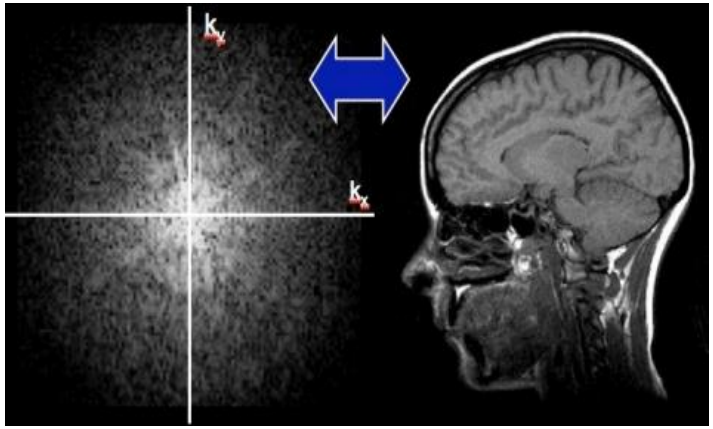
Recurrent Generative Adversarial Networks for Compressive Image Recovery

Morteza Mardani
Research Scientist
Stanford University, Electrical Engineering and Radiology Depts.

March 26, 2018

Motivation

- High resolution Image recovery from (limited) raw sensor data



- Medical imaging critical for diseases diagnosis
  - MRI is very slow due to the physical and physiological constraints
  - High dose CT is harmful
- Natural image restoration
  - Image super-resolution, inpainting, denoising
- Seriously ill-posed linear inverse tasks

Challenges

- Real-time tasks need rapid inference
  - Real-time visualization for interventional neurosurgery tasks
  - Interactive tasks such as image super-resolution on a cell phone
- Robust against measurement noise and image hallucination
  - Data fidelity controls the hallucination; critical for medical imaging!
  - Often happens due to memorization (or overfitting)
- Plausible images with high perceptual quality
  - Radiologists need to see sharp images with high level of details for diagnosis
  - Conventional methods usually rely on SNR as a figure of merit (e.g., CS)
- **Objective:** rapid and robust recovery of plausible images from limited sensor data by leveraging training information

Roadmap

- Problem statement
- Prior work
- GANCS
  - Network architecture design
  - Evaluations with pediatric MRI patients
- Recurrent GANCS
  - Proximal learning
  - Convergence claims
  - Evaluations for MRI recon. and natural image super-resolution
- Conclusions and future directions

Problem statement

- Linear inverse problem ($M << N$)

$$\mathbf{y} \approx \mathbf{\Phi x} \qquad \mathbf{\Phi} \in \mathbb{C}^{M \times N}$$

- $\mathbf{x}$ lies in a low-dimensional manifold $\mathcal{M}$
- About $\mathcal{M}$ only know the training samples $\mathcal{X} := \{\mathbf{x}_k\}_{k=1}^K$, $\mathcal{Y} := \{\mathbf{y}_k\}_{k=1}^K$
- Non-linear inverse map (given the manifold)

$$(\mathrm{P1}) \qquad \min_{\mathbf{x}} \ \|\mathbf{y} - \mathbf{\Phi x}\|^2 \qquad \text{s.to} \quad \mathbf{x} \in \mathcal{M}$$

$$\mathbf{x}^* = f_{\mathcal{M}}(\mathbf{\Phi}; \mathbf{y})$$

- Given $(\mathcal{X}, \mathcal{Y})$ design a neural net that approximates the inverse map $f_{\mathcal{M}}$

Prior art

- Sparse coding (l_1-regularization)
  - Compressed sensing (CS) for sparse signals [Donoho-Elad'03], [Candes-Tao'04]
  - Stable recovery guarantees with ISTA, FISTA [Beck-Teboulle'09] $M = O(s \log(N))$
- LISTA automates ISTA, shrinkage with single-layer FC layer [Gregor-LeCun'10]
- Data-driven regularization enhances robustness to noise
- Natural image restoration (local)
  - Image super-resolution; perceptual loss [Johnson et al'16], GANs [Leding et al'16]
  - Image de-blurring; CNN [Xu et al'16]; [Schuler et al'14]
- Medical image reconstruction (global)
  - MRI; denoising auto-encoders [Majumdar'15], Automap [Zhu et al'17]
  - CT; RED-CNN, U-net [Chen et al'17]
- The main success has been on improving the speed; training entails many parameters, and no guarantees for data fidelity (post-processing)

Cont'd

- Learning priors by unrolling and modifying the optimization iterations
  - Unrolled optimization with deep CNN priors [Diamond et al'18]
  - ADMM-net; CS-MRI; learns filters and nonlinearities (iterative) [Sun et al'16]
  - LDAMP: Learned denoising based approximate message passing [Metzler et al'17]
  - Learned primal-dual reconstruction, forward and backward model [Adler et al'17]

- Inference; given a pre-trained generative model
  - Risk minimization based on generator representation [Bora et al'17], [Paul et al'17]
  - Reconstruction guarantees; Iterative and time intensive inference; no training

- High training overhead for multiple iterations (non-recurrent); pixel-wise costs

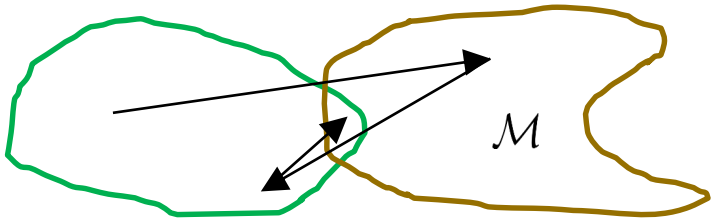
- Novelty: design and analyze architectures with low training overhead
  - Offer fast & robust inference
  - Against noise and hallucination

GANCS

$$\mathbf{y} = \mathbf{\Phi x} \qquad \mathbf{x} \in \mathcal{M}$$

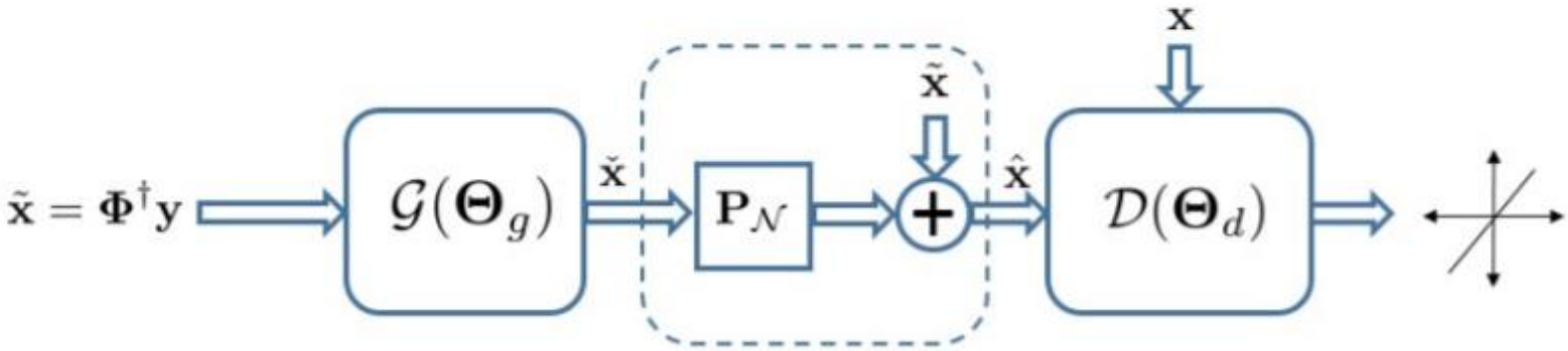
- Alternating projection (noiseless scenario)

$$\mathbf{x}[k+1] = \mathcal{P}_{\mathcal{M}}\left\{ \mathbf{\Phi}^{\dagger}\mathbf{y} + (\mathbf{I} - \mathbf{\Phi}^{\dagger}\mathbf{\Phi})\mathbf{x}[k] \right\}$$



data-consistent images

- Network architecture



Mixture loss

- LSGAN + \ell_1/\ell_2 loss

$$\text{(P1.1)} \qquad \min_{\boldsymbol{\Theta}_d} \mathbb{E}_{\mathbf{x}}\left[\left(1-\mathcal{D}(\mathbf{x};\boldsymbol{\Theta}_d)\right)^2\right] + \mathbb{E}_{\mathbf{y}}\left[\left(\mathcal{D}(\mathcal{G}(\boldsymbol{\Phi}^\dagger\mathbf{y};\boldsymbol{\Theta}_g);\boldsymbol{\Theta}_d)\right)^2\right]$$

$$\text{(P1.2)} \qquad \min_{\boldsymbol{\Theta}_g} \mathbb{E}_{\mathbf{y}}\left[\left\|\mathbf{y}-\boldsymbol{\Phi}\mathcal{G}(\boldsymbol{\Phi}^\dagger\mathbf{y};\boldsymbol{\Theta}_g)\right\|^2\right] + \eta\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\left\|\mathbf{x}-\mathcal{G}(\boldsymbol{\Phi}^\dagger\mathbf{y};\boldsymbol{\Theta}_g)\right\|_1\right] + \lambda\mathbb{E}_{\mathbf{y}}\left[\left(1-\mathcal{D}(\mathcal{G}(\boldsymbol{\Phi}^\dagger\mathbf{y};\boldsymbol{\Theta}_g);\boldsymbol{\Theta}_d)\right)^2\right]$$

- GAN hallucination
  - Data consistency
  - Pixel-wise cost (ℓ_1/ℓ_2) avoids high-frequency noise, especially in low sample complexity regimes

GAN equilibrium

$$\mathbf{x} \sim p_d(\mathbf{x})$$
$$\hat{\mathbf{x}} \sim \hat{p}_d(\mathbf{x})$$

Proposition 1. If G and D have infinite capacity, then for the given generator net G, the optimal D admits

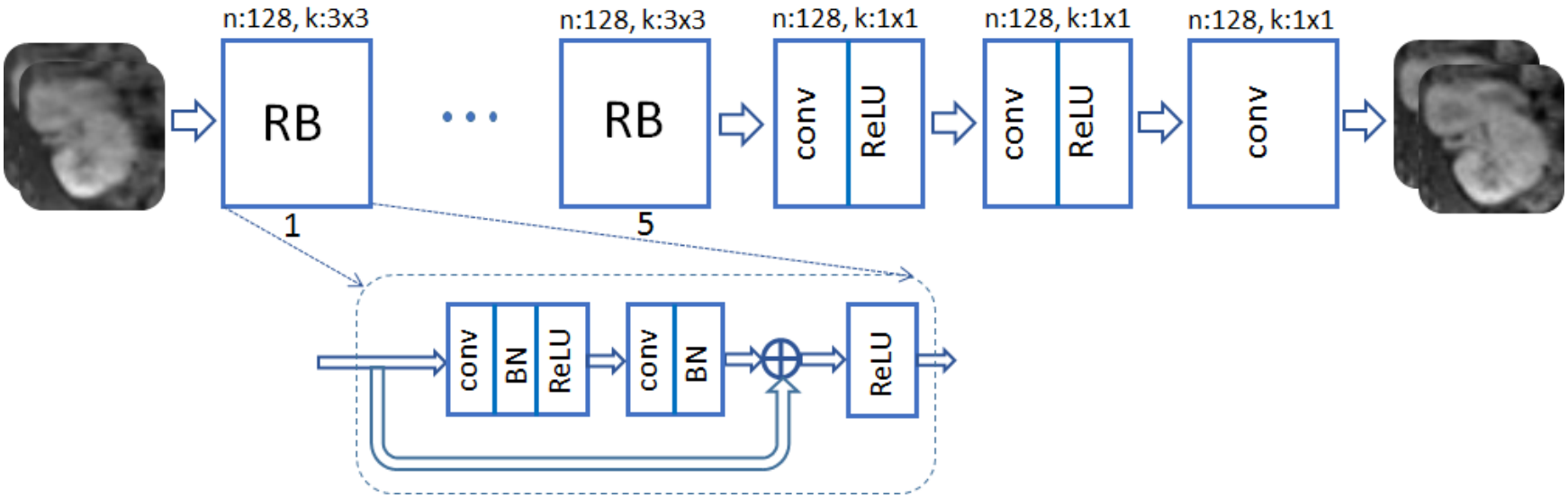
$$\mathcal{D}^*(\boldsymbol{\Theta}, \mathbf{x}) = p_d(\mathbf{x})/(p_d(\mathbf{x}) + \hat{p}_d(\mathbf{x}))$$

Also, the equilibrium of the game is achieved when

$$p_d(\mathbf{x}) = \hat{p}_d(\mathbf{x})$$

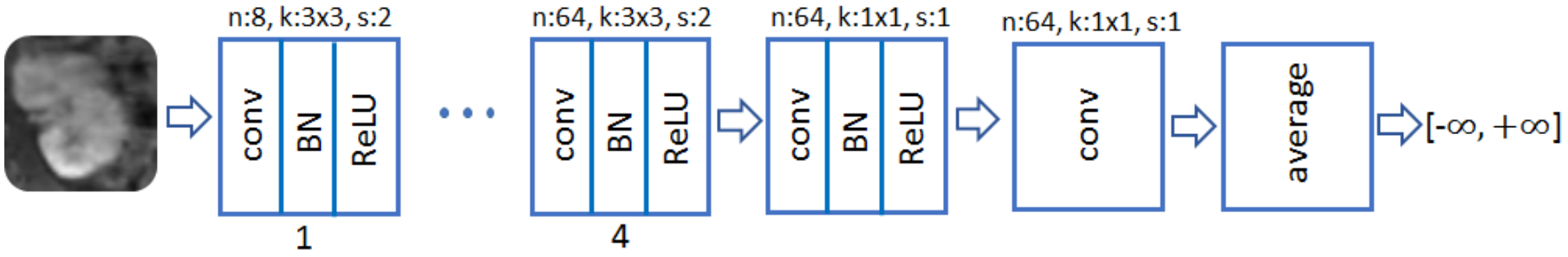
- Solving (P1.1)-(P1.2) yields minimizing the Pearson-χ^2 divergence
- At equilibrium $\mathcal{D}^*(\boldsymbol{\Theta}, \mathbf{x}) = 1/2$

Denoiser net (G)

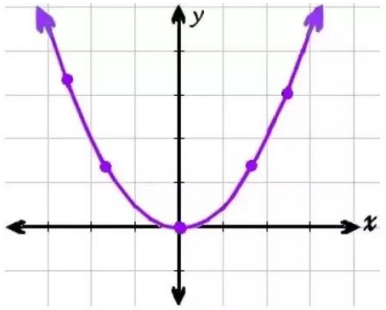


- No pooling, 128 feature maps, 3x3 kernels
- Complex-valued images considered as real and imaginary channels

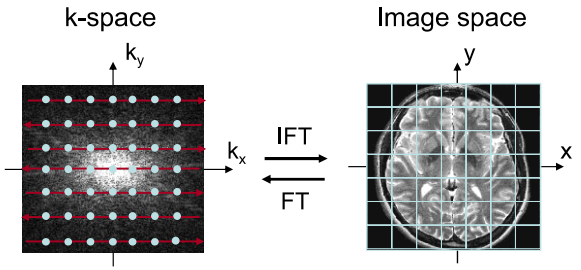
Discriminator net (D)



- 8 CNN layers, no pooling, no soft-max (LSGAN)
- Input: magnitude image



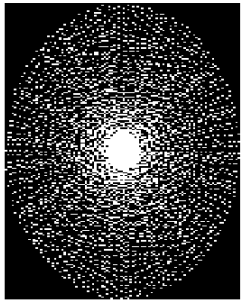
Experiments



- MRI acquisition model

$$y_{i,j}^{(n)} = [\mathcal{F}(\mathbf{X}_n)]_{i,j} + v_{i,j}^{(n)}, \quad (i,j) \in \Omega$$

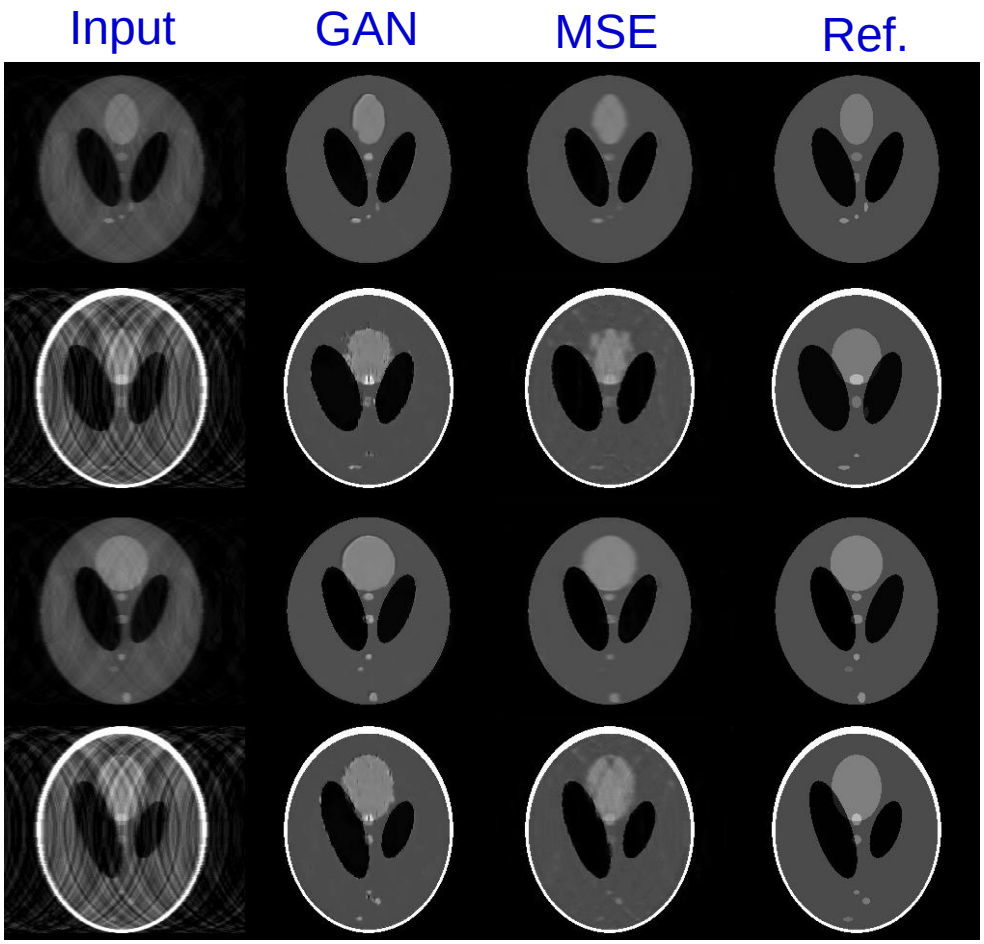
- Synthetic Shepp-Logan phantom dataset
  - 1k train, 256 x 256 pixel resolution magnitude images
  - 5-fold variable density undersampling trajectory



- T1-weighted contrast-enhanced abdominal MRI
  - 350 pediatric patients, 336 for train, and 14 for test
  - 192 axial image slices of 256 x 128 pixels
  - Gold-standard is the fully-sampled one aggregated over time (2 mins)
  - 5-fold variable density undersampling trajectory with radial-view ordering

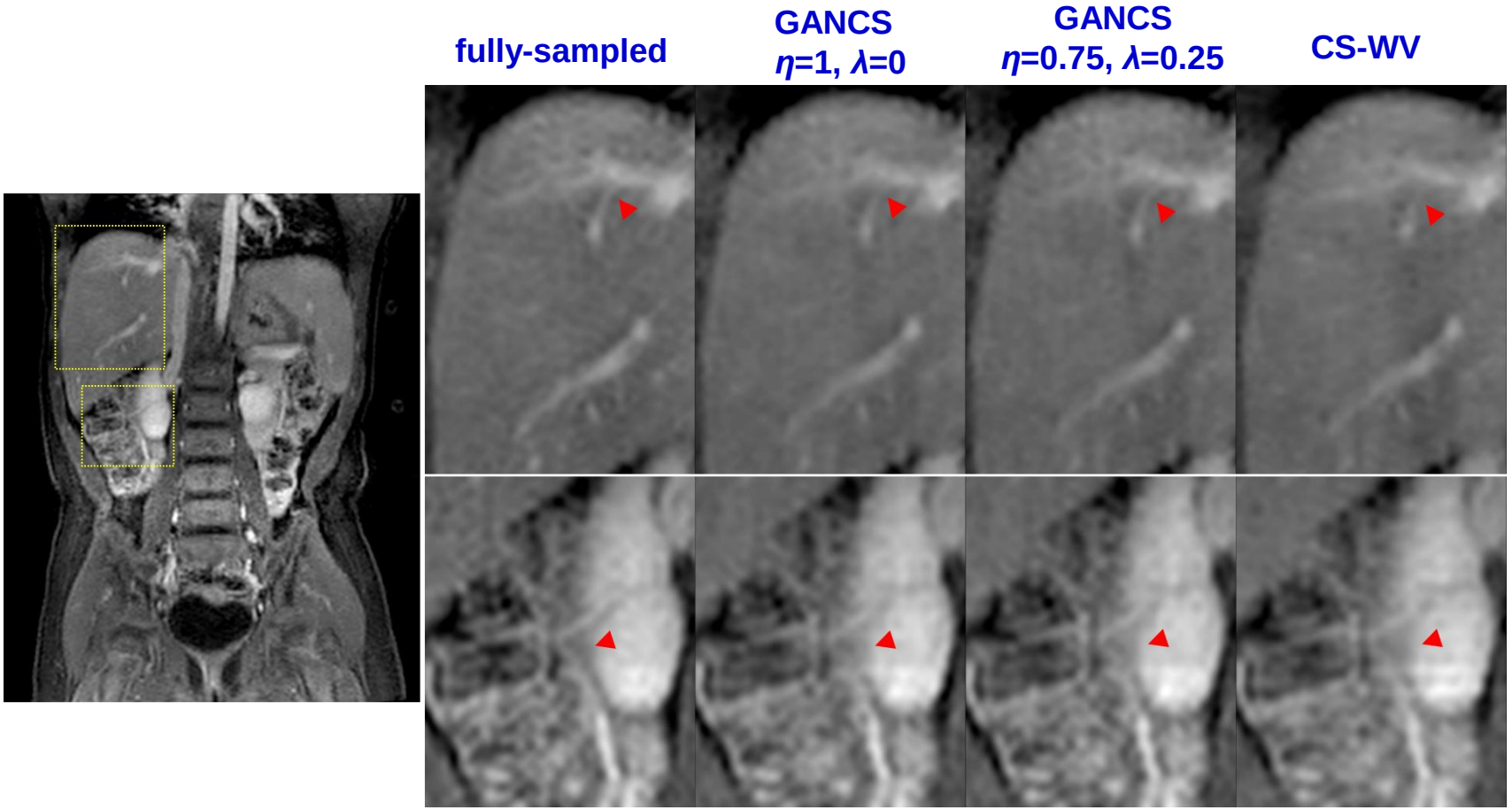
- TensorFlow, NVIDIA Titan X Pascal GPU with 12GB RAM

Phantom training



- Sharper images than pairwise MSE training

Abdominal MRI



- GANCS reveals tiny liver vessels and sharper boundaries for kidney

Quantitative metrics

Quantitative metrics (single copy, and 5-RBs)

Scheme	GANCS (ℓ_1) $\eta = 0.975, \lambda = 0.025$	GANCS (ℓ_1) $\eta = 1, \lambda = 0$	GANCS (ℓ_2) $\eta = 1, \lambda = 0$	GANCS (ℓ_1) $\eta = 0.75, \lambda = 0.25$	CS-WV	CS-TV	ZF
5-fold acceleration							
SNR	20.56	21.40	21.21	18.03	18.22	16.75	13.66
SSIM	0.82	0.84	0.84	0.76	0.80	0.77	0.64
Time	0.03	0.03	0.03	0.03	12.28	1.81	9×10^{-4}

> 100 times faster

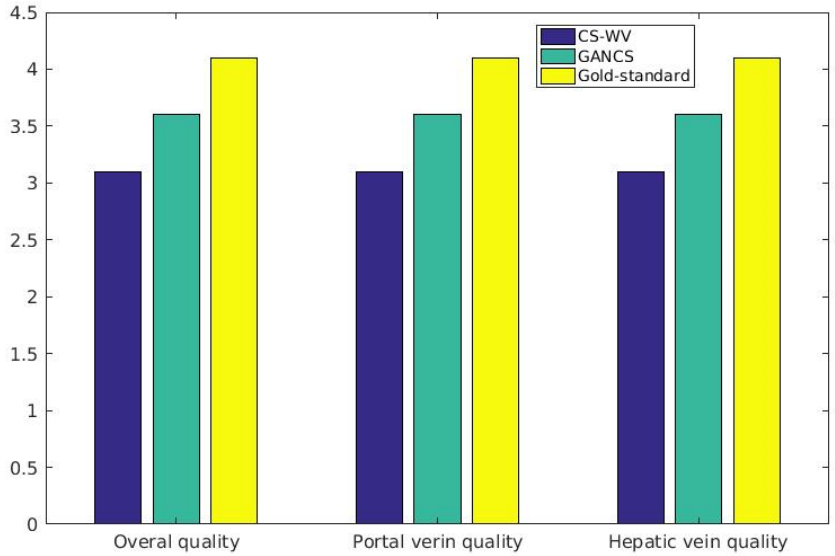
proposed

- CS-MRI runs using the optimized BART toolbox

Diagnostic quality assessment

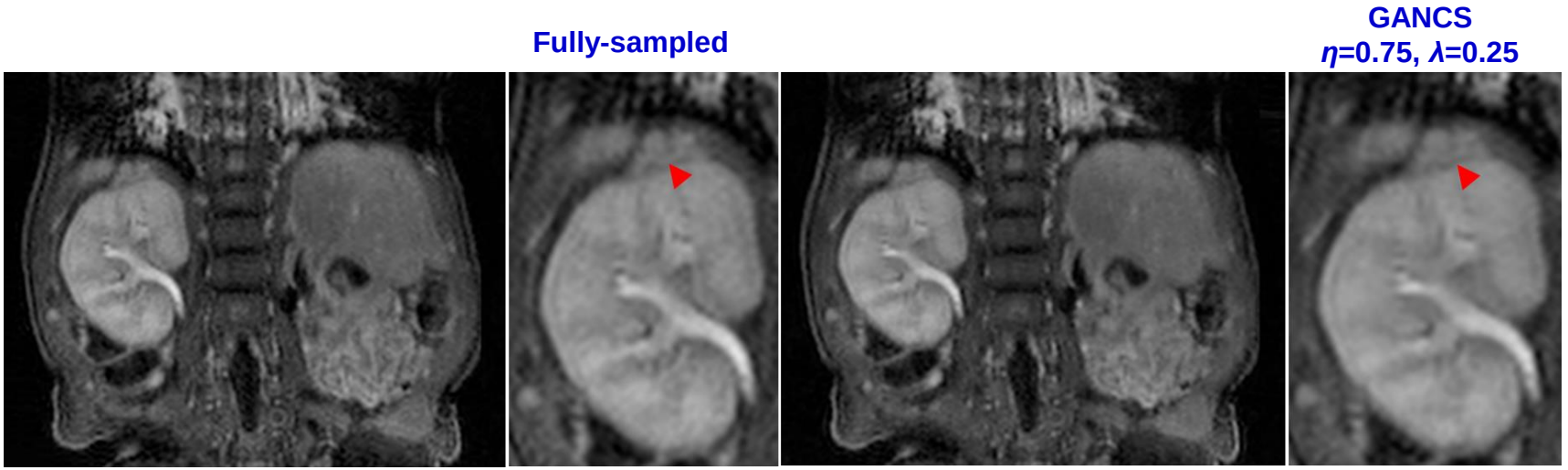
- Two pediatric radiologists independently rate the images

Score	overall quality	portal veins	hepatic veins
nondiagnostic (1)	no structures assessed	PV blurred	RHV blurred
limited (2)	limited assessment of several structures	1st-order branches of PV blurred	1st-order branches of RHV blurred
diagnostic (3)	all but $1-2$ structures assessed	sharp first order branches of PV	sharp first order branches of RHV
good (4)	all structures assessed	sharp 2nd-order branches of PV	sharp 2nd-order branches of RHV
excellent (5)	sharp delineation of all structures	braches seen to within 1cm of periphery	braches seen to within 1cm of periphery

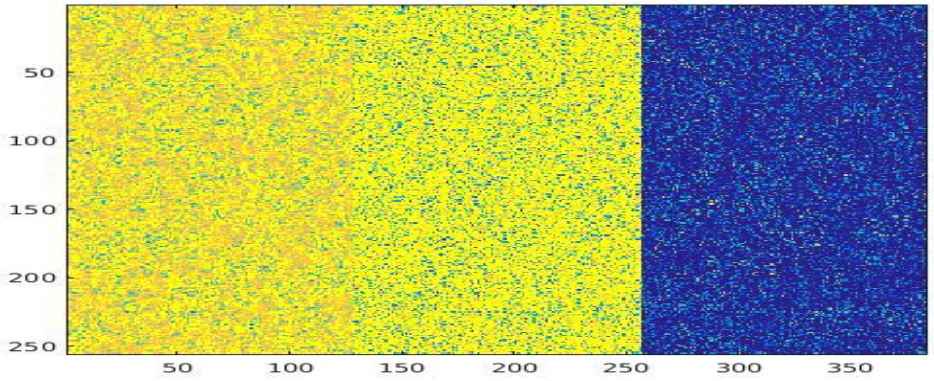


- No sign of hallucination observed

Generalization

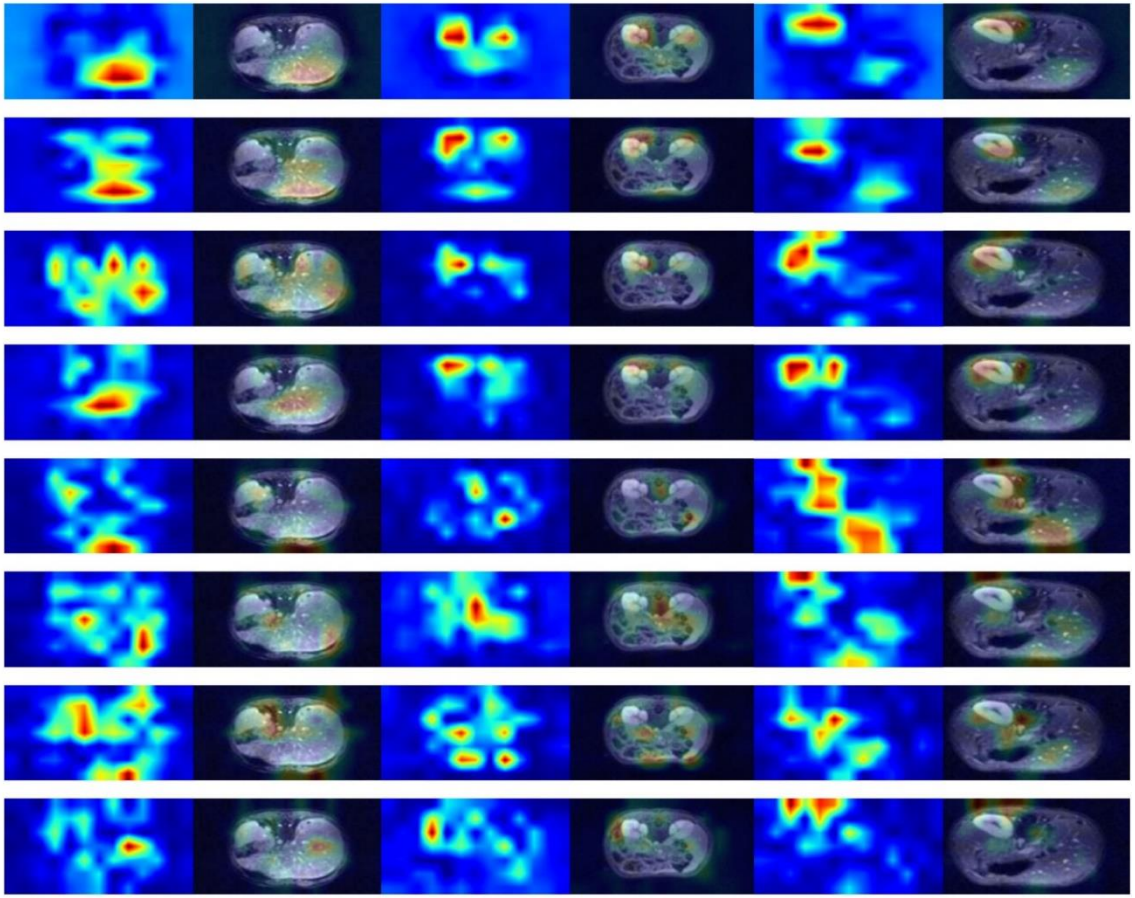


- Memorization tested with Gaussian random inputs



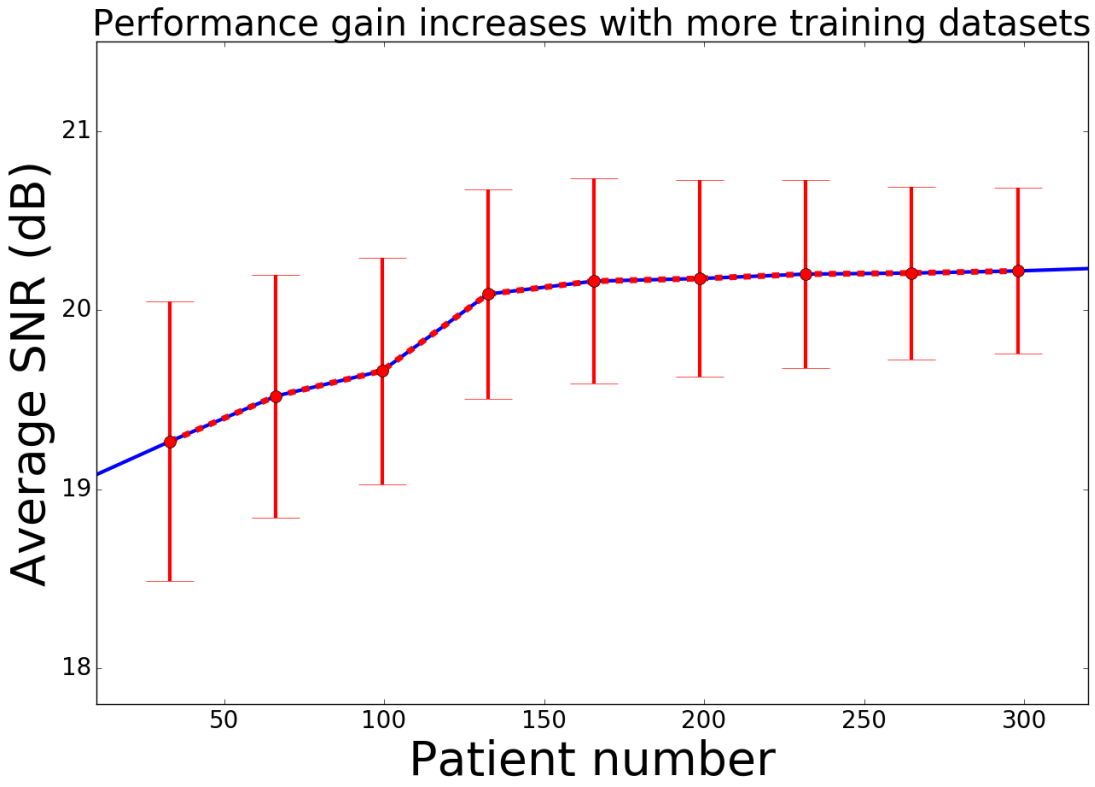
No structures picked up!

Saliency maps



- Picks up the regions that are more susceptible to artifacts

Patient count



- 150 patients suffices for training with acceptable inference SNR

Caveats

- Noisy observations $\mathbf{y} = \mathbf{\Phi x} + \mathbf{v}$
- The exact affine projection is costly e.g., for image super-resolution
- Training deep nets is resource intensive (1-2 days)
- Training deep nets also may lead to overfitting and memorization that causes hallucination

Proximal gradient iterations

- Regularized LS

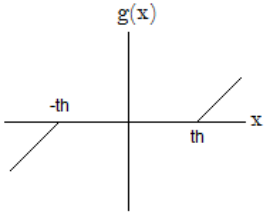
$$\min_{\mathbf{x}} \ \left\|\mathbf{y}-\mathbf{\Phi x}\right\|^2 + \psi(\mathbf{x};\mathbf{\Theta})$$

- For instance, if $\mathbf{x} \in \mathcal{M}$, then $\psi(\mathbf{x};\mathbf{\Theta}) := \mathbf{1}_{\{\mathbf{x}\in\mathcal{M}\}}$

- Proximal gradient iterations

$$\mathbf{x}[k+1] = \mathcal{P}_{\psi}\Big\{\mathbf{x}[k] + \alpha\mathbf{\Phi}^{\mathsf{H}}(\mathbf{y}-\mathbf{\Phi x}[k])\Big\}$$

- Sparsity regularizer leads to iterative soft-thresholding (ISTA)



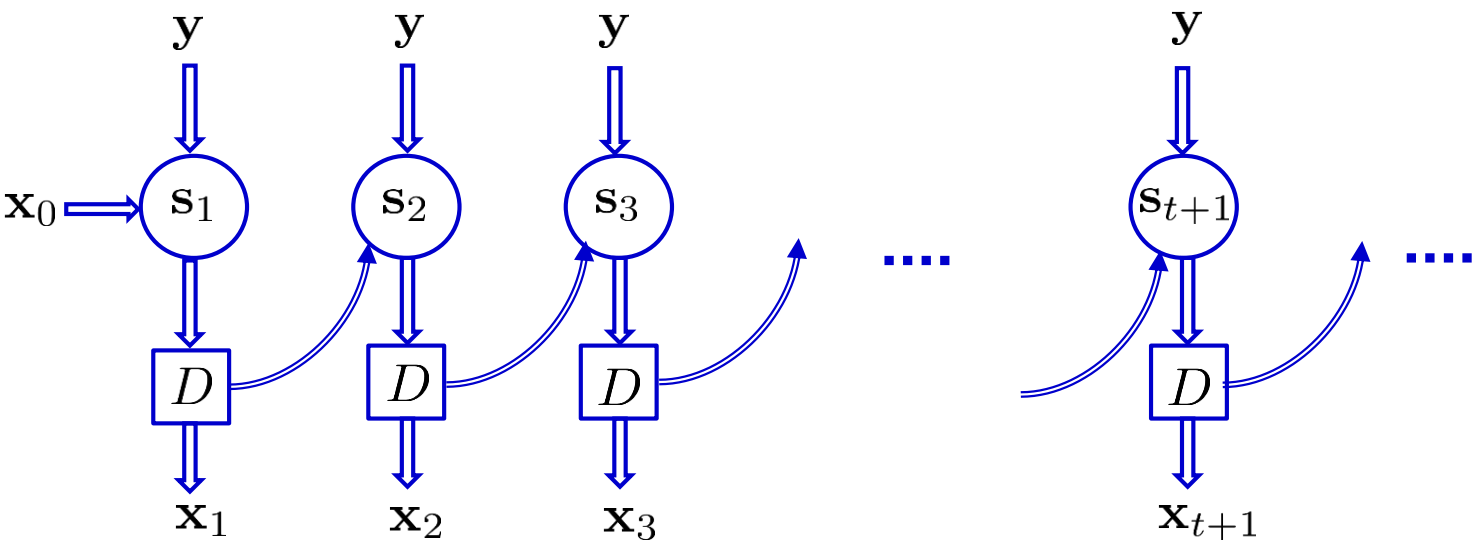
Recurrent proximal learning

$$D(\cdot) := \mathcal{P}_{\Psi}(\cdot)$$

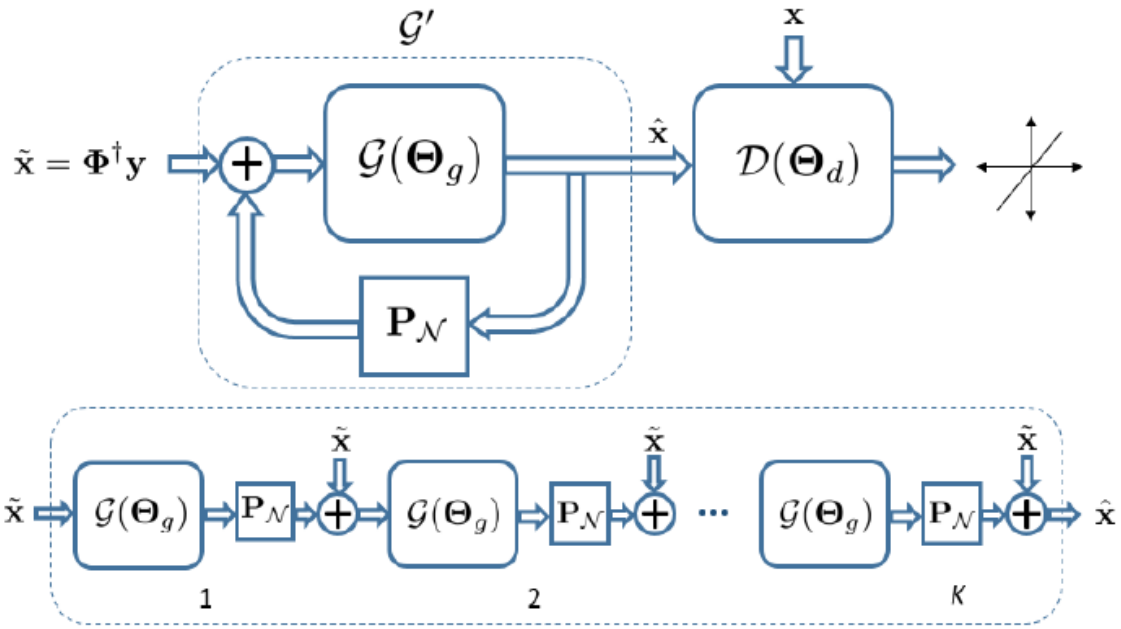
$$\mathbf{x}_{t+1} = D\Big(\underbrace{\alpha\boldsymbol{\Phi}^{\mathsf{H}}\mathbf{y} + (\mathbf{I} - \alpha\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi})\mathbf{x}_t}_{g(\mathbf{x}_t;\mathbf{y})}\Big)$$

- State-space evolution model

$$\mathbf{s}_{t+1} = g(\mathbf{x}_t;\mathbf{y})$$
$$\mathbf{x}_{t+1} = D(\mathbf{s}_{t+1})$$



Recurrent GANCS



Truncated
K iterations

- Training cost

$$\text{(P1.1)} \quad \min_{\mathbf{\Theta}} \mathbb{E}_{\mathbf{x}}\left[\left(1-\mathcal{D}(\mathbf{x};\mathbf{\Theta}_d)\right)^2\right]+\mathbb{E}_{\mathbf{y}}\left[\left(\mathcal{D}(\hat{\mathbf{x}};\mathbf{\Theta}_d)\right)^2\right]$$

$$\text{(P1.2)} \quad \min_{\mathbf{\Theta}} \mathbb{E}_{\mathbf{y}}\left[\sum_{k=1}^{K}\left\|\mathbf{y}-\mathbf{\Phi}\check{\mathbf{x}}_k\right\|^2\right]+\lambda\mathbb{E}_{\mathbf{y}}\left[\left(1-\mathcal{D}(\hat{\mathbf{x}};\mathbf{\Theta}_d)\right)^2\right]$$
$$+\eta\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_{1,2}\right] \qquad (3)$$

Empirical validation

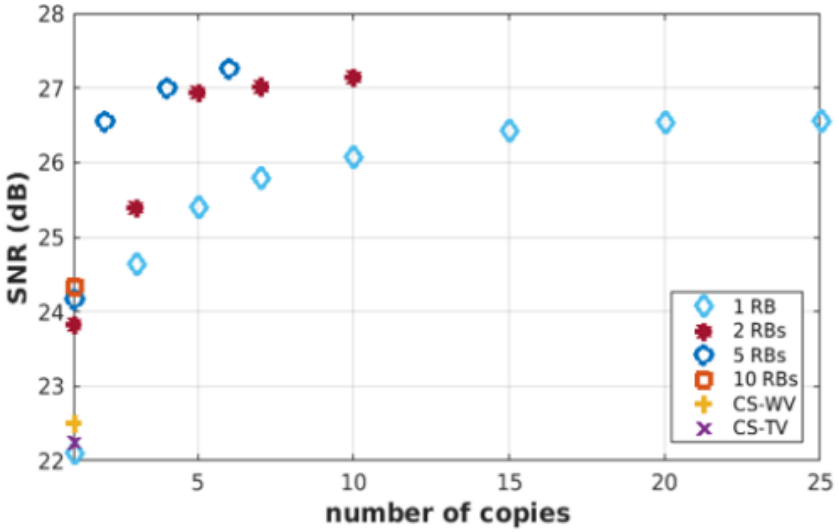
Q1. proper combination of iterations and denoiser net size?

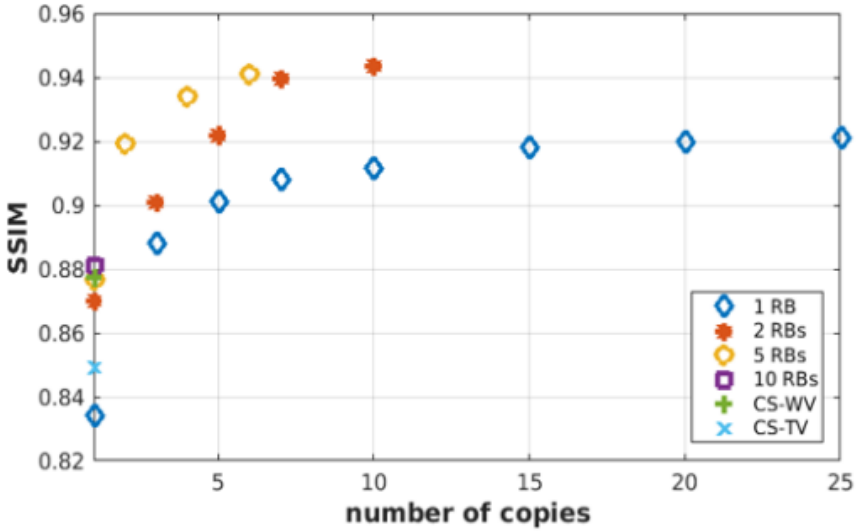
Q2. trade-off between PSNR/SSIM and inference/training complexity?

Q3. performance compared with conventional sparse coding?

- T1-weighted contrast-enhanced abdominal MRI
  - 350 pediatric patients, 336 for train, and 14 for test
  - 192 axial image slices of 256 x 128 pixels
  - Gold-standard is the fully-sampled one aggregated over time (2 mins)
  - 5-fold variable density undersampling trajectory with radial-view ordering

SNR/SSIM

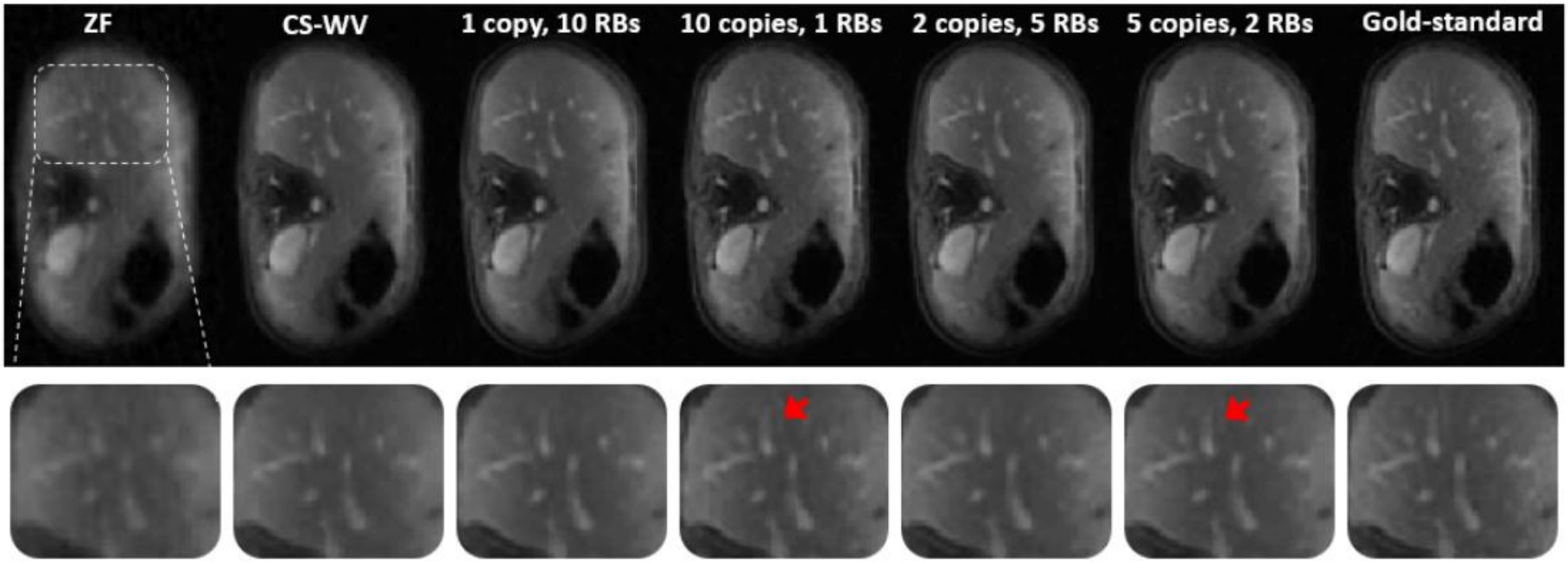




- For a single iteration depth does not matter after some point
- Significant SNR/SSIM gain when using more than a single copy

copies	RBs	inference time (sec)	SNR (dB), independent	SSIM, independent	SNR (dB), shared	SSIM, shared
10	1	0.04	27.03	0.923	26.07	0.9117
5	2	0.10	27.01	0.9258	26.94	0.9221
2	5	0.12	28.14	0.944	26.55	0.9194
1	10	0.0522	24.33	0.8810	24.33	0.8810
CS-TV	n/a	1.30	22.20	0.82	22.20	0.82
CS-WV	n/a	1.16	22.51	0.86	22.51	0.86

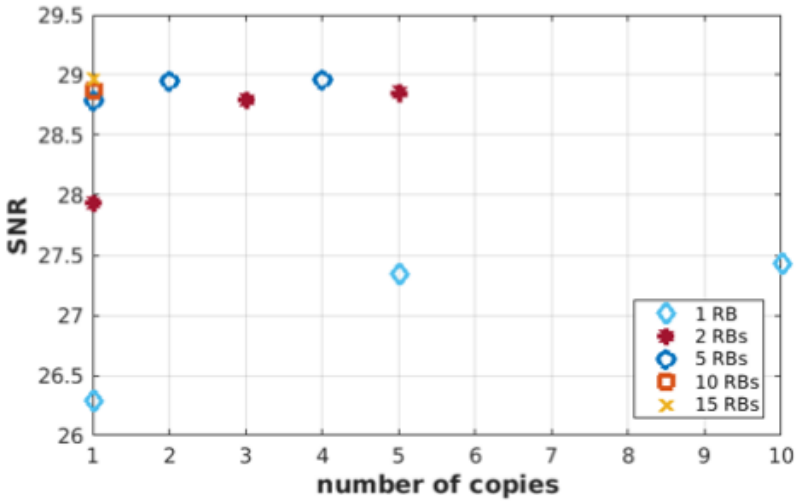
Reconstructed images

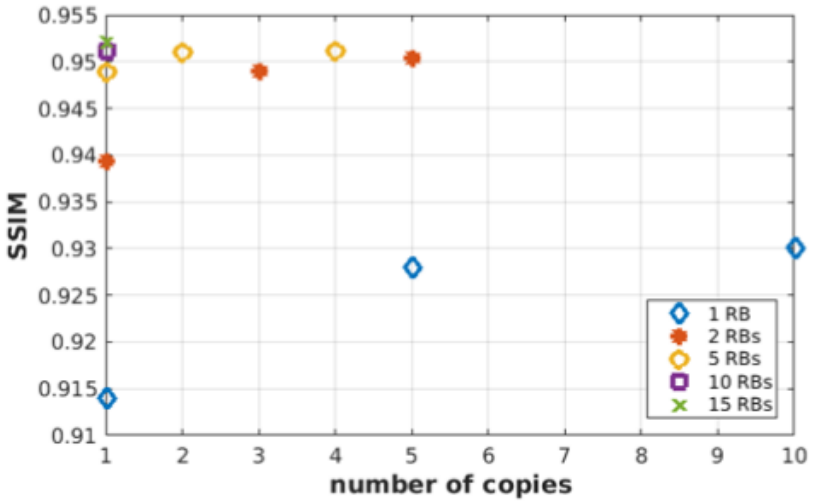


- Train time: 10 copies,1RB needs 2-3 h; 1 copy, 10RBs 10-12h
- Better to use 1-2 RBs with 10-15 iterations!

Image super-resolution

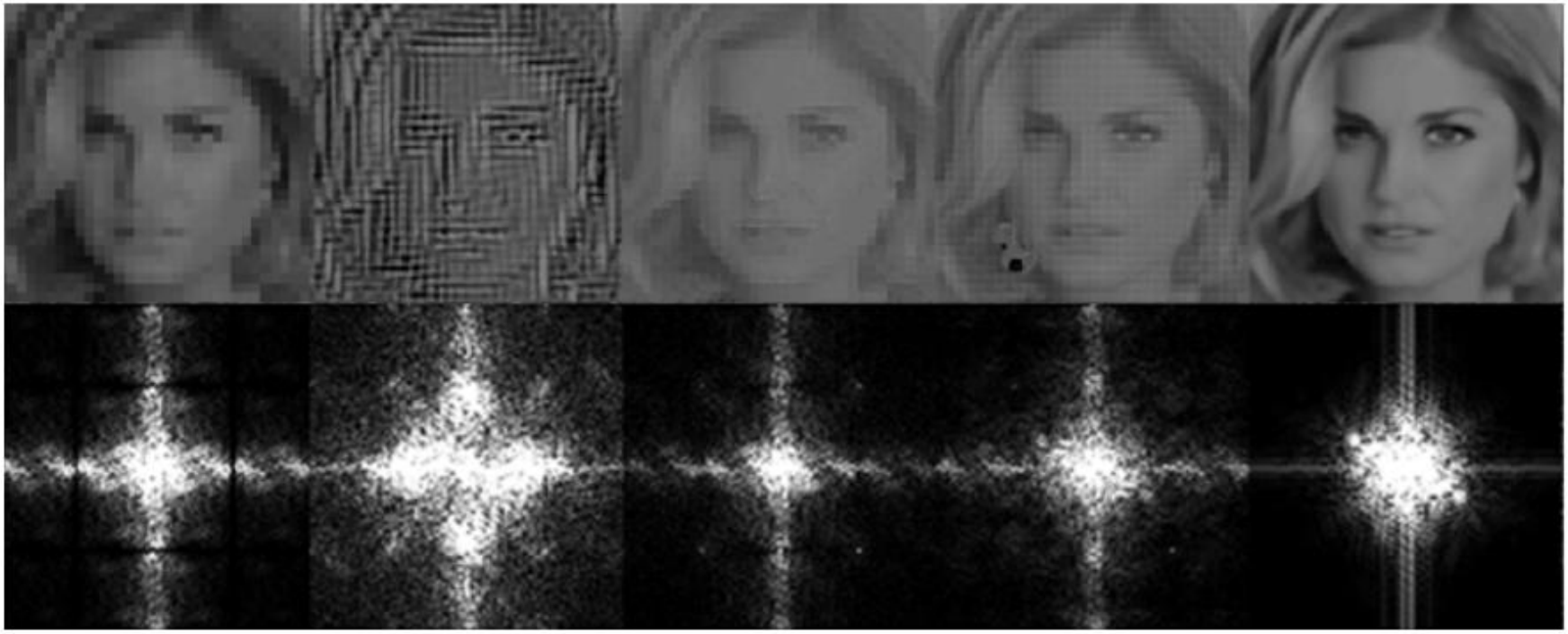
- Image super-resolution (local),
  - CelebA Face dataset 128x128, 10k images for train, and 2k for test
  - 4x4 constant kernel with stride 4
  - Independent weights are chosen





- Proximal learning needs a deeper net rather than more iterations

Independent copies



- 4 independent copies & 5 RBs
- Overall process alternates between image sharpening and smoothing

Convergence

$$\mathbf{s}_{t+1} = g(\mathbf{x}_t; \mathbf{y})$$
$$\mathbf{x}_{t+1} = \mathcal{D}(\mathbf{s}_{t+1})$$

Proposition 2. For a single-layer neural net with ReLU, i.e., $D(\mathbf{z}) = \sigma(\mathbf{W}\mathbf{z})$, suppose there exists a fixed-point $\mathbf{x}_0 = D(\mathbf{x}_0)$. Define $\mathbf{D}_{\mathbf{x}} := \mathrm{diag}(\mathrm{D}(\mathbf{x}))$, $\boldsymbol{\Delta}_{\boldsymbol{\delta}} := \mathbf{D}_{\mathbf{x}_0+\boldsymbol{\delta}} - \mathbf{D}_{\mathbf{x}_0}$, $\mathbf{M} := \mathbf{W}(\mathbf{I} - \alpha\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi})$, and assume the following holds

$$\text{(i)}\ \|(1-\frac{1}{\beta})\mathbf{I} + \frac{1}{\beta}\mathbf{D}_{\mathbf{x}_0}\mathbf{M}\| \le \eta_1$$

$$\text{(ii)}\ \frac{1}{\beta}\|\Delta_{\boldsymbol{\delta}}(\mathbf{W}\mathbf{x}_0 + \mathbf{M}\boldsymbol{\delta})\| \le \eta_2\|\boldsymbol{\delta}\|, \quad \forall \boldsymbol{\delta} \in \mathcal{S}_{\mathrm{n}}$$

For some $\eta_1, \eta_2 > 0$, with the step size $\alpha > 0$ and $\beta \in (0,1)$. If $\eta_1 + \eta_2 \le 1$, the iterates $\{\mathbf{x}_t, \mathbf{s}_t\}$ converge to a fixed point.

- Low-dimensionality taken into account $\mathbf{x}_0 = D(\mathbf{x}_0)$

Implications

- Random Gaussian ReLU with bias $b \sim \mathcal{N}(0, \epsilon^2)$

Lemma 1. For Gaussian ReLU, the mask is Lipschitz continuous w.h.p

$$\|\mathbf{1}_{\{\mathbf{W}\mathbf{z}+\epsilon n \geq 0\}} - \mathbf{1}_{\{\mathbf{W}\mathbf{z}'+\epsilon n \geq 0\}}\| \leq \frac{C}{\epsilon}\|\mathbf{W}\|\|\mathbf{z}-\mathbf{z}'\|$$

- For a small perturbation $\|\boldsymbol{\delta}\| \ll$

$$\eta_2 = (C/\epsilon) \max_{\boldsymbol{\delta} \in \mathcal{S}_n} \frac{|\langle \mathbf{W}(\mathbf{I} - \alpha \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{H}})\boldsymbol{\delta}, \mathbf{W}\mathbf{x}_0 \rangle|}{\|\boldsymbol{\delta}\|}$$

- Deviation from the tangent space

$$\langle \boldsymbol{\delta}, (\mathbf{I} - \alpha \boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi})\mathbf{W}^{\mathsf{H}}\mathbf{W}\mathbf{x}_0 \rangle = 0$$

Multi-layer net

$$\mathbf{s}_{t+1} = g(\mathbf{x}_t; \mathbf{y})$$
$$\mathbf{x}_{t+1} = \mathcal{D}(\mathbf{s}_{t+1})$$

Proposition 3. For a *L*-layer neural net with $D(\mathbf{z}) = \sigma(\mathbf{W}_L\sigma(\mathbf{W}_{L-1}\ldots\sigma(\mathbf{W}_1\mathbf{z})))$ suppose there exists a fixed-point $\mathbf{x}_0 = D(\mathbf{x}_0)$. Define feature maps $\mathbf{h}_\ell(\boldsymbol{\delta}) = \mathbf{W}_\ell\sigma(\mathbf{W}_{\ell-1}\mathbf{h}_{\ell-1}(\boldsymbol{\delta}))$ where $\mathbf{x}_0 = D(\mathbf{x}_0)$, and $\boldsymbol{\Delta}_{\mathbf{h}_\ell(\boldsymbol{\delta})} := \mathbf{D}_{\mathbf{h}_\ell(\mathbf{x}_0)+\boldsymbol{\delta}_\ell} - \mathbf{D}_{\mathbf{h}_\ell(\mathbf{x}_0)}$. Then if

$$\text{(i)}\ \|(1-\frac{1}{\beta})\mathbf{I} + \frac{1}{\beta}\mathbf{D}_{\mathbf{h}_\ell(\mathbf{x}_0)}\mathbf{W}_\ell\mathbf{M}_\ell\| \le \eta_1^{(\ell)}, \quad \ell \ge 1$$

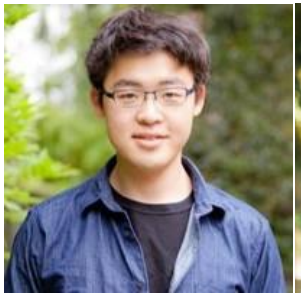
$$\text{(ii)}\ \|\Delta_{\mathbf{h}_\ell(\boldsymbol{\delta}_\ell)}\mathbf{W}_\ell(\mathbf{h}_{\ell-1}(\mathbf{x}_0)+\boldsymbol{\delta}_\ell)\| \le \eta_2^{(\ell)}\|\boldsymbol{\delta}_\ell\|, \quad \forall\boldsymbol{\delta}_\ell \in \mathcal{S}_{\mathrm{n}}^{(\ell)}, \quad \ell \ge 1$$

where $\mathbf{M}_\ell := \mathbf{I}-\alpha\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi}$, $\ell = 1$, $\mathbf{M}_\ell = \mathbf{I}$, $\ell \ge 2$, $\mathcal{S}_n^{(1)} := \{(\mathbf{I}-\alpha\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi})\boldsymbol{\delta} \ : \ \boldsymbol{\delta} \in \mathbb{R}^n\}$, and $\mathcal{S}_n^{(\ell)} := \{\boldsymbol{\delta} \ : \ \|\boldsymbol{\delta}\|_0 \le p_\ell n\}, \ell \ge 2$, and if for some $\alpha > 0$ and $\beta \in (0,1)$, it satisfies $\eta := \prod_{\ell=1}^{L}(\eta_1^{(\ell)} + \eta_2^{(\ell)}) \le 1$, the iterations $\{\mathbf{x}_t, \mathbf{s}_t\}$ converge to a fixed point.

Concluding summary

- A novel data-driven CS framework
  - Learning proximal from historical data
  - Mixture of adversarial (GAN) and pixel-wise costs
- ResNet for the denoiser (G) and a deep CNN used for the discriminator
- Recurrent implementation leads to low training overhead
  - The physical model is taken into account
  - Avoids overfitting that improves the generalization
- Evaluations on abdominal MRI scans of pediatric patients
  - GANCS achieves Higher diagnostic score that CS-MRI
  - RGANCS leads to 2dB better SNR (SSIM) than GANCS
  - 100x faster inference
- Proximal learning for (local) MRI task with 1-2 RBs (several iterations)
- While for (global) SR use a deep ResNet (couple of iterations)

Acknowledgments



Dr. Enhao Gong
EE

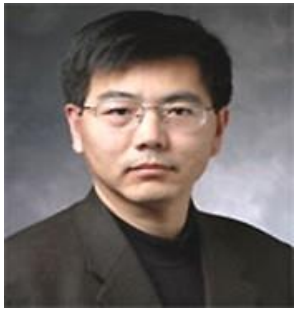
Dr. Joseph Cheng
Radiology/EE

Dr. Shreyas Vasawanala
Radiology

Dr. Greg Zaharchuk
Radiology

Dr. David Donoho
Statistics

Dr. John Pauly
EE



Dr. Lei Xing
Medical physics/EE

Dr. Hatef Monajemi
Statistics

Dr. Vardan Papayan
Statistics

Dr. Marcus Alley
Radiology

Grants: NIH T32121940, NIH R01EB009690

Further details

[1] Morteza Mardani, Enhao Gong, Joseph Cheng, Shreays Vasanawala, Greg Zaharcuk, Lei Xing, and John Pauly, ``Deep generative adversarial networks for compressive sensing (GANCS) automates MRI,'' *arXiv preprint arXiv:1706.00051*, May 2017.

[2] Morteza Mardani, Hatef Monajemi, Vardan Papyan, Shreyas Vasanawala, David Donoho, and John Pauly, ``Recurrent generative adversarial networks for proximal learning and compressive image recovery,'' *arXiv preprint arXiv:1711.10046*, November 2017.

TensorFlow code available at: https://github.com/gongenhao/GANCS

Email: Morteza@Stanford.edu

Thank you!

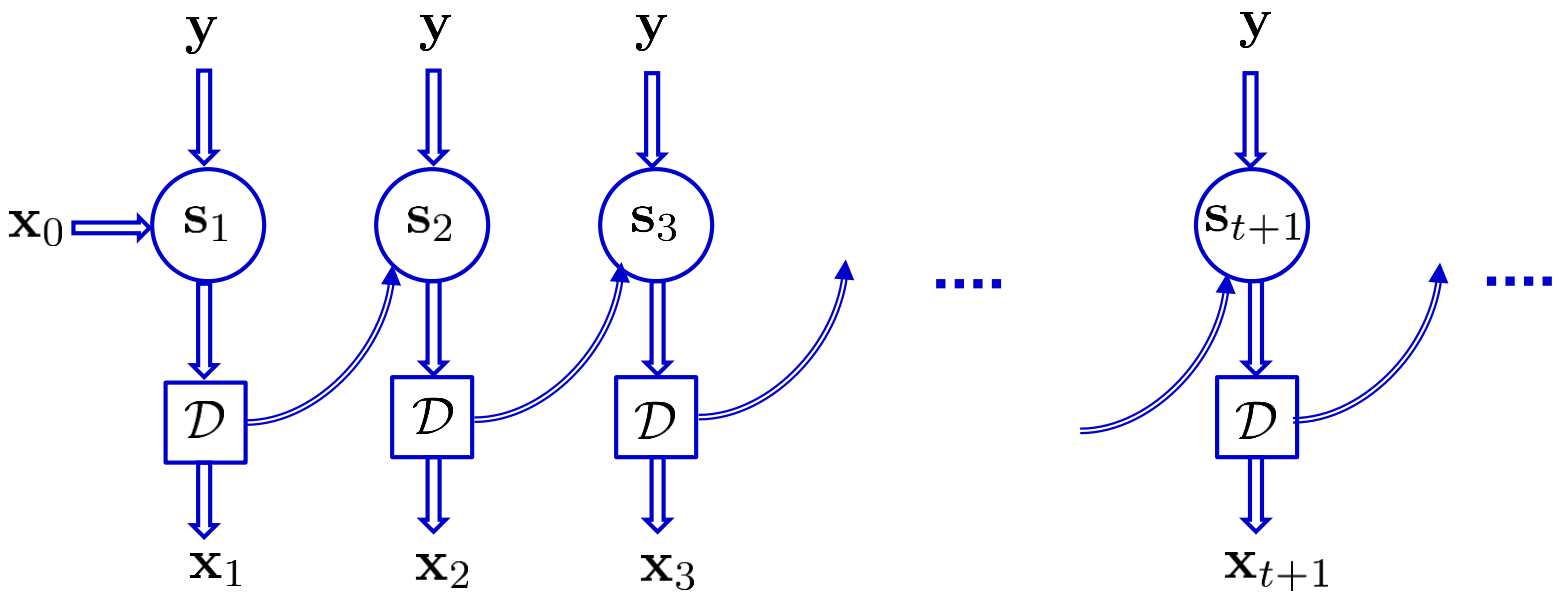
Questions

Q1. How is GANCS compared with the CS-MRI and pixel-wise training?

Q2. How much inference speed up one can achieve relative to CS-MRI?

Q3. What MR image features derive the network to learn the manifold and remove the aliasing artifacts?

Q4. How many samples/patients are needed to achieve an acceptable image quality?


y
y
y
y
$\mathbf{x}_0$
$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_3$
$\mathbf{s}_{t+1}$
$\mathcal{D}$
$\mathcal{D}$
$\mathcal{D}$
$\mathcal{D}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_{t+1}$