Evaluating Adversarial Robustness of Low dose CT Recovery

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Low dose computer tomography (CT) acquisition using reduced radiation or sparse angle measurements is recommended to reduce the harmful effects of X-ray radiation. However, such low dose acquisition can result in poor quality reconstructions. Recently, deep networks have been applied to the problem of low dose and sparse view CT recovery. These methods have demonstrated high quality reconstructions, surpassing classical approaches on benchmark CT recovery datasets. However, robustness of CT recovery approaches needs be carefully evaluated before they can be used in clinical applications. In this paper, we evaluate the robustness of CT reconstruction approaches to adversarial attacks. We evaluate different end-to-end trained neural networks and model inspired architectures, in addition to classical approaches of filtered back projection and variational approach using total variation regularization. We consider untargeted attacks and localized attacks which aim to change the visual appearance of a localized lesion in the reconstructed CT image. We demonstrate that imperceptible distortion can significantly degrade the performance of several CT reconstruction approaches.

Keywords: Computer tomography, robustness, adversarial attacks, image reconstruction.

1. Introduction

Computer tomography (CT), is a non-invasive medical imaging technique used to diagnose a wide range of medical conditions, such as cancer, heart disease, and internal injuries. A scanner rotates around a target and projects X-ray radiation from different angles and the attenuated radiation is recorded on the other side by detectors. This is arranged into a sinogram, from which a CT image is reconstructed. The amount of radiation that is attenuated varies depending on the density of the tissue, and recording this attenuation at different angles aids in creating a detailed image of internal structures of the target. While the accuracy and resolution of CT images improves with number of X-ray beams used, exposure of patients to X-rays poses serious health risks. To reduce the effect of ionizing X-ray radiation, different solutions to low-dose CT acquisition have been proposed under the ALARA (as low as reasonably achievable) principle (Slovis, 2002; Newman and Callahan, 2011). These protocols can be broadly classified into two categories- i) adjusting the settings on the CT scanner tube to reduce total number of X-ray photons ii) recording measurements from fewer projection angles. However, there exists a trade-off between dose reduction during CT acquisition and diagnostic quality. Lower number of X-ray photons degrades reconstruction quality due to increased image noise level. On the other hand, CT recovery using measurements from fewer projection angles can suffere from severe artefacts. Further, sparse-view CT is an ill-posed problem, and therefore there can be many valid solutions for the same measurement.

© 2023 Anonymous.

Traditional approaches to ill-posed inverse problems impose suitable priors such as total variation (Sidky et al., 2006; Chen et al., 2013) in a variational reconstruction algorithm. Recently deep learning approaches are being applied in many image reconstruction tasks, including sparse view CT reconstruction (Chen et al., 2017; He et al., 2020). While end to end trained deep networks achieve impressive performance, they lack convergence guarantees provided by classical approaches to inverse problems. Further, deep neural networks are known to be susceptible to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015), which is a serious concern when using such deep networks for clinical applications. In this paper, we investigate the robustness of deep neural networks trained for CT reconstruction to norm bounded additive adversarial perturbations. We evaluate different neural network based approaches, including model-inspired architectures. In addition, we also evaluate the robustness of classical approaches, filtered back projection and energy minimization using TV regularization. We evaluate robustness to both targeted and untargeted adversarial examples. Our experiments demonstrate.....

2. Related Work

Deep learning for Image and CT Reconstruction Recently deep learning approaches have achieved impressive performance in many image reconstruction tasks. These encompass a wide array of methods including i) fully learned methods to directly invert the forward imaging model (Zhu et al., 2018; Kupyn et al., 2018), examples include iRadonmap (He et al., 2020), ADAPTIVE-Net (Ge et al., 2020) for CT recovery, which also learn filtered back projection operation ii) learning deep neural network denoisers to post-process an initial reconstruction, for instance, using a post-processing network to recover high quality CT images from filtered-back-projection outputs (Chen et al., 2017; Jin et al., 2017; Yang et al., 2018) iii) methods that incorporate the forward imaging model into end-to-end trained neural networks, for example, by unrolling iterative algorithms such as gradient descent, primal-dual hybrid gradient for a fixed number of steps (Adler and Öktem, 2017; Aggarwal et al., 2018; Adler and Öktem, 2018) iv) use of trained or untrained neural networks as priors in a variational inference (Bora et al., 2017; Rick Chang et al., 2017; Meinhardt et al., 2017). In context of CT recovery (Baguer et al., 2020) use untrained neural network prior to improve over initial reconstruction from trained neural network, (Gupta et al., 2018) train a CNN to perform projection step in an iterative projected gradient descent algorithm. Instead of recovering a single solution, another line of work (Bahat and Michaeli, 2020; Dröge et al., 2022) propose to explore the space of solutions which satisfy data consistency, as the solution to ill-posed problems is not unique.

Adversarial Attacks on Image Reconstruction Adversarial attacks refers to a phenomenon where a carefully crafted imperceptible change in the input causes a catastrophic failure of neural networks. Starting from (Szegedy et al., 2014; Goodfellow et al., 2015) a lot of research was done on creating strong adversarial examples and defense mechanisms to improve adversarial robustness of networks, mainly in the context of classification networks. Recent works (Antun et al., 2019; Raj et al., 2020) have demonstrated the susceptibility of image reconstruction networks to adversarial attacks. (Antun et al., 2019) investigate instabilities to perturbations in the image domain mainly consider models doing direct inversion or using post processing networks. (Raj et al., 2020) investigate robustness of fully

learned reconstruction networks to additive adversarial perturbation in the measurement domain and propose to perform adversarial training to improve robustness. However, these works consider only non-targeted attacks. Similarly, recent works (Gandikota et al., 2022; Choi et al., 2019) investigated the adversarial robustness of image restoration methods. In the context of CT recovery, (Huang et al., 2018) perform preliminary investigations if some additive perturbations will mislead a neural network to fail to detect an existing lesion. Closely related to our work, (Genzel et al., 2022) investigate the adversarial robustness of different reconstruction approaches for CT and MRI recovery. However, this is not evaluated on medical CT data, and this work mainly considered untargeted attacks. In contrast, we investigate susceptibility of reconstruction methods to both untargeted and localized adversarial attacks. We perform our experiments on LoDoPaB dataset (Leuschner et al., 2019) which has simulated low-intensity measurements for thoracic CT scans from(Armato III et al., 2011).

3. CT Acquisition and Reconstruction

CT acquisition involves projecting X-ray beams from different angles and recording the attenuation of X-ray as a a sinogram. The forward operator is given by the 2D Radon transform (Radon, 1986) which models the attenuation of the radiation passing through the target by calculating line integral along the path of X-ray beam. The measurement, which is a sinogram consists of the recorded integrals for different distances and measurement angles. Since the Radon transform is linear, the measurement process can be written as:

$$f = Au + n \tag{1}$$

where f, A, u, n represent the sinogram, forward Radon transform, the ground truth image and measurement noise respectively. The aim of CT reconstruction is to recover a CT image $\hat{u}$ from the sinogram f. Filtering the linear pseudo inverses, commonly referred to as filtered back projection (FBP) (Feldkamp et al., 1984) is one standard classical approach to CT recovery. Variational approaches (Sidky et al., 2006; Chen et al., 2013) find minimizer of energy function

$$\hat{u} = \arg\min_u \frac{1}{2}\|Au - f\|^2 + R(u) \tag{2}$$

for a suitable regularizer $R(u)$ such as total variation. Recent approaches employ deep learning for recovering CT images in single feed forward pass using :
i) end-to-end trained networks (He et al., 2020; Ge et al., 2020; Adler and Öktem, 2017; Aggarwal et al., 2018; Adler and Öktem, 2018) to directly predict clean CT image

$$\hat{u} = \mathcal{N}(f) \tag{3}$$

ii) learning post-processors (Chen et al., 2017; Jin et al., 2017; Yang et al., 2018) to denoise outputs from FBP operation $B^\dagger(\cdot)$

$$\hat{u} = \mathcal{N}(B^\dagger(f)) \tag{4}$$

4. Analyzing Stability of (CT) Image Recovery

Ideally, the reconstruction algorithm or network $\mathcal{N}$ should have a small Lipschitz constant L to ensure that small changes in the input produce only small bounded changes in the reconstruction.

$$\|\mathcal{N}(f_1) - \mathcal{N}(f_2)\| \leq L\|f_1 - f_2\|$$

However, exactly computing Lipschitz constant for neural networks has extremely high computational complexity (Jordan and Dimakis, 2020) even for moderately sized neural networks, and recent works (Combettes and Pesquet, 2020; Jordan and Dimakis, 2020; Huang et al., 2021) instead estimate an upper bound on Lipschitz constant. On the other hand, it is easier to analyze the stability of classical approaches. Stability of classical FBP recovery depends on the spectral radius of the filtered pseudo-inverse operator $B^\dagger$. For variational energy minimization approaches, error estimates can be derived in terms of Bregman distance (Burger et al., 2007):

$$\|f_1 - f_2\|^2 \geq \|Au_1 - Au_2\|^2 + 2\langle \nabla R(u_1) - \nabla R(u_2), u_1 - u_2\rangle \tag{5}$$

where the term $\langle \nabla R(u_1) - \nabla R(u_2), u_1 - u_2\rangle$ is called the 'symmetric Bregman distance' with respect to the regularizer R.

Further, in this work we empirically analyze the robustness of CT recovery approaches to adversarial attacks.

4.1. Adversarial Attacks

Adversarial attacks on image reconstruction methods make small changes the inputs to cause unpredictable large changes in the output. In this work, we consider robustness to tiny norm bounded additive perturbations in the measurement domain. We assume that the parameters of the neural network $\mathcal{N}$ or the recovery algorithm is fully known to the attacker.
Untargeted Attacks: Here the aim is to find an additive perturbation in the measurement domain that maximizes the reconstruction error subject to L_p constraints on the perturbation,

$$\delta_{adv} = \underset{\delta \in \mathrm{R}^m}{\operatorname{argmax}} \|\mathcal{N}(f+\delta) - \mathcal{N}(f)\|_2 \text{ s.t. } \|\delta\|_p \leq \epsilon. \tag{6}$$

Localized Attacks: Here the goal is to find an additive L_p norm constrained perturbation in the that produces a localized change in the reconstruction. Specifically, we investigate whether a tiny perturbation changes the reconstruction in a localized region to cause a classifier to change the predicted malignancy of a nodule. As this classification is also susceptible to adversarial attacks without actual change in the malignancy level in the region of interest, we utilize an adversarially trained classification network $\mathcal{N}_\theta$ pre-trained for classifying nodules in chest CT. Our localized attack can be formulated as:

$$\delta_{adv} = \underset{\delta \in \mathrm{R}^m}{\operatorname{argmax}} E(\mathcal{N}_\theta\left(g_c\left(\mathcal{N}(f+\delta)\right)\right), \mathcal{N}_\theta\left(g_c\left(\mathcal{N}(f)\right)\right)) \text{ s.t. } \|\delta\|_p \leq \epsilon. \tag{7}$$

where $g_c(\cdot)$ crops the region of interest, and $E(\cdot)$ refers to the energy function (loss) to be maximized, for binary classification of nodules, this corresponds to the binary cross entropy loss. Further, in addition to flipping the label of the malignancy prediction, we can also

modify the attack such that the reconstructed output region corresponds to a specific target malignancy level m. This can be formalized as:

$$\delta_{adv} = \underset{\delta \in \mathrm{R}^m}{\operatorname{argmin}} E(\mathcal{N}_\theta \left(g_c \left(\mathcal{N}(f+\delta)\right)\right), m)) \text{ s.t. } \|\delta\|_p \leq \epsilon. \tag{8}$$

In this work we consider L_∞ norm bounded perturbations. Similar attacks can also be defined for other L_p norm bounded perturbations. We solve the constrained optimization problems (6), (7) using the projected gradient descent (PGD) algorithm (Madry et al., 2018), with gradient updates using the Adam optimizer (Kingma and Ba, 2015).

5. Experiments

We conduct experiments with low-dose parallel beam (LoDoPaB) CT dataset (Leuschner et al., 2019), consisting data pairs of simulated low-intensity measurements and corresponding ground truth human chest CT reconstructions from the LIDC/IDRI dataset(Armato III et al., 2011). We consider two version of this dataset in our experiments, LODoPaB and LoDoPaB_200 which correspond to sampling 513 and 200 projection beams respectively from 1000 parallel projection beams. We evaluate the robustness of the following approaches: i) Filtered back projection(FBP) ii) FBP-Unet,using UNet to post-process FBP outputs, iii)iRadonmap, a fully learned approach, which also learns back projection in addition to pre-processing, iv) LearnedGD, learned gradient descent v) LearnedPD, learned Primal Dual method vi) Variational reconstruction with Total Variation regularization. For the learned methods, we use the pretrained models[1] from (Baguer et al., 2020) trained on the full training set, with exception to iRadonmap which we trained ourselves.

Method	LoDoPaB				LoDoPaB_200			
	clean	$\epsilon = 0.01$	$\epsilon = 0.025$	$\epsilon = 0.05$	clean	$\epsilon = 0.01$	$\epsilon = 0.025$	$\epsilon = 0.05$
FBP	30.37/0.738	25.18/0.448	18.68/0.194	13.02/0.074	28.38/0.649	25.26/0.465	19.76/0.233	14.36/0.096
TV								
FBP-Unet	35.47/0.837	18.39/0.287	12.19/0.095	7.38/0.034	33.55/0.799	19.37/ 0.384	12.82/0.115	8.38/0.036
iRadonMap	33.94/0.810	17.98/0.326	10.85/0.084	6.24/0.026				
LearnedPD	35.73/0.843	9.47/0.164	3.38/0.030	0.361/0.008	33.64/0.802	17.75/0.412	10.56/0.153	5.94/0.053
LearnedGD	34.55/0.815	21.14/0.504	13.90/0.291	8.64/0.180	32.49/0.776	22.44/0.583	15.66/0.418	10.89/0.301

Table 1: Comparison of robustness to untargeted attacks on different CT reconstruction methods when sampling 513 and 200 projection beams out of 1000 parallel beams using 20 attack iterations. Average PSNR/SSIM values on 100 samples LoDoPAB testset are reported.

Acknowledgments

Acknowledgments withheld.

1. https://github.com/oterobaguer/dip-ct-benchmark

Method	clean $\hat{x}$			$\epsilon = 0.01$				$\epsilon = 0.025$				$\epsilon = 0.05$		
	PSNR/SSIM	e_{diff}	$\hat{x}_{adv}$	e_{diff}	y_{adv}	$A\hat{x}_{adv}$	$\hat{x}_{adv}$	e_{diff}	y_{adv}	$A\hat{x}_{adv}$	$\hat{x}_{adv}$	e_{diff}	y_{adv}	$A\hat{x}_{adv}$
FBP														
TV														
FBP-Unet														
iRadonMap														
LearnedPD														
LearnedGD														

Table 2: **BUG: Rerun for new nodule locations** Comparison of robustness to localized attacks on different CT reconstruction method using 20 attack iterations. Average PSNR/SSIM values of reconstructions on 100 samples LoDoPAB testset are reported. Difference between interior and exterior errors is scaled by 10^4, i.e. $e_{diff} = (e_i - e_o) \times 10^4$

Algorithm 1: Computing Net Activation

Input: $x_1, \ldots, x_n, w_1, \ldots, w_n$
Output: y, the net activation
$y \leftarrow 0$;
for $i \leftarrow 1$ **to** n **do**
  $y \leftarrow y + w_i * x_i$;
end

References

Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on medical imaging*, 37(6):1322–1332, 2018.

H. K. Aggarwal, M. P. Mani, and M. Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on medical imaging*, 38(2):394–405, 2018.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction-does ai come at a cost? *arXiv preprint arXiv:1902.05300*, 2019.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, 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*, 38(2):915–931, 2011.

Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9):094004, 2020.

Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2716–2725, 2020.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.

M. Burger, E. Resmerita, and L. He. Error estimation for bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2):109–135, 2007.

H. Chen, Y. Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE Transactions on medical imaging*, 36(12):2524–2535, 2017.

Zhiqiang Chen, Xin Jin, Liang Li, and Ge Wang. A limited-angle ct reconstruction method based on anisotropic tv minimization. *Physics in Medicine & Biology*, 58(7):2119, 2013.

J. H Choi, H. Zhang, J. H. Kim, C. J Hsieh, and J. S. Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proc. IEEE/CVF International Conference on Computer Vision*, 2019.

P. L Combettes and J. C Pesquet. Lipschitz certificates for layered network structures driven by averaged activation operators. *SIAM Journal on Mathematics of Data Science*, 2(2): 529–557, 2020.

Hannah Dröge, Yuval Bahat, Felix Heide, and Michael Möller. Explorable data consistent ct reconstruction. In *Proc. British Machine Vision Conference*, 2022.

L. A. Feldkamp, L. C. Davis, and J. W. Kress. Practical cone-beam algorithm. *J. Opt. Soc. Am. A*, 1(6):612–619, Jun 1984. doi: 10.1364/JOSAA.1.000612. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-1-6-612.

Kanchana Vaishnavi Gandikota, Paramanand Chandramouli, and Michael Moeller. On adversarial robustness of deep image deblurring. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3161–3165. IEEE, 2022.

Yongshuai Ge, Ting Su, Jiongtao Zhu, Xiaolei Deng, Qiyang Zhang, Jianwei Chen, Zhanli Hu, Hairong Zheng, and Dong Liang. Adaptive-net: deep computed tomography reconstruction network with analytical domain transformation knowledge. *Quantitative Imaging in Medicine and Surgery*, 10(2):415, 2020.

M. Genzel, J. Macdonald, and M. März. Solving inverse problems with deep neural networks-robustness included. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

I. J Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *Proc. ICLR*, 2015.

Harshit Gupta, Kyong Hwan Jin, Ha Q Nguyen, Michael T McCann, and Michael Unser. Cnn-based projected gradient descent for consistent ct image reconstruction. *IEEE Transactions on medical imaging*, 37(6):1440–1453, 2018.

Ji He, Yongbo Wang, and Jianhua Ma. Radon inversion via deep learning. *IEEE Transactions on medical imaging*, 39(6):2076–2087, 2020.

Yixing Huang, Tobias Würfl, Katharina Breininger, Ling Liu, Günter Lauritsch, and Andreas Maier. Some investigations on robustness of deep learning in limited angle tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 145–153. Springer, 2018.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J. Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22745–22757. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c055dcc749c2632fd4dd806301f05ba6-Paper.pdf.

Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7344–7353. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/5227fa9a19dce7ba113f50a405dcaf09-Paper.pdf.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of. International Conference on Learning Representations*, 2015.

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maaß. The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods. *arXiv preprint arXiv:1910.01113*, 2019.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

T. Meinhardt, M. Moller, C. Hazirbas, and D. Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proc. IEEE/CVF International Conference on Computer Vision*, 2017.

Beverley Newman and Michael J Callahan. Alara (as low as reasonably achievable) ct 2011—executive summary. *Pediatric radiology*, 41(2):453–455, 2011.

J. Radon. On the determination of functions from their integral values along certain manifolds. *IEEE Transactions on Medical Imaging*, 5(4):170–176, 12 1986. ISSN 0278-0062. doi: 10.1109/TMI.1986.4307775.

A. Raj, Y. Bresler, and B. Li. Improving robustness of deep-learning-based image reconstruction. In *Proc. ICML*, volume 119 of *PMLR*, pages 7932–7942, 2020. URL http://proceedings.mlr.press/v119/raj20a.html.

JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 5888–5897, 2017.

Emil Y Sidky, Chien-Min Kao, and Xiaochuan Pan. Accurate image reconstruction from few-views and limited-angle data in divergent-beam ct. *Journal of X-ray Science and Technology*, 14(2):119–139, 2006.

Thomas L Slovis. The alara concept in pediatric ct: myth or reality? *Radiology*, 223(1):5–6, 2002.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *Proc. ICLR*, 2014.

Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K. Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging*, 37(6):1348–1357, 2018.

Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.

Appendix A. Additional Results

Method	LoDoPaB				LoDoPaB_200			
	clean	$\epsilon = 0.01$	$\epsilon = 0.025$	$\epsilon = 0.05$	clean	$\epsilon = 0.01$	$\epsilon = 0.025$	$\epsilon = 0.05$
FBP	30.52/ 0.737	24.99/0.432	18.44/ 0.182	12.78/0.069	28.45/0.645	25.16/0.453	19.58/0.223	14.17/0.091
TV								
FBP-Unet	35.84/0.844	18.14/0.271	12.04/0.09	7.23/0.03	33.89/0.810	19.18/0.375	12.66/0.113	8.30/0.036
iRadonMap								
LearnedPD	36.12/ 0.850	9.32/0.153	3.51/0.031	0.355/0.008	33.98/0.813	17.61/0.405	10.49/0.147	5.95/0.051
LearnedGD	34.88/0.824	20.92/0.499	13.71/0.285	8.53/0.173	32.78/0.787	22.36/0.587	15.59/0.421	10.94/0.302

Table 3: Comparison of robustness to untargeted attacks on different CT reconstruction methods when sampling 513 and 200 projection beams out of 1000 parallel beams using 20 attack iterations. Average PSNR/SSIM values on LoDoPAB testset are reported.

Appendix B. Proof of Theorem 1

This is a boring technical proof of

$$\cos^2\theta + \sin^2\theta \equiv 1. \tag{9}$$

Appendix C. Proof of Theorem 2

This is a complete version of a proof sketched in the main text.