AAPM 2021
JULY 25–29 | VIRTUAL
63RD ANNUAL MEETING & EXHIBITION

FDA U.S. FOOD & DRUG ADMINISTRATION

Frontiers in AI and Its Applications in Medical Physics

GROWING SIGNIFICANCE OF DEEP LEARNING IN MODERN MEDICAL IMAGING

Rongping Zeng

Division of Imaging, Diagnostics and Software Reliability (**DIDSR**)

Office of Science and Engineering Laboratories

Center for Devices and Radiological Health

U.S. Food and Drug Administration

July 2021

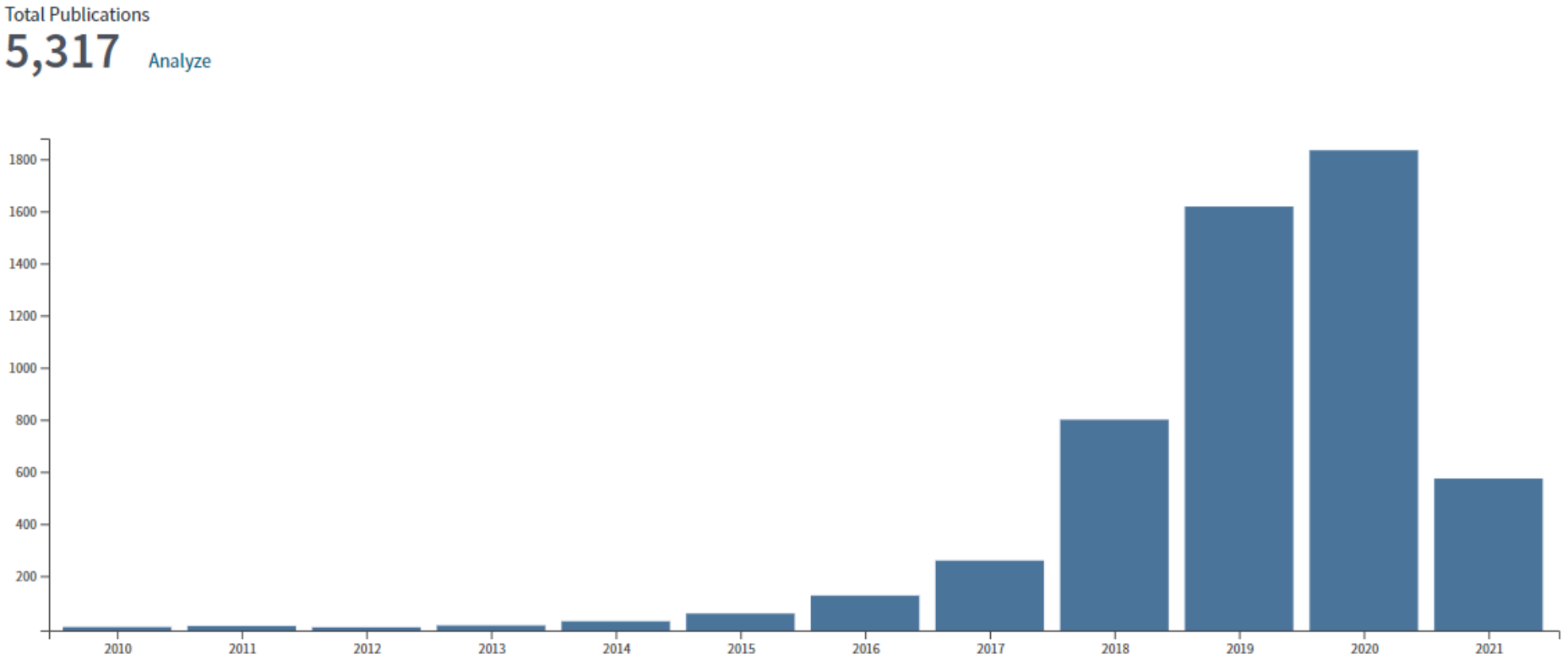
Disclaimer

The mention of commercial products, their sources, or their use in connection with materials reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

Outline

- Deep learning (DL) and medical imaging
- DL in image reconstruction/processing
  - Applications
  - Performance evaluation
- A few DIDSR research studies
  - MRI reconstruction
  - Synthetic digital mammography
  - Low-dose CT image denoising
- Summary

Growth of DL in medical imaging

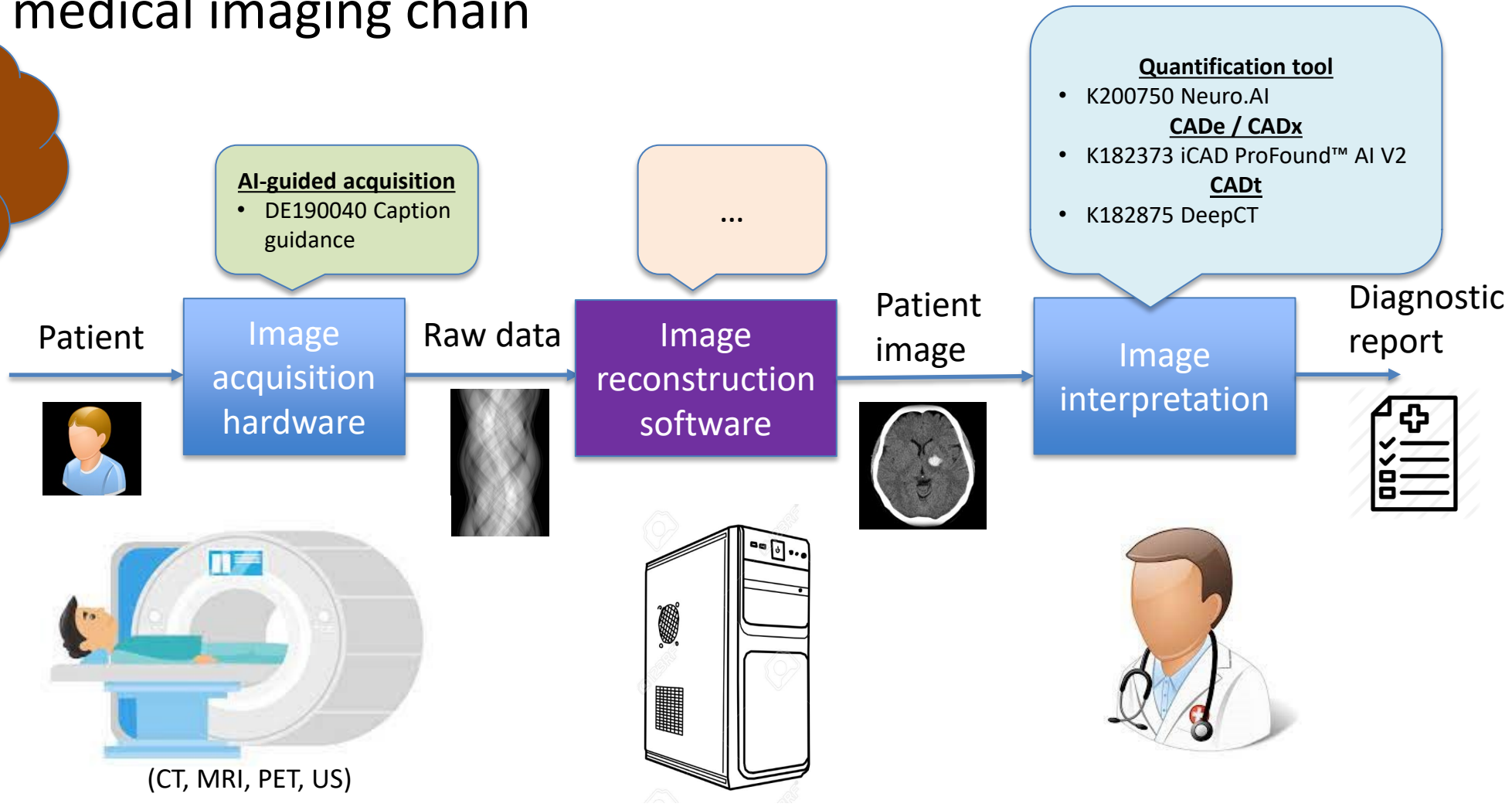


Web of science: **deep learning**, **radiology nuclear medicine medical imaging, 2010 - May 2020**

DL in modern medical imaging

- A modern medical imaging chain

130+ AI-enabled medical imaging products cleared in the United States*



*Wu et.al., Nature Medicine **volume 27**, pages582–584 (2021)

DL in image reconstruction/processing

- Canon: “AiCE” for CT, MRI, DE-CT
- GE: “TrueFidelity“ for CT, “AIR Recon DL” for MRI
- Siemens: “Deep Resolve” for MRI
- UI: “uAI HYPER DLR” for CT

Raw data

Raw data pre-processing
- Denoising
- Inpainting/interpolation
- Scatter correction

Image reconstruction
- Unrolling
- Learned priors

Image post - processing
- Denoising
- Artifact reduction

Patient image

Image quality directly affects the clinical decision.

DL in image reconstruction/processing

*Table 1 | Recent deep learning tomographic reconstruction methods for various biomedical imaging modalities based on their technical contributions

Modality	Architecture	Applications and citations
X-ray CT	Image domain	Low-dose[15–17,48,70,97], sparse view[29,113], limited angle[50], metal artefact[52], dual-energy[143]
	Unrolling	Sparse view[29,58,59,61,62], low-dose[60]
	Sensor domain	Metal artefact[85,86], sinogram analysis[87]
	Model-based/plug-and-play	Low-dose[60,68,70,82,144], sparse view[58,59], interior[83], cone-beam artefact[84]
	Domain transform	Low-dose and sparse view[77–79], interior tomography[83]
	GAN, unsupervised	Low-dose[48,97,110], metal artefact[130]
MRI	Image domain	Accelerated MRI[24,30,49]
	Unrolling	Accelerated MRI[23,25,53–56]
	Sensor domain	Accelerated MRI[88,90,91], artefact removal[89]
	Model-based/plug-and-play	Accelerated MRI[26,63,72], dynamic MRI[65,75]
	Domain transform	Accelerated MRI[76]
	GAN, unsupervised	Accelerated MRI[49,109], contrast synthesis[106,107]
PET/SPECT	Image domain	Low-dose PET[31,32,145,146]
	Model-based/plug-and-play	Low-dose PET[31,67–69]
	Domain transform	Low-dose PET[80,81]
	GAN, unsupervised	Low-dose PET[111], attenuation correction[104,105]
Ultrasound	Image domain	Artefact removal[45], photoacoustic artefact removal[46,47]
	Sensor domain	Radiofrequency interpolation[96]
	Model-based/plug-and-play	Photo-acoustic imaging[64]
	Domain transform	Beamformer[92,94,95]
Optical	Image domain	Super resolution[36–38,103], ghost imaging[42], scattering medium imaging[40,41], mobile phone microscopy[43]
	GAN, unsupervised	Deconvolution microscopy[100,102,103]

*Wang, G., Ye, J.C. & De Man, B. **Deep learning for tomographic image reconstruction**. *Nat Mach Intell* **2,** 737–748 (2020).

- *Potential to achieve comparable image performance as model-based iterative image reconstruction methods*
- *Faster computation speed*
- *New challenges on image quality assessment*

Image quality (IQ)

Performance testing
- Diagnostic accuracy

- Global IQ metrics:
  - MSE, PSNR, SSIM etc.

Less predictive

- Standard system IQ metrics
  - MTF, NPS, CNR
  - Non-uniformity, CT number accuracy, etc.

Predictive for linear systems

- Task-based IQ metrics
  - Lesion detectability, estimation of lesion shape and lesion size, etc.

Most clinically relevant

Generalizability Performance

- DL performance highly depends on the training data
- Real-world data is much broader
  - Imaging system/acquisition protocol
  - Patient population: sex, body size, disease, etc.
  - Data site
- Understanding of the generalizability performance improves regulatory review
  - More accurate definition of the intended use of a device
  - Appropriate design of the testing data variety

Key Considerations for Authors, Reviewers, and Readers of AI/ML Manuscripts in *Radiology*

Key Considerations
Are all three image sets (training, validation, and test sets) defined? (non overlap, independent)
Is an *external* test set used for final statistical reporting?
Have multivendor images been used to evaluate the AI algorithm?
Are the sizes of the training, validation, and test sets justified?
Was the AI algorithm trained using a standard of reference that is widely accepted in our field?
Was preparation of images for the AI algorithm adequately described?
Were the results of the AI algorithm compared with radiology experts and/or pathology?
Was the manner in which the AI algorithm makes decisions demonstrated?
Is the AI algorithm publicly available?

Note.—AI = artificial intelligence, ML = machine learning.

Assessing Radiology Research on Artificial Intelligence, Bluemke et. al., Radiology, Dec.,2019

A few DIDSR research studies

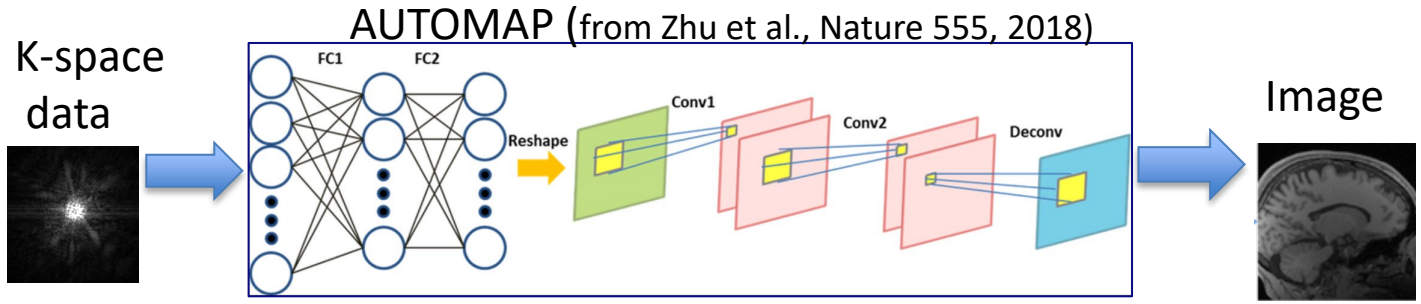
- MRI reconstruction
  - Effect of training data on MRI AUTOMAP reconstruction
- Synthetic digital mammography
  - Tasked-based evaluation of a DL-based synthetic digital mammography algorithm
- Low-dose CT image denoising
  - 1. Framework for DL model optimization
  - 2. Generalizability performance Evaluation

MRI reconstruction

Comparison of Quality Assessment Methods for Deep-Learning-Based MR Image Reconstruction

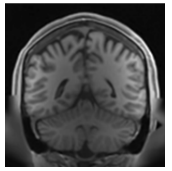
Yazdanabady M, Fujimoto K, Berman B, Rosen M, Koonjoo N, Zhu B, Graff CG, Pezeshk A, ISMRM 2019

Purpose: Evaluate the influence of the choice of training set composition on AUTOMAP image reconstruction.

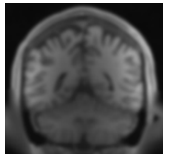


- Performance testing
  - Testing set: 9,600 coronal brain MRI slices from 30 HCP subject datasets (not used in training)
  - Metrics:
    - Global IQ (e.g., MAE)
    - Task IQ (e.g., Cortical thickness)

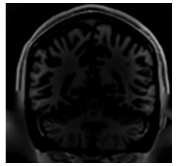
- Five training datasets



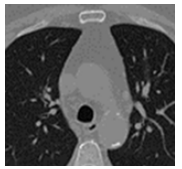
Human connectome project (HCP) : 50,000 T_1W brain MRI slices



Gaussian low-pass filtered HCP (HCP-LP)



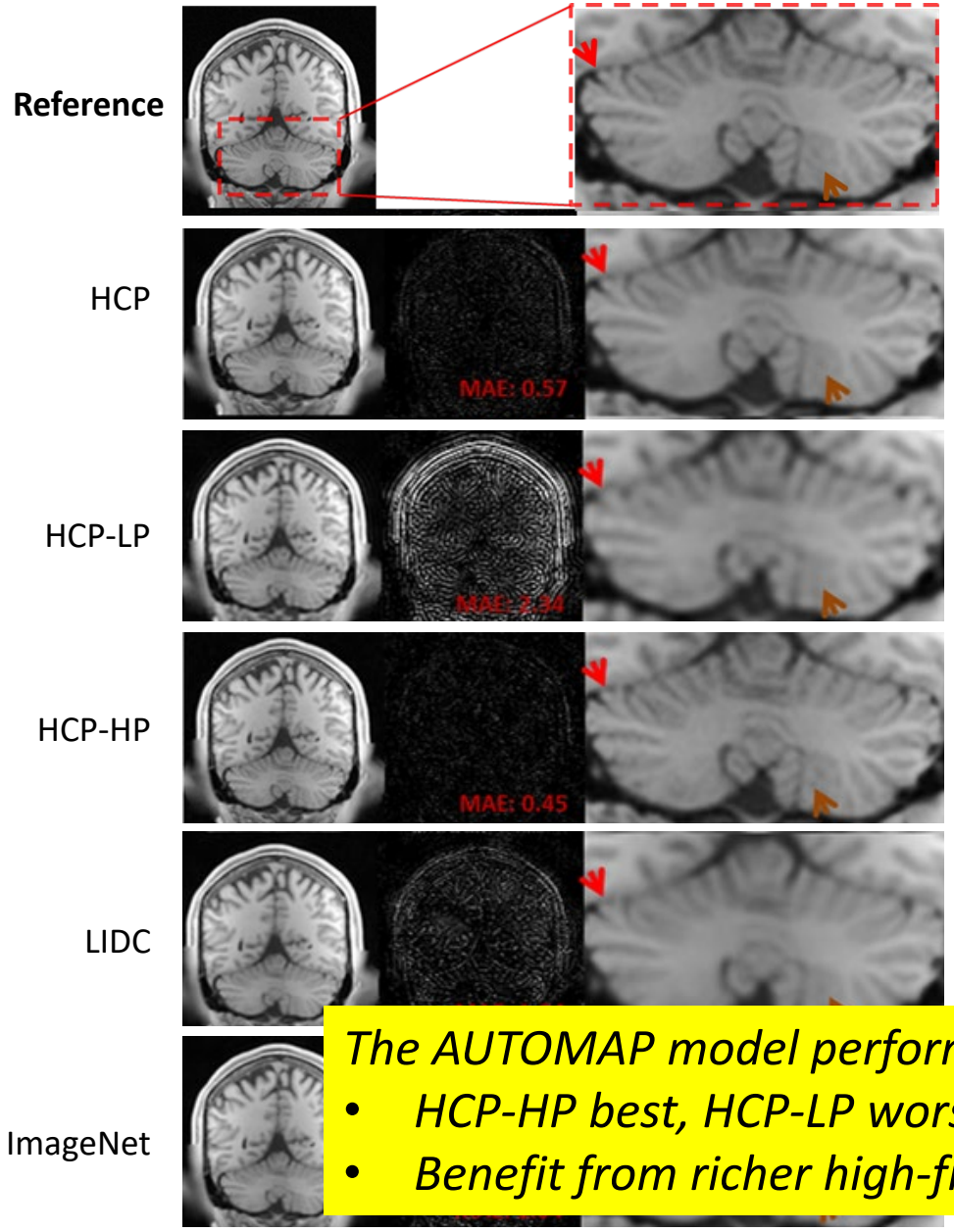
Gaussian high-pass filtered HCP (HCP-HP)



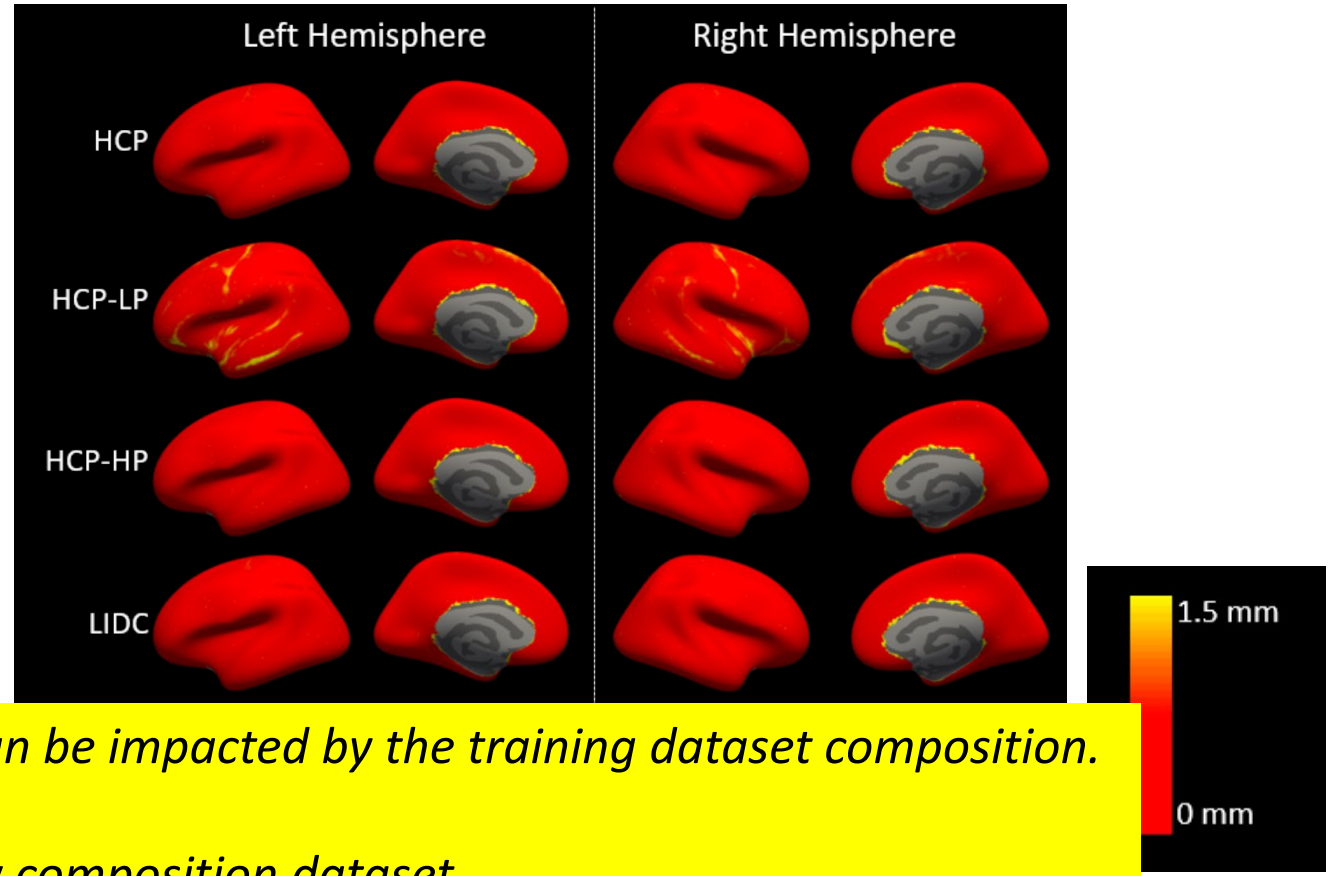
lung image database consortium (LIDC): 50,000 lung CT slices

ImageNet: 50,000 natural images

MRI reconstruction
Results



Absolute Cortical Thickness Difference*



The AUTOMAP model performance can be impacted by the training dataset composition.

- *HCP-HP best, HCP-LP worst.*
- *Benefit from richer high-frequency composition dataset.*

B *et al.*, 1999]

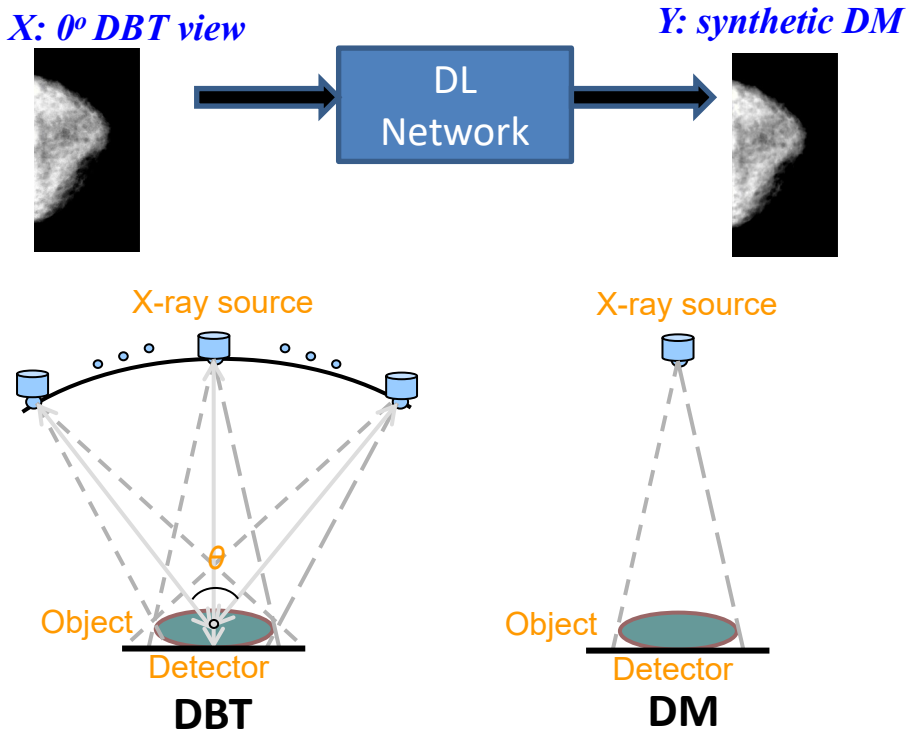
Synthetic Digital Mammography (DM)

Virtual clinical trial for task-based evaluation of a deep learning synthetic mammography algorithm

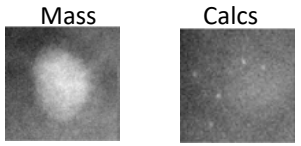
Badal A, Cha K, Divel S, Graff CG, Zeng R, Badano A, SPIE medical imaging conference 2020

Purpose: Evaluate a synthetic DM method through a DL-based denoising of the digital breast tomosynthesis (DBT) 0º projection view.

- Network structures: CNN3, DnCNN, REDCNN
- IQ metrics
  - Global IQ: PSNR;
  - Task IQ: Lesion detectability (spiculated mass/calcification cluster)



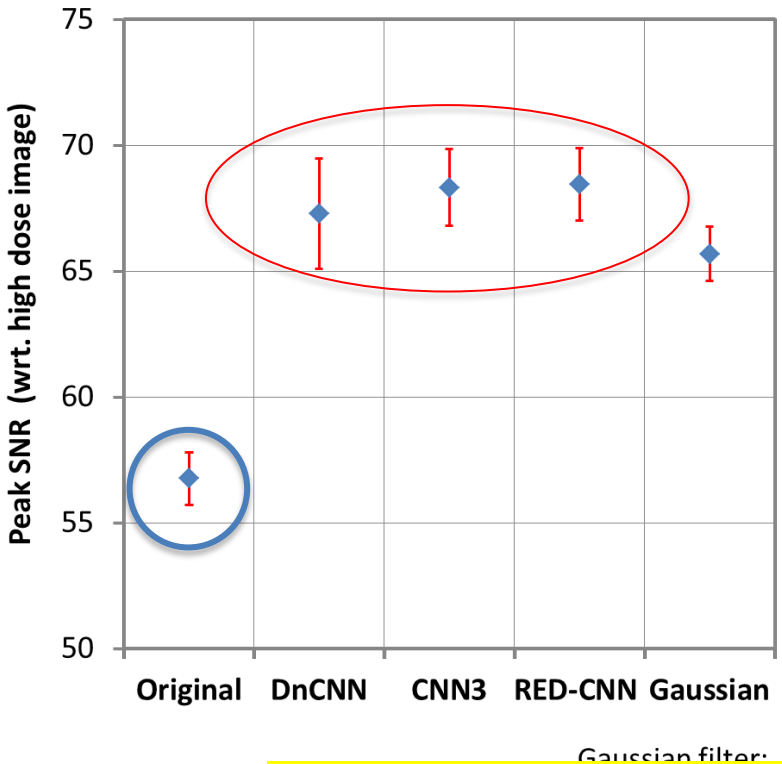
- Using the VICTRE* pipeline to create 120 breast phantom data
  - Training: 100
    - 25-projection DBT and high-dose DM (x25 dose)
    - Not contain lesions
  - Testing: 3000
    - Contain lesions



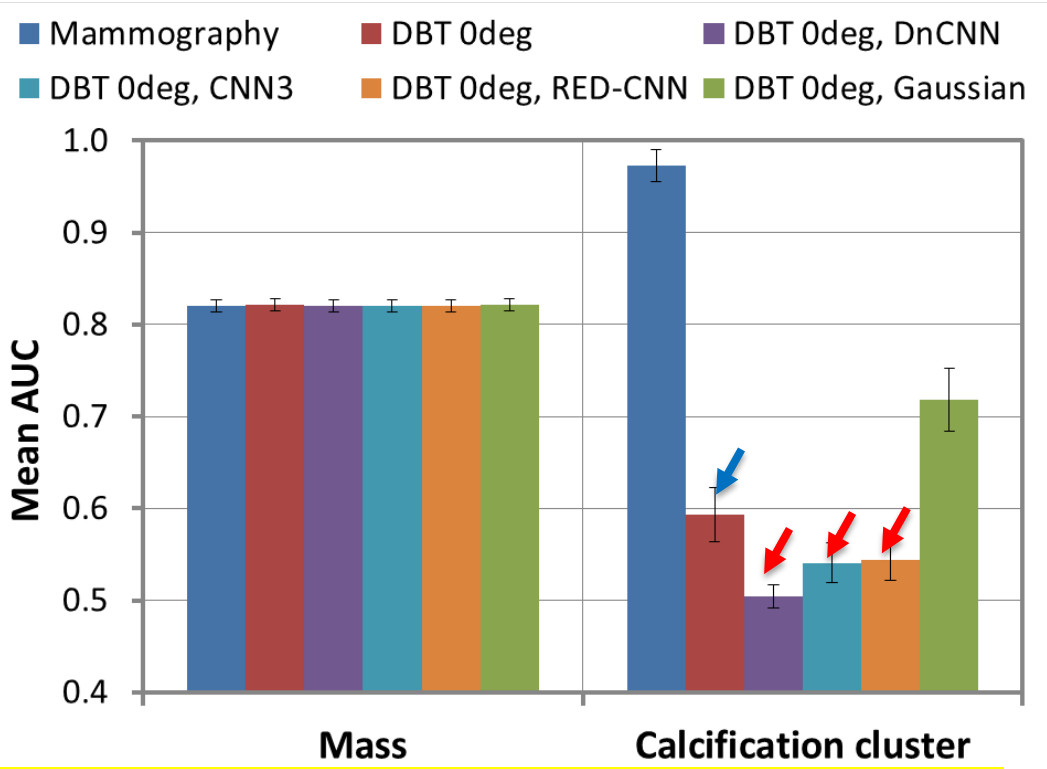
* https://github.com/DIDSR/VICTRE

Synthetic Digital Mammography (DM)
Results

Global IQ metric: PSNR



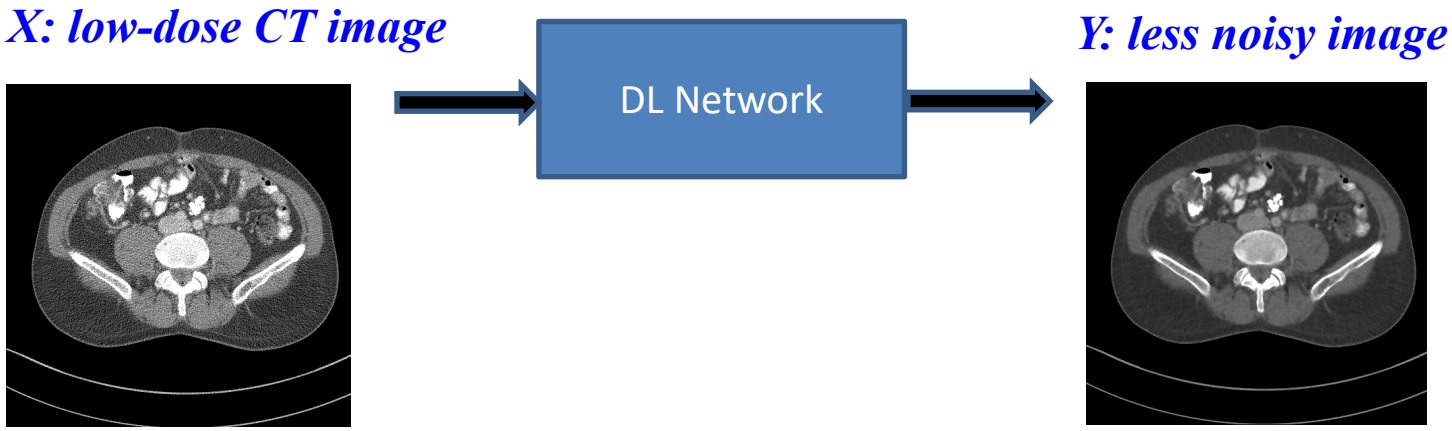
Task-based IQ metrics



- *Diagnostic performance of post-processed images does not always correlate with PSNR.*
- *Lesion detectability might decrease even if PSNR in image increases.*

Low-dose CT (LDCT) image denoising

- The DL network is trained to process a low-dose CT image slice (e.g., reconstructed by FBP) with the intend to output a cleaner image.



  - Training data
    - Data resource: The low-dose CT grand challenge* (7 out of 10 for training)
    - Pairs of image patches extracted from full and quarter-dose image slices

LDCT - Study 1

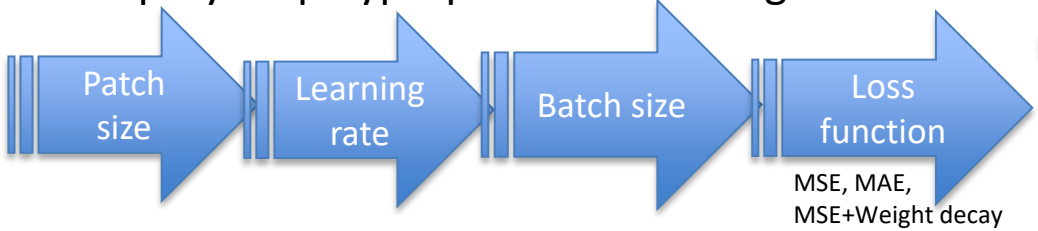
Deep neural networks-based denoising models for CT imaging and their efficacy

Kc P, Zeng R, Farhangi M, Myers KJ, SPIE Medical Imaging 2021

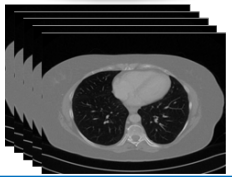
Purpose: Holistic framework for optimizing DL denoising models using both global IQ and bench performance IQ metrics.

- CNN3, REDCNN, DnCNN, Du-DnCNN, GAN

- Step-by-step hyperparameter tuning

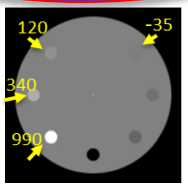


PSNR, SSIM calculated using a tuning dataset

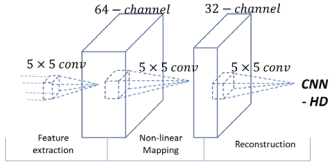


- Preprocessing choices
  - Normalization
    - Normalize image intensity to the range of [0,1].
  - No normalization
    - Use the original dynamic range [0, 2048].

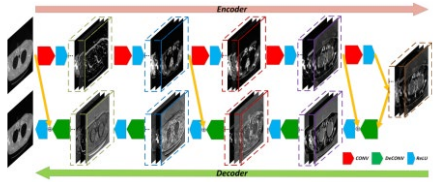
HU accuracy estimated using a standard CT IQ phantom image



- **CNN3**: 3-layer CNN, from Chen et el., BMOE 2017

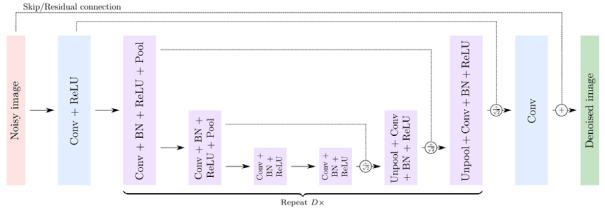


- **REDCNN**: 10-layer residual encoder-decoder CNN, from Chen et al., IEEE-TMI, 2017



- **DnCNN**: 17-layer CNN, from Zhang et al., IEEE_TIP, 2017

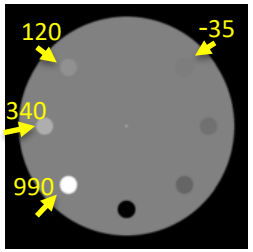
- **Du-DnCNN**: Dilated U-shaped DnCNN, form Cheng et al., "Image-denoising-with-deep-cnns", 2020 https://github.com/lychengr3x/Image-Denoising-with-Deep-CNNsImage-Denoising-with-Deep-CNNs.



- **GAN**: RESNET as generator, DnCNN as discriminator, based on Goodfellow et al., "Generative adversarial nets", 2014

LDCT - Study 1
Results

- HU accuracy and PSNR



HU disk	-35	120	340	990	PSNR
FBP	**-34.91**	**119.97**	**338.94**	**987.88**	
CNN3, Norm	-70.46	74.80	293.09	925.85	**35.99**
CNN3, No Norm	**-35.16**	**119.45**	**339.38**	**991.85**	35.61
REDCNN, Norm	-59.47	96.87	336.36	959.66	35.26
REDCNN, No Norm	**-33.83**	**120.02**	**339.89**	**992.39**	**35.76**
DnCNN, Norm	**-41.01**	**115.67**	**334.25**	**982.51**	**35.13**
DnCNN, No Norm	-55.21	99.87	318.53	967.52	34.99
Du-DnCNN, Norm	-49.36	102.99	316.01	960.31	34.82

- *Data normalization affects the HU accuracy performance differently for different network structures.*
- *Consider using bench performance to complement the global metrics when optimizing the DL denoising models.*

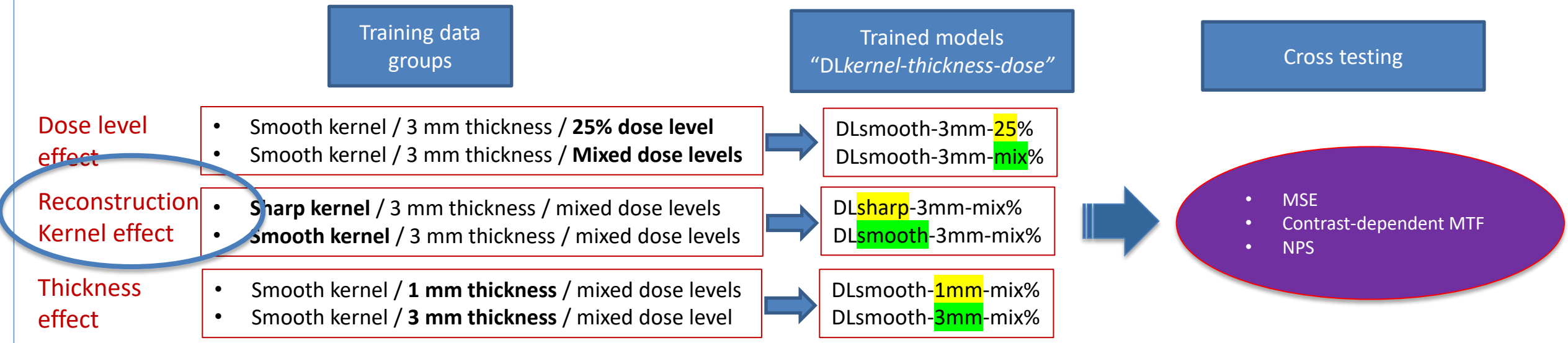
LDCT - Study 2

Generalizability performance of a deep learning-based CT image denoising method

Zeng R, Lin Y, Li Q, Lu J, Skopec M, Fessler JA, Myers KJ, The 6th X-ray CT conference 2020

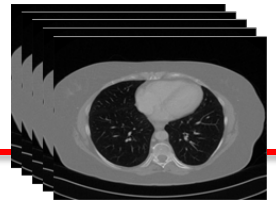
Purpose: Investigate the generalizability performance of a CT image denoising network (REDCNN) in data acquired with different scan parameters

- Dose, reconstruction filter, slice thickness
- Identify underlying data property shift that may affect the generalizability



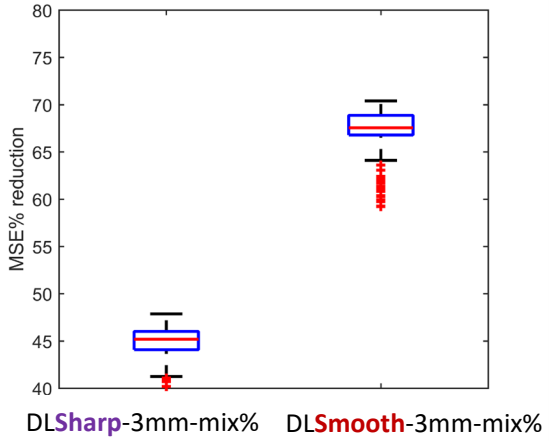
LDCT - Study 2

Results: reconstruction kernel effect

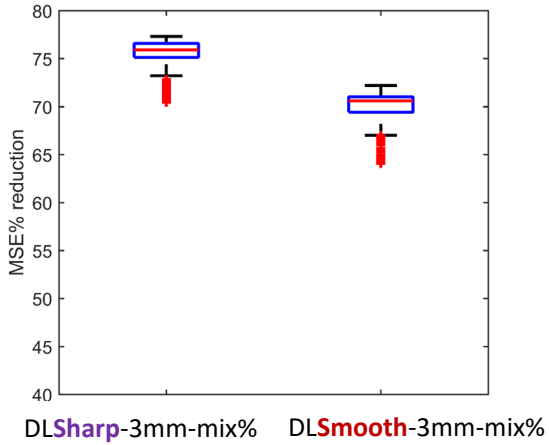


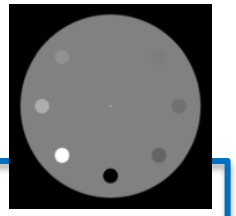
MSE

Test: FBP-smooth

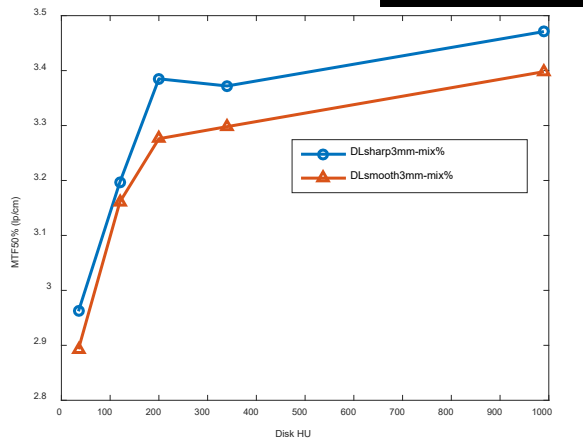


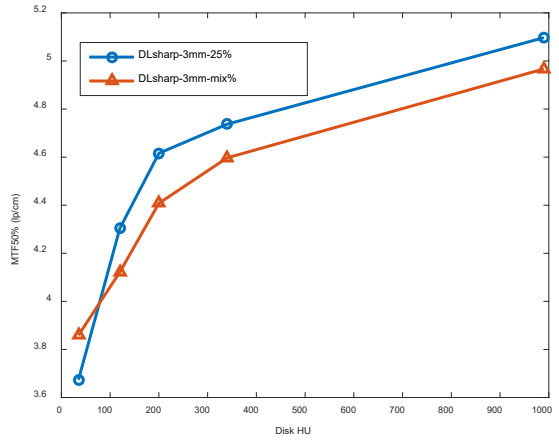
Test: FBP-sharp

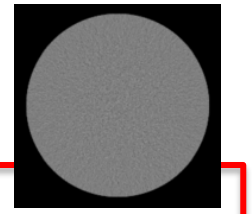




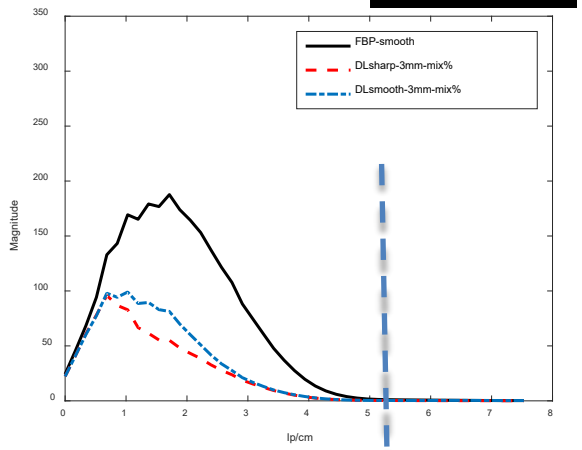
Contrast-dependent MTF

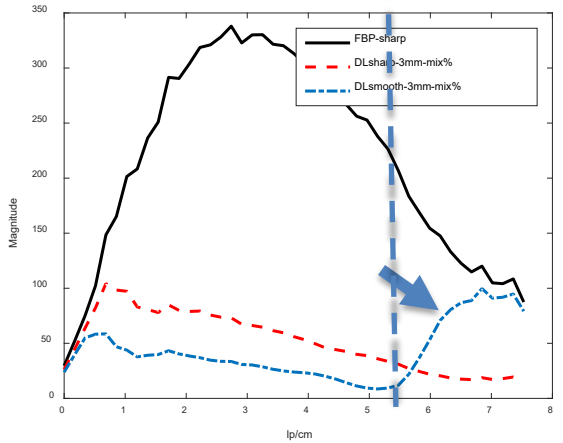




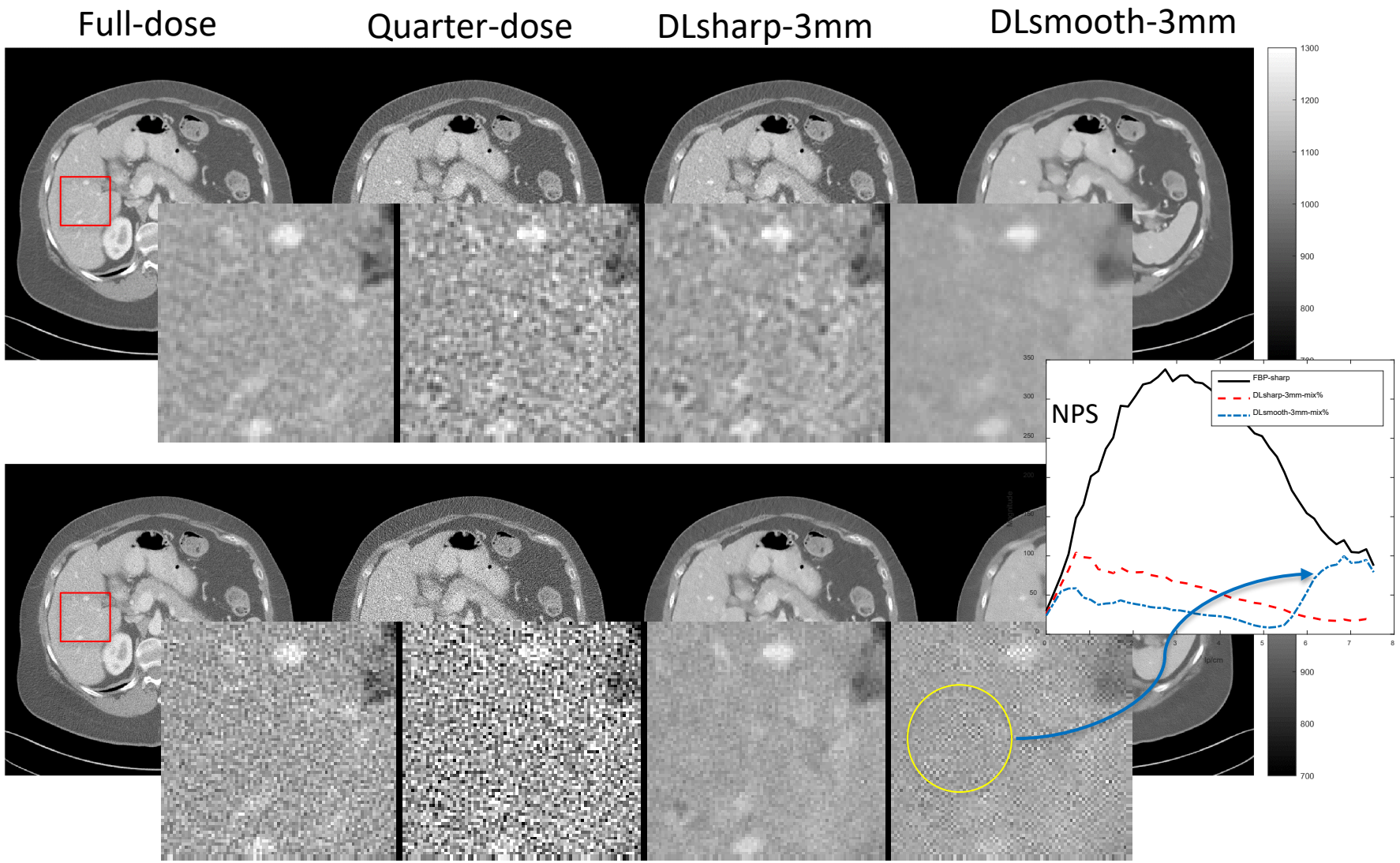


NPS



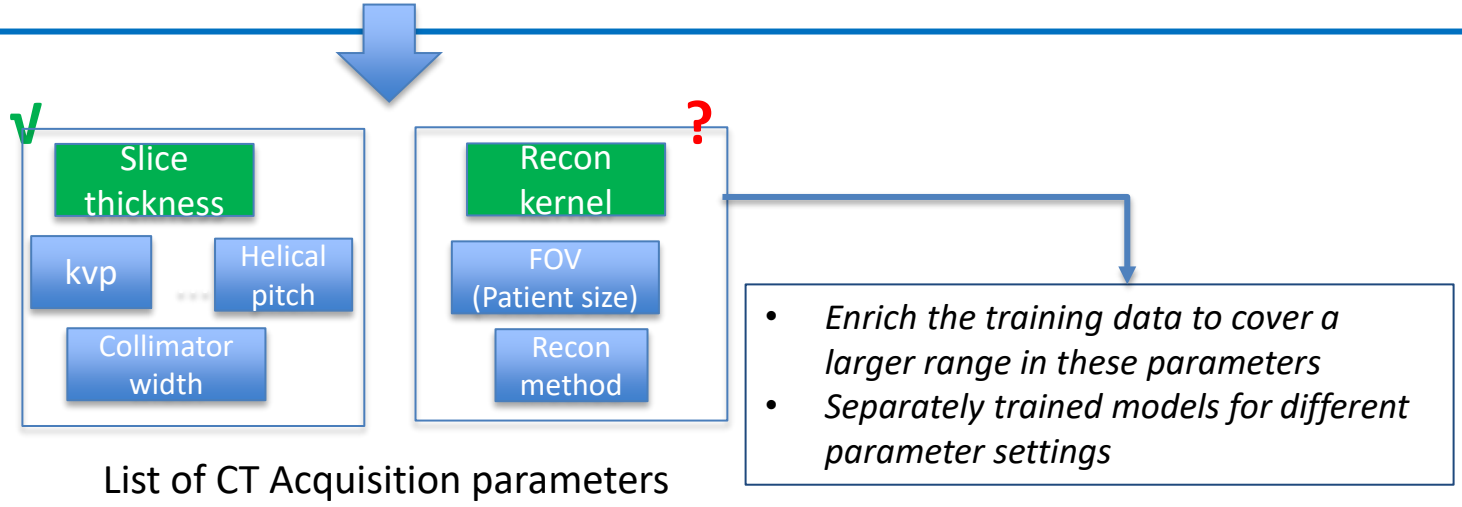


LDCT - Study 2
Results: reconstruction kernel effect



LDCT - Study 2
Results: generalizability

- Performance of REDCNN, a slice-wise CT image denoising network REDCNN
  - Generalized poorly between the sharp and smooth reconstruction kernel;
  - Not different between the 1mm and 3mm slice thickness data;
  - More robust when trained with mixed-dose data.
- The NPS property in a CT image may be used to as an underlying data property to predict whether the denoising performance is generalizable to the data or not.



Summary

- DL applies to every stage in a modern medical imaging chain
  - From image acquisition, image reconstruction, to image interpretation
- DL-based image reconstruction/processing
  - Consider multiple types of IQ metrics in algorithm development and performance evaluation
  - Investigate the generalizability performance
    - Help with more efficient training and testing data design
    - Ensure a DL method is used within its generalizable range

Acknowledgements

- My FDA colleagues
  - Dr. Andreu Badal
  - Dr. Berkman Sahiner
  - Dr. Jiang Lu
  - Dr. Kyle J Myers
  - Dr. Kyoko Fujimoto
  - Dr. Prabhat Kc
- Host
  - Dr. Lei Xing

Rongping.zeng@fda.hhs.gov
FDA/CDRH/OSEL/DIDSR