Some Investigations on Robustness of Deep Learning in Limited Angle Tomography

Yixing Huang[1](✉), Tobias Würfl[1], Katharina Breininger[1], Ling Liu[1], Günter Lauritsch[2], and Andreas Maier[1,3]

[1] Friedrich-Alexander Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
yixing.yh.huang@fau.de
[2] Siemens Healthcare GmbH, 91301 Forchheim, Germany
[3] Erlangen Graduate School in Advanced Optical Technologies (SAOT), 91058 Erlangen, Germany

Abstract. In computed tomography, image reconstruction from an insufficient angular range of projection data is called limited angle tomography. Due to missing data, reconstructed images suffer from artifacts, which cause boundary distortion, edge blurring, and intensity biases. Recently, deep learning methods have been applied very successfully to this problem in simulation studies. However, the robustness of neural networks for clinical applications is still a concern. It is reported that most neural networks are vulnerable to adversarial examples. In this paper, we aim to investigate whether some perturbations or noise will mislead a neural network to fail to detect an existing lesion. Our experiments demonstrate that the trained neural network, specifically the U-Net, is sensitive to Poisson noise. While the observed images appear artifact-free, anatomical structures may be located at wrong positions, e.g. the skin shifted by up to 1 cm. This kind of behavior can be reduced by retraining on data with simulated Poisson noise. However, we demonstrate that the retrained U-Net model is still susceptible to adversarial examples. We conclude the paper with suggestions towards robust deep-learning-based reconstruction.

Keywords: Deep learning · Limited Angle Tomography · Adversarial example.

1 Introduction

In practical applications of computed tomography (CT), the gantry rotation of a CT system, particularly an angiographic C-arm device, might be restricted by other system parts or external obstacles. In this case, only limited angle data are acquired. Image reconstruction from an insufficient angular range of data is called limited angle tomography. Due to missing data, artifacts will occur, including distorted boundaries, blurred edges, and biased intensities. These artifacts may lead to misinterpretation of the images. Therefore, artifact reduction in limited angle tomography has important clinical value.

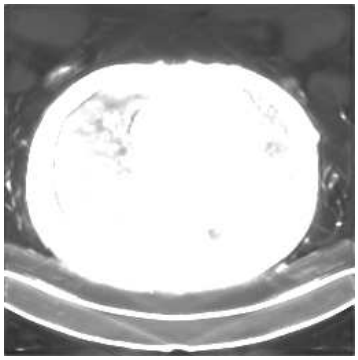


Fig. 1. The fake organ-like structures created by the U-Net, window: [-1000, -760] HU.

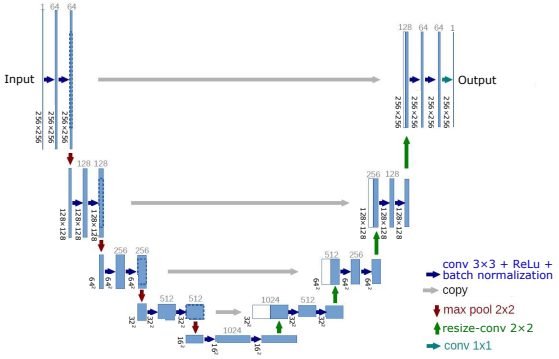


Fig. 2. The modified U-Net architecture for artifact reduction in limited angle tomography with an example of 256×256 input images (modified from [12]).

Many approaches have been investigated to reduce artifacts in limited angle tomography. One approach is to restore missing data using extrapolation/interpolation methods based on the band-limitation property [10] or data consistency conditions [6]. These methods can improve the image quality of simple data, but are not suited for clinical data consisting of complex structures. Another popular approach is iterative reconstruction with total variation [1, 13]. Iterative algorithms can reduce artifacts effectively, but are computationally expensive.

Recently, deep learning has achieved impressive success in various fields including limited angle tomography [15, 5, 4]. Würfl et al. [15] propose a neural network to learn the compensation weights for limited angle data based on [11]. Hammernik et al. further add a variational network to eliminate coherent streak artifacts [5]. Gu and Ye adapt the U-Net architecture [12] to learn artifacts from streaky images in the multi-scale wavelet domain [4]. Their work shows a promising prospect of the clinical application of deep learning into limited angle tomography in the near future.

However, the robustness of neural networks in practice is still a concern. It is reported that most neural networks are vulnerable to adversarial examples [16], which are typically generated by adding small perturbations [14, 7]. In some cases, the perturbations are too small to be noticed by human eyes. Nevertheless, they will cause a neural network to predict entirely wrong labels. For example, robust physical adversarial examples can be generated to attack an autonomous driving system such that it misclassifies a stop sign as a speed limited sign [2].

Like autonomous driving, clinical applications of deep learning also require a high level of safety and security. In our preliminary experiments on fan-beam limited angle tomography, we observed that the U-Net occasionally creates fake organ-like structures in the background without any attacker model (Fig. 1). This motivates us to look into the robustness of deep learning. In this paper, we aim to investigate whether a trained neural network for limited angle tomography is vulnerable to perturbations. Particularly, the influence of projection-domain Poisson noise, the most common noise existing in real CT data, is inves-

tigated. Taking this a step further, we look into trained adversarial examples. We conclude the paper by giving recommendations on how to benchmark deep learning-based approaches.

2 Materials and Methods

2.1 U-Net architecture

Based on [4, 12], we adapt the popular U-Net architecture for artifact reduction in limited angle tomography, as displayed in Fig. 2. The left part is a contraction path which follows the typical architecture of a convolutional network. Each blue arrow represents a 3×3 zero-padded convolution operation, a ReLU operation, and a batch normalization operation. The right side is an expansion path. The green arrow represents an up-sampling operation where we replace the original deconvolution operation by a resize-convolution to avoid checkerboard artifacts [9]. The copy (grey arrow) operation concatenates the up-sampled features with the corresponding features from the contraction path. The last 1×1 convolution operation maps the 64-channel features to a desired output image.

For limited angle tomography, the input images are reconstructed from limited angle data while the output images are the artifact-free images. The Hounsfield scaled images (input and target) are normalized to ensure stable training. An L_2 loss function is used.

2.2 Adversarial examples

Given a neural network classifier $\mathcal{C}$, an input image $\boldsymbol{f}$, and its true label l, an adversarial example can be described as the following,

$$\text{find } \boldsymbol{f}' \text{ s.t. } ||\boldsymbol{f}' - \boldsymbol{f}|| < \epsilon \text{ such that } \mathcal{C}(\boldsymbol{f}') = l' \text{ and } l' \neq l,$$

where $\boldsymbol{f}'$ is the adversarial example of $\boldsymbol{f}$, l' is the label of $\boldsymbol{f}'$ which is different from l, and ϵ is a parameter to control the difference between $\boldsymbol{f}$ and $\boldsymbol{f}'$. The perturbation is denoted by $\boldsymbol{e}$ where $\boldsymbol{e} = \boldsymbol{f}' - \boldsymbol{f}$. When the new label l' is specified, it is a targeted attack. Otherwise, it is a non-targeted attack.

To the best of our knowledge, adversarial examples have exclusively been reported for classification and segmentation tasks. We intend to investigate the robustness of the U-Net for limited angle tomography, which is a regression neural network. Since no discrete category labels are assigned to the outputs, the influence of a perturbation is evaluated by checking whether the U-Net is able to solve a specific task. In our case, we aim to reconstruct an existing lesion.

We pick a reference image, denoted by $\boldsymbol{f}_{\text{ref}}$. An image reconstructed from its limited angle projection data is denoted by $\boldsymbol{f}_{\text{limited}}$. The U-Net predicts an estimation of $\boldsymbol{f}_{\text{ref}}$ from $\boldsymbol{f}_{\text{limited}}$. The predicted image is denoted by $\boldsymbol{f}_{\text{est}}$. To check the robustness of the U-Net to perturbations, a simulated lesion is added to the reference image. The new reference image is denoted by $\boldsymbol{f}_{\text{ref,L}}$ where L is short for "lesion". Its limited angle reconstruction image and the predicted image by the U-Net are denoted by $\boldsymbol{f}_{\text{limited,L}}$ and $\boldsymbol{f}_{\text{est,L}}$, respectively.

Non-targeted attack: For non-targeted attacks, the fast gradient sign (FGS) method [7] is the most popular method to generate adversarial examples. However, the perturbations found by the FGS are like "salt-and-pepper" noise, which we do not expect to appear in real CT data. Instead, the most common noise in CT is Poisson noise. Therefore, it is worthwhile to investigate the influence of Poisson noise as the perturbation.

Targeted attack: For a targeted attack, we try to find a certain perturbation that misleads the U-Net to predict a target image where the lesion is missing. As the target, we use the estimated image without the lesion $\boldsymbol{f}_{\text{est}}$. The perturbation can be generated by the following optimization problem,

$$\arg\min_{\boldsymbol{e}} J(\boldsymbol{e}) = \arg\min_{\boldsymbol{e}} ||\boldsymbol{w}_1 \cdot \left(\mathcal{U}(\boldsymbol{f}_{\text{limited, L}} + \boldsymbol{e}) - \boldsymbol{f}_{\text{est}}\right)||_2^2 + \lambda||\boldsymbol{w}_2 \cdot \boldsymbol{e}||_2^2, \quad (1)$$

where $J(\boldsymbol{e})$ is the objective function to minimize, $\mathcal{U}$ is the trained U-Net model, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are weight vectors which have large weight elements at the lesion area, and λ is a relaxation parameter for the L_2 regularizer. The purpose of $\boldsymbol{w}_1$ is to penalize the error at the lesion area more than other areas. $\boldsymbol{w}_2$ further constrains the magnitude of $\boldsymbol{e}$ at the lesion area, otherwise the optimization may result in removing the lesion in the input image. The iterative least-likely class method in [7] is adapted to solve the above optimization problem:

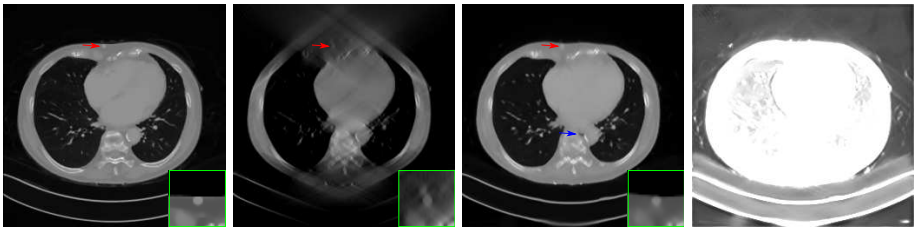
$$\boldsymbol{e}_0 = \boldsymbol{0}, \qquad \boldsymbol{e}_{i+1} = \boldsymbol{e}_i - \alpha \nabla_{\boldsymbol{e}} J(\boldsymbol{e}_i), \quad (2)$$

where $\boldsymbol{e}_i$ is an approximation of $\boldsymbol{e}$ at the i-th iteration, $\nabla_{\boldsymbol{e}} J(\boldsymbol{e}_i)$ is the gradient of $J(\boldsymbol{e}_i)$ w. r. t. the perturbation $\boldsymbol{e}$ and is obtained by back-propagation, and α is the step size for the update.

2.3 Experimental setup

Experiment (Exp.) 1: In the first experiment, we evaluate the U-Net on lesion detection in cone-beam limited angle tomography without any perturbation as commonly performed in deep learning CT papers. We pick 17 patients from the AAPM Low-Dose CT Grand Challenge data for training and one patient for testing. The limited angle projections are simulated in a 120° angular range scan with an angular step of 1°. The source-to-isocenter distance is 600 mm and the source-to-detector distance is 1200 mm. The detector size is 620×480 with an isotropic element size 1.0 mm. Images are reconstructed using FDK with the Ram-Lak kernel from the limited angle projections. The size of the reconstructed images is $256 \times 256 \times 256$ with a pixel size of 1.25 mm, 1.25 mm, and 1 mm in the X, Y, and Z direction, respectively. For each patient we pick 13 slices from its reconstructed volume. As a result, 221 slices are used as training set. The slices have a distance of 2 cm in depth between neighbouring slices. Although different slices have different cone angles, the artifacts are mainly caused by the limitation in the scan angle. Therefore, we train on slices instead of volumes.

The U-Net is trained on the above noise-free data using the Adam optimizer. The learning rate is 10^{-3} for the first 100 epochs, 10^{-4} for the $101-130_{\text{th}}$ epochs,



(a) $\boldsymbol{f}_{\mathrm{ref,L}}$ (b) $\boldsymbol{f}_{\mathrm{limited,L}}$ (c) $\boldsymbol{f}_{\mathrm{est,L}}$ (d) $\boldsymbol{f}_{\mathrm{est,L}}$

Fig. 3. The result of lesion detection from 120° cone-beam limited angle reconstruction, perturbation/noise-free, window: [-1000, 1000] HU. The lesion position is marked by the red arrow and the mispredicted cavity is marked by the blue arrow. A region-of-interest (ROI) at the lesion area is shown at the right bottom corner with a window width of 1000 HU. $\boldsymbol{f}_{\mathrm{est,L}}$ in (c) is re-displayed at a narrow window [-1000, -760] HU in (d).

and 10^{-5} for the $131 - 150_{\mathrm{th}}$ epochs. The L_2-norm is applied to regularize the network weights. The regularization parameter is 10^{-4}.

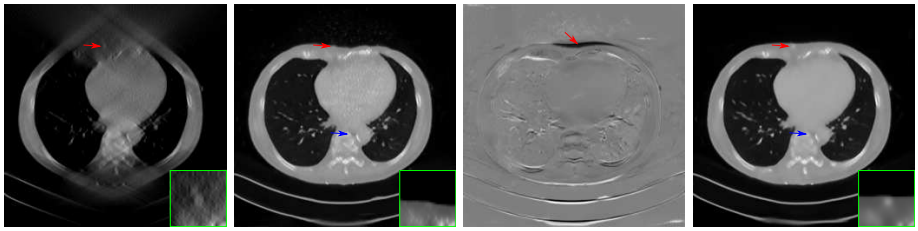
A simulated lesion cylinder is added to the ground truth testing volume. The lesion has a radius of 3 mm and a contrast of 200 HU. The volume with the lesion is forward projected and reconstructed from its limited angle data. For our lesion attack, we investigate a slice which is 13 cm away from the center plane for evaluation.

Exp. 2: For the non-targeted attack, Poisson noise is simulated considering an initial exposure of 10^5 photons at each detector pixel before attenuation. A linear attenuation coefficient of 0.02/mm is chosen as 0 HU. Poisson noise is added to the testing volume. The U-Net trained either without or with Poisson noise is evaluated on the selected noisy testing slice.

Exp. 3: For the targeted attack, the weight vectors are set $\boldsymbol{w}_1 = \boldsymbol{w}_2$ in Eqn. (1) with a value of 100 at a 15×15 patch (cf. Fig. 5) covering the lesion and a value of 1 for other areas. The L_2 regularizer parameter λ is set to 10. The step size α in Eqn. (2) is set to 10^{-3}. 32 iterations are used for the perturbation.

3 Results and Discussion

Exp.1: The results of the lesion detection in the perturbation free case are displayed in Fig. 3. Fig. 3(b) shows that the limited angle reconstruction $\boldsymbol{f}_{\mathrm{limited,L}}$ suffers from severe artifacts. The body outline is highly distorted at the top and bottom parts. The heart is obscured by streak artifacts. Many vessels in the lung are missing. The lesion (marked by the red arrow) is located at a position where many artifacts appear. $\boldsymbol{f}_{\mathrm{est,L}}$, the estimation of $\boldsymbol{f}_{\mathrm{ref,L}}$ predicted by the U-Net trained from the noise-free data, is shown in Fig. 3(c). The body outline is well restored. Most streaks at the heart are reduced. In addition, most vessel structures in the lung are also recovered. Importantly, the lesion can be clearly seen. These observations indicate a promising prospect of the clinical application of deep learning in the near future.



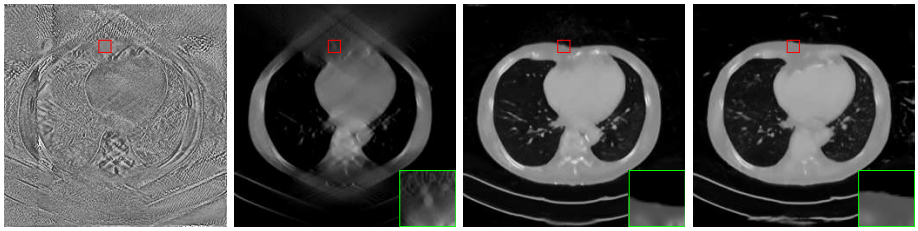
(a) $\boldsymbol{f}_{\text{limited,L,Poi}}$ (b) $\boldsymbol{f}_{\text{est,NL,ori}}$ (c) error of $\boldsymbol{f}_{\text{est,NL,ori}}$ (d) $\boldsymbol{f}_{\text{est,L,retrain}}$

Fig. 4. The influence of Possion noise in lesion detection: (a) is the reconstruction from the 120° limited angle sinogram with Poisson noise; (b) is the prediction of (a) by the original U-Net model, where the lesion cannot be detected; (c) is the difference image between (b) and the reference image $\boldsymbol{f}_{\text{ref,L}}$ with a window width of 2000 HU; (d) is the prediction of (a) by a retrained U-Net model from the data with Poisson noise, where the lesion is detected again. The lesion position is marked by the red arrow and the "cavity" area is marked by the blue arrow. The window for (a), (b), and (d) is [-1000, 1000] HU and the ROIs have a window width of 1000 HU.

In contrast to the preliminary fan-beam experiment from Fig. 1, fake organ-like structures are not observed, as shown in Fig. 3(d). However, still not all structures predicted by the U-Net are reliable. For example, the U-Net mispredicts a cavity structure in $\boldsymbol{f}_{\text{est,L}}$, marked by the blue arrow in Fig. 3(c), since this area has a low intensity in $\boldsymbol{f}_{\text{limited,L}}$.

Exp. 2: The influence of projection-domain Poisson noise is shown in Fig. 4. Fig. 4(a) is a reconstruction from the 120° limited angle sinogram with Poisson noise, denoted by $\boldsymbol{f}_{\text{limited,L,Poi}}$. Fig. 4(b) is an estimation of $\boldsymbol{f}_{\text{ref,L}}$ from $\boldsymbol{f}_{\text{limited,L,Poi}}$ using the original trained U-Net model, denoted by $\boldsymbol{f}_{\text{est,NL,ori}}$. Because of the Poisson noise, the lesion is hardly seen at $\boldsymbol{f}_{\text{est,NL,ori}}$. Although the patient top surface looks realistic, it is severely incorrect, shifting by up to 1 cm. The surface shift area is clearly indicated by the arrow at the difference image between $\boldsymbol{f}_{\text{est,NL,ori}}$ and $\boldsymbol{f}_{\text{ref,L}}$ displayed in Fig. 4(c). These observations demonstrate that the U-Net is sensitive to Poisson noise. In order to make the model robust to Poisson noise, we retrain the U-Net using the data with Poisson noise. The prediction of $\boldsymbol{f}_{\text{limited,L,Poi}}$ using the retrained model, denoted by $\boldsymbol{f}_{\text{est,L,retrain}}$, is shown in Fig. 4(d). The lesion is detected again, although it is smoothed. Interestingly, the "cavity" area marked by the blue arrow in Fig. 4 is predicted well by both U-Nets trained with and without Poisson noise.

Exp. 3: The results of the targeted attack are displayed in Fig. 5. Fig. 5(a) is the found perturbation which has small magnitude at the marked patch due to weight $\boldsymbol{w}_2$. The limited angle reconstruction with the perturbation is shown in Fig. 5(b). We can still notice the existence of the lesion, and to some extend the perturbation outside the patch. However, the lesion disappears at the predicted image by the original U-Net model (denoted by $\boldsymbol{f}_{\text{est,ori,pert}}$ in Fig. 5(c)). The U-Net retrained with Poisson noise also fails to reconstruct the lesion at the predicted image (denoted by $\boldsymbol{f}_{\text{est,retrain,pert}}$ in Fig. 5(d)).



(a) perturbation $\boldsymbol{e}$ (b) $\boldsymbol{f}_{\text{limited,L}} + \boldsymbol{e}$ (c) $\boldsymbol{f}_{\text{est,ori,pert}}$ (d) $\boldsymbol{f}_{\text{est,retrain,pert}}$

Fig. 5. The results of the targeted attack: (a) is the found perturbation $\boldsymbol{e}$; (b) is the adversarial example — the 120° limited angle reconstruction with the perturbation; (c) is the prediction of the adversarial example by the original U-Net model; (d) is the prediction of the adversarial example by the retrained U-Net model with Poisson noise. The patch covering the lesion position is marked by the red box. (a) is displayed at a window width of 200 HU, the window for (b)-(d) is [-1000, 1000] HU, and the ROIs have a window width of 1000 HU.

The nonlinearity and the linear behavior of high-dimensional spaces are the potential causes of adversarial examples [3]. They allow some tiny perturbations or noise to change the outputs of the U-Net drastically. The U-Net has a large perceptive field due to the contraction and expansion path. Therefore, although the perturbation has very small magnitude elements inside the lesion patch, its elements outside the patch still have an influence on the predicted values inside the patch through convolutional layers and make the lesion vanish.

4 Conclusion

In this paper, we investigate the application of the U-Net to limited angle tomography. The U-Net is able to reduce most artifacts. In the predicted image, distorted body outlines are restored, biased intensities are corrected, and missing vessels in the lung come back. The experiments on the robustness of the U-Net to perturbations indicate that training with projection-domain Poisson noise is mandatory for a limited angle reconstruction neural network. However, the retrained neural network is still vulnerable to non-local adversarial examples, despite its resistance to Poisson noise. We believe that the appearance of such adversarial examples in real clinical applications is unlikely, yet their non-localness has to be discussed.

Based on the presented experiments, we suggest that the following recommendations on how to benchmark deep learning CT (DLCT) algorithms should be followed. 1) **DLCT algorithms need to be exposed to accurate physical modelling and evaluated on real measured data.** Evaluation on synthetic data only delivers overly optimistic results. 2) Due to the dependency on training data, we believe that **many DLCT algorithms will be tailored towards specific applications** and not suited for generic image reconstruction. Claims of generality cannot be based on evaluation using a finite dataset. The inclusion of known operators can potentially remedy these problems [8]. 3)

DLCT reconstructions appear visually artifact-free. This prevents differentiation between the true signal and image completion solely based on prior knowledge. We demonstrate this quite drastically in our results that produce realistically looking patient surfaces that move by up to 1 cm, simply because the necessary data in the area was not measured. Still these reconstructions may be superior for a specific clinical task. As such **DLCT algorithms must be evaluated task-based**. 4) Additional exploration of **adversarial examples** might be useful to **explore limits of the trained algorithms**. As long as such effects are not sufficiently studied, deep learning-based reconstruction techniques are not yet ready for clinical applications.

Disclaimer: The concepts and information presented in this paper are based on research and are not commercially available.

References

1. Chen, Z., Jin, X., Li, L., Wang, G.: A limited-angle CT reconstruction method based on anisotropic TV minimization. Phys Med Biol **58**(7), 2119–2141 (2013)
2. Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., Song, D.: Robust physical-world attacks on deep learning models. arXiv preprint **1** (2017)
3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint (2014)
4. Gu, J., Ye, J.C.: Multi-scale wavelet domain residual learning for limited-angle CT reconstruction. Procs Fully3D pp. 443–447 (2017)
5. Hammernik, K., Würfl, T., Pock, T., Maier, A.: A deep learning architecture for limited-angle computed tomography reconstruction. Procs BVM pp. 92–97 (2017)
6. Huang, Y., Huang, X., Taubmann, O., Xia, Y., Haase, V., Hornegger, J., Lauritsch, G., Maier, A.: Restoration of missing data in limited angle tomography based on Helgason-Ludwig consistency conditions. Biomed Phys & Eng Express **3**(3), 035015 (2017)
7. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint (2016)
8. Maier, A., Schebesch, F., Syben, C., Würfl, T., Steidl, S., Choi, J.H., Fahrig, R.: Precision Learning: Towards Use of Known Operators in Neural Networks. Int Conf Pattern Recogn (2018), https://arxiv.org/abs/1712.00374, to appear
9. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill **1**(10), e3 (2016)
10. Qu, G.r., Lan, Y.s., Jiang, M.: An iterative algorithm for angle-limited three-dimensional image reconstruction. Acta Math Appl Sin **24**(1), 157–166 (2008)
11. Riess, C., Berger, M., Wu, H., Manhart, M., Fahrig, R., Maier, A.: TV or not TV? That is the Question. Procs Fully3D pp. 341–344 (2013)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Procs MICCAI pp. 234–241 (2015)
13. Sidky, E., X., P.: Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Med phys **53**(17), 4777–4807 (2008)
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint (2013)
15. Würfl, T., Hoffmann, M., Christlein, V., Breininger, K., Huang, Y., Unberath, M., Maier, A.K.: Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems. IEEE Trans Med Imaging **37**(6), 1454–1463 (2018)
16. Yuan, X., He, P., Zhu, Q., Bhat, R.R., Li, X.: Adversarial examples: Attacks and defenses for deep learning. arXiv preprint (2017)