Self Supervised Learning Methods for Imaging

Seminario modelos generativos, Montevideo, Uruguay

Julián Tachella, CNRS, École Normale Supérieure de Lyon

Work with Brett Levac, Jon Tamir (UTAustin, USA) and Marcelo Pereyra (Heriot-Watt University, UK)

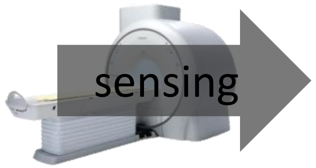
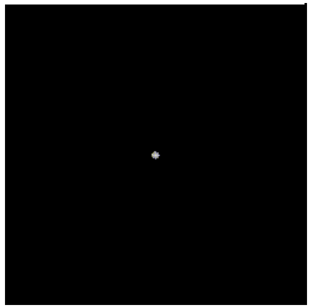
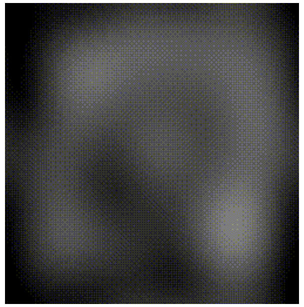
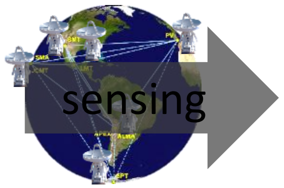
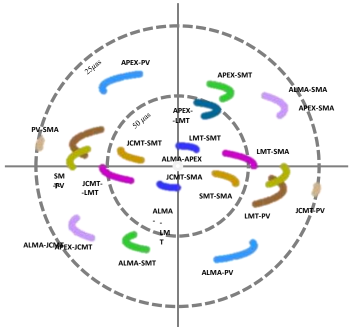
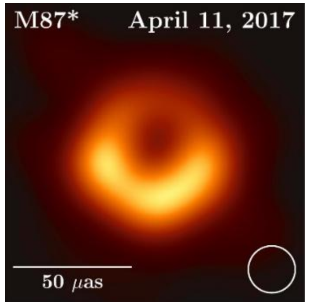
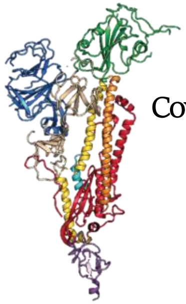
The Inverse problem

Goal: estimate signal $\boldsymbol{x}$ from $\boldsymbol{y}$

$$\underset{\in \mathbb{R}^m}{\text{measurements}} \rightarrow \boldsymbol{y} = \underbrace{A}_{\text{Physics}}(\overset{\text{signal} \in \mathbb{R}^n}{\boldsymbol{x}}) + \boldsymbol{\epsilon} \leftarrow \text{noise/error}$$

We will focus on linear problems where the forward operator A is a matrix

Examples

	x	A	y	reconstruction
Magnetic Resonance Imaging (MRI) A: undersampled Fourier models		 		recover  Source: Brian Hargreaves
Black Hole Imaging A: spatial-frequency e.g. Event Horizon Telescope (EHT)	?	 	 25 µas 50 µas APEX-PV APEX-SMT ALMA-SMA APEX-SMA PV-SMA APEX-LMT LMT-SMT JCMT-SMT ALMA-APEX LMT-SMA JCMT-SMA SM-PV JCMT-LMT SMT-SMA ALMA LMT-PV JCMT-PV ALMA-JCMT APEX-JCMT ALMA-LMT ALMA-SMT ALMA-PV	recover  The Astrophysical Journal Letters, vol. 875, no. L1, 2019.
Cryogenic electron microscopy (Cryo-EM) A: 2D projections of protein particles		 		recover  Covid-19 virus' structure D. Wrapp et al. *Science*, vol. 367, no. 6483, 2020.

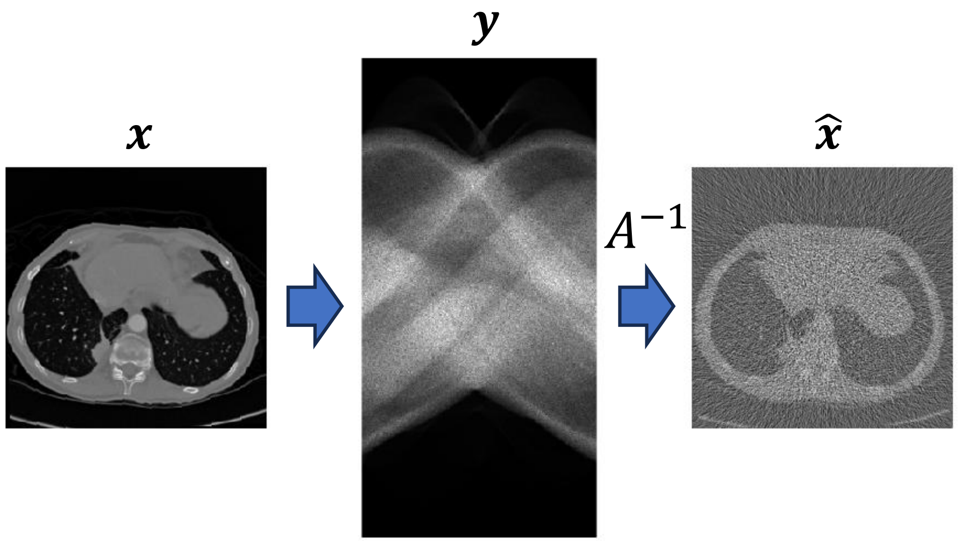
Why it is hard to invert?

Measurements are usually corrupted by noise, e.g.

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

Can be additive, as above, or more complex, e.g. Poisson.

- Often, we do not know the exact noise distribution
- The forward operator may be poorly conditioned



with noise

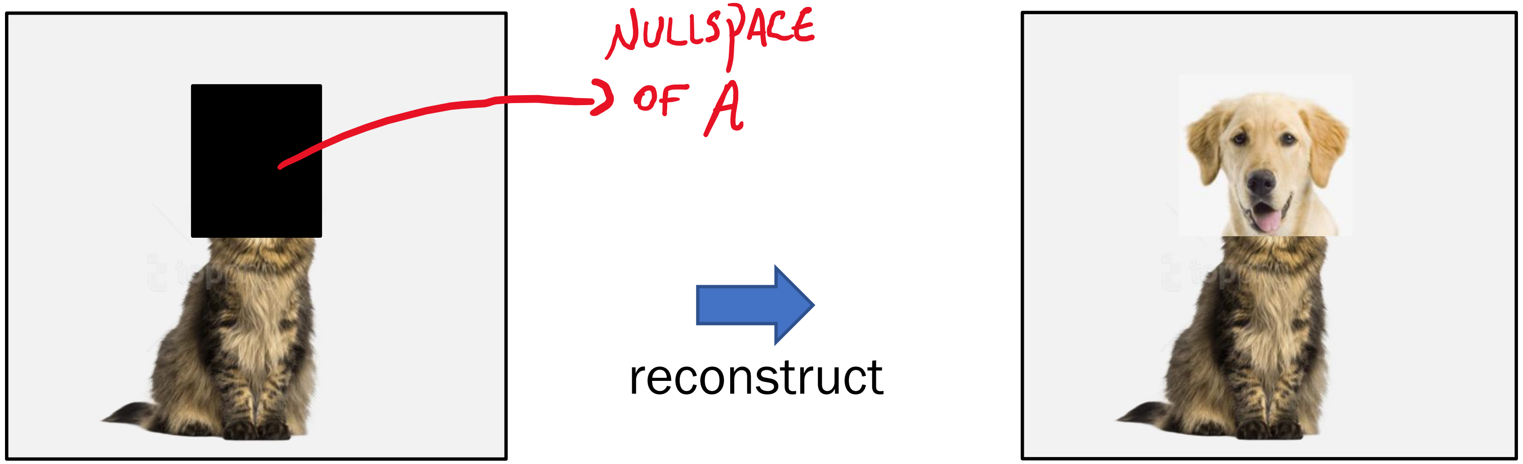
Why it is hard to invert?

Even in the absence of noise, A may not be invertible, giving infinitely many $\hat{\boldsymbol{x}}$ consistent with $\boldsymbol{y}$:

$$\hat{\boldsymbol{x}} = A^{\dagger}\boldsymbol{y} + \boldsymbol{v}$$

where $A^{\dagger}$ is the pseudo-inverse of A and $\boldsymbol{v}$ is any vector in nullspace of A

Unique solution only possible if set of signals $\mathcal{X}$ is low-dimensional

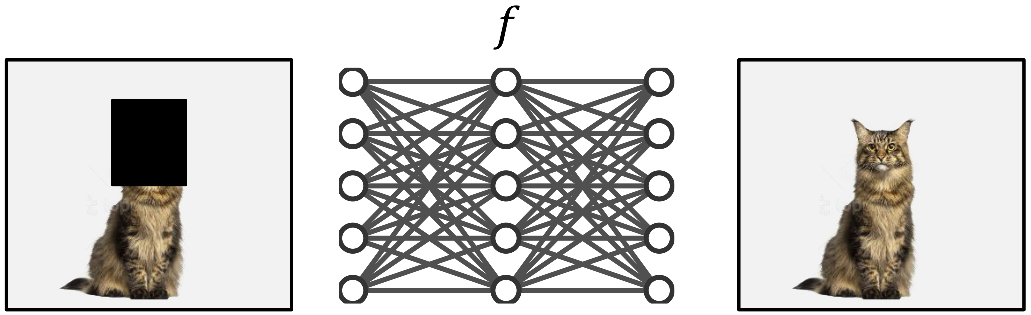


NULLSPACE OF A

reconstruct

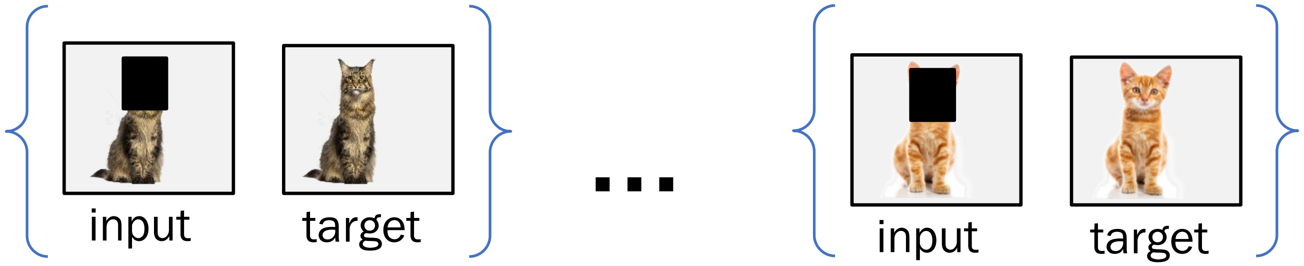
Learning approach

Idea: use training pairs of signals and measurements to directly learn the inversion function



$$\underset{f}{\operatorname{argmin}} \sum_{i=1}^{N} || f(\boldsymbol{y}_i) - \boldsymbol{x}_i||^2$$

supervised dataset



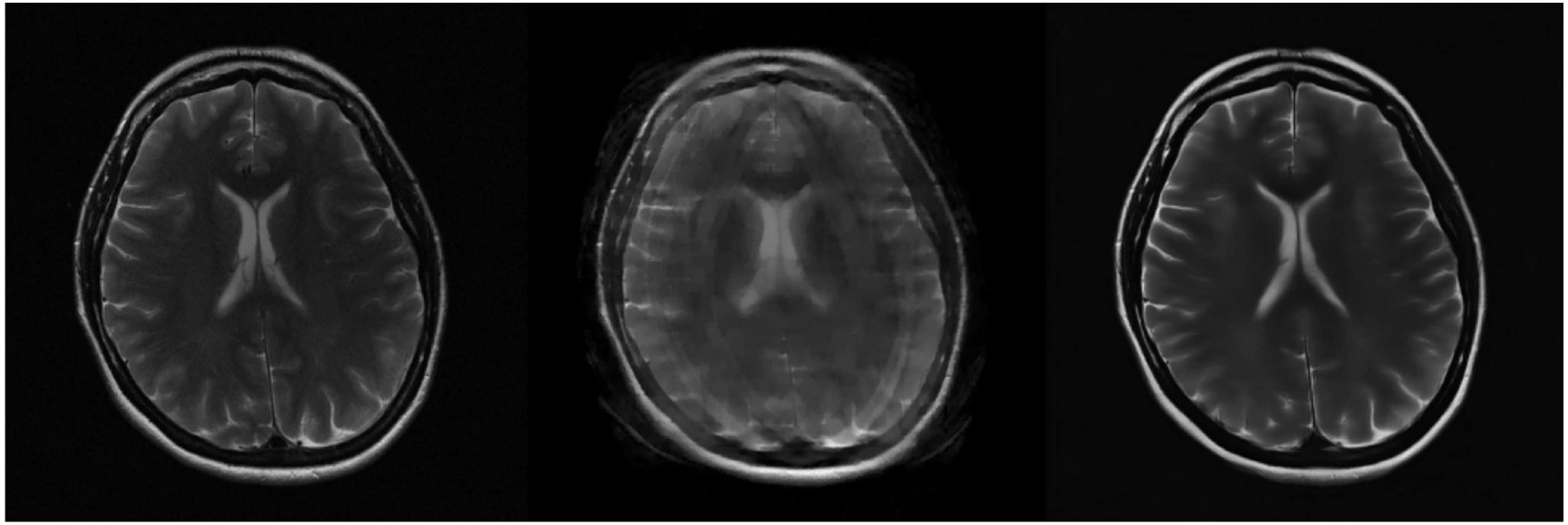
Learning approach

Advantages:

- State-of-the-art reconstructions
- Once trained, f_θ is easy to evaluate

fastMRI

Accelerating MR Imaging with AI



Ground-truth

Total variation (28.2 dB)

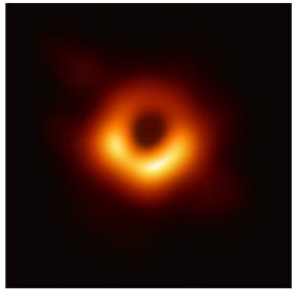
Deep network (**34.5 dB**)

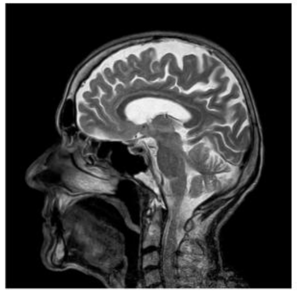
x8 accelerated MRI [Zbontar et al., 2019]

Learning approach

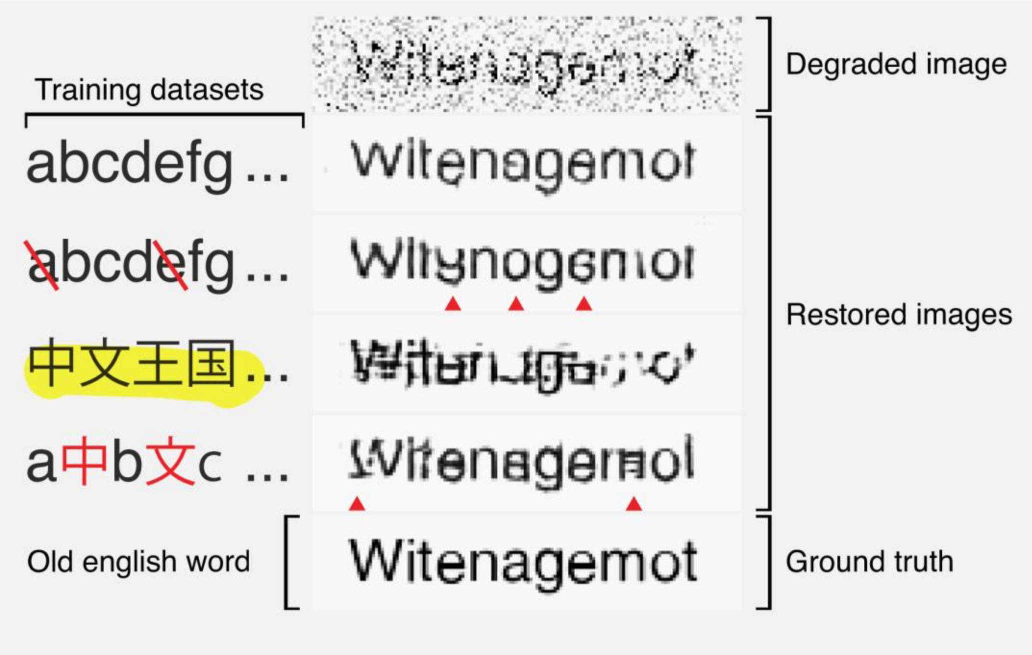
Main disadvantage: Obtaining training signals $\boldsymbol{x}_i$ can be expensive or impossible.

- Medical and scientific imaging





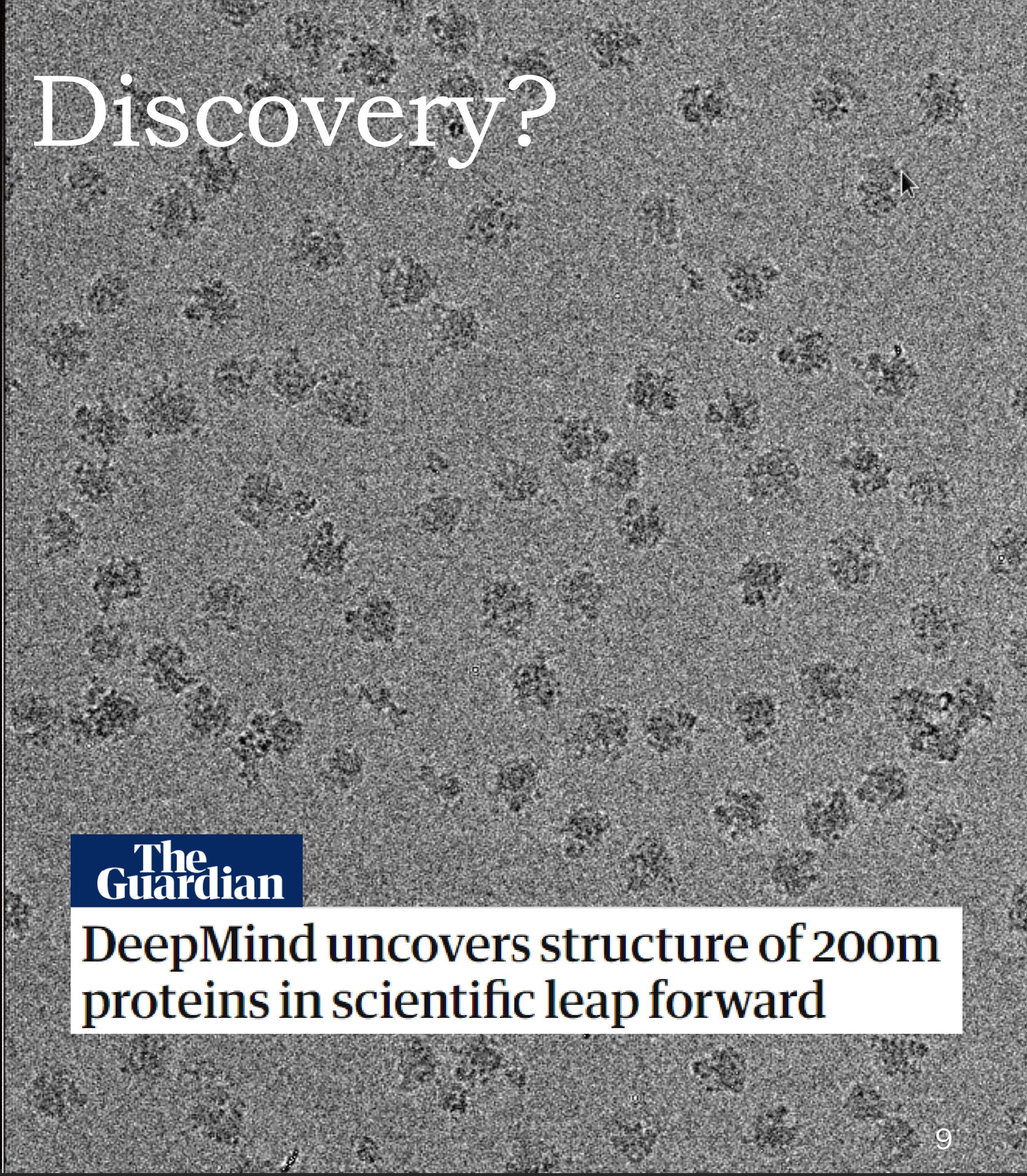
- Distribution shift [Belthangady & Royer, 2019]



AI for Knowledge Discovery?

The Guardian

Black hole picture captured for first time in space breakthrough



The Guardian

DeepMind uncovers structure of 200m proteins in scientific leap forward

Purpose of this talk

How can we **learn diffusion models** from measurement $\{\boldsymbol{y}_i\}_{i=1}^{N}$ data alone?

1. Noisy: $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$

2. Incomplete and noisy: $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$

Estimators

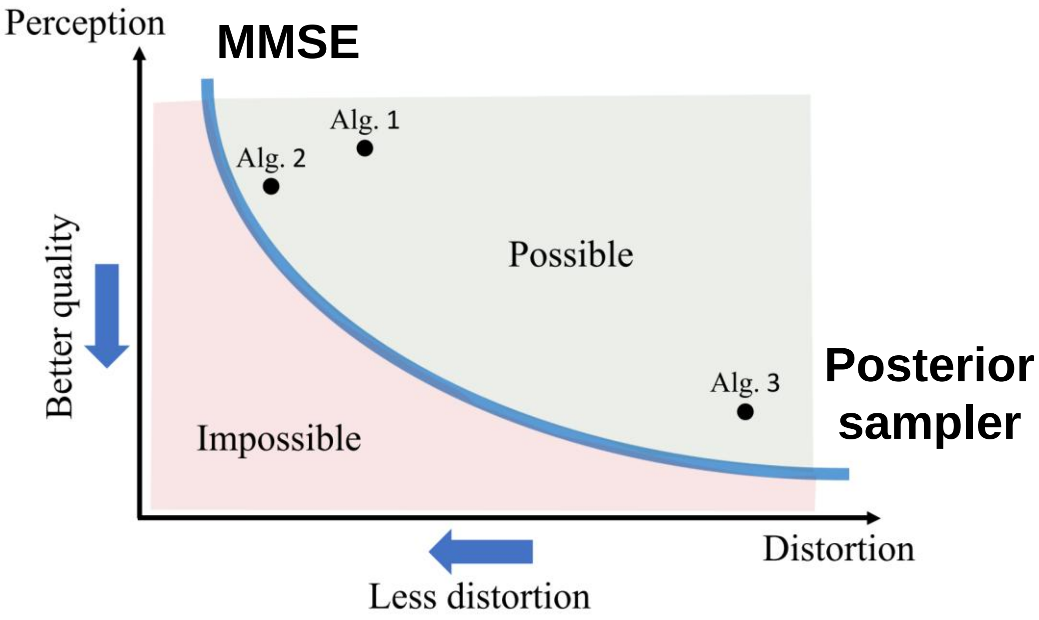
We will focus on two estimators

- **MMSE** $f^*(\boldsymbol{y}) = \mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}\}$

obtained via $f^* = \arg\min_f \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \, ||\boldsymbol{x} - f(\boldsymbol{y})||^2$

- **Posterior sampler** $f^*(\boldsymbol{y}) \sim p(\boldsymbol{x}|\boldsymbol{y})$

we can approximate with diffusion models



MMSE

Posterior sampler

Distortion-perception trade-off
[Blau and Michaeli, 2018]

Part 2: Learning MMSE estimators from noisy data

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$$

Self-supervised risk estimators

Supervised loss

$$\mathcal{L}_{\text{sup}}(\boldsymbol{x},\boldsymbol{y},f) = ||\boldsymbol{x}-f(\boldsymbol{y})||^2 = \underbrace{||\boldsymbol{y}-f(\boldsymbol{y})||^2}_{\substack{\text{Measurement}\\ \text{consistency}}} + \underbrace{2f(\boldsymbol{y})^\top(\boldsymbol{y}-\boldsymbol{x})}_{\substack{\text{key term to approximate!}\\ = \boldsymbol{f}(\boldsymbol{y})^\top\boldsymbol{\epsilon}}} + \text{const.}$$

Naïve loss doesn’t work!

$$\mathcal{L}_{\text{MC}}(\boldsymbol{y},f) = ||\boldsymbol{y}-f(\boldsymbol{y})||^2$$

$$\Rightarrow f^*(\boldsymbol{y}) = \boldsymbol{y}$$

Stein’s Unbiased Risk Estimator

- **Stein’s lemma** [Stein 1974] **:** Let $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{x}, I\sigma^2\right)$, f be weakly differentiable, then

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\,(\boldsymbol{y}-\boldsymbol{x})^\top f(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\,\sigma^2 \sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})$$

$$\boxed{\mathcal{L}_{\text{SURE}}(\boldsymbol{y}, f) = \underbrace{||\,\boldsymbol{y} - f(\boldsymbol{y})||^2}_{\text{Measurement consistency}} + \underbrace{2\sigma^2 \sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})}_{\text{Degrees of freedom [Efron, 2004]}}}$$

- **Hudson’s lemma** [Hudson 1978] extends this result for the exponential family (eg. **Poisson Noise**)
- Beyond exponential family: **Poisson-Gaussian noise** [Le Montagner et al., 2014]

[Raphan and Simoncelli, 2011]

Stein’s Unbiased Risk Estimator

Monte Carlo SURE [Efron 1975, Breiman 1992, Ramani et al., 2007]

SURE’s divergence is generally approximated as

$$\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) \approx \frac{\boldsymbol{\omega}^\top}{\alpha}\left(f(\boldsymbol{y}) - f(\boldsymbol{y} + \boldsymbol{\omega}\alpha)\right)$$

- **Recorrupted2Recorrupted** [Pang et al. CVPR 2021] [Monroy Bacca and Tachella, CVPR 2025].

$$\mathcal{L}_{\mathrm{R2R}}(\boldsymbol{y}, f) = ||\boldsymbol{y} + \alpha\boldsymbol{\omega} - f\left(\boldsymbol{y} - \frac{\boldsymbol{\omega}}{\alpha}\right)||^2$$

where $\alpha > 0$ small, $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, I)$

Stein’s Unbiased Risk Estimator

The solution to SURE is **Tweedie’s Formula**

$$\arg\min_{f}\ \mathbb{E}_{\boldsymbol{y}}||\,\boldsymbol{y}-f(\boldsymbol{y})||^2+2\sigma^2\sum_i\frac{\delta f_i}{\delta y_i}(\boldsymbol{y})$$

Integration by parts

$$\arg\min_{f}\ \mathbb{E}_{\boldsymbol{y}}\,||\,\boldsymbol{y}-f(\boldsymbol{y})||^2-2\sigma^2\sum_i f(\boldsymbol{y})\frac{\delta\log p_{\boldsymbol{y}}(\boldsymbol{y})}{\delta y_i}$$

Complete squares

$$\arg\min_{f}\ \mathbb{E}_{\boldsymbol{y}}\,||\,f(\boldsymbol{y})-\boldsymbol{y}-\sigma^2\nabla\log p_{\boldsymbol{y}}(\boldsymbol{y})\,||^2$$

$$\Rightarrow\quad f(\boldsymbol{y})=\boldsymbol{y}+\sigma^2\nabla\log p_{\boldsymbol{y}}(\boldsymbol{y})$$

- Key formula behind diffusion models, which can be trained self-supervised

Learning posterior samplers from noisy data

Model Identification

- Can we actually learn a clean distribution from noisy samples?
- Model identification is a **linear** inverse problem in **infinite** dimensions

$$p_y(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x}) p_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}$$

$$\boxed{p_{\boldsymbol{y}} = \boldsymbol{\mathcal{A}}(p_{\boldsymbol{x}})}$$

- Here we assume access to p_y, however, in practice we only have finite observations $\hat{p}_{\boldsymbol{y}} = \sum_{i=1}^{N} \delta_{\boldsymbol{y}_i}$

Can we learn with noise?

Noisy measurement setting $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$

- For additive noise $p(\boldsymbol{y}|\boldsymbol{x}) = g(\boldsymbol{x} - \boldsymbol{y})$:

$$p_{\boldsymbol{y}} = \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\sigma^2\right) * p_{\boldsymbol{x}}$$

- This is a **deconvolution** problem!
- In Fourier we have, $\phi_{\boldsymbol{y}}(\boldsymbol{\omega}) = \phi_{\boldsymbol{x}}(\boldsymbol{\omega})\ \hat{g}(\boldsymbol{\omega})$ where $\phi_{\boldsymbol{x}}$ and $\phi_{\boldsymbol{y}}$ are the characteristic functions of $p_{\boldsymbol{x}}$ and $p_{\boldsymbol{y}}$, and $\hat{g}$ is the Fourier transform of g.

Can we learn with noise?

- Since $\mathcal{N}(\mathbf{0}, I\sigma^2)$ is an invertible kernel $\hat{g}(\boldsymbol{\omega}) \neq 0$ for all $\boldsymbol{\omega}$, we can identify $p_{\boldsymbol{x}}$ from $p_{\boldsymbol{y}}$

Proposition [T. et al., 2023]: For additive noise with nowhere zero characteristic function, it is possible to uniquely identify $p_{\boldsymbol{x}}$ from $p_{\boldsymbol{y}}$.

- For non-additive noise (eg. Poisson), the problem is slightly harder

Diffusion models

Diffusion model SDE:

We need this!

$$d\boldsymbol{x} = -2\dot{\sigma}_t \frac{\mathbb{E}\{\boldsymbol{x}|\boldsymbol{x} + \sigma\boldsymbol{\epsilon}\} - \boldsymbol{x}}{\sigma_t} dt + \sqrt{2\dot{\sigma}_t\sigma_t}\, d\boldsymbol{\omega}_t$$

where $\boldsymbol{\omega}_t$ is a Brownian noise process and $t \in (0,1)$

- We need the MMSE estimator $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{x} + \sigma\boldsymbol{\epsilon}\}$ for all $\sigma > 0$
- With dataset $\{\boldsymbol{y}_i = \boldsymbol{x}_i + \sigma_n\boldsymbol{\epsilon}_i\}_{i=1:N}$ self-sup methods learn estimator for $\sigma \geq \sigma_n$ only!

Consistent diffusion

- We need the MMSE estimator $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{x}+\sigma\boldsymbol{\epsilon}\}$ for all $\sigma > 0$

First solution proposed by [Daras et al., ICML 2024]

- **Idea in a nutshell:**
  1. Learn self-sup denoiser for $\sigma \geq \sigma_n$
  2. Run diffusion steps up to $\sigma = \sigma_n - \Delta$ for small Δ to generate less noisy data
  3. Train the model with this less noisy data
  4. Iterate

- **Problem:** requires a small supervised dataset to work [Daras, ICLR 2025]

Normalization equivariance

- We need the MMSE estimator $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{x}+\sigma\boldsymbol{\epsilon}\}$ for all $\sigma > 0$

- **Idea:** if signal distribution is scale invariant $p(\alpha\boldsymbol{x}+\boldsymbol{1}\mu) \approx p(\boldsymbol{x})$ for $\alpha > 0, \mu > 0$ [Levac et al., 2025]

$$\boldsymbol{y} = \boldsymbol{x} + \sigma_n\boldsymbol{\epsilon}$$

$$\alpha\boldsymbol{y} + \boldsymbol{1}\mu = \alpha\boldsymbol{x} + \alpha\sigma_n\boldsymbol{\epsilon} + \boldsymbol{1}\mu$$

$$\boldsymbol{y}' = \boldsymbol{x}' + \sigma'\boldsymbol{\epsilon}$$

where $\sigma' = \alpha\sigma_n < \sigma$ is a smaller noise level and $\boldsymbol{x}' = \alpha\boldsymbol{x} + \boldsymbol{1}\mu$ is a valid signal

- Has been used to improve generalization of denoisers [Mohan 2020, Herbreteau 2024]

Normalization equivariance

- We need the MMSE estimator $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{x}+\sigma\boldsymbol{\epsilon}\}$ for all $\sigma > 0, \mu$

We look for normalization equivariant denoisers

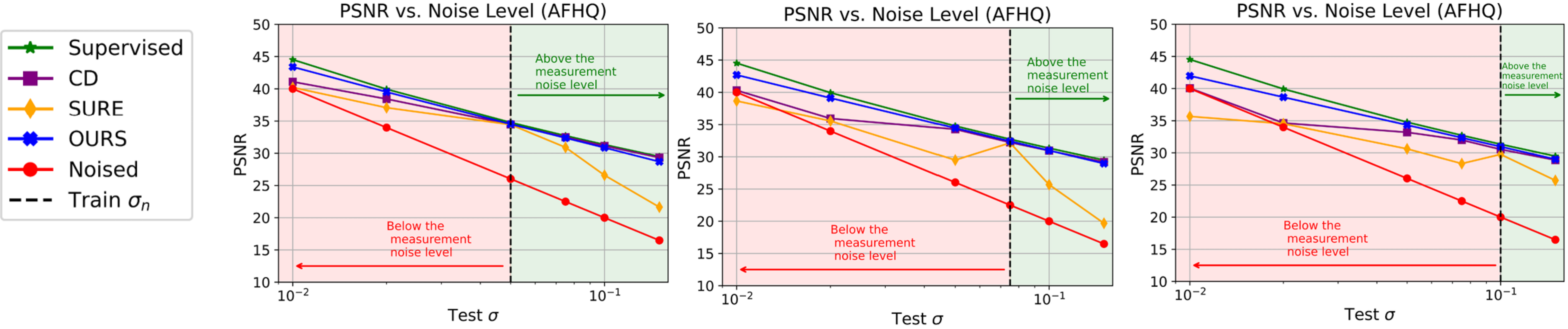
$$f_{\alpha\sigma}(\alpha\boldsymbol{y}+\mathbf{1}\mu) = \alpha f_{\sigma}(\boldsymbol{y}) + \mathbf{1}\mu$$

- We can achieve this property by
  - 1. Normalization equivariant architectures [Herbreteau, NeurIPS 2024]
  - 2. Adapting the loss (our work) [Levac et al., 2025]

$$\boxed{\mathcal{L}_{\mathrm{N-SURE}}(\boldsymbol{y}, f) = \mathbb{E}_{\alpha,\mu}||\,\alpha\boldsymbol{y}+\mathbf{1}\mu - f(\alpha\boldsymbol{y}+\mathbf{1}\mu)||^2 + 2(\alpha\sigma)^2\sum_i \frac{\delta f_i}{\delta y_i}(\alpha\boldsymbol{y}+\mathbf{1}\mu)}$$

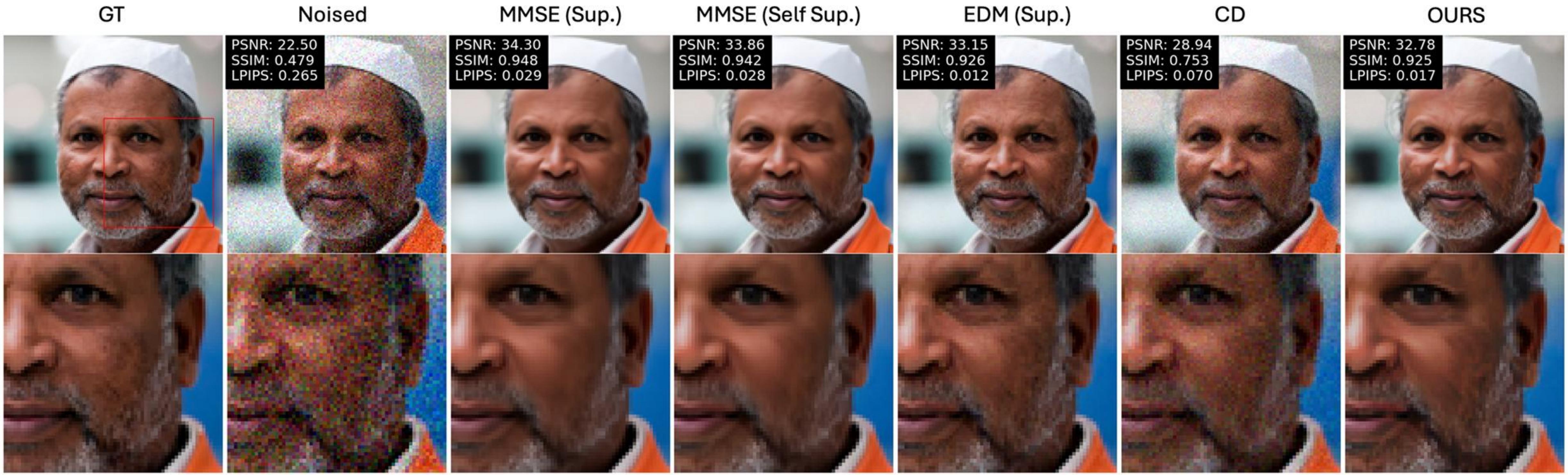
Experiments

- Learned denoiser for $f_\sigma(\boldsymbol{x} + \sigma\boldsymbol{\epsilon})$ for $\sigma > 0$



Experiments

- FFHQ dataset, $\sigma_n = 0.075$



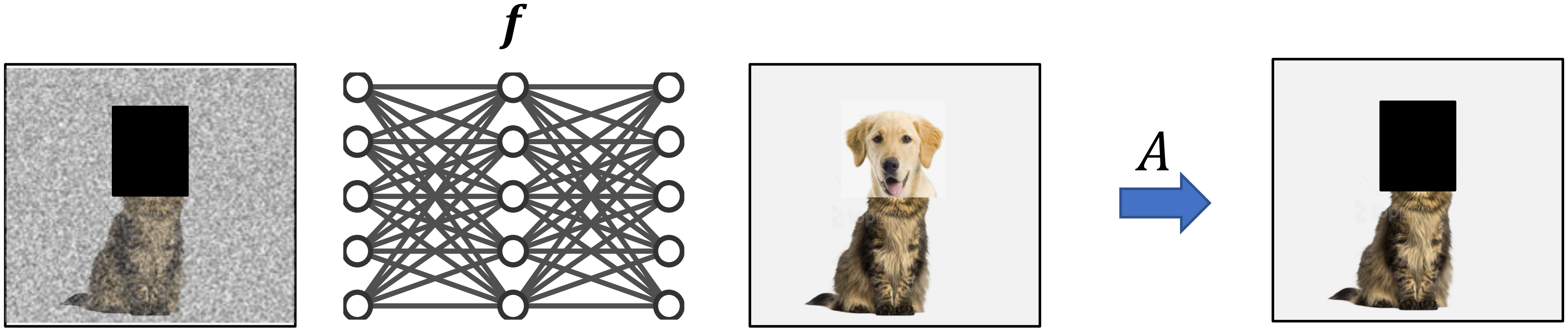
Part 2: Learning from incomplete data

Incomplete measurements?

For $A \neq I$, most estimators can be adapted to approximate

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\ ||A(\boldsymbol{x} - f(\boldsymbol{y}))||^2$$

In this case, the risk does not penalise $f(\boldsymbol{y})$ in the **nullspace** of A!



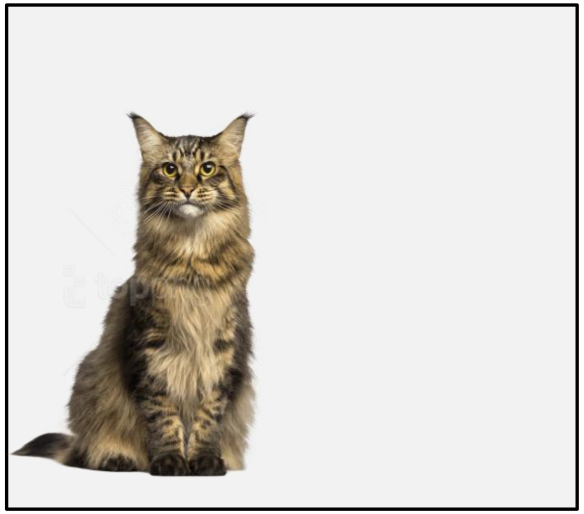
Symmetry Prior

Idea: Most natural signals sets $\mathcal{X}$ are invariant to groups of transformations.

Example: natural images are translation invariant

- Mathematically, a set $\mathcal{X}$ is invariant to $\{T_g \in \mathbb{R}^{n\times n}\}_{g\in G}$ if

$$\forall \boldsymbol{x} \in \mathcal{X},\ \ \forall g \in G,\ T_g\boldsymbol{x} \in \mathcal{X}$$



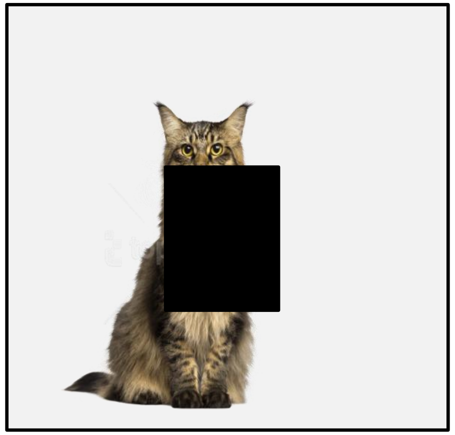
Other symmetries: rotations, permutation, amplitude

Symmetry prior

Equivariant Imaging [Chen, Davies and Tachella, ICCV 2021]

For all $g \in G$ we have

$$\boldsymbol{y} = A\boldsymbol{x} = \underbrace{AT_g}_{A_g}\overbrace{T_g^{-1}\boldsymbol{x}}^{x'} = A_g\boldsymbol{x}'$$



- We get multiple virtual operators $\{A_g\}_{g\in G}$ ‘for free’!
- Each AT_g might have a different nullspace

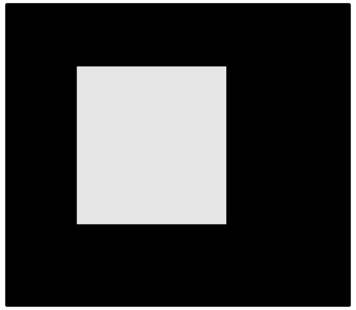
Necessary condition

Proposition [T. et al., 2023]: Learning reconstruction mapping f from observed measurements possible only if

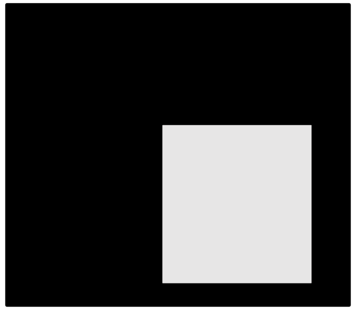
$$\mathrm{rank}\left(\mathbb{E}_g\, T_g^\top A^\top A T_g\right) = n,$$

and thus if $m \geq \max \frac{c_j}{s_j} \geq \frac{n}{|G|}$ where s_j and c_j are dimension and multiplicity of irreps.

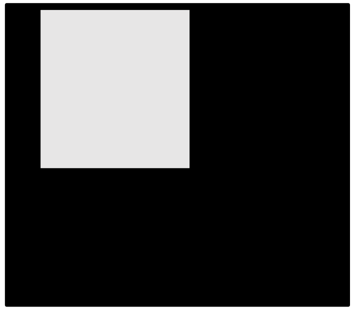
A_1



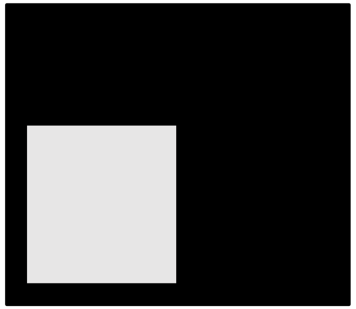
A_2



A_3



A_4



(Non)-Equivariant operators

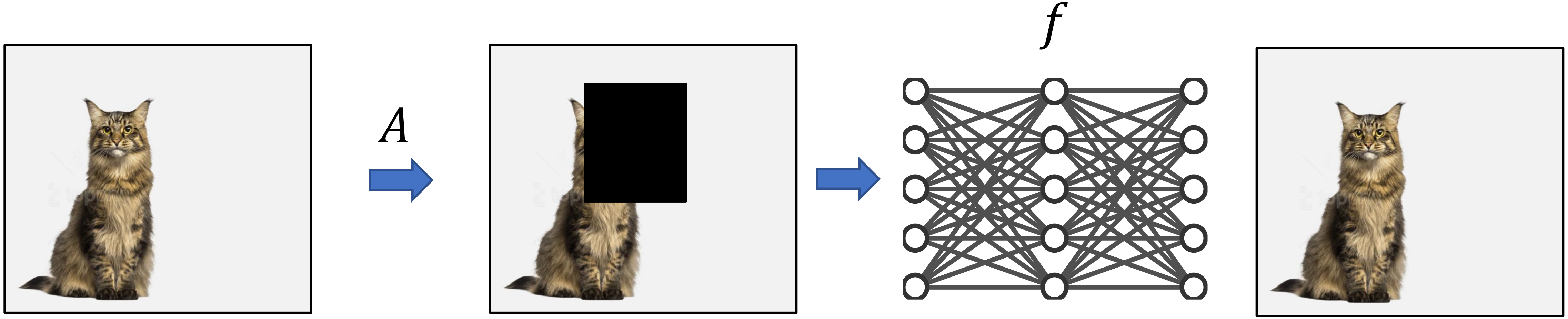
Theorem [T. et al., 2023]*:* The full rank condition requires that A **is not equivariant:** $AT_g \neq \tilde{T}_g A$

$$\mathrm{rank}\big(\mathbb{E}_g\, T_g^\top A^\top A T_g\big) = \mathrm{rank}\big(A^\top (\mathbb{E}_g \tilde{T}_g^\top \tilde{T}_g) A\big) = \mathrm{rank}\big(A^\top A\big) = m < n$$

Equivariant imaging

How can we enforce equivariance in practice?

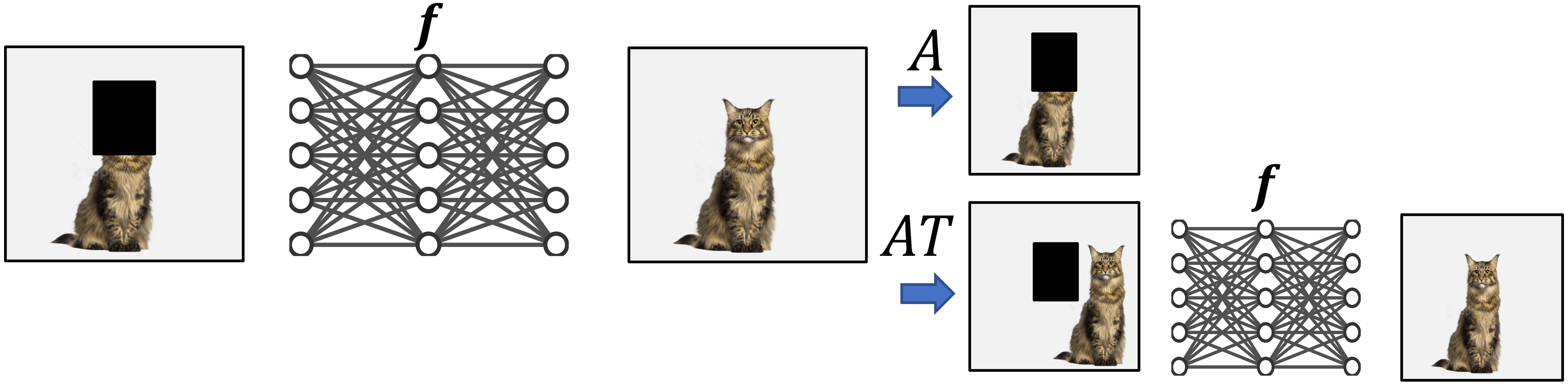
Idea: we should have $f\left(AT_g\boldsymbol{x}\right) = T_g f(A\boldsymbol{x})$, i.e. $f \circ A$ should be G-equivariant



Equivariant imaging

$$\mathcal{L}(\boldsymbol{y},f) = \underbrace{\mathcal{L}_{\mathrm{N-SURE}}(\boldsymbol{y},f)}_{\text{Measurement consistency}} + \underbrace{\mathbb{E}_g\, ||\, T_g\hat{\boldsymbol{x}} - f\left(AT_g\hat{\boldsymbol{x}}\right)||^2}_{\text{enforces equivariance of } f\circ A}$$

where $\hat{\boldsymbol{x}} = f(\boldsymbol{y})$ is used as reference



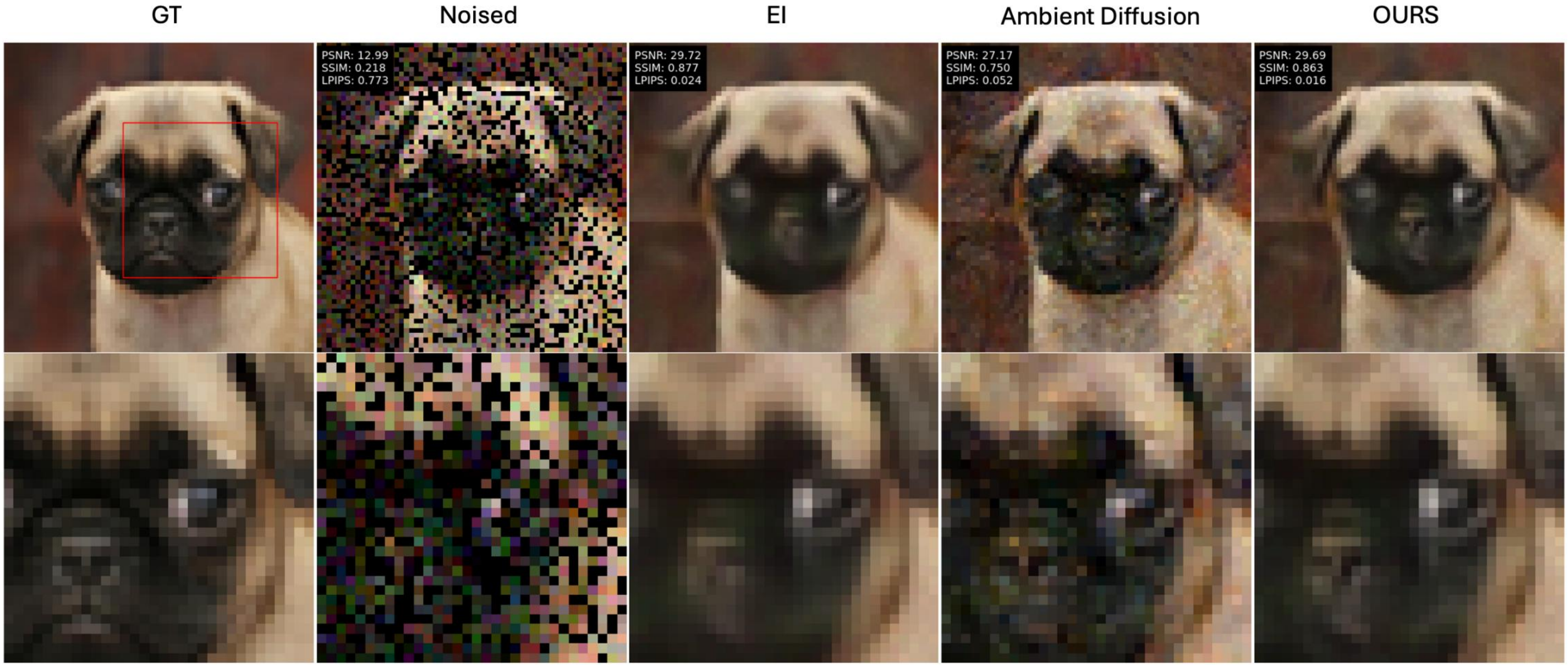
Diffusion model

- Using the previous loss, we learn $f_\sigma(A\boldsymbol{x} + \sigma\boldsymbol{\epsilon}) \approx \mathbb{E}\{\boldsymbol{x} | A\boldsymbol{x} + \sigma\boldsymbol{\epsilon}\}$
- How can we run a diffusion with this estimator?
- MMSE denoiser in measurement space $Af_\sigma(A\boldsymbol{x} + \sigma\boldsymbol{\epsilon}) \approx \mathbb{E}\{A\boldsymbol{x} | A\boldsymbol{x} + \sigma\boldsymbol{\epsilon}\}$

- Idea:
  1. Using $Af_\sigma,$ run diffusion measurement space from σ_n to $\sigma = 0$
  2. Using $f_\sigma,$ reconstruct sampled measurement
  3. If A is one-to-one over the set of signals, we obtain a true posterior sample

Experiments

- AFHQ dataset, inpainting problem $\sigma_n = 0.1$



Conclusions

We address one of the main challenges in self-supervised learning:

Learning to posterior samplers without ground-truth data

Key ideas:

- Use existing theory for learning MMSE estimators
- Leverage invariance to scaling (and transformations)

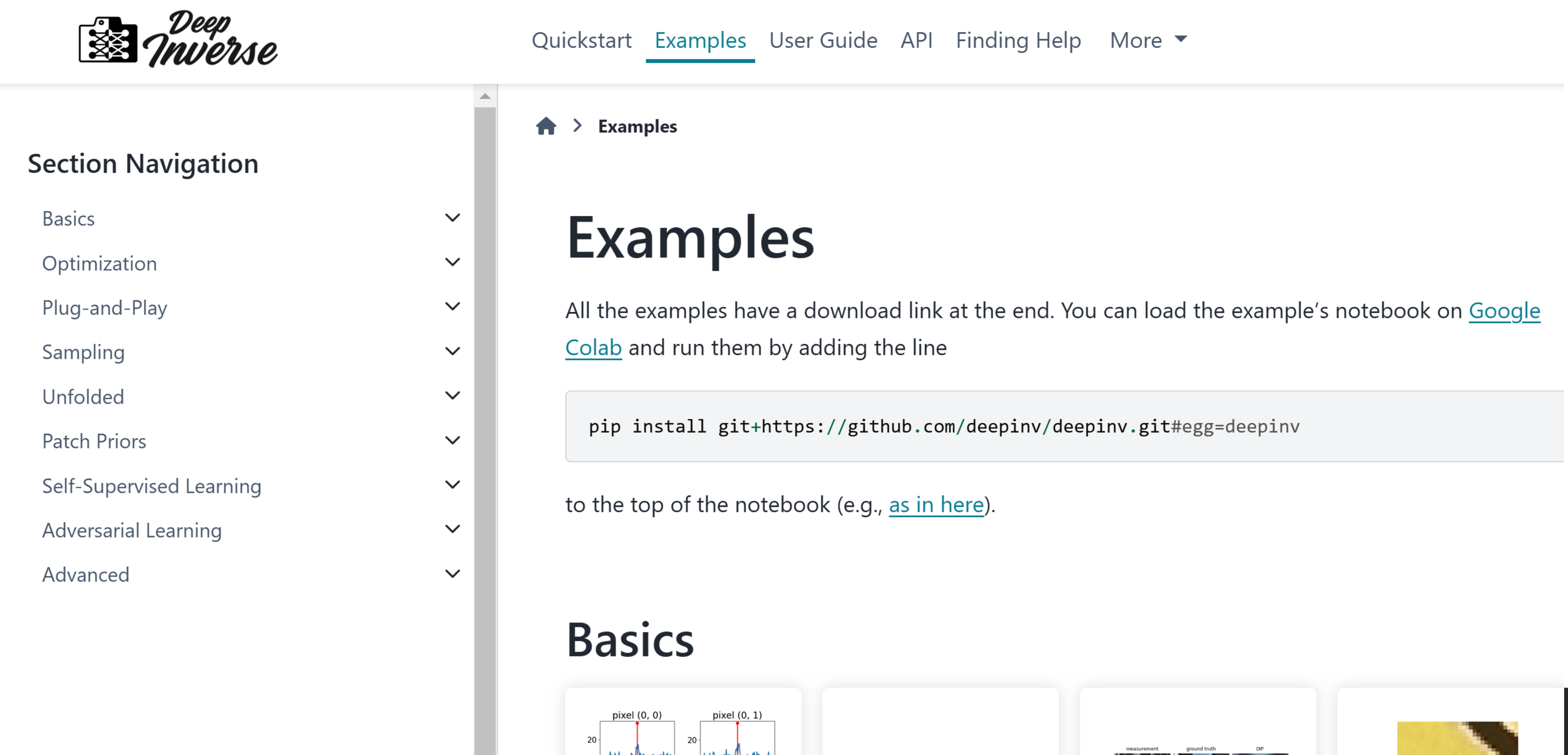
Remaining challenges:

Can we still learn samplers in highly incomplete and/or highly noisy cases?

- Requires ground-truth
- Not useful scientific/medical imaging

nty

Deep Inverse



Section Navigation

- Basics
- Optimization
- Plug-and-Play
- Sampling
- Unfolded
- Patch Priors
- Self-Supervised Learning
- Adversarial Learning
- Advanced

Examples

Examples

All the examples have a download link at the end. You can load the example's notebook on Google Colab and run them by adding the line

```
pip install git+https://github.com/deepinv/deepinv.git#egg=deepinv
```

to the top of the notebook (e.g., as in here).

Basics

Deep Inverse

cnrs Inria THE UNIVERSITY OF EDINBURGH 1794 ENS PSL ENS DE LYON EPFL

References

Paper: https://arxiv.org/abs/2510.11964

Self-sup references:

https://tachella.github.io/projects/selfsuptutorial/

Code examples:

https://deepinv.github.io/deepinv/auto_examples/self-supervised-learning/index.html

YouTube version (3 hours):

https://youtu.be/gf-WCHXAdfk?si=bRC6Pq0WpZHNrRLU