Generalization in diffusion models arises from geometry-adaptive harmonic representations

By Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, Stéphane Mallat

Thodoris Kouzelis

NTUA & Archemides AI

CV & Robotics reading group
14 October 2024

Main Question addressed in the Paper

Q: High dimension density estimation is difficult! Yet DMs manage to generate HQ samples even when trained on limited data. How?

1) Memorizing the training set?

2) Interpolating between training images? (Generalize)

3) If they do are there any inductive biases that restrict the hypothesis space?

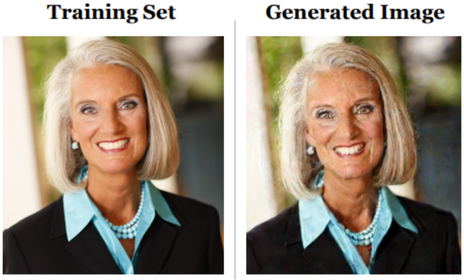
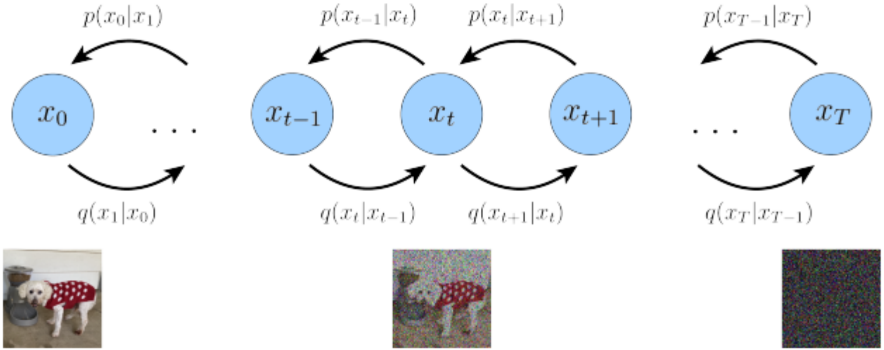


Figure: Figure from Carlini et al 2024

But how do Diffusion Models work in the first place ?



But how do Diffusion Models work in the first place ?

- A forward Gaussian process adds noise to the data
  - $q(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}) = \mathcal{N}(\boldsymbol{x_t}; \sqrt{1-\beta_t}\boldsymbol{x_{t-1}}, \beta_t \boldsymbol{I})$
  - $q(\boldsymbol{x_t}|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x_t}; \sqrt{\bar{a}}\boldsymbol{x}_0, (1-\bar{a})\boldsymbol{I})$, where $\bar{a}_t = \prod_{i=1}^{T} a_i$ and $a_t = 1 - \beta_t$

- A *learned* reverse Gaussian process p_θ generated data from noise
  - $p_\theta(\boldsymbol{x_T}) = \mathcal{N}(\boldsymbol{x_T}; 0, \boldsymbol{I})$
  - $p_\theta(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}) = \mathcal{N}(\boldsymbol{x_{t-1}}; \boldsymbol{\mu_\theta}(\boldsymbol{x_t}, \boldsymbol{t})), \boldsymbol{\sigma}(\boldsymbol{x_t}, \boldsymbol{t}))$

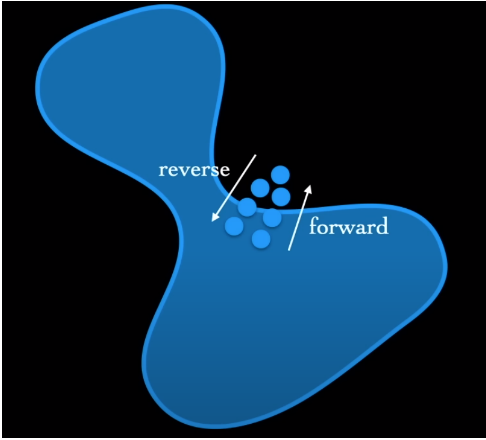
- The goal is to learn the reverse process i.e. $\boldsymbol{\mu_\theta}(\boldsymbol{x_t}, \boldsymbol{t}), \boldsymbol{\sigma}(\boldsymbol{x_t}, \boldsymbol{t})$ from data

Diffusion Models Intro

What objective will be optimized?

Maximize $p_\theta(\boldsymbol{x}_0)$?

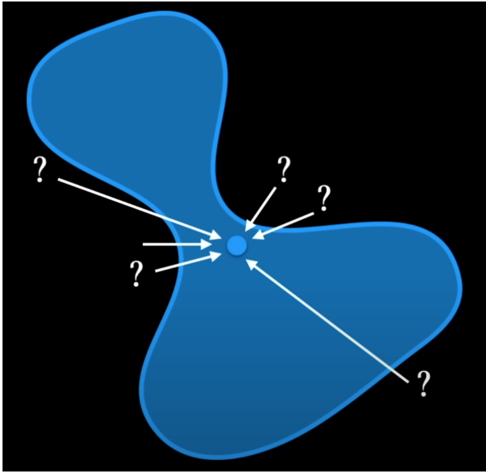
$$p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$$



What objective will be optimized?

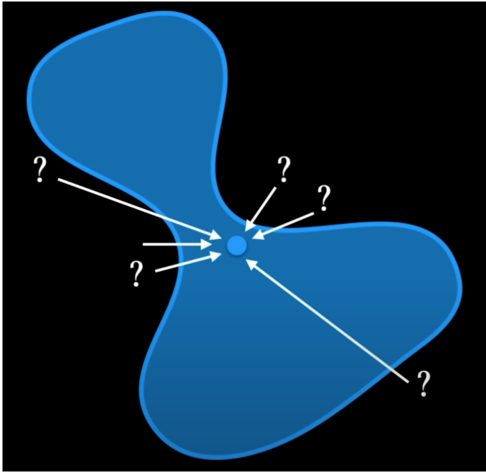
Marginalizing over all possible trajectories is intractable.

$$p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_{0:T})\underline{d\boldsymbol{x}_{1:T}}$$



What objective will be optimized?

- View $x_1, x_2, \ldots x_T$ as latent varables
- And x_0 as the observed variable
- Maximize an Evidence Lower Bound (ELBO)



Evidence Lower Bound :

$$\log p(x) \geq \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right]}_{\text{ELBO}}$$

$$\text{ELBO} = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{how good the "decoder" is}} \quad - \quad \underbrace{D_{KL}(q(z|x)\|p(z))}_{\text{how good the "encoder" is}}$$

- ELBO: $\log p_\theta(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$
- $\text{argmax}_\theta \left\{\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]\right\}$ as a proxy for maximizing $p(x_0)$

...

- $\text{argmin}_\theta \; D_{KL}(q(x_{t-1}|x_t, x_0) \mid\mid p_\theta(x_{t-1}|x_t))$

...

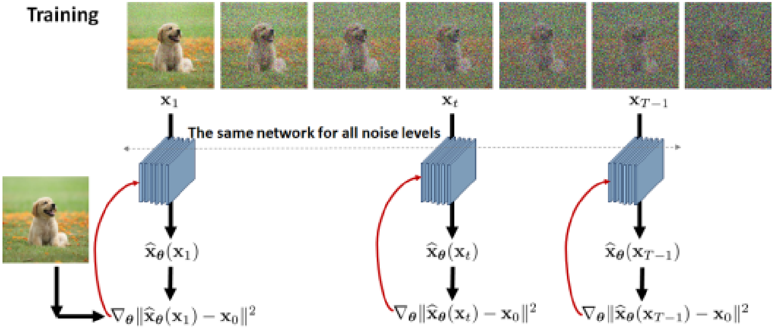
- $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \underbrace{\frac{\sqrt{a_t}(1-\bar{a}_{t-1})x_t + \sqrt{\bar{a}_{t-1}}(1-a_t)x_0}{1-\bar{a}_t}}_{\mu_q(x_t, x_0)}, \underbrace{\frac{(1-a_t)(1-\bar{a}_{t-1})}{(1-\bar{a}_t)}I}_{\sum_q(t)})$

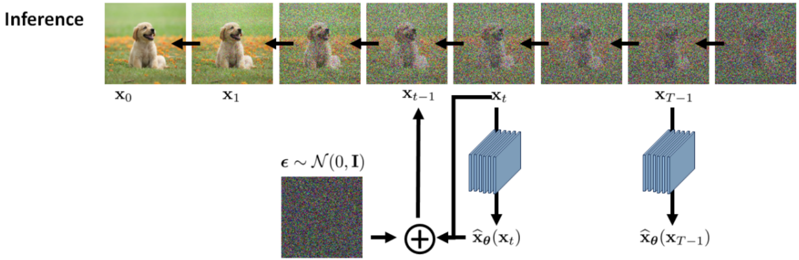
Diffusion Models Intro

- Now we can derive an objective for the mean of the reverse process:
  - $\text{argmin}_\theta \frac{1}{2\sigma_q^2(t)} \mathbb{E}_{q(x_t|x_0)} \left[||\mu_\theta(x_t) - \mu_q||_2^2 \right]$
- Which can be reformulated to:
  - $\text{argmin}_\theta \frac{1}{2\sigma_q^2(t)} \lambda_t \mathbb{E}_{q(x_t|x_0)} \left[||x_\theta(x_t) - x_0||_2^2 \right]$ (where λ_t is dependent on the noise schedule)
- x_θ is an MSE denoiser parameterized with a DNN.
- And is trained to perform denoising at all noise levels

Diffusion Models Intro

Training the denoiser at all noise levels





Sampling is performed according to: $x_{t-1} = \frac{(1-\bar{a}_{t-1})\sqrt{a_t}}{1-\bar{a}_t} x_t + \frac{(1-a_t)\sqrt{\bar{a}_{t-1}}}{1-\bar{a}_t} x_\theta(x_t) + \sigma_q(t)\epsilon$

Density estimation error

How good is the estimation of the density p_θ the denoiser learns (implicitly)

- $D_{KL}(p(x)||p_\theta(x)) \leq \int_0^\infty \Big(MSE(x_\theta, \sigma^2) - MSE(x^*, \sigma^2) \Big) \sigma^{-3} d\sigma$
  (x^* is the optimal denoiser)

- The density estimation error is bounded
  by the denoiser error (integrated over all noise levels)

- Thus, learning the true density model
  is equivalent to performing optimal denoising at all noise levels

Density estimation error

However the optimal denoiser x^* for photographic images is unknown.

- Separate two factors that contribute to sub-optimal performance:
  - Model Variance
    - Low variance implies the model generalizes well across different datasets
    - Can be evaluated without knowledge of x^*
  - Model Bias
    - Model bias measures the distance of the true denoiser to the approximated
    - Cannot be evaluated without knowledge of x^*

Transition from Memorization to Generalization

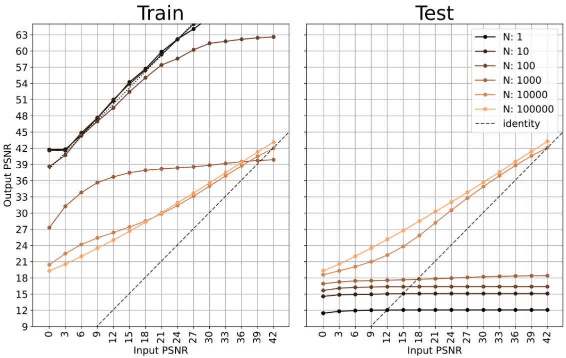


Figure: PSNR $= 10\log_{10}\frac{255^2}{MSE}$ [Kadkhodaie et al. 2024]

Transition from Memorization to Generalization

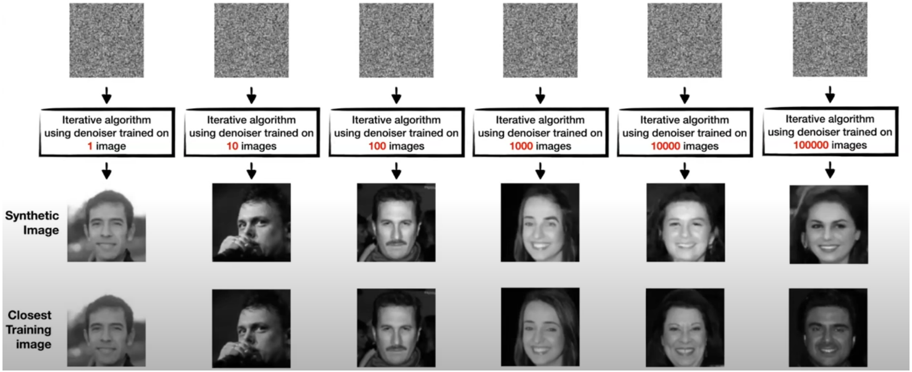


Figure: Kadkhodaie et al. 2024

Transition from Memorization to Generalization

Model Variance is tending to zero!

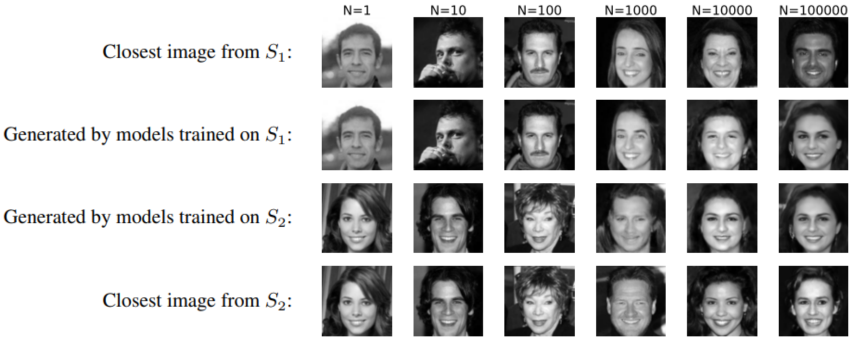


Figure: Kadkhodaie et al. 2024

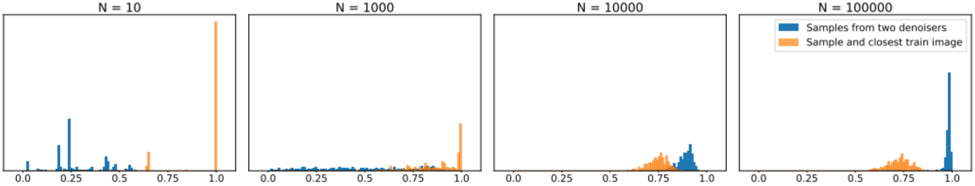


Figure: Kadkhodaie et al. 2024

Transition from Memorization to Generalization

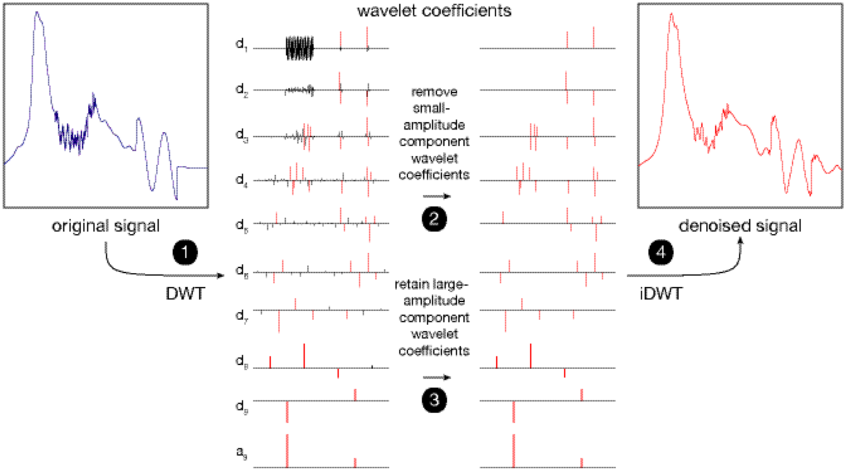
What are the inductive Biases of the denoiser

Denoising as shrinkage in a basis

Classical denoising framework:

- Transform the image to a **new basis** where noise and signal are seperable
- Suppress the noise (perform shrinkage)
- Transform back to pixel space

Denoising as shrinkage in a basis



Denoising as shrinkage in a basis

Wavelet based be - Fixed basis e_k, and **adaptive** shrinkage λ_k:

- $f(y) = \sum_k \lambda_k < y, e_k > e_k$
- Sparse representation in the Wavelet basis
- Adaptive thresholding:

$$\lambda_k = \begin{cases} 1 & \text{if } | < y, e_k > | > \alpha\sigma \\ 0 & \text{otherwise} \end{cases}$$

Denoising as shrinkage in adaptive basis

What if we could also make the basis adaptive?

Denoising as shrinkage in adaptive basis

A deep denoiser without bias terms written as:

- $f(y) = W_L R(W_{L-1} \ldots R(W_1 y)) = A_y y$

- $f(y) = J_y y$, where J_y is the jacobian of the denoiser w.r.t y

- Is nearly symmetric $\rightarrow$ Eigendecomposition!

Denoising as shrinkage in adaptive basis

The eigendecomposition of the Jacobian:

- $J_y = V \Lambda V^T = \sum_i \lambda_i v_i v_i^T$, where λ_i, v_i are the eigenvalues and eigenvectors
- Thus $f(y) = \sum_i \lambda_i < x, v_i > v_i$
- Both shrinkage factors and basis are adaptive!

Denoising as shrinkage in adaptive basis

Observations:

- Small eigenvalues $\lambda_k(y)$ reveal local invariances of the denoising function (effective null space) i.e. $f(y+v) \approx f(y)$
- Such invariances are a desirable property for a denoiser

Denoising as shrinkage in adaptive basis

On the low-rankness of the Jacobian:

- The optimal denoiser is the conditional mean of the posterior
  $f^* = \mathbb{E}_x[x|y]$

- The jacobian of the optimal denoiser is proportional to the posterior covariance matrix
  $J_y^* = \nabla f^*(y) = \sigma^{-2}\mathrm{Cov}[x|y]$

- Thee optimal denoising error is then given by:
  $\mathrm{MSE}(f^*, \sigma^2) = \mathbb{E}_y[\mathrm{tr}(\mathrm{Cov}[x|y])] = \sigma^2\mathbb{E}_y[\mathrm{tr}(\nabla f^*(y)] = \sigma^2\mathbb{E}_y\left[\sum_k \lambda_k^*(y)\right]$

- A small denoising error thus implies an approximately low-rank Jacobian

Geometry Adaptive Harmonic Basis (GAHB)

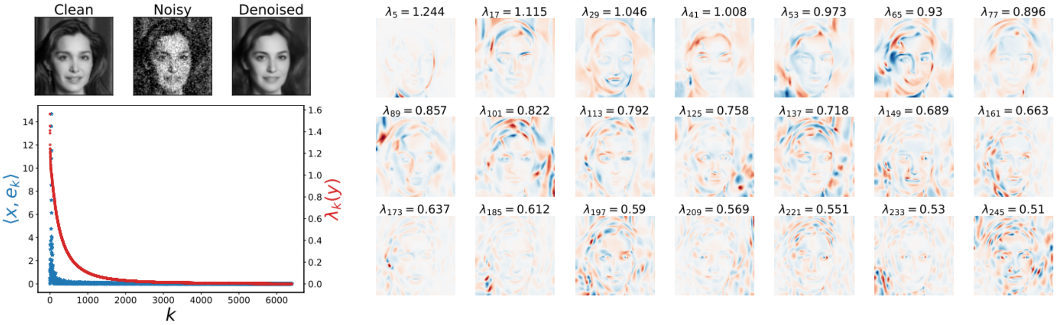
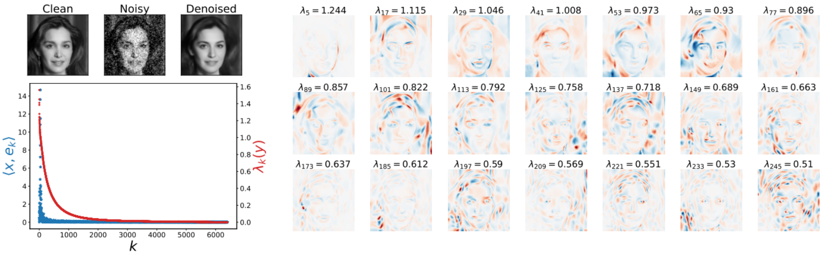


Figure: DNN trained on 10^5 images from CelebA, [Kadkhodaie et al. 2024]

ARCHIMEDES

Characteristics of GAHB

- Oscillating patterns both along the contours and in uniformly regular regions
- Thus adapt to the geometry of the input image
- The coefficients are sparse in this basis,
  and the fast rate of decay of eigenvalues exploits this sparsity



Denoising as shrinkage in adaptive basis

Conjecture: DNN denoisers have inductive biases towards learning GAHBs

Aligned inductive bias and optimality

Test on C_a images [Peyre & Mallat, 2008]:

- Regular contours on regular backgrounds
- As α is increases images become more regulars (smooth & without high-frequency variations)
- Known optimal denoiser

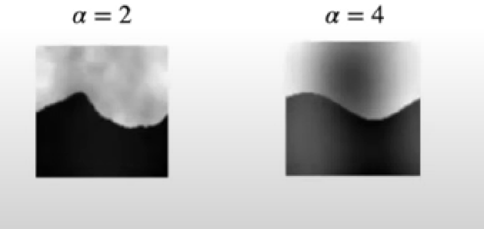


Figure: Kadkhodaie et al. 2024

Aligned inductive bias and optimality

DNN Denoiser trained on 10^5 C_a images

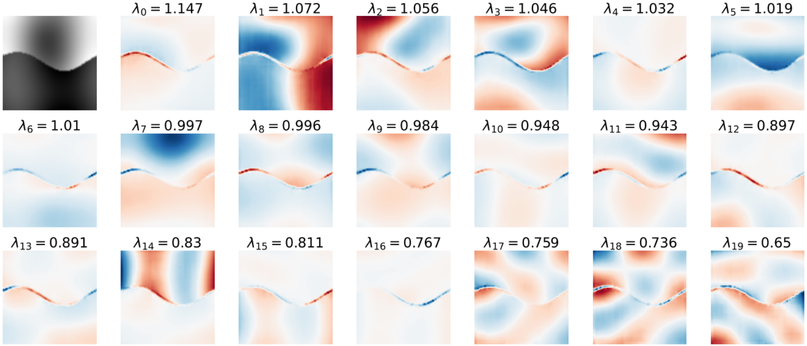


Figure: Kadkhodaie et al. 2024

Aligned inductive bias and optimality

- Optimal denoiser has slope $\frac{a}{a+1}$ Korostelev & Tsybacov, 1993
- The optimal slope is obtained by denoising with"bandlet" basis
- Larger $\alpha \to$ more regular $\to$ $\to$ sparser representation $\to$ larger slope

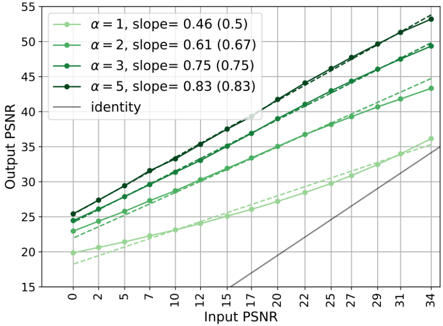


Figure: PSNR curves for various regularity levels [Kadkhodaie et al. 2024]

(Mis)-Aligned inductive bias

- If DNN denoisers are inductively biased towards GAHBs then we expect these bases to emerge even in cases where they are suboptimal
- Dataset of disk images with varying positions, sizes, and foreground/background intensities.
- 5-dimentional manifold (5 degrees of freedom)
- Known optimal basis

Figure: Examples from disk dataset [Kadkhodaie et al. 2024]

(Mis)-Aligned inductive bias

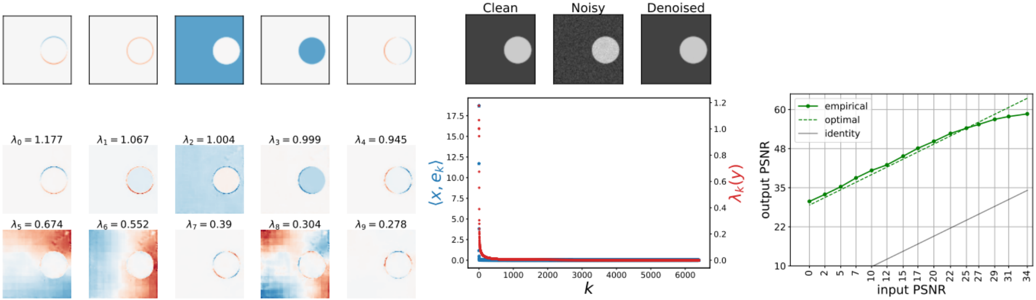


Figure: Kadkhodaie et al. 2024

The End

Thank You !!!