Multiscale Geometry of Images and Physics with Score Diffusion

N. Cuvelle-Magar, F. Guth, Z. Kadhokaie,
E. Lempereur, G. Biroli, M. Ozawa,
E. Simoncelli, S. Mallat

Collège de France
École Normale Supérieure

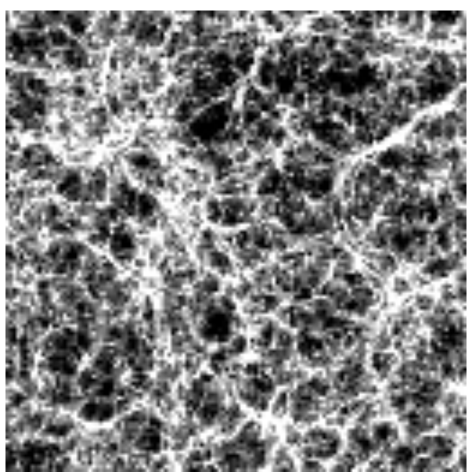
Learning Physics and Image Geometry

- Learning systems at equilibrium: estimate the probability $p(x)$

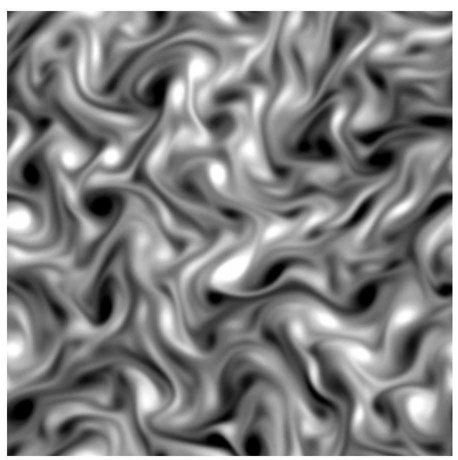
$$p(x) = \mathcal{Z}^{-1}\, e^{-U(x)} \quad \text{for } x \in \mathbb{R}^d$$

Curse of dimensionality if $d \gg 1$.

Statistical physics



Cosmic web



Turbulences

long-range geometry

since 1940's

Image generation by score denoising

Does it memorise or generalise ?

How does it circumvent the curse ?

Overview

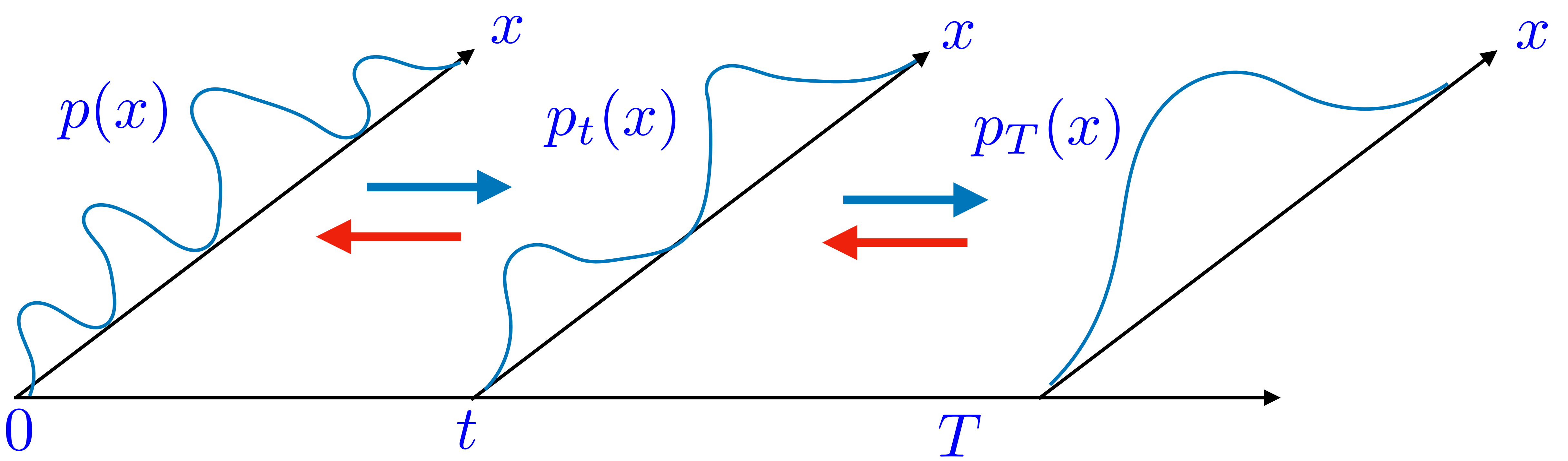
- 1. Generation by denoising score matching with deep networks . Generalisation or memorisation ? What prior ?

- 2. Renormalisation group with long-range geometric interactions for turbulences.

Transport of Probabilities

- Define a transport from p to a simple p_T

  Learn the inverse transport from data



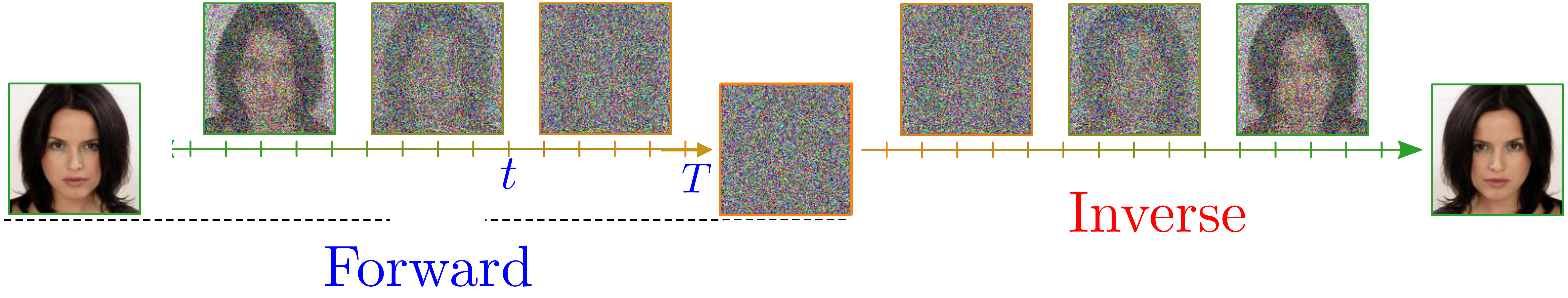
- Transport is learned from data: what type ?
  - Markov chains (1906): too general in high dimension
  - Physics Wilson renormalisation group (1970): along scales
  - AI score diffusion generation (2020): along noise variance

Score Diffusion Generation

Yang Song et. al.

- Forward diffusion: add noise with Ornstein-Uhlenbeck equation

$$dx_t = -x_t\,dt + \sqrt{2}\,dB_t$$



- The diffusion is inverted with a damped-Langevin equation:

$$dx_{T-t} = \big(x_{T-t} + 2\nabla \log p_{T-t}(x_{T-t})\big)\,dt + \sqrt{2}dB_t$$

- The score $\nabla \log p_t$ is estimated with a deep neural network.

Score Based Denoising

Noisy signal: $x_t = x + z$ with $z \sim \mathcal{N}(0, \sigma_t \, Id)$

The estimator $\hat{x}$ of x given x_t which minimises

$$\mathbb{E}_{z,x}\left(\|\hat{x} - x\|^2\right)$$

is the conditional expectation : $\hat{x} = \mathbb{E}[x|x_t]$

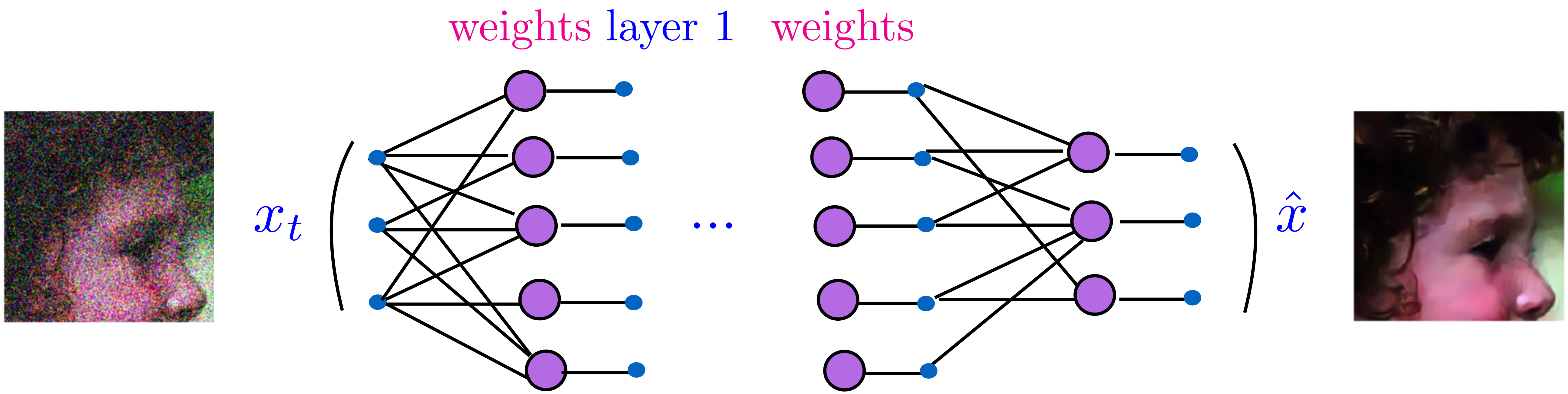
Score denoising: Tweetie, Robbins, Misayawa identity

$$\mathbb{E}[x|x_t] = x_t + \sigma_t^2 \, \nabla_{x_t} \log p_t(x_t)$$

The score $\nabla \log p_t$ is estimated with a denoising neural network

which computes $\hat{x}$ by minimising $\mathbb{E}(\|\hat{x} - x\|^2)$.

Score Estimation by Denoising



Trained by minimising $\mathbb{E}_{x_t}(\|\hat{x} - x\|^2)$ on the training set

$$\nabla \log p_t(x_t) \approx \hat{s}_t(x_t) = \frac{\hat{x} - x_t}{\sigma_t^2}$$

Can it estimate the score in high dimension ? Why ?

Image Generation by Score Diffusion

from large databases with N examples of images with score based diffusions.

Does it learn an underlying probability distribution ?

Estimation Error

Z. Kadkhodaie, F. Guth,, E. Simoncelli, S. M.

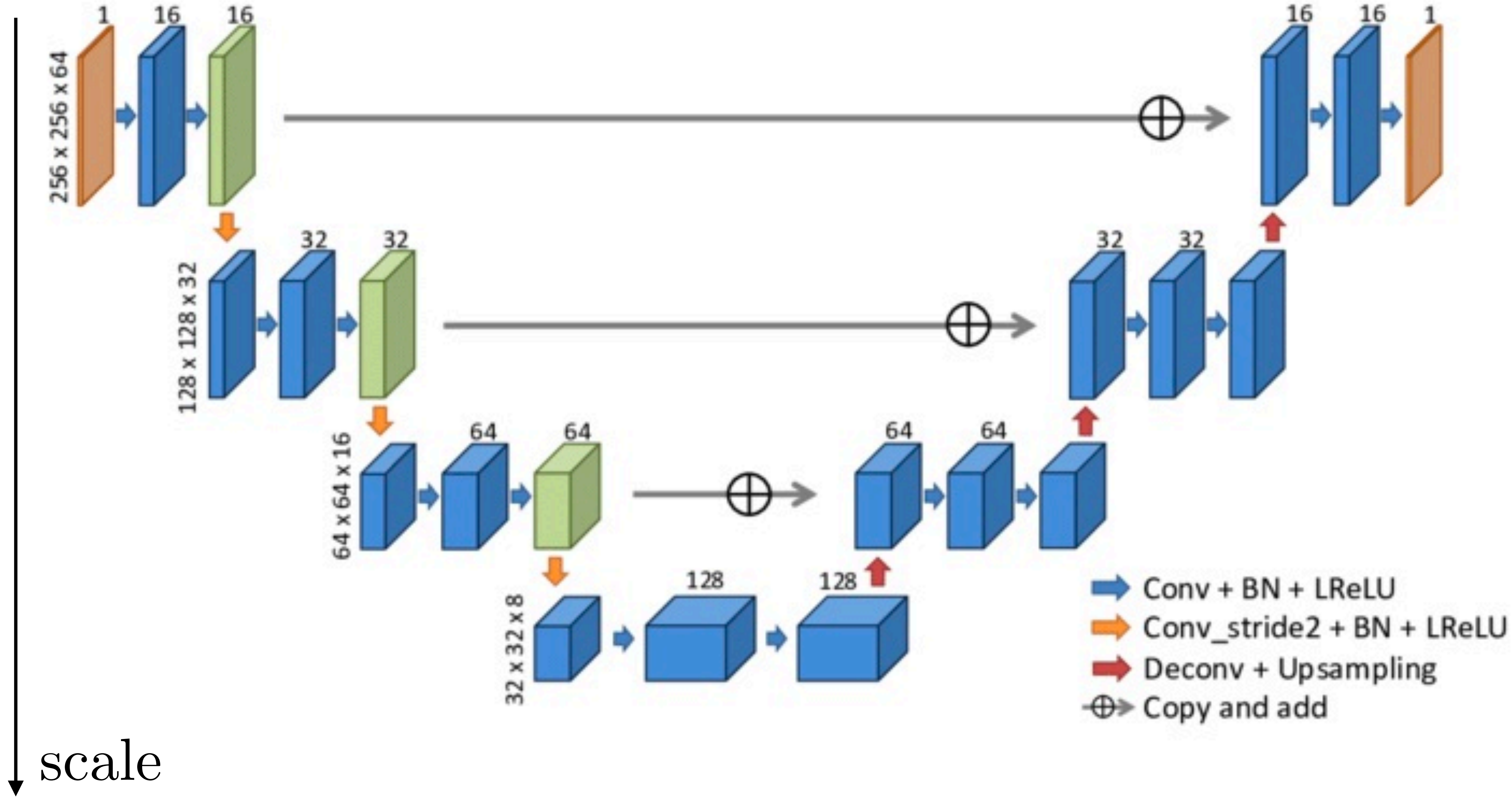
- **Variance:** does the estimation vary with the choice of training sample ? Does it memorise or generalise from the training ?

- **Bias:** does the model converge to the "true" underlying probability distribution ?

Convolutional U-Nets

7 million parameters for small 80×80 images

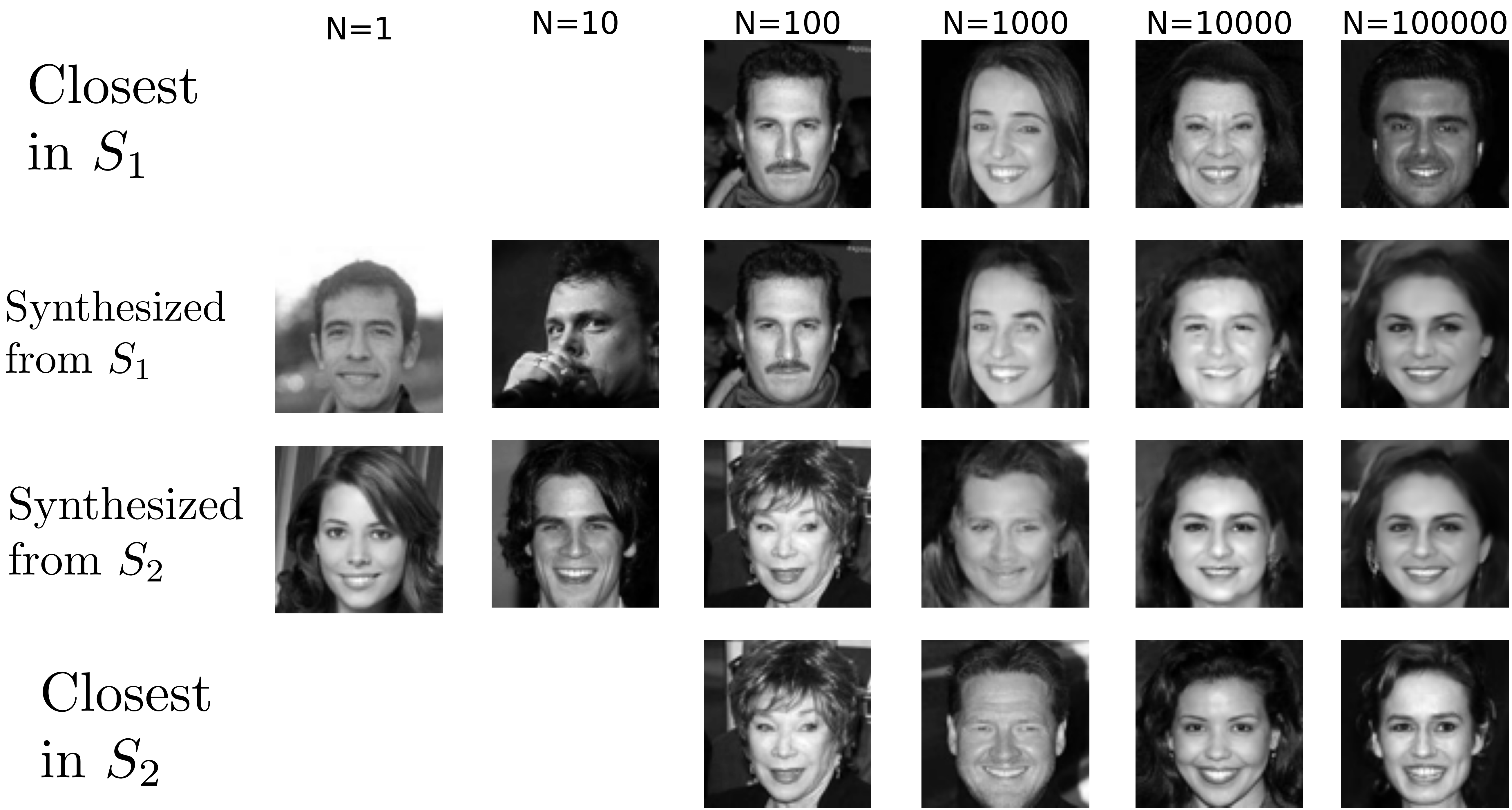
Linear convolutions and $ReLU(v) = \max(v, 0)$



Generalises or Memorises ?

Images reconstructed from the same noise with 2 scores estimated from 2 different train sets S_1 and S_2 of N images of 80×80 pixels

Generalises!



Generalisation Test

Z. Kadkhodaie, F. Guth, S.M., E. Simoncelli

Images reconstructed from the same noise with 2 scores estimated from 2 different train sets S_1 and S_2 of N images of 80×80 pixels

N = 100,000

Synthesized from S_1

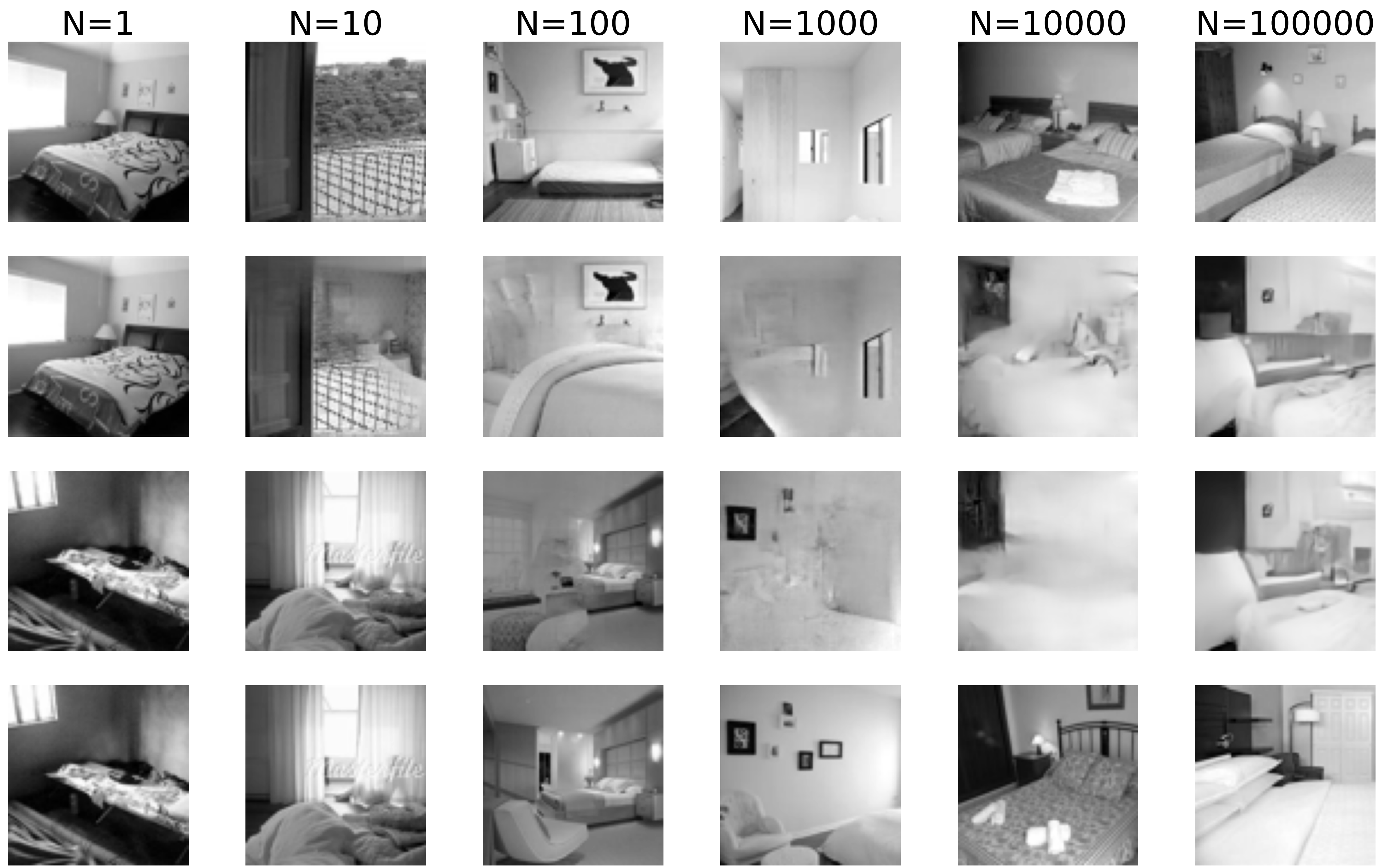
Synthesized from S_2

The estimation variance is small for N large enough

Generalisation Test: Memorise ?

Images reconstructed from the same noise with 2 scores estimated from 2 different train sets S_1 and S_2 of N images of 80×80 pixels

Generalises!



Closest in S_1

Synthesized from S_1

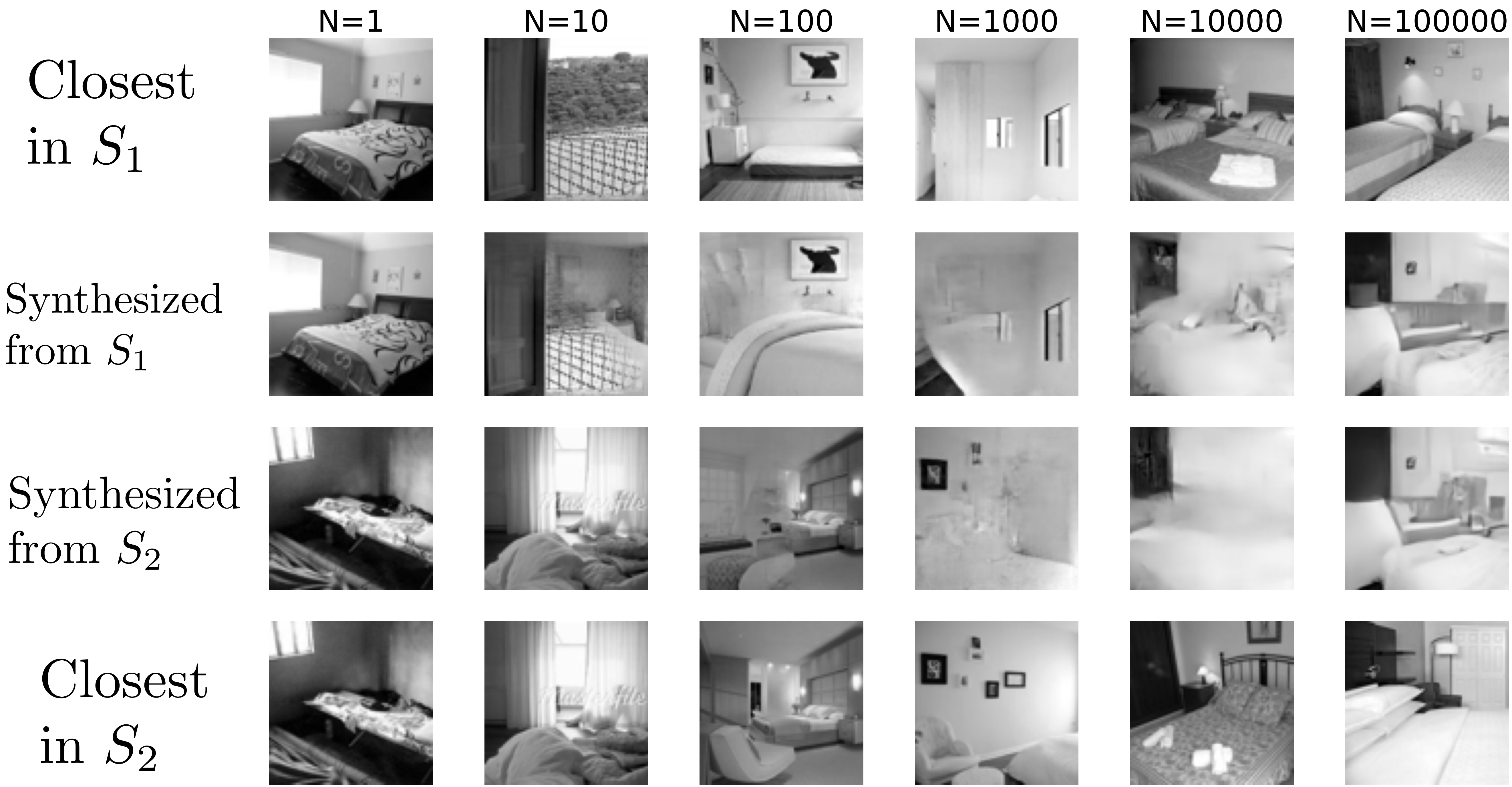
Synthesized from S_2

Closest in S_2

Generalisation Test: Memorise ?

The number N for generalisation depends on the number of parameters of the network.

Generalises!



Analysis of Bias

What is the bias when estimating p ?

$\Leftrightarrow$ bias when estimating $\nabla \log p_t$?

$\Leftrightarrow$ optimality of denosing estimator $\hat{x}$?

Sparse Denoising in Adapted Basis

A CNN provides a score estimation

$$\hat{x} = x_t + \hat{s}_t(x_t) \quad \text{with } \hat{s}_t(x) = -\nabla \hat{U}_t(x)$$

It is locally linear $\Rightarrow \quad \hat{s}_t(x) = \nabla \hat{s}_t(x)\, x.$

$$\hat{x} = (Id - \nabla^2 \hat{U}_t)\, x_t$$

In the basis $\{\psi_k\}_k$ which diagonalises the energy hessian $\nabla^2 \hat{U}_t$:

$$\hat{x} = \sum_k \lambda_k \langle x_t, \psi_k \rangle \psi_k$$

signal + noise

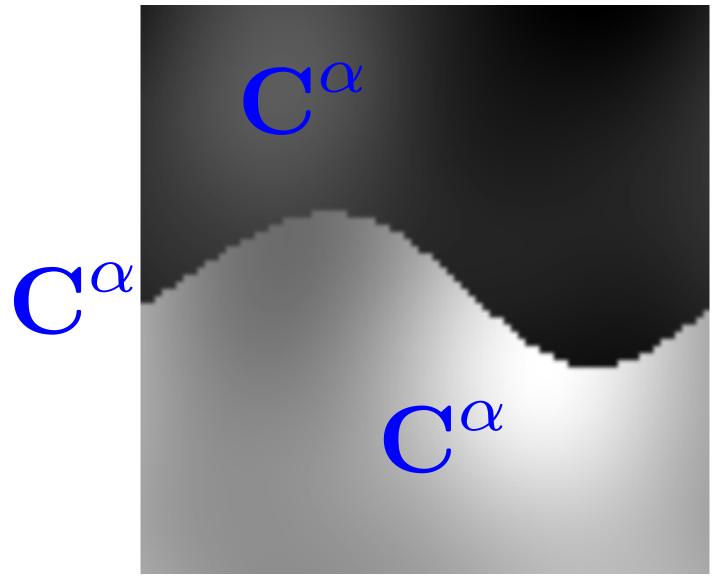
with $\langle x_t, \psi_k \rangle = \langle x, \psi_k \rangle + \langle z_t, \psi_k \rangle$

Shrinks $\langle x_t, \psi_k \rangle$ in an orthonormal basis $\{\psi_k\}_k$ adapted to x_t

Minimise error $\Leftrightarrow \{\langle x, \psi_k \rangle\}_k$ is a sparse representation of x

Optimal Denoising of Geometry ?

random $\mathbf{C}^\alpha$ curve in a random $\mathbf{C}^\alpha$ background



Optimal estimator: $\mathbb{E}(\|\hat{x} - x\|^2) \sim \sigma^{2\alpha/(\alpha+1)}$

C. Dossal, E. LePennec, G. Peyre, S. M(2005).

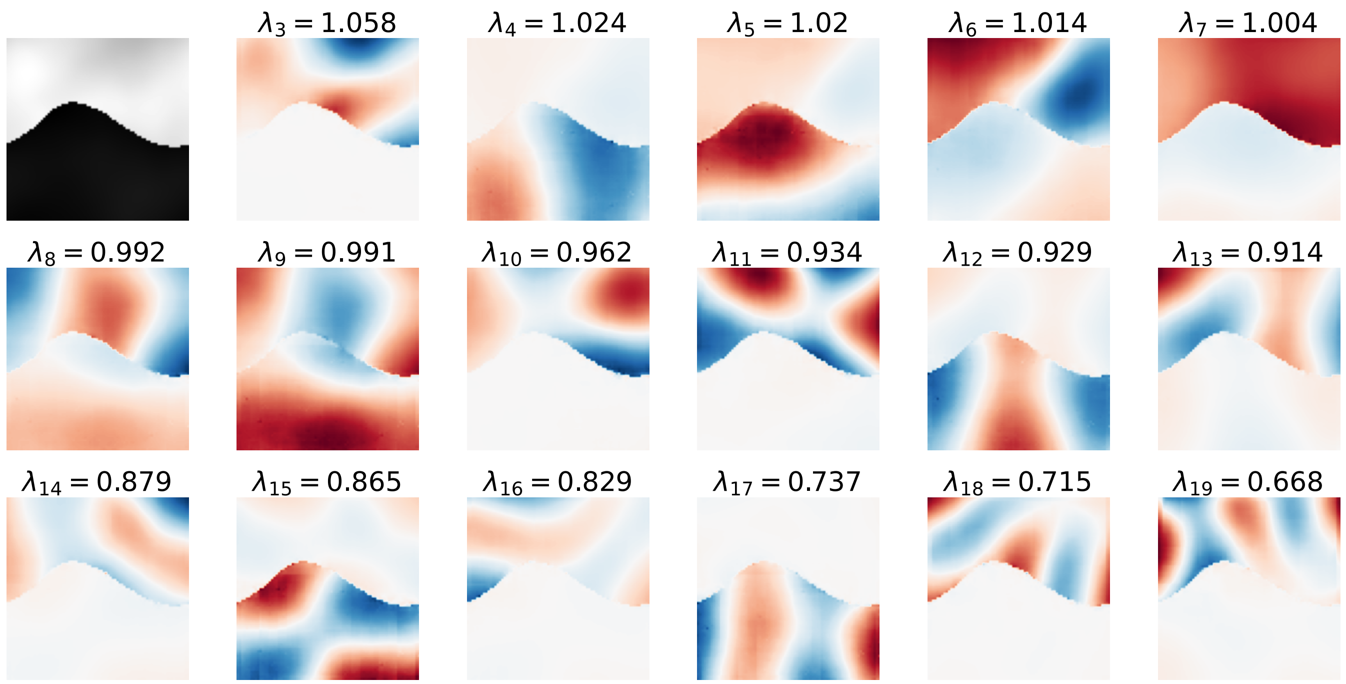
by shrinking coefficients in geometric harmonic bases

adapted to the estimated geometry from x_t

Is a CNN able to reach this optimal denoising rate ?

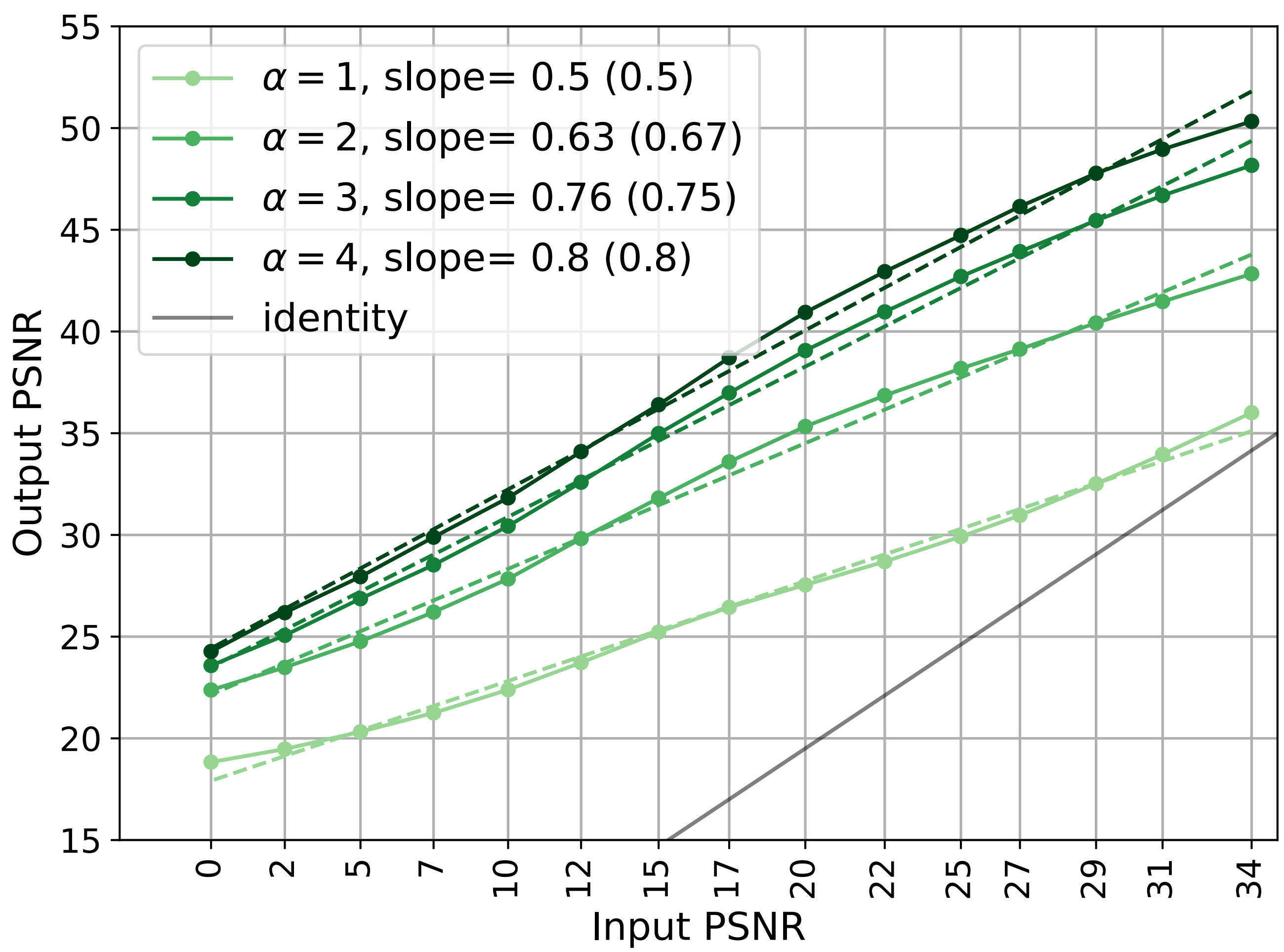
Optimal Denoising in GAHB

Z. Kadkhodaie, F. Guth, S.M., E. Simoncelli



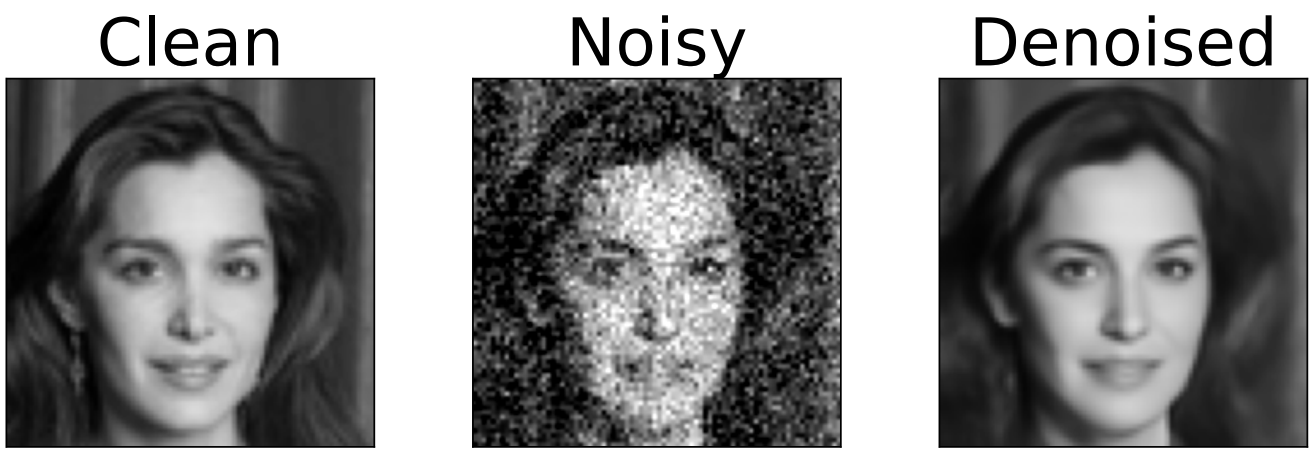
Eigenvectors of Hessian

Geometrically Adapted Harmonic Bases

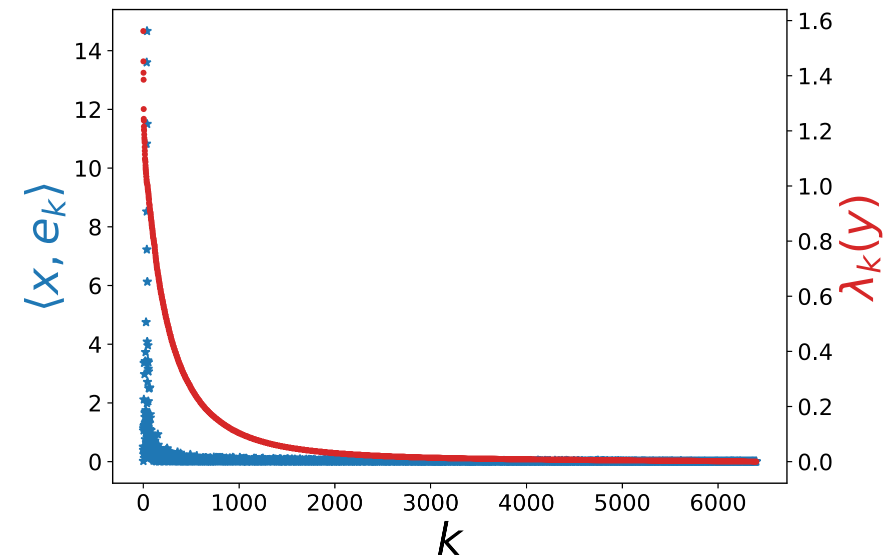


Optimal denoising rate

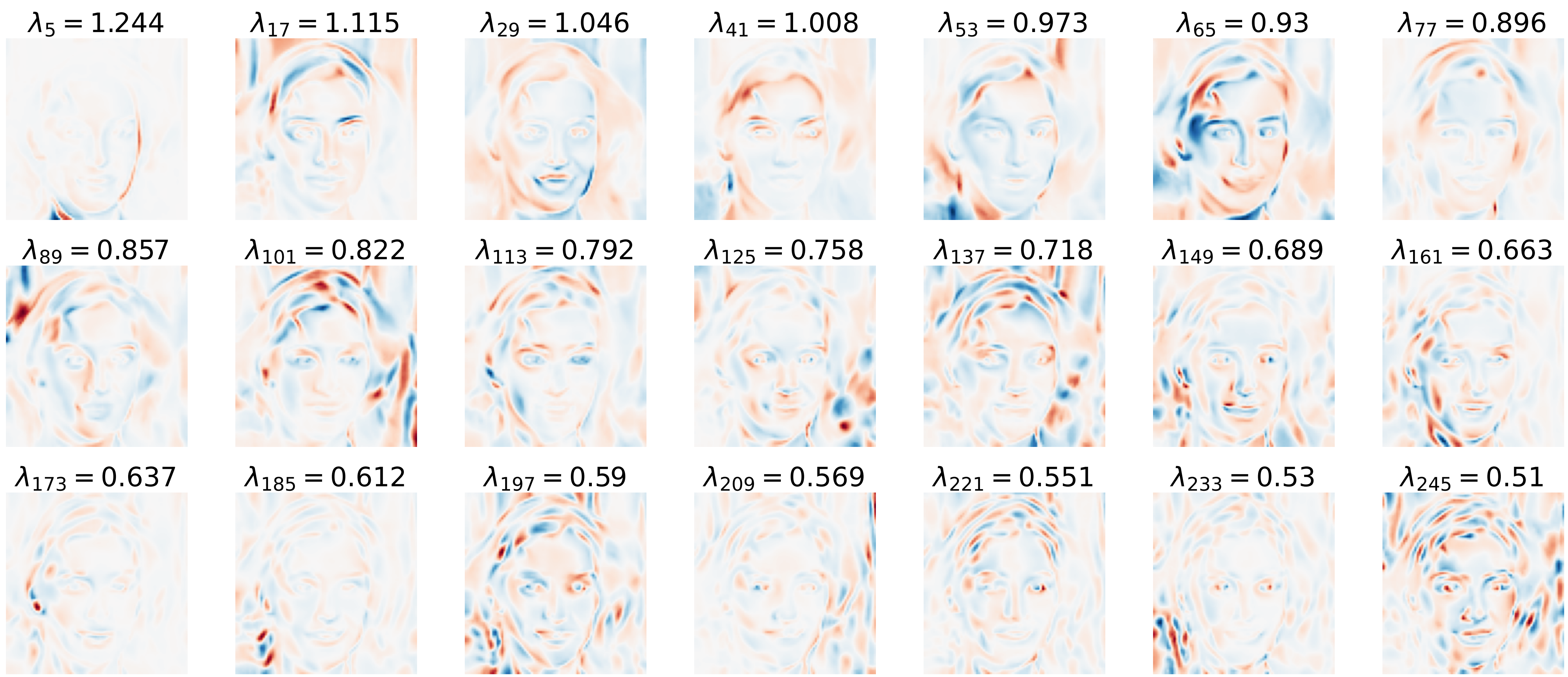
Geometrically Adapted Basis



Eigenvalues of the Hessian



Eigenvectors of the Hessian

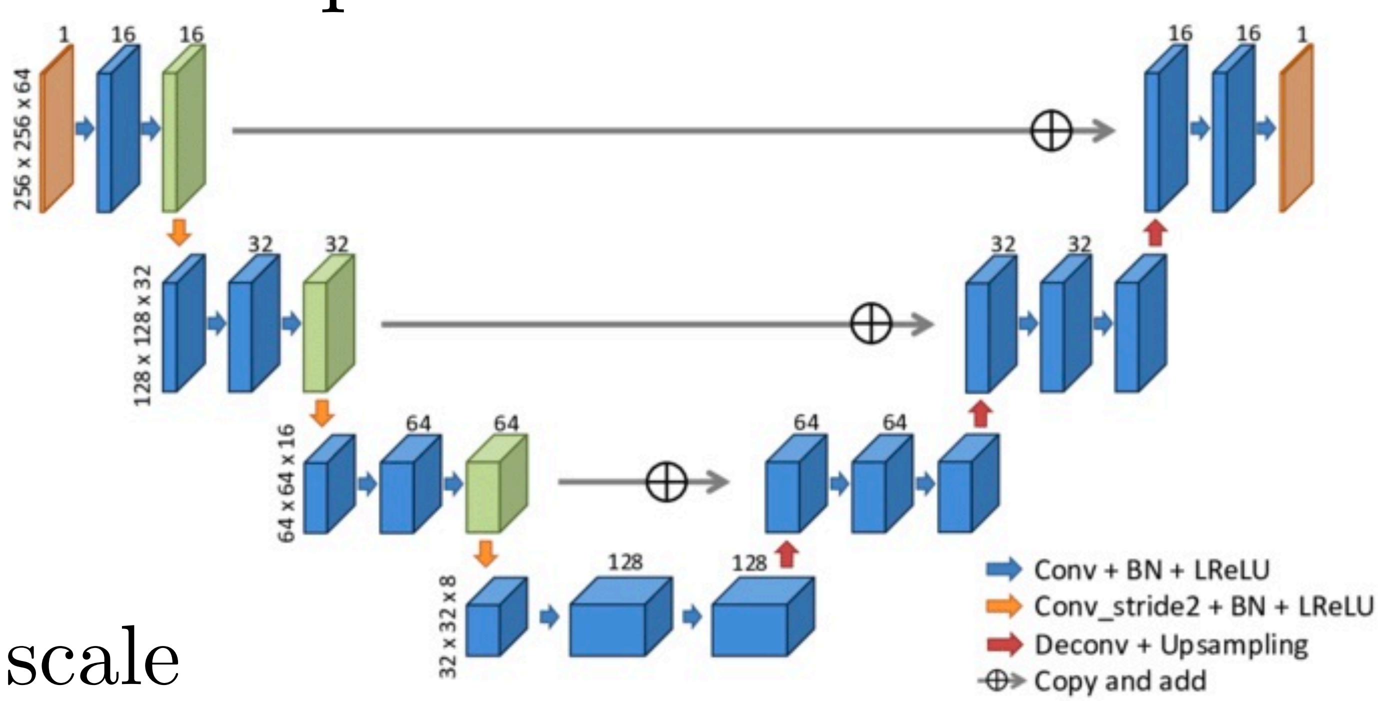


2. High Dimensional Models

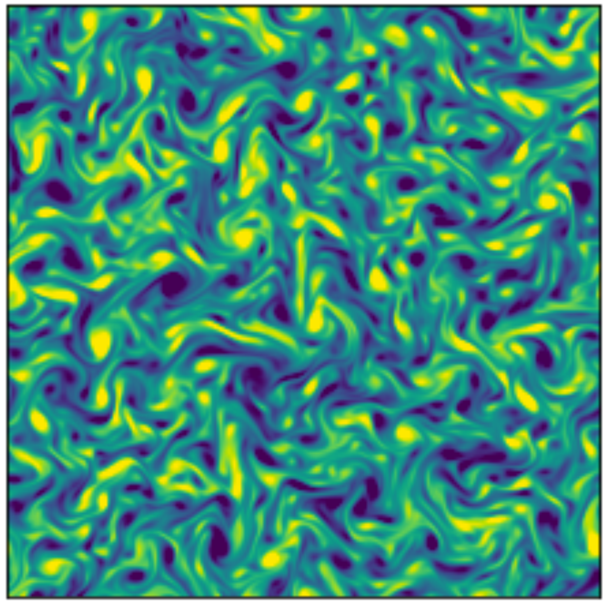
- Score diffusion generalises with enough training examples
- Generalisation depends upon the number of network parameters
- Circumvents the curse of dimensionality: how ?
  Symmetries over geometric groups are not enough.

Can we build accurate models with fewer examples ?

Open the black box



scale

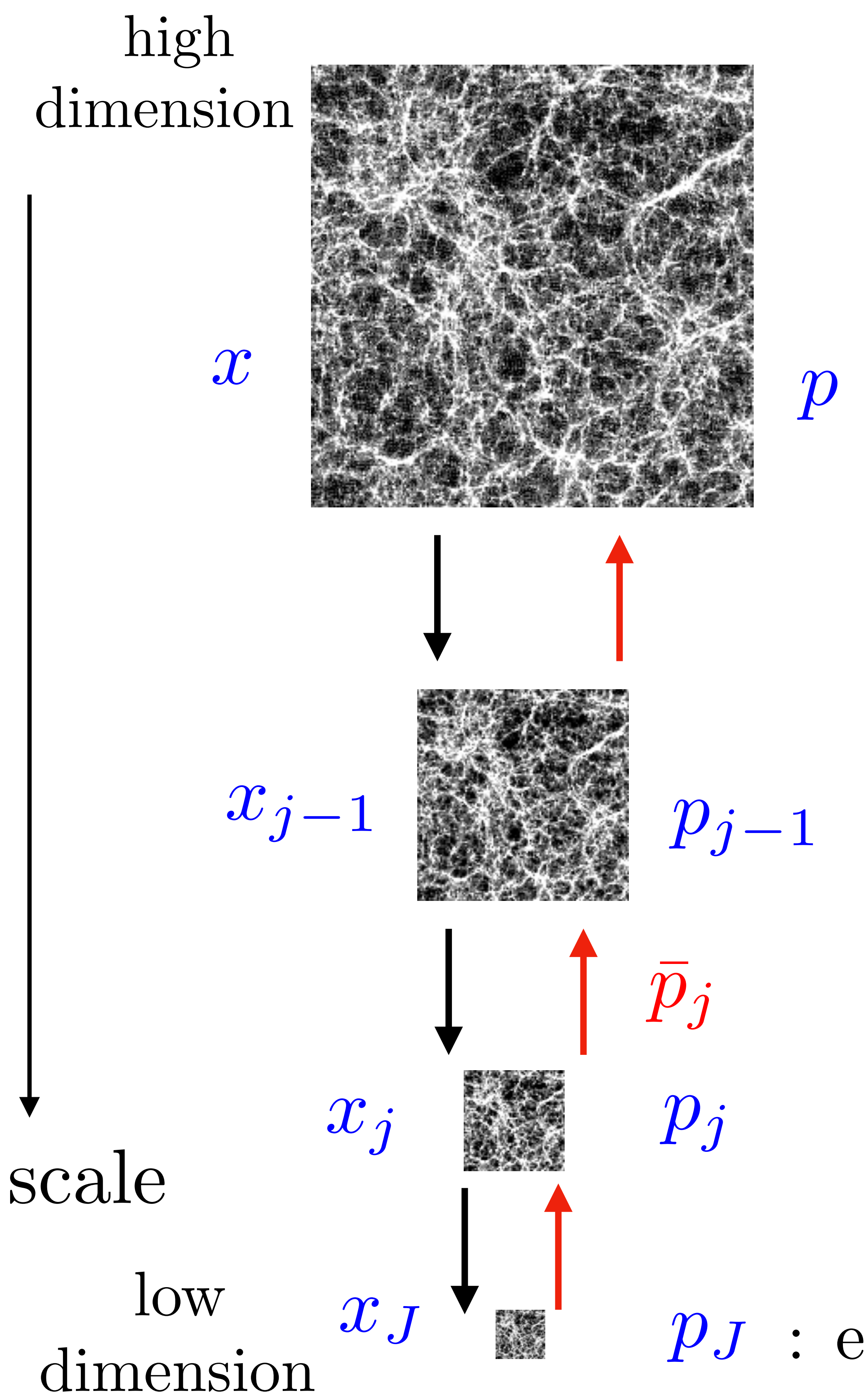


How to capture an image geometry ?

Can we model physical turbulences ?

Renormalisation Group : Hierachy

Kadanoff, Wilson 1970



Probability transport across scales

Inverse Markov chain

$$p_{j-1}(x_{j-1}) = p_j(x_j)\,\bar{p}_j(x_{j-1}|x_j)$$

G. Biroli, E. Lempereur
T. Marchand, M. Ozawa, S. M.

Need to estimate each $\bar{p}_j(x_{j-1}|x_j)$

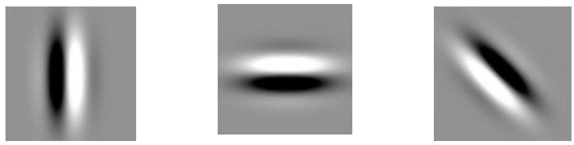
having long range interactions:

p_J : easy to estimate and sample

Transition Probabilities Across Scales

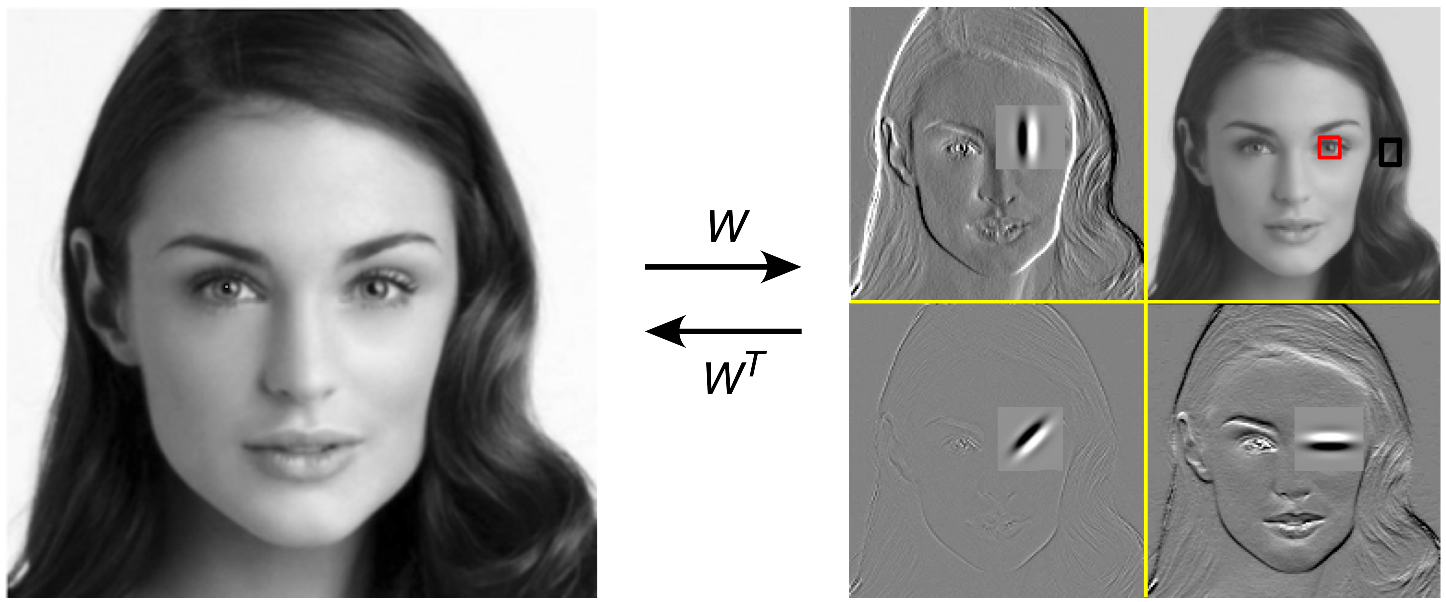
Wavelet orthogonal basis : $x_{j-1} \leftrightarrow (x_j, \bar{x}_j)$

ψ^1 ψ^2 ψ^3



$$\bar{x}_j = \{x * \psi_j^k\}_{k \leq 3} \quad \text{with } \psi_j^k(u) = 2^{-j}\psi^k(2^{-j}u)$$

x_{j-1} $\bar{x}_j$ x_j



$$\bar{p}_j(x_{j-1}|x_j) = \bar{p}_j(\bar{x}_j|x_j)$$

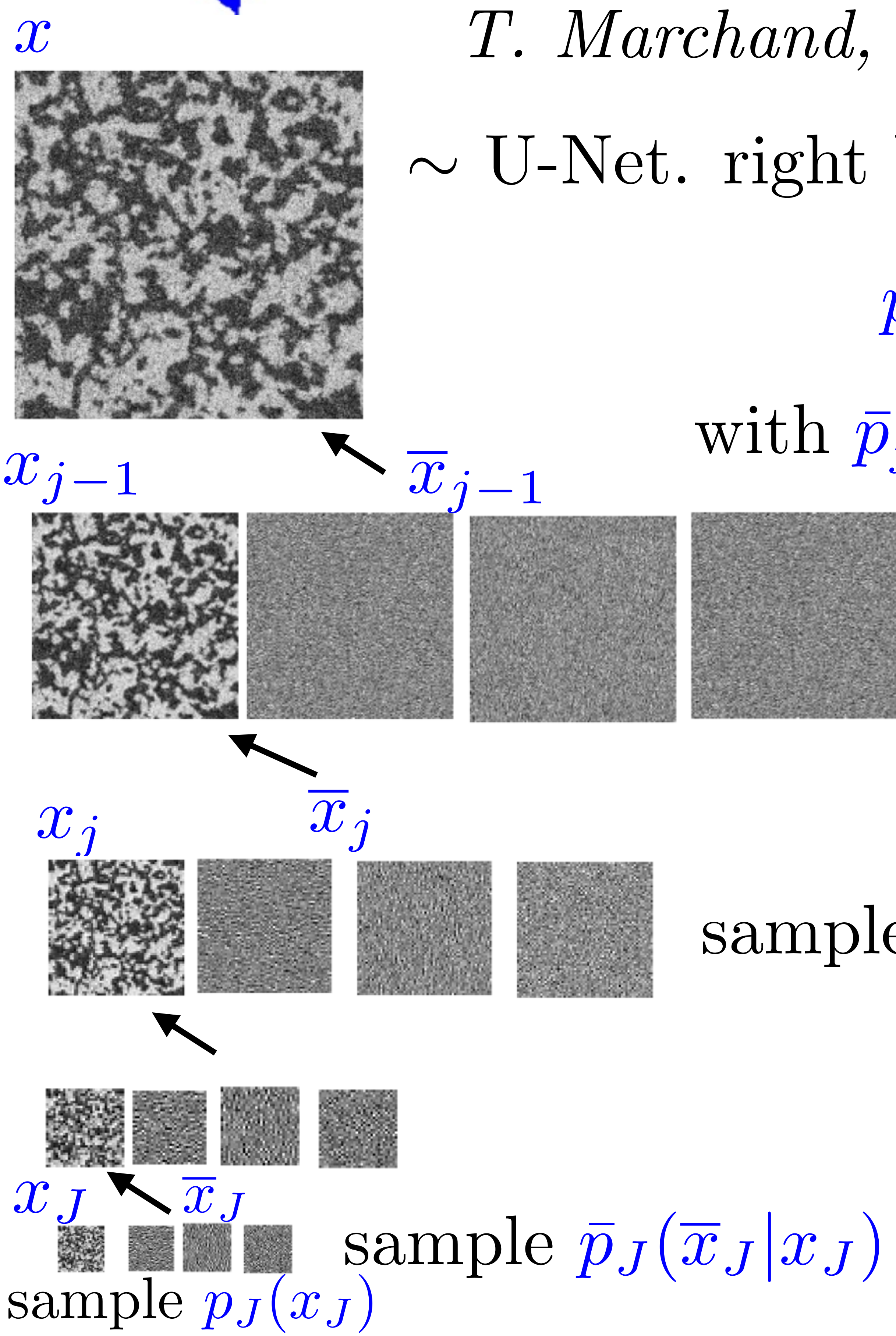
Can we build low-dimensional exponential model ?

$$\bar{p}_j(\bar{x}_j|\bar{x}_{j+1}\ldots) = \mathcal{Z}_j^{-1}\, e^{-\theta_j{}^T \Phi(\bar{x}_j, \bar{x}_{j+1}\ldots)}$$

Multiscale representation of geometry

Hierarchic Sampling

T. Marchand, M. Ozawa, G. Biroli, S. M.



$\sim$ U-Net. right branch

$$p(x) = p(x_J) \prod_{j=1}^{J} \bar{p}_j(\overline{x}_j | x_j)$$

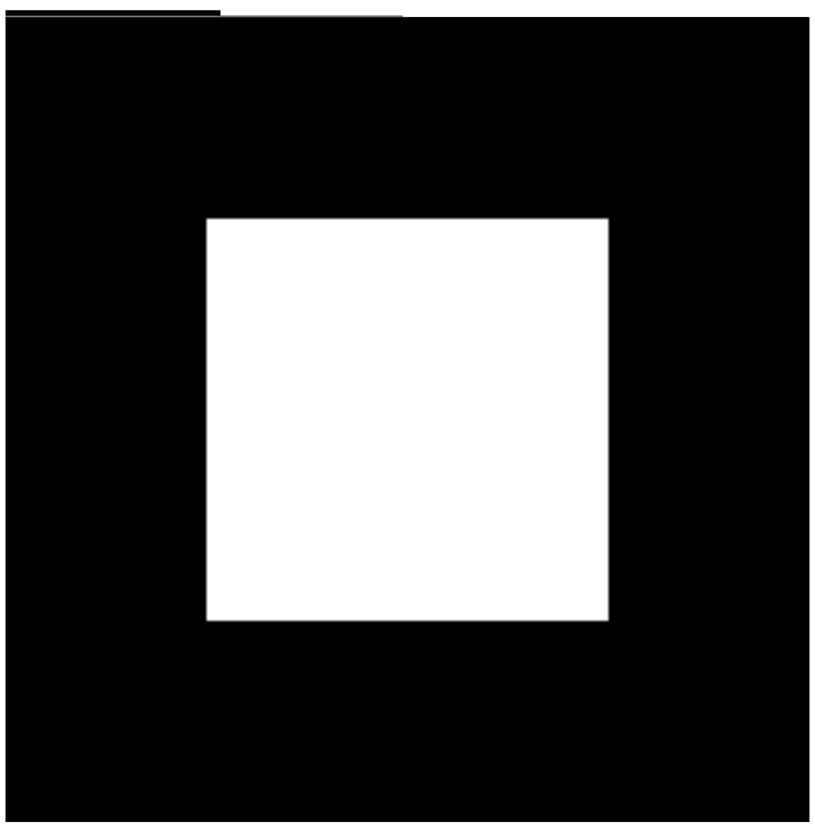
with $\bar{p}_j(\bar{x}_j | x_j) = \mathcal{Z}_j^{-1} \, e^{-\theta_j{}^T \Phi(\bar{x}_j, \bar{x}_{j+1} \ldots)}$

sample $\bar{p}_j(\overline{x}_j | x_j)$

sample $\bar{p}_J(\overline{x}_J | x_J)$

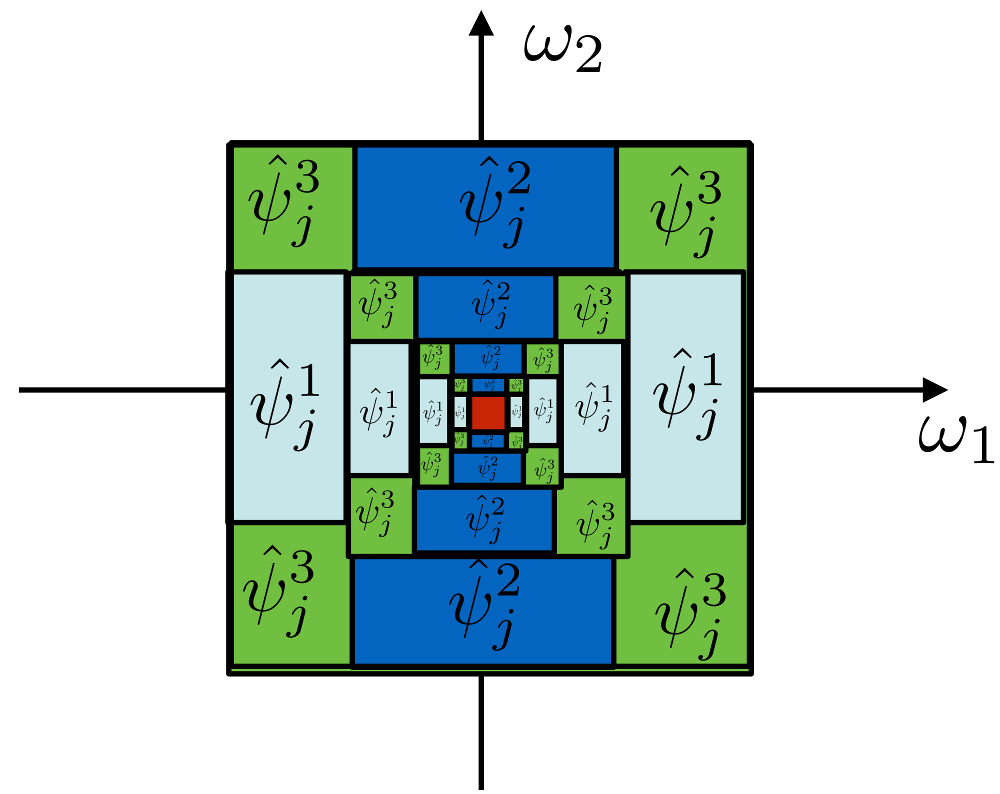
Wavelet Subdivision in Fourier

Orthogonal wavelets decomposes in different frequency bands



Wavelet coefficients

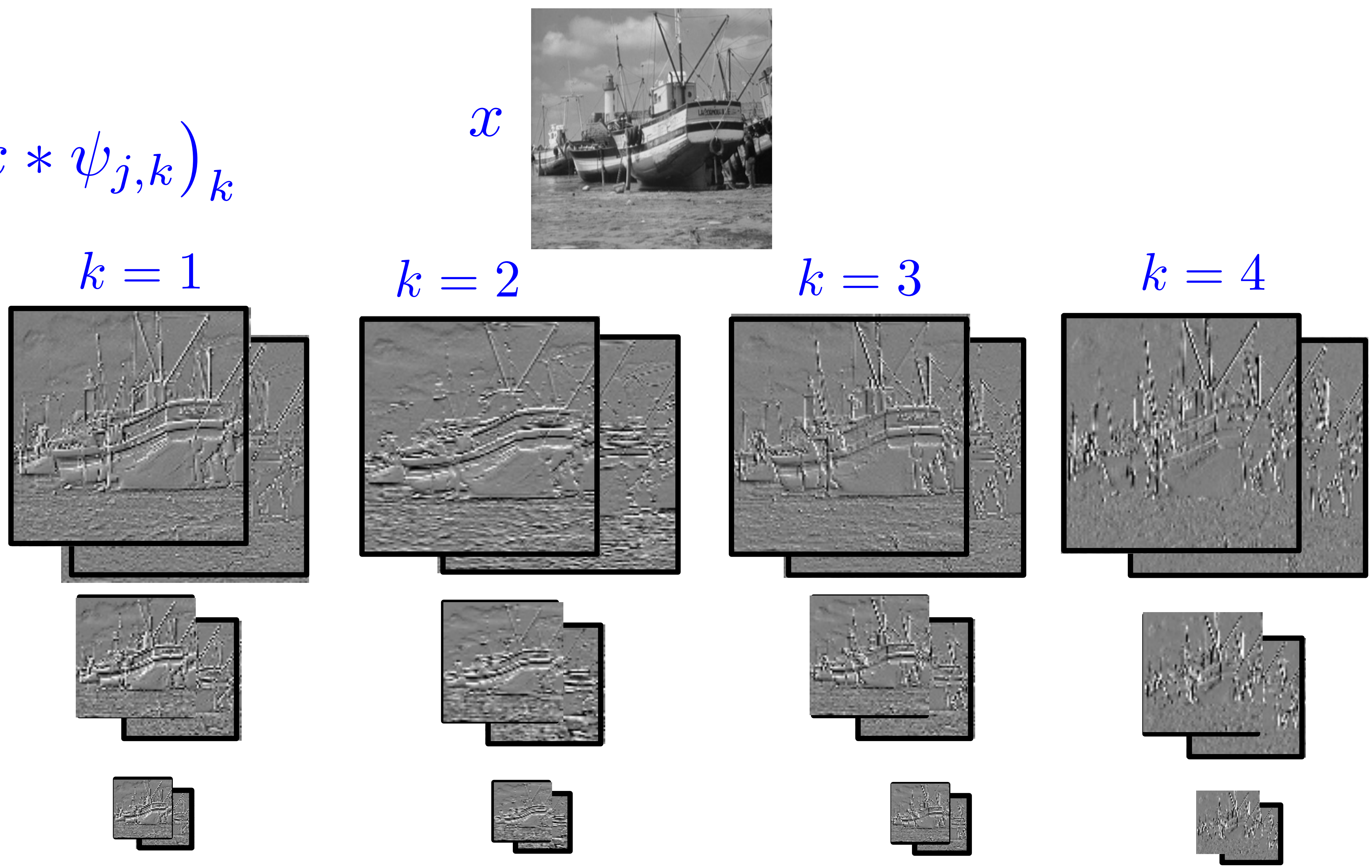
Frequency (Fourier) domain



Long-Range Dependencies

- Need to build models of $p(\bar{x}_j | x_j) = p(\bar{x}_j | \bar{x}_{j+1} ... \bar{x}_\ell ...)$

$$\bar{x}_j = \left(x * \psi_{j,k}\right)_k$$



Nearly no correlation at different positions, scales, orientations because phases of wavelets coefficients oscillate at different frequencies.

How to capture dependencies across scales ?

Wavelet Modulus

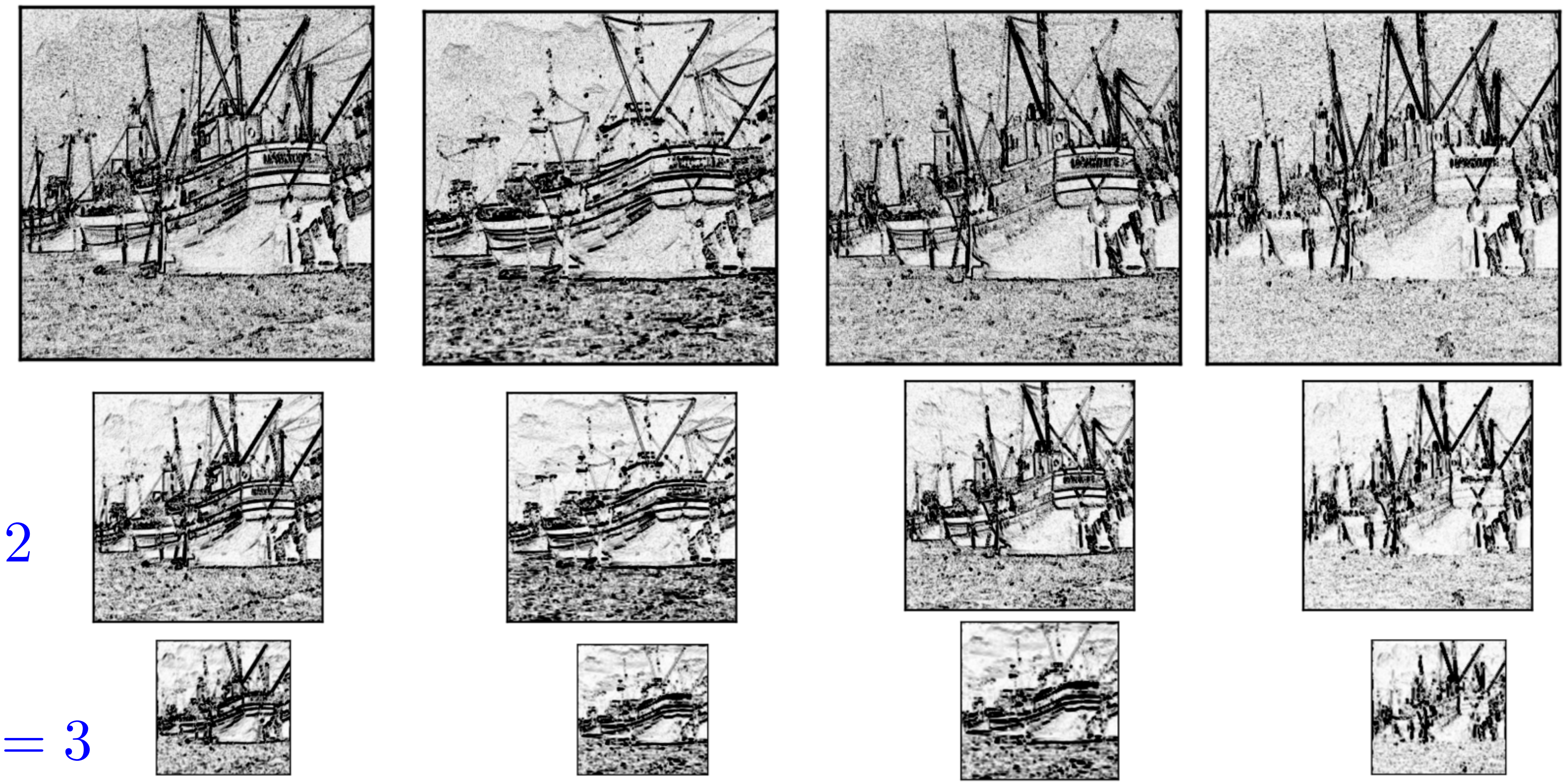
$$|\bar{x}_j| = \left(|x * \psi_j^k|\right)_k$$

"Edge detection"

$j = 1$

$j = 2$

$j = 3$

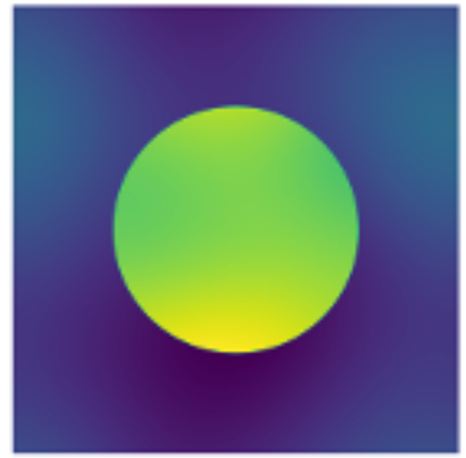


Long-range correlations across positions, scales, orientations

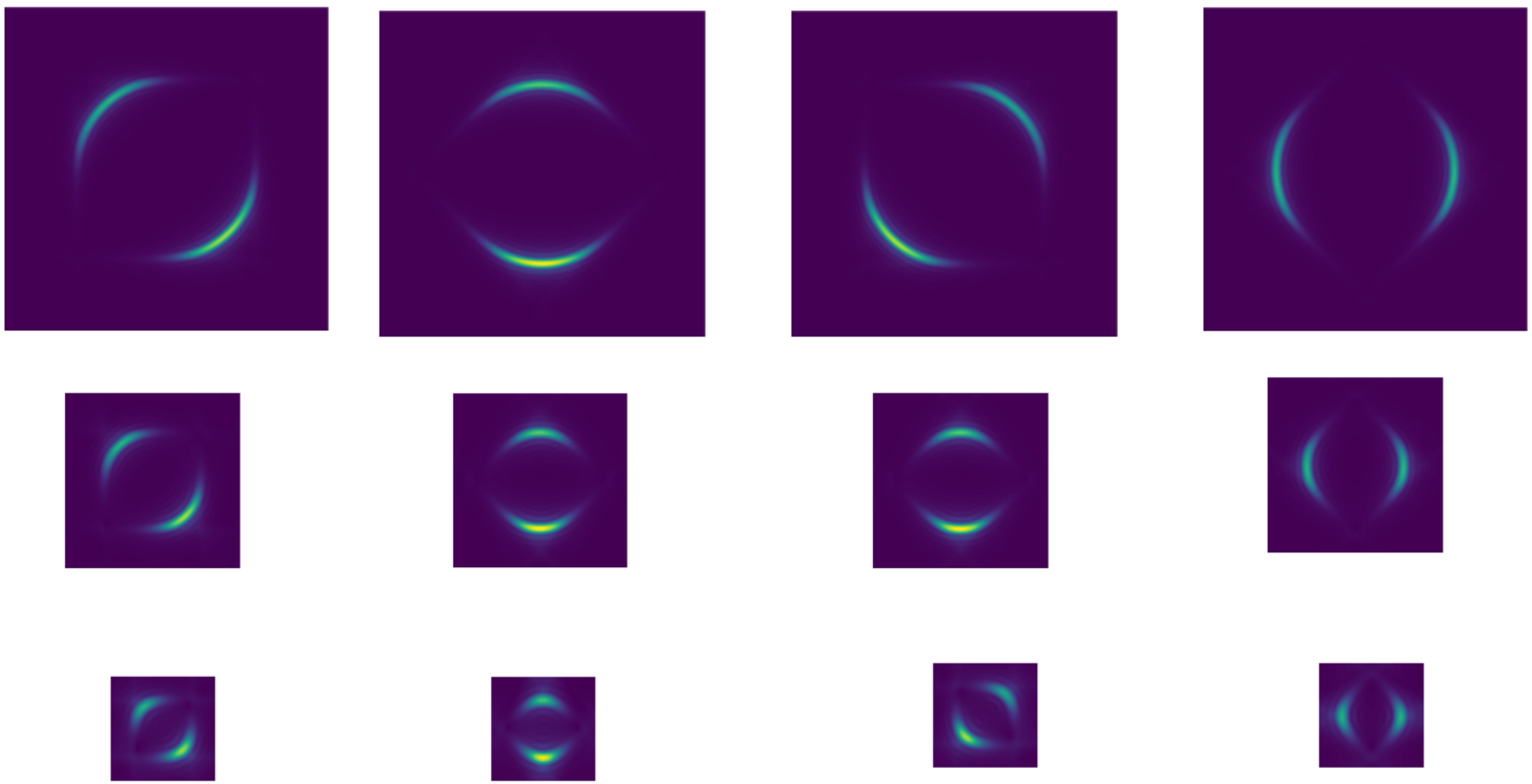
Geometry

1st wavelet transform

$$|\bar{x}_j| = \left(|x * \psi_j^k|\right)_k$$

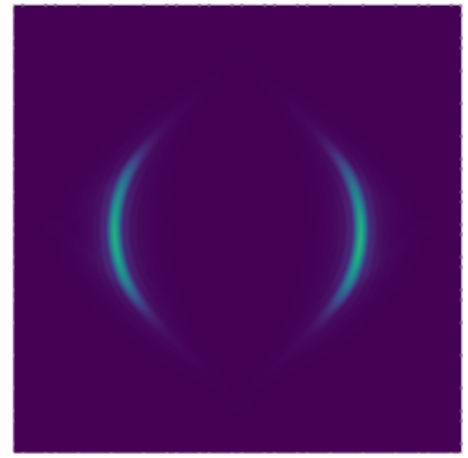


"Edge detection"



Directional Regularity

$|x * \psi_j^k|$



for $k = 4, j = 1$

2nd wavelet transform

second wavelet perpendicular to first wavelet

$|x * \psi_{j,k}| * \psi_{\ell,k^\perp}$

$\ell = j + 1$

$\ell = j + 2$

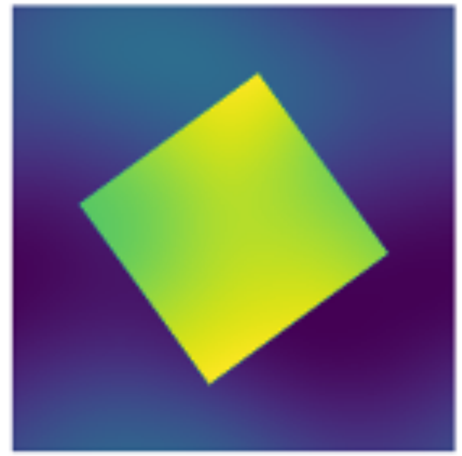
$\ell = j + 3$

Sparse

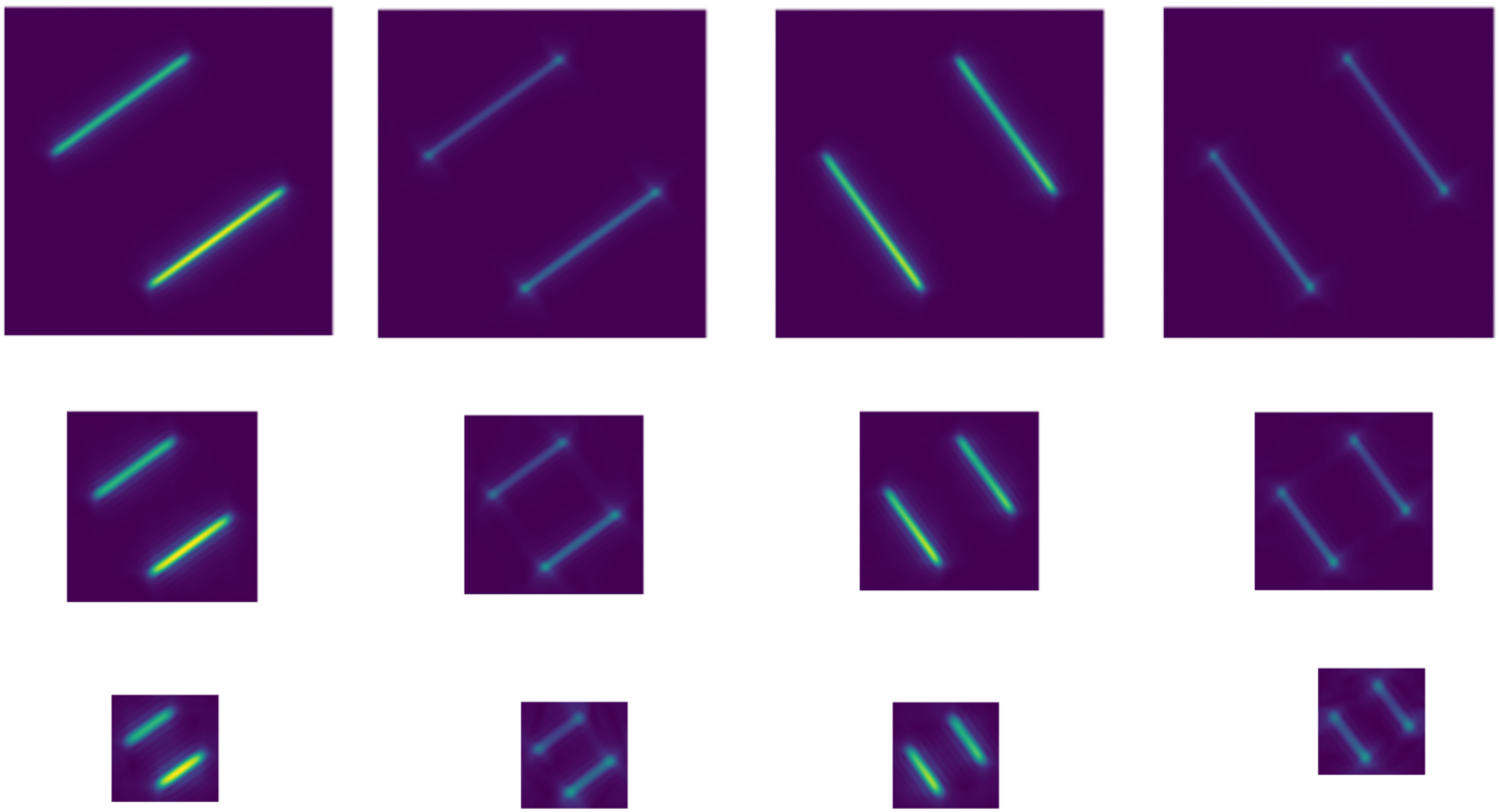
Multiscale Image Geometry

1st wavelet transform

$$|\bar{x}_j| = \left(|x * \psi_j^k|\right)_k$$



"Edge detection"



$j = 1$

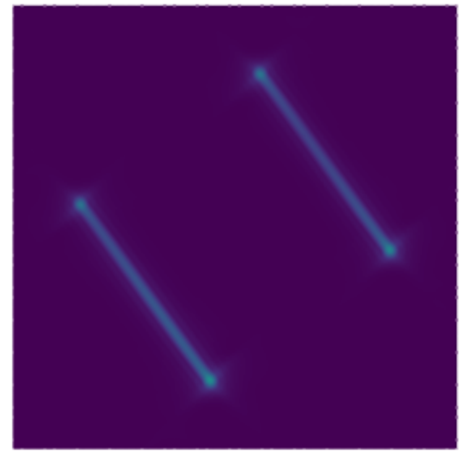
$j = 2$

$j = 3$

Long-range correlations across positions, scales, orientations

Directional Regularity

$|x * \psi_j^k|$

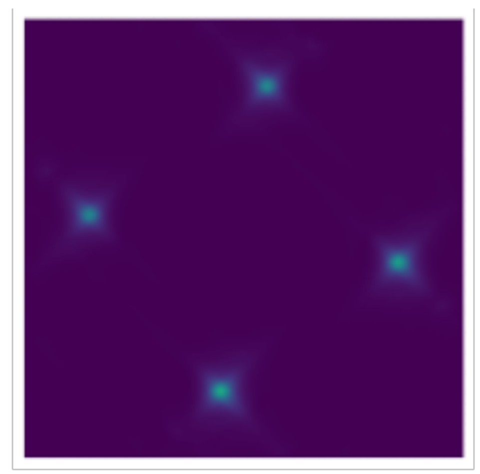


for $k = 4, j = 1$

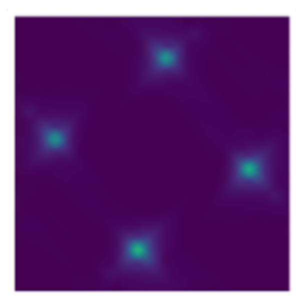
2nd wavelet transform

second wavelet perpendicular to first wavelet

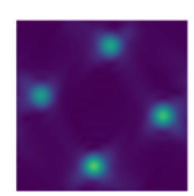
$|x * \psi_{j,k}| * \psi_{\ell,k^\perp}$



$\ell = j + 1$



$\ell = j + 2$



$\ell = j + 3$

Sparse

Geometric Directional Regularity

Theorem *(N. Cuvelle-Magar, S. M.)*

If x is a $\mathbf{C}^2$ image besides piecewise $\mathbf{C}^2$ edges curves then

for all $k, j' \geq j$ and $\alpha < 2$ there exists $C > 0$ with

$$\| |x * \psi_{j,k}| * \psi_{j',k^\perp} \|_1 \leq C\, 2^{\alpha j'} .$$

Scattering Covariance Model

- Wavelet coefficients at scale 2^j: $\bar{x}_j(u) = \big(x * \psi_{j,k}(u)\big)_k$
- Exponential models: $p(\bar{x}_j | \bar{x}_{j+1} \ldots) = \mathcal{Z}_j^{-1} e^{-\theta_j^T \Phi(\bar{x}_j | \bar{x}_{j+1} \ldots)}$
- Scattering: $S_j = \big(\bar{x}_j \,,\, |\bar{x}_j| * \psi_{\ell,k}\big)_{\ell > j, k}$
- Scattering covariance: $\Phi(\bar{x}_j | \bar{x}_{j+1} \ldots) = \big(S_j\, S_{j'}^T\big)_{j' \geq j}$
- Scattering covariance model: *Etienne Lempereur*

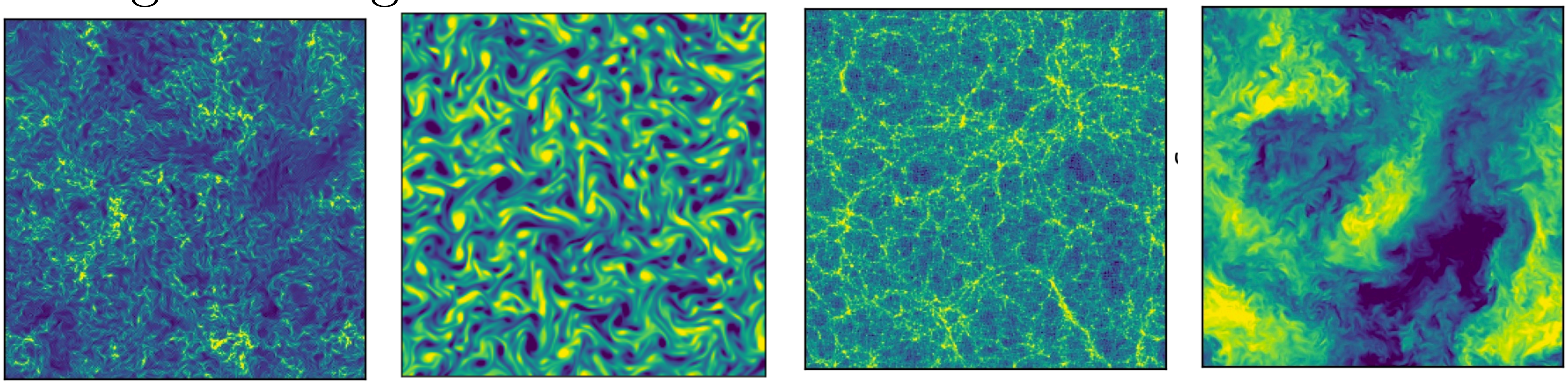
$$\theta_j^T \Phi(\bar{x}_j | \bar{x}_{j+1} \ldots) = \sum_{j' \geq j} S_j^T\, K_{j'} S_{j'}$$

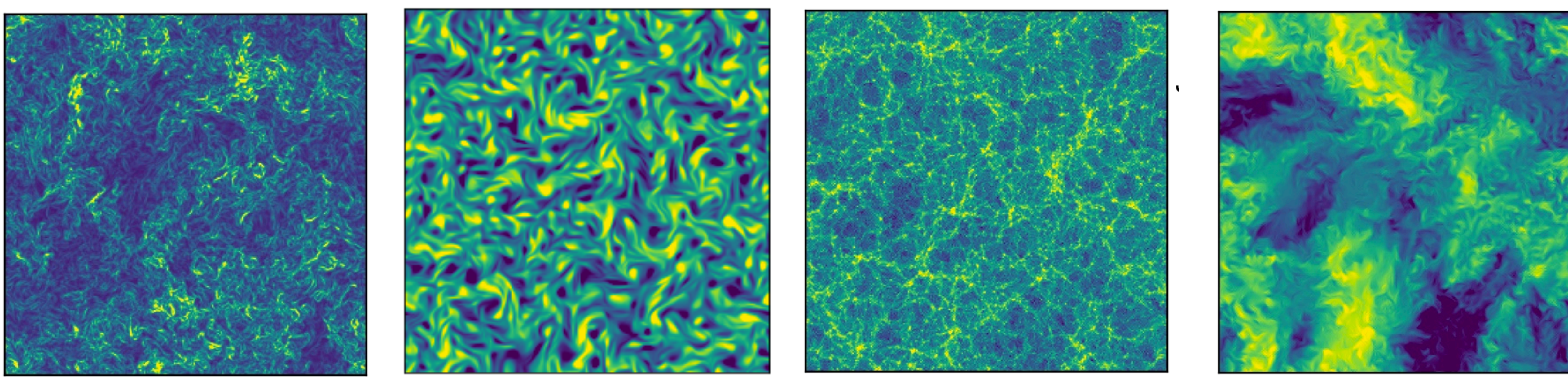
Spatially local interaction matrices $K_{j'}$ but across scales, with $O(\log^3 d)$ non-zero interaction coefficients across scales.

Generation from Scattering Models

E. Allys, S. Cheng, E. Lempereur, B. Ménard, R. Morel, S. M.

Original images of dimension $d = 5\,10^4$





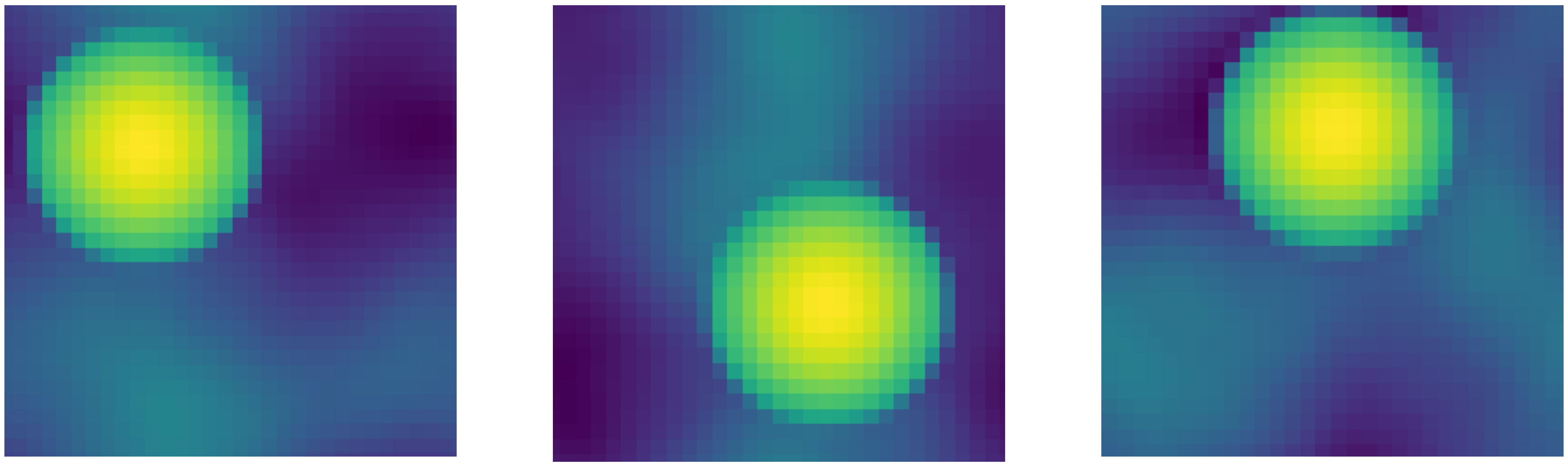
Generated with models having 500 parameters

Reproduces moments of order 3 (bispectrum) and 4 (trispectrum)

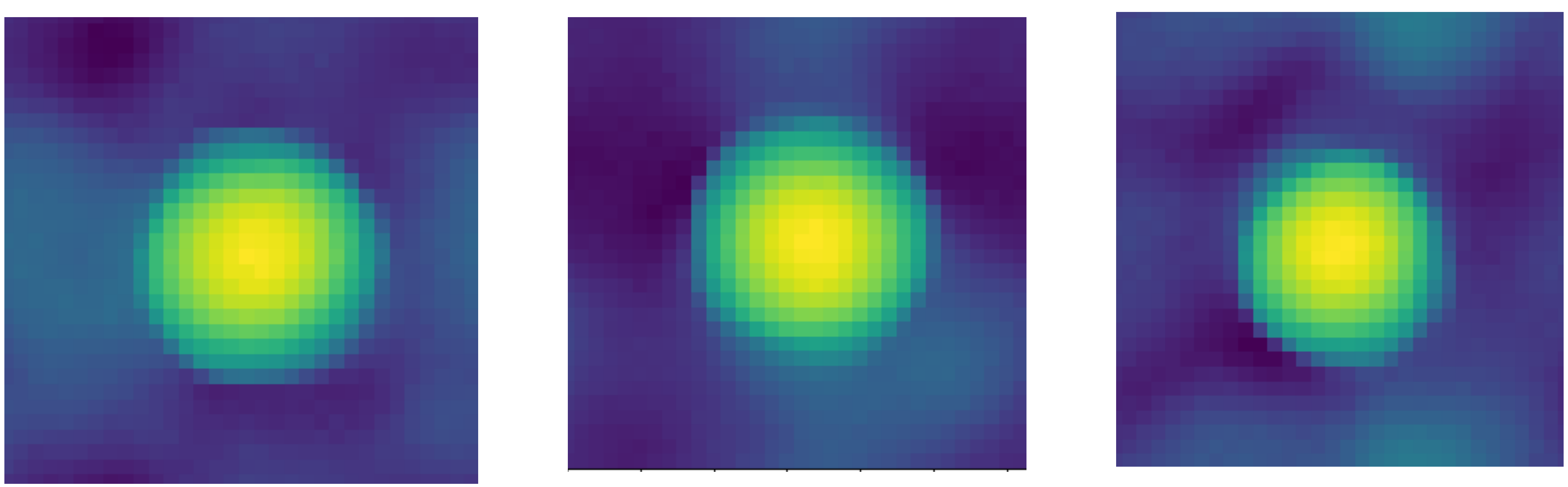
Generation from Scattering Models

N. Cuvelle-Magar, E. Lempereur

Original images of dimension $d = 32 \times 32$



Generated by sampling scattering models



Scattering interactions can model regular geometries

Equivalent to a network with 2 hidden layers.

Conclusion

- Neural network score generation do generalise: they do not just memorise if the data set is large enough: very large…
- They define geometrically adapted harmonic bases

Generalisation in diffusion models arises from geometry-adaptive harmonic representation, ICLR 2024 *Z. Kadkhodaie, F. Guth, , E. Simoncelli, S. M.*

- Learning the geometry of complex physics is possible with much fewer parameters, within the renormalisation group framework:

Multiscale Data-Driven Energy Estimation and Generation
Phys. Rev X 13, *T. Marchand, M. Ozawa, G. Biroli, S. M.*

Scattering Spectra Models for Physics, arXiv:2306.17210
S. Cheng, R. Morel, E. Allys, B. Menard, S. M.

Hierarchic flows to estimate and sample high-dimensional probabilities
arXiv:2405.03468, *E. Lempreur, S. M.*