Self Supervised Learning Methods for Imaging

Part 2: Learning from noisy data

Mike Davies, University of Edinburgh

Julián Tachella, CNRS, École Normale Supérieure de Lyon

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

$$\boxed{\mathbb{E}_{\boldsymbol{y}}\, \mathcal{L}_{\text{self}}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathcal{L}_{\text{sup}}(\boldsymbol{x},\boldsymbol{y},f) + \text{const.}} \xrightarrow{\text{optimal solution}} \underset{\text{MMSE estimator}}{\boxed{f(\boldsymbol{y}) = \mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}\}}}$$

Noise2Noise

Mallows $\boldsymbol{C_p}$ [Mallows, 1973], **Noise2Noise** [Lehtinen, 2018]

- **Independent** pairs $\boldsymbol{y}_a = \boldsymbol{x} + \boldsymbol{\epsilon}_a$ and $\boldsymbol{y}_b = \boldsymbol{x} + \boldsymbol{\epsilon}_b$ with $\boldsymbol{\epsilon}_a, \boldsymbol{\epsilon}_b$ **independent**
- $\mathbb{E}_{\boldsymbol{\epsilon_b}|\boldsymbol{x}} \boldsymbol{\epsilon_b} = \boldsymbol{0}$

$$\mathbb{E}_{\boldsymbol{y}_b|\boldsymbol{x}}\, f(\boldsymbol{y}_a)^\top(\boldsymbol{y_b} - \boldsymbol{x}) = f(\boldsymbol{x} + \underbrace{\boldsymbol{\epsilon}_a)\, \mathbb{E}\, \boldsymbol{\epsilon}_b}_{=0} = 0$$

$$\boxed{\mathcal{L}_{N2N}(\boldsymbol{y}, f) = ||\, \boldsymbol{y}_b - f(\boldsymbol{y}_a)||^2}$$

- Also works for any noise distribution with $\mathbb{E}_{\boldsymbol{y_b}|\boldsymbol{x}}\, \boldsymbol{y_b} = \boldsymbol{x}$
- **Limitation:** observing independent copies is often impossible

Noisier2Noise

Recorrupted2Recorrupted [Pang et al., 2021], **Coupled Bootstrap** [Oliveira et al., 2022], **Noisier2Noise** [Moran et al., 2020].

Proposition: Let $\boldsymbol{y} \sim N(\boldsymbol{x}, I\sigma^2)$ and define

$$\boldsymbol{y}_a = \boldsymbol{y} + \alpha\boldsymbol{\omega}$$
$$\boldsymbol{y}_b = \boldsymbol{y} - \boldsymbol{\omega}/\alpha$$

where $\boldsymbol{\omega} \sim N(\boldsymbol{0}, I\sigma^2)$ and $\alpha \in \mathbb{R}$, then $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$ are **independent** random variables (fixed $\boldsymbol{x}$).

$$\boxed{\mathcal{L}_{R2R}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{\omega}}||\ \boldsymbol{y}_b - f(\boldsymbol{y}_a)\ ||^2}$$

- Price to pay: SNR($\boldsymbol{y}_a$) < SNR($\boldsymbol{y}$)
- Trick can be extended to Poisson noise [Oliveira et al., 2023]
- At **test time**, $f^{\text{test}}(\boldsymbol{y}) = \frac{1}{N}\sum_i f(\boldsymbol{y} + \alpha\boldsymbol{\omega}_i)$ with $\boldsymbol{\omega}_i \sim \mathcal{N}(\boldsymbol{0}, I\sigma^2)$
- **Extensions to the exponential family** [Monroy, Bacca & T., 2025]

Stein's Unbiased Risk Estimator

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

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

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

- **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

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

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

- Noisier2Noise is equivalent to SURE when $\alpha \to 0$ [Oliveira, 2022].

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

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

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

- **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]

Summary So Far

	Train Eval	Test Eval	Single $\boldsymbol{y}$	MMSE optimal	Unknown noise
Noise2Noise	1	1		✓	✓
Noisier2Noise	1	>1	✓	✓	
SURE	2	1	✓	✓	

If we have a single $\boldsymbol{y}$ and don't know the noise distribution?

Zero Expected Divergence

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

UNSURE [Tachella 2025]

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

- SURE's perspective: $\mathcal{L}_{\text{ZED}}(\boldsymbol{y}, f) = ||\, \boldsymbol{y} - f(\boldsymbol{y})||^2 + 2\sigma^2 \cancel{\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})}$

- Not MMSE optimal (but almost)

- Generalises to other partially known noise models

Cross-Validation Methods

Assumption: f_i does not depend on y_i, that is $\frac{\delta f_i}{\delta y_i} = 0$. Decomposable noise $p(\boldsymbol{y}|\boldsymbol{x}) = \prod p(y_i|x_i)$

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \sum_{i=1}^{n} f_i(\boldsymbol{y})(y_i - x_i) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}_{-i}|\boldsymbol{x}}\, f_i(\boldsymbol{y}_{-i})\, \underbrace{\mathbb{E}_{y_i|x_i}\,(y_i - x_i)}_{=0} = 0$$

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

- SURE's perspective: $\mathcal{L}_{SURE}(\boldsymbol{y}, f) = ||\,\boldsymbol{y} - f(\boldsymbol{y})||^2 + 2\sigma^2 \cancel{\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y})}$

- These methods are not MMSE optimal
- How to remove dependence on y_i: training or architecture

Measurement Splitting

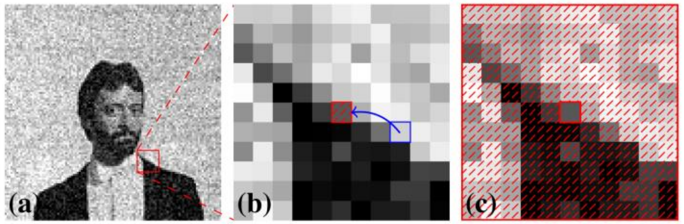
Cross-validation [Efron, 2004]: random split $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_a \\ \boldsymbol{y}_b \end{bmatrix}$ at each iteration

$$\boxed{\mathcal{L}_{N2V}(\boldsymbol{y}, f) = \mathbb{E}_{a,b} || \, \boldsymbol{y}_b - \operatorname{diag} \boldsymbol{m}_b \; f(\boldsymbol{y}_a) ||^2}$$

where $\boldsymbol{m}_b \in \{0,1\}^n$ masks out the pixels in $\boldsymbol{y}_a$.

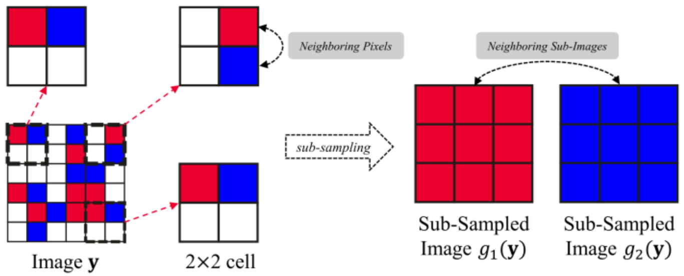
Noise2Void [Krull et al., 2019], **Noise2Self** [Batson, 2019]

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



Neighbor2Neighbor [Huang, 2023]

- Use different subsampling as input and target
- Assumes scale invariance



Measurement Splitting

At **test time**, $f(\boldsymbol{y})$ is evaluated as

1. Test f as trained (expensive)

$f^{\text{test}}(\boldsymbol{y}) = \frac{1}{N}\sum_i M\, f\left(\boldsymbol{y}_{\boldsymbol{a}_i}\right)$ with $\boldsymbol{y}_{\boldsymbol{a}_i} \sim p(\boldsymbol{y}_{\boldsymbol{a}}|\boldsymbol{y})$ and $M = \left(\sum_i^N \operatorname{diag}\left(\boldsymbol{m}_{b,i}\right)\right)^{-1}$

2. Assume good generalization of f (cheap)

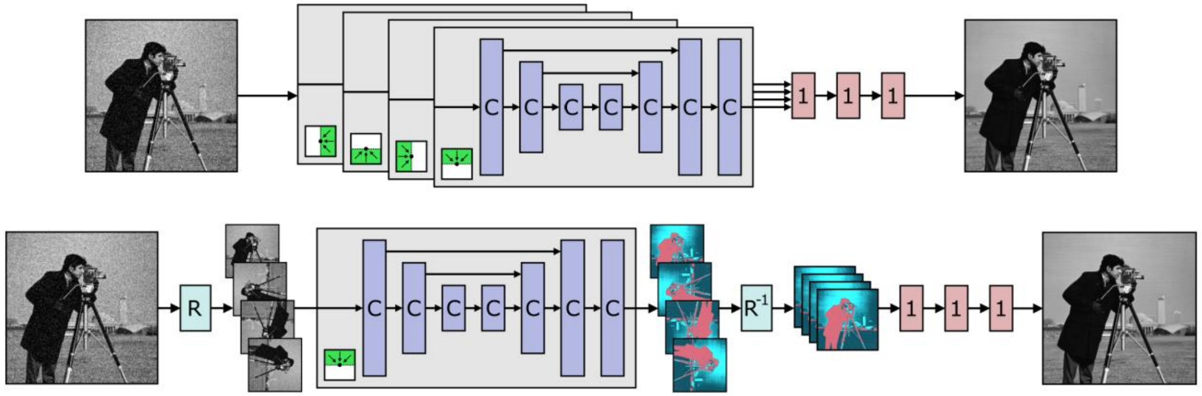
- $f^{\text{test}}(\boldsymbol{y}) = f(\boldsymbol{y}_{\boldsymbol{a}})$ with $\boldsymbol{y}_{\boldsymbol{a}} \sim p(\boldsymbol{y}_{\boldsymbol{a}}|\boldsymbol{y})$
- $f^{\text{test}}(\boldsymbol{y}) = f(\boldsymbol{y})$

Blind Spot Networks

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

- Convolutional architecture that doesn't 'see' centre pixel by construction

$$\boxed{\mathcal{L}_{\mathrm{BS}}(\boldsymbol{y}, f_{\mathrm{BS}}) = ||\,\boldsymbol{y} - f_{\mathrm{BS}}(\boldsymbol{y})||^2}$$



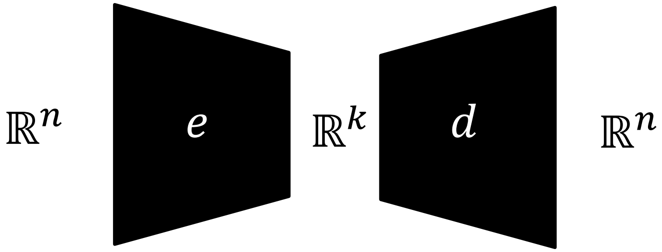
Autoencoders

Autoencoders

Assume

- f has a strong bottleneck

$f_{\mathrm{AE}} = d \cdot e$



$\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) \approx O(k) \ll n$ is small

$$\boxed{\mathcal{L}_{AE}(\boldsymbol{y}, f) = ||\, \boldsymbol{y} - f_{\mathrm{AE}}(\boldsymbol{y})||^2}$$

- Noise distribution is ‘high-dimensional’ whereas signal distribution is ‘low-dimensional’
- **Example:** linear ortho denoiser $f(\boldsymbol{y}) = M\boldsymbol{y}$, then $\sum_i \frac{\delta f_i}{\delta y_i}(\boldsymbol{y}) = \mathrm{tr}\, M = k$

Summary

	Train Eval	Test Eval	Single y	MMSE optimal	Unknown separable noise	Unknown coloured noise
Noise2Noise	1	1		✓	✓	✓
R2R	1	>1	✓	✓		
SURE	2	1	✓	✓		
UNSURE	2	1	✓	✓	✓	✓
Noise2Void	1	1	✓		✓	
Blind Spot	1	>1	✓		✓	
Autoencoders	1	1			✓	✓

No free lunch: less assumptions about noise = less optimal estimator

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

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

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

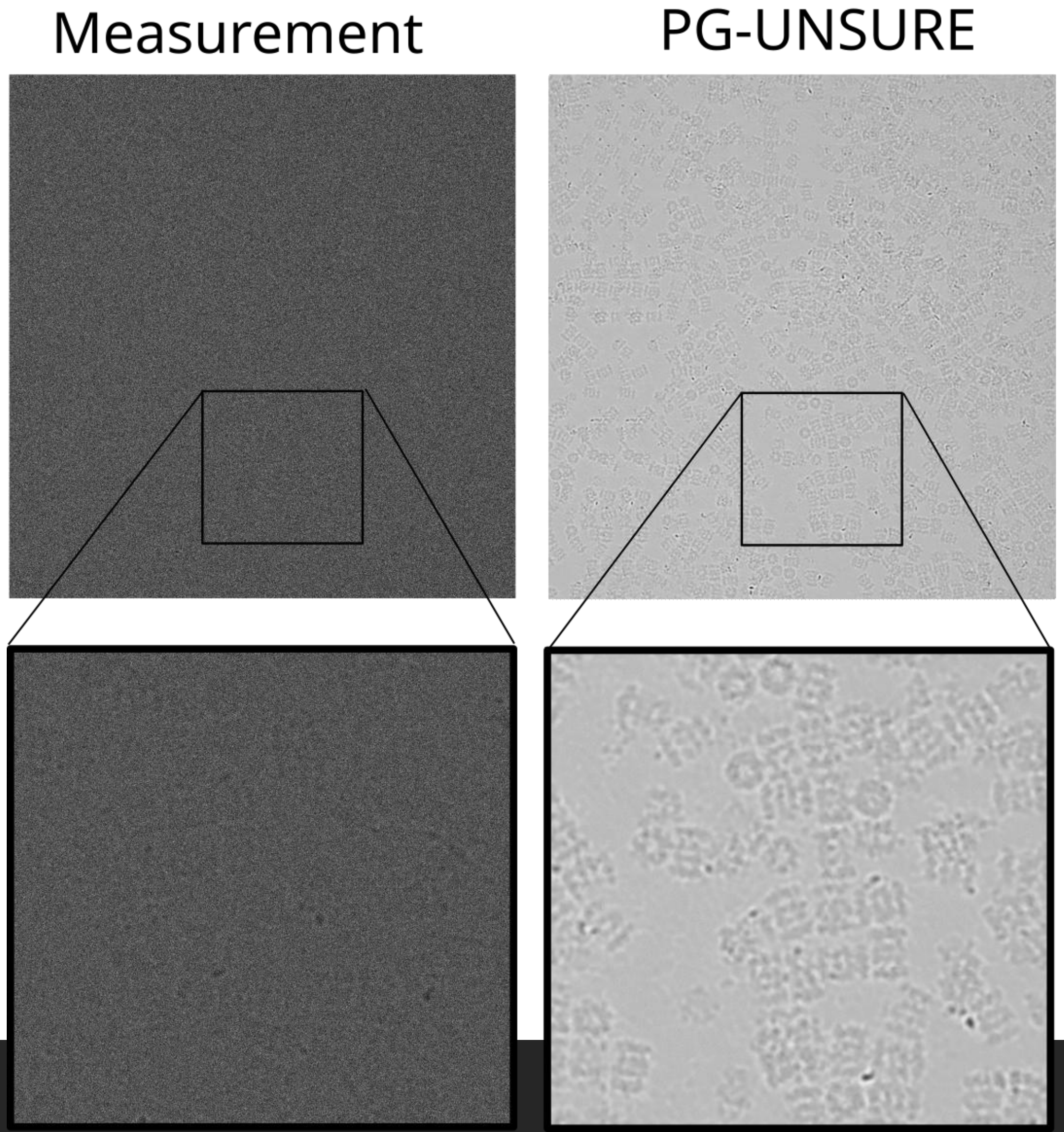
unknown p_i

Less expressive estimators

Experiments

Real data experiments

- Cryo electron microscopy images
- Extremely low SNR
- Approx. Poisson-Gaussian noise



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

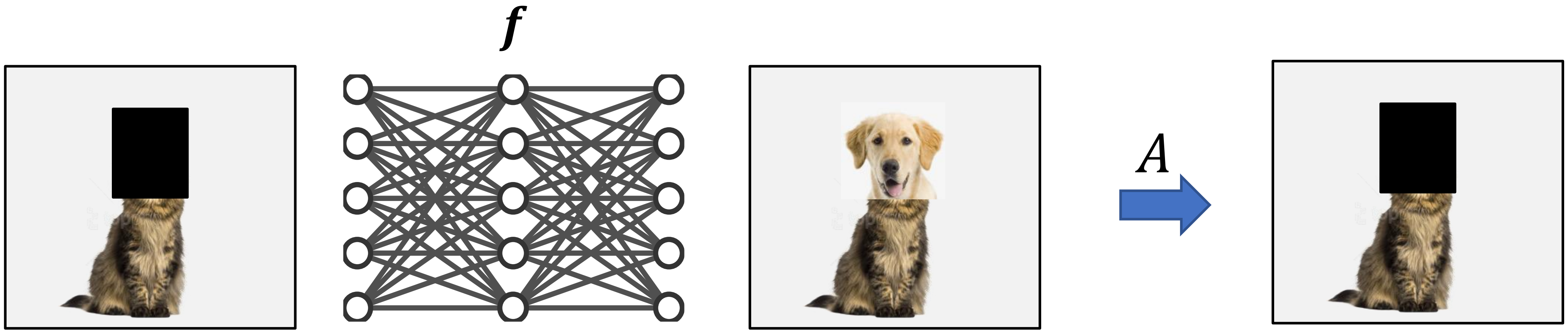
$$\boxed{\mathcal{L}_{\mathrm{GSURE}}(\boldsymbol{y}, f) = ||A^\dagger \boldsymbol{y} - A^\dagger A f(\boldsymbol{y})||^2 + 2\sigma^2 \sum_i \frac{\delta\left[A^\dagger A \cdot f\right]_i}{\delta y_i}(\boldsymbol{y})}$$

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 $\boldsymbol{A}$!



References

The full reference list for this tutorial can be found here:

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