Structure-Aware Metrics for the Evaluation of Deep Learning-Based Image Reconstruction Algorithms

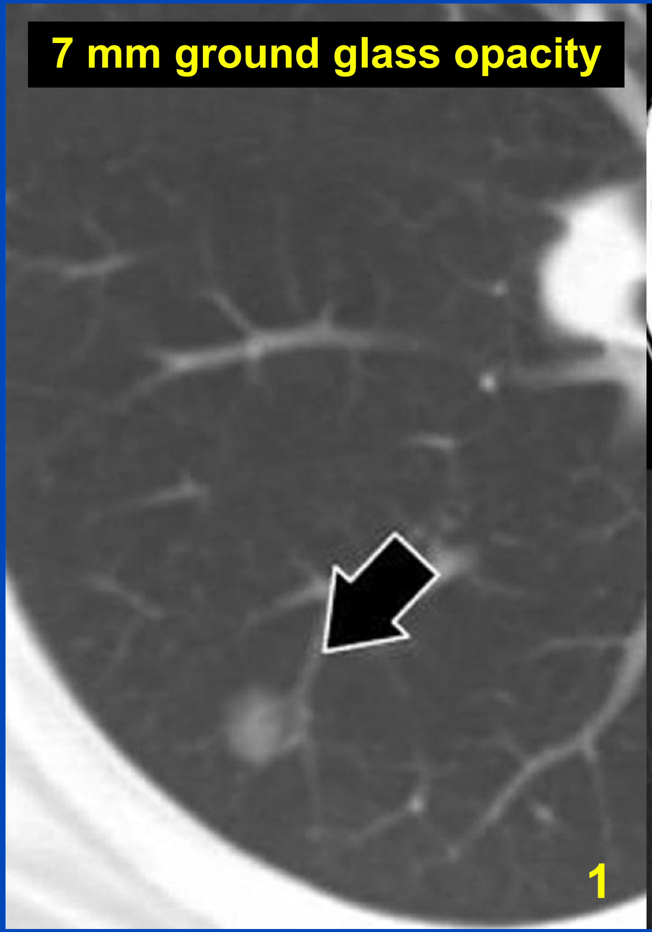
Elias Eulig[1,2], Joscha Maier[1], and Marc Kachelrieß[1,2]

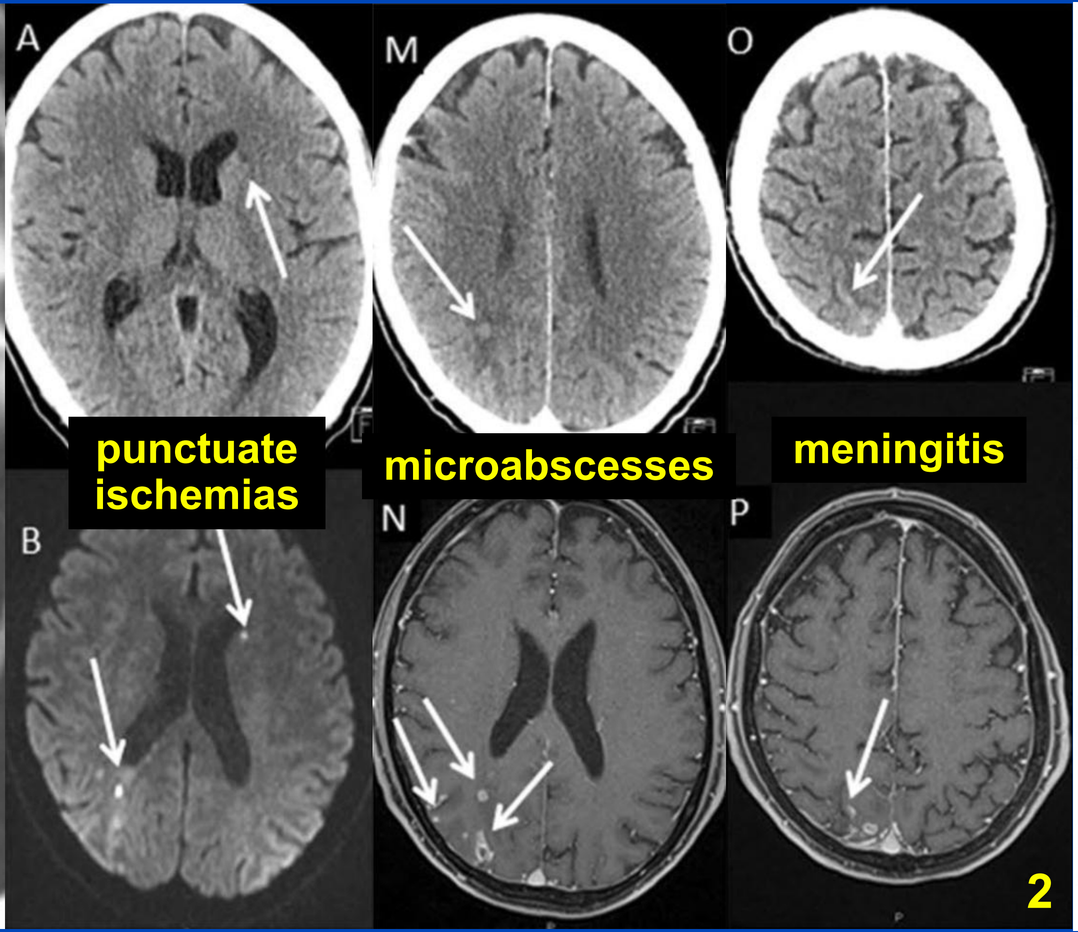
[1]German Cancer Research Center (DKFZ), Germany
[2]Heidelberg University, Germany

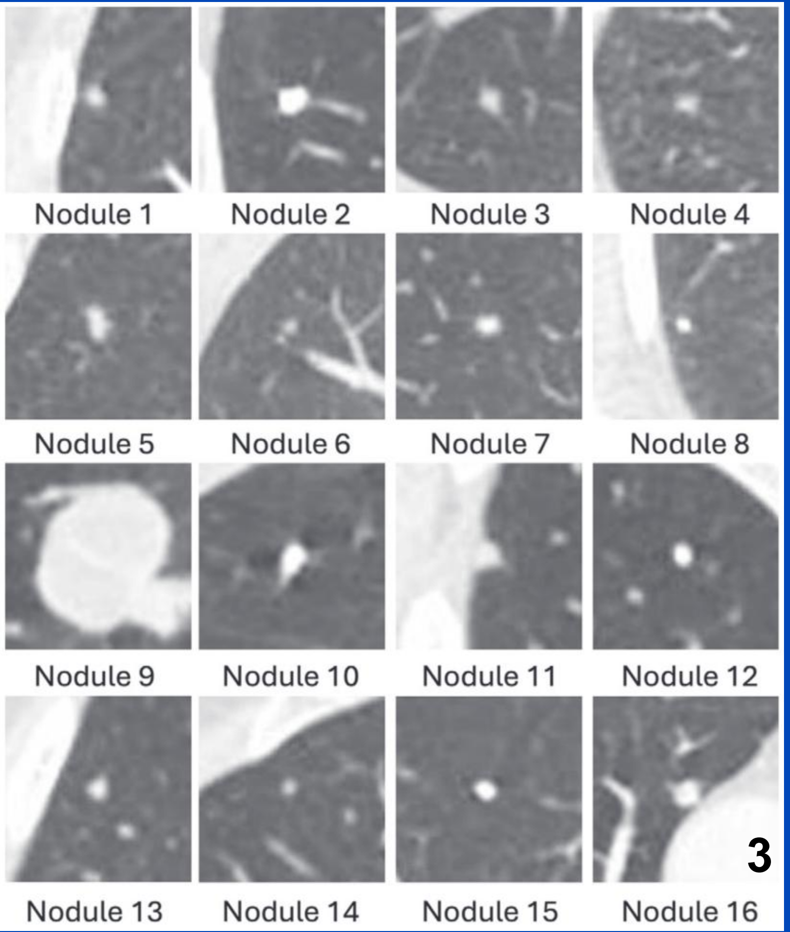
DEUTSCHES
KREBSFORSCHUNGSZENTRUM
IN DER HELMHOLTZ-GEMEINSCHAFT

Motivation

In medical imaging (CT, MRI) pathological features are often present as small, potentially low-contrast structures





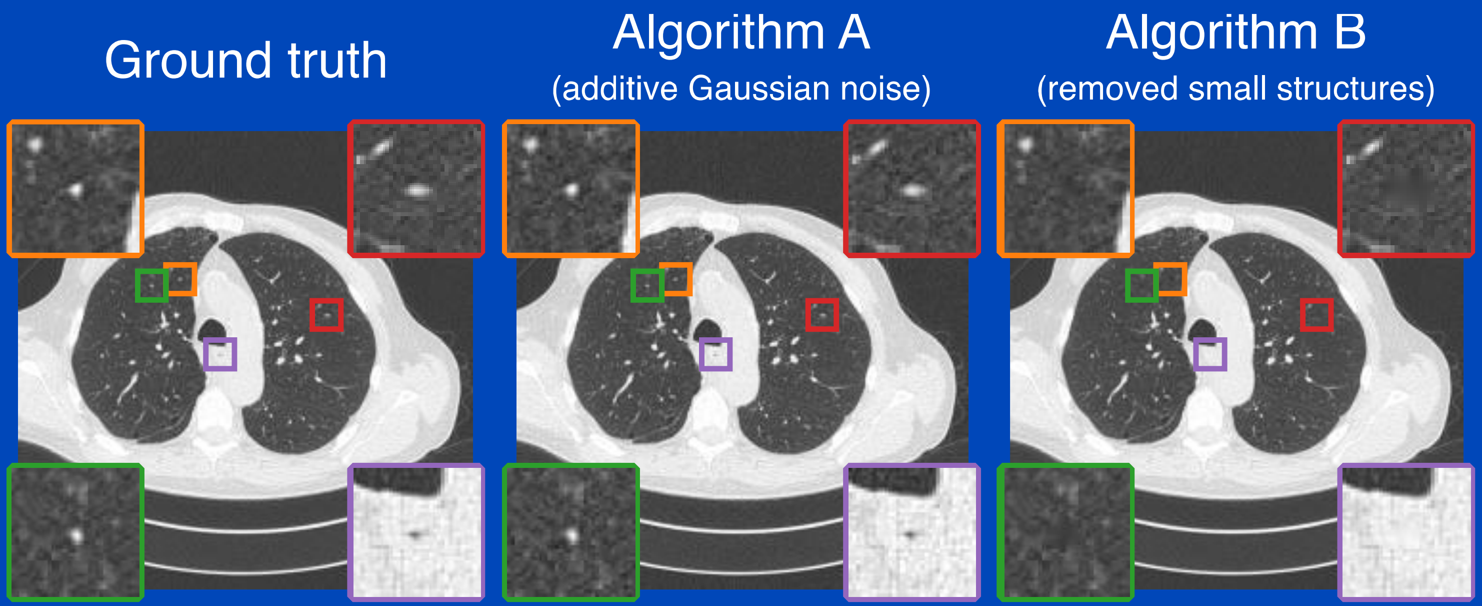


[1]H. K. Kim *et al.*, "Management of Multiple Pure Ground-Glass Opacity Lesions in Patients with Bronchioloalveolar Carcinoma," *Journal of Thoracic Oncology*, vol. 5, no. 2, 2010.
[2]P. Vitali *et al.*, "MRI versus CT in the detection of brain lesions in patients with infective endocarditis before or after cardiac surgery," *Neuroradiology*, vol. 64, no. 5, 2022.
[3]G. J. DiGirolamo *et al.*, "Non-conscious Detection of 'Missed' Lung Nodules by Radiologists: Expanding the Boundaries of Successful Processing during the Visual Assessment of Chest CT Scans," *Radiology*, vol. 314, no. 2, 2025.

Motivation

Common IQA metrics weigh pixels of differently sized structures equally

→ These metrics are not sensitive to the removal of small structures



	Algorithm A	Algorithm B
RMSE (↓)	35.028 HU	31.157 HU
SSIM (↑)	0.936	0.996
PSNR (↑)	33.797 dB	34.814 dB

Aim

Develop a novel metric that explicitly penalizes the removal of small structures

Methods

General setup

Denote with $x \in \mathbb{R}^{H \times W}$ an image reconstructed using some algorithm and $y \in \mathbb{R}^{H \times W}$ the corresponding (aligned) ground truth image

1. Generate a set $\mathcal{S}$ of binary segmentations $s \in \{0,1\}^{H \times W}$, each representing a structure present in y
2. For each of the segments $s \in \mathcal{S}$ compute a traditional image metric between x and y
3. Aggregate the per-segment metrics to compute a single metric for the entire image x

Methods

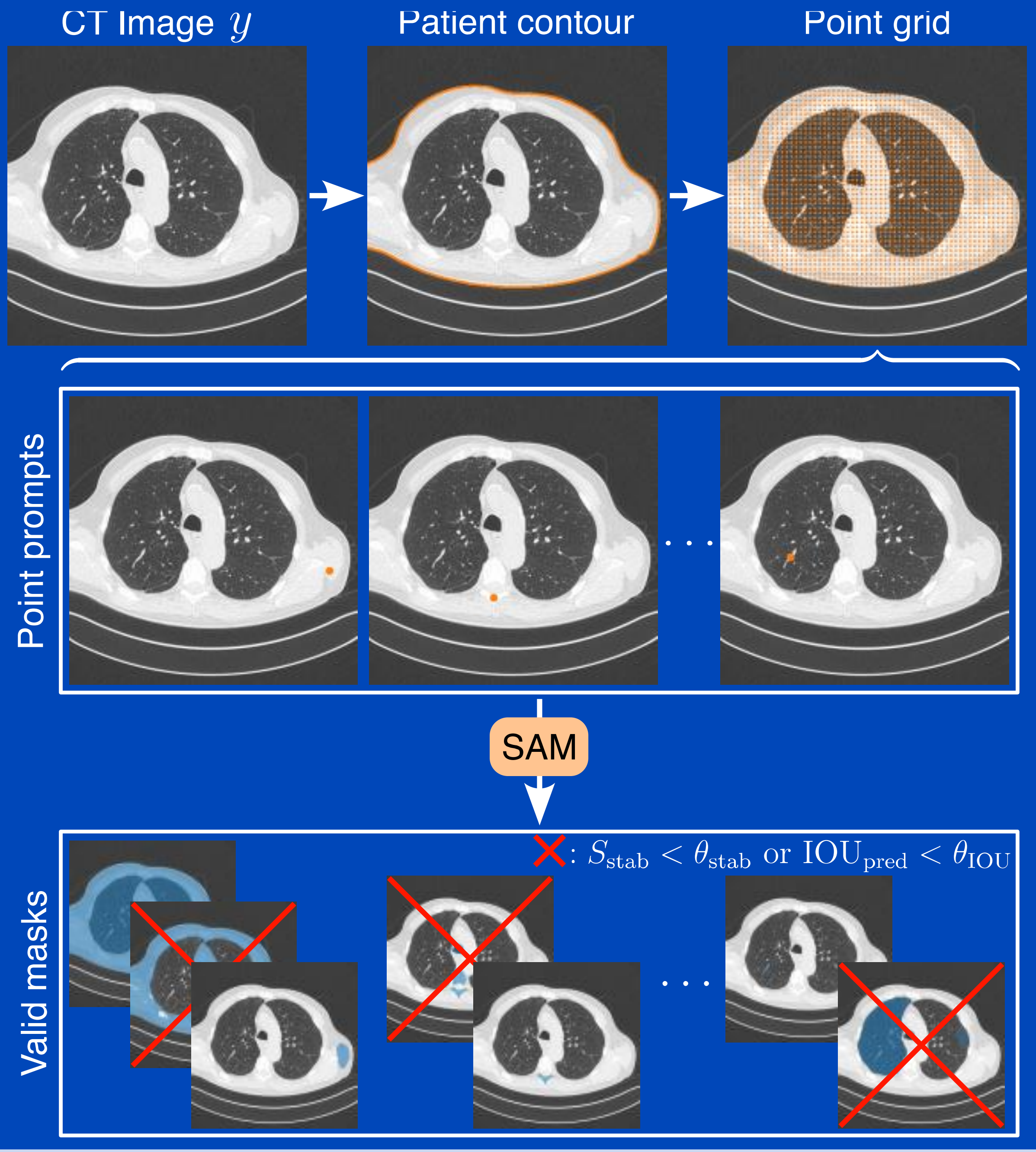
1. Segment arbitrary structures using SAM

1. **Segment patient via thresholding and finding largest contour**
2. **Define a grid of point prompts over the previously found patient segmentation**
3. **Query SAM[1] using these point prompts and filter masks with low S_{stab} and IoU_{pred}**

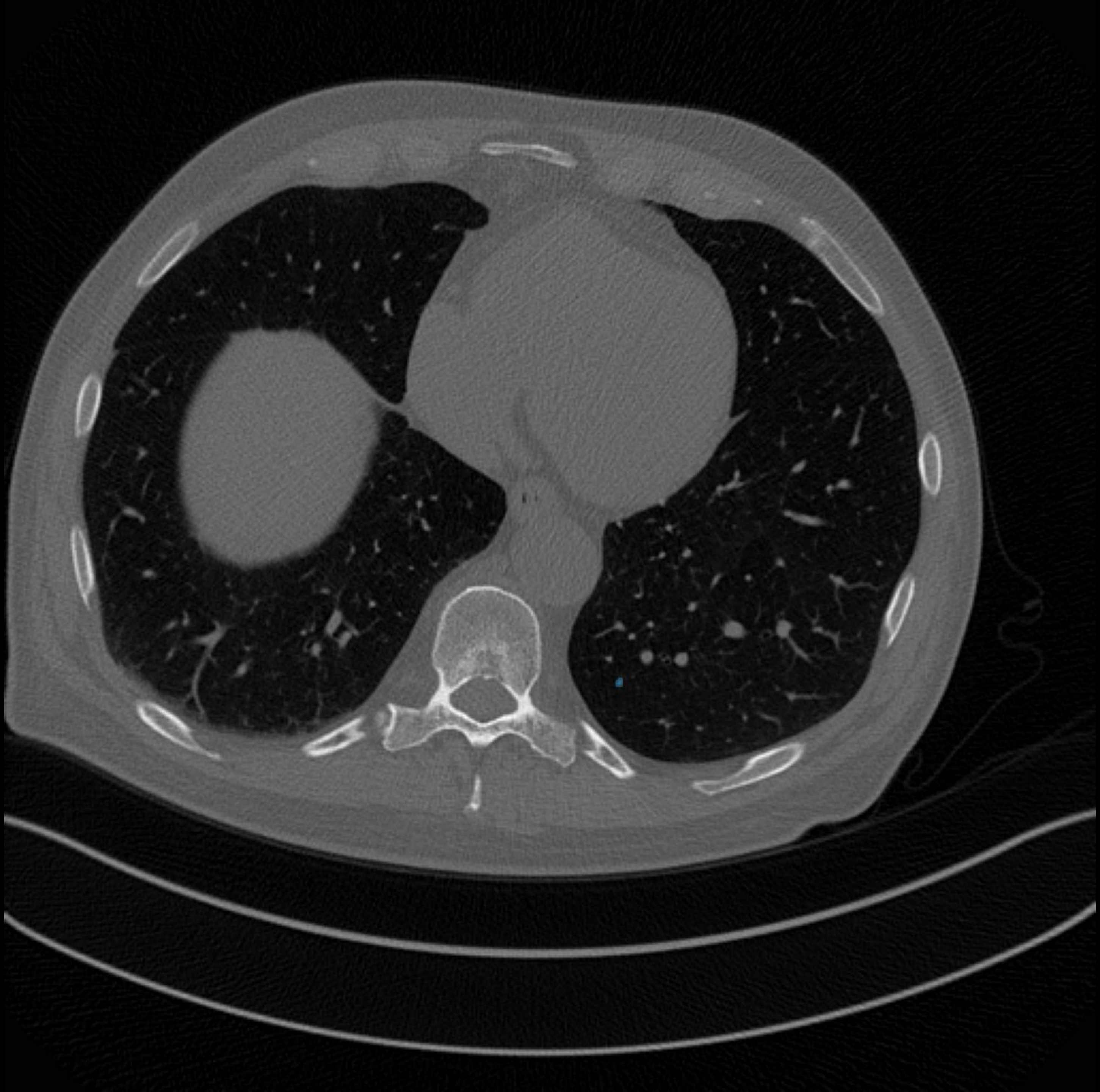
$$S_{\text{stab}}(l, \theta_0, \theta_1) = \frac{|l > \theta_1|}{|l > \theta_0|}, \theta_0 < \theta_1 \qquad \text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

l : Logits predicted by network

4. **Sort segments by increasing area. Starting with smallest segment, remove intersections with any segment already in $\mathcal{S}$ to ensure that each pixel is only assigned to one segment.**



[1]A. Kirillov et al., “Segment Anything,” arXiv preprint, Apr. 2023.



Methods

General setup

Denote with $x \in \mathbb{R}^{H \times W}$ an image reconstructed using some algorithm and $y \in \mathbb{R}^{H \times W}$ the corresponding (aligned) ground truth image

1. Generate a set $\mathcal{S}$ of binary segmentations $s \in \{0, 1\}^{H \times W}$, each representing a structure present in y
2. For each of the segments $s \in \mathcal{S}$ compute a traditional image metric between x and y
3. Aggregate the per-segment metrics to compute a single metric for the entire image x

Methods

2. Segment-wise metric computation

- **Compute traditional image metric for each $m \in \mathcal{M}$, here root mean squared error (RMSE)**
- ***Segment-wise root mean squared error (SRMSE)* between x and y for some segment s:**

$$\mathrm{SRMSE}(x, y; s) = \sqrt{\frac{1}{\sum_{(i,j)} s_{ij}} \sum_{(i,j)} s_{ij} (x_{ij} - y_{ij})^2}$$

- **Other metrics, including ones based on perceptual similarity or mutual information are possible**

Methods

General setup

Denote with $x \in \mathbb{R}^{H \times W}$ an image reconstructed using some algorithm and $y \in \mathbb{R}^{H \times W}$ the corresponding (aligned) ground truth image

1. Generate a set $\mathcal{S}$ of binary segmentations $s \in \{0, 1\}^{H \times W}$, each representing a structure present in y
2. For each of the segments $s \in \mathcal{S}$ compute a traditional image metric between x and y
3. Aggregate the per-segment metrics to compute a single metric for the entire image x

Methods

3. Metric aggregation

After computing SRMSE for each segment, aggregate these values to obtain a single metric for the entire image

Average SRMSE over all segments:

$$\text{Mean-SRMSE}(x, y) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \text{SRMSE}(x, y; s)$$

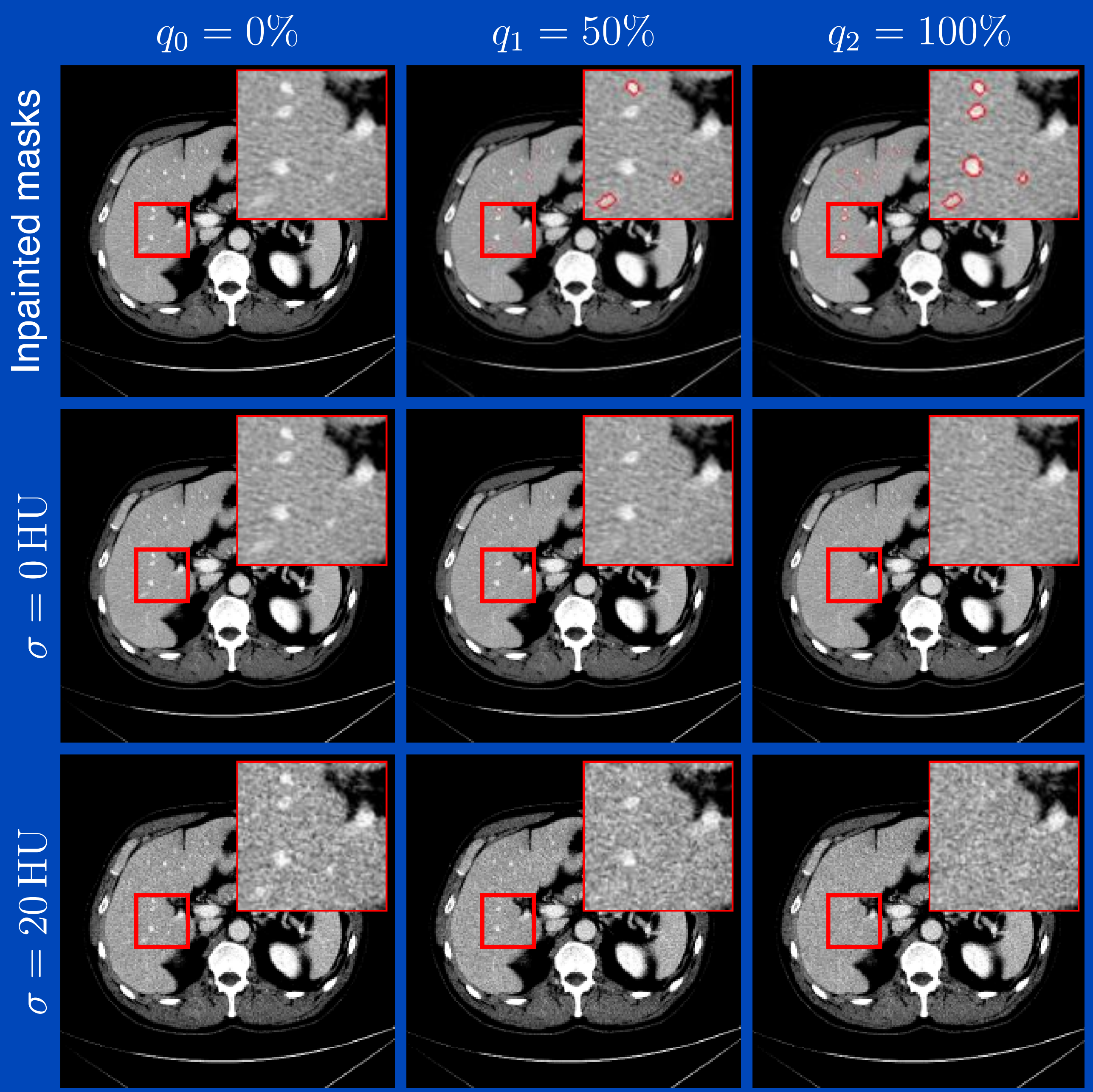
Maximum SRMSE over all segments:

$$\text{Max-SRMSE}(x, y) = \max_{s \in \mathcal{S}} \text{SRMSE}(x, y; s)$$

Experiments

Dataset

- Evaluate sensitivity of our metric to alterations of small structures
- Use abdominal CT dataset with ground truth segmentation of hepatic vessels[1]
- For each patient, remove F increasing fractions $Q = \{q_1, \dots, q_F\}$ of hepatic vessels via inpainting
  → Simulates increasing amount of anatomical changes
- Emulate other deviations by adding Gaussian noise with varying σ
  → Unstructured noise may overshadow the small anatomical differences



[1]M. Antonelli *et al.*, "The Medical Segmentation Decathlon," *Nat Commun*, vol. 13, no. 1, p. 4128, 2022.

Experiments

Evaluation details

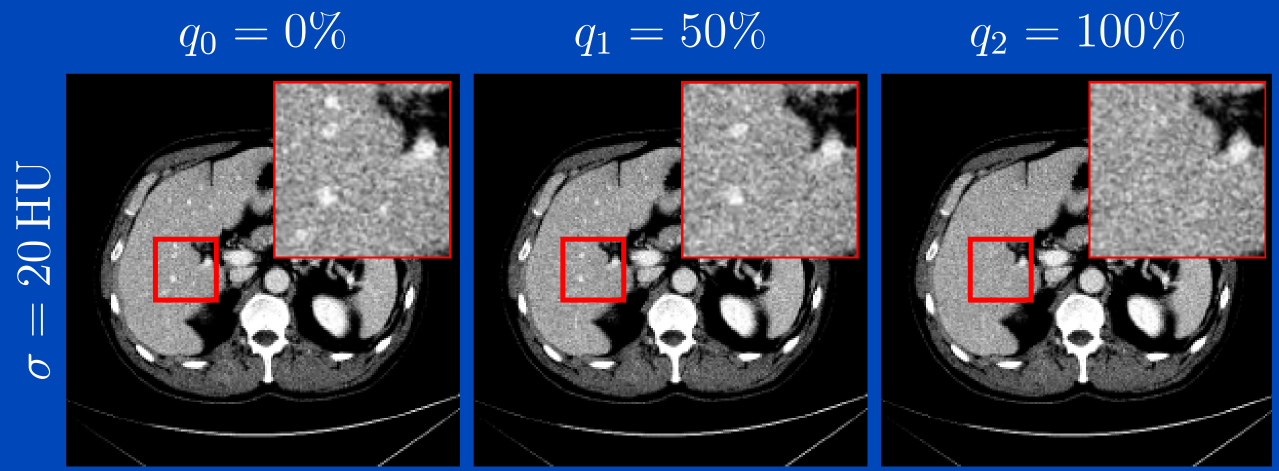
- **Quantify whether a metric detects that in some image more structures were removed than in another image**
- **Normalized Kendall-Tau rank distance[1]:**

$$\tau(A,B) = \frac{2\sum_{(i,j):i<j} t_{i,j}(A,B)}{n(n-1)} \qquad t_{i,j}(A,B) = \begin{cases} 1 & \text{if } \operatorname{sgn}(A_i - A_j)\operatorname{sgn}(B_i - B_j) < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$A_i, B_i : \text{ Ranking of element } i \leq n$$

- **Example:**

$$R_{\text{true}} = [0, 50, 100] \ \%$$
$$R_{\text{RMSE}} = [20, 18, 24] \ \text{HU}$$
$$\tau(R_{\text{true}}, R_{\text{RMSE}}) = 1/3$$

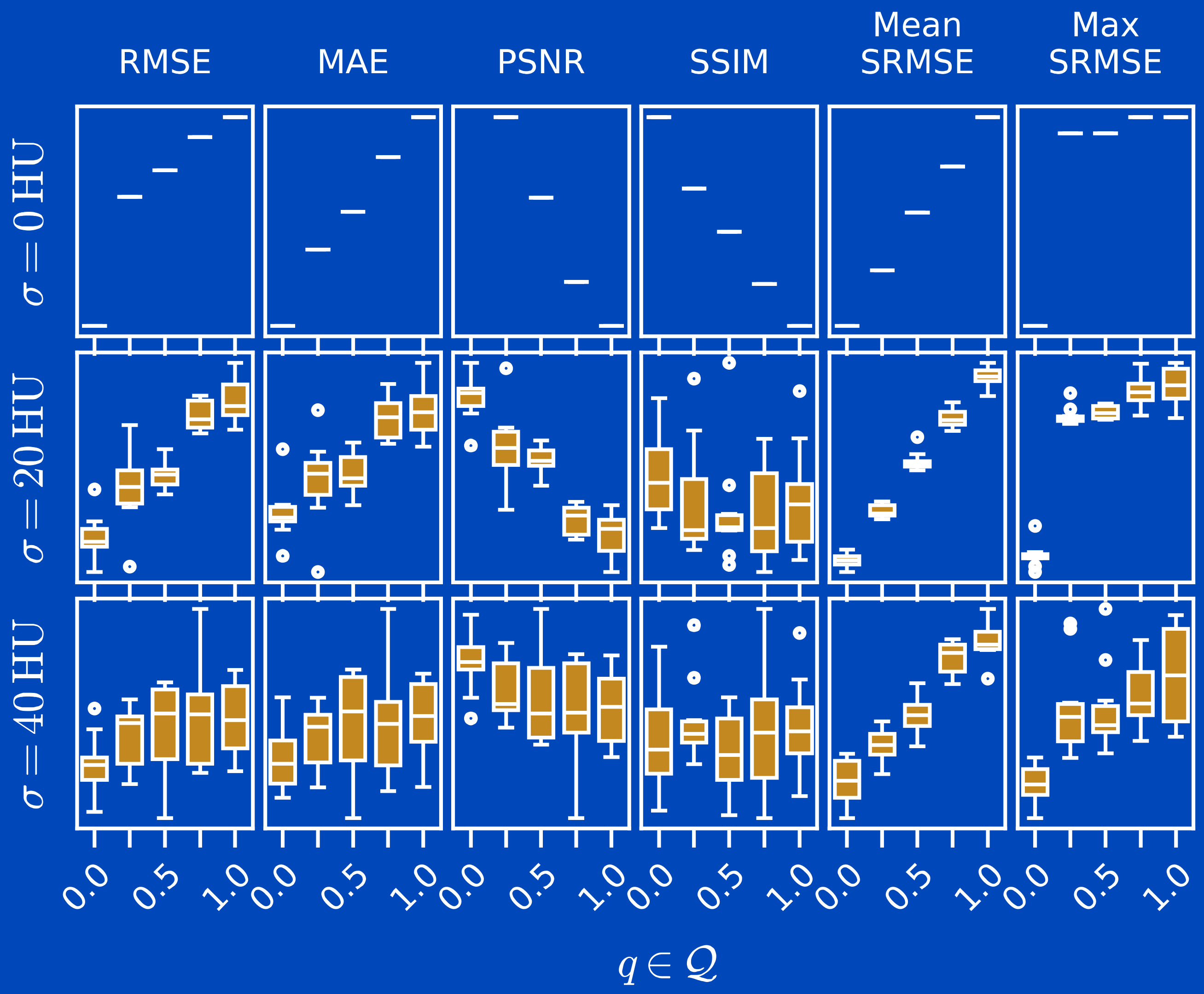


[1]M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

Results

Example for one patient, boxplots correspond to 10 trials of random noise

- **RMSE, PSNR, MAE:** Differences caused by the removal of small structures are overshadowed by noise for large σ
- **SSIM:** Performs worse for small $\sigma = 20$ HU already
- **Mean-SRMSE:** Outperforms other metrics in this regard
- **Max-SRMSE:** Sensitive to removal of any structure, not sensitive to amount of removed structures



Results

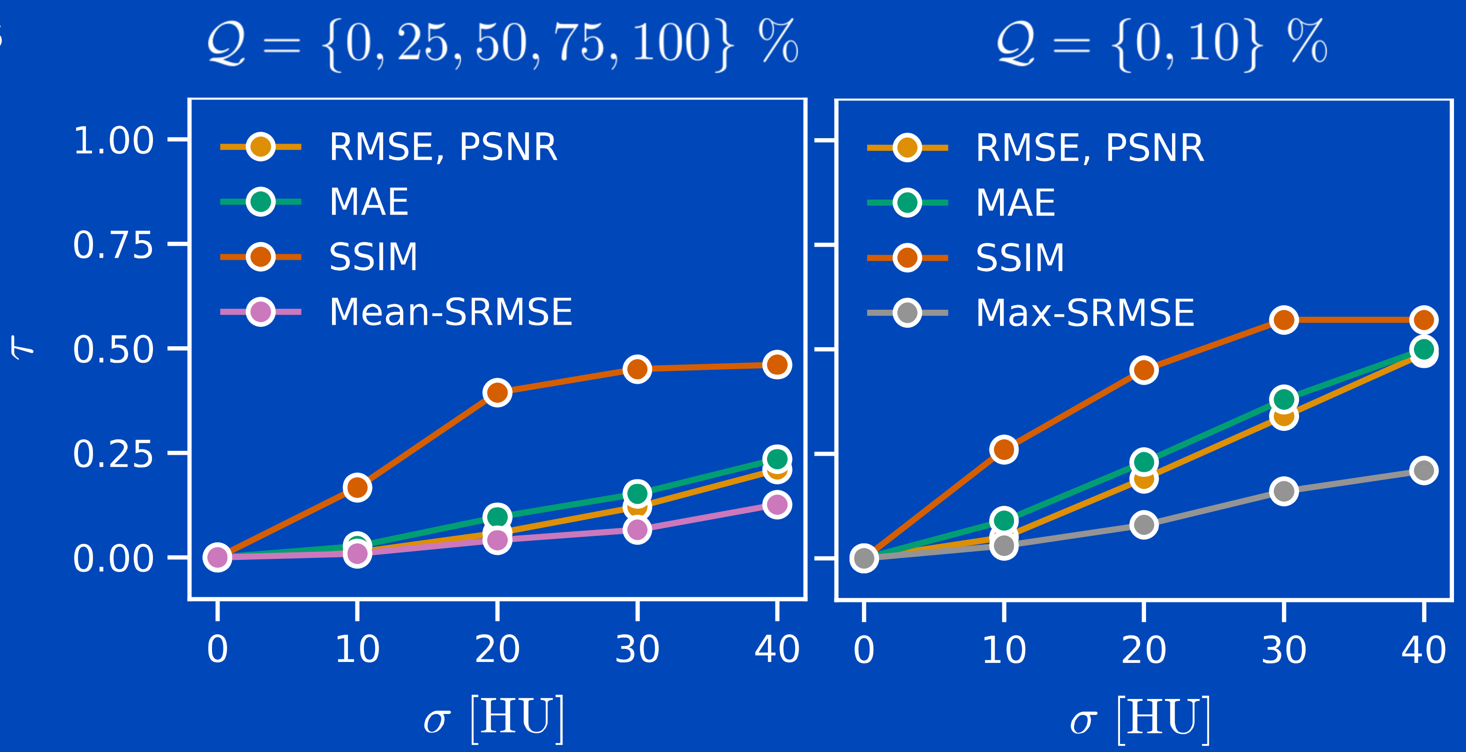
Evaluation of Mean-SRMSE

- Mean-SRMSE is more sensitive to removal of small structures than comparison metrics
- SSIM performs as bad as random guessing for large σ

Evaluation of Max-SRMSE

- Max-SRMSE is more sensitive to removal of very few structures compared to other metrics
- Again, SSIM performs exceptionally bad

For our data, this alters only $10^{-3} - 10^{-2}\%$ of voxels of a patient



Conclusions & Outlook

Conclusions

- Increasing amount of random deviations overshadow systematic removal of small structures for common IQA metrics
- SSIM performs exceptionally bad in this regard
- Preliminary experiments suggest that proposed metrics are more sensitive

Limitations & Outlook

- Proposed metric can only detect removal of structures
- Validate our findings using additional experiments with different modalities
- Explore use of other per-segment metrics and other aggregation methods
- Go Fully3D by using a general-purpose 3D segmentation model
- Use the new metric to train networks

Thank You!

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

APPENDIX

Sparse View Restoration Example

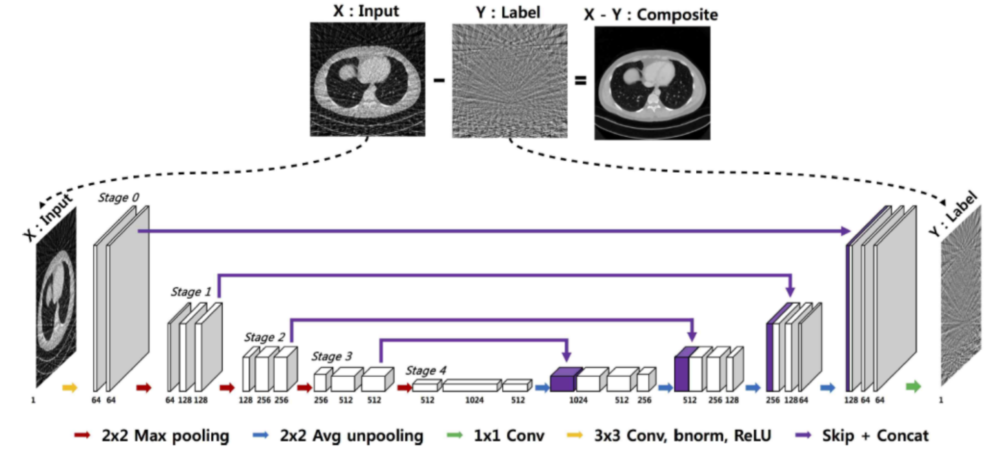
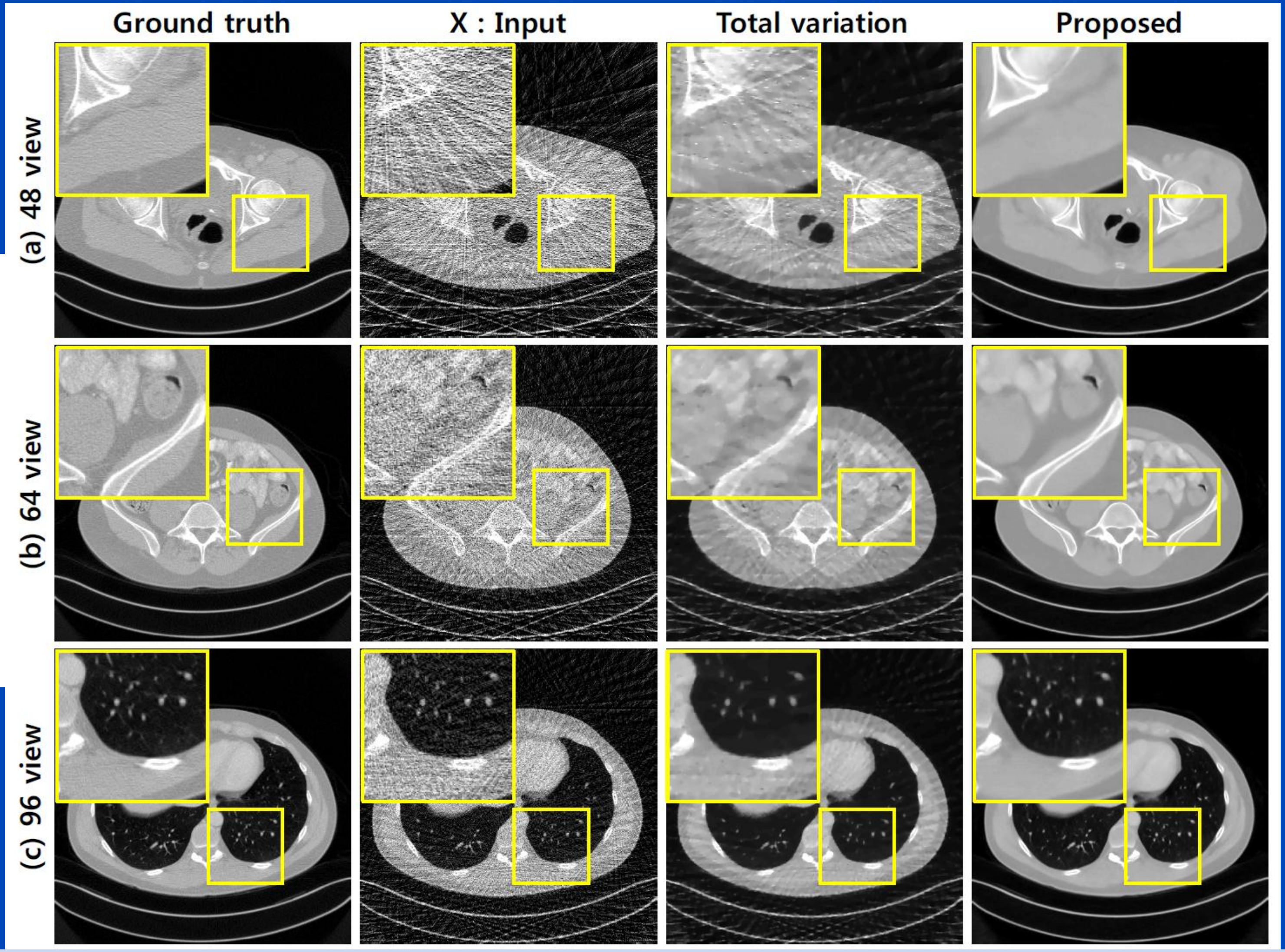
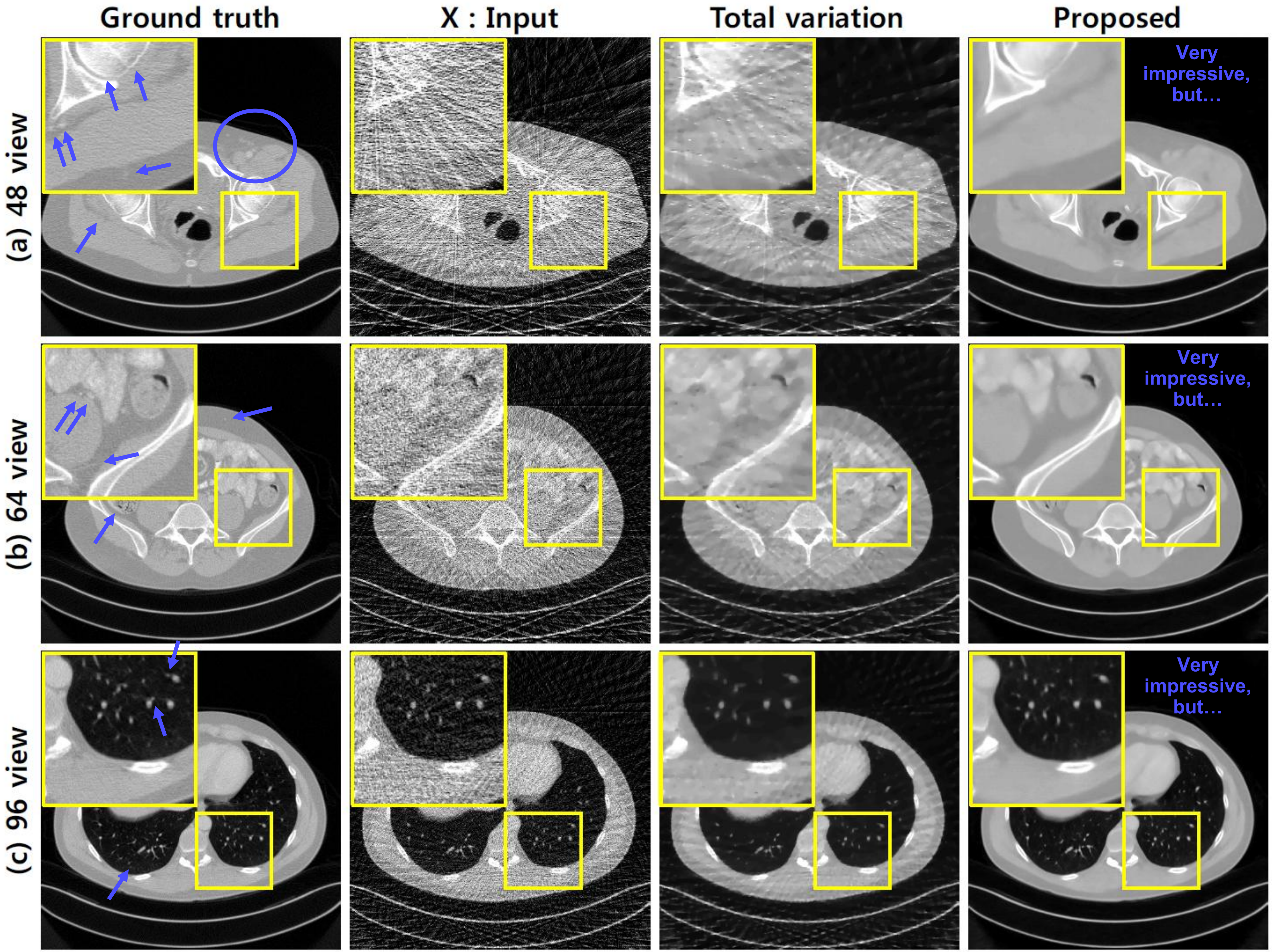


Figure 1. The proposed deep residual learning architecture for sparse view CT reconstruction.



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…

Resolution Improvement Example

- 2D U-net to converts 5 mm thick images into 1 mm ones.
- E.g. to “replace a scanning protocol for a 1 mm slice with a 5 mm protocol”