Problems of Assessing AI-Based CT Image Reconstruction, Denoising or Artifact Reduction

Marc Kachelrieß

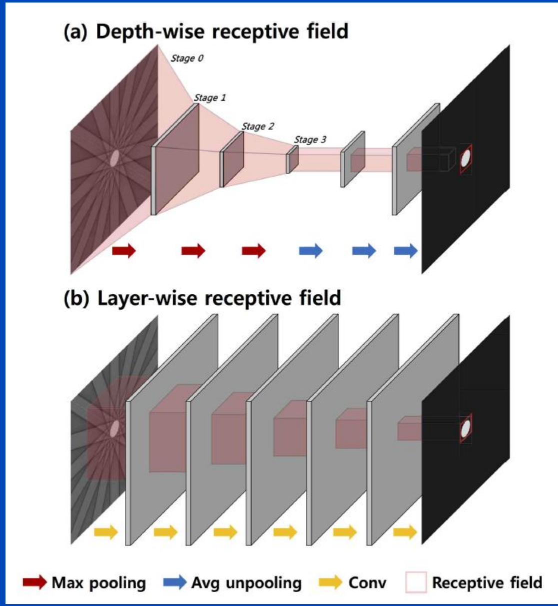
German Cancer Research Center (DKFZ)
Heidelberg, Germany
www.dkfz.de/ct

DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

Unmeasured information is often faked

PROBLEMS WITH AI-BASED RECON

Sparse View Restoration Example



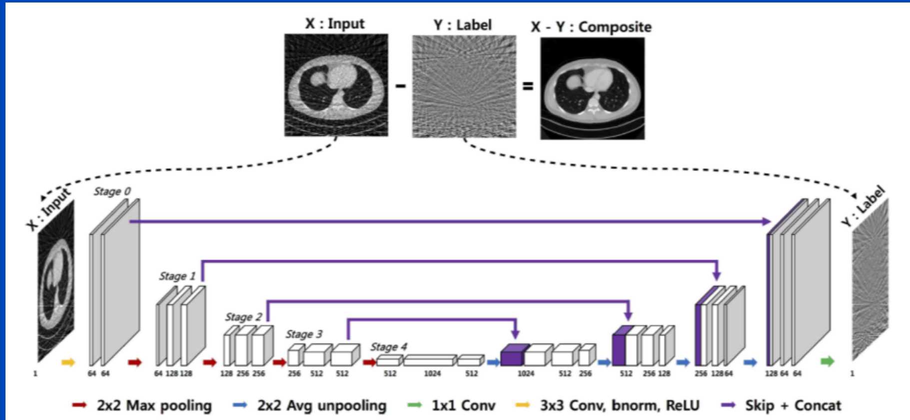
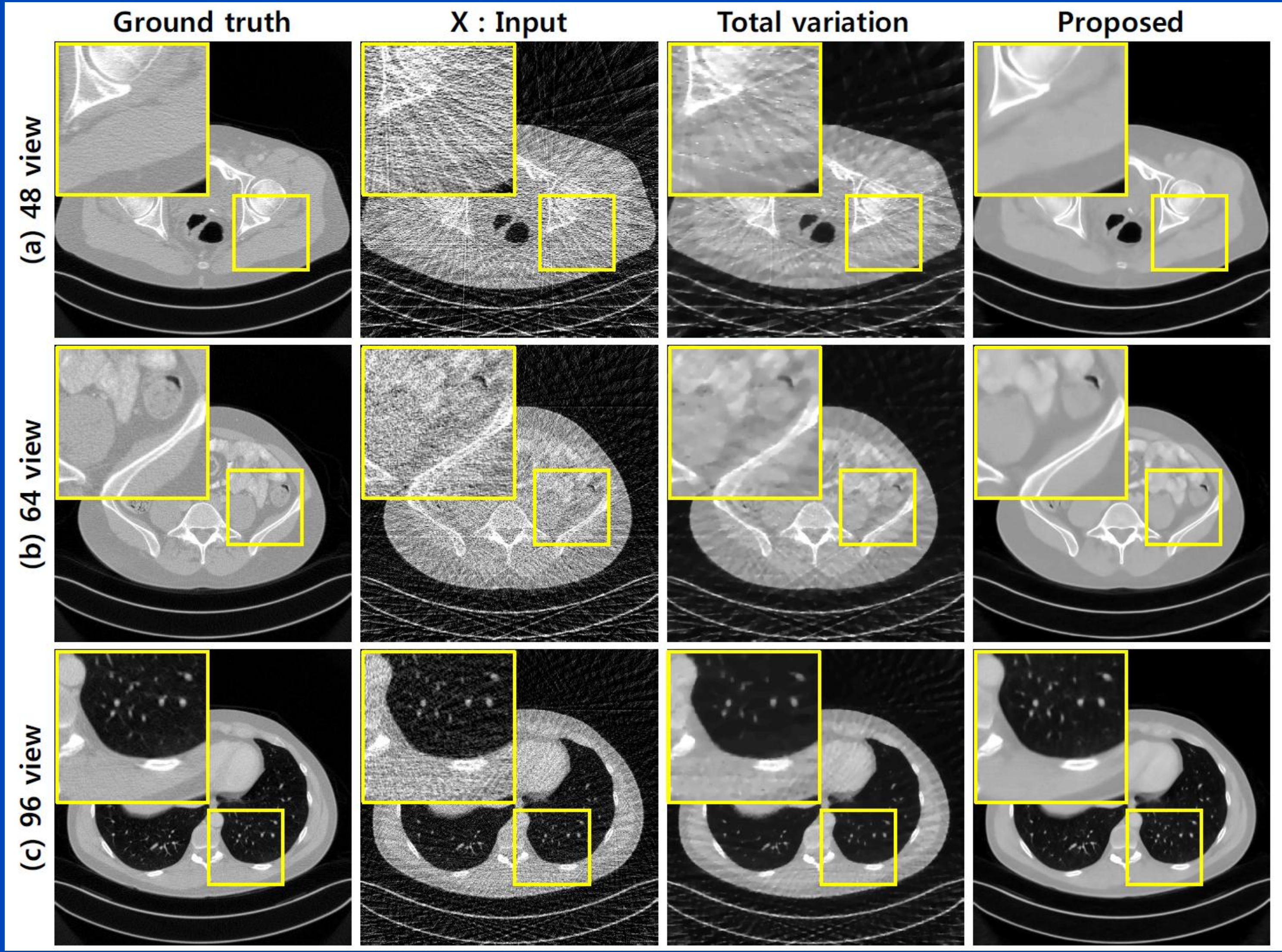
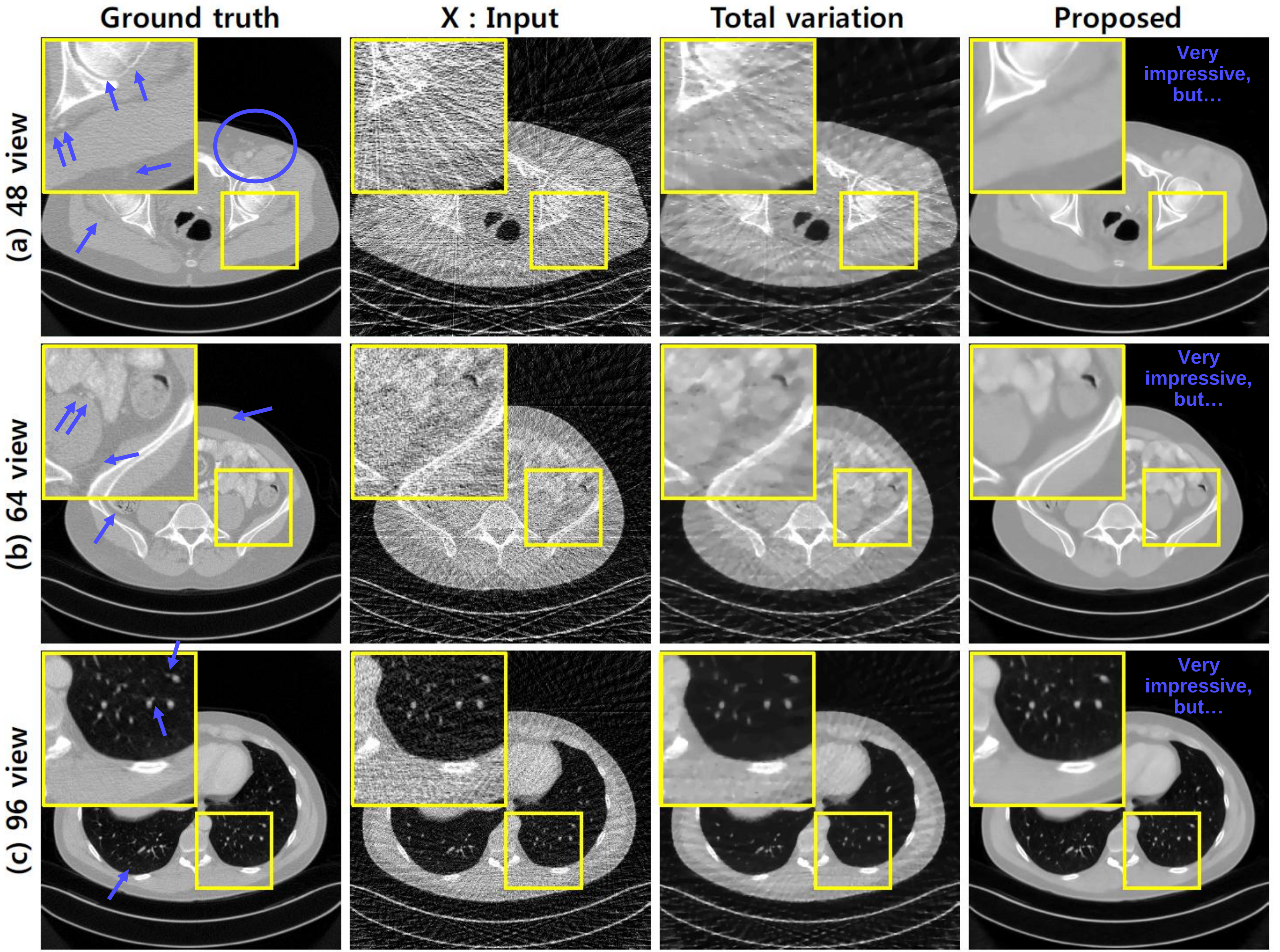


Figure 1. The proposed deep residual learning architecture for sparse view CT reconstruction.



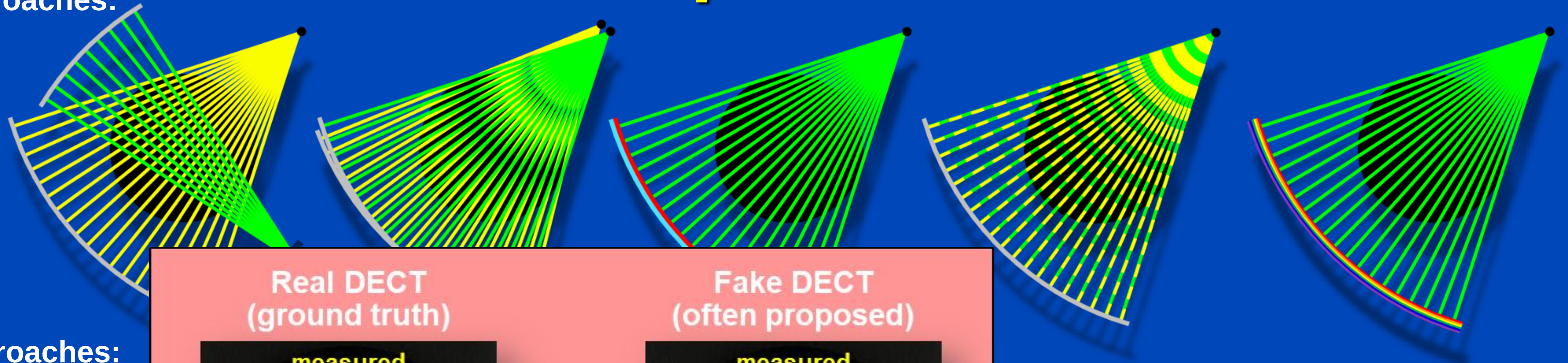
Yo Seob Han, Jaejun Yoo and Jong Chul Ye. Deep Residual Learning for Compressed Sensing CT Reconstruction via Persistent Homology Analysis. ArXiv 2016.


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…

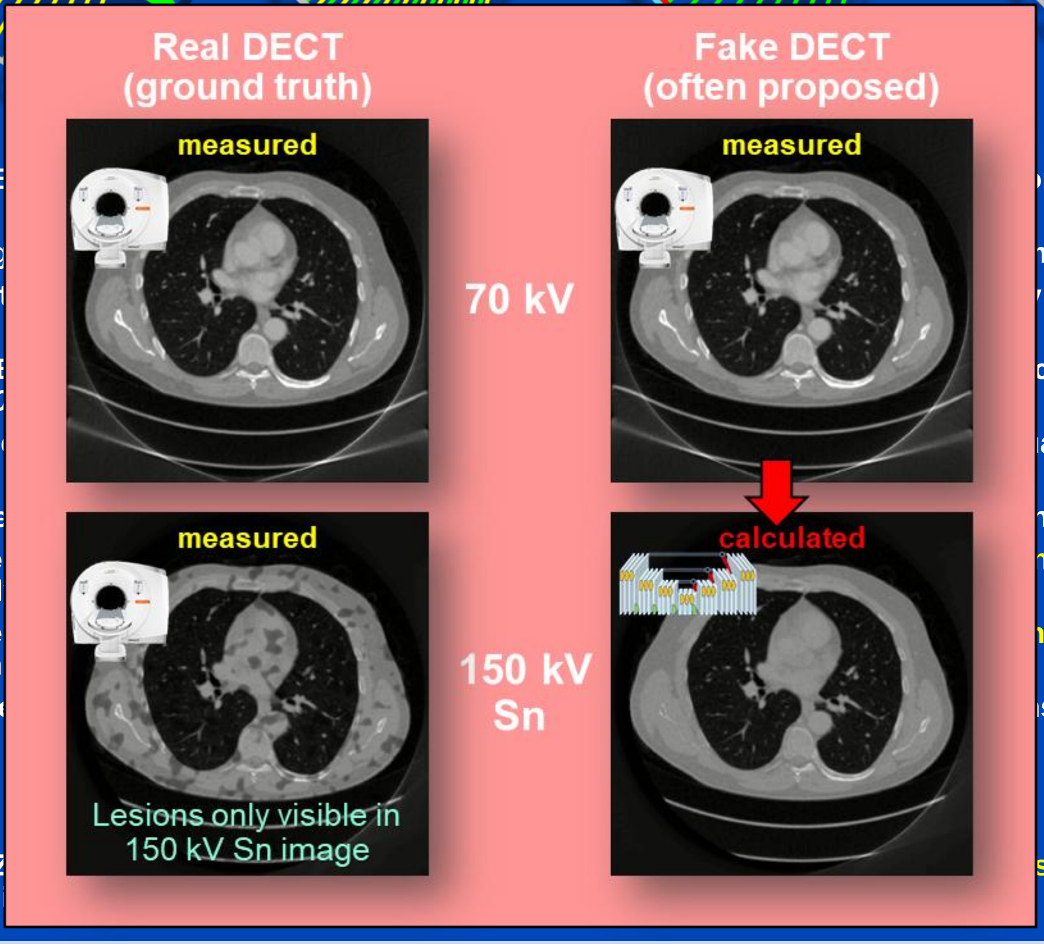
dkfz.

True and Fake Spectral CT

Existing <u>true</u> spectral CT approaches:



Existing <u>fake</u> spectral CT approaches:



[1] J. Ma, Y. Liao, Y. Wang, S. Li, J. He, D. Zeng, Z. B … rk: basic material estimation“, *SPIE Medical Imaging 2018*.

[2] W. Zhao, T. Lv, P. Gao, L. Shen, X. Dai, K. Cheng … naging using a single-energy CT data”, *Fully3D 2019.*

[3] D. Lee, H. Kim, B. Choi, H. J. Kim, “Development … images with single x-ray exposure”, PMB 64(11), 2019.

[4] L. Yao, S. Li, D. Li, M. Zhu, Q. Gao, S. Zhang, Z. B … ct energy-resolving CT imaging via existing energy-integrating CT images”, *SPIE Medical Imaging 2020*

[5] D. P. Clark, F. R. Schwartz, D. Marin, J. C. Ramir … al-source, dual-energy x-ray CT”, Med. Phys. 47 (9): 4150–4163, 2020.

[6] C. K. Liu, C. C. Liu, C. H. Yang, H. M. Huang, “Ge … ng”, Journal of Digital Imaging 34(1):149–161, 2021.

[7] T. Lyu, W. Zhao, Y. Zhu, Z. Wu, Y. Zhang, Y. Che … nergy CT data with material decomposition convolutional neural network”, Medical Image Anal

[8] F. R. Schwartz, D. P. Clark, Y. Ding, J. C. Ramire … ning based extension of dual-energy FoV in dual-source CT—A retrospective pilot study”, European

[9] Y. Li, X. Tie, K. Li, J. W. Garrett, G.-H. Chen, “Dee … s using energy integration detectors”, *SPIE Medical Imaging 2022*.

…

[18] T. Wang, C. Jiang, W. Ding, Q. Chen, D. Shen, Z … s based on single-energy CT to differentiate intracranial hemorrhage and contrast staining withi … 2025.

J. Maier, J. Erath, S. Sawall, E. Fournié, K. Stierstorfer, and M. Kachelrieß. Raw data consistent deep learning-based field of view extension for dual-source dual-energy CT. Med. Phys. 51(3):1822-1831, March 2024.

Fake Contrast Enhancement

[1] G. Santini, L. M. Zumbo, N. Martini, G. Valvano, A. Leo, A. Ripoli, F. Avogliero, D. Chiappino, D. D. Latta, “Synthetic contrast enhancement in cardiac CT with deep learning,” arXiv 1807:01779, 2018.

[2] J. Liu, Y. Tian, A. M. Ağıldere, K. M. Haberal, M. Coşkun, C. Duzgol, and O. Akin, “DyeFreeNet: Deep virtual contrast CT synthesis,” Lecture Notes in Computer Science. Springer International Publishing, pp. 80–89, 2020.

[3] A. Chandrashekar, A. Handa, N. Shivakumar, P. Lapolla, V. Grau, R. Lee, “A deep learning approach to generate contrast-enhanced computerised tomography Angiography without the use of intravenous contrast agents,” arXiv 2003.01223, 2020.

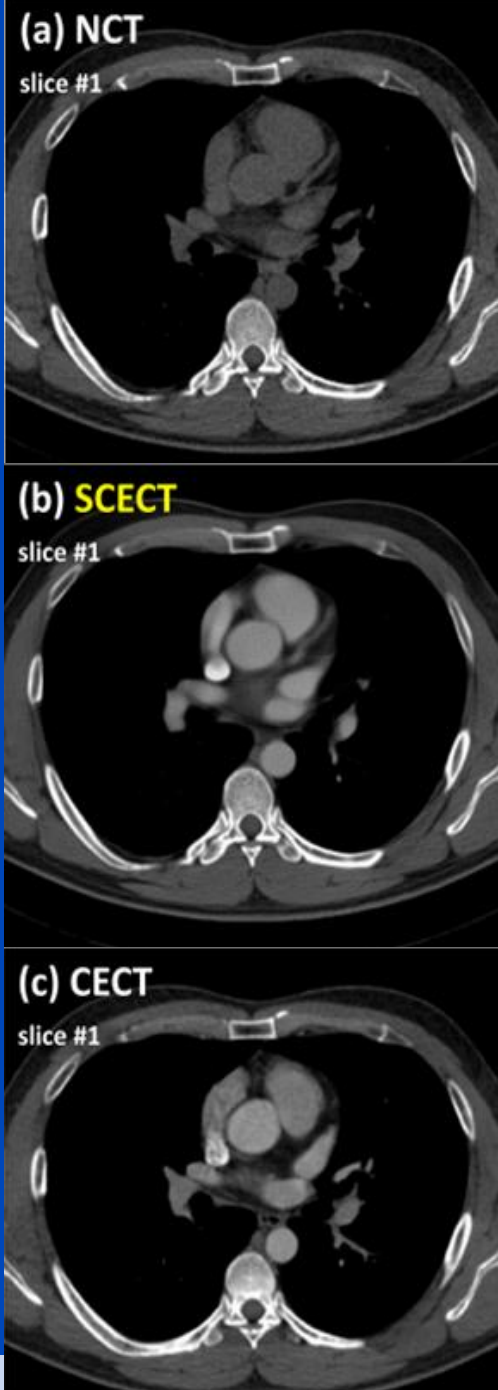
[4] J. W. Choi, Y. J. Cho, J. Y. Ha, S. B. Lee, S. Lee, Y. H. Choi, J.-E. Cheon, and W. S. Kim, “Generating synthetic contrast enhancement from non-contrast chest computed tomography using a generative adversarial network,” Scientific Reports, vol. 11, no. 1, 2021.

[5] S. W. Kim, J. H. Kim, S. Kwak, M. Seo, C. Ryoo, C.-I. Shin, S. Jang, J. Cho, Y.-H. Kim, and K. Jeon, “The feasibility of deep learning-based synthetic contrast-enhanced CT from non-enhanced CT in emergency department patients with acute abdominal pain,” Scientific Reports, vol. 11, 2021.

[6] J. Chun, J. S. Chang, C. Oh, I. Park, M. S. Choi, C.-S. Hong, H. Kim, G. Yang, J. Y. Moon, S. Y. Chung, Y. J. Suh, and J. S. Kim, “Synthetic contrast-enhanced computed tomography generation using a deep convolutional neural network for cardiac substructure delineation in breast cancer radiation therapy: a feasibility study,” Radiation Oncology, vol. 17, no. 1, 2022.

[7] Y. Gao, H. Xie, C. Chang, J. Peng, S. Pan, R. L. J. Qiu, T. Wang, B. Ghavidel, J. Roper, J. Zhou, and X. Yang, “CT-based synthetic iodine map generation using conditional denoising diffusion probabilistic model,” Medical Physics, vol. 51, no. 9, pp. 6246–6258, 2024.

[8] S. Han, J.-M. Kim, J. Park, S. W. Kim, S. Park, J. Cho, S.-J. Park, H.-J. Chung, S.-M. Ham, S. J. Park, and J. H. Kim, “Clinical feasibility of deep learning based synthetic contrast-enhanced abdominal CT in patients undergoing non-enhanced CT scans,” Scientific Reports, vol. 14, no. 1, 2024.

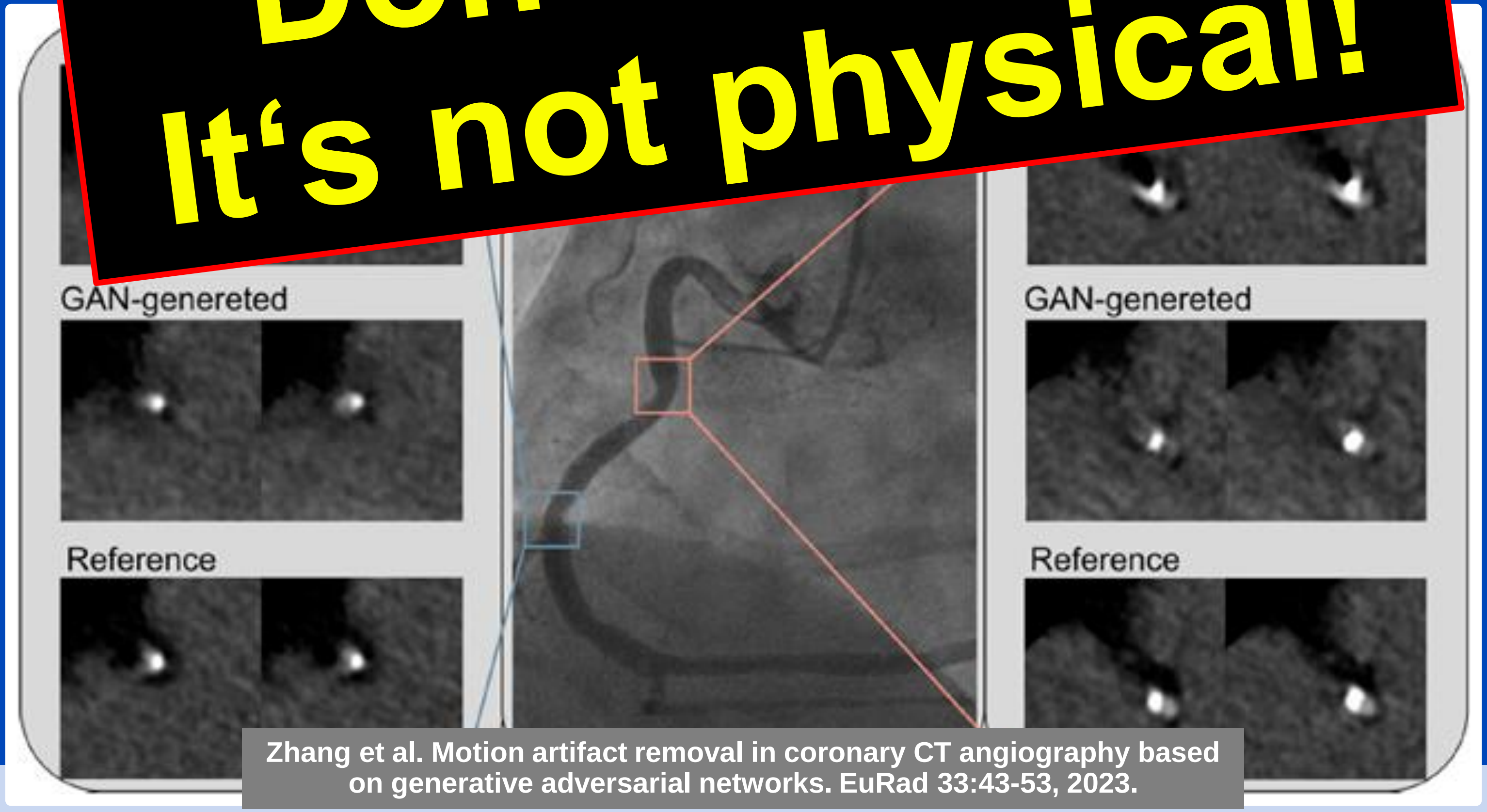


From [4]

Deep Cosmetic Motion Artifact Reduction

- Image-based correction
  = cosmetic correction
  = similar to pic beauty and others
- May not be the

Don‘t do that!
It‘s not physical!



Zhang et al. Motion artifact removal in coronary CT angiography based on generative adversarial networks. EuRad 33:43-53, 2023.

Denoising benchmark with surprising results

IS NEWER ALWAYS BETTER?

LDCT Benchmark

- **Algorithms used for our benchmark:**
  - CNN-10 (2017)
  - RED-CNN (2017)
  - ResNet (2018)
  - WGAN-VGG (2017)
  - QAE (2019)
  - DU-GAN (2021)
  - TransCT (2021)
  - Bilateral (2022)

Standard CNNs trained with pixelwise losses

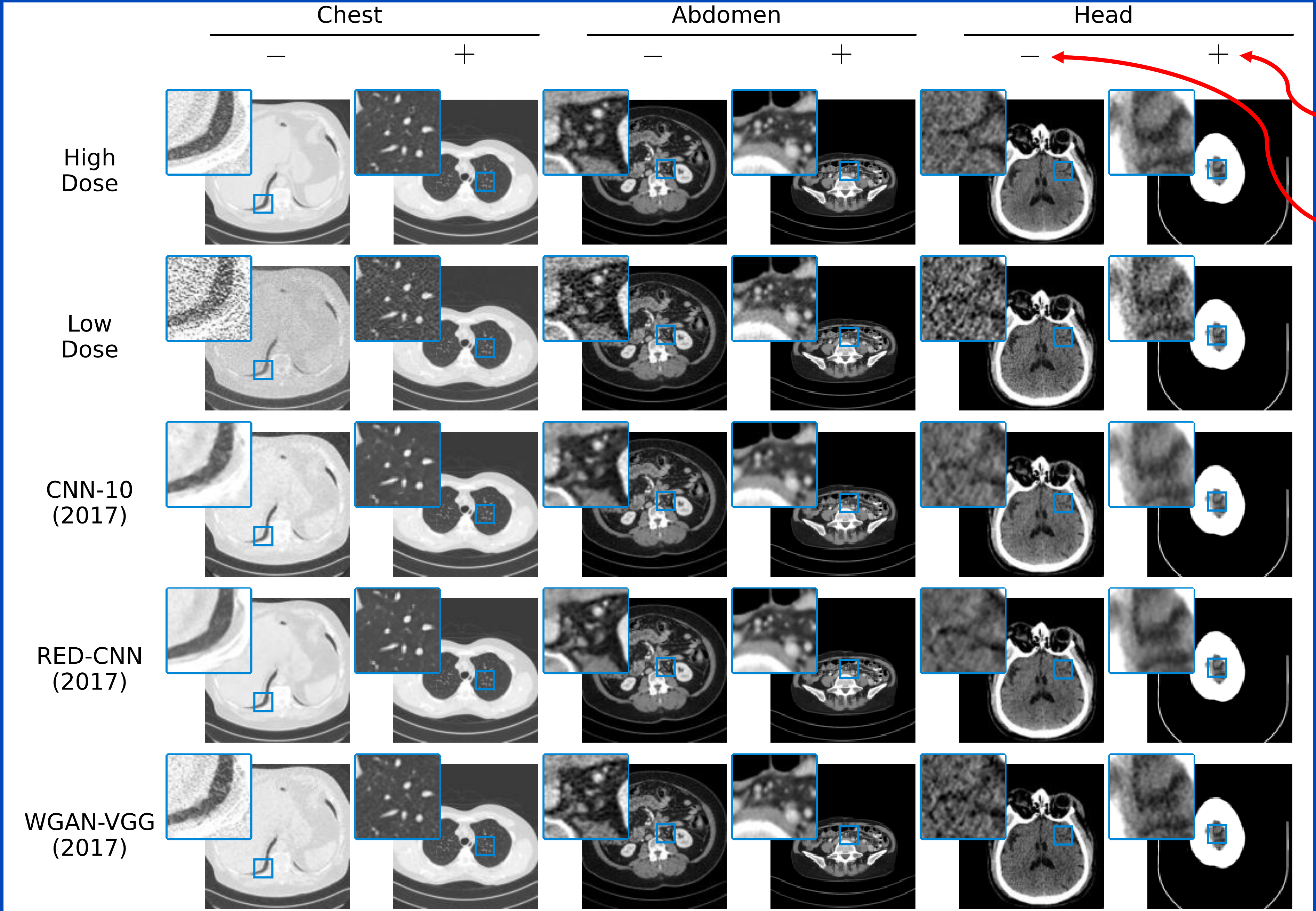
CNNs trained with adversarial losses

Specialized architectures trained with pixelwise losses

- **All tested methods**
  - do the same hyperparameter optimization
  - use the same train/validation set
  - were evaluated on the same test set

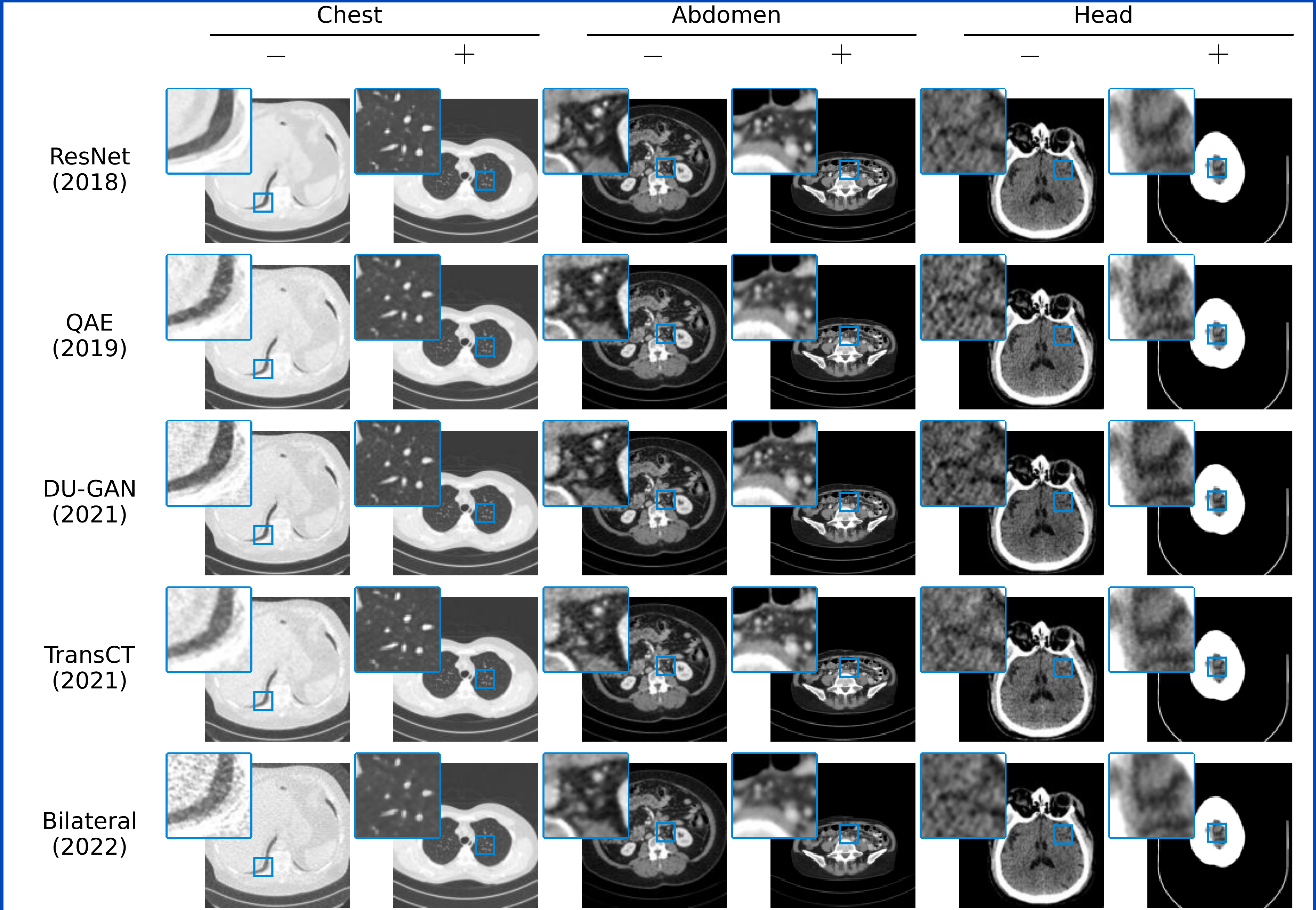
github.com/eeulig/ldct-benchmark

E. Eulig, B. Ommer, and M. Kachelrieß. Benchmarking deep learning-based low-dose CT image denoising algorithms. Med. Phys. 51(12):8776-8788, December 2024.



Slice where the average SSIM across all head slices and methods is **highest**.

Slice where the average SSIM across all head slices and methods is **lowest**.



E. Eulig, B. Ommer, and M. Kachelrieß. Benchmarking deep learning-based low-dose CT image denoising algorithms. Med. Phys. 51(12):8776-8788, December 2024.

Image Quality Metrics

Given reference image *x* and test image *y*, *N* pixels each.

Peak signal-to-noise ratio (PSNR)

$$\mathrm{RMSE}(x,y) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2}$$

$$\mathrm{PSNR}(x,y) = 20\log\left(\frac{I_{\max}}{\mathrm{RMSE}(x,y)}\right)$$

Visual information fidelity (VIF)

Compare information *I* extracted by a human visual system (HVS) model of the test image *y* with that of the reference image *x*.

$$\mathrm{VIF}(x,y) = \frac{\sum_k I(G_k * x)}{\sum_k I(G_k * y)}$$

G_k are Gaussians of different scale

Structural similarity index measure (SSIM)

$$\mathrm{SSIM}(u,v) = \frac{(2\mu_u\mu_v + C_1)(2\sigma_{uv} + C_2)}{(\mu_u^2 + \mu_v^2 + C_1)(\sigma_u^2 + \sigma_v^2 + C_2)}$$

$$\mathrm{SSIM}(x,y) = \operatorname*{mean}_{u\in x, v\in y} \mathrm{SSIM}(u,v)$$

$$\left.\begin{array}{llll} \mu_u & : \text{Mean of } u & \sigma_u^2 & : \text{Variance of } u \\ \mu_v & : \text{Mean of } v & \sigma_v^2 & : \text{Variance of } v \\ \sigma_{uv} & : \text{Covariance of } u \text{ and } v & & \end{array}\right\}$$

in a sliding 7×7 window

Radiomic feature similarity (RFS)

1. Extract radiomic features R_x and R_y from segmentations in *x* and *y*
2. Compute cosine similarity between R_x and R_y

$$\mathrm{RFS}(x,y) = \frac{R_x \cdot R_y}{\|R_x\|\,\|R_y\|}$$

Quantitative Results

PSNR units are decibel (dB)	Head (25% dose)				Chest (10% dose)				Abdomen (25% dose)			
	SSIM	PSNR	VIF	RFS	SSIM	PSNR	VIF	RFS	SSIM	PSNR	VIF	RFS
Low dose scan	26.40	0.55	0.71	0.34	18.77	0.09	0.70	0.84	28.67	0.34	0.75	0.88
CNN-10 (2017)	28.86	0.62	0.94	0.59	27.71	0.19	0.80	0.90	32.39	0.45	0.88	0.90
RED-CNN (2017)	30.41	0.69	0.95	0.61	28.36	0.22	0.76	0.90	33.22	0.49	0.80	0.90
WGAN-VGG (2017)	25.36	0.53	0.86	0.51	25.54	0.15	0.98	0.88	30.51	0.38	0.92	0.88
ResNet (2018)	29.64	0.67	0.91	0.61	28.42	0.22	0.75	0.90	33.15	0.49	0.79	0.90
QAE (2019)	28.51	0.59	0.95	0.58	27.62	0.19	0.83	0.89	32.02	0.42	0.96	0.90
DU-GAN (2021)	28.76	0.62	0.94	0.57	26.68	0.17	0.96	0.89	32.13	0.43	0.97	0.90
TransCT (2021)	24.65	0.44	0.88	0.56	26.99	0.17	0.83	0.88	30.53	0.37	0.92	0.85
Bilateral (2022)	26.60	0.50	0.87	0.55	25.59	0.16	0.64	0.86	27.13	0.36	0.87	0.87

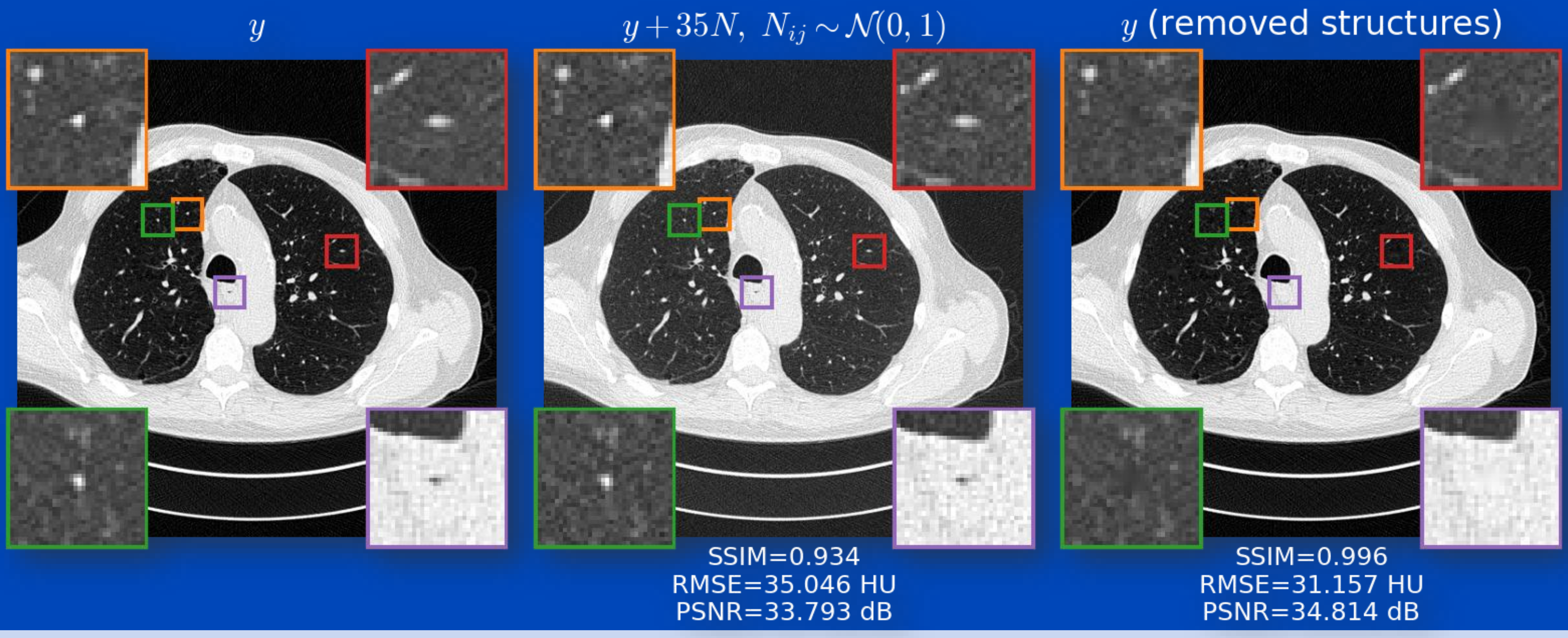
Green numbers indicate that a method is significantly better than the previously published best method.
Red numbers indicate that it is significantly worse.

Let small structures be just as important as large structures

A NEW METRIC FOR SUBTLE DETAILS

Attention: Each Pixel May be Significant!

- MAE, PSNR, RMSE and SSIM* are often used to quantify image quality, e.g. in loss functions or to rank algorithms.
- Alteration of a few pixels may mislead diagnosis.



*SSIM also accounts in parts for the human visual system by using luminance, contrast and structure to estimate perceptual quality.

Detecting Small Structures Using SAM[1,2]

Step 1: Segment patient via simple thresholding and finding largest contour

Step 2: Define a point grid over the previously found patient segmentation

Step 3: Generate masks using SAM and previously defined point prompts

a) Sort masks by their area

b) Starting with smallest mask:

- Remove masks with low stability score or low predicted IoU

$$\mathrm{Stab}(l, \theta_0, \theta_1) = \frac{|l > \theta_1|}{|l > \theta_0|}, \theta_0 < \theta_1 \qquad \mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

l : Logits predicted by network

- Remove intersections with any previous masks
- Only add mask if it is fully within the patient

Mask generated for single point prompt

[1]Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, et al. "Segment Anything." arXiv, 2023.
[2]Ma, Jun, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment Anything in Medical Images." *Nature Communications* 15 (1): 654, 2024.

dkfz.

Detecting Small Structures Using SAM[1,2]

Step 1: Segment patient via simple thresholding and finding largest contour

Step 2: Define a point grid over the previously found patient segmentation

Step 3: Generate masks using SAM and previously defined point prompts

a) Sort masks by their area

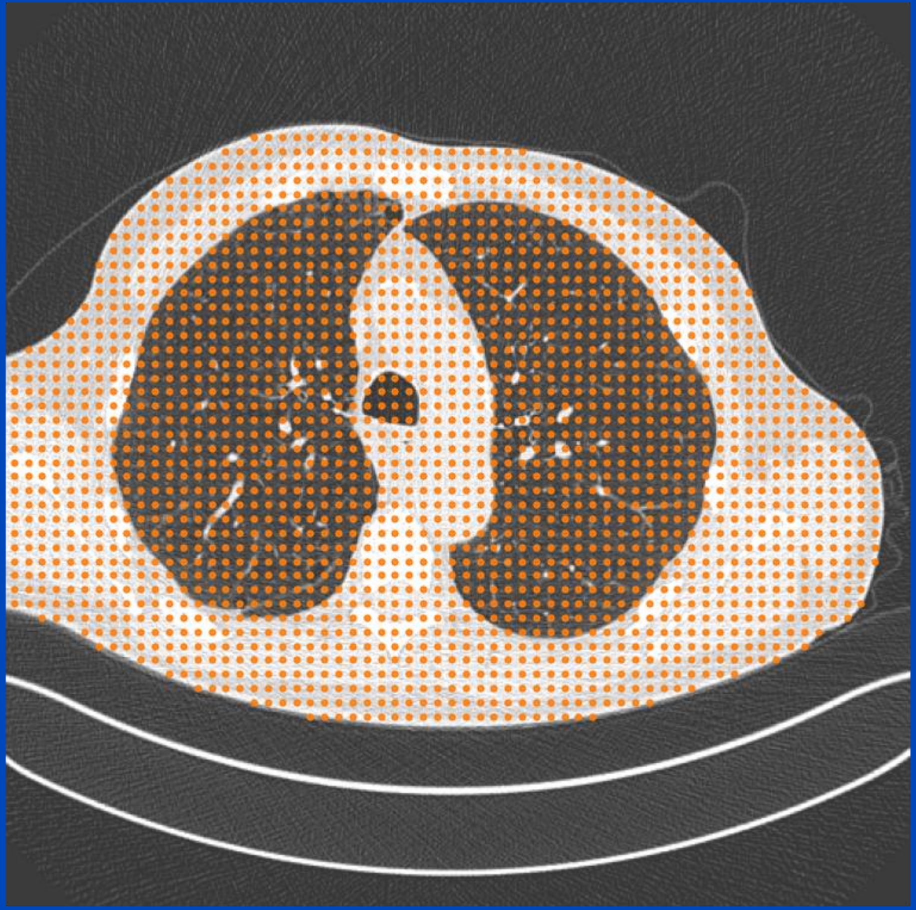
b) Starting with smallest mask:

- Remove masks with low stability score or low predicted IoU

$$\mathrm{Stab}(l, \theta_0, \theta_1) = \frac{|l > \theta_1|}{|l > \theta_0|}, \theta_0 < \theta_1 \qquad \mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

l : Logits predicted by network

- Remove intersections with any previous masks
- Only add mask if it is fully within the patient



Mask generated for single point prompt

[1]Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, et al. "Segment Anything." arXiv, 2023.
[2]Ma, Jun, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment Anything in Medical Images." *Nature Communications* 15 (1): 654, 2024.

dkfz.

Detecting Small Structures Using SAM[1,2]

Step 1: Segment patient via simple thresholding and finding largest contour

Step 2: Define a point grid over the previously found patient segmentation

Step 3: Generate masks using SAM and previously defined point prompts

a) Sort masks by their area

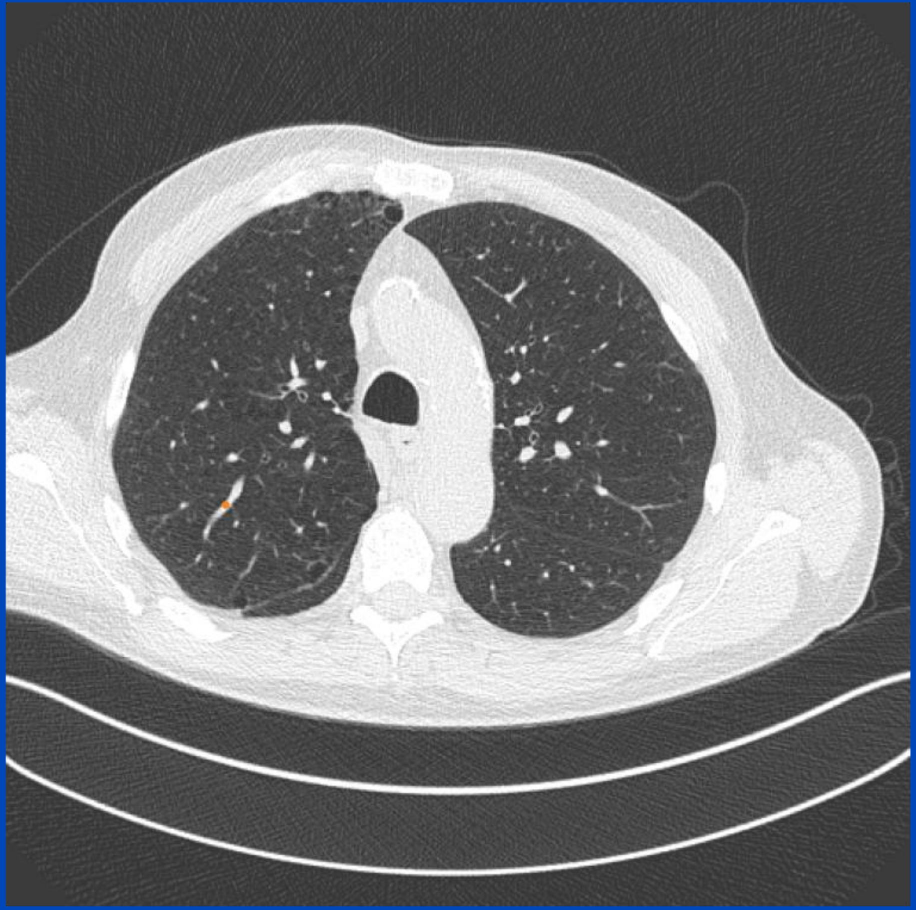
b) Starting with smallest mask:

- Remove masks with low stability score or low predicted IoU

$$\mathrm{Stab}(l, \theta_0, \theta_1) = \frac{|l > \theta_1|}{|l > \theta_0|}, \theta_0 < \theta_1 \qquad \mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

l : Logits predicted by network

- Remove intersections with any previous masks
- Only add mask if it is fully within the patient



Mask generated for single point prompt

[1]Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, et al. "Segment Anything." arXiv, 2023.
[2]Ma, Jun, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment Anything in Medical Images." *Nature Communications* 15 (1): 654, 2024.

Detecting Small Structures Using SAM[1,2]

Step 1: Segment patient via simple thresholding and finding largest contour

Step 2: Define a point grid over the previously found patient segmentation

Step 3: Generate masks using SAM and previously defined point prompts

a) Sort masks by their area

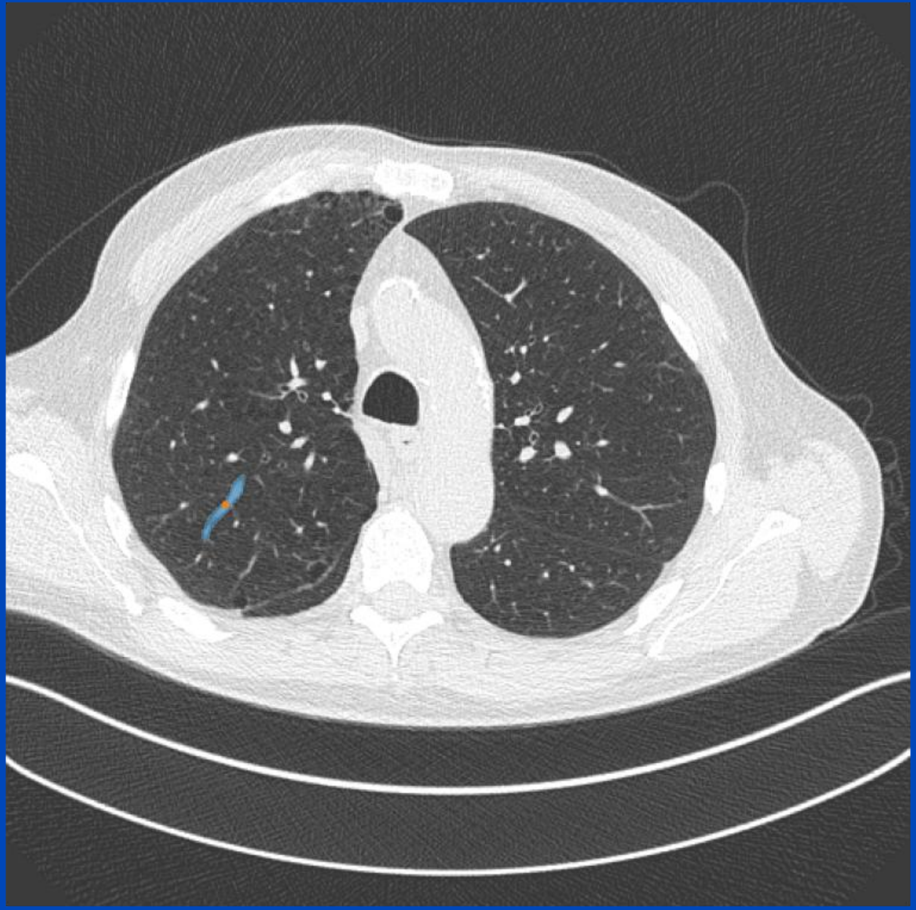
b) Starting with smallest mask:

- Remove masks with low stability score or low predicted IoU

$$\mathrm{Stab}(l, \theta_0, \theta_1) = \frac{|l > \theta_1|}{|l > \theta_0|}, \theta_0 < \theta_1 \qquad \mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

l : Logits predicted by network

- Remove intersections with any previous masks
- Only add mask if it is fully within the patient



Mask generated for single point prompt

[1]Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, et al. "Segment Anything." arXiv, 2023.
[2]Ma, Jun, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment Anything in Medical Images." *Nature Communications* 15 (1): 654, 2024.

Detecting Small Structures Using SAM[1,2]

Step 1: Segment patient via simple thresholding and finding largest contour

Step 2: Define a point grid over the previously found patient segmentation

Step 3: Generate masks using SAM and previously defined point prompts

a) Sort masks by their area

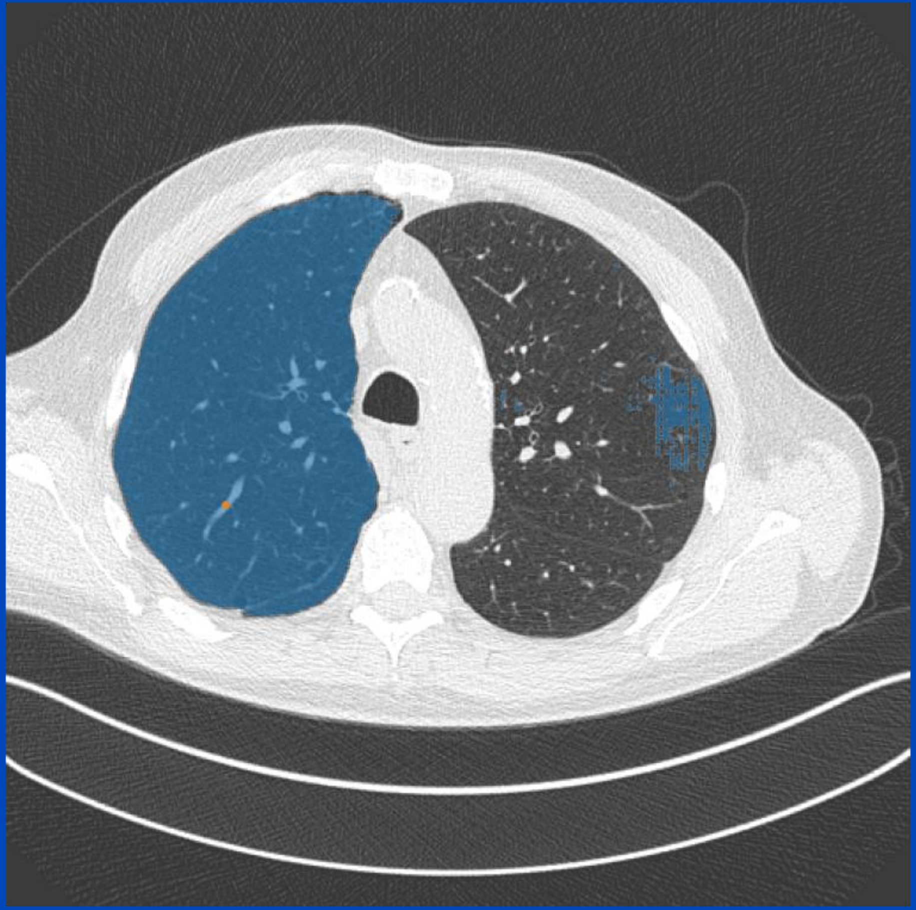
b) Starting with smallest mask:

- Remove masks with low stability score or low predicted IoU

$$\mathrm{Stab}(l, \theta_0, \theta_1) = \frac{|l > \theta_1|}{|l > \theta_0|}, \theta_0 < \theta_1 \qquad \mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

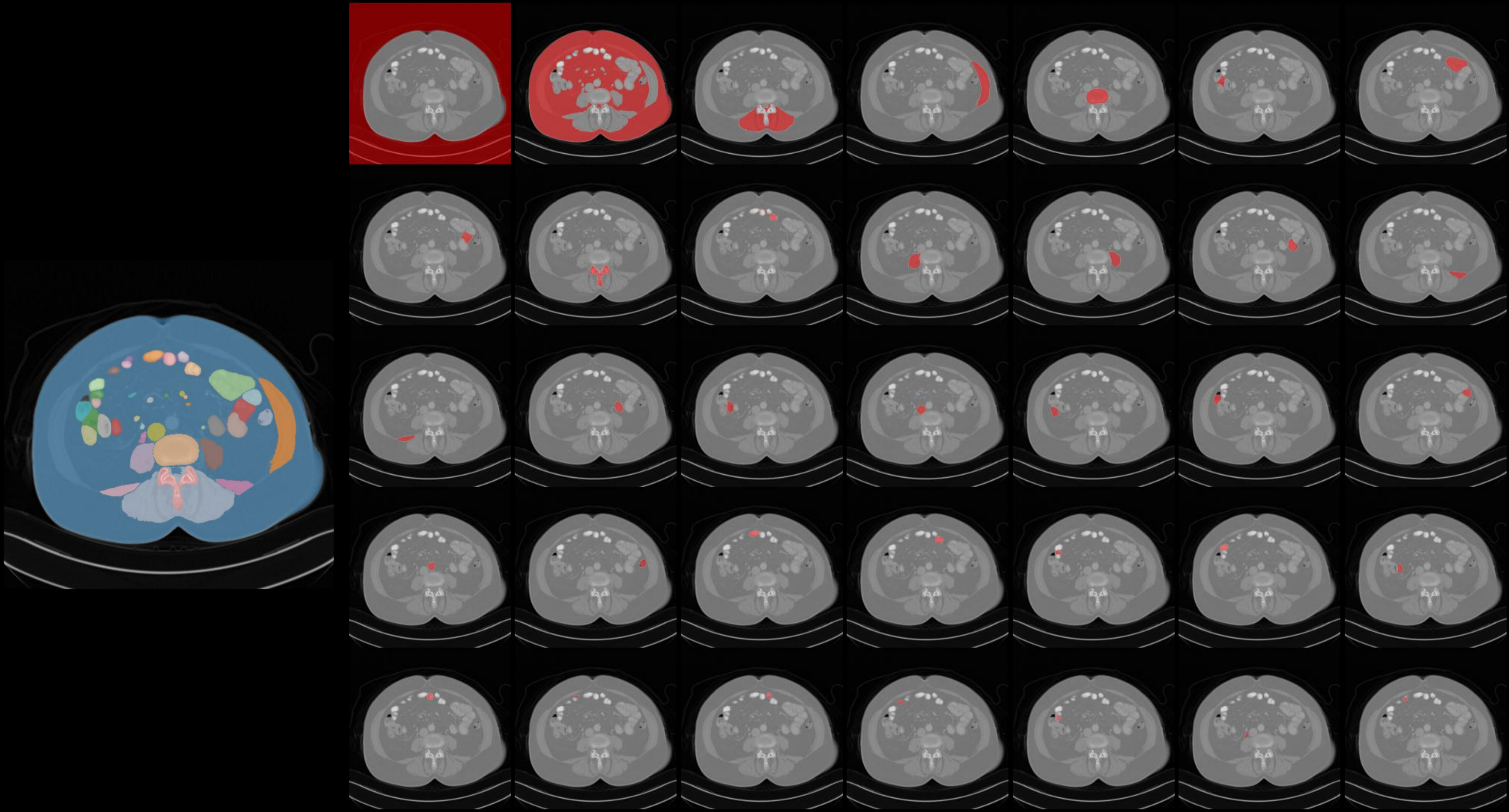
l : Logits predicted by network

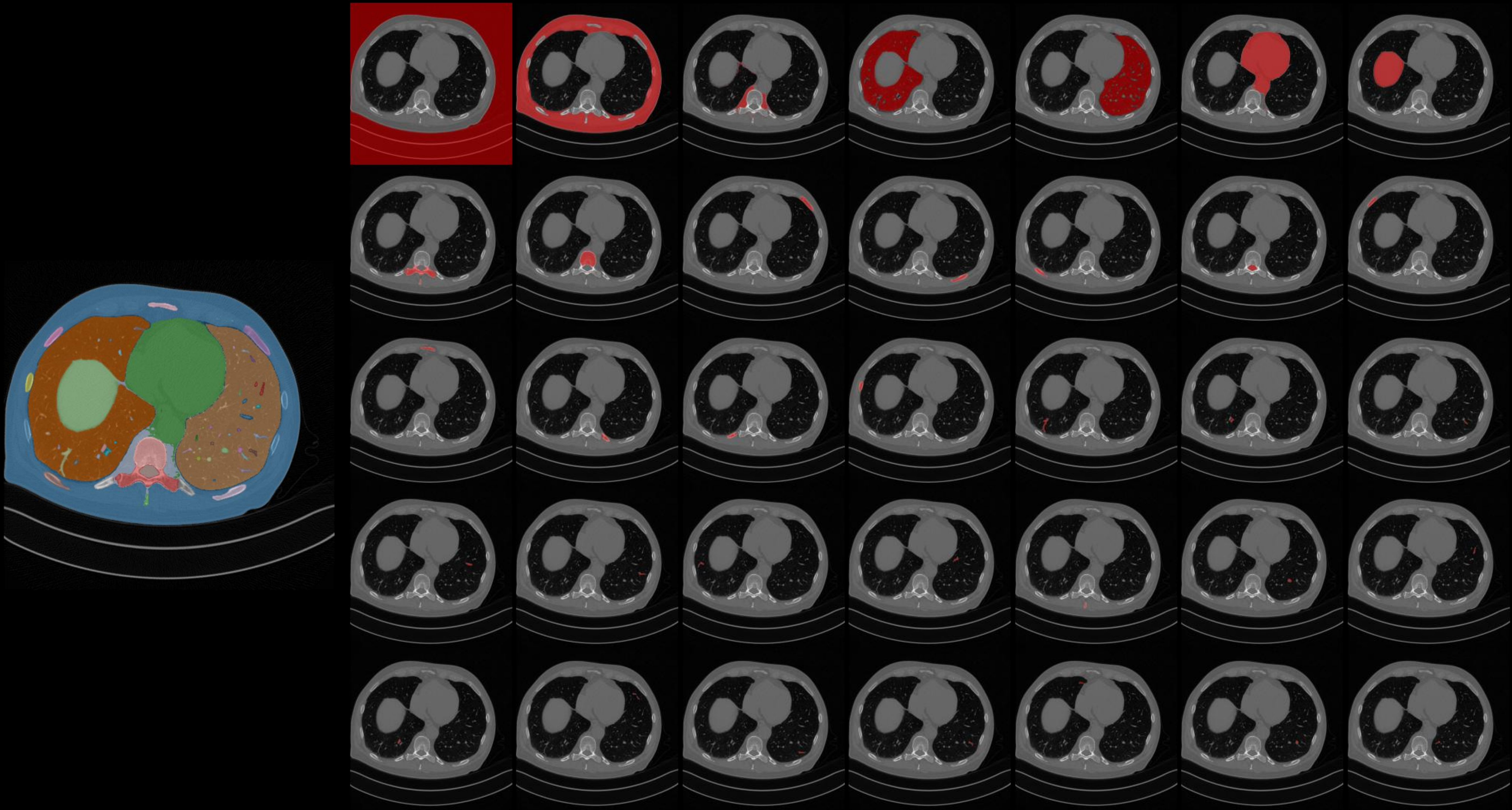
- Remove intersections with any previous masks
- Only add mask if it is fully within the patient



Mask generated for single point prompt

[1]Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, et al. "Segment Anything." arXiv, 2023.
[2]Ma, Jun, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment Anything in Medical Images." *Nature Communications* 15 (1): 654, 2024.





Methods

Segment RMSE (SRMSE)

Given a set of SAM-segmented masks $\mathcal{M} = \{\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \ldots, \boldsymbol{m}^{(M)}\}$, where each mask $\mathrm{m}^{(i)} \in \{0,1\}^N$ with $N = H \times W$, define with SRMSE the mask-wise root mean square error (RMSE) for two images $\boldsymbol{x}, \boldsymbol{y}$ and mask $\boldsymbol{m}$

$$\mathrm{SRMSE}(x, y; m) = \sqrt{\frac{\sum_{i=1}^{N} m_i (x_i - y_i)^2}{\sum_{i=1}^{N} m_i}}$$

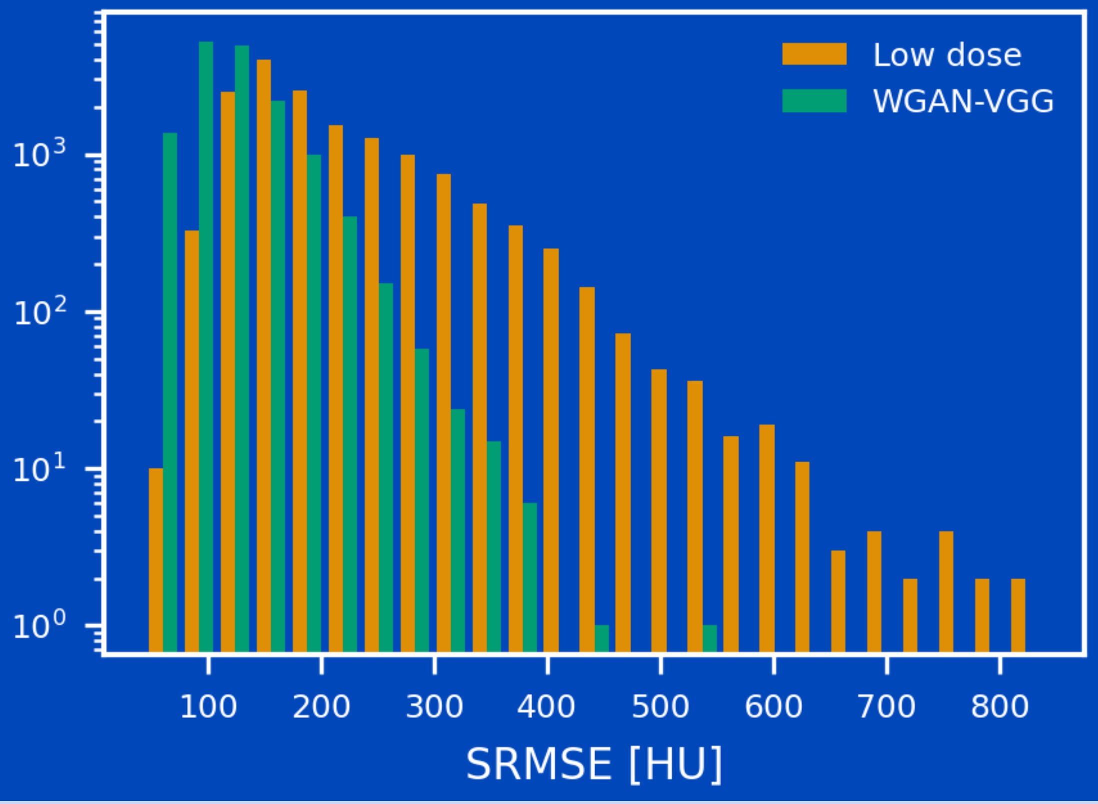
Using the set of all SRMSEs $\left\{\mathrm{SRMSE}\left(x, y; m^{(i)}\right)\right\}_{i=1}^{M}$, define the

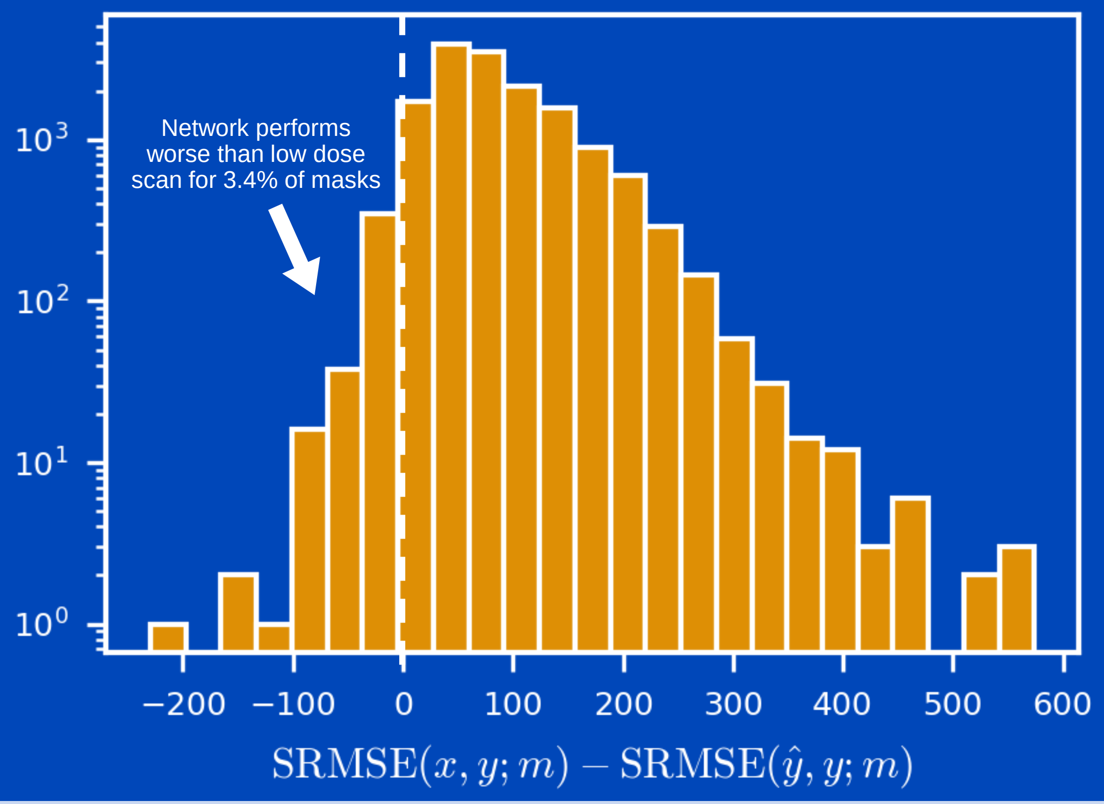
$$\text{Mean-SRMSE}(x, y) = \frac{1}{M} \sum_{i=1}^{M} \mathrm{SRMSE}(x, y; m^{(i)})$$

$$\text{Max-SRMSE}(x, y) = \max \left\{\mathrm{SRMSE}(x, y; m^{(i)})\right\}_{i=1}^{M}$$

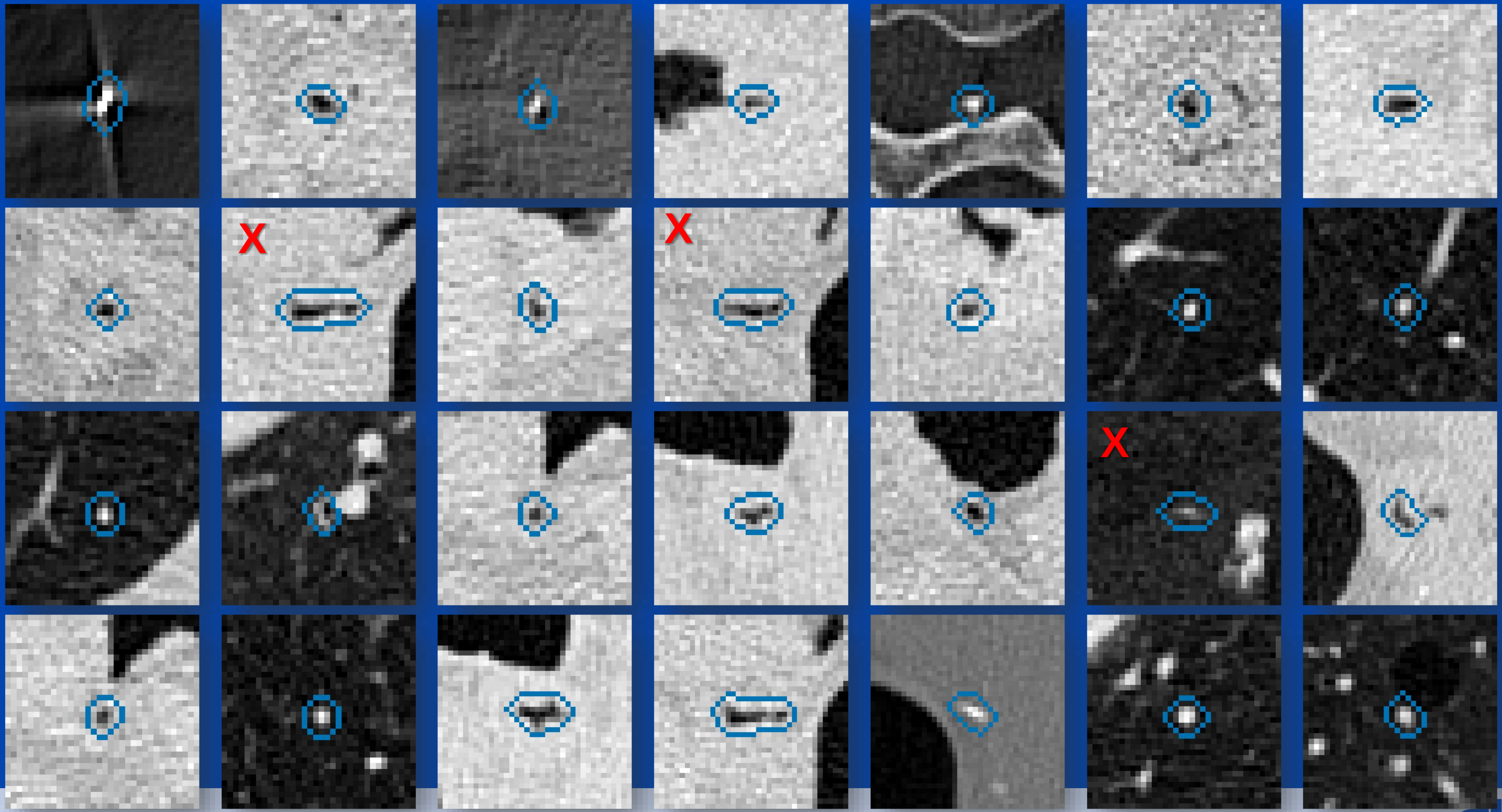
Detecting Hallucinations

- Compare SRMSE of low dose scan (x) with network prediction ($\hat{y}$).
- On a chest scan with 392 axial slices we have a total of 15,547 masks.

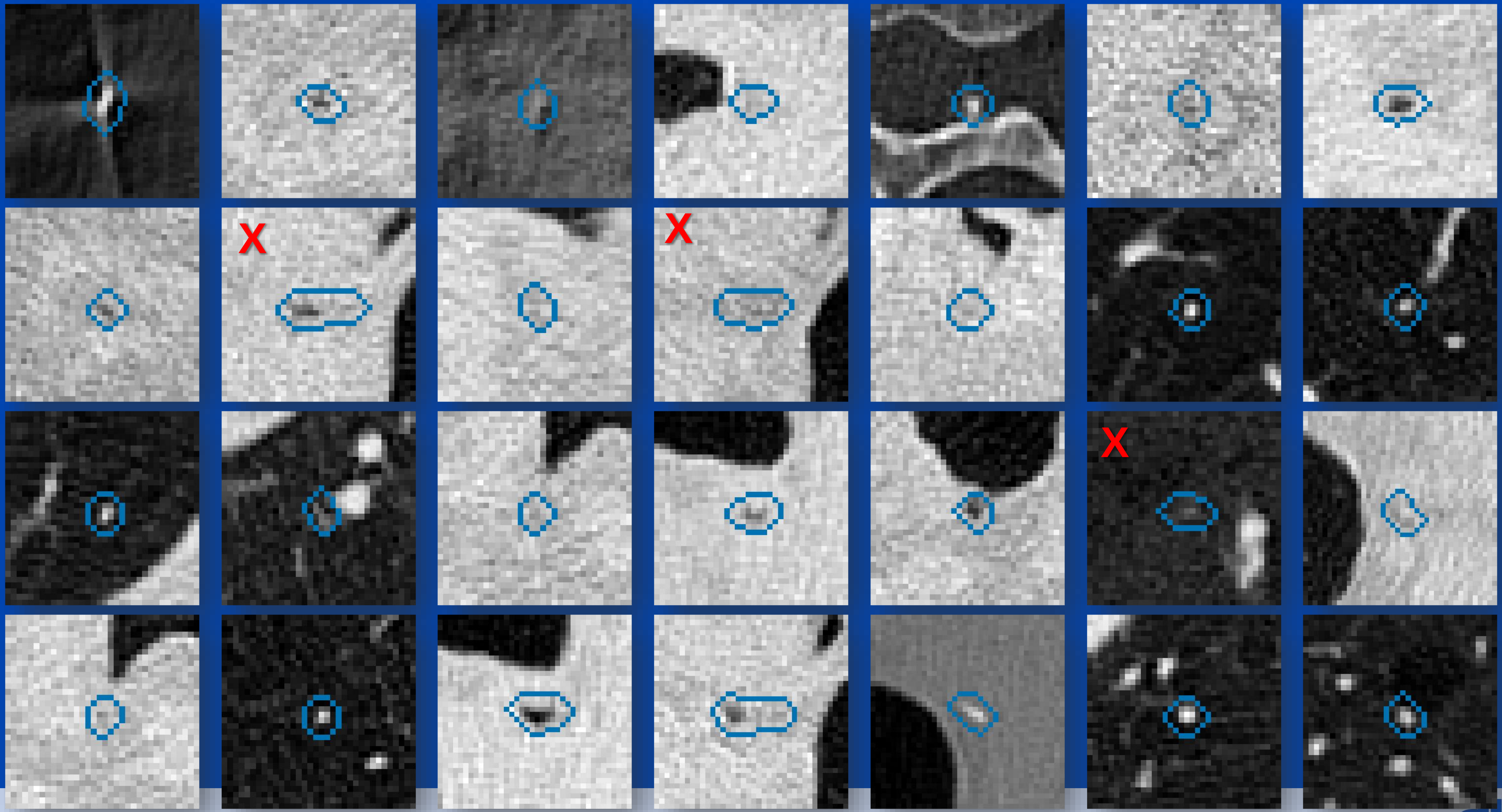




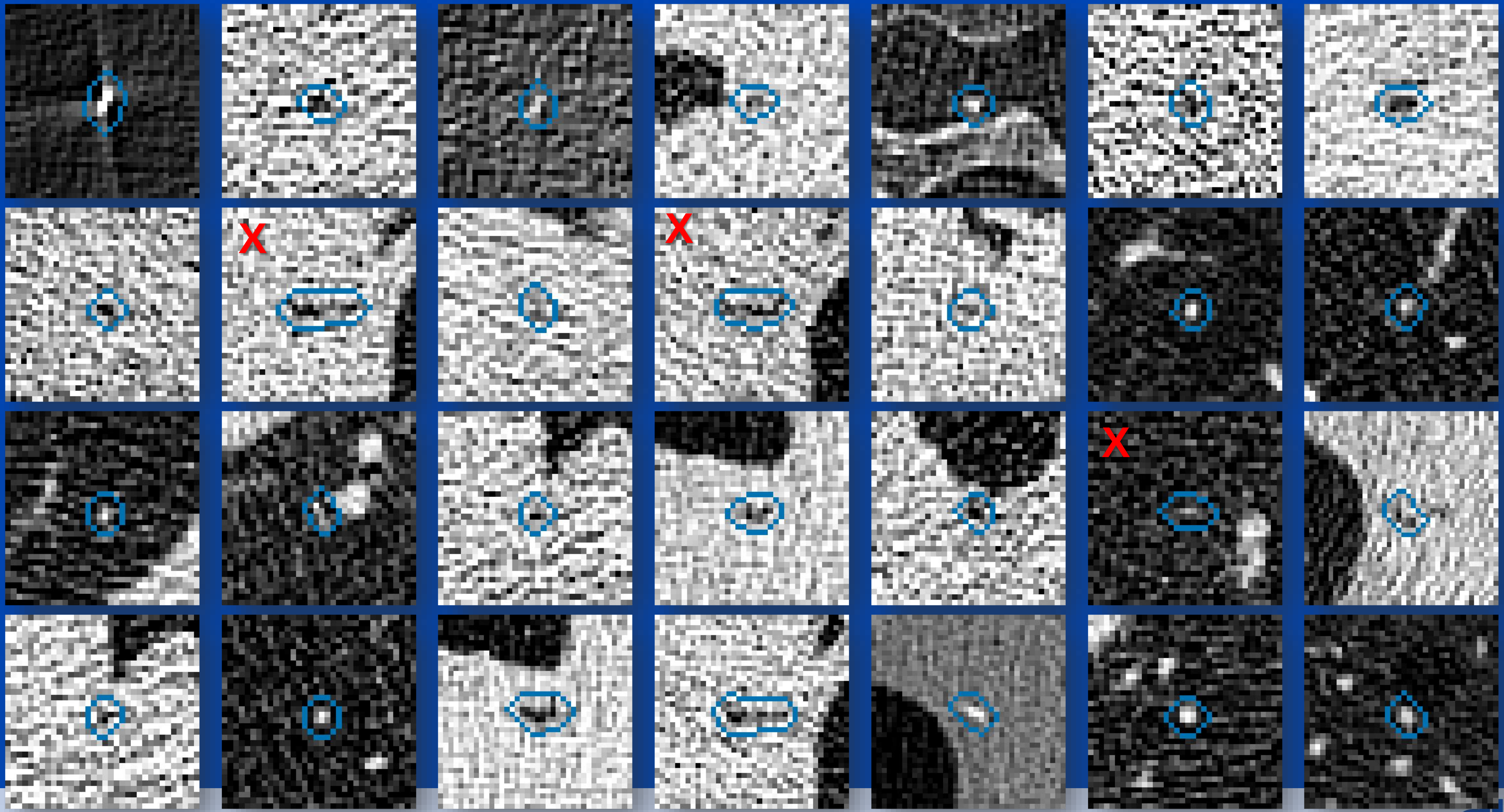
High Dose Images



Network Predictions (WGAN-VGG)

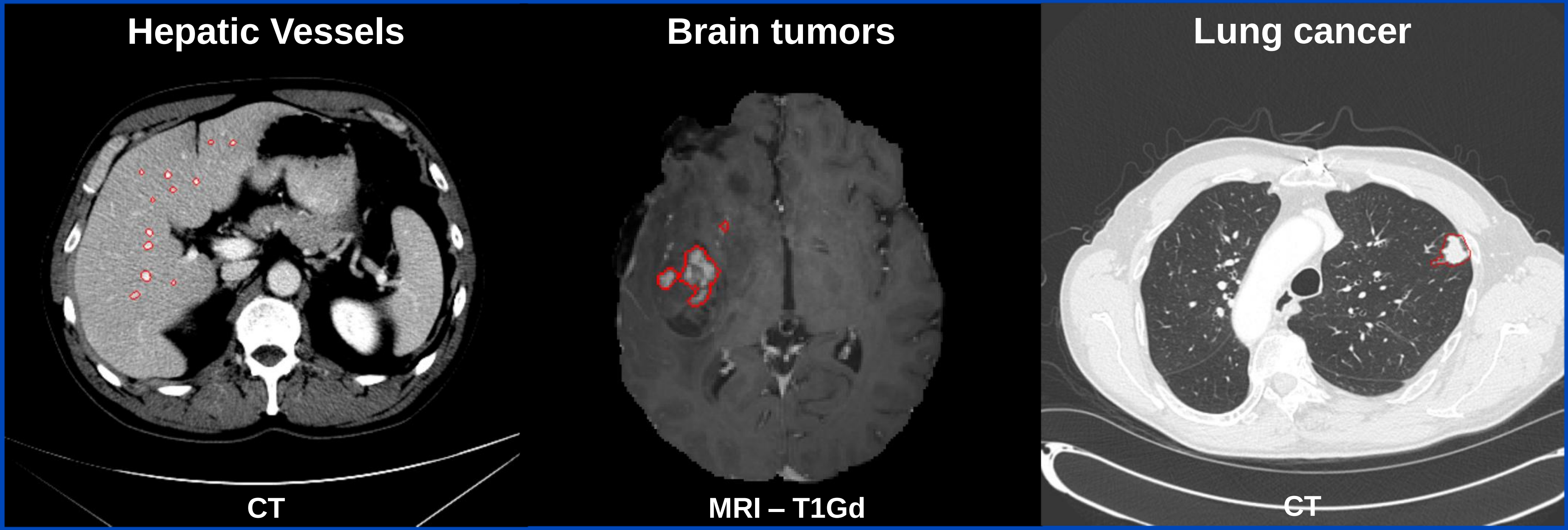


Low Dose Images



Experiments
Evaluation

- **Evaluate the proposed metric on synthetic datasets where the amount of removed structures is known**
- **Utilize three datasets from the *Medical Decathlon*[1], a collection of ten medical image segmentation tasks with ground truth annotations**

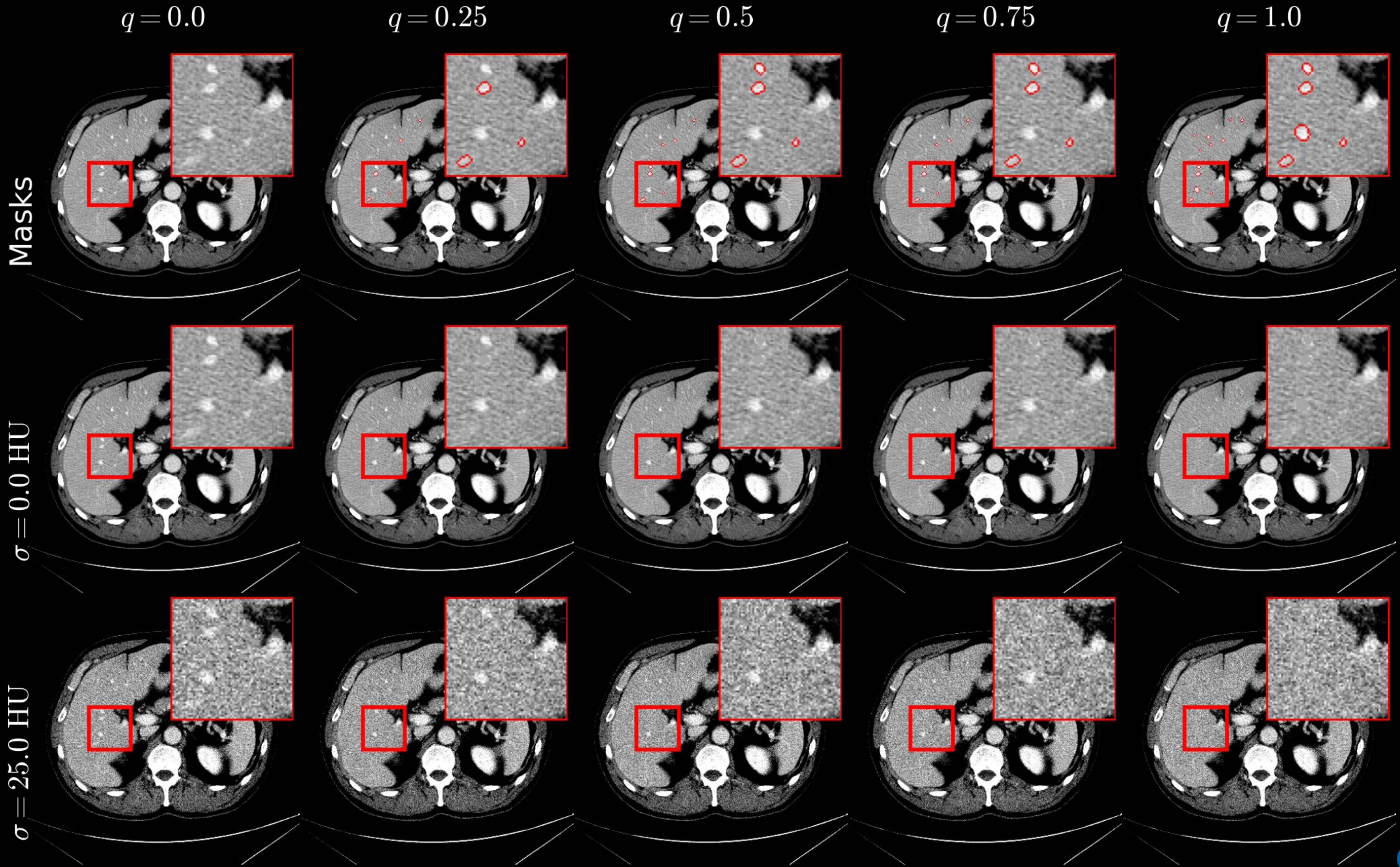


[1]Simpson, Amber L., Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, et al. 2019. “A Large Annotated Medical Image Dataset for the Development and Evaluation of Segmentation Algorithms.” arXiv.

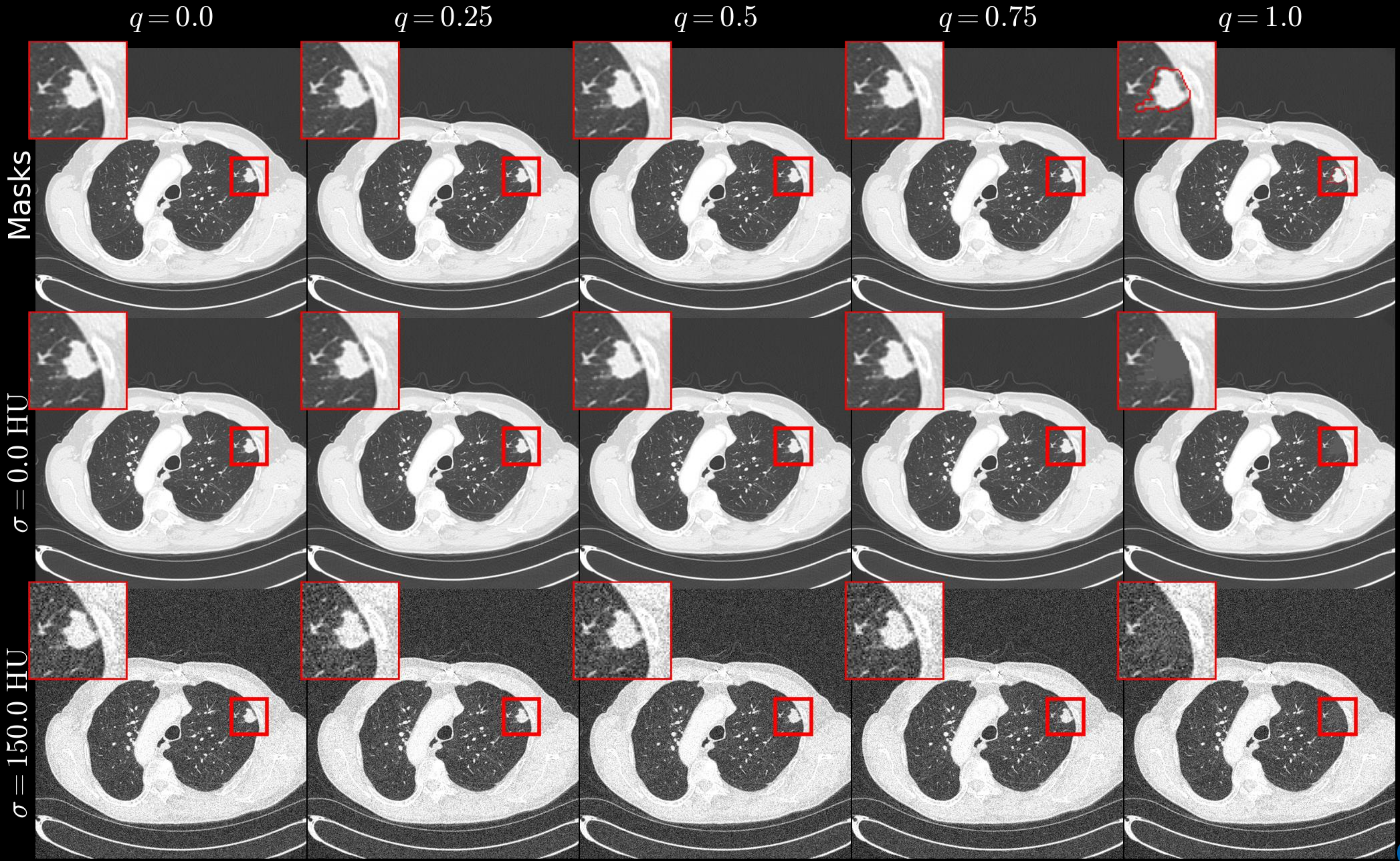
Evaluation

- For each scan in a dataset we can randomly remove fractions *q* of the ground truth (manually segmented) structures by means of inpainting.
- Fraction *q* refers to the whole patient and not just to a single slice!
- Here we simply replace pixels with
  - Hepatic vessel: 130 HU
  - Lung: -800 HU
  - Brain tumor: median pixel value
- Add Gaussian noise with various standard deviations
- Then evaluate how well different metrics
  a) can rank images with different q
  b) can detect that an algorithm removed very few, e.g. $q \ll 1\%$, structures

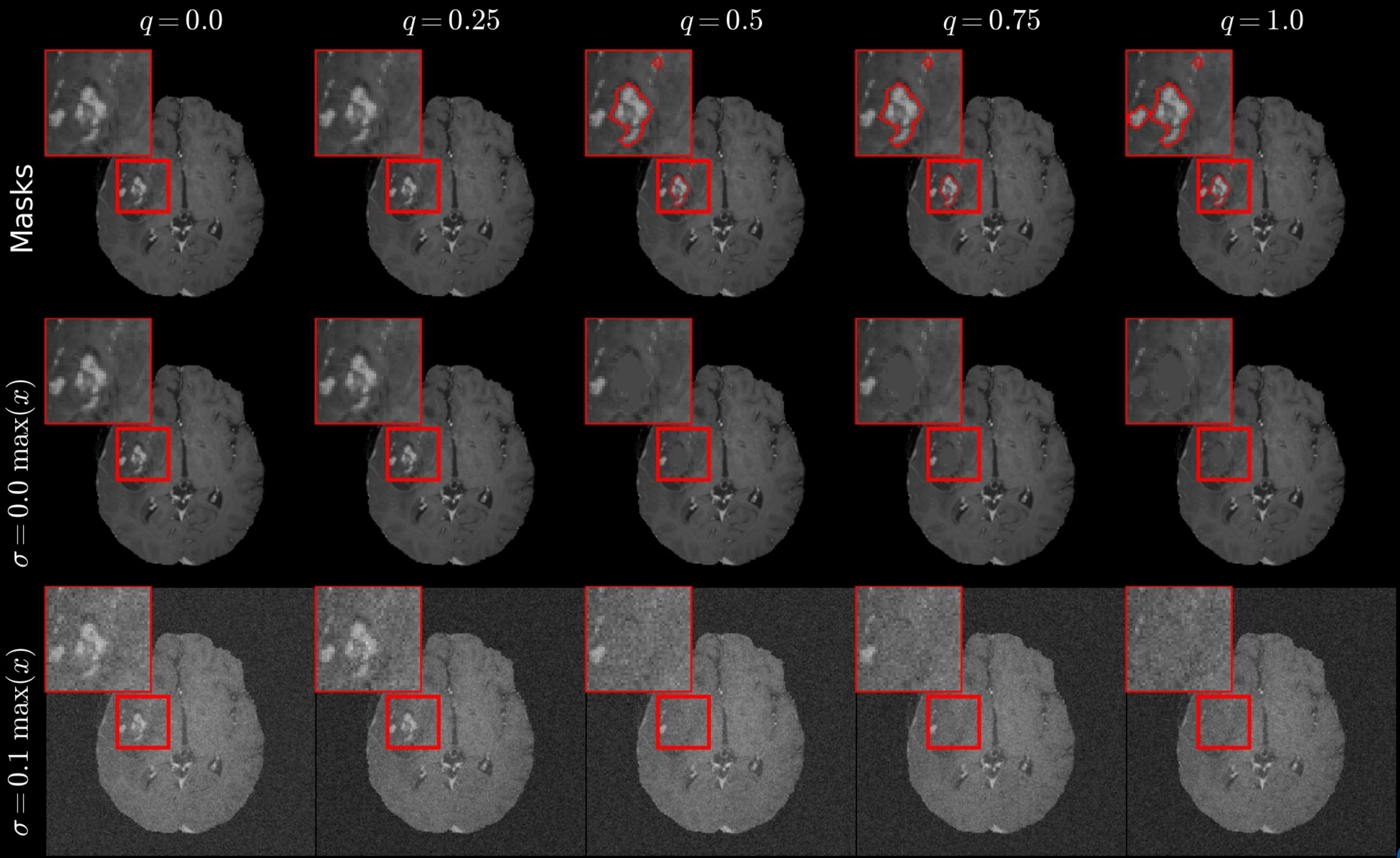
Hepatic Vessels



Lung cancer

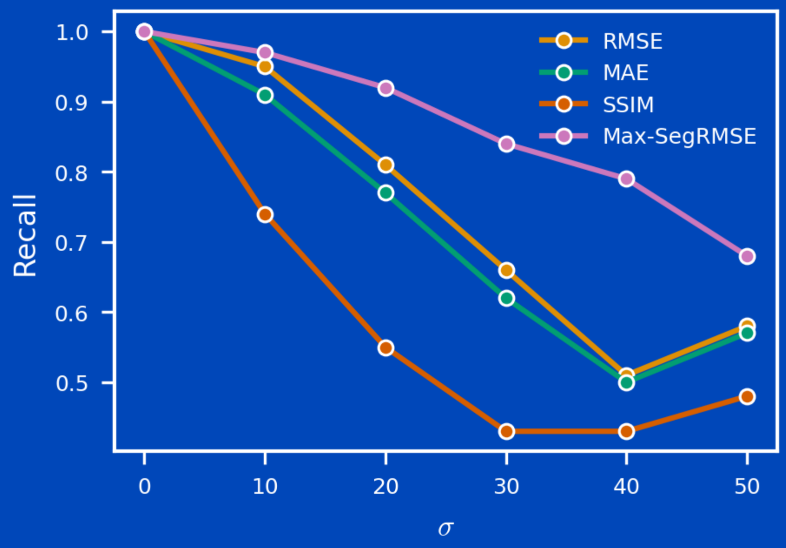


Brain tumor

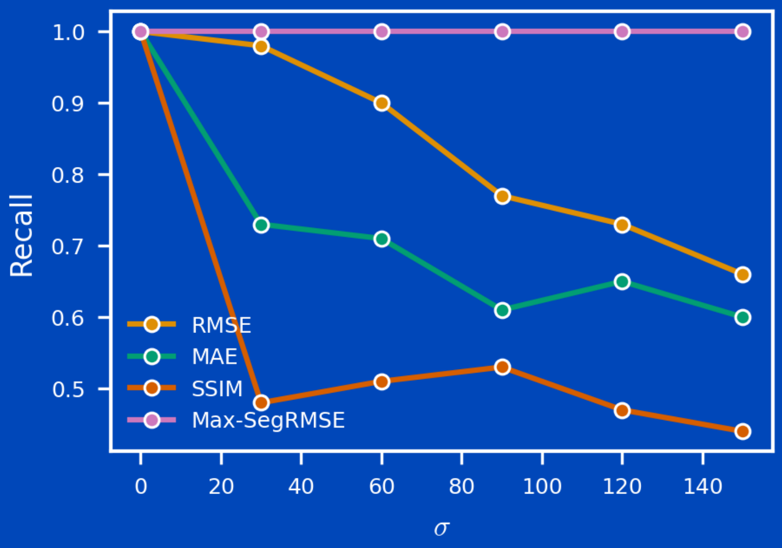


Results: True Positive Fraction

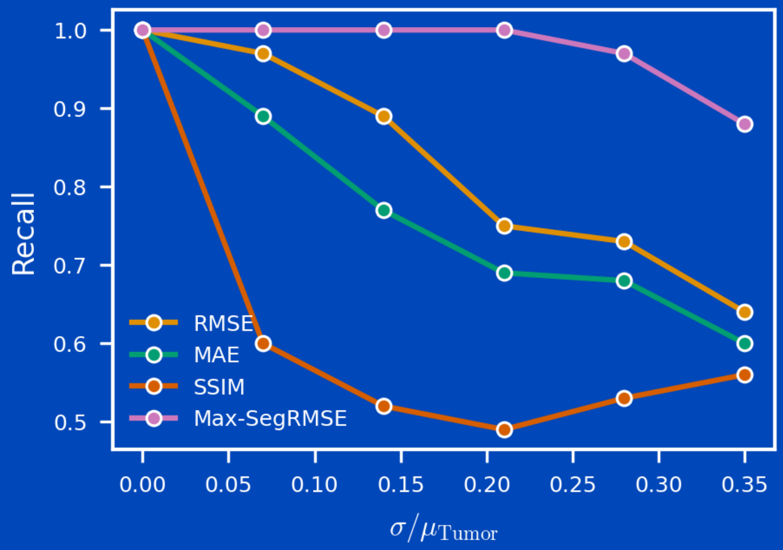
Hepatic Vessels



Lung Cancer



Brain Tumor



The plots are for q = 0.1, i.e. for about 0.007% to 0.03% modified voxels.

Summary

- New metrics are needed to quantify changes in subtle details.
- Needed to evaluate the quality of AI-based algorithms.
- Could become part of the loss function to train networks.
- May help to determine the amount of dose reduction possible for a given algorithm.

Thank You!

- This presentation will soon be available at www.dkfz.de/ct.
- Job opportunities through marc.kachelriess@dkfz.de or through DKFZ's PhD program.
- Parts of the reconstruction software were provided by RayConStruct® GmbH, Nürnberg, Germany.

Low dose CT benchmark:

github.com/eeulig/ldct-benchmark

E. Eulig, B. Ommer, and M. Kachelrieß. Benchmarking deep learning-based low-dose CT image denoising algorithms. Med. Phys. 51(12):8776-8788, December 2024.