UNSURE: Unknown Noise level Stein’s Unbiased Risk Estimator

Julián Tachella, CNRS, École Normale Supérieure de Lyon

Joint work with Mike Davies (University of Edinburgh) and Laurent Jacques (UCLouvain)

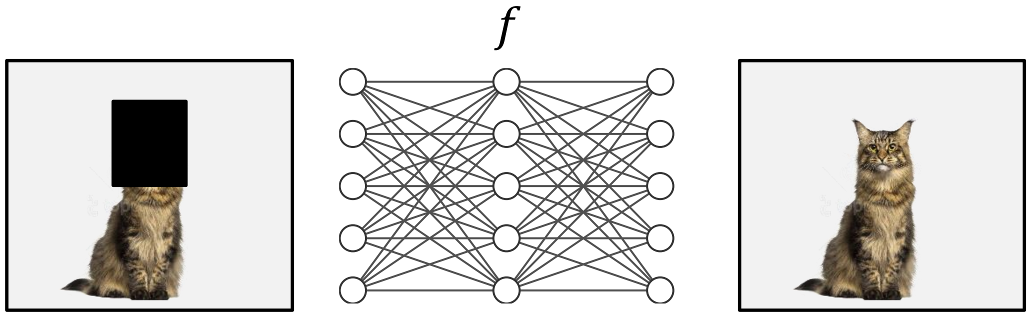
Inverse Problems

Goal: recover signal x from y

$$\underbrace{\boldsymbol{y}}_{\text{MEASUREMENTS} \in \mathbb{R}^{m}} = \overbrace{A}^{\text{MEASUREMENT OPERATOR} \in \mathbb{R}^{m \times n}}\overbrace{\boldsymbol{x}}^{\text{SIGNAL} \in \mathbb{R}^{n}} + \underbrace{\boldsymbol{\epsilon}}_{\text{NOISE} \in \mathbb{R}^{m}}$$

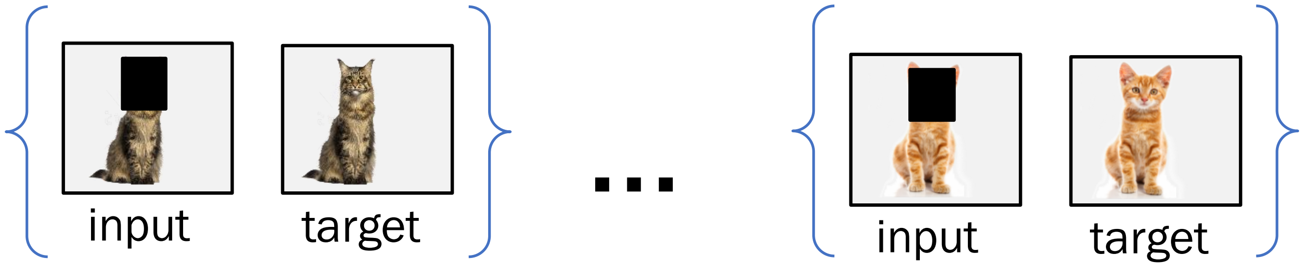
Learning approach

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



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

supervised dataset



Learning approach

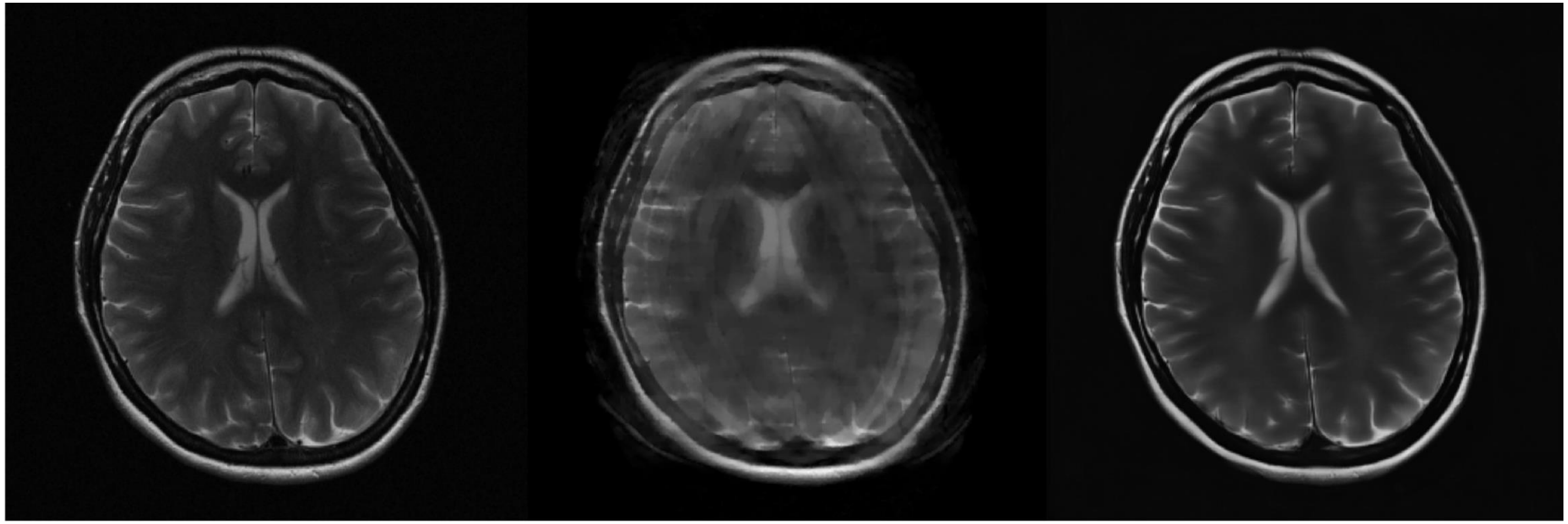
Advantages:

- 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**)

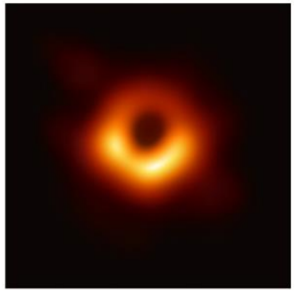
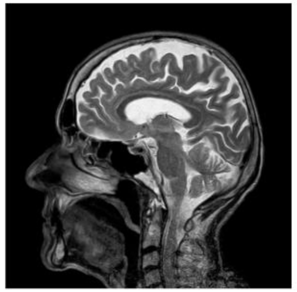


x8 accelerated MRI [Zbontar et al., 2019]

Learning approach

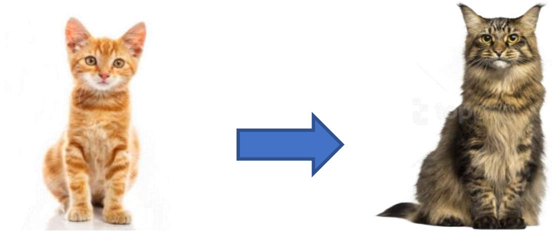
Main disadvantage: Obtaining training signals x_i can be expensive or impossible.

- Medical and scientific imaging

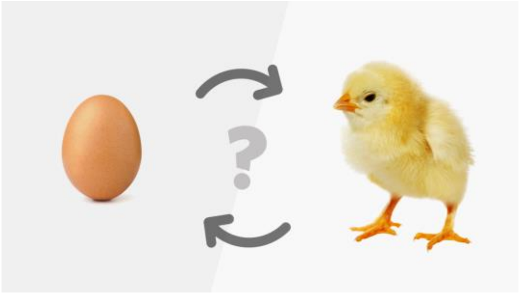



- Only solves inverse problems which we already know what to expect

- Risk of training with signals from a different distribution



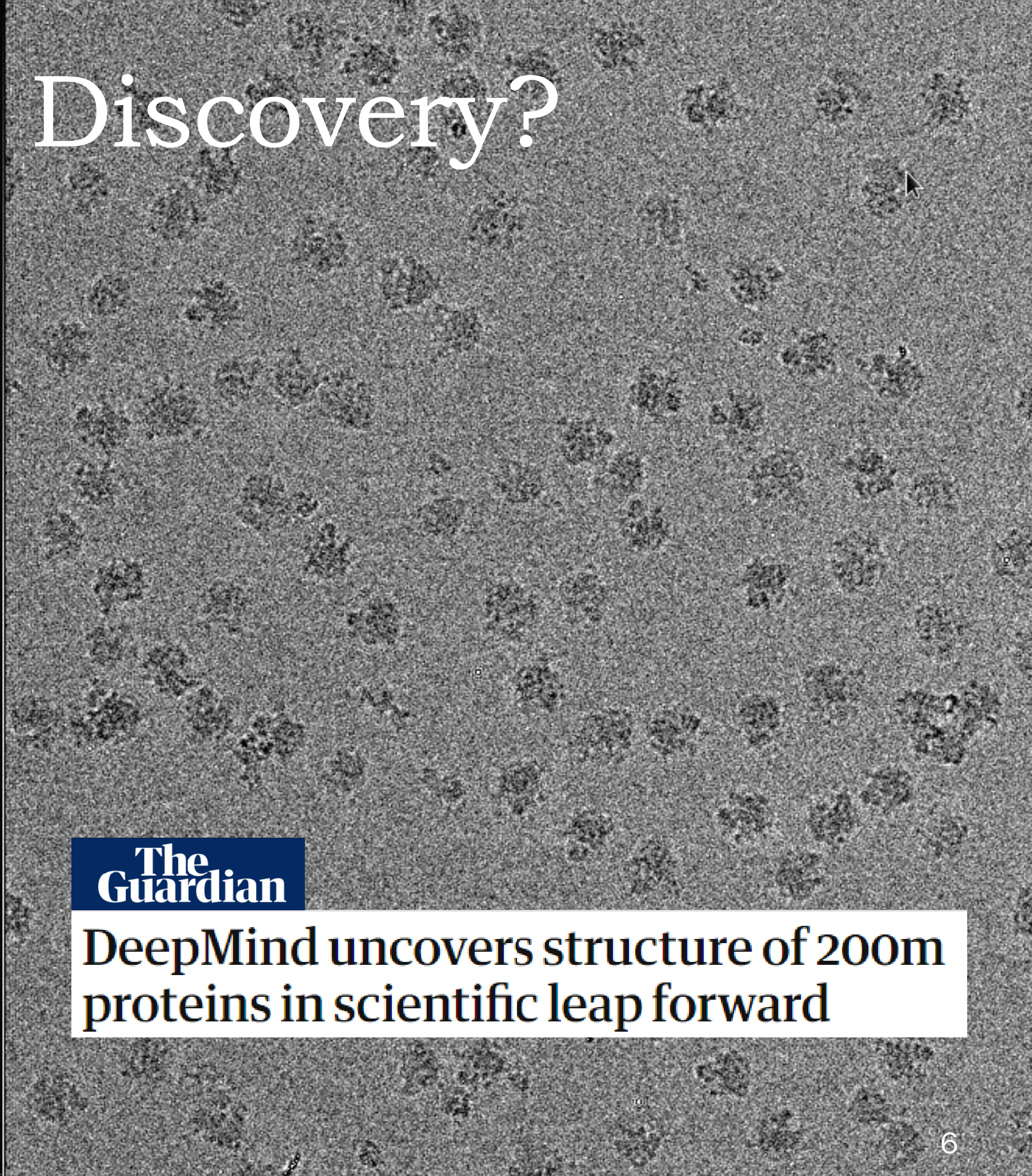
train *test?*



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

Denoising problems

In this first part, we will focus on ‘denoising’ problems

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

where $A \in \mathbb{R}^{m \times n}$ is invertible (and thus $m \geq n$).

- We focus on $\boldsymbol{A} = \boldsymbol{I}$ for simplicity.
- All methods in this part can be extended to any invertible $\boldsymbol{A}$.

MMSE estimators

We focus on ℓ_2 loss and minimum mean squared error estimators (MMSE)

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

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

- Other estimators might be preferred, eg. perceptual [Blau and Michaeli, 2018]

Unsupervised Risk Estimators

Supervised loss

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

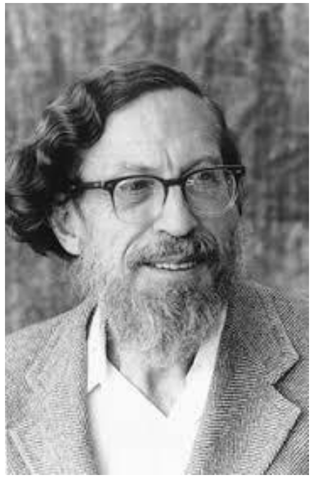
Goal: build a self-supervised loss $\mathcal{L}_{\text{self}}$ such that

$$\boxed{\mathbb{E}_{\boldsymbol{y}}\, \mathcal{L}_{\text{self}}(\boldsymbol{y},f) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathcal{L}_{\text{SUP}}(\boldsymbol{x},\boldsymbol{y},f) + \text{const.}}$$

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

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



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

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

Stein's Unbiased Risk Estimator

- **Hudson's lemma** [Hudson 1978] **:** Let $\boldsymbol{y}|\boldsymbol{x}$ exponential family, f be weakly differentiable, then

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

- Gaussian noise: $a(y) = \sigma^2$
- Poisson noise: $a(y) \approx y$

Tweedie's Formula

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

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

Integration by parts

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

Complete squares

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

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

- **Noise2Score** [Kim and Ye, 2021] learns $\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$ from noisy data + denoises with Tweedie.
- Key formula behind diffusion models, which can be trained self-supervised [Daras et al., 2024]

Cross-Validation Methods

What happens if we only know that $p(\boldsymbol{y}|\boldsymbol{x}) = \prod p(y_i|x_i)$?

$$\boxed{\min_f || \boldsymbol{y} - f(\boldsymbol{y})||^2 \text{ subject to } \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0 \ \forall i, \forall \boldsymbol{y}}$$

- SURE's perspective:

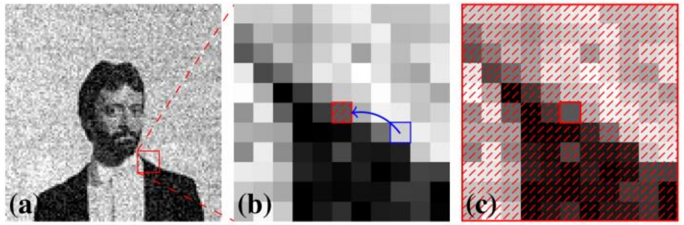
$$\min_f || \boldsymbol{y} - f(\boldsymbol{y})||^2 + \cancel{2\sigma^2 \sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})}$$

- These methods are not MMSE optimal!
- f_i shouldn't depend on y_i: training or architecture

Cross-Validation Methods

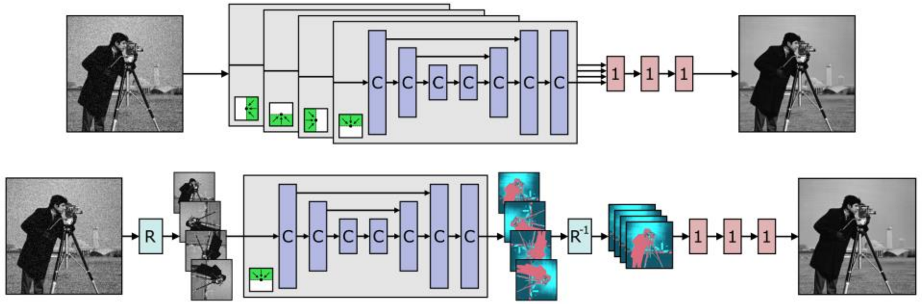
Noise2Void [Krull et al., 2019], **Noise2Self** [Batson, 2019], **Neighbor2Neighbor** [Huang, 2023]

- During training flip centre pixel
- Computes loss only on flipped pixels



Blind spot networks [Laine et al., 2019], [Lee et al., 2022]

- Convolutional architecture that doesn’t ‘see’ centre pixel by construction



Are We Missing Something?

SURE, R2R

???

Noise2Self, Noise2Inverse, BlindSpot Nets

$\frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0 \ \forall \boldsymbol{y}$

Less expressive estimators

$\boldsymbol{y} = \boldsymbol{x} + \sigma \boldsymbol{\epsilon}$
known σ

$\boldsymbol{y} = \boldsymbol{x} + \sigma \boldsymbol{\epsilon}$
unknown σ

$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$
unknown $\boldsymbol{\sigma}$

$\boldsymbol{y} = \gamma \mathcal{P}(\frac{\boldsymbol{x}}{\gamma}) + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$
unknown $\boldsymbol{\sigma}, \gamma$

$\boldsymbol{y}|\boldsymbol{x} \sim \prod_{i=1}^{n} p_i(y_i|x_i)$, $\mathbb{E}\, \boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}$
unknown p_i

Zero Expected Divergence

Assumption: Let $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, I\sigma^2)$.

$$\boxed{\mathcal{L}_{\mathrm{ZED}}(\boldsymbol{y}, f) = ||\,\boldsymbol{y} - f(\boldsymbol{y})||^2 \text{ subject to } \mathbb{E}_{\boldsymbol{y}} \textstyle\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0}$$

- SURE's perspective: $$\mathcal{L}_{\mathrm{ZED}}(\boldsymbol{y}, f) = ||\,\boldsymbol{y} - f(\boldsymbol{y})||^2 + 2\sigma^2 \cancel{\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})}$$
- Not MMSE optimal (but almost)

ZED Denoisers

Assumption: Let $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, I\sigma^2)$.

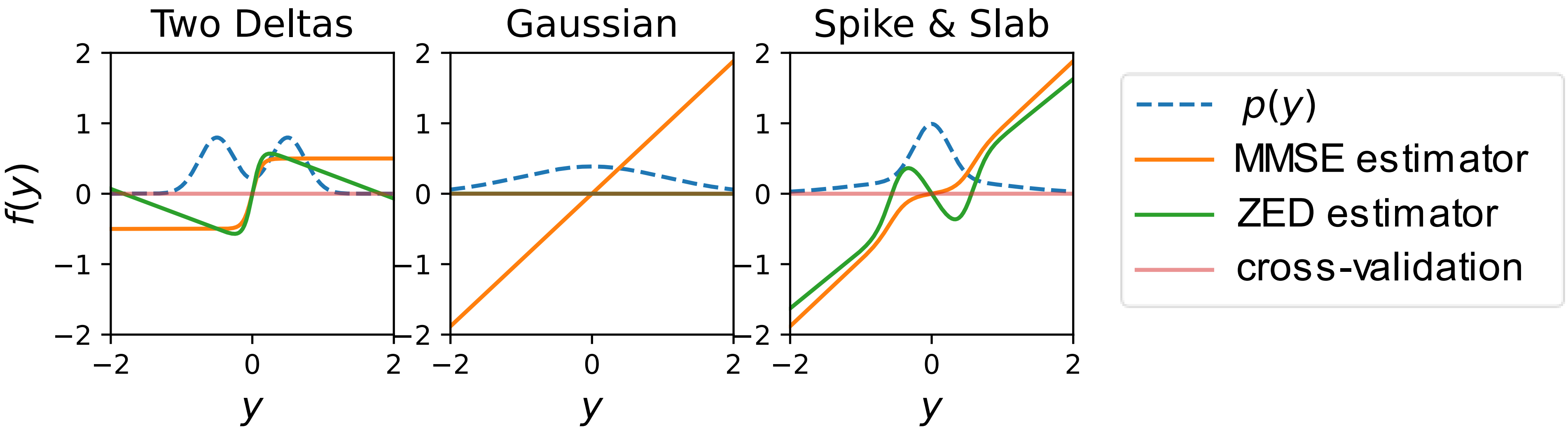
$$\min_f \mathbb{E}_{\boldsymbol{y}} || \boldsymbol{y} - f(\boldsymbol{y})||^2 \text{ subject to } \textstyle\sum_i \mathbb{E}_{\boldsymbol{y}} \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0$$

$$\max_{\eta \geq 0} \min_f \mathbb{E}_{\boldsymbol{y}} || \boldsymbol{y} - f(\boldsymbol{y})||^2 + 2\eta \sum_i \mathbb{E}_{\boldsymbol{y}} \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})$$

$\Rightarrow$ $f^{\mathrm{ZED}}(\boldsymbol{y}) = \boldsymbol{y} + \hat{\eta}\, \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$ $\quad \hat{\eta} = \left(\frac{1}{n} \mathbb{E}_{\boldsymbol{y}} ||\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})||^2\right)^{-1}$

ZED Denoisers

Examples: separable prior $p(\boldsymbol{x}) = \prod_i q(x_i)$



ZED Denoisers

How far is f^{ZED} from MMSE optimality?

Theorem:

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}||\, f^{\mathrm{ZED}}(\boldsymbol{y}) - \boldsymbol{x}\,|| = \sigma^2\left(\frac{1}{1-\frac{\mathrm{MMSE}}{\sigma^2}} - 1\right) \approx \mathrm{MMSE} + \underbrace{\frac{\mathrm{MMSE}^2}{\sigma^2}}$$

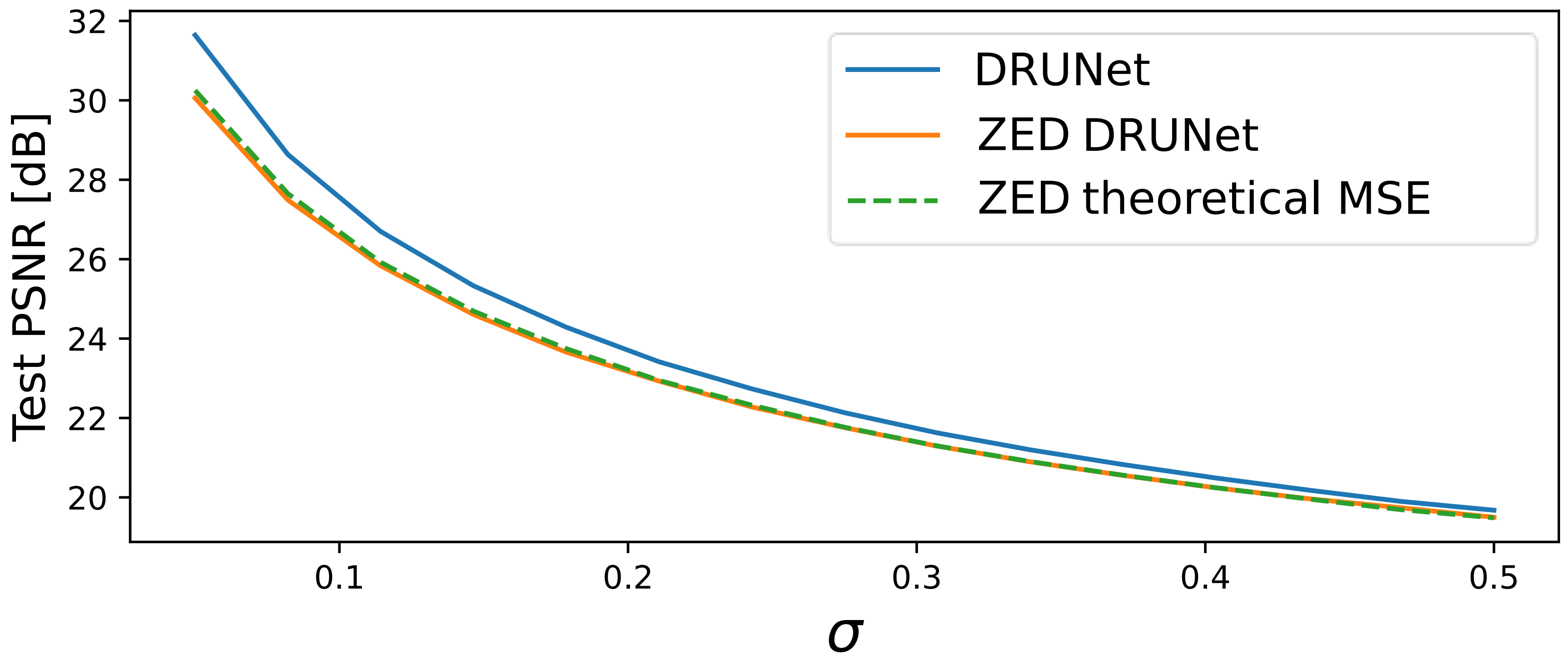
If supp. $p(\boldsymbol{x})$ is k-dimensional

$$\approx \left(\frac{k\sigma}{n}\right)^2$$

- $\hat{\eta} = \frac{\sigma^2}{1-\frac{\mathrm{MMSE}}{\sigma^2}}$ which serves as an estimator of the noise level

ZED Denoisers

Theorem is well verified by a pre-trained denoiser



Poisson-Gaussian Noise

- Noise model $\boldsymbol{y} = \gamma \boldsymbol{z} + \sigma \boldsymbol{\epsilon}$, $\boldsymbol{z} \sim \mathcal{P}(\boldsymbol{x}/\gamma)$ unknown σ, γ

$$\min_f \mathbb{E}_{\boldsymbol{y}} || \, \boldsymbol{y} - f(\boldsymbol{y}) ||^2 \text{ subject to } \textstyle\sum_i \mathbb{E}_{\boldsymbol{y}} \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0 \text{ and } \textstyle\sum_i \mathbb{E}_{\boldsymbol{y}} y_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0$$

$$\max_{\eta,\gamma} \min_f \mathbb{E}_{\boldsymbol{y}} || \, \boldsymbol{y} - f(\boldsymbol{y}) ||^2 + 2\eta \, \mathrm{div} f(\boldsymbol{y}) + 2\gamma \, \nabla f(\boldsymbol{y})^\top \boldsymbol{y}$$

- Closed-form solution for f in the paper

Unknown Correlations

Non-isotropic noise $\boldsymbol{y}|\boldsymbol{x} \sim N(\mathbf{0}, \Sigma)$ with unknown Σ

Define a set of possible covariances $R = \{\Sigma_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathbb{R}_+^p\}$

$$\max_{\boldsymbol{\eta} \geq \mathbf{0}} \min_f \mathbb{E}_{\boldsymbol{y}} || \boldsymbol{y} - f(\boldsymbol{y}) ||^2 + 2 \operatorname{tr}(\Sigma_{\boldsymbol{\eta}} \frac{\delta f}{\delta \boldsymbol{y}})$$

$$\Rightarrow \quad f(\boldsymbol{y}) = \boldsymbol{y} + \Sigma_{\widehat{\boldsymbol{\eta}}} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \qquad \widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta} \geq \mathbf{0}} \operatorname{tr}\left(\Sigma_{\boldsymbol{\eta}} \Sigma_{\boldsymbol{\eta}}^\top H\right) - 2\operatorname{tr}\left(\Sigma_{\boldsymbol{\eta}}\right)$$

with $H = \mathbb{E}_{\boldsymbol{y}} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})^\top$

Unknown Correlations

Practical examples:

- Unknown per-pixel noise level $[\Sigma_{\hat{\boldsymbol{\eta}}}]_{i,i} = \frac{1}{\mathbb{E}_{\boldsymbol{y}} \frac{\delta \log p_{\boldsymbol{y}}(\boldsymbol{y})}{\delta y_i}}$

- Unknown spatial correlation $\Sigma_{\hat{\boldsymbol{\eta}}} = \mathrm{circ}\, \hat{\boldsymbol{\eta}}$ with $\hat{\boldsymbol{\eta}} = F^{-1}(\frac{1}{F\boldsymbol{h}})$ and $\boldsymbol{h}$ the autocorrelation of the score up to $\pm r$ pixels

Summary

SURE, R2R

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

known σ

UNSURE

$\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}} \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0$

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

unknown σ

$\mathbb{E}_{\boldsymbol{y}} \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0$

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

unknown $\boldsymbol{\sigma}$

$\mathbb{E}_{\boldsymbol{y}} \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0$

$\mathbb{E}_{\boldsymbol{y}}\, y_i \frac{\delta f_i}{\delta y_i} = 0$

$\boldsymbol{y} = \gamma \mathcal{P}(\frac{\boldsymbol{x}}{\gamma}) + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$

unknown $\boldsymbol{\sigma}, \gamma$

Noise2Self, Noise2Inverse, BlindSpot Nets

$\frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = 0 \; \forall \boldsymbol{y}$

$\boldsymbol{y}|\boldsymbol{x} \sim \prod_{i=1}^{n} p_i(y_i|x_i)$, $\mathbb{E}\, \boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}$

unknown p_i

Less expressive estimators

Implementation

We parameterize f as a deep neural network, small $\alpha > 0$, $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, I)$

UNSURE: Solve Lagrangian problem, approximating divergence as [Ramani et al., 2007]

$$\mathrm{tr}(\Sigma \frac{\delta f}{\delta \boldsymbol{y}}) \approx \frac{(\Sigma \boldsymbol{\omega})^\top}{\alpha} \left(f(\boldsymbol{y}) - f(\boldsymbol{y} + \boldsymbol{\omega}\alpha) \right)$$

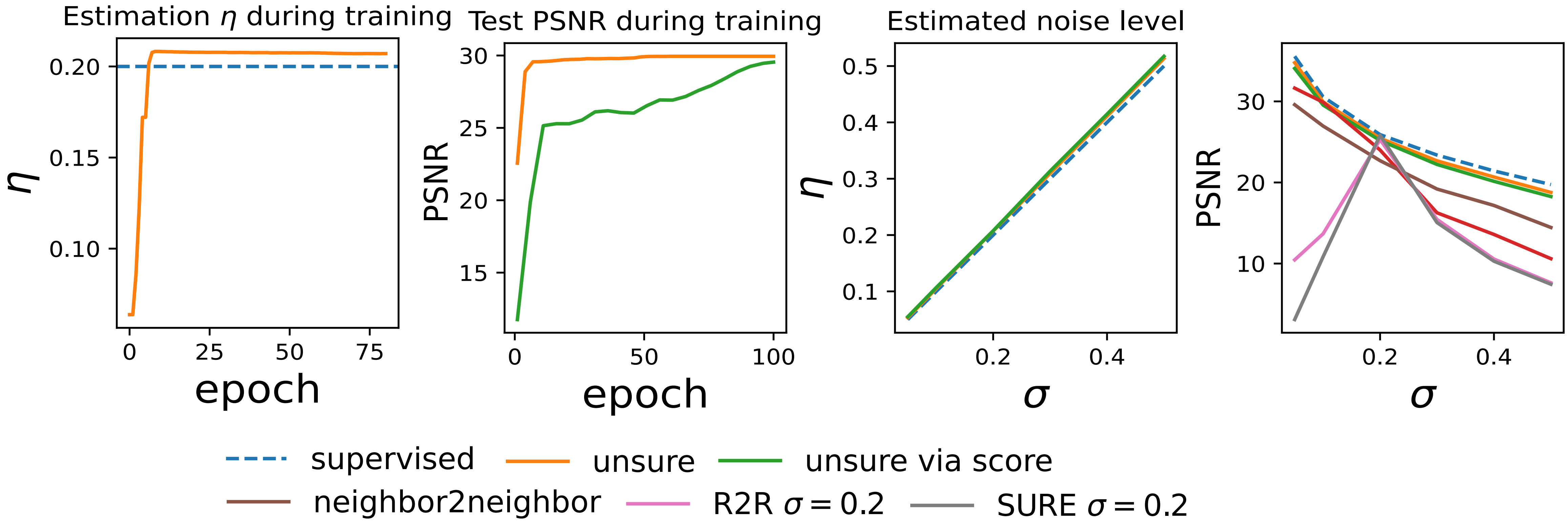
UNSURE via score: Learn score $s(\boldsymbol{y}) \approx \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$ via

$$\arg\min_{s} \mathbb{E}_{\boldsymbol{y},\boldsymbol{\omega}} ||\boldsymbol{\omega} - \alpha\, s(\boldsymbol{y} + \boldsymbol{\omega}\alpha)||^2$$

and use $f(\boldsymbol{y}) = \boldsymbol{y} + \Sigma_{\hat{\boldsymbol{\eta}}}\, s(\boldsymbol{y})$ to denoise at test time.

Experiments

- MNIST dataset
- Isotropic Gaussian noise



Experiments

- DIV2K dataset (320x320 RGB)
- Spatially correlated Gaussian noise

Method	Noise2Void	Neighbor2Neighbor	UNSURE (unknown Σ)	SURE (known Σ)	Supervised
PSNR [dB]	19.09 ± 1.79	23.61 ± 0.13	28.72 ± 1.03	29.77 ± 1.22	29.91 ± 1.26

Kernel size $\boldsymbol{\eta}$	1×1	3×3	5×5
PSNR [dB]	23.62	28.72 ± 1.03	27.38 ± 0.88

Ground-truth Measurement Noise2Void Neighbor2Neighbor UNSURE SURE knownΣ Supervised

Beyond Denoising

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

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

where $A^{\dagger}$ is the pseudoinverse of A.

For example, **GSURE** [Eldar, 2008] writes for Gaussian noise

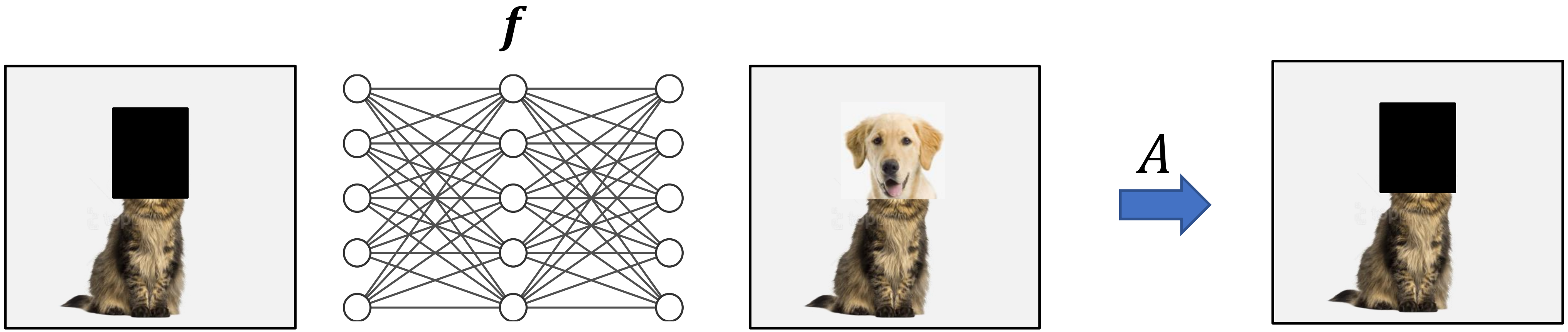
$$\boxed{\mathcal{L}_{GSURE}(\boldsymbol{y}, f) = ||A^{\dagger}\boldsymbol{y} - A^{\dagger}Af(\boldsymbol{y})||^2 + 2\sigma^2 \sum_i \frac{\delta\left[A^{\dagger}A \cdot f\right]_i}{\delta y_i}(\boldsymbol{y})}$$

Incomplete Measurements?

1. If A is invertible, we have $A^{\dagger}A = I$

2. If A is not invertible, $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,||A^{\dagger}A\big(\boldsymbol{x} - f(\boldsymbol{y})\big)||^2 \neq \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,||\boldsymbol{x} - f(\boldsymbol{y})||^2$

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

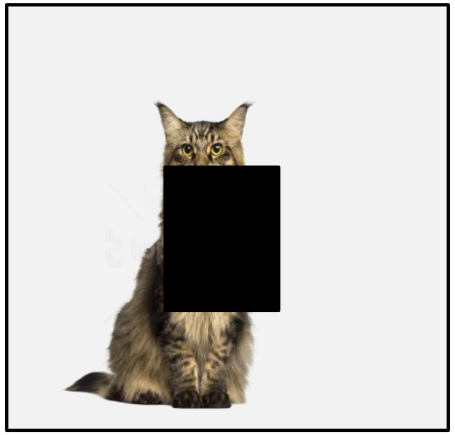


Symmetry Prior

Equivariant Imaging [Chen et al., 2021]

For all $g \in G$ we have

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

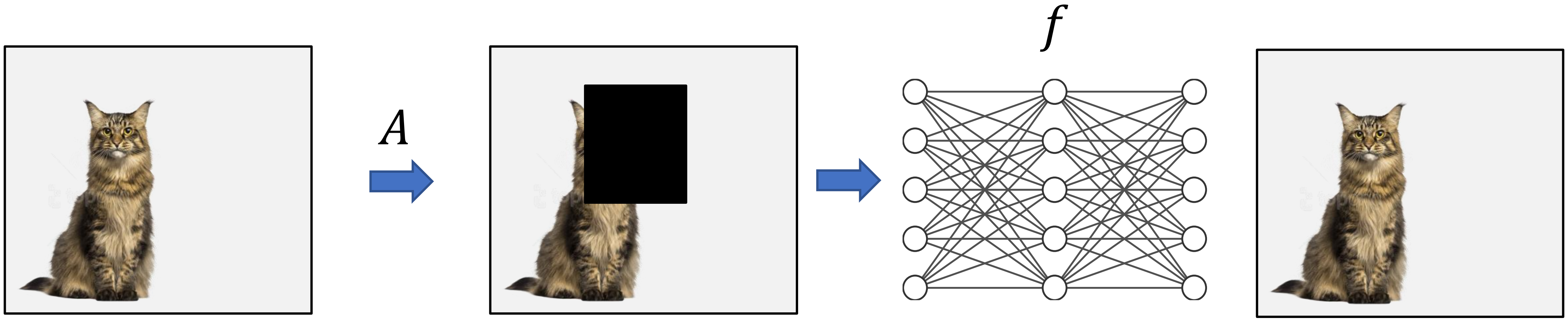


- We get multiple virtual operators $\{A_g\}_{g\in G}$ ‘for free’!
- Each AT_g might have a different nullspace

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

We can leverage invariance of X to transformations T_g to learn in the nullspace [Chen, 2021]

$$\boxed{\mathcal{L}_{EI}(\boldsymbol{y}, f) = \mathbb{E}_g \, || \, T_g\hat{\boldsymbol{x}} - f\left(AT_g\hat{\boldsymbol{x}}\right)||^2}$$

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

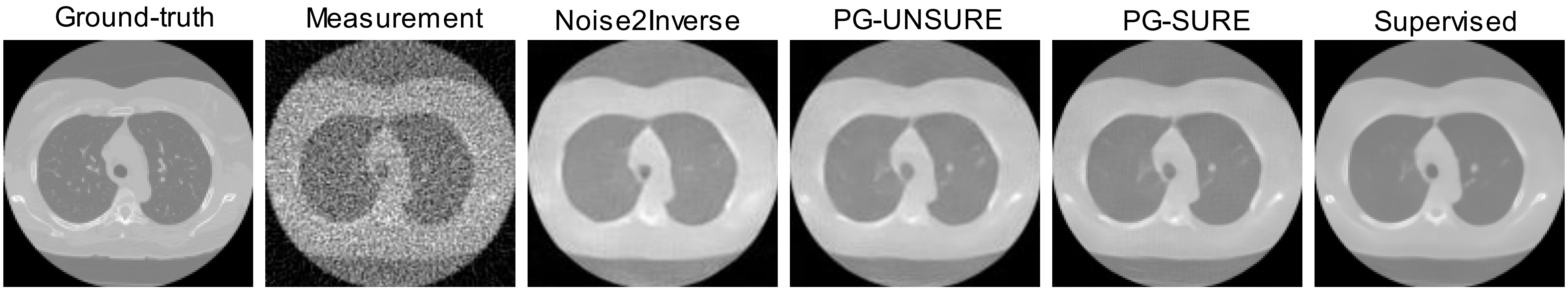
Robust Equivariant Imaging++

$$\mathcal{L}_{\text{REI}}(\boldsymbol{y}, f) = \underbrace{\mathcal{L}_{\text{UNSURE}}(\boldsymbol{y}, f)}_{\text{Handles noisy measurements of unknown noise level}} + \overbrace{\mathcal{L}_{\text{EI}}(\boldsymbol{y}, f)}^{\text{enforces equivariance of } f \circ A}$$

Experiments

- LIDC-IDRI dataset
- Poisson-Gaussian noise
- Uses equivariant imaging with rotations

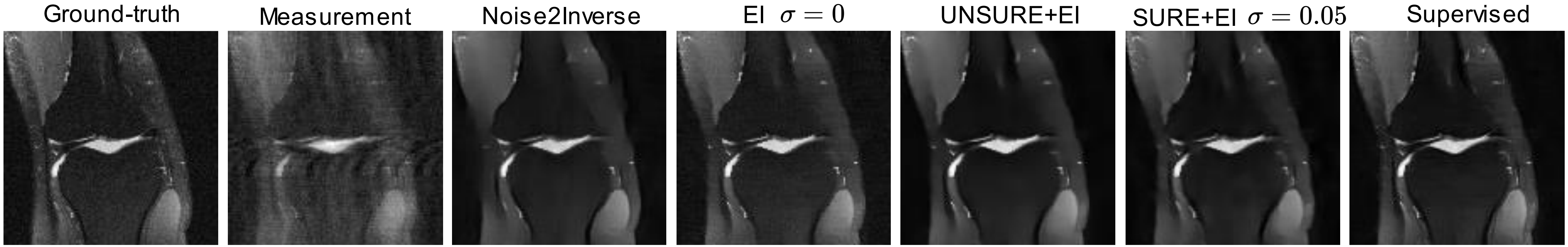
Method	Noise2Inverse	PG-UNSURE (unknown σ, γ)	PG-SURE (known σ, γ)	Supervised
PSNR [dB]	32.54 ± 0.71	33.31 ± 0.57	33.76 ± 0.61	34.67 ± 0.68

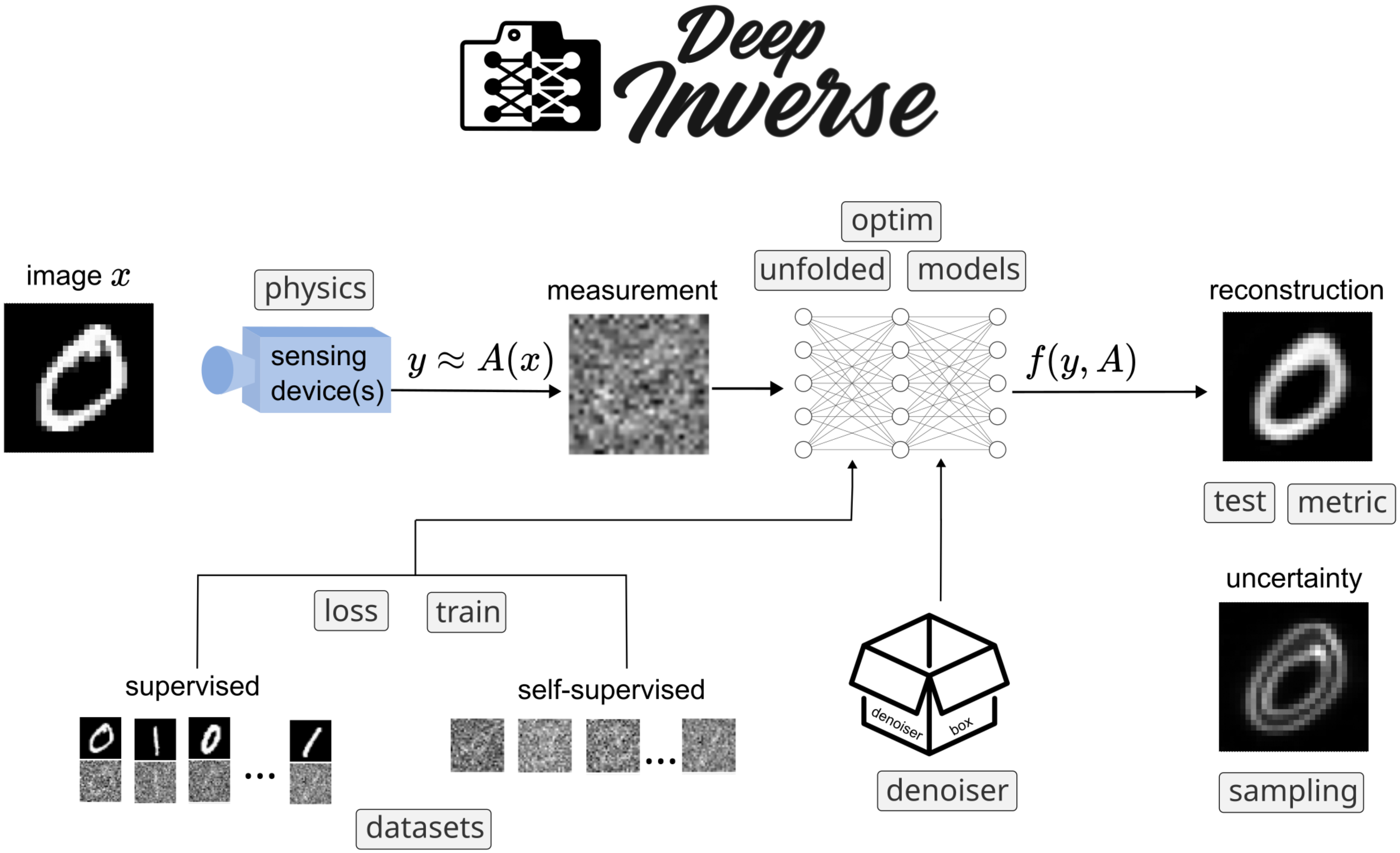


Experiments

- FastMRI dataset
- Gaussian noise

Method	CV + EI	EI (assumes $\sigma = 0$)	UNSURE + EI (unknown σ)	SURE + EI (assumes $\sigma = 0.05$)	Supervised
PSNR [dB]	33.25 ± 1.14	34.32 ± 0.91	35.73 ± 1.45	28.05 ± 4.73	36.63 ± 1.38





Deep Inverse
image x
physics
sensing device(s)
$y \approx A(x)$
measurement
optim
unfolded
models
$f(y, A)$
reconstruction
test
metric
loss
train
supervised
self-supervised
datasets
denoiser box
denoiser
uncertainty
sampling

Uncertainty Quantification

Can we measure the uncertainty of the reconstructions?

Self-supervised losses can also be used for uncertainty quantification!

- SURE can be used to assess reconstruction error in denoising
- SURE4SURE [Bellec et al., 2021] gives error variance estimates.
- EI loss can be seen as a bootstrapping technique

[T. & Pereyra, 2024] with well calibrated uncertainty estimates

References

Slides and codes of a recent 3-hour tutorial can be found here:

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