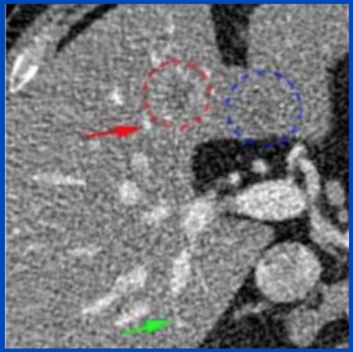
Reconstructing Invariances of CT Image Denoising Networks using Invertible Neural Networks

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

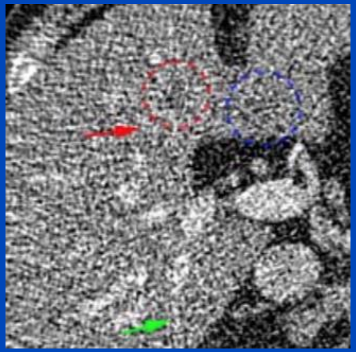
[1]German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]IWR, Heidelberg University, Germany
[3]Ludwig Maximilian University of Munich, Germany

Motivation

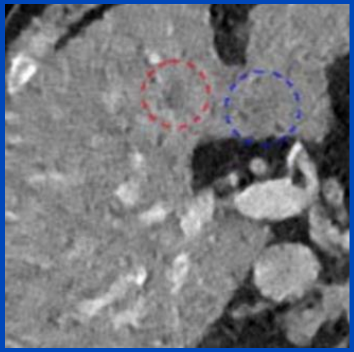
- Deep learning methods are employed for many problems in medical image formation, e.g.
  - Reconstruction
  - Scatter estimation
  - Image-based noise reduction
- Results of DNN-based methods often excel those of conventional algorithms qualitatively and quantitatively
- They lack interpretability due to black-box nature of DNNs → recent advancement in generative modelling signal false confidence



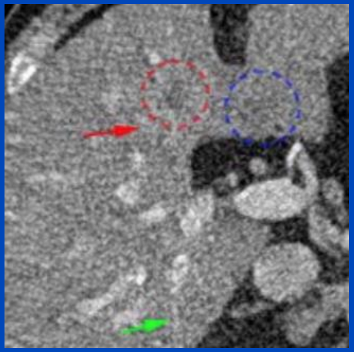
Full-dose reconstruction



Quarter-dose reconstruction



CNN trained with MSE



CNN trained as WGAN with VGG Loss

Examples for Low-dose CT denoising[1]

Here:

- Not focusing on denoising performance
- Lay fundamentals for post-hoc interpretability and robustness analysis of denoising DNNs
- Investigate what networks learned to represent and to ignore → Their invariances

[1]Q. Yang et al. (2018). Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein […]. IEEE TMI.

Methods

Deep-learning based CT Denoising

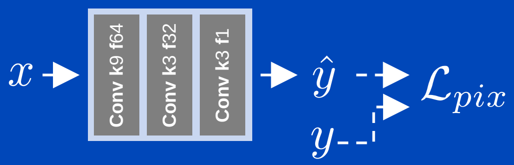
Deep learning-based CT denoising methods aim to find a function $f(\cdot\,;\theta)$ (realized by a CNN with parameters θ), s.t.

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

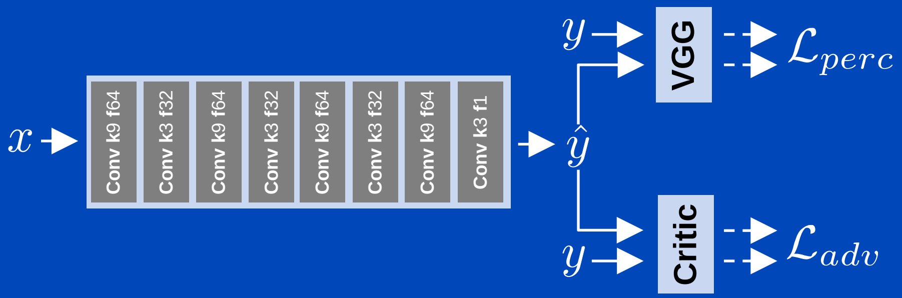
where x is the low dose input image and y is the high dose target image.

Recover invariances of two denoising methods:

- **Chen et al.[1]:**
  - **Simple 3-layer CNN**
  - **Trained to with $\mathcal{L}_2$ loss**
  - **Trained on patches of size 33 × 33 px²**



- **Yang et al.[2]:**
  - **8-layer CNN as generator**
  - **Trained as Wasserstein GAN (WGAN)**
  - **Additional perceptual loss**
  - **Trained on patches of size 64 × 64 px²**



[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.

dkfz.

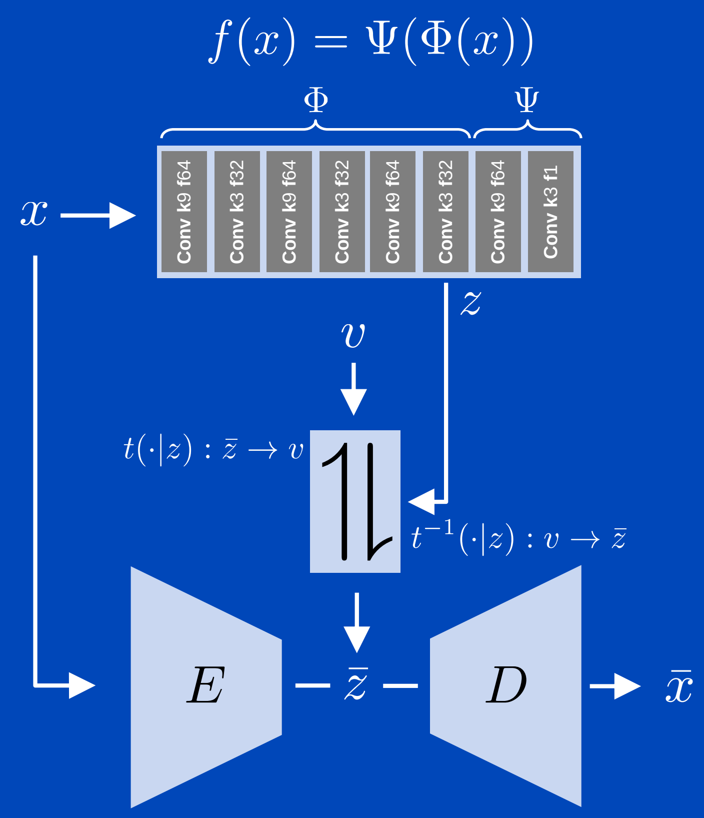
Methods

Recovering Invariances

- **Our work is based on Rombach et al.[1]**
- **Given a denoising network $f(\cdot\,;\theta)$ we can analyze internal latent representations z by decomposing $f(x) = \Psi(z) = \Psi(\Phi(x))$**
- **To reconstruct which information of x is captured in z we train a VAE to learn a complete data representation $\bar{z} = E(x)$**
- **To improve reconstruction quality, $G = D \circ E$ is trained together with critic C as a Wasserstein GAN**

$$\mathcal{L}(E, D) = \mathbb{E}_{\epsilon \sim \mathcal{N}(\epsilon,0,1)} \left[-C(\bar{x}) + \frac{1}{2} \sum_{i}^{N_{\bar{z}}} \mu_i^2 + \sigma_i^2 - \log(\sigma_i^2) \right]$$

- **Train G on 128×128 px[2] patches**
- **A similar VAE can be trained to learn a complete data representation of high-dose images y**



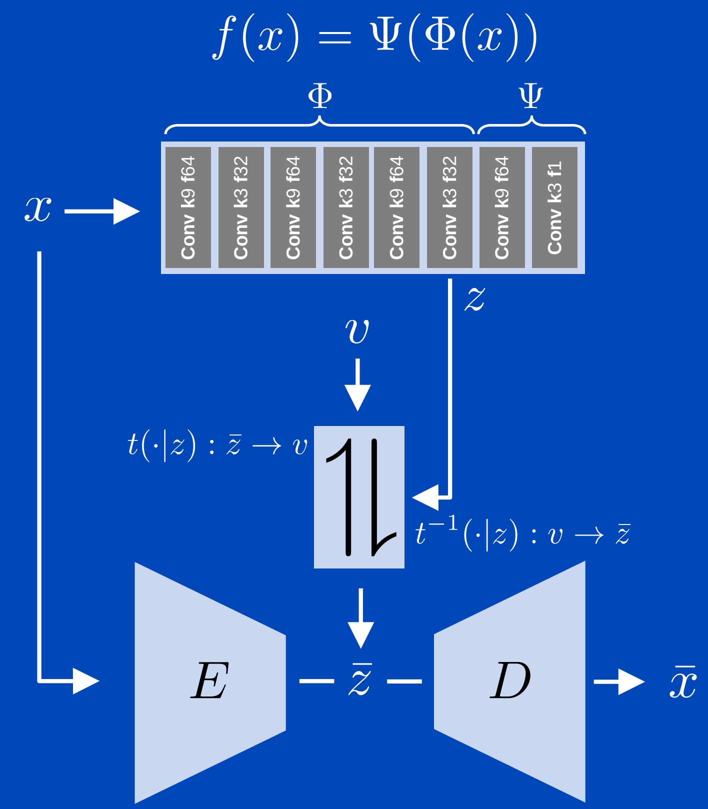
[1]Rombach, Robin, Patrick Esser, and Björn Ommer. "Making sense of CNNs: Interpreting deep representations and their invariances with INNs", ECCV 2020, 2020.

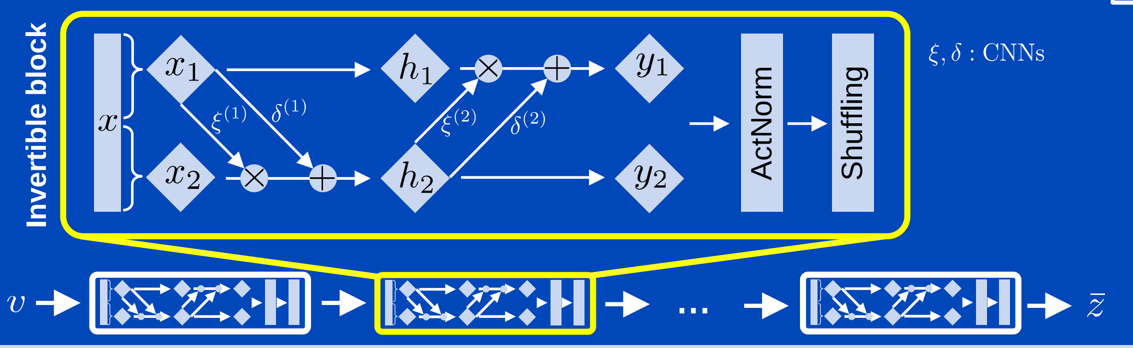
dkfz.

Methods

Recovering Invariances

- **Disentangle information captured in z and invariances v by learning a mapping**
  $$t(\cdot|z) : \bar{z} \rightarrow v = t(\bar{z}|z)\,, \quad p(v) = \mathcal{N}(v|0,1)$$
- $t(\cdot|z)$ **is realized by a conditional invertible neural network[1] (cINN)**
- **Generate new $\bar{z}$ by sampling $v \sim p(v)$ and then applying the inverse mapping**
  $$t^{-1}(\cdot|z) : v \rightarrow \bar{z} = t^{-1}(v|z)$$
- **Generate new images that only vary in their realization of invariances by applying the decoder** $\bar{x} = D(t^{-1}(v|z))$

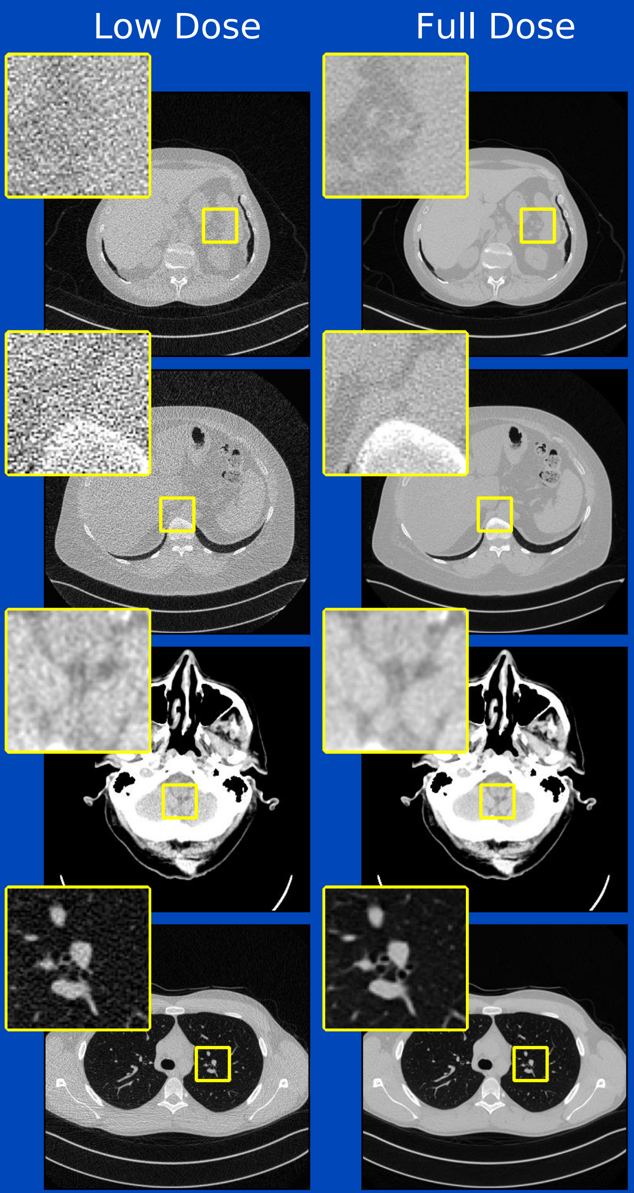




[1]Kingma, Durk P, and Prafulla Dhariwal. “Glow: Generative Flow with Invertible 1x1 Convolutions.” NeurIPS, Vol. 31,2018.

Methods
Dataset

- **Low Dose CT Image and Projection Dataset[1]**
  - **50 {head, chest, abdomen} scans**
  - **Reconstructions of size 512 × 512 px^2**
  - **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**
- **Train/validate/test each denoising method and our invariance reconstruction method on the same data splits**
  → Comparable results between different methods

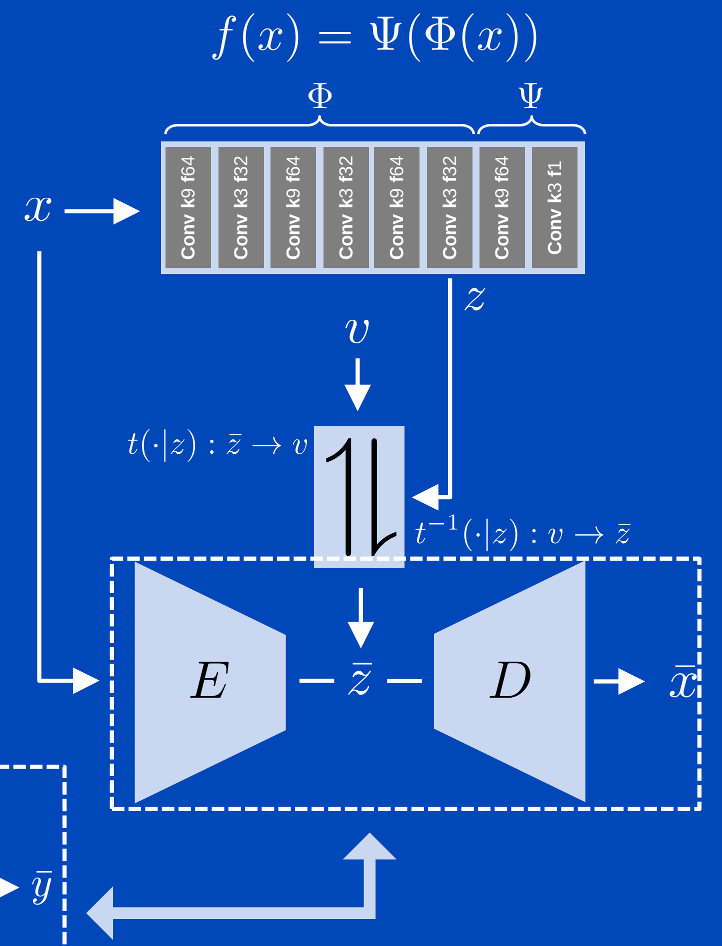


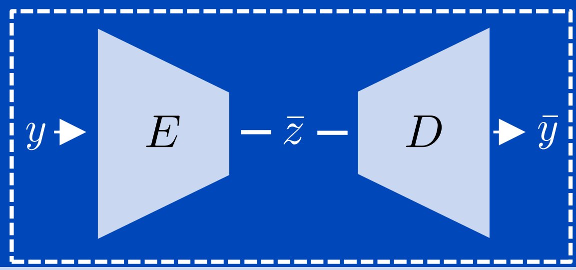
[1]C. McCollough, et al., "Data from Low Dose CT Image and Projection Data [Data Set]," *The Cancer Imaging Archive*, 2020.

Methods

Summary

1. Train denoising methods Chen et al. & Yang et al.
2. Train VAE to learn a complete data representation of the low dose images
3. For each denoising method and layer in the network we wish to evaluate, train a cINN to recover the invariances
4. For a given test image, sample N invariances (here $N = 250$), apply the inverse mapping t^{-1} and apply the pretrained decoder.
5. Train a second VAE which learns a complete data representation of the high dose images

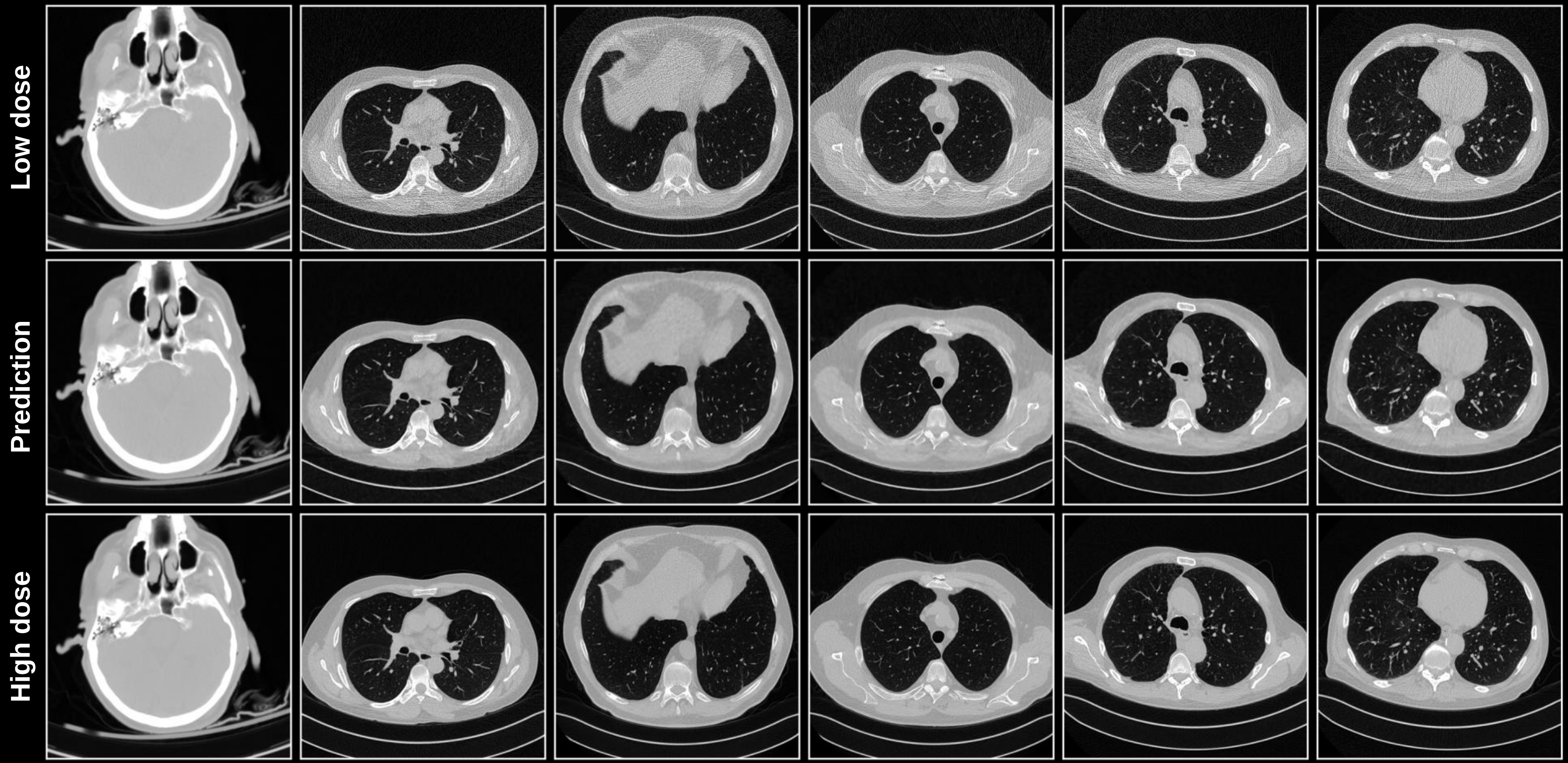




dkfz.

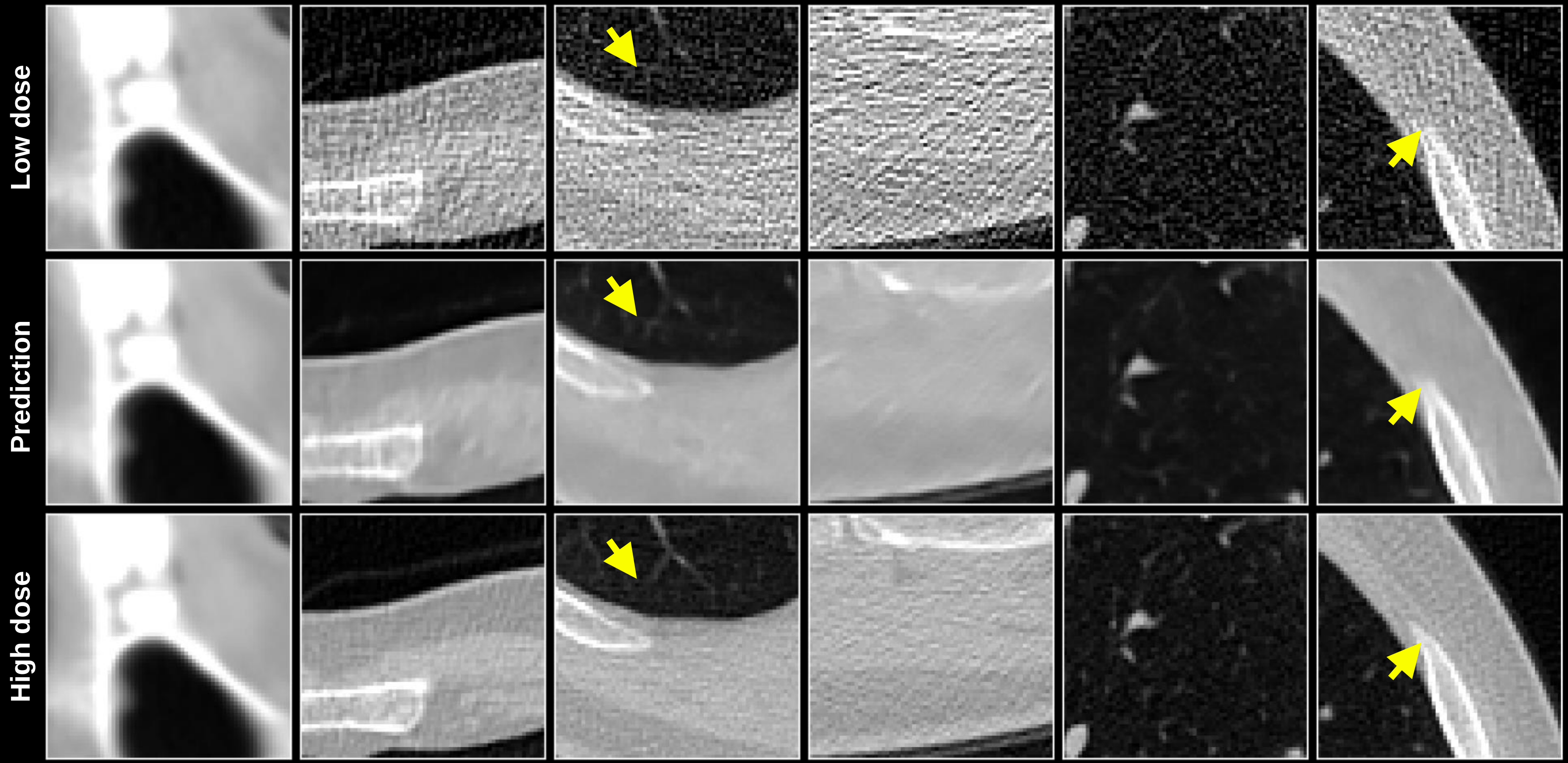
Results

Denoising (Chen et al.) $f = \Psi \circ \Phi$



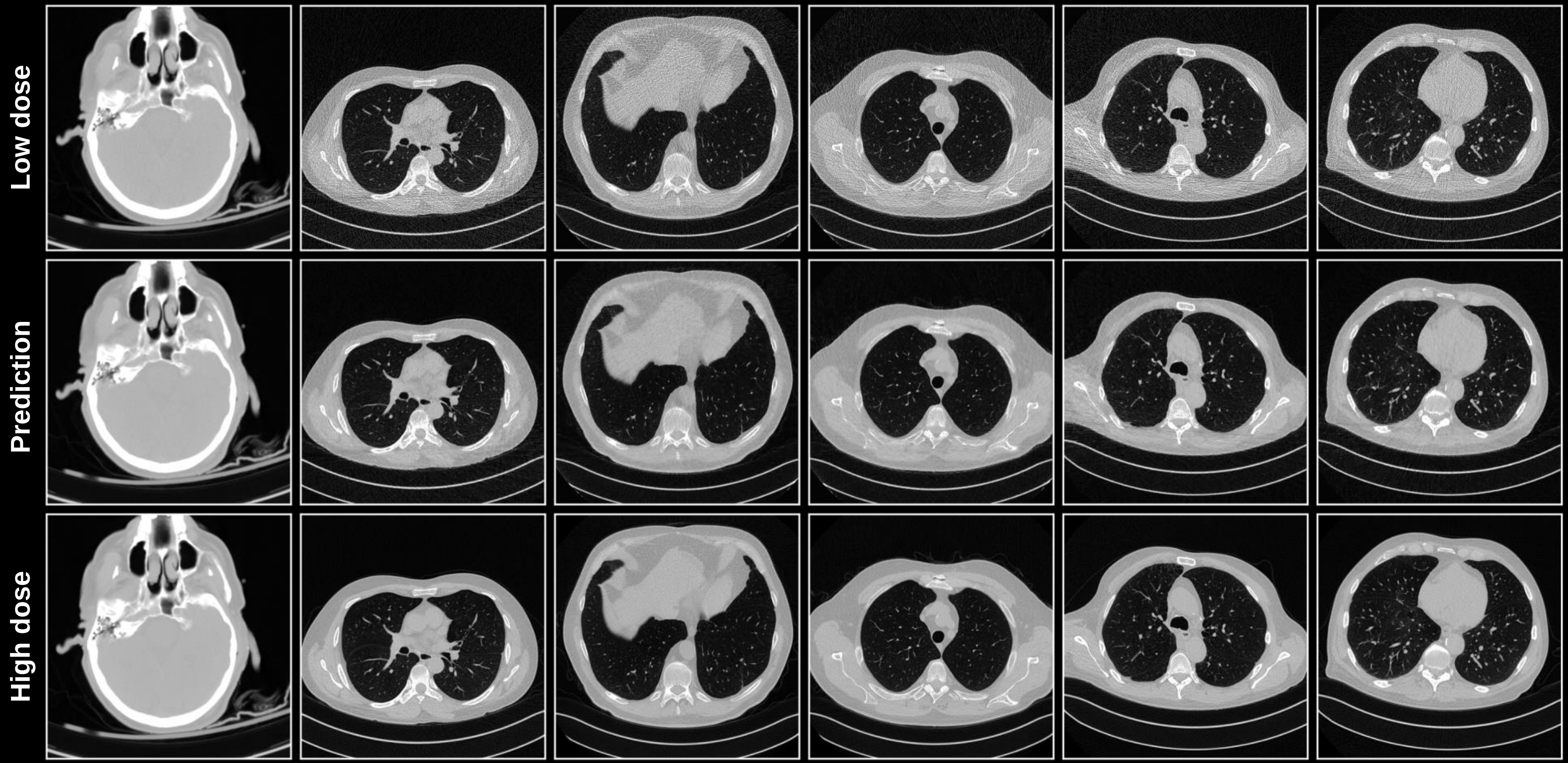
Results

Denoising (Chen et al.) $f = \Psi \circ \Phi$



Results

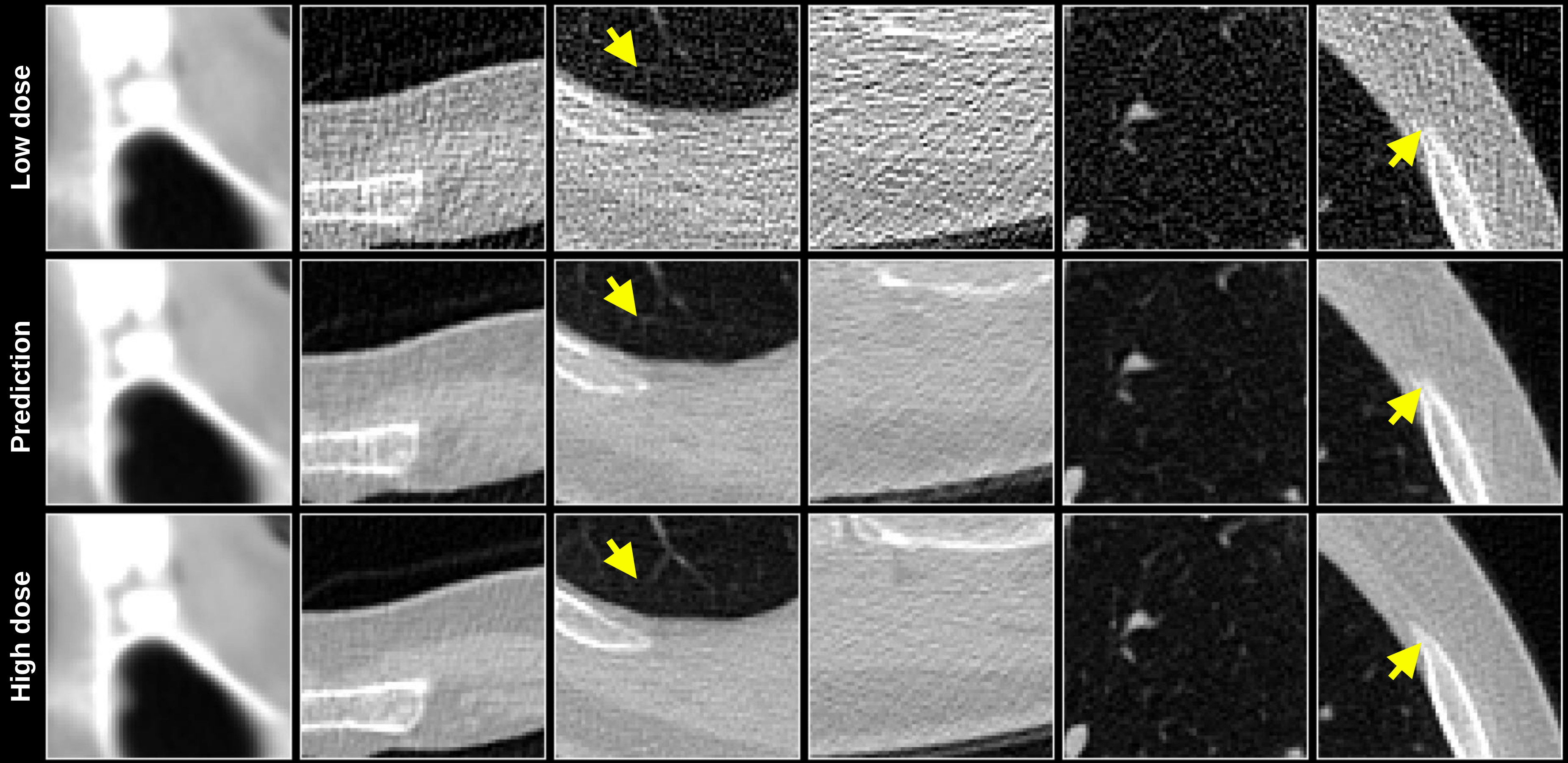
Denoising (Yang et al.) $f = \Psi \circ \Phi$

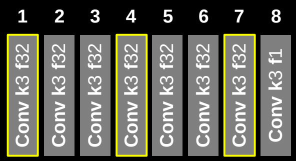


dkfz.

Results

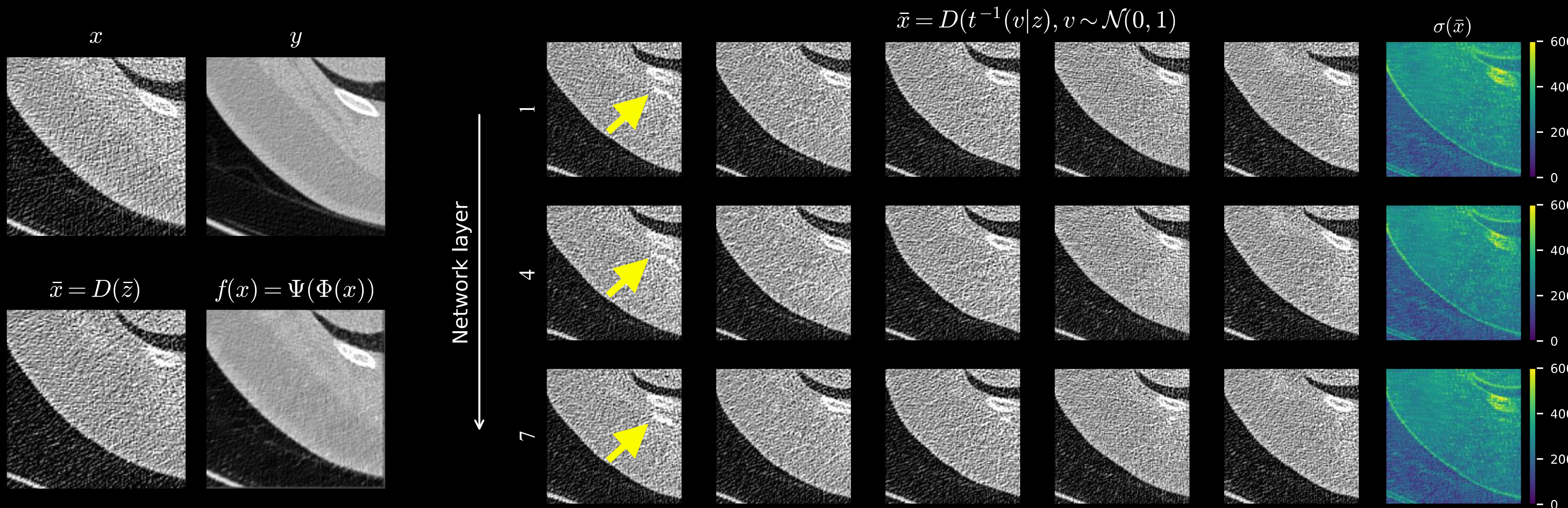
Denoising (Yang et al.) $f = \Psi \circ \Phi$

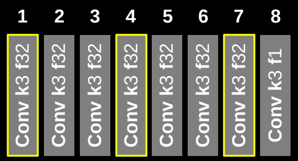




Results

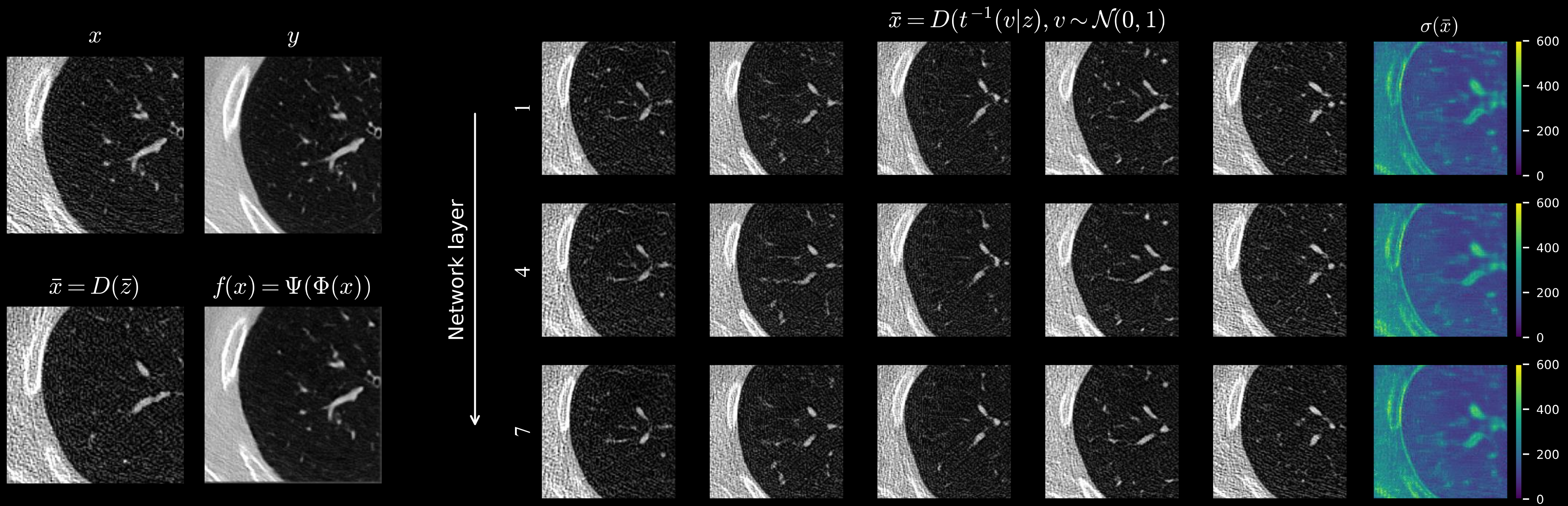
Sampling Invariances (Yang et al.)

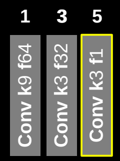




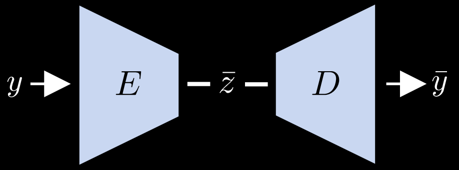
Results

Sampling Invariances (Yang et al.)

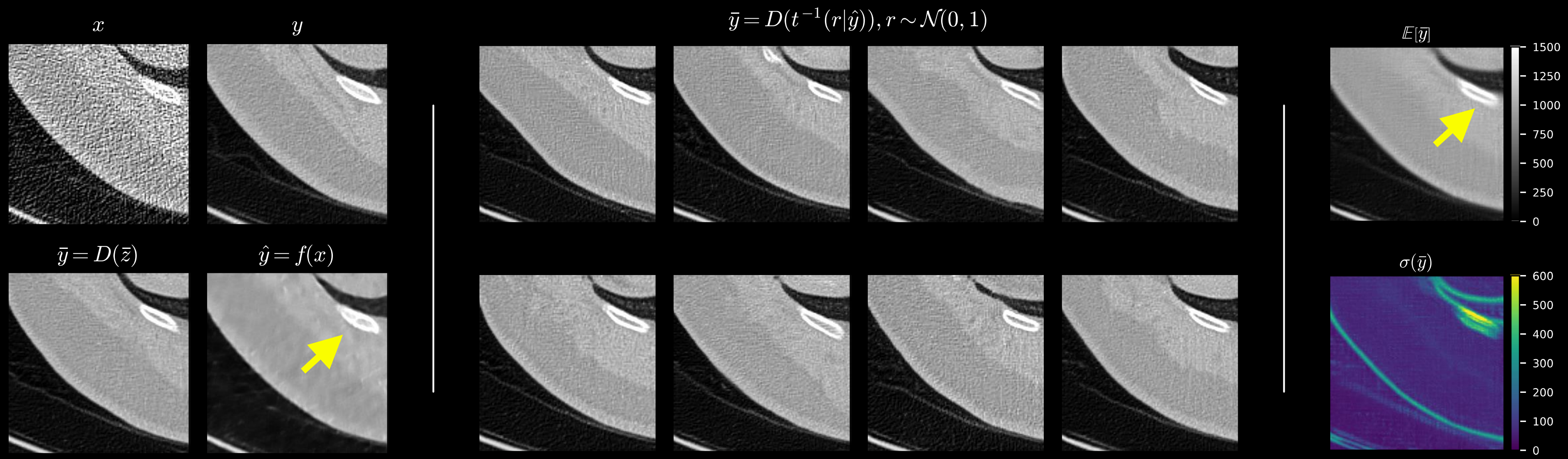


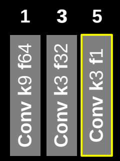


Results

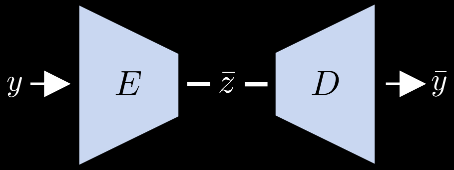


Sampling Invariances in Target Domain (Chen et al.)

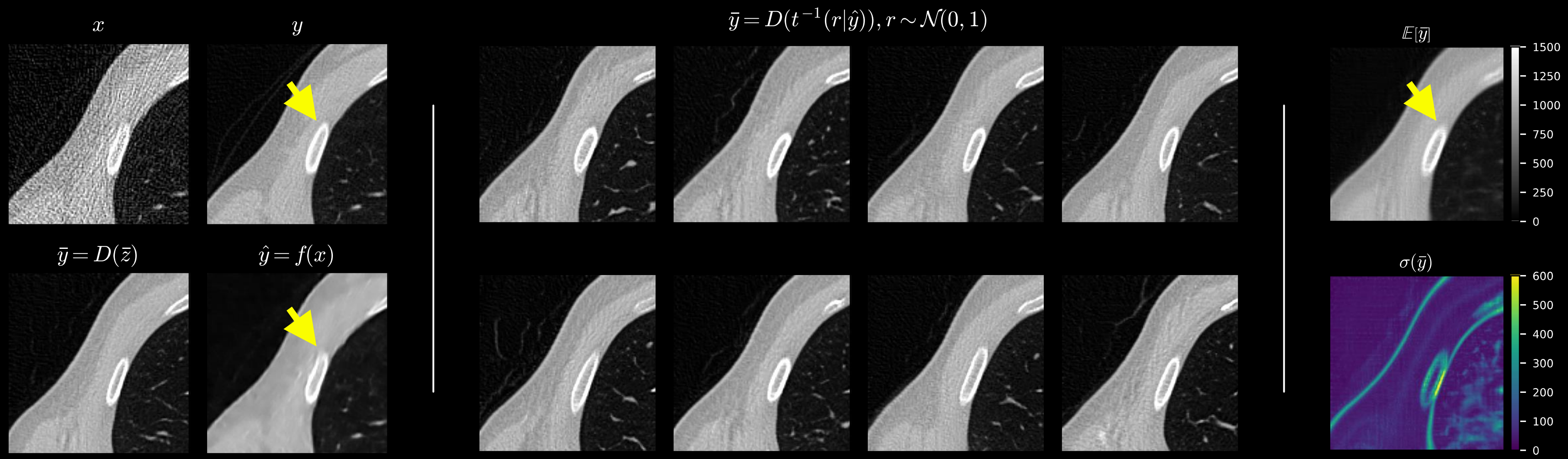




Results



Sampling Invariances in Target Domain (Chen et al.)



Conclusion & Outlook

Conclusion

- Both denoising networks perform similar as reported in their respective papers
- Yang et al. produces more realistic results compared to Chen et al. due to training in an adversarial setting
- Both denoising methods are invariant to some anatomical features to some extent
- Incomplete data representation learned by the VAE may explain some of the invariances

Outlook

- Improve interpretability by
  - Improving the embedding $\bar{z}$
  - Mapping sampled invariance images $\bar{x} = D(t^{-1}(v|z)$ to semantically meaningful space
- Minimize “undesired” invariances through a finetuning of the pretrained denoising methods

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

Job opportunities through DKFZ's international Fellowship programs (marc.kachelriess@dkfz.de).