May Denoising Remove Structures? How to Reconstruct Invariances of CT Denoising Algorithms

Elias Eulig[1,2], Joscha Maier[1],
Björn Ommer[3], and Marc Kachelrieß[1,4]

[1]German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Faculty of Physics and Astronomy, Heidelberg University, Germany
[3]LMU Munich, Germany
[4]Medical Faculty, Heidelberg University, Germany

DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

Motivation

Deep neural networks (DNNs) are powerful tools to reduce artifacts caused by

- Motion
- Scatter
- …
- Metal
- Noise

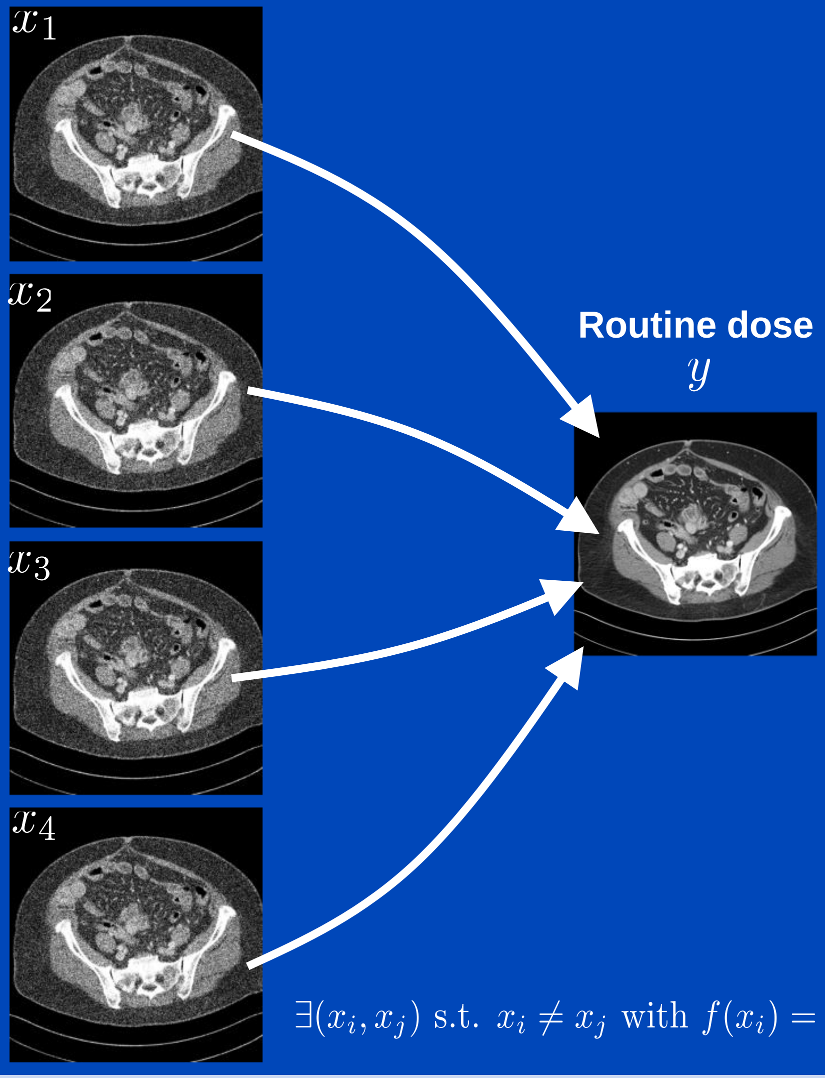
Networks have to be invariant in order to model non-injectivity of the data

- Invariances come from non-injective layers, e.g., max-pooling, ReLU

Aim: Reconstruct those invariances

Low dose

$x_i = y + n_i$



Routine dose

y

$\exists (x_i, x_j) \text{ s.t. } x_i \neq x_j \text{ with } f(x_i) = f(x_j)$

Introduction

Deep Learning-Based CT Image Denoising

- **Supervised in projection domain:**

$$\varphi^* = \arg\min_{\varphi} \mathbb{E}_{p',p\sim\mathcal{D}^{\mathrm{train}}} \|g_\varphi(p') - p\|$$

p' **: low dose projections**

p **: high dose projections**

- **Supervised in image domain:**

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{x',x\sim\mathcal{D}^{\mathrm{train}}} \|f_\theta(x') - x\|$$

x' **: low dose images**

x **: high dose images**

- **Supervised dual domain:**

$$\theta^*, \varphi^* = \arg\min_{\theta,\varphi} \mathbb{E}_{p',p\sim\mathcal{D}^{\mathrm{train}}} \|g_\varphi(p') - p\| + \|f_\theta(\mathrm{X}^{-1} g_\varphi(p')) - \mathrm{X}^{-1} p\|$$

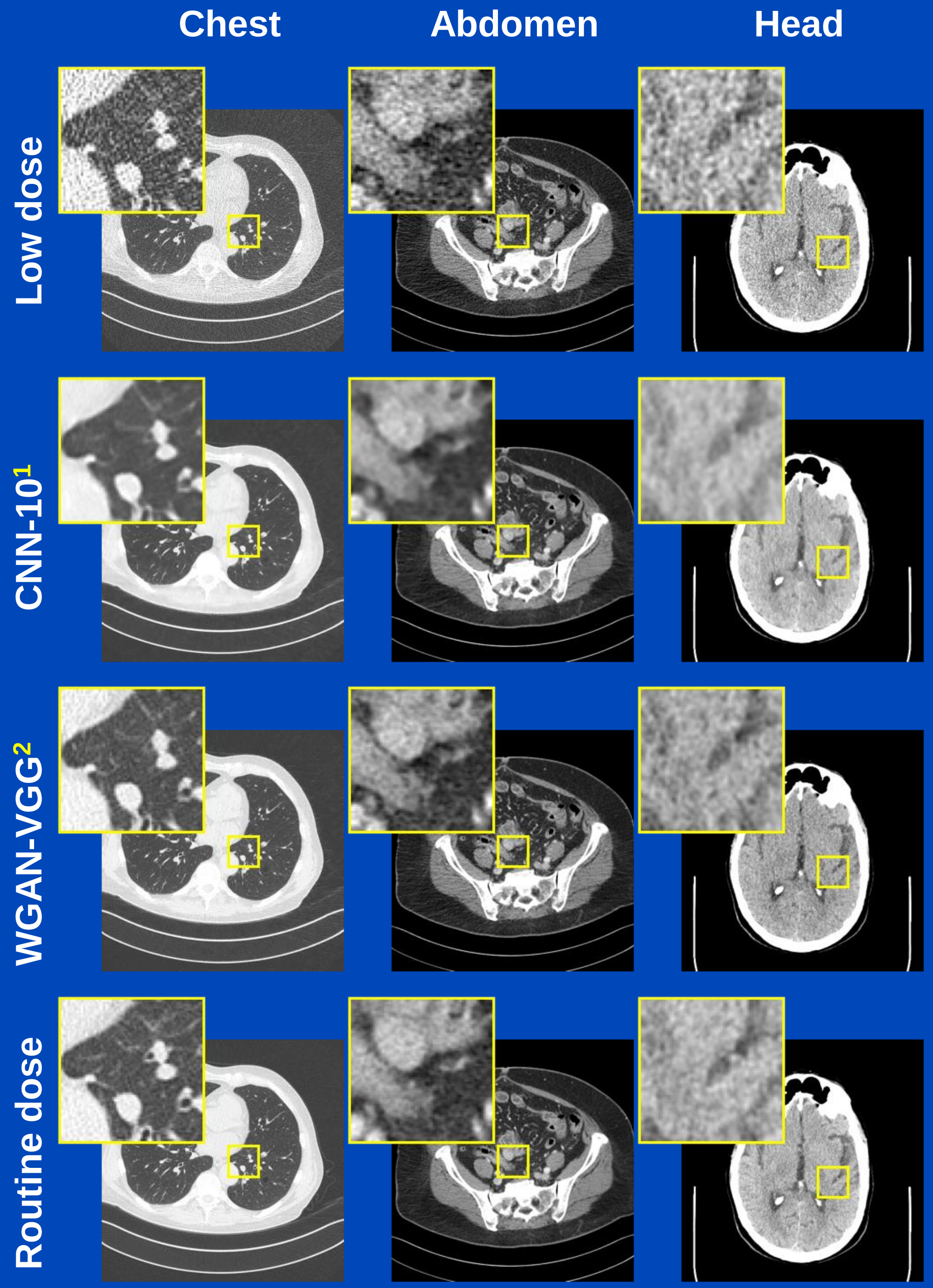
Introduction

Deep Learning-Based CT Image Denoising

Most work on improving f_θ focused on finding better

1. distance functions $\|\cdot\|$
2. architectures
3. training schemes

In particular, training f_θ as GAN with $\|\cdot\|$ being an adversarial loss leads to visually impressive results[2]

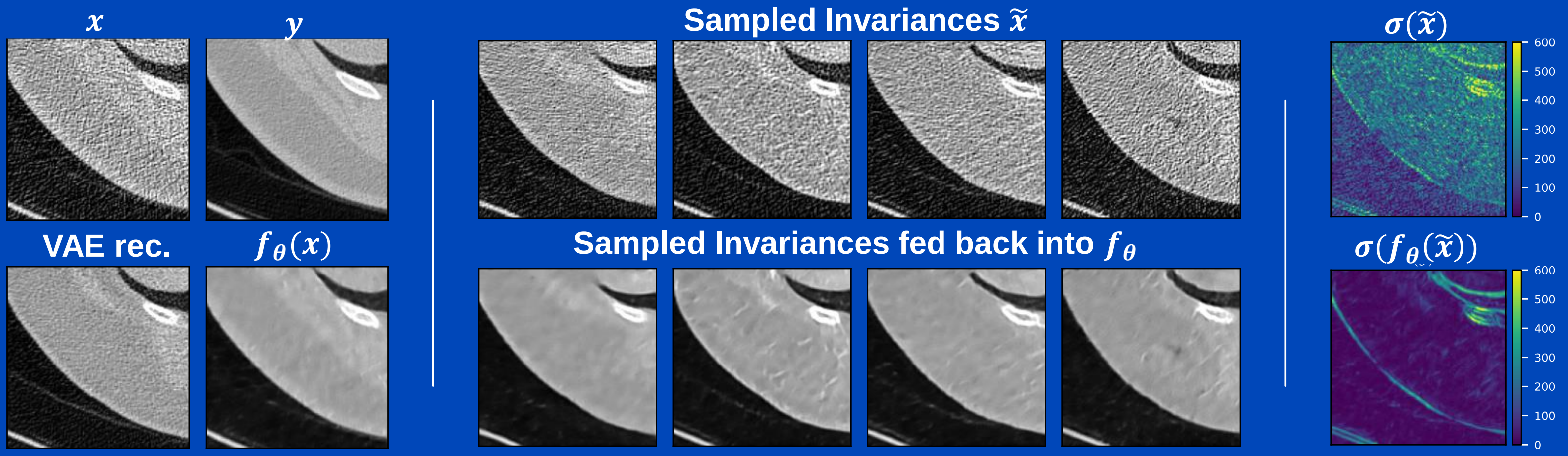


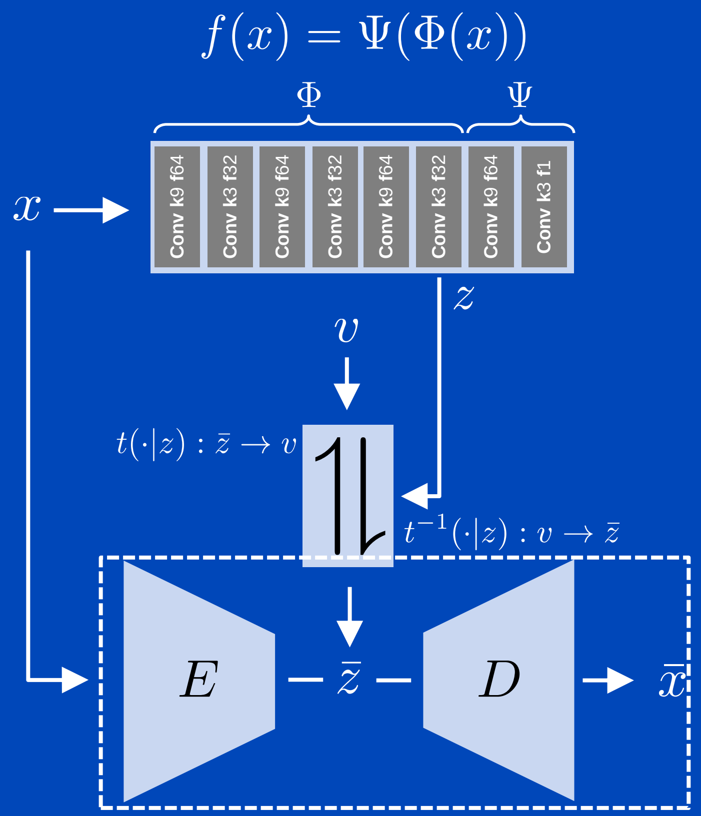
[1]Chen, Hu, Yi Zhang, Weihua Zhang, Peixi Liao, Ke Li, Jiliu Zhou, and Ge Wang. 2017. “Low-Dose CT via Convolutional Neural Network.” Biomedical Optics Express 8 (2): 679–94.
[2]Yang, Qingsong, […], Ge Wang. 2018. “Low-Dose CT Image Denoising Using a Generative Adversarial Network with Wasserstein Distance and Perceptual Loss.” IEEE TMI 37 (6): 1348–57.

Introduction

Invariances of DNNs

- Rombach et al.[1] reconstructed invariances of classifiers.
- Was later adapted to reconstruct invariances of DNNs for CT image denoising.[2]
- **Idea:** Learn complete data representation using VAE. Disentangle what the network learned and what it ignores using cINN.
- Problem: VAE may introduce invariances (bottom)





[1]Rombach, Robin, Patrick Esser, and Björn Ommer. "Making sense of CNNs: Interpreting deep representations and their invariances with INNs", ECCV, 2020.
[2]Eulig, Elias, Björn Ommer, and Marc Kachelrieß. "Reconstructing Invariances of CT Image Denoising Networks Using Invertible Neural Networks." CT Meeting, 2020.

Introduction
Adversarial Attacks

DNNs are vulnerable to adversarial examples

Example: Image classification:

Given: Classifier c_ϕ

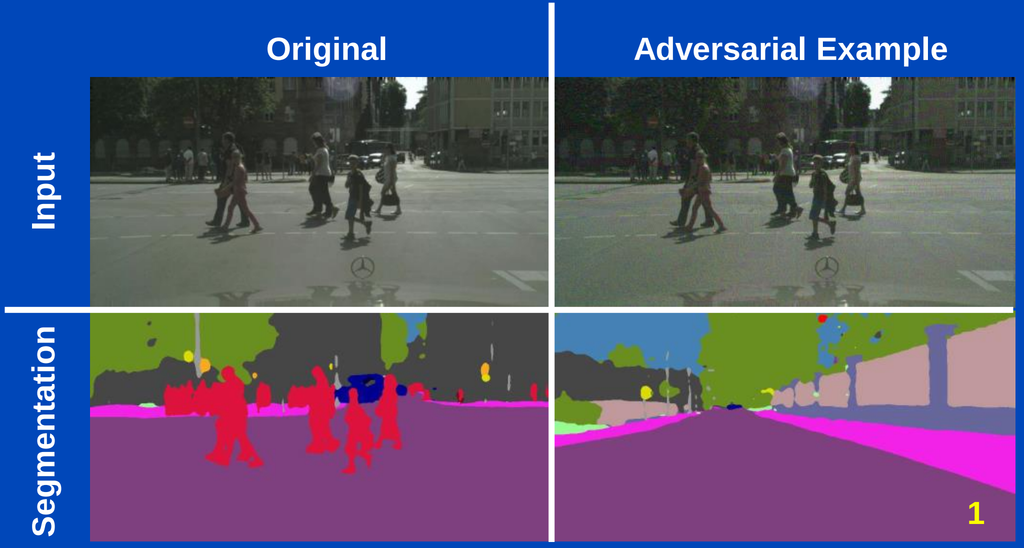
Input image x

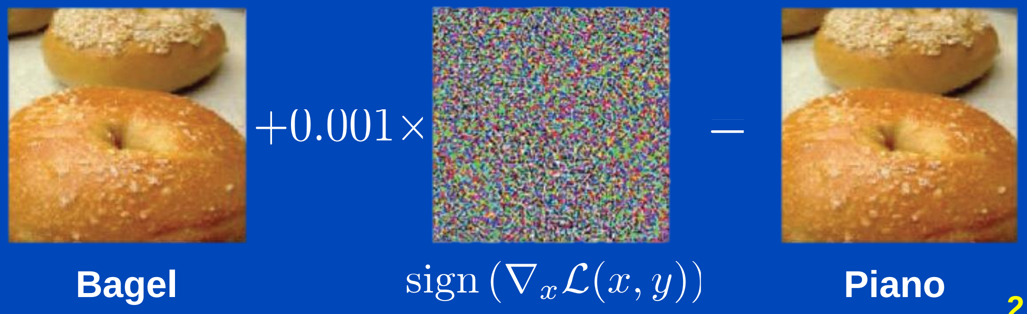
True class: $y \in \{1, ..., k\}$

(Target class: $\tilde{y} \in \{1, ..., k\}, \tilde{y} \neq y$**)**

Untargeted: $$\arg\min_{x^{\mathrm{adv}}} \left\{ \|x - x^{\mathrm{adv}}\| : c_\phi(x^{\mathrm{adv}}) \neq y \right\}$$

Targeted: $$\arg\min_{x^{\mathrm{adv}}} \left\{ \|x - x^{\mathrm{adv}}\| : c_\phi(x^{\mathrm{adv}}) = \tilde{y} \right\}$$





x **: Input image**

y **: Target label**

[1]Hendrik Metzen, J., Chaithanya Kumar, M., Brox, T., & Fischer, V. (2017). Universal adversarial perturbations against semantic image segmentation. In *ICCV*.

[2]Chen, Pin-Yu, […], Cho-Jui Hsieh. 2017. "ZOO: Zeroth Order Optimization Based Black-Box […]" In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 15–26. AISec '17

Introduction
Adversarial Attacks

Finding adversarial perturbations

Fast gradient sign method (FGSM)

If network was trained with loss function $\mathcal{L}$ (e.g., cross-entropy), we can backpropagate to x

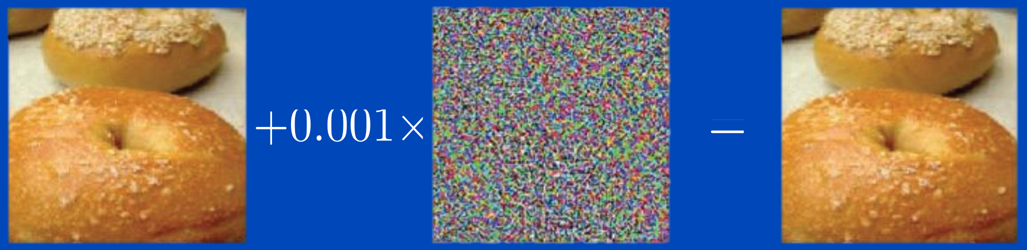
$$x^{\mathrm{adv}} = x + \epsilon \ \mathrm{sign}\left(\nabla_x \mathcal{L}(x, y)\right)$$

Basic iterative method (BIM)

Apply FGSM iteratively

$$x_0^{\mathrm{adv}} = x$$

$$x_{i+1}^{\mathrm{adv}} = \mathrm{Clip}_{x,\epsilon}\left\{x_i^{\mathrm{adv}} + \alpha \ \mathrm{sign}\left(\nabla_x \mathcal{L}(x_i^{\mathrm{adv}}, y)\right)\right\}$$



x : Input image

y : Target label

[1]Hendrik Metzen, J., Chaithanya Kumar, M., Brox, T., & Fischer, V. (2017). Universal adversarial perturbations against semantic image segmentation. In *ICCV*.
[2]Chen, Pin-Yu, […], Cho-Jui Hsieh. 2017. “ZOO: Zeroth Order Optimization Based Black-Box […]” In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 15–26. AISec ’17

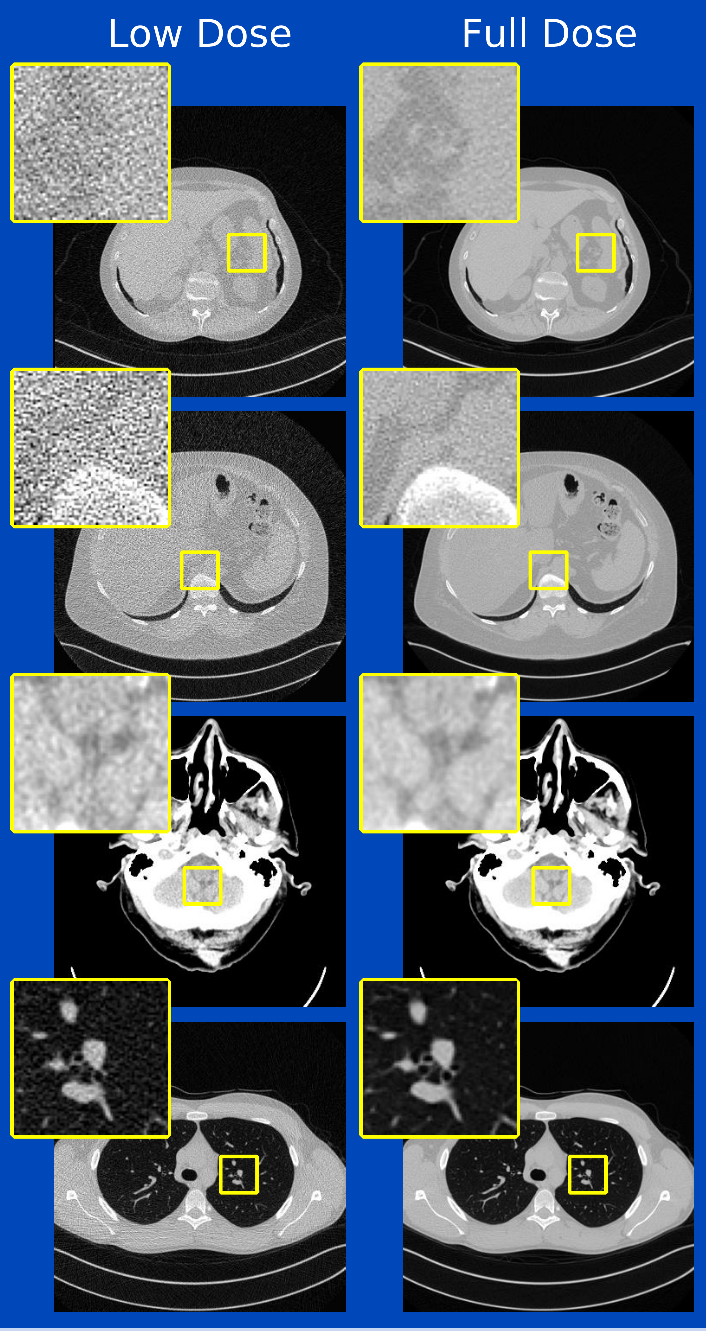
Methods

Dataset

Low dose CT image and projection dataset[1]

- 50 {head, chest, abdomen} scans
- Reconstructions of size 512×512 px
- Acquired with SOMATOM Definition Flash
- For each scan, simulated low dose acquisitions are available (25% dose for abdomen/head, 10% for chest)

Use weighted sampling scheme, such that slices from each patient were sampled with equal probability



[1]C. McCollough, et al., "Data from Low Dose CT Image and Projection Data [Data Set]," *The Cancer Imaging Archive*, 2020.

Methods

Denoising DNNs

CNN-10[1]:

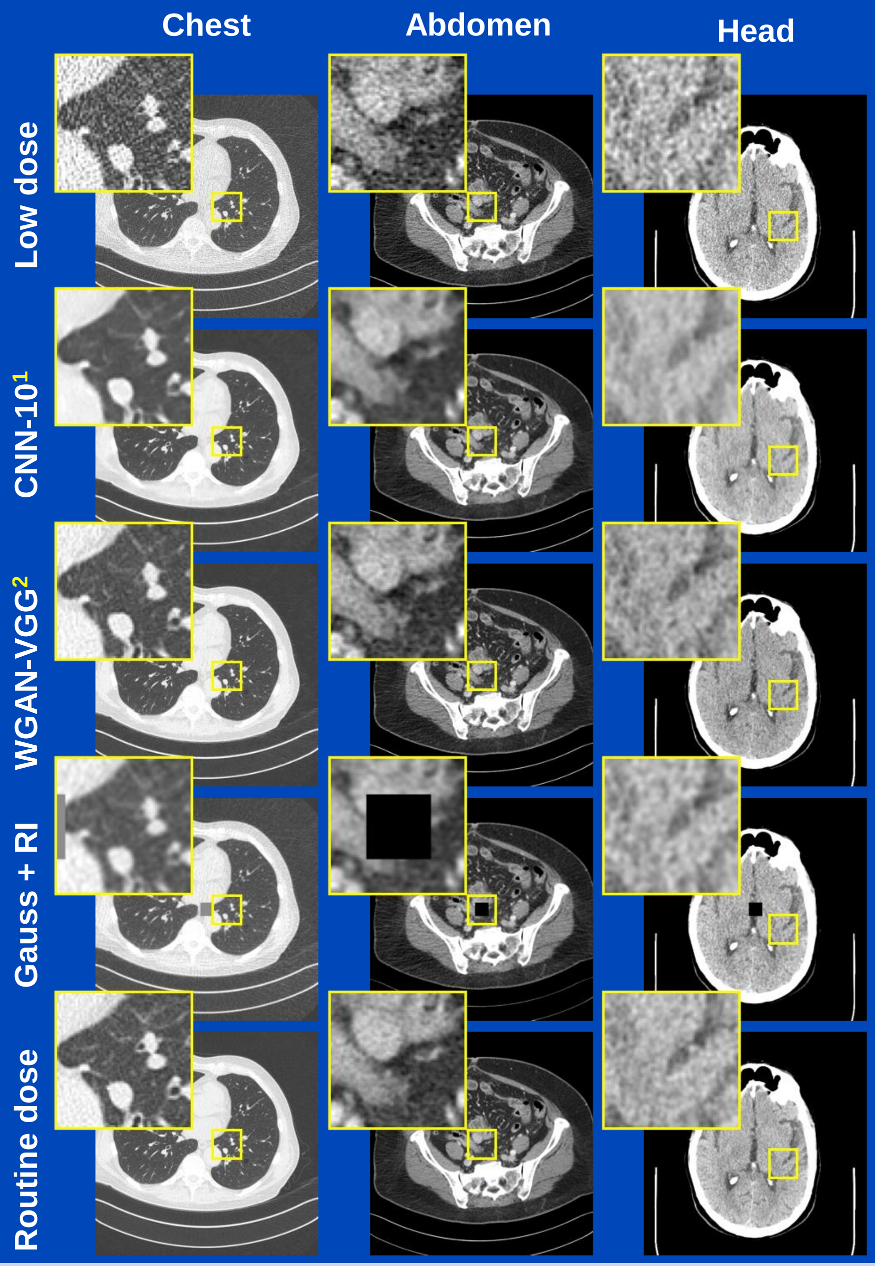
- Simple 3-layer CNN
- Trained to with $\mathcal{L}_2$ loss

WGAN-VGG[2]:

- 8-layer CNN as generator
- Trained as Wasserstein GAN (WGAN)
- Additional perceptual loss using ImageNet-pretrained VGG

Gaussian filter with rectangular invariances (Gauss + RI)

- Gaussian filter with unit standard deviation
- Returns zero for all pixels inside a center 30×30 px square



[1]H. Chen et al., "Low-dose CT denoising with convolutional neural network”, ISBI 2017, 2017.
[2]Q. Yang *et al*., "Low-Dose CT Image Denoising Using a Generative Adversarial Network […]”, in *IEEE TMI*, vol. 37, no. 6, 2018.

Methods

Reconstructing Invariances

Adversarial perturbations
Small perturbations in the input that lead to large alterations in predictions

Invariances
Large perturbations in the input that leave network predictions unaffected

Find invariances x^{inv} via

$$\arg\min_{x^{\text{inv}}} \left(\| f_\theta(x) - f_\theta(x^{\text{inv}}) \| - \alpha \| x - x^{\text{inv}} \| \right)$$

$$\alpha \in \mathbb{R}^+$$
$$x_0^{\text{inv}} = x + n$$
$$n \sim \mathcal{N}(0, 10^{-2})$$

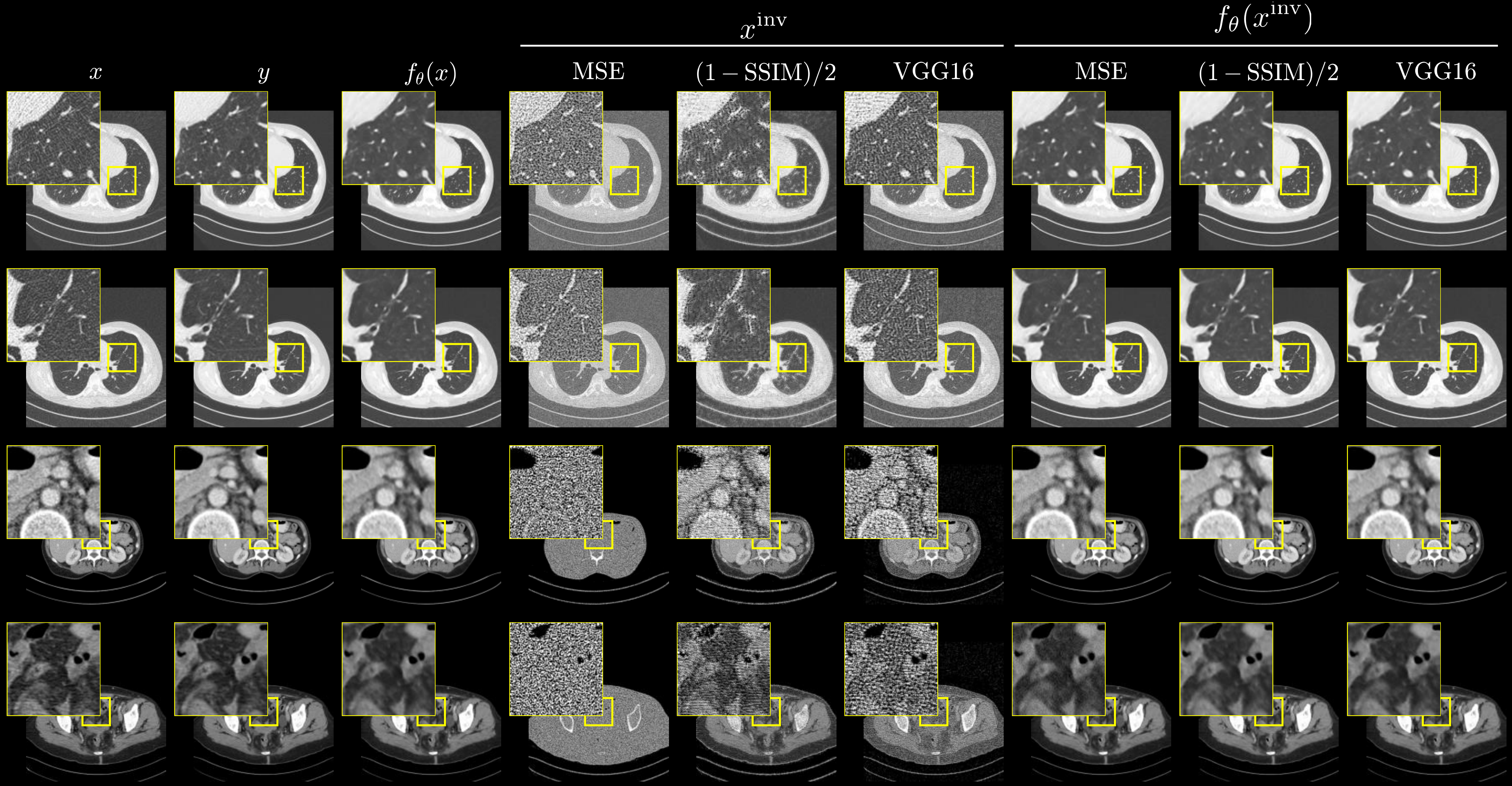
In our experiments we use for $\|\cdot\|$:

- **Mean-squared-error (MSE), bounded by +1**
- **Structural dissimilarity: $(1 - SSIM)/2$**
- **Perceptual loss using ImageNet-pretrained VGG16**

Optimize x^{inv} using Adam optimizer for 3k iterations

Results

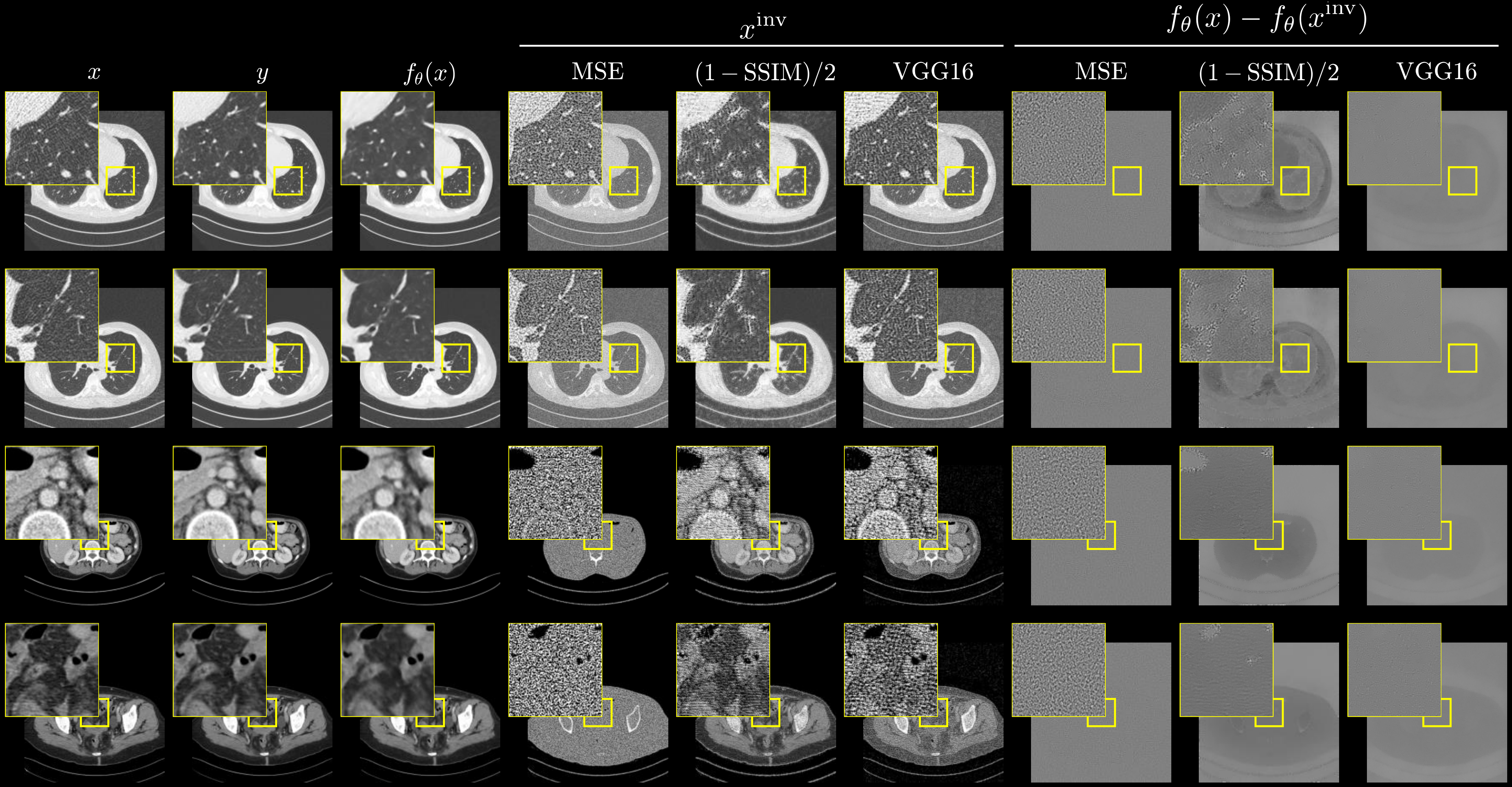
CNN-10



Lung: *C* = -600 HU, *W* = 1500 HU. Abdomen: *C* = 50 HU, *W* = 400 HU.

Results

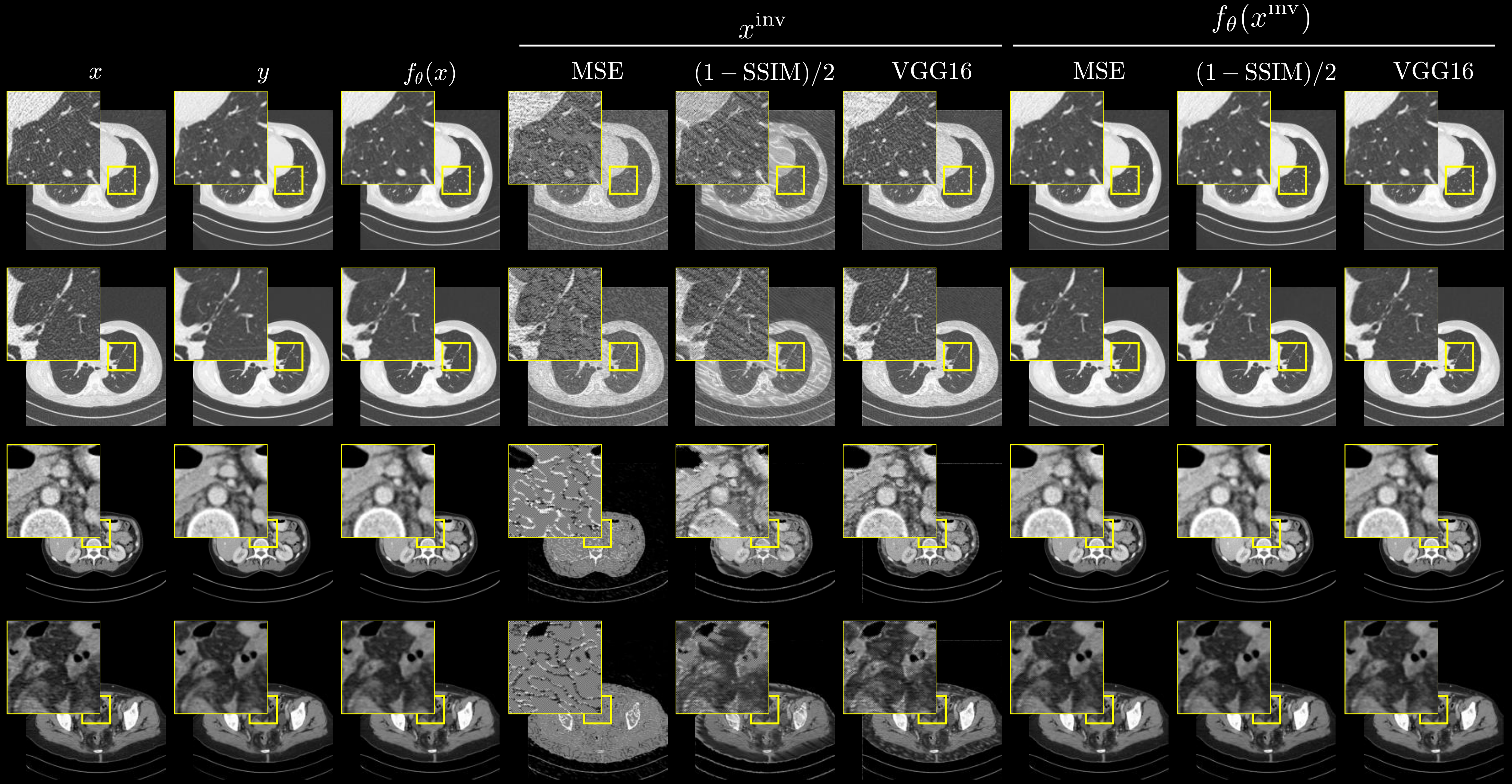
CNN-10



Lung: *C* = -600 HU, *W* = 1500 HU. Abdomen: *C* = 50 HU, *W* = 400 HU. Differences: *C* = 0 HU, *W* = 100 HU.

Results

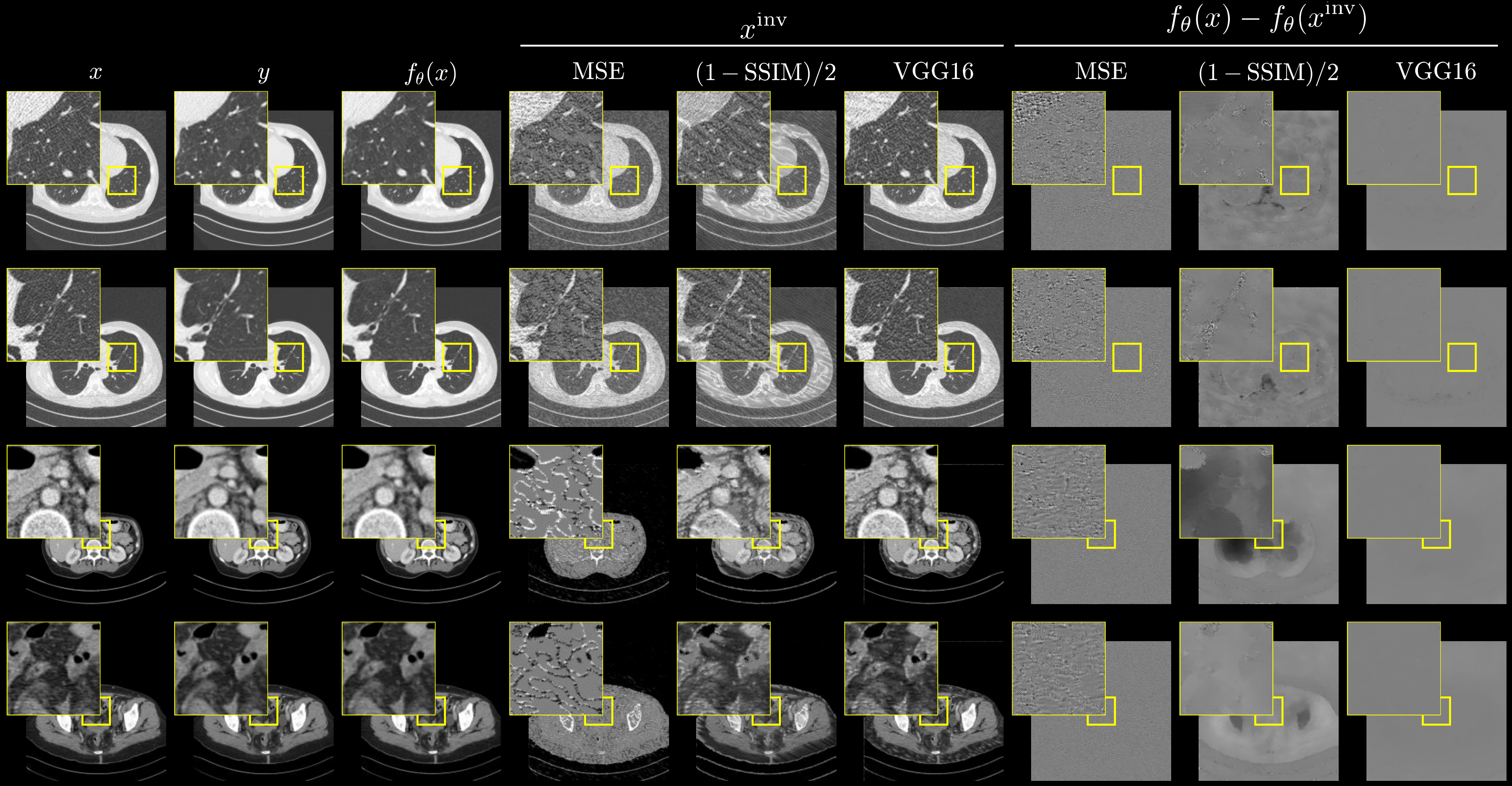
WGAN-VGG



Lung: *C* = -600 HU, *W* = 1500 HU. Abdomen: *C* = 50 HU, *W* = 400 HU.

Results

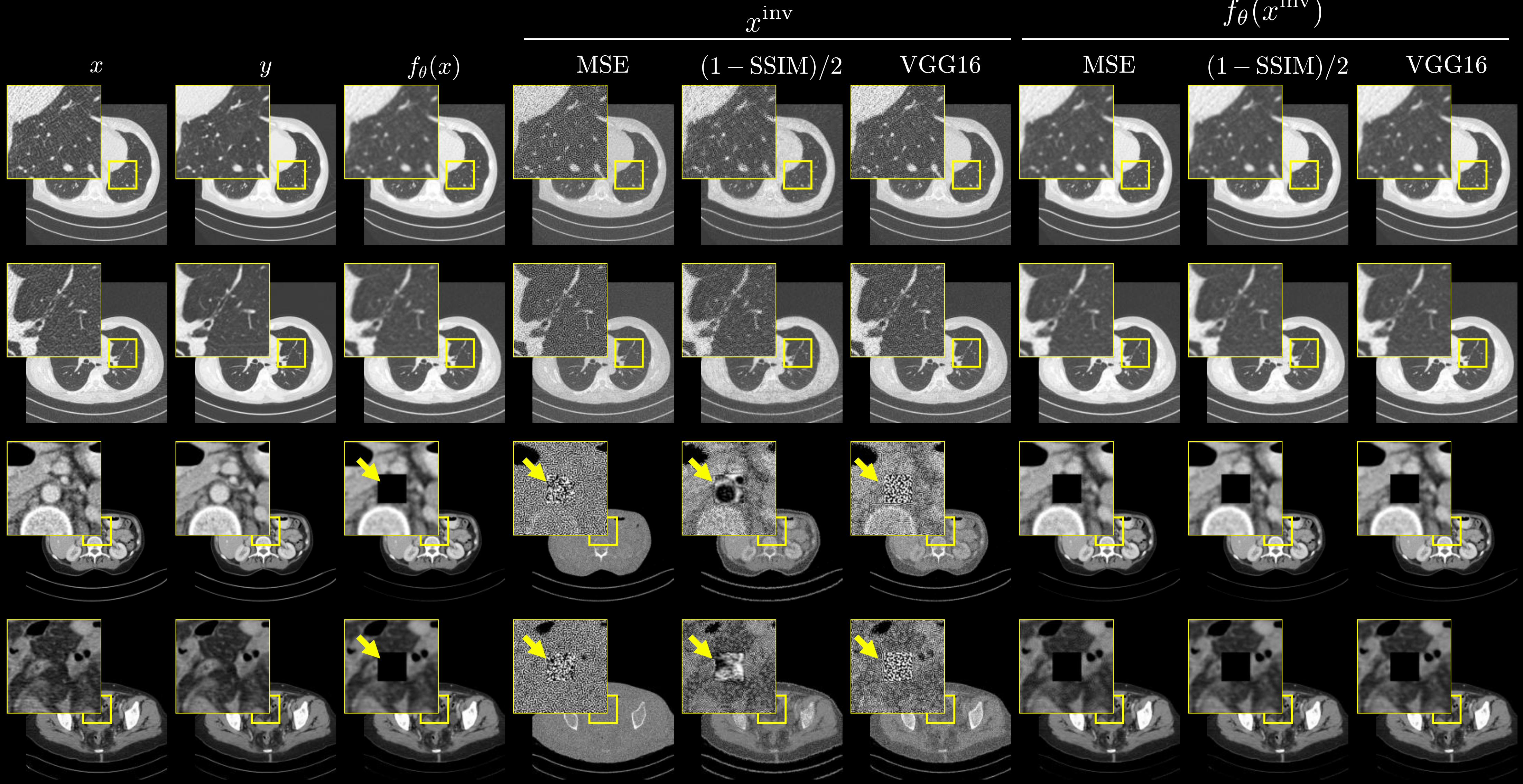
WGAN-VGG



Lung: *C* = -600 HU, *W* = 1500 HU. Abdomen: *C* = 50 HU, *W* = 400 HU. Differences: *C* = 0 HU, *W* = 100 HU.

Results

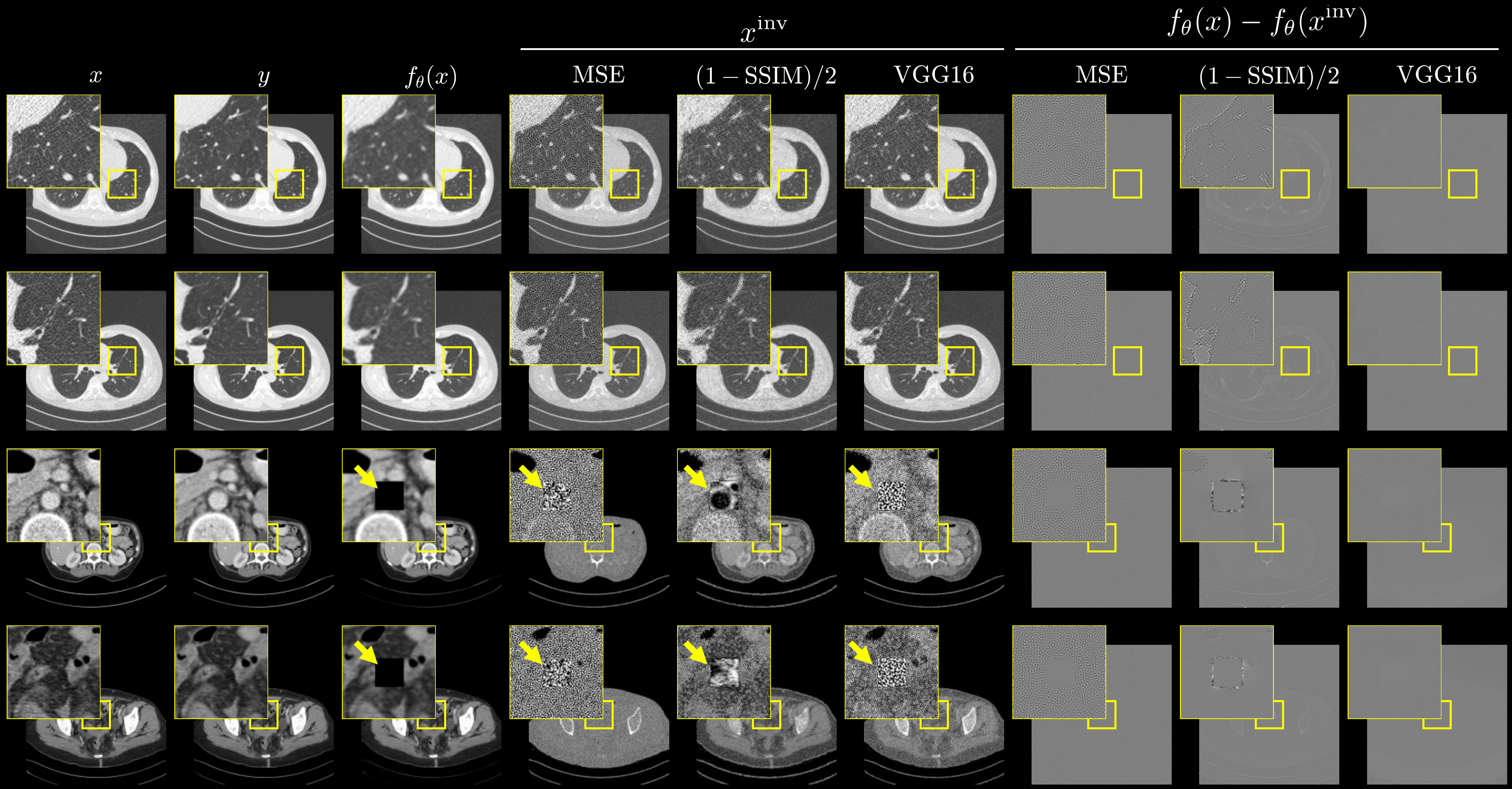
Gaussian + RI



Lung: *C* = -600 HU, *W* = 1500 HU. Abdomen: *C* = 50 HU, *W* = 400 HU.

Results

Gaussian + RI



Lung: *C* = -600 HU, *W* = 1500 HU. Abdomen: *C* = 50 HU, *W* = 400 HU. Differences: *C* = 0 HU, *W* = 100 HU.

Methods

Generate natural x^{inv}

Drawbacks of previous approach

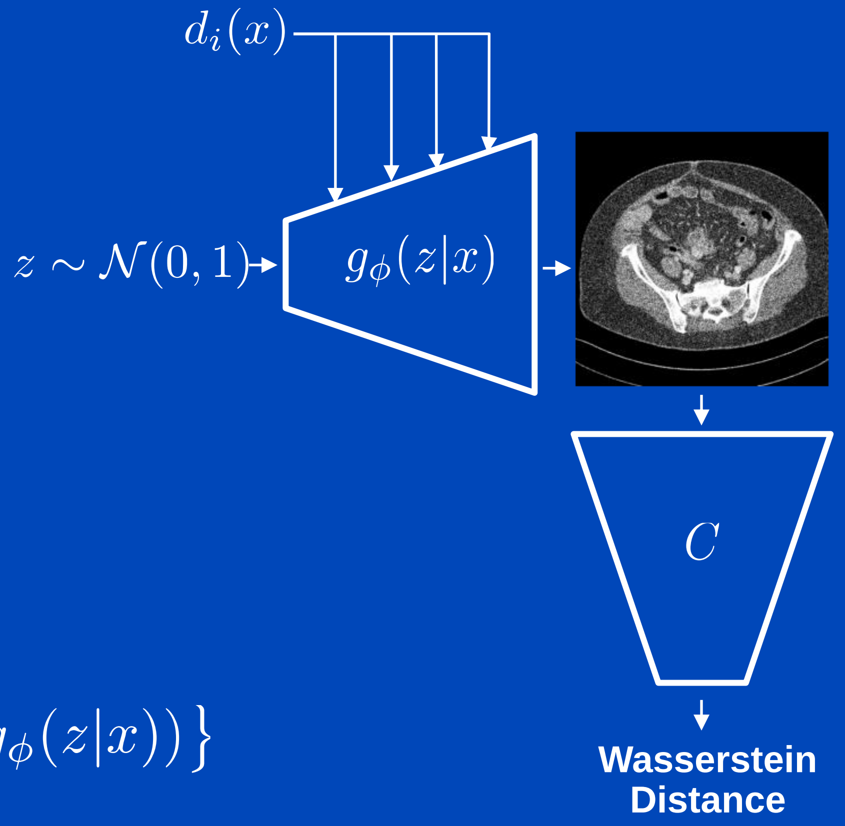
- Generated x^{inv} generally do not lie on the data manifold of low dose images $p_{x^{\text{inv}}} \neq p_x$
- Sampling new invariances requires new x_0^{inv}

Natural invariances

Generate natural (on-manifold) x^{inv} by training a conditional generator $g_\phi(z \sim \mathcal{N}(0,1)|x)$ together with a critic C

$$\arg\min_{\phi} \{ \|f_\theta(x) - f_\theta(g_\phi(z|x))\| - \alpha\|x - g_\phi(z|x)\| + \beta C(g_\phi(z|x)) \}$$

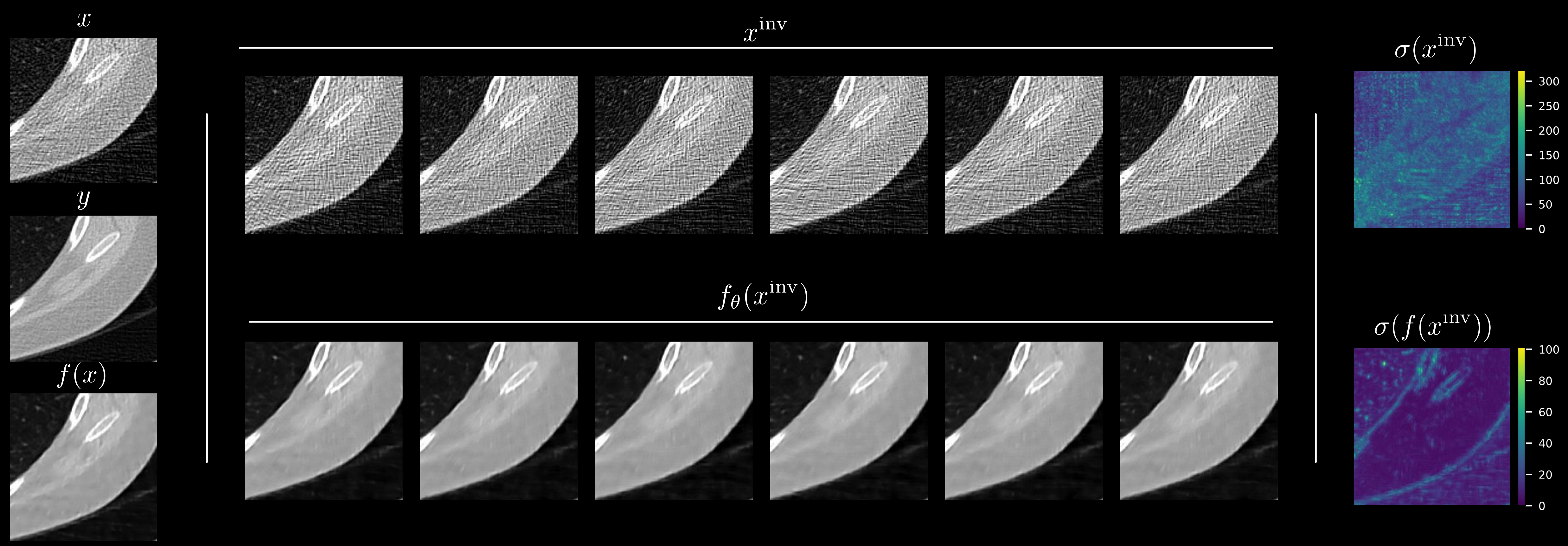
- Train on 80×80 px patches



$d_i(x)$: Downsampling at stage i

Results

CNN-10



Summary & Conclusions

Image-domain optimization

- The presented approach can generate perturbations to which common DNNs for CT image denoising are invariant.
- Magnitude of invariances is dependent on network structure and training scheme.

Adversarial training scheme

- To ensure natural (on-manifold) perturbations, we need to introduce an adversarial training scheme.
- Preliminary results indicate that natural perturbations mostly alter noise structure and much less the anatomical structure.

Thank You!

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

- This presentation will soon be available at www.dkfz.de/ct.
- Job opportunities through DKFZ’s international PhD or Postdoctoral Fellowship programs (marc.kachelriess@dkfz.de).