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 2011.

[2] D. O. Baguer et al. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9), 2020.

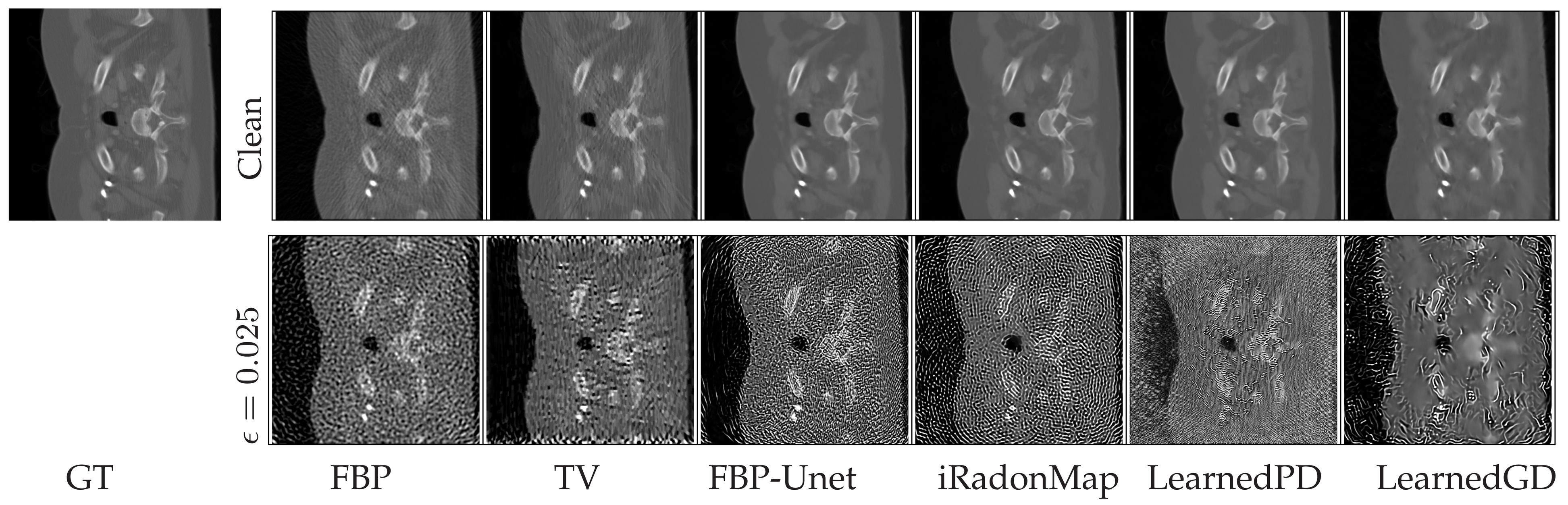
Adversarial Robustness Evaluation of CT Reconstruction

Robustness to Untargeted Attacks

Untargeted attacks find an additive image perturbation that maximizes the reconstruction error subject to L_∞ constraints on the perturbation.

$$\delta_{adv} = \arg\max_\delta \|\mathcal{N}_\theta(f+\delta) - \mathcal{N}_\theta(f)\|_2 \text{ s.t. } \|\delta\|_\infty \le \epsilon.$$

Method	$\hat{u}$ PSNR	$(A\hat{u}, f)$ PSNR	$\hat{u}_\delta$, $\epsilon = 0.01$ PSNR	$(A\hat{u}_\delta, f)$ PSNR	L_b Empir
FBP	30.37	33.82	25.18	33.36	**15.03**
TV	31.62	36.52	25.20	35.62	16.52
FBP-Unet	35.47	36.47	18.39	35.06	46.71
iRadonMap	33.94	36.03	17.98	29.62	43.80
LearnedPD	**35.73**	36.46	**9.47**	25.27	**143.39**
LearnedGD	34.55	36.43	21.14	35.18	30.48



Lipschitz lower-bound $L_b(\mathcal{N}_\theta) = \max_i \left(\frac{\|\mathcal{N}_\theta(f_i + \delta_i) - \mathcal{N}_\theta(f_i)\|}{\|\delta_i\|} \right)$

- Classical approaches FBP & TV are slightly more robust than neural networks.
- TV is better than FBP in terms of SSIM and Bregman distance.
- Consistency with the original sinogram is less affected than accuracy.

Universal Attacks & Transferability

Universal Attacks we find an adversarial perturbation that maximizes the reconstruction error of a recovery method $\mathcal{N}_\theta$ for any input subject to L_∞ norm constrains on the perturbation.

$$\delta_{uniadv} = \arg\max_\delta \sum_{\text{examples i}} \|\mathcal{N}_\theta(f_i + \delta) - \mathcal{N}_\theta(f_i)\|_2 \text{ s.t. } \|\delta\|_\infty \le \epsilon.$$

- Universal attacks are both feasible and transferable (ϵ=0.05 in below)

Source Noise	FBP	FBP-Unet	iRadonMap	LearnedGD	LearnedPD
Clean	30.53/0.714	35.67/0.824	34.19/0.799	34.74/ 0.802	35.92/0.829
FBP	**10.34/0.036**	9.90/0.031	8.74/0.025	7.68/0.021	10.62/0.041
FBP-Unet	14.42/0.098	**4.95/0.022**	9.06/0.035	9.26/ 0.095	7.77/0.042
iRadonMap	13.02/0.0706	9.61/0.049	**3.82/0.0108**	7.38/0.042	10.99/0.057
LearnedGD	15.60/0.188	13.52/0.220	10.38/0.112	**4.32/0.183**	9.69/0.109
LearnedPD	23.07/0.358	21.42/0.444	19.45/0.232	23.54/0.453	**-2.95/0.003**

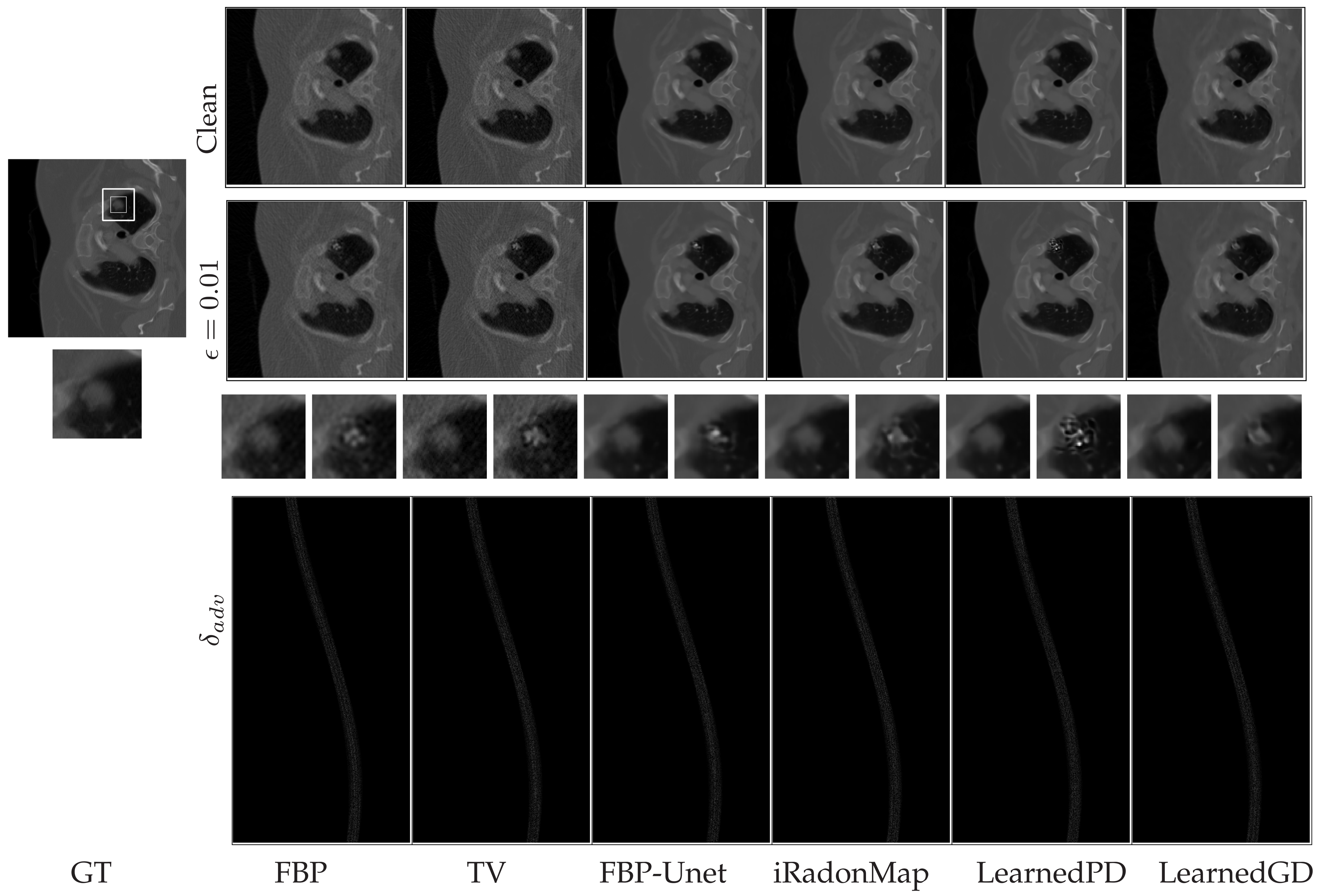
Robustness to Localized Attacks

Localized attacks find an additive perturbation that produces a change in the visual appearance within a localized clinically relevant region g_c of the reconstruction using an *adversarially trained* classifier $\mathcal{G}_\phi$.

$$\delta_{adv} = \arg\max_\delta E(\mathcal{G}_\phi (g_c (\mathcal{N}_\theta(f+\delta))), y) \text{ s.t. } \|\delta\|_\infty \le \epsilon.$$

- Change in malignancy prediction of robust classifier $\mathcal{G}_\phi$ requires visible change in reconstruction within g_c.
- Apply a smoothed Gaussian mask to the adversarial noise to localize degradation, and avoid boundary artifacts.

Method	Clean $\hat{u}$ PSNR	Clean $\hat{u}_i \mid \hat{u}_e$ PSNR	Clean $(A\hat{u}, f)$ PSNR	$\epsilon = 0.01$ $\hat{u}_\delta$ PSNR	$\epsilon = 0.01$ $\hat{u}_{\delta_i} \mid \hat{u}_{\delta_e}$ PSNR	$\epsilon = 0.01$ $(A\hat{u}_\delta, f)$ PSNR
FBP	30.86	31.45|30.86	33.81	30.60	22.29 | 30.83	33.79
TV	32.36	31.84|32.37	36.52	32.00	22.70 | 32.32	36.48
FBP-Unet	36.94	35.67|36.95	36.50	34.85	19.43 | 36.61	36.46
iRadonMap	35.25	34.07|35.27	36.09	33.70	18.85 | 35.12	36.03
LearnedPD	37.22	35.97|37.23	36.49	33.15	18.34 | 35.08	36.28
LearnedGD	35.80	34.86|35.82	36.49	34.86	22.02 | 35.71	36.46



- Local attacks preserve consistency with the original sinogram.
- Changes visual appearance in the local region without affecting exterior region.
- Attack also changes predicted malignancy of robust classifier $\mathcal{G}_\phi$.
- Both classical approaches & neural networks are susceptible to localized attacks.