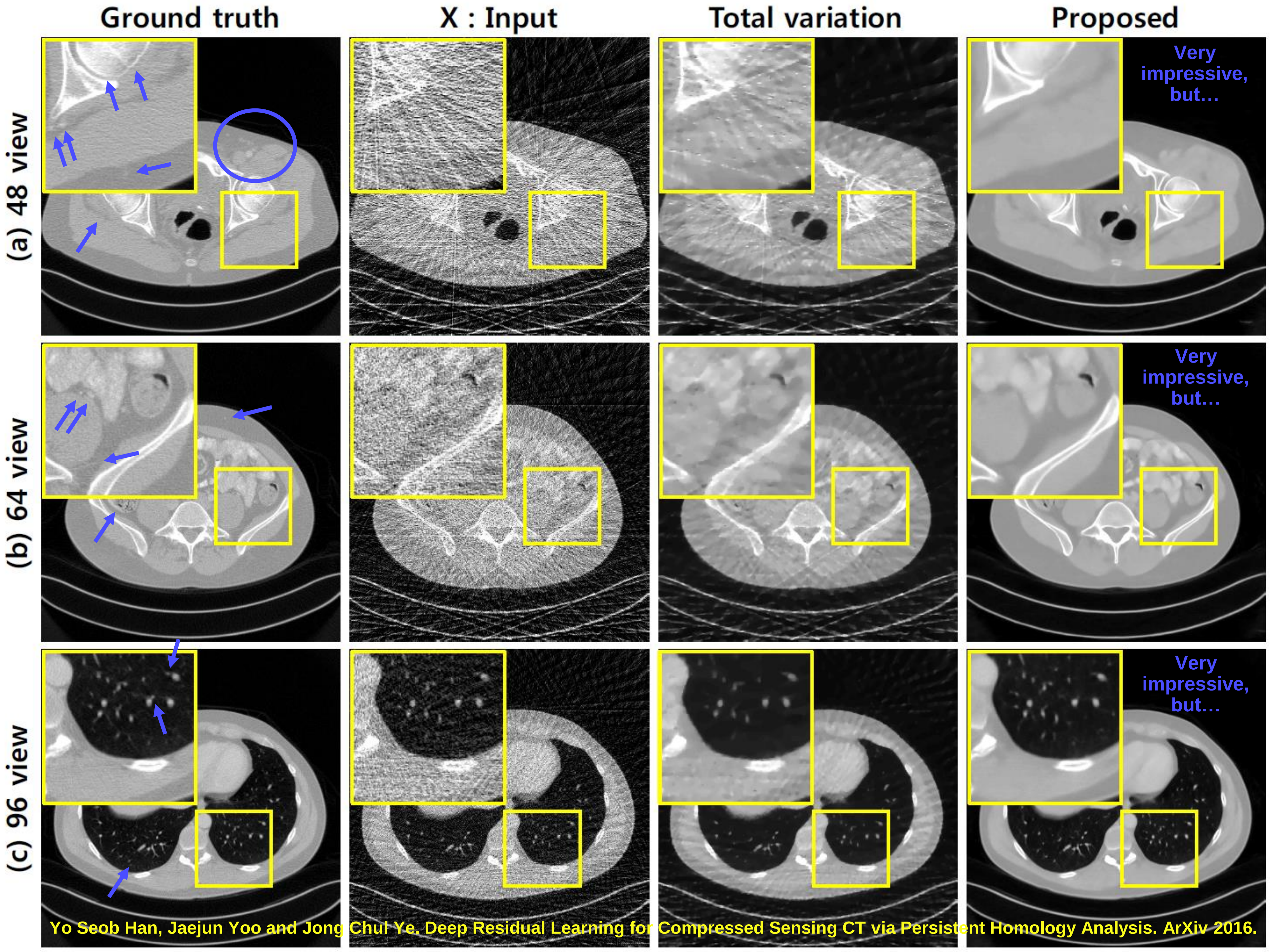
Explainable AI for CT: Analyzing CT Image Denoising Networks by Reconstructing their Invariances

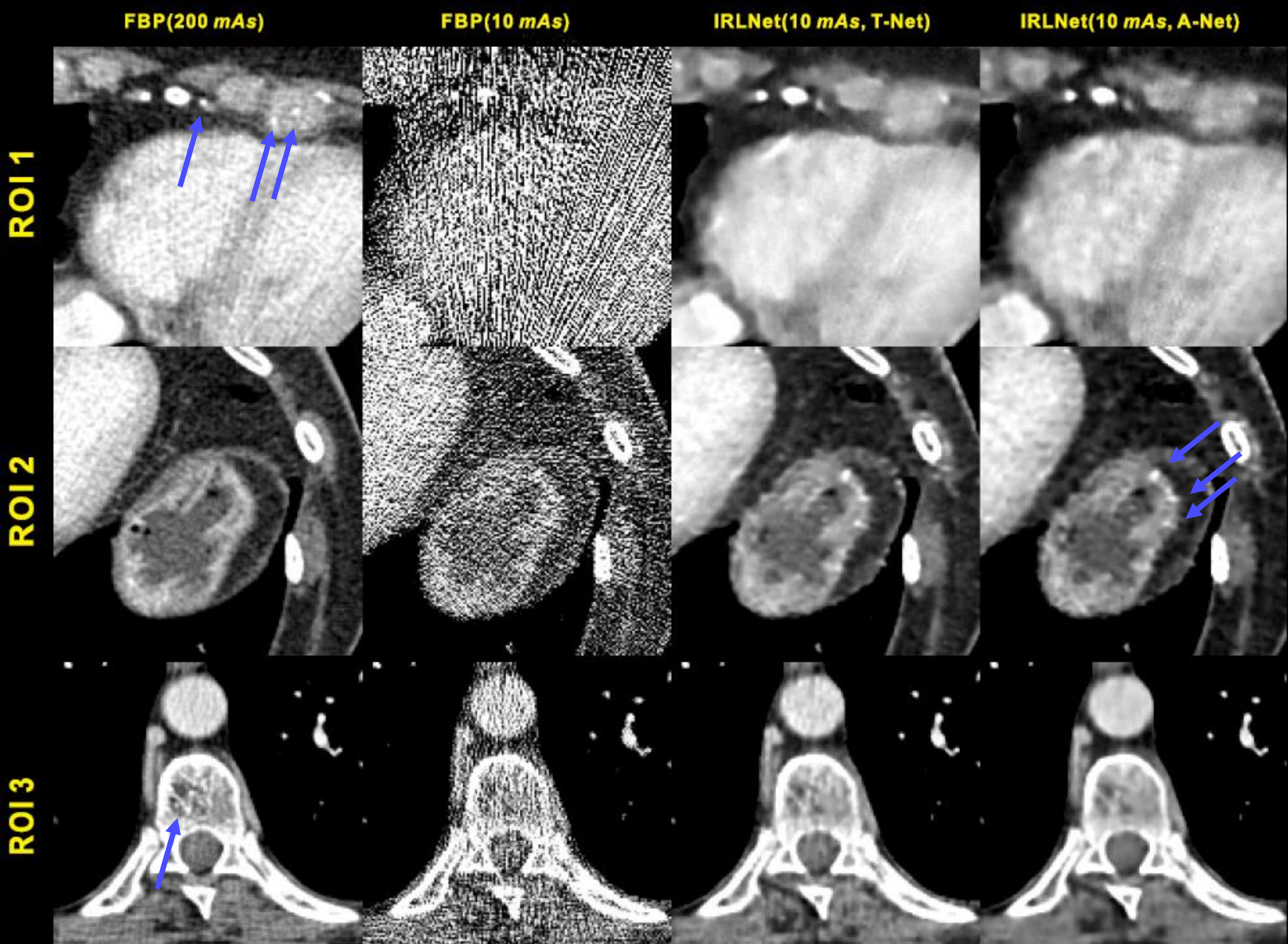
Elias Eulig[1], Björn Ommer[2,3], and Marc Kachelrieß[1]

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

DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT



Ground truth
X : Input
Total variation
Proposed
(a) 48 view
(b) 64 view
(c) 96 view
Very impressive, but…
Very impressive, but…
Very impressive, but…
Yo Seob Han, Jaejun Yoo and Jong Chul Ye. Deep Residual Learning for Compressed Sensing CT via Persistent Homology Analysis. ArXiv 2016.



Y. Wang et al. Iterative quality enhancement via residual-artifact learning networks for low-dose CT. Phys. Med. Biol. 63:215004, 2018.

Motivation

In general:

- **Deep learning methods are employed for many problems in medical image formation, including image-based noise reduction.**
- **However, they lack interpretability due to black-box nature of DNNs. Recent advancement in generative modelling signal false confidence.**

Aim of this work:

- **Lay fundamentals for post-hoc interpretability and robustness analysis of denoising DNNs.**
- **Use two simple denoising networks *f* as initial examples:**
  - **Chen‘s simple 3-layer CNN trained with $\mathcal{L}_2$ loss[1]**
  - **Yang‘s Wasserstein GAN with additional perceptual loss[2]**
- **See what they have learned to represent and what to ignore: For a given output *x´* there are many inputs *x* that produce the same output *x´* = *f*(*x*).**
- **Employ low dose CT image and projection dataset for all studies.[3]**

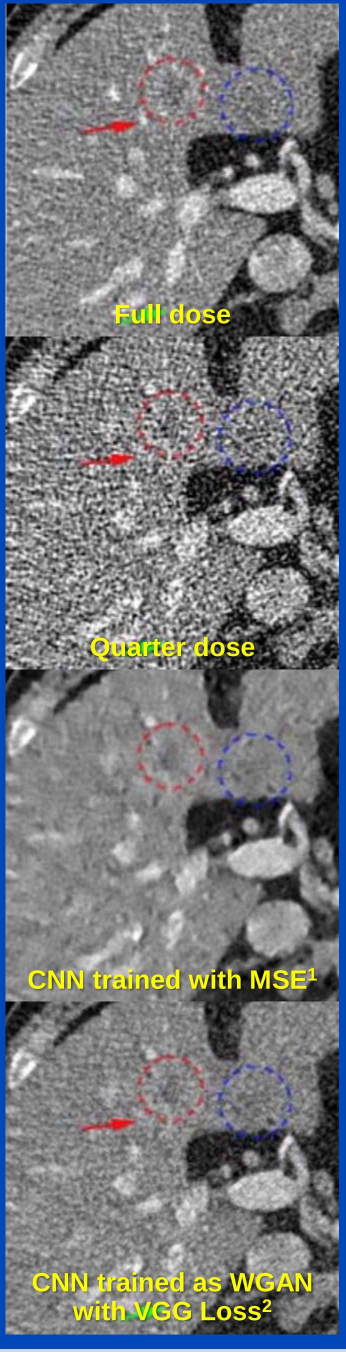


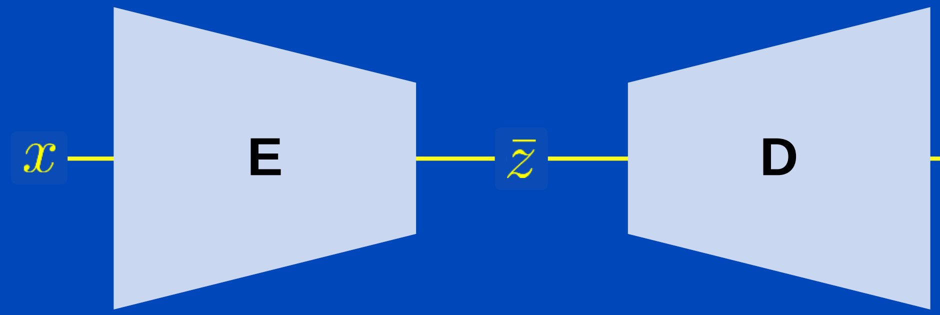
Figure from reference [2]

[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.
[3]C. McCollough et al., “Data from low dose CT image and projection data [data set],” The Cancer Imaging Archive, 2020.

dkfz.

Recap 1: What is an Autoencoder (AE)?

- **In and output domain are the same, here *x*.**
- **Bottleneck *z* enforces the encoder and decoder to do a good job.**

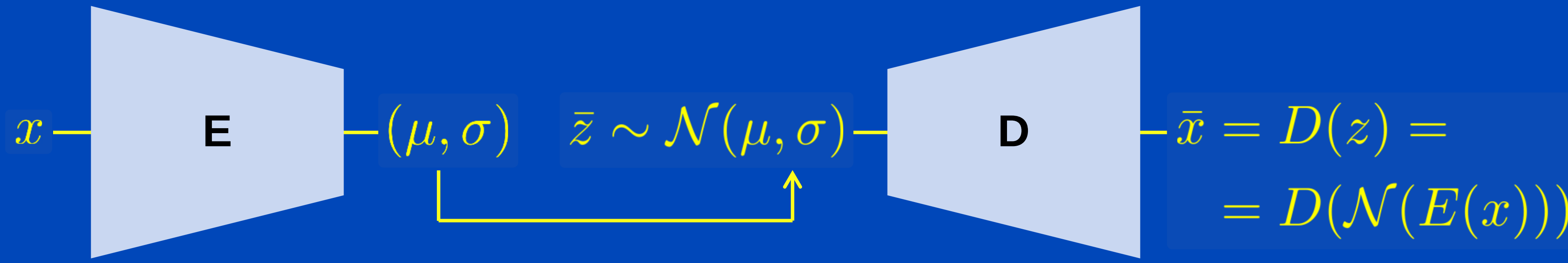


$$\begin{aligned} \bar{x} &= D(z) = \\ &= D(E(x)) \end{aligned}$$

- **Examples:**
  - **Principal component analysis (linear autoencoder), lossless**
  - **PCA with dimensionality reduction (nonlinear due to clipping), lossy**
  - **Image compression and decoding, e.g. jpeg, lossy**
- **Latent space typically not interpretable.**

Recap 2: What is a Variational AE (VAE)?

- Make latent space regular.
- Allow to sample in latent space from a given distribution, here: normal distribution.



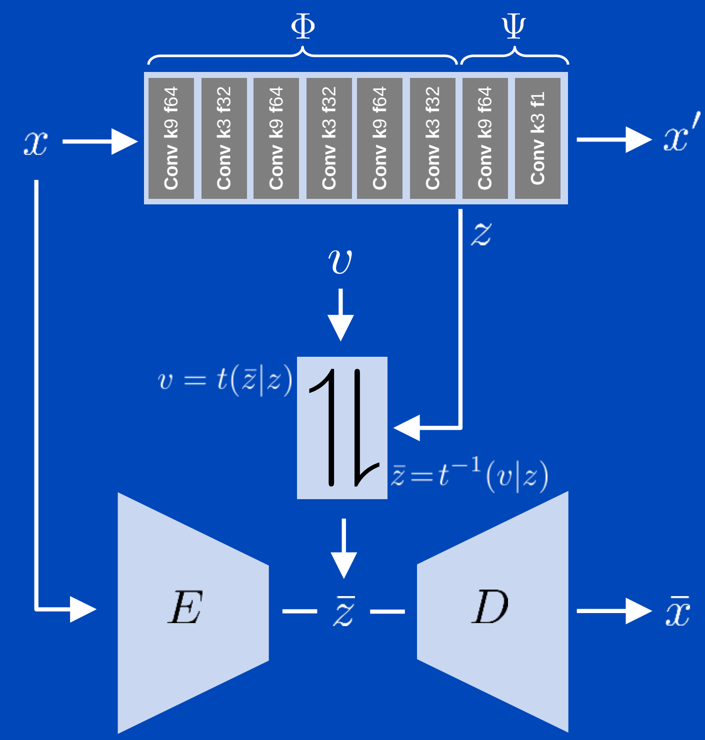
- The VAE is a generative model.
- It allows to generate new data by sampling new values from the normal distribution.

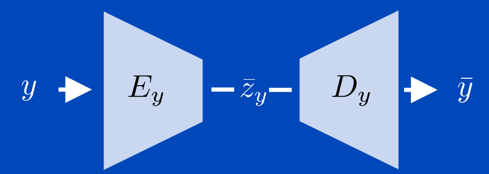
Method

Recovering Invariances

- Our work is based on Rombach et al.[1]
- Given a function or network $f(x) = \Psi(\Phi(x))$ we analyze its internal latent representations
  $z = \Phi(x)$.
- Train a VAE to learn a complete data representation $\bar{z} = E(x)$ of low dose images.
- Disentangle information captured in z and invariances v by learning a mapping
  $$v = t(\bar{z}|z),\ \mathcal{L}(v) = \mathcal{N}(0,1)$$
- $t(\cdot|z)$ is realized by a conditional invertible neural network (cINN).
- Generate new images varying only by their invariances
  $$\bar{x} = D(t^{-1}(v|z)) \qquad v \sim \mathcal{N}(0,1)$$

$$f(x) = \Psi(\Phi(x))$$





Alternative: Use VAE in high dose domain, i.e. VAE_y, to visualize the invariances.

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

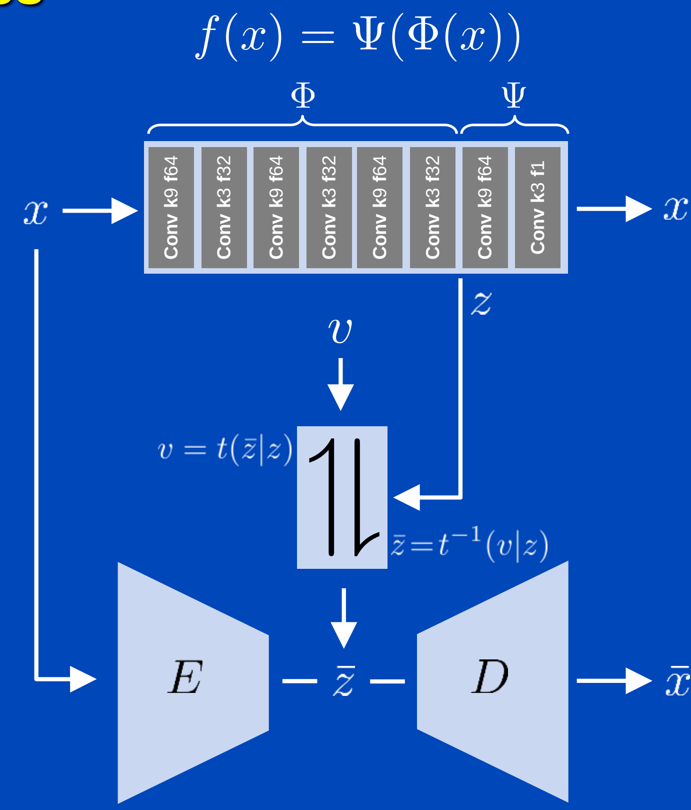
Method

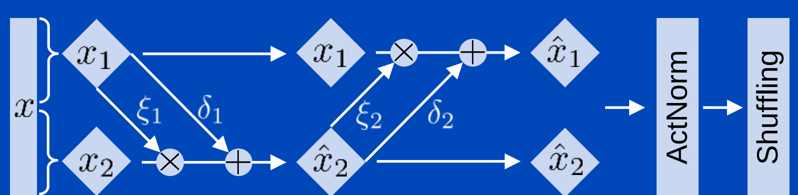
Recovering Invariances

1. Our work is based on Rombach et al.[1]
2. **Train denoising methods** Chen et al. & Yang et al.
3. Train VAE to learn a complete data representation of the **low dose** images *x*.
4. For each denoising method and layer in the network we wish to evaluate, train a **cINN to recover the invariances.**
5. For a given test image, sample 250 invariances *v*, apply the inverse mapping t^{-1} and apply the pretrained decoder *D*.

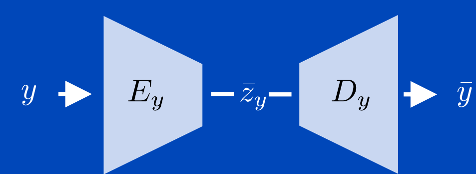
$$t^{-1} \text{ maps } \mathcal{N}(0,1) \text{ onto } p(\bar{z}|z).$$

Thus it produces only images that are likely under the training distribution of the AE.





Building block of INN: Invertible block, ξ_{12} and δ_{12} are CNNs or NNs



Alternative: Use VAE in high dose domain, i.e. VAE_y, to visualize the invariances.

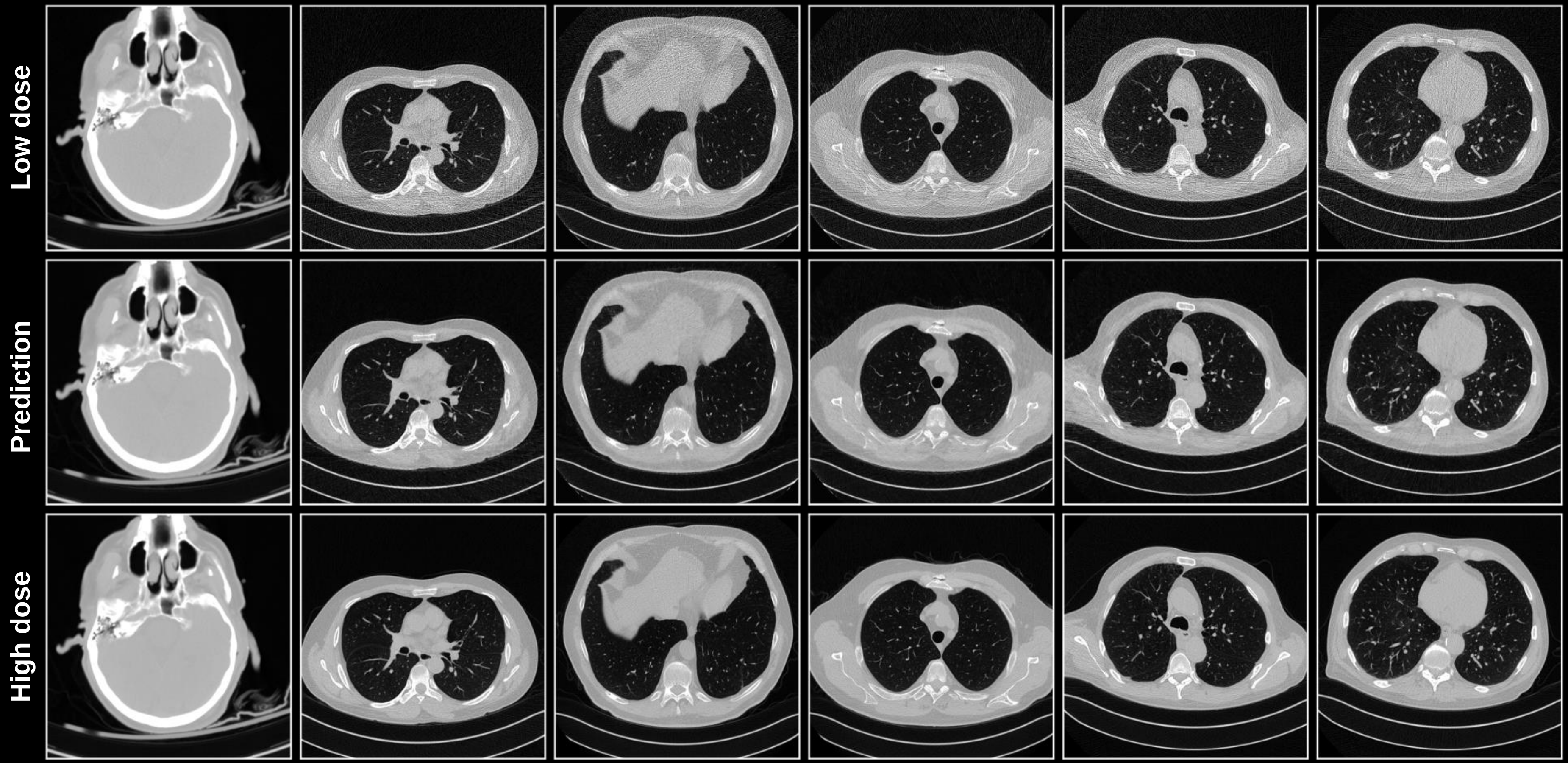
$$x_1 \exp(\xi_2(\hat{x}_2)) + \delta_2(\hat{x}_2) = \hat{x}_1$$
$$x_2 \exp(\xi_1(x_2)) + \delta_1(x_1) = \hat{x}_2$$

[1]Rombach, et al. "Making sense of CNNs: Interpreting deep representations and their invariances with INNs", ECCV 2020, 2020.

dkfz.

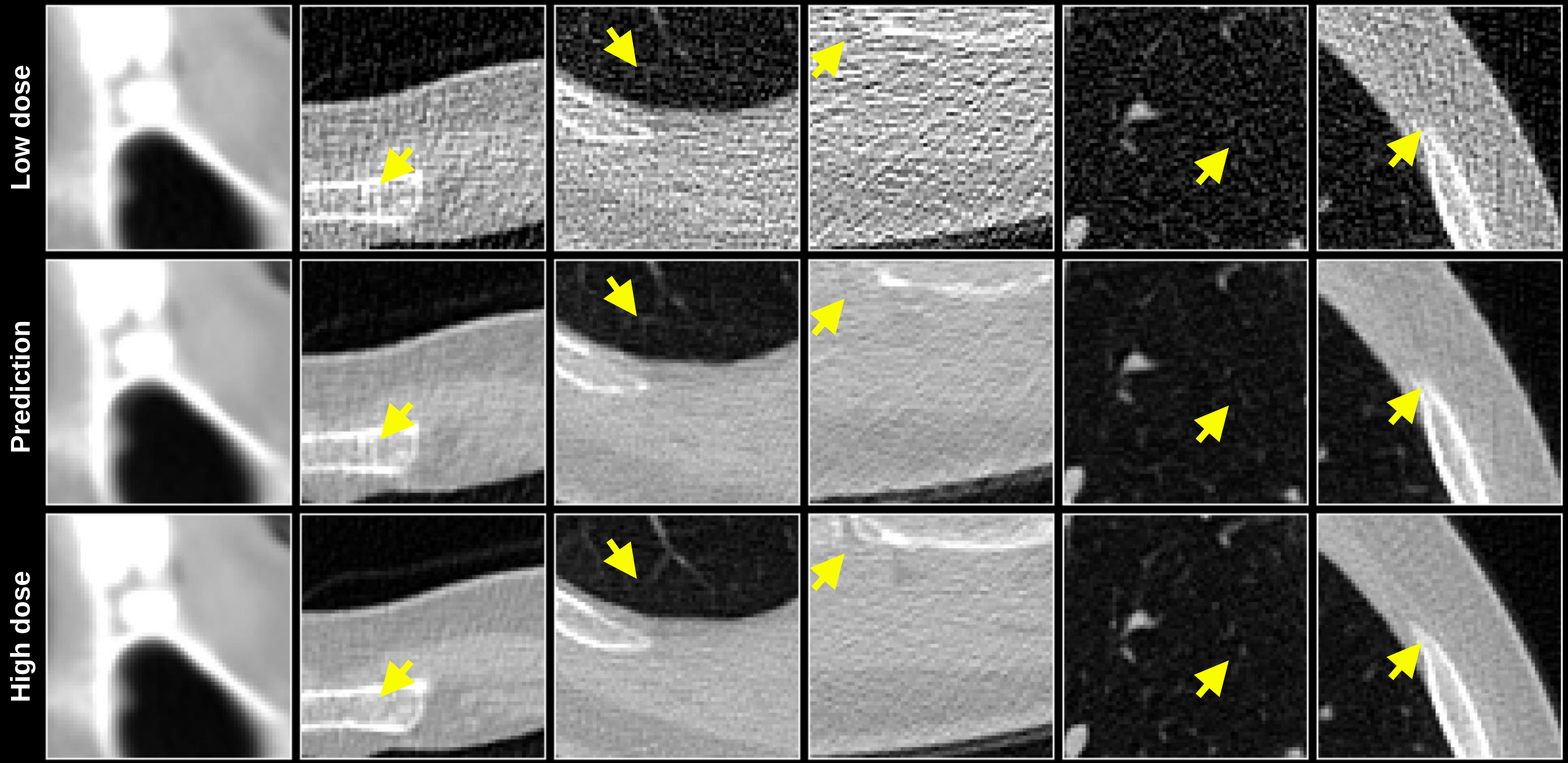
Results

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

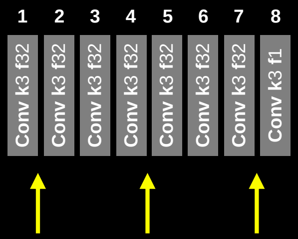


Results

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

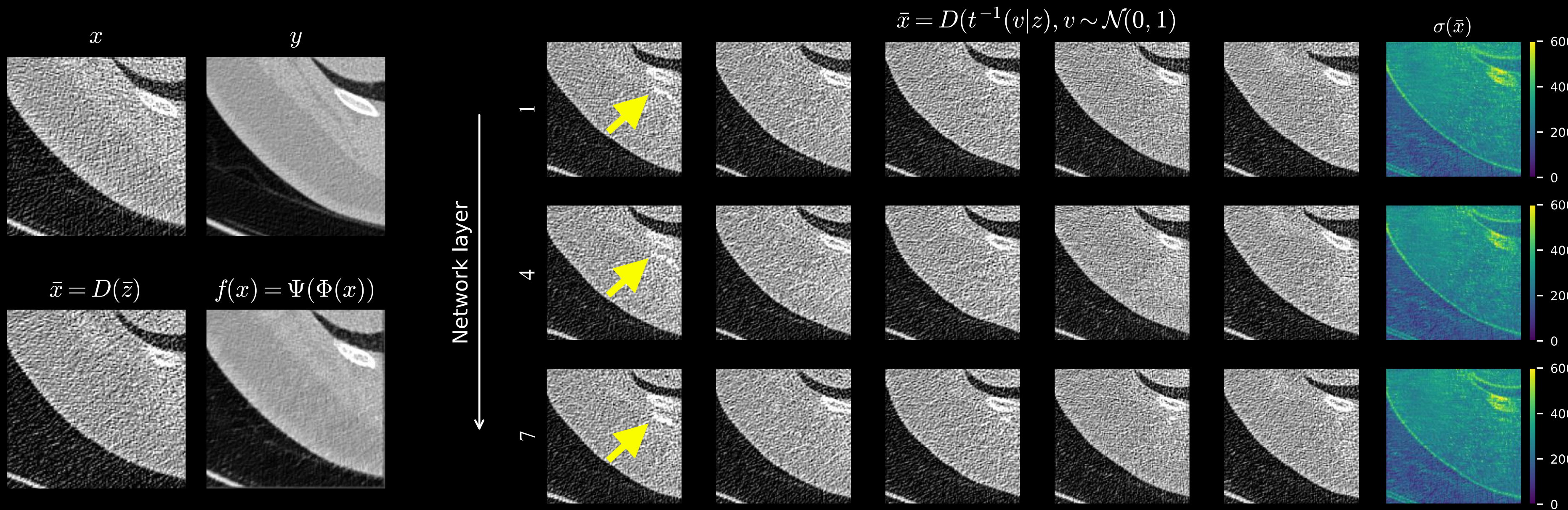


Arrows point at selected differences between prediction and ground truth.



Results

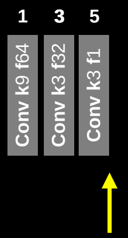
Sampling Invariances in Yang et al.'s Net

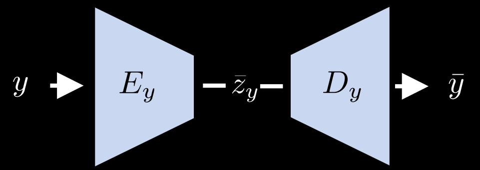


$$\Phi(x) = \Phi(\bar{x}) \quad \forall\ \bar{x}$$

$$x' = f(x) = f(\bar{x}) \quad \forall\ \bar{x}$$

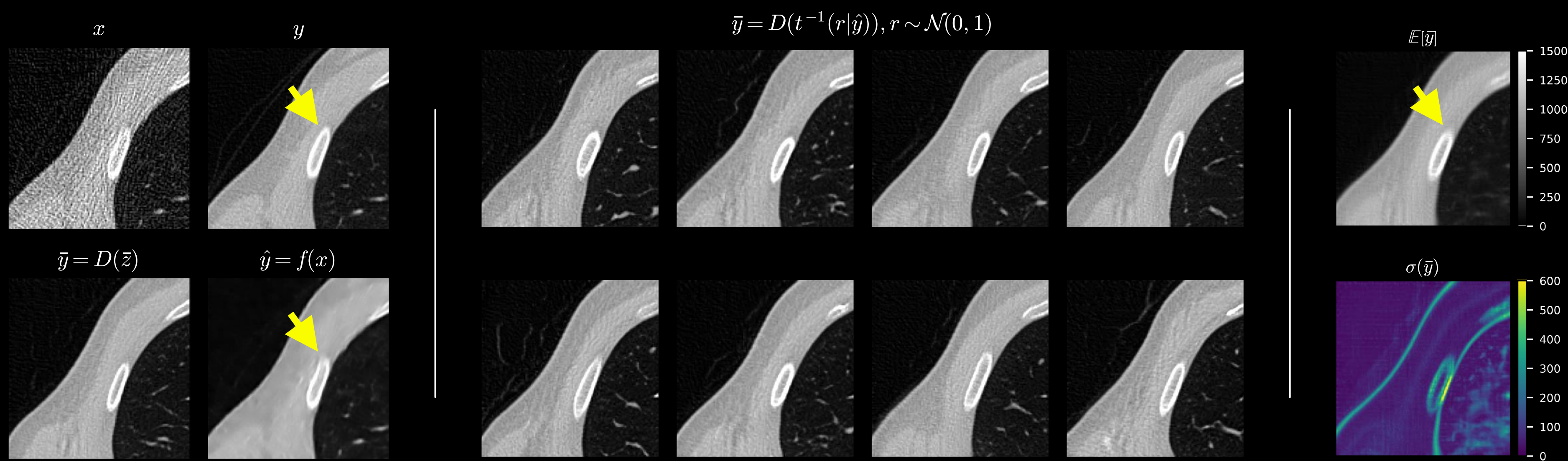
Same samples of *v* used for the rows corresponding to wiretapping after layers 1, 4 and 7.





Results

Sampling Invariances in Target Domain in Chen et al.’s Net



$$x' = f(x) = f(\mathsf{R}^{-1}\mathcal{P}\,\mathsf{R}\,\bar{y}) \quad \forall\ \bar{y}$$

forward project
add Poisson noise
reconstruct

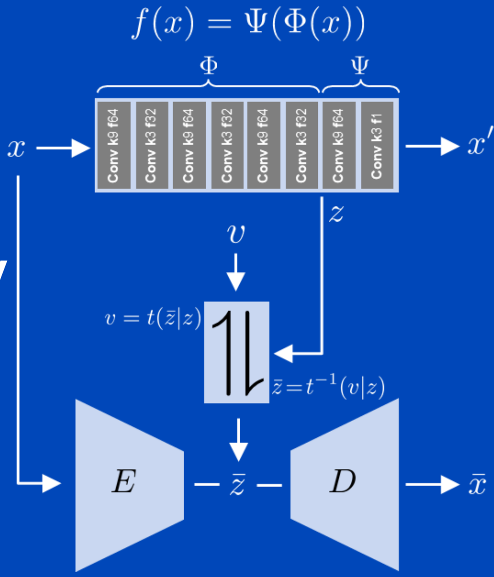
Conclusions & Outlook

Conclusions

- Designed a method to highlight invariances of a given network.
- Algorithm agnostic, not restricted to denosing .
- Architecture agnostic, not restricted to CT.
- Both denoising methods are invariant to some anatomical features to some extent.

Outlook

- Improve interpretability by
  - improving the embedding of the VAEs,
  - mapping sampled invariance images to semantically meaningful space (disentangled representations of e.g. tumors).
- One could use the undesired invariances to finetune the denoising methods.



Thank You!

This presentation will soon be available at www.dkfz.de/ct.
This work was supported by the Helmholtz International Graduate School for Cancer Research.
Job opportunities through DKFZ’s international Fellowship programs (marc.kachelriess@dkfz.de).