

Smale's 18th Problem and the Barriers of Deep Learning

Matthew Colbrook
University of Cambridge

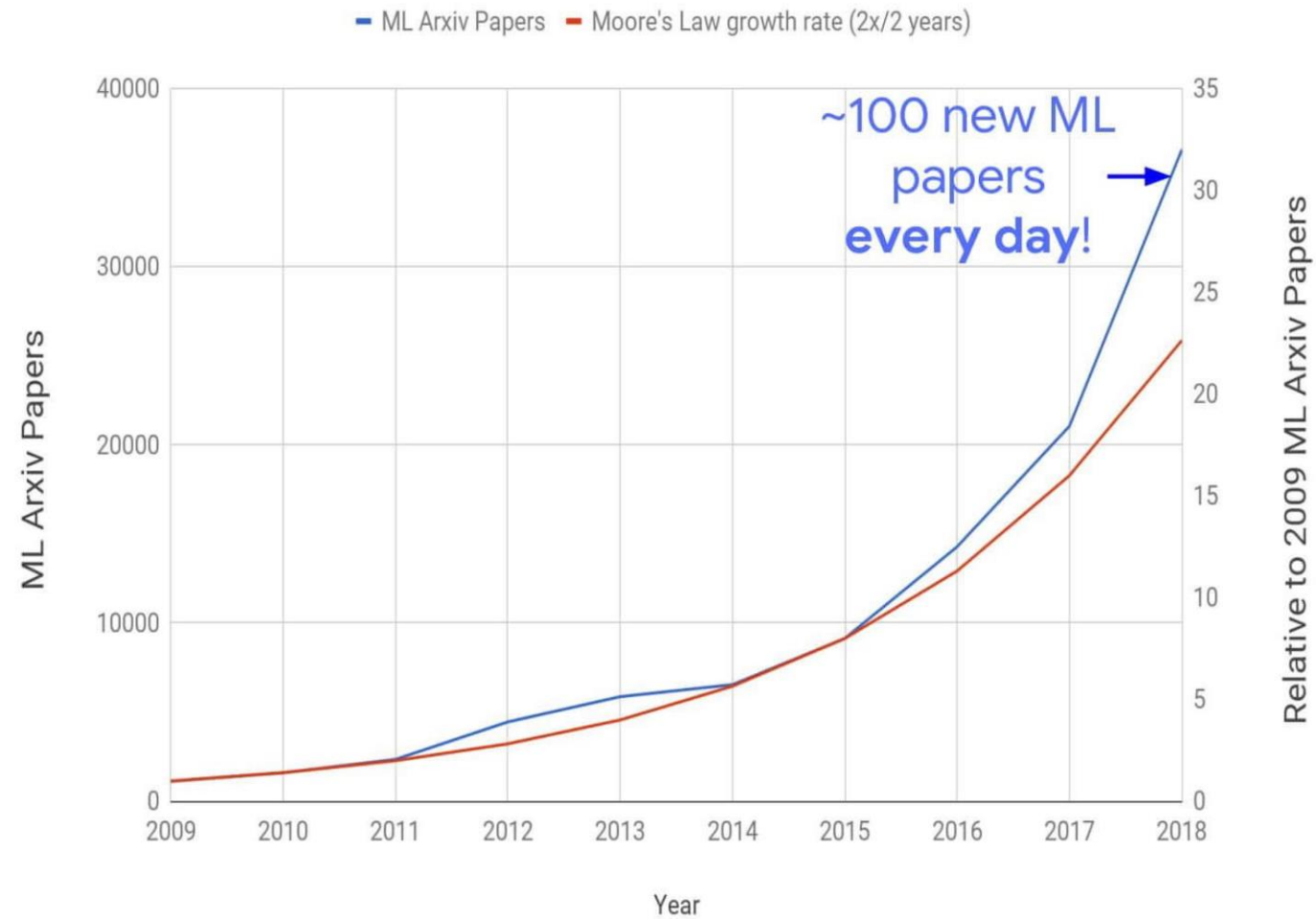
Smale's 18th problem*: *What are the limits of artificial intelligence?*

M. Colbrook, V. Antun, A. Hansen, "*The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem*" (PNAS, 2022)

*Steve Smale's list of problems for the 21st century (requested by Vladimir Arnold), inspired by Hilbert's list.

<http://www.damtp.cam.ac.uk/user/mjc249/home.html>: slides, papers, and code

A fun stat!

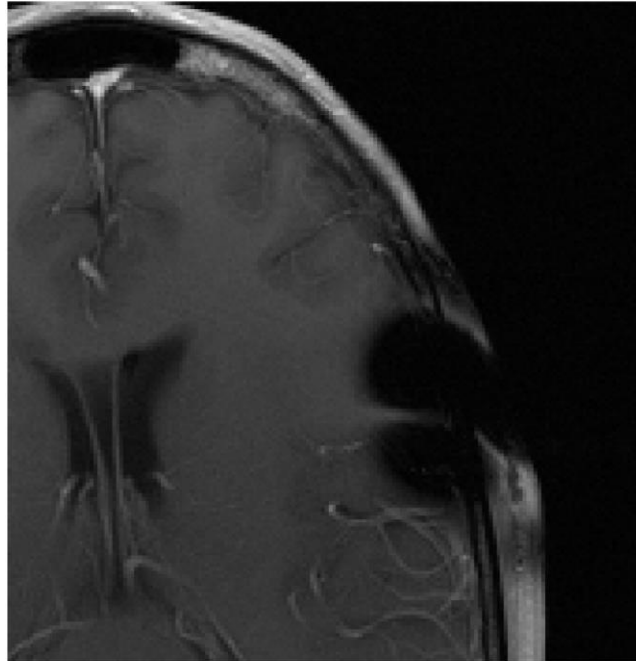


To keep up during first lockdown, would need to continually read a paper every 4 mins!

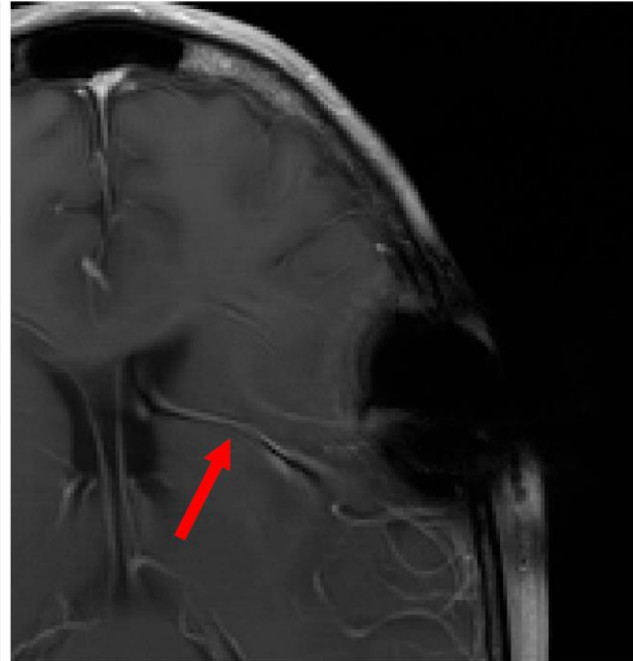
Problem: hallucinations and instability

Hallucinations in image reconstruction

Original image



AI reconstruction



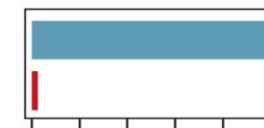
“AI generated hallucination”, from Facebook and NYU’s *FastMRI challenge* 2020

Instabilities in medical diagnosis

Original Mole

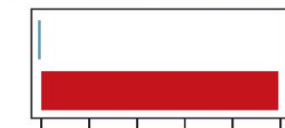


Perturbed Mole



Benign
Malignant

Model confidence



Benign
Malignant

Model confidence

From Finlayson et al., “Adversarial attacks on medical machine learning,” *Science*, 2019.

When can we make AI robust and trustworthy?

Smale's 18th problem: “What are the limits of AI?”

*“Very often, the creation of a technological artifact precedes the science that goes with it. The steam engine was invented before thermodynamics. Thermodynamics was invented to explain the steam engine, essentially the **limitations** of it. **What we are after is the equivalent of thermodynamics for intelligence.**”*

Yann LeCun

*“2021 was the year in which the wonders of artificial intelligence stopped being a story. Many of this year's top articles grappled with the **limits of deep learning** (today's dominant strand of AI).”*

IEEE Spectrum, 2021's Top Stories About AI (Dec. 2021)

Example of the limits of deep learning

Paradox: “Nice” linear inverse problems where a *stable* and *accurate* neural network for image reconstruction exists, but it can never be trained!

E.g., suppose we want to solve (holds for much more general problems)

$$\min_{x \in \mathbb{C}^N} \|x\|_{l^1} + \lambda \|Ax - y\|_{l^2}^2$$

$$A \in \mathbb{C}^{m \times N} \text{ (modality, } m < N), \quad S = \{y_K\}_{K=1}^R \text{ (samples)}$$

Arises when given $y \approx Ax + e$.

Enforce condition numbers bounded by 1.

Data

$$A \in \mathbb{C}^{m \times N} \text{ (modality, } m < N), \quad S = \{y_k\}_{k=1}^R \text{ (samples)}$$

In practice, A is not known exactly or cannot be stored to infinite precision.

Assume access to $\{y_{n,k}\}_{k=1}^R$ and A_n (rational approx, e.g., floats) such that

$$\|y_{n,k} - y_k\| \leq 2^{-n}, \quad \|A_n - A\| \leq 2^{-n}, \quad n \in \mathbb{N}.$$

Training set for $(A, S) \in \Omega$:

$$\iota_{A,S} = \{(y_{n,k}, A_n) : k = 1, \dots, R \text{ and } n \in \mathbb{N}\}.$$

In a nutshell: allow access to arbitrary precision training data.

Question: Given a collection Ω of (A, S) , does there exist a neural network approximating the solution map, and can it be trained by an algorithm?

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{l^1} + \lambda \|Ax - y\|_{l^2}^2$$

What could go wrong?

- 1. Non-existence:** No neural network approximates solution map.

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{l^1} + \lambda \|Ax - y\|_{l^2}^2$$

What could go wrong?

~~1. **Non-existence:** No neural network approximates solution map.~~

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{l^1} + \lambda \|Ax - y\|_{l^2}^2$$

What could go wrong?

- ~~1. **Non-existence:** No neural network approximates solution map.~~
- 2. Non-trainable:** \exists a neural network that approximates solution map, but it cannot be trained.

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{l^1} + \lambda \|Ax - y\|_{l^2}^2$$

What could go wrong?

- ~~1. **Non-existence:** No neural network approximates solution map.~~
- 2. Non-trainable:** \exists a neural network that approximates solution map, but it cannot be trained.
- 3. Not practical:** \exists a neural network that approximates solution map, and an algorithm training it. However, the algorithm needs prohibitively many samples.

Example of the limits of deep learning

Paradox: “Nice” linear inverse problems where a *stable* and *accurate* neural network for image reconstruction exists, but it can never be trained!

Theorem: Pick positive integers $n \geq 3$ and M . Class of problems such that:

- **(Not trainable)** No algorithm (even random) can train a neural network with n digits of accuracy over the dataset with probability greater than $1/2$.
- **(Not practical)** $n - 1$ digits of accuracy possible over the dataset, but any training algorithm requires **arbitrarily large training data**.
- **(Trainable and practical)** $n - 2$ digits of accuracy possible over the dataset via training algorithm using M training data.

Holds for any architecture, any precision of training data.

⇒ Classification theory telling us what can and cannot be done

-
- C., Antun, Hansen, "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem," **PNAS**, 2022.
 - Antun, C., Hansen, "Proving Existence Is Not Enough: : Mathematical Paradoxes Unravel the Limits of Neural Networks in Artificial Intelligence," **SIAM News**, May 2022.
 - Choi, "Some AI Systems May Be Impossible to Compute," **IEEE Spectrum**, March 2022.

Numerical example: fails with training methods

$\text{dist}(\Psi_{A_n}(y_n), \Xi(A, y))$	$\text{dist}(\Phi_{A_n}(y_n), \Xi(A, y))$	$\ A_n - A\ \leq 2^{-n}$ $\ y_n - y\ _{\ell^2} \leq 2^{-n}$	10^{-K}
0.2999690	0.2597827	$n = 10$	10^{-1}
0.3000000	0.2598050	$n = 20$	10^{-1}
0.3000000	0.2598052	$n = 30$	10^{-1}
0.0030000	0.0025980	$n = 10$	10^{-3}
0.0030000	0.0025980	$n = 20$	10^{-3}
0.0030000	0.0025980	$n = 30$	10^{-3}
0.0000030	0.0000015	$n = 10$	10^{-6}
0.0000030	0.0000015	$n = 20$	10^{-6}
0.0000030	0.0000015	$n = 30$	10^{-6}

$A \in \mathbb{C}^{19 \times 20}$ from discrete cosine transform, $R = 8000$, solutions 6-sparse. LISTA (learned iterative shrinkage thresholding algorithm) Ψ_{A_n} and FIRENETs Φ_{A_n} . The table shows the shortest l_2 distance between the output and the true minimizer of the problem for different values of n, K .

A paradox relevant to applications

For engineers

For scientists

IEEE Spectrum FOR THE TECHNOLOGY INSIDER

NEWS | ARTIFICIAL INTELLIGENCE

Some AI Systems May Be Impossible to Compute >

New research suggests there are limitations to what deep neural networks can do

BY CHARLES Q. CHOI | 30 MAR 2022 | 4 MIN READ

GETTY IMAGES/IEEE SPECTRUM

NEWS RELEASE 17-MAR-2022

Mathematical paradoxes demonstrate the limits of AI

Peer-Reviewed Publication

UNIVERSITY OF CAMBRIDGE

HOME NEWS RELEASES MULTIMEDIA MEETINGS

For numerical analysts

Newsjournal of the Society for Industrial and Applied Mathematics

siam news

Volume 55/ Issue 4
May 2022

Privacy with Synthetic Data

Differential Privacy Made Simple!

Figure 1: In addition to adding noise to the data set, Sebastian Gombis' differential privacy-based method processes it with an information theory algorithm to ensure synthetic data that—in principle—protects the privacy of the people involved. Figure courtesy of the author.

Figure 2: Researchers can protect privacy by performing a full statistical analysis on the original data set, then using a missing data algorithm called multiple imputation to construct a synthetic data set that has exactly the same statistical characteristics. Figure courtesy of the author.

Proving Existence Is Not Enough: Mathematical Paradoxes Unravel the Limits of Neural Networks in Artificial Intelligence

By Yegor Antun, Matthew J. Colbrook, and Anders C. Hansen

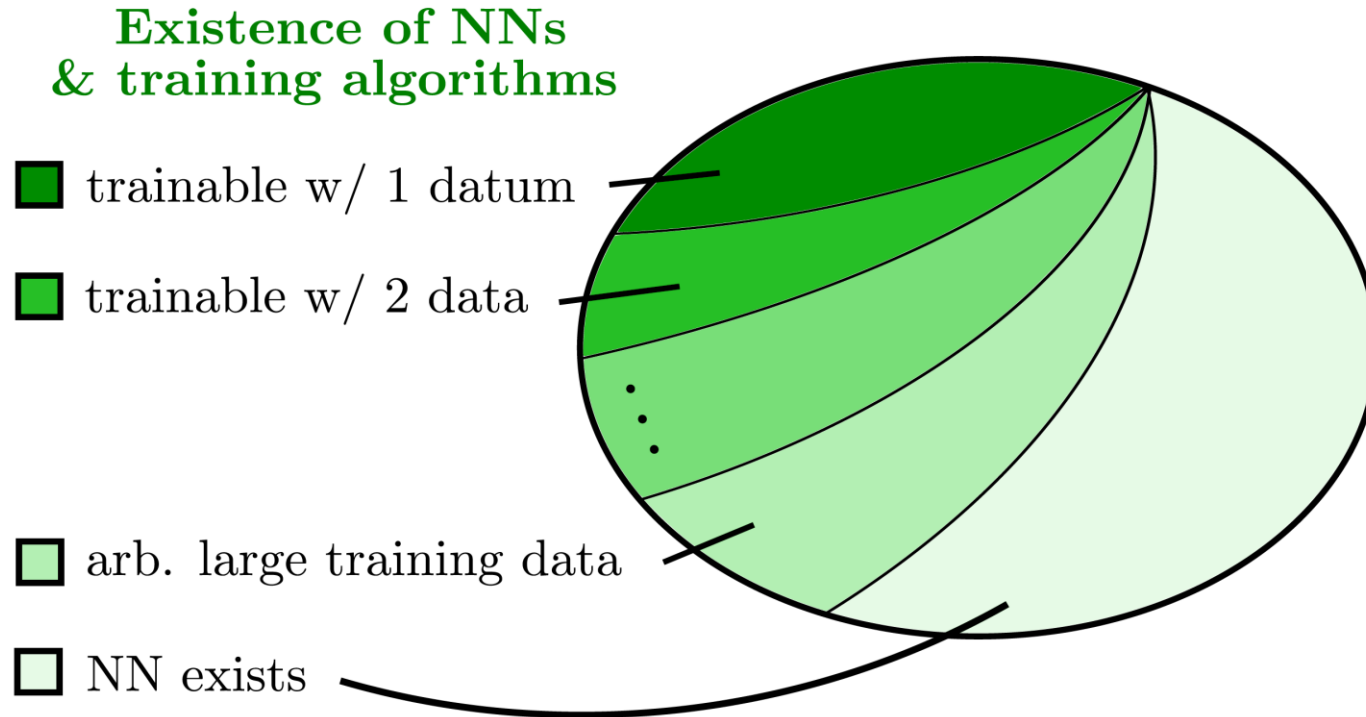
The impact of deep learning (DL), neural networks (NNs), and artificial intelligence (AI) over the last decade has been profound. Advances in computer vision and natural language processing have yielded smart speakers in our homes, driving assistance in our cars, and automated diagnoses in medicine. AI has also rapidly entered scientific computing. However, overwhelming amounts of empirical evidence [3, 8] suggest that modern AI is often non-robust (unstable), may generate hallucinations, and can produce nonsensical output with high levels of prediction confidence (see Figure 1). These issues present a serious concern for AI use within legal frameworks. As stated by the European Commission's Joint Research Centre, "In the light of the recent advances in AI, the serious negative consequences of its use for EU citizens and organizations have led to multiple initiatives [...] Among the identified requirements, the concepts of robustness and explainability of AI systems have emerged as key elements for a future regulation."

Robustness and trust of algorithms lie at the heart of numerical analysis [9]. The lack of robustness and trust in AI is hence the "Achilles' heel" of DL and has become a serious political issue. Classical approximation theorems show that a continuous function can be approximated arbitrarily well by a NN [5]. Therefore, stable problems that are described by stable functions can be solved stably with a NN. These results inspire the following fundamental question: Why does DL lead to unstable methods and AI-generated hallucinations, even in scenarios where we can prove that stable and accurate NNs exist?

The strong opinion that surrounds AI is evident in computer scientist Geoffrey Hinton's 2017 quote: "They should stop training multi-layered nets." Such optimism is comparable to the confidence that surrounded mathematics in the early 20th century, as summed up in David Hilbert's sentiment: "We know more. We wonder what we know." Hilbert believed that mathematics could prove or disprove any statement, and that there were no restrictions on which problems algorithms could solve. The seminal contributions of Kurt Gödel [11] and Alan Turing [12] turned Hilbert's idealism upside down by establishing paradoxes that demonstrate the impossibility of such a goal.

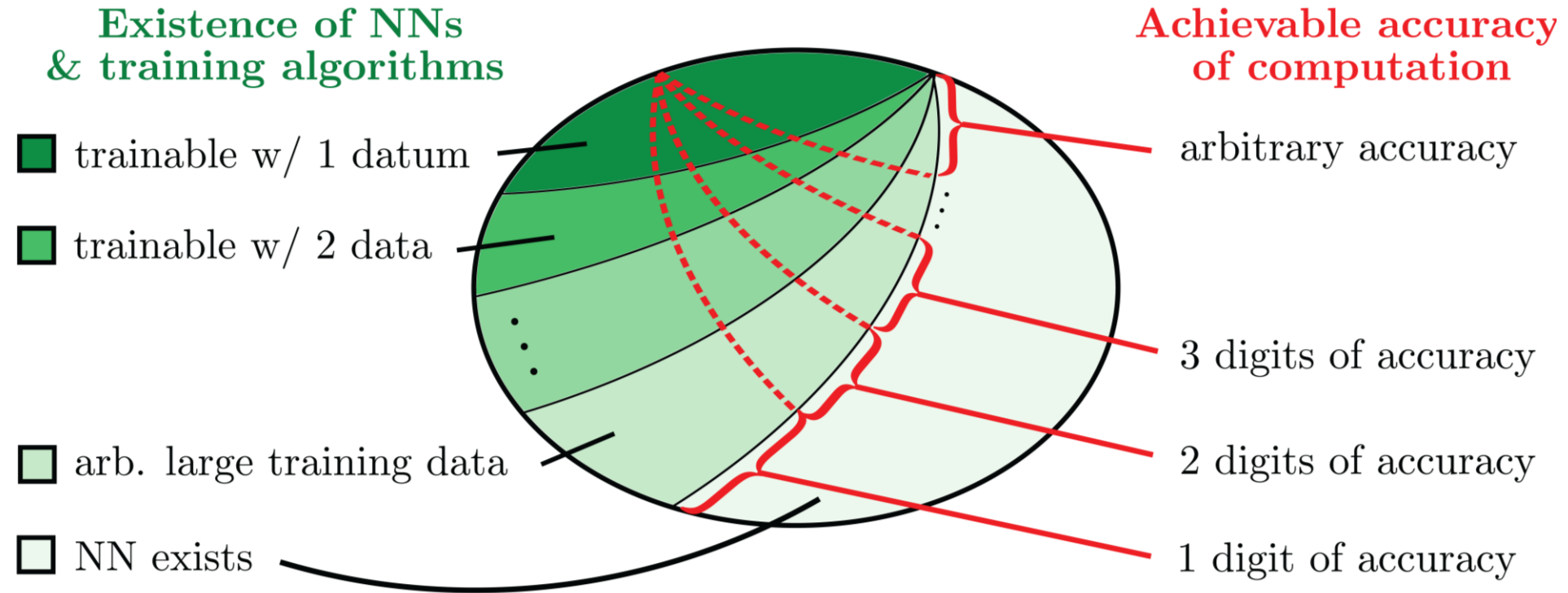
Figure 1: Hallucinations in image reconstruction and medical diagnoses. **1a**, The correct, original image from the 2020 fastMRI Challenge. **1b**, Reconstruction by an artificial intelligence (AI) method that produces an incorrect detail (dispergated hallucination). **1c**, Dermatoscope image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network (DNN). **1d**, Confirmed image of the nevus with a slight perturbation and the diagnostic probability from the same DNN. One diagnosis is clearly incorrect, but can an algorithm determine which one? Figures 1a and 1b are courtesy of the 2020 fastMRI Challenge [10], and 1c and 1d are courtesy of [6].

The world of neural networks



Given a problem and conditions, where does it sit in this diagram?

The world of neural networks



Given a problem and conditions, where does it sit in this diagram?

Example counterpart theorem

Certain conditions: stable neural networks trained with exponential accuracy.
E.g., *approximate Łojasiewicz-type inequality*:

$$(1) \quad \min_{x \in \mathbb{C}^N} f(x) \quad \text{s.t.} \quad \|Ax - y\| \leq \varepsilon$$

$$\text{dist}(x, \text{solution}) \leq \alpha([f(x) - f^*] + [\|Ax - y\| - \varepsilon] + \delta)$$

Fast Iterative REstarted NETworks (FIRENETs)
(unrolled primal-dual with novel restart scheme)

Theorem: Training algorithm that, under above assumption, produces *stable* neural networks φ_n of width $O(N)$, depth $O(n)$, guaranteed worst bound

$$\text{dist}(\varphi_n(y), \text{solution}) \lesssim e^{-n} + \delta$$

- C., Antun, Hansen, "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem," **PNAS**, 2022.
- C., "WARPd: A linearly convergent first-order method for inverse problems with approximate sharpness conditions," **SIIMS**, 2022.

Demonstration of convergence

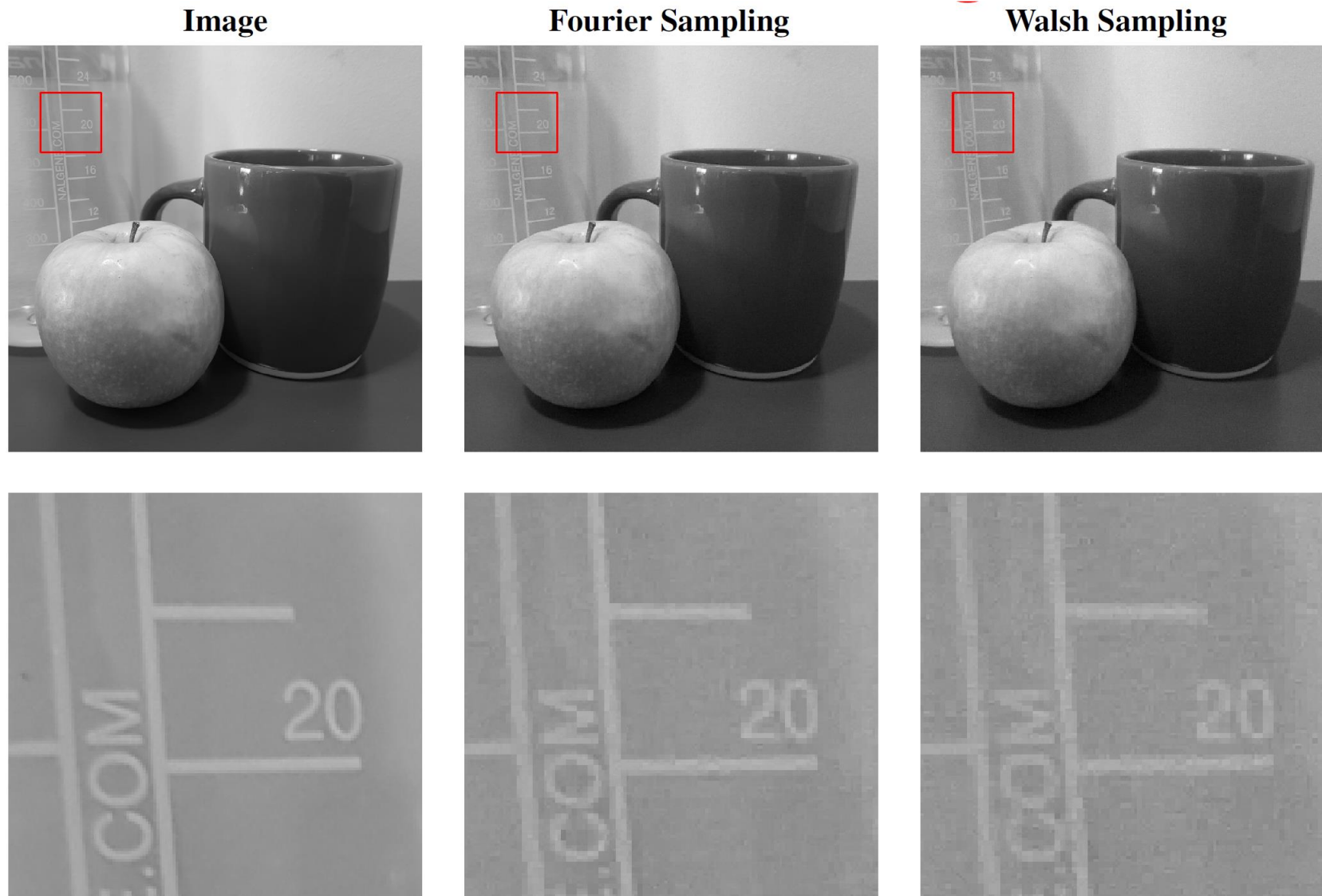
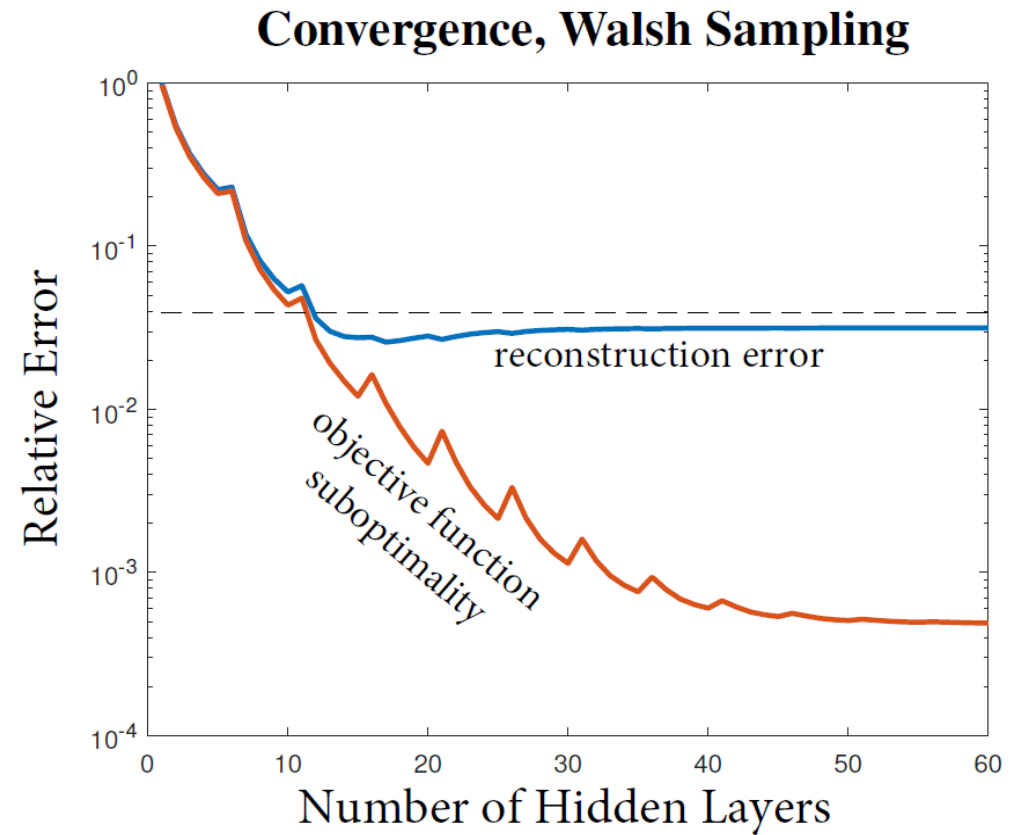
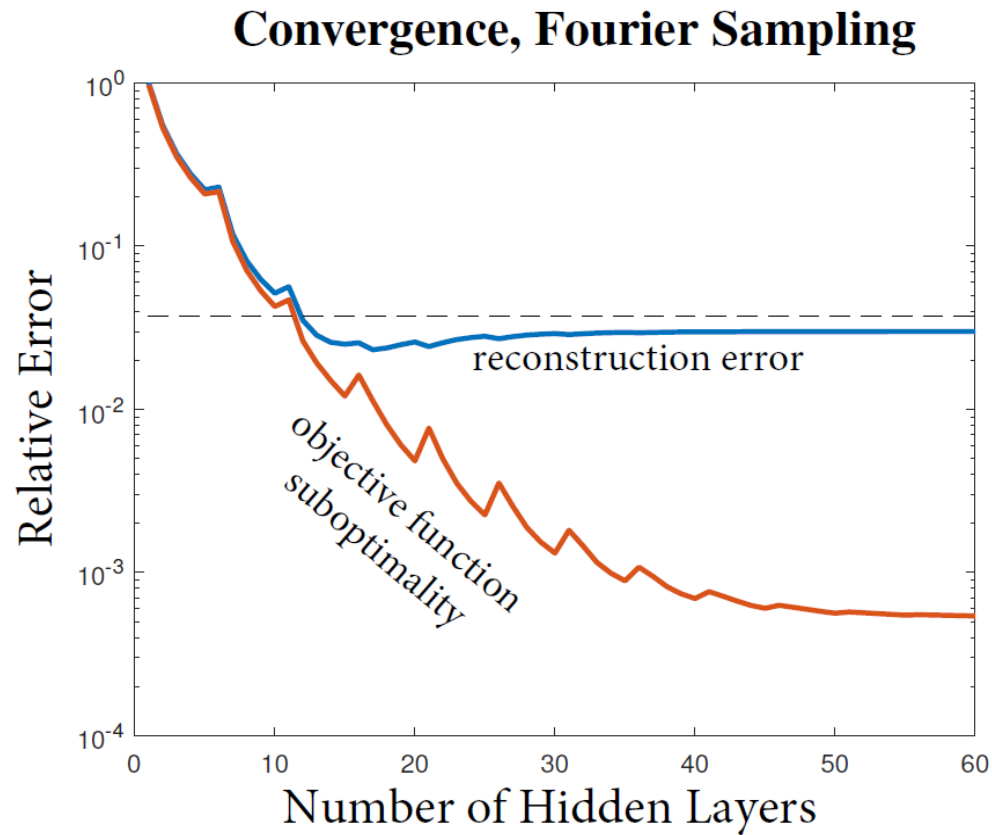


Figure: Images corrupted with 2% Gaussian noise and reconstructed using 15% sampling.

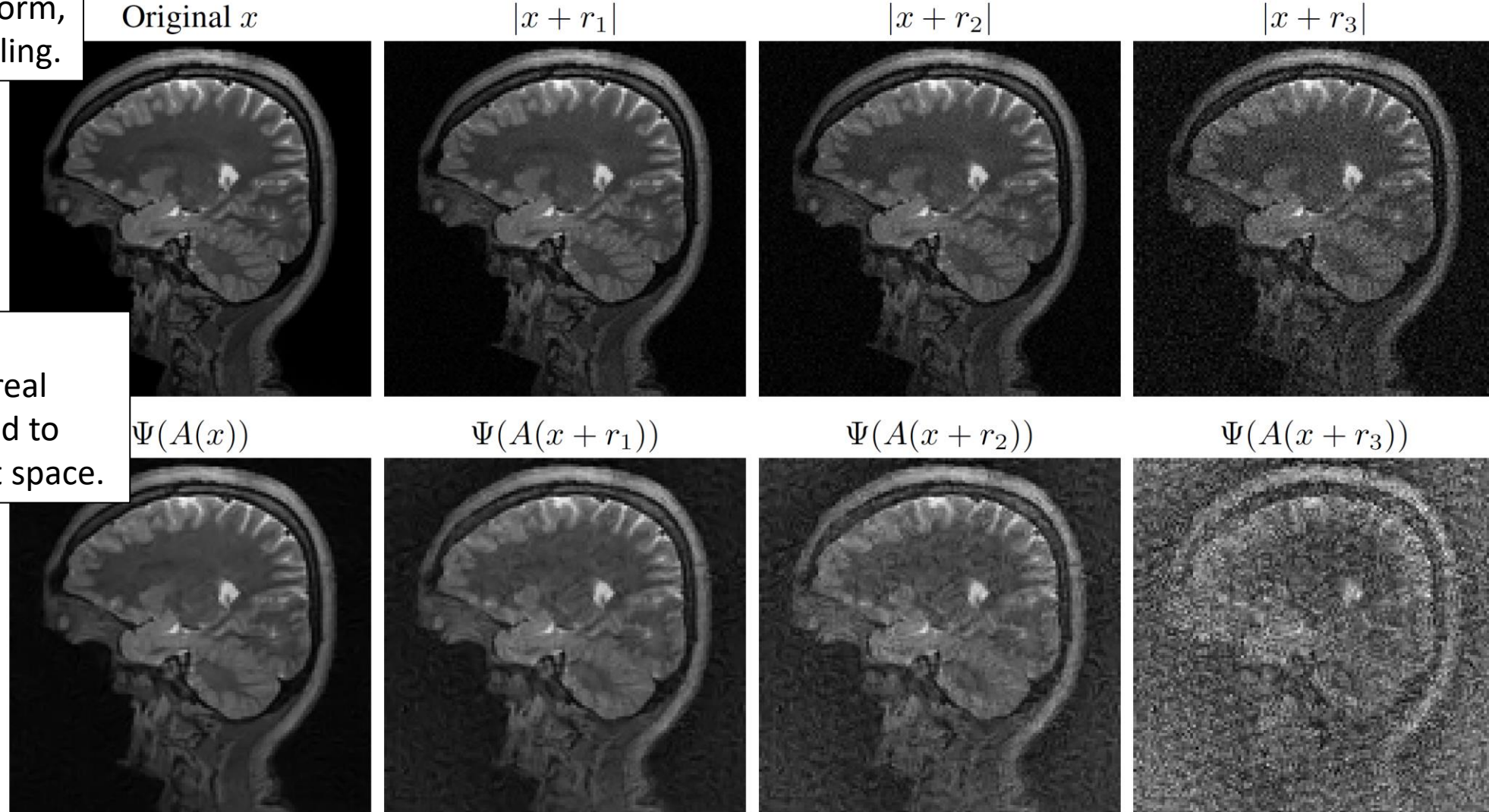
Demonstration of convergence



Example of severe instability

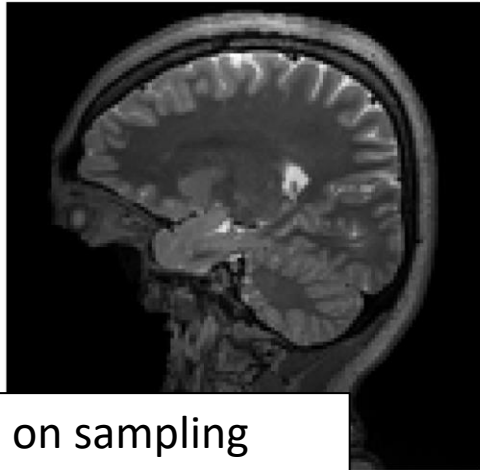
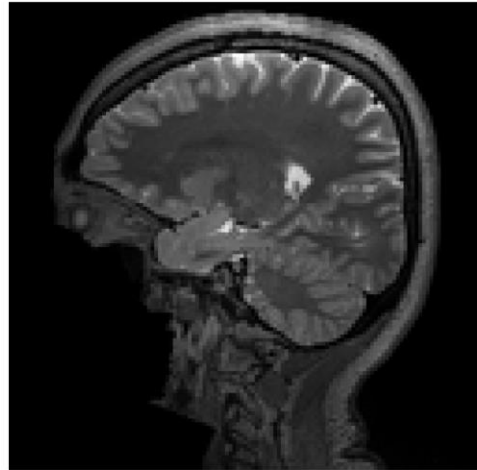
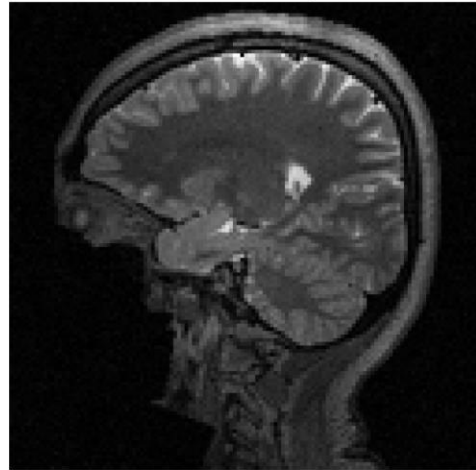
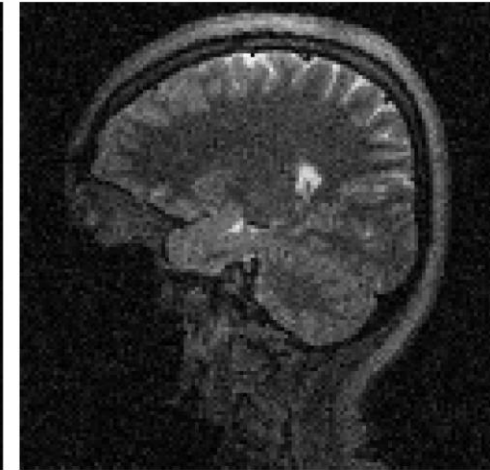
MRI: discrete 2D
Fourier transform,
60% subsampling.

Perturbations
computed in real
space, mapped to
measurement space.

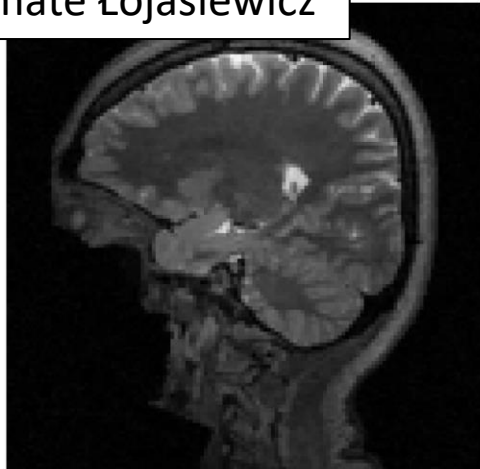
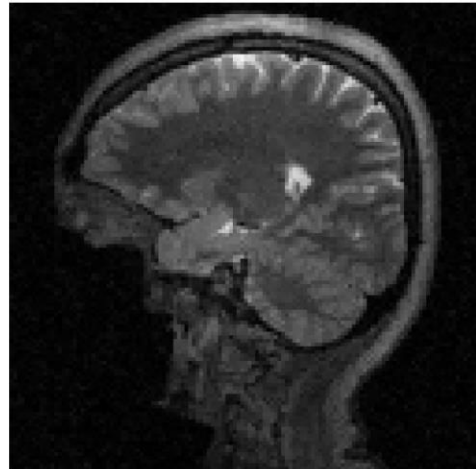
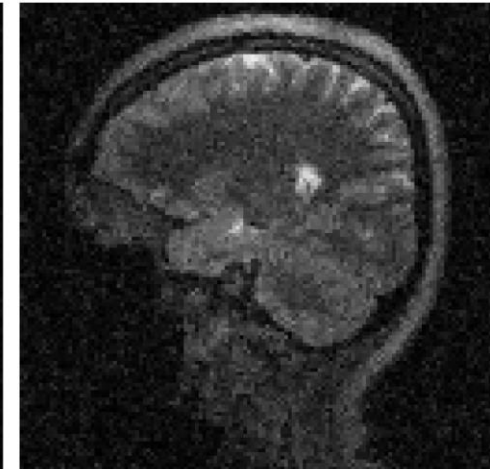


- Zhu et al., “Image reconstruction by domain-transform manifold learning,” **Nature**, 2018.
- Antun et al., “On instabilities of deep learning in image reconstruction and the potential costs of AI,” **PNAS**, 2020.

FIRENET: provably stable (even to adversarial examples) and accurate

Original x  $|x + v_1|$  $|x + v_2|$  $|x + v_3|$ 

Assumptions on sampling
and approximate sparseness
give approximate Łojasiewicz

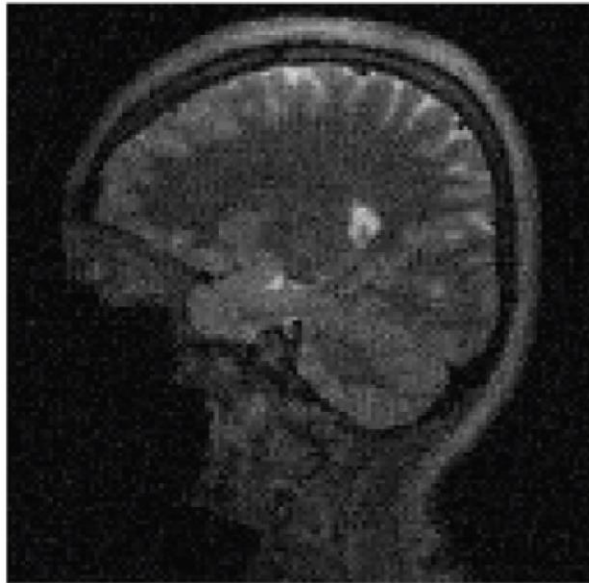
 $\Phi(A(x + v_1))$  $\Phi(A(x + v_2))$  $\Phi(A(x + v_3))$ 

Stabilising unstable neural networks

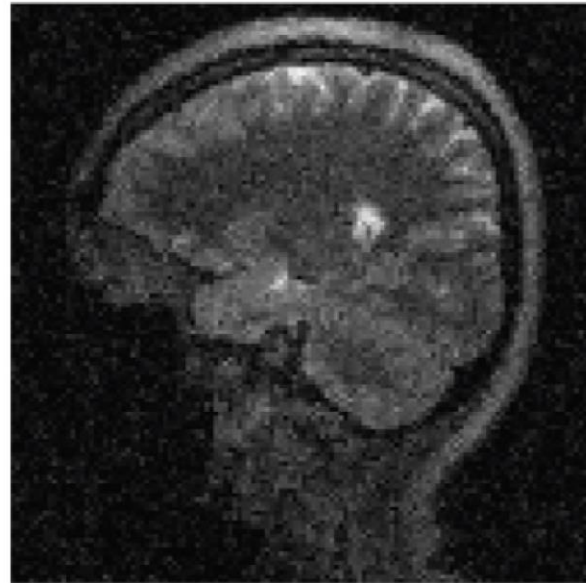
$\Psi(\tilde{y}), \tilde{y} = Ax + e_3$



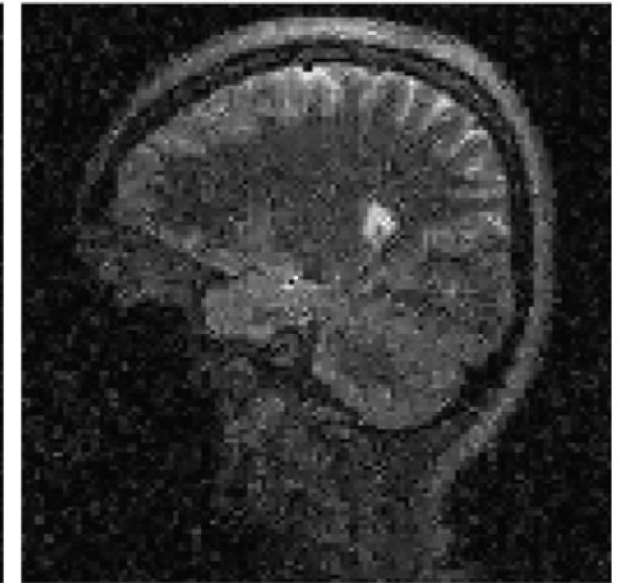
$\Phi(\tilde{y}, \Psi(\tilde{y}))$



FIRENET rec. from $y = Ax + \tilde{e}_3$



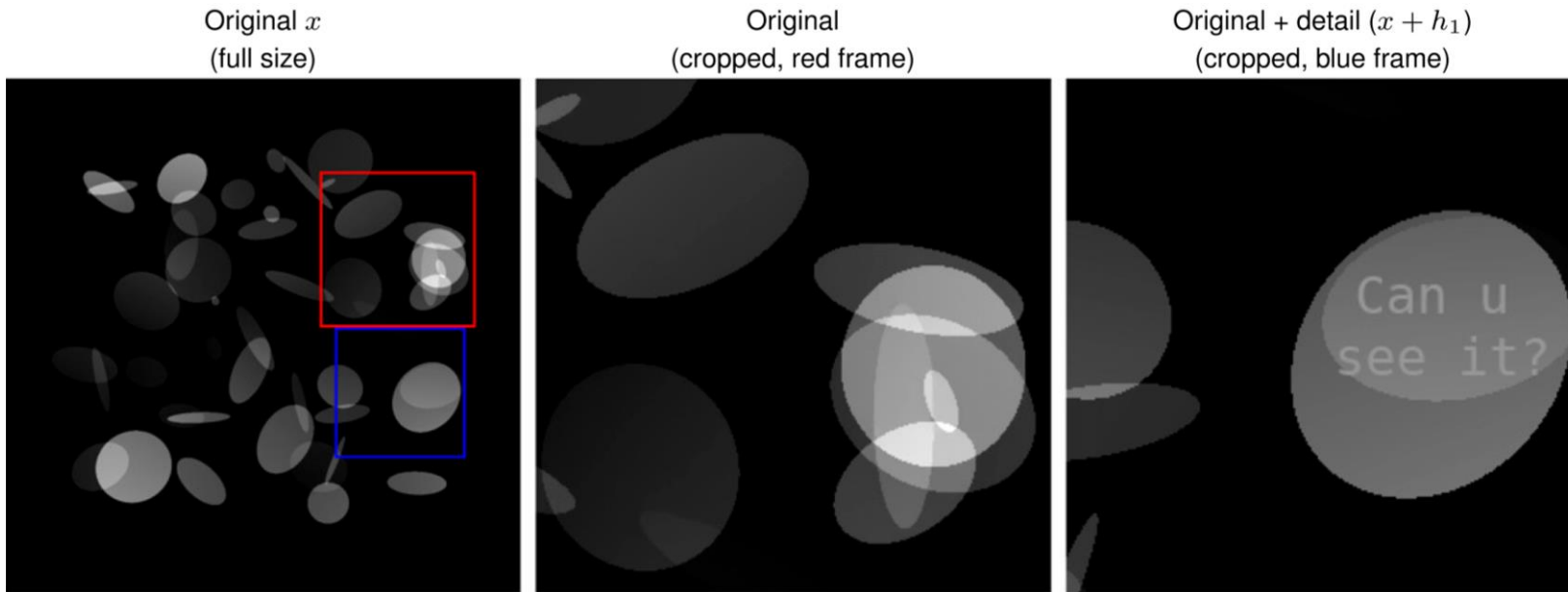
AUTOMAP+FIRENET rec. from
 $y = Ax + \hat{e}_3$



Key pillars: stability and accuracy

MRI: discrete 2D
Fourier transform,
15% subsampling.

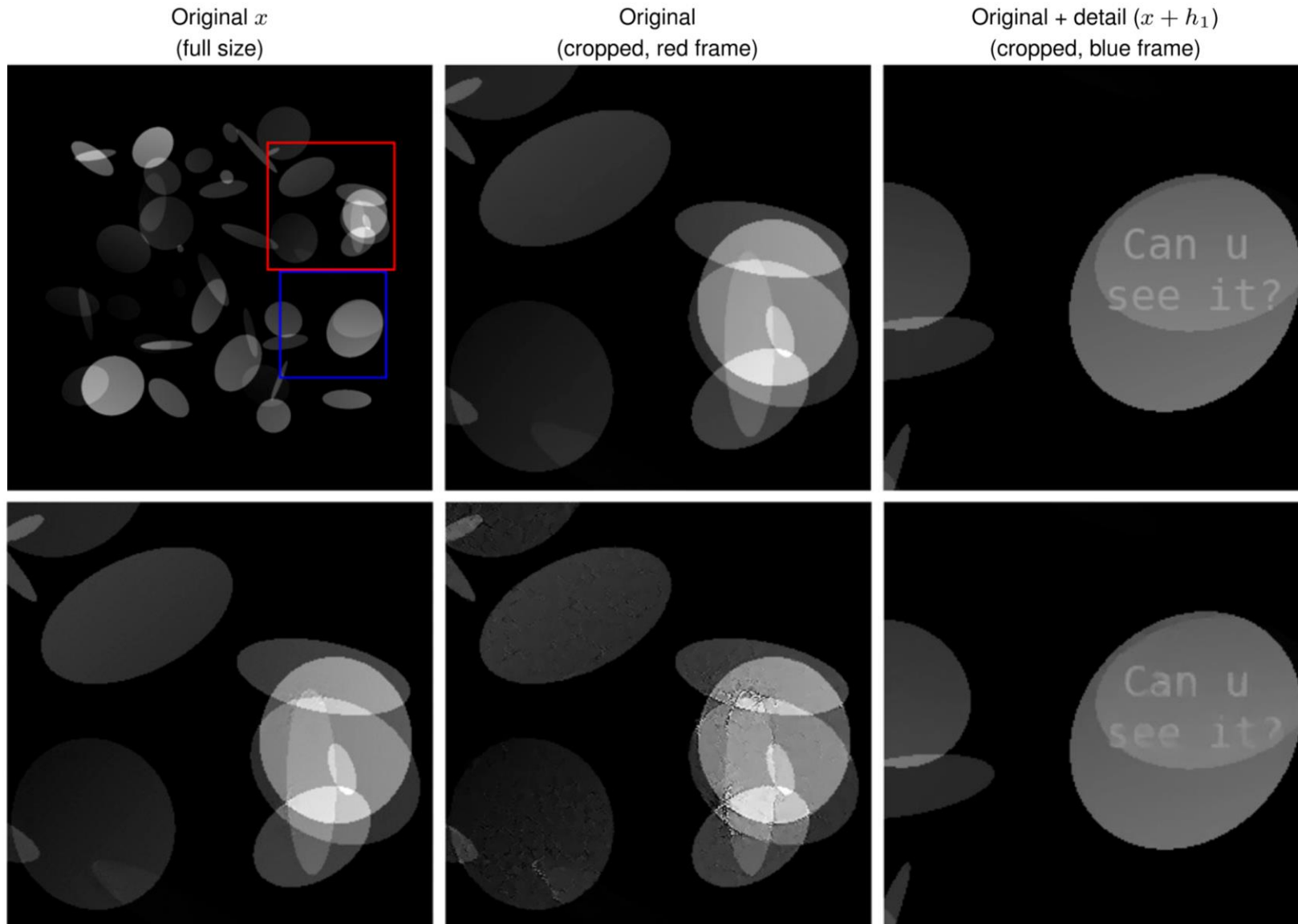
All networks
trained on 5000
images of ellipses



U-Net with no noise: accurate but unstable

15/16

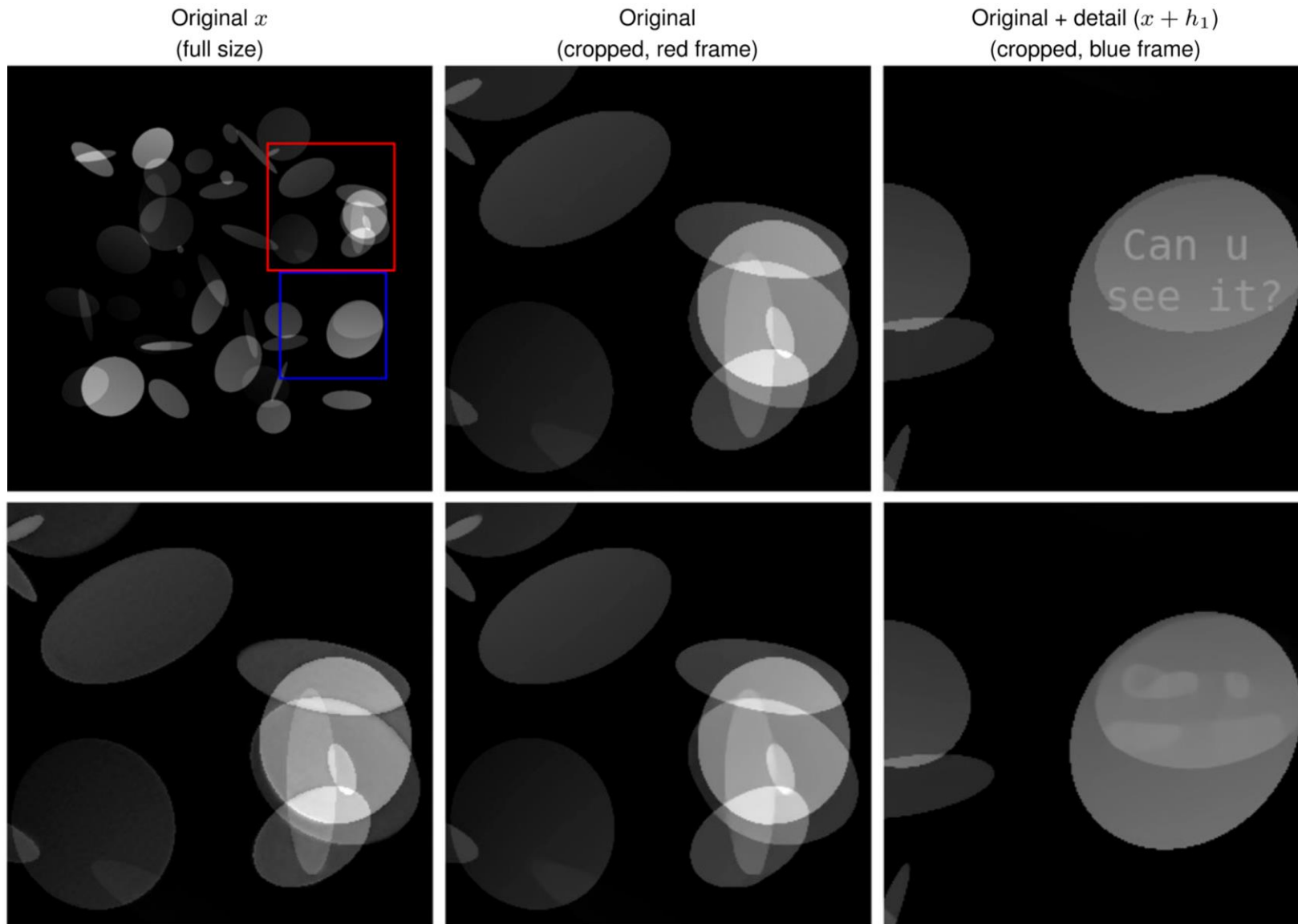
U-Net: standard neural network architecture for imaging. Approx 4 million parameters.



- C., Antun, Hansen, "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem," **PNAS**, 2022.

U-Net with noise: stable but inaccurate

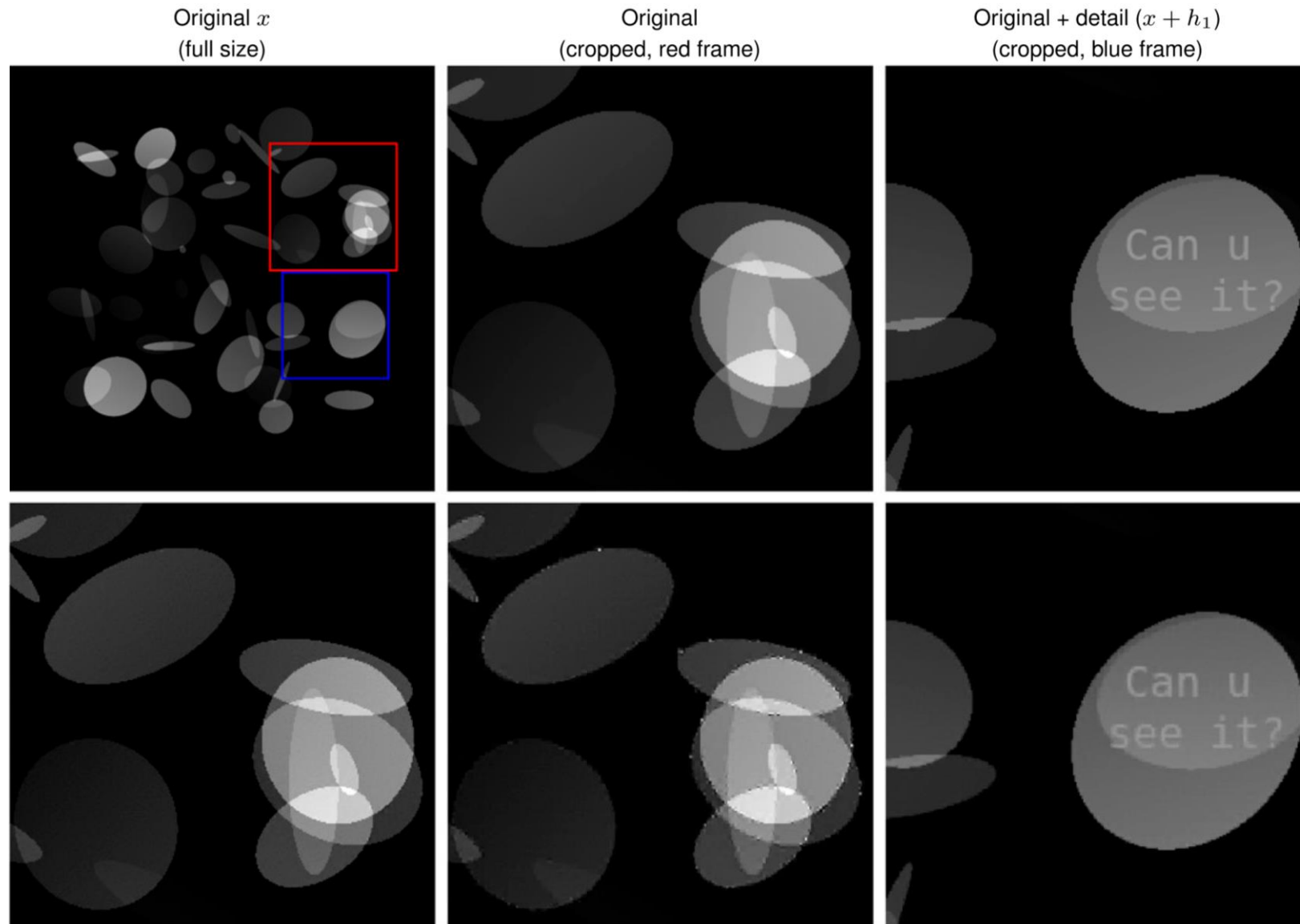
15/16



- C., Antun, Hansen, "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem," **PNAS**, 2022.

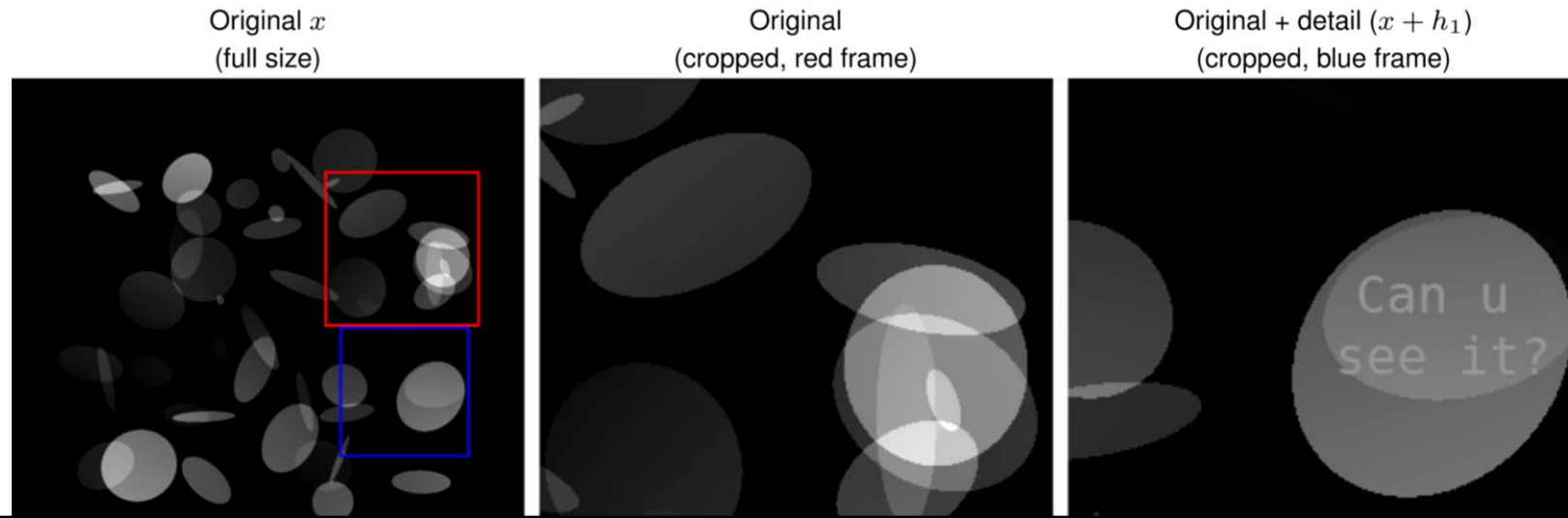
FIRENET: balances stability and accuracy?

15/16

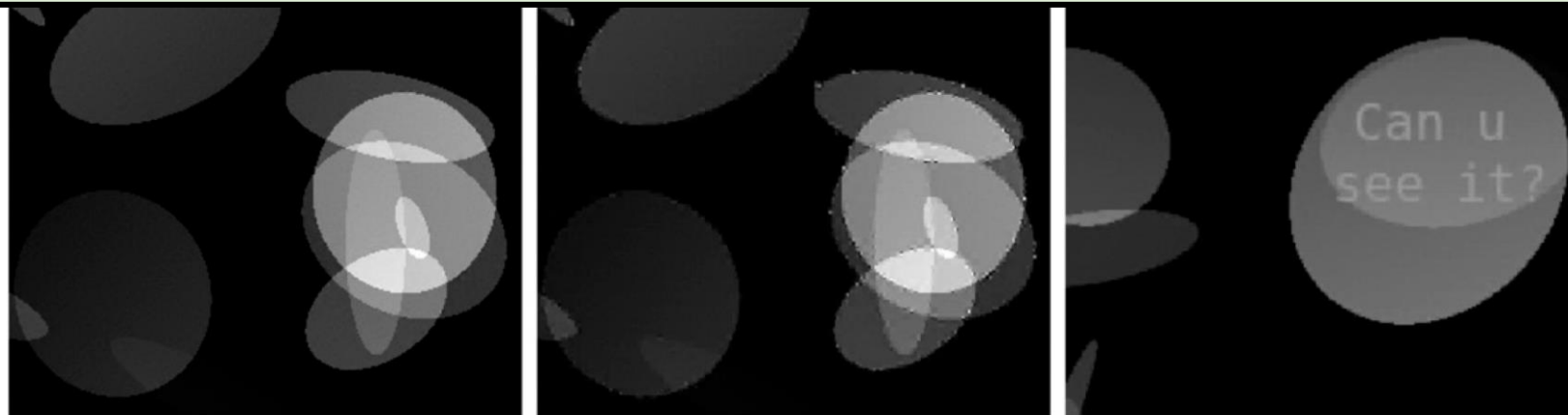


FIRENET: balances stability and accuracy?

15/16



Open problem: use the toolkit to precisely prove theorems about *optimal* trade-offs.



Summary

Need for foundations in AI/deep learning!

- **Paradox:** Nice linear inverse problems where stable and accurate neural network exists but cannot be trained!
- Trainability depends on
 - Accuracy desired.
 - Amount of training data.
- Specific conditions \Rightarrow FIRENETs exp. convergence
+ withstand adversarial attacks.
- Trade-off between stability and accuracy in deep learning.