Reconstructing Invariances of CT Image Denoising Networks using Invertible Neural Networks

Elias Eulig, Björn Ommer, and Marc Kachelrieß

***Abstract*—Long lasting efforts have been made to reduce radiation dose and thus the potential radiation risk to the patient for CT acquisitions without severe deterioration of image quality. To this end, different reconstruction and noise reduction algorithms have been developed, many of which are based on iterative reconstruction techniques, incorporating prior knowledge in the image domain. Recently, deep learning-based methods have shown impressive performance, outperforming many of the previously proposed CT denoising approaches both visually and quantitatively. However, with most neural networks being black boxes they remain notoriously difficult to interpret and concerns about the robustness and safety of such denoising methods have been raised. In this work we want to lay the fundamentals for a post-hoc interpretation of existing CT denoising networks by reconstructing their invariances.**

***Index Terms*—low dose, denoising, explainable, invariances.**

I. Introduction

In recent years, deep learning methods have been employed for many problems in medical image formation, including image-based and projection-based noise reduction, image reconstruction, scatter estimation, and artifact reduction. While the results of deep neural-network (DNN) based methods often excel those of conventional algorithms both qualitatively and quantitatively, they lack interpretability due to most DNNs being black boxes. Particularly for low dose CT imaging, recent advancements in generative methods such as generative adversarial networks (GANs) [1] and variational autoencoders (VAEs) [2] demonstrated impressive performance, providing competitive image quality compared to commercial iterative reconstruction techniques [3].

In this work, instead of focusing on the actual denoising performance of DNN-based methods for CT imaging, we want to lay the fundamentals for a post-hoc analysis of such networks in terms of their interpretability and robustness. To this end, we investigate what they have learned to represent and to ignore (i.e. their invariances) at different layers and argue that robust and non-robust denoising networks are invariant to different input features. Note, that this type of analysis is not restricted to CT and similar methods can be applied to denoising networks for other imaging modalities (e.g. magnetic resonance imaging or positron emission tomography).

Elias Eulig is with the German Cancer Research Center (DKFZ), Heidelberg, Germany.

Prof. Dr. Marc Kachelrieß is with the German Cancer Research Center (DKFZ), Heidelberg, Germany and with the Heidelberg University, Heidelberg, Germany.

Prof. Dr. Björn Ommer is with the University of Munich and with the Interdisciplinary Center for Scientific Computing at Heidelberg University.

Corresponding author: Elias Eulig (elias.eulig@dkfz.de)

II. Background

A. CT Image Denoising with DNNs

In this work we assume to have high-dose images $y \in \mathbb{R}^{m \times n}$ as well as low dose images $x \in \mathbb{R}^{m \times n}$ during training time. The aim of any deep-learning based denoising method is then to find a function $f(\cdot\,;\theta)$ with parameters θ, such that

$$\arg\min_{\theta} \|f(x;\theta) - y\| \,, \tag{1}$$

where f is realised by a DNN. In recent years most improvements on finding an optimal f focused on alterations of the architecture and training scheme. While earlier work utilized pixelwise losses (in image or feature space) which lead to smooth predictions and lack high-frequency information [4, 6], many recent methods are being trained as GANs, leading to extremely realistic denoising results [3, 5].

B. Invariances of DNNs

Our work is based on reference [7], where the authors seek to reconstruct and interpret the invariances of image classification DNNs using invertible neural networks (INNs).

Given a network $f(x)$ we can analyze any internal latent representation z thereof by decomposing f into $f(x) = \Psi(z) = \Psi(\Phi(x))$. To then explain z we need to know what information of the input x is captured in z and to what information Φ is invariant to (and is thus missing in z). To this end, the authors of [7] employ a VAE comprised

7th International Conference on Image Formation in X-Ray Computed Tomography,
edited by Joseph Webster Stayman, Proc. of SPIE Vol. 12304, 123040S
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2647170

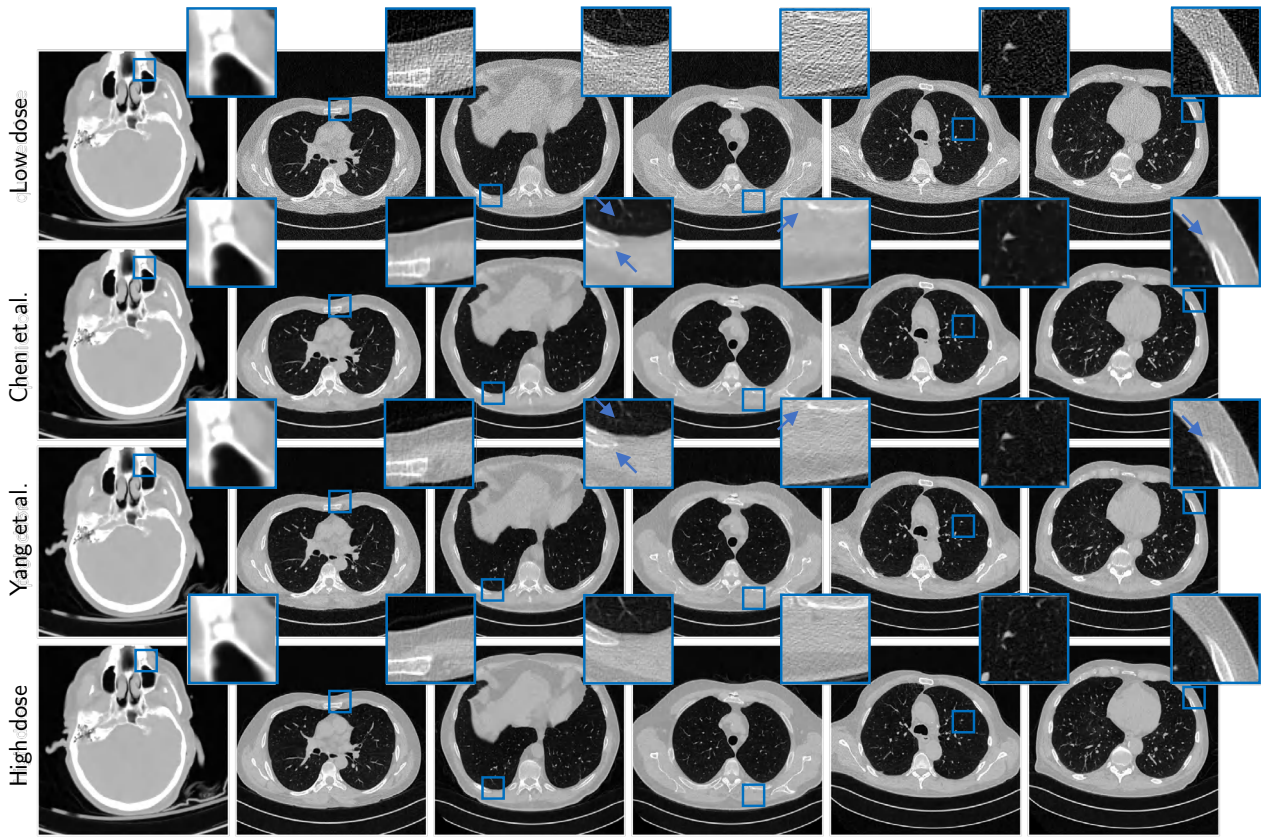

Fig. 1: Denoising performance of Chen et al. [4] and Yang et al. [5] for six different dataset samples (columns). Blue arrows indicate regions where the networks produced errors in the reconstruction of anatomical details.

of an encoder E and a decoder D that is trained to learn a complete data representation $\bar{z} = E(x)$ by reconstructing the input from $\bar{z}$ s.t. $\|D(E(x)) - x\|$ is minimized.

Since the complete data representation $\bar{z}$ now not only contains the information captured in z but also its invariances v, we need to disentangle v and z by learning a mapping

$$t(\cdot|z) : \bar{z} \rightarrow v = t(\bar{z}|z)\,. \tag{2}$$

Here, it is assumed that invariances v can be sampled from a Gaussian distribution, i.e. $p(v) = \mathcal{N}(v|0,1)$, and the mapping t is realized through a normalizing flow [8–10], a sequence of INNs between the simple (normal) distribution $p(v)$ and the complex distribution $p(\bar{z})$.

Since t is invertible, we can generate new $\bar{z}$ that only differ in their realization of the invariances by first sampling $v \sim p(v)$ and then applying the inverse mapping of t

$$t^{-1}(\cdot|z) : v \rightarrow \bar{z} = t^{-1}(v|z)\,. \tag{3}$$

To visualize $\bar{z}$ in the low dose image space we can reconstruct them using the previously trained decoder D: $\bar{x} = D(\bar{z})$.

III. Methods

A. Dataset

For all our studies the Low Dose CT Image and Projection dataset [11] is employed. The dataset comprises 50 head scans, 50 chest scans, and 50 abdomen scans acquired at routine dose levels with a SOMATOM Definition Flash (Siemens Healthineers, Forchheim, Germany) CT scanner. Additionally the dataset provides simulated low dose reconstructions (at 25 % dose for abdomen/head and at 10 % dose for chest scans) which were used as input to the denoising networks. We split the dataset into 70 %/20 %/10 % for training/validation/test across all patients and trained with a weighted sampling scheme such that slices from each patient were sampled with equal probability.

To make results between different methods comparable we trained and validated all denoising networks as well as the invariance reconstruction method on the same training/validation split of our data.

B. Denoising Methods

While our method can be used to provide post-hoc invariance analysis for any (trained) DNN-based denoising method, for simplicity, we here focus on interpreting the invariances of two well-known denoising methods:

Chen et al. [4] proposed a simple three-layer convolutional neural network which was trained to minimize (1) using an L_2 loss. The authors trained their network on patches of size 33×33 using an SGD optimizer and showed that their method can outperform conventional state-of the art methods.

Yang et al. [5] improved on previous works by training a Wasserstein GAN (WGAN) [12] in combination with a perceptual loss [13] in feature space. Furthermore, they utilize a deeper generator compared to [4] and train the network on larger patches of size 64×64.

We trained both [4] and [5] on the dataset described in Sec. III-A using the hyperparameters as described in the original papers. Whenever hyperparameters were not stated by the authors, we ran a grid-search and used the parameters that result in the lowest validation loss.

C. Recovering Invariances

Similar to reference [7] we first learn a complete data representation $\bar{z} = E(x)$ for a given low dose image x by training a VAE $g(x) = D(E(x))$. Our encoder is based on a ResNet-101 [14] and our decoder on a BigGAN [15] where the conditioning on the class is replaced by a conditioning on the latent representation $\bar{z}$. To improve reconstruction quality the VAE is trained together with a critic C as a WGAN and instead of training it on entire 512×512 pixels images we train it on 128×128 pixels patches.

For both of the two denoising networks evaluated, we train three conditional INNs (cINNs) to learn to reconstruct invariances at three different layers in the network.

TABLE I: Overview of generator architectures used in Chen et al. [4] and Yang et al. [5]. Kernel sizes of the 2D convolutions are indicated by k and their number of filters by f. Final nonlinearities of the original architectures were omitted to accommodate for the normalization of our data.

Layer	Chen et al. [4]	Yang et al. [5]
1	Conv k9 f64	Conv k3 f32
	ReLU	ReLU
3	Conv k3 f32	Conv k3 f32
	ReLU	ReLU
5	Conv k3 f1	Conv k3 f32
⋮		⋮
15		Conv k3 f1

For Chen et al. [4] we do so at layer 1, 3, and 5 and for Yang et al. [5] at layer 1, 7, and 13 (refer Tab. I). Each of the cINNs, t is composed of four invertible blocks, where each block is composed of coupling blocks [16], actnorm layers [17], and shuffling layers. For each invertible block, the conditioning on the denoising network representation z is realized by concatenating an embedding $h = H(z)$, where H is a shallow network, with the input to the respective block.

For each network and layer we then reconstruct different samples of the invariances $\bar{x} = D(t^{-1}(v, z)), v \sim \mathcal{N}(0, 1)$. Additionally, we can compute the standard deviation over a large set (here 250) of samples to highlight regions with high variation across the reconstructed invariances.

IV. Results

A. Denoising Methods

We find that the results from both denoising networks are similar to those reported in the respective original papers (Fig. 1). Due to the L_2 loss in image space the results from [4] appear smooth and lack structural fidelity. This is alleviated by training with an adversarial loss and consequently our results for [5] look much more realistic with higher details and noise structures very similar to those present in the high dose images.

However, we find that both methods are unable to correctly reconstruct anatomical details in several cases (refer Fig. 1, blue arrows). This is particularly problematic when the network is trained in an adversarial setting, where those false anatomies can look very convincing to the radiologist.

B. Reconstructed Invariances

The reconstructed invariances for both networks and two different samples (ref. Sec. III-C) are provided in Fig. 2. For each sample we also show the low dose input image x, the high dose ground truth image y, the reconstruction of the complete data representation $\bar{x} = D(\bar{z})$, and the denoised image $f(x)$.

From this we find that both denoising methods are invariant to several anatomical features to some extent (Fig. 2; blue arrows). We also find a higher overall variance of the invariances in homogeneous regions of the image for [4], indicating that it is more invariant to the specific realization of noise in the low dose input image. However, when inspecting the VAE reconstructions $\bar{x} = D(\bar{z})$ we also find major deviations from the original low dose image x (Fig. 2, red arrows), which may explain some of the differences between the reconstructed invariances and x.

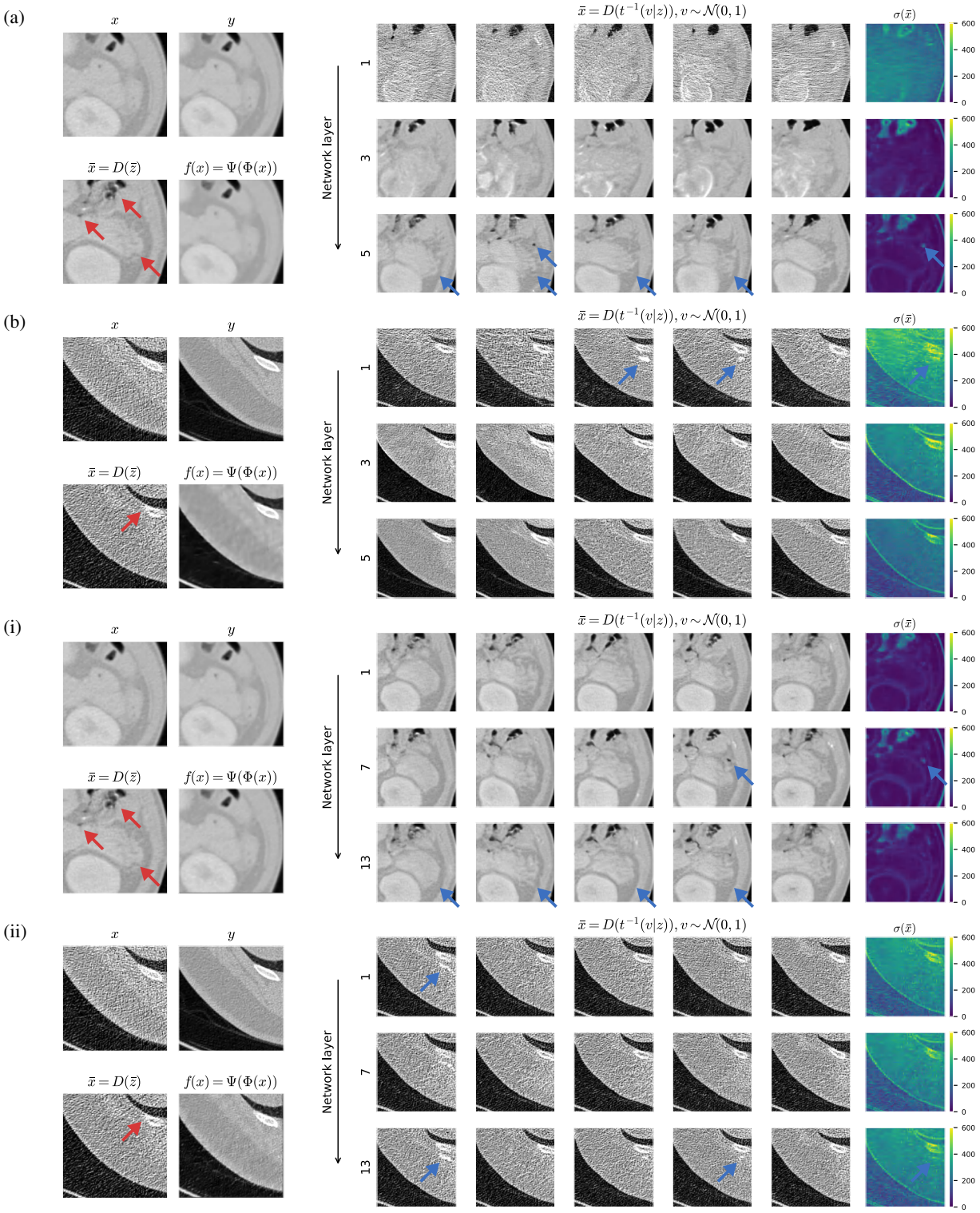

Fig. 2: Best viewed in color. Analysis of Chen et al. [4], (a) & (b), and Yang et al. [5], (i) & (ii). Provided are low dose input image x, high dose ground truth image y, VAE network reconstruction $\bar{x}$ (Sec. III-C), denoised image $f(x)$, five reconstructed samples from the space of invariances, and the standard deviation over 250 invariance samples. Red arrows highlight errors in the VAE reconstruction and blue arrows highlight regions in the reconstructed invariances.

V. Conclusion

In this work we analyzed deep neural networks for CT image denoising regarding their invariances to anatomical features in the low dose image domain. To reconstruct those invariances we adapted a method from prior work on interpretable AI and sampled reconstructions of invariances for two CT denoising networks. Upon analysis of the reconstructed invariances, we find that the representations of both networks at different layers are invariant to several anatomical features.

While this work demonstrated the potential of an invariance-based analysis of DNNs for CT image denoising, the ability to interpret those invariances is currently limited due to reconstruction errors from the embedding $\bar{z}$ and the complex, high-dimensional structure of the invariance images $\bar{x}$. Overcoming this drawback by improving the embedding $\bar{z}$ as well as mapping the sampled invariances to a semantically meaningful space remains part of future work.

Acknowledgment

This work was supported in part by the Helmholtz International Graduate School for Cancer Research, Heidelberg, Germany.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 2, pp. 2672–2680, 2014.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *ICLR*, 2014.

[3] H. Shan, A. Padole, F. Homayounieh, U. Kruger, R. D. Khera, C. Nitiwarangkul, M. K. Kalra, and G. Wang, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction," *Nature Machine Intelligence*, vol. 1, no. 6, pp. 269–276, 2019.

[4] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose CT denoising with convolutional neural network," *International Symposium on Biomedical Imaging (ISBI)*, pp. 143–146, 2017.

[5] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE TMI*, vol. 37, no. 6, pp. 1348–135, 2018.

[6] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE TMI*, vol. 36, no. 12, pp. 2524–2535, 2017.

[7] R. Rombach, P. Esser, and B. Ommer, "Making sense of CNNs: Interpreting deep representations & their invariances with INNs," *ECCV*, 2020.

[8] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *ICML*, pp. 1530–1538, 2015.

[9] L. Dinh, D. Krueger, and Y. Bengio, "NICE: non-linear independent components estimation," *ICLR*, 2015.

[10] J. S.-D. Dinh, Laurent and S. Bengio, "Density estimation using real NVP," *ICLR*, 2017.

[11] C. McCollough, B. Chen, D. Holmes, X. Duan, Z. Yu, L. Yu, S. Leng, and J. Fletcher, "Data from low dose ct image and projection data [data set]," *The Cancer Imaging Archive*, 2020.

[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *ICML*, vol. 15, pp. 214–223, 2017.

[13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *ECCV*, pp. 214–223, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.

[15] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *ICLR*, 2019.

[16] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Köthe, "Analyzing inverse problems with invertible neural networks," *ICLR*, 2019.

[17] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *NeurIPS*, vol. 31, 2018.