De l'IA Générative à la Physique Statistique

S. Mallat

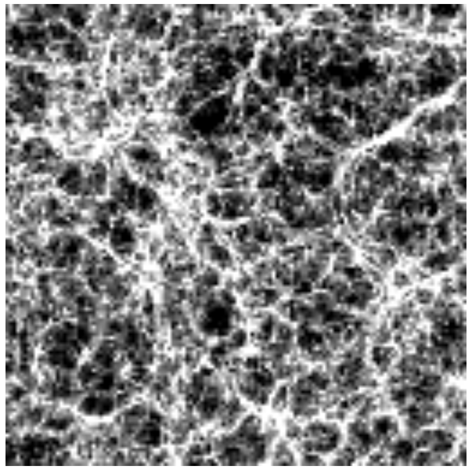
Collège de France
École Normale Supérieure

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

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

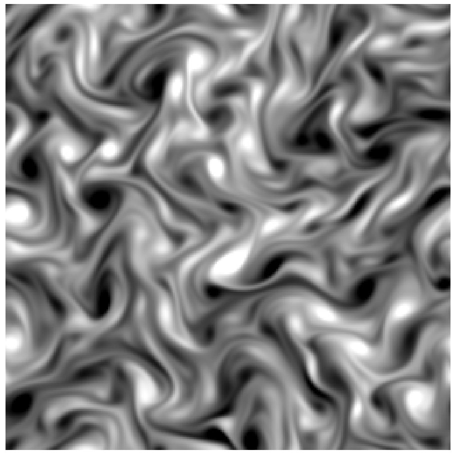
Curse of dimensionality if $d \gg 1$.

Statistical physics



Cosmic web

long-range



Turbulences

geometry

since 1940's

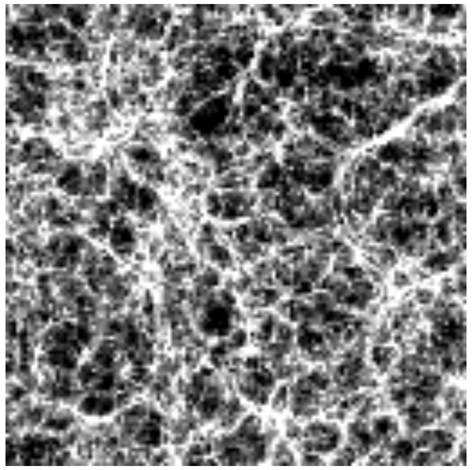
Learning Physics and Image Generation

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

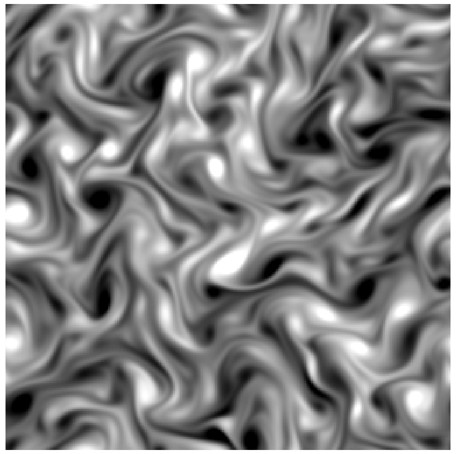
$$p(x) = \mathcal{Z}^{-1}\, e^{-U(x)} \quad \text{for } x \in \mathbb{R}^d$$

Curse of dimensionality if $d \gg 1$.

Statistical physics



Cosmic web



Turbulences

long-range geometry

since 1940's

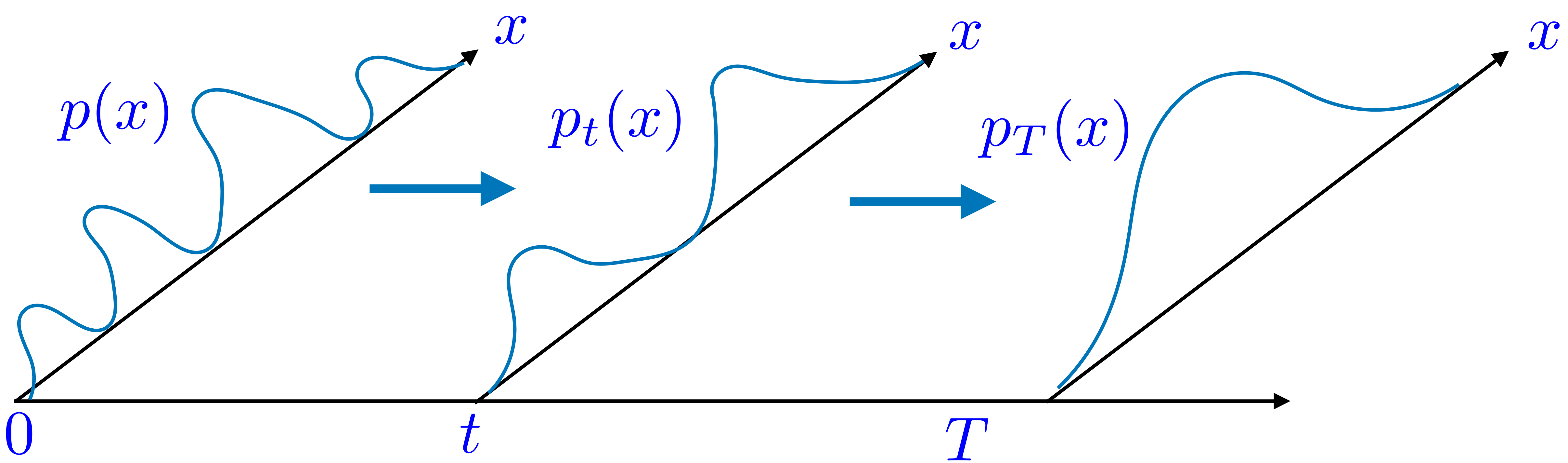
Image generation by score denoising

Does it memorise or generalise ?

How does it circumvent the curse ?

Transport of Probabilities

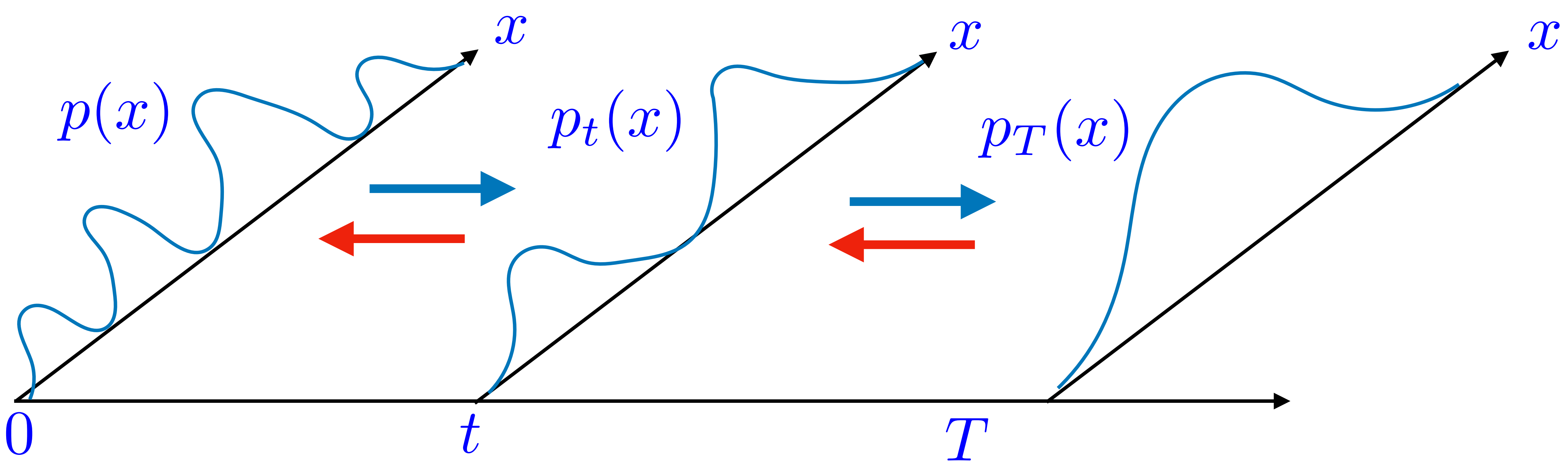
- Define a transport from p to a simple p_T



Transport of Probabilities

- Define a transport from p to a simple p_T

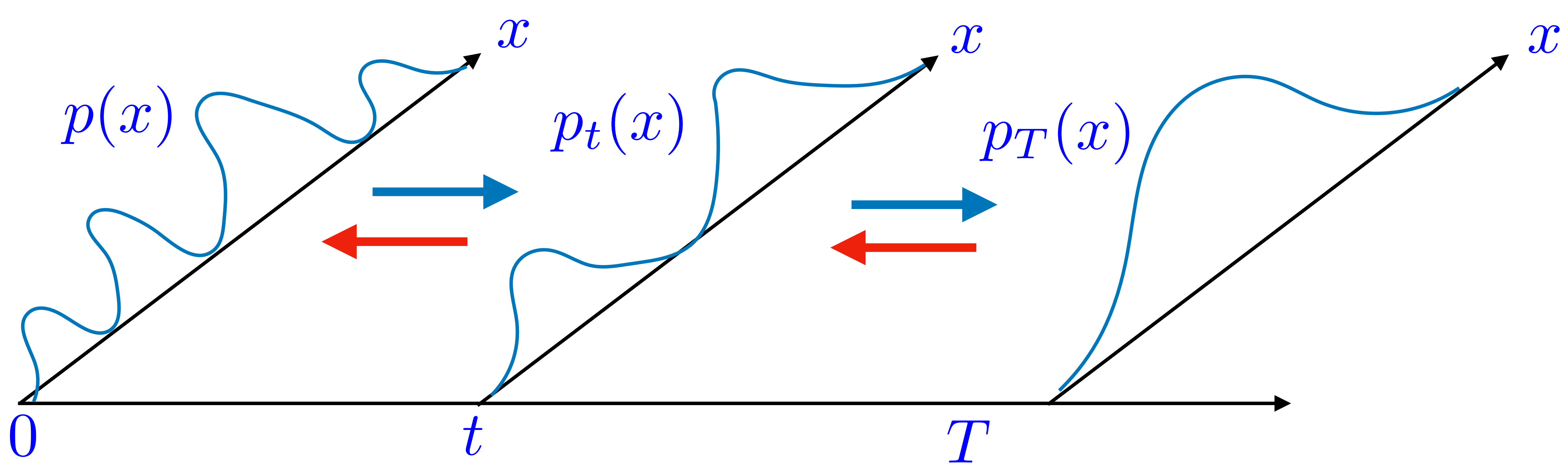
Learn the inverse transport from data



Transport of Probabilities

- Define a transport from p to a simple p_T

  Learn the inverse transport from data

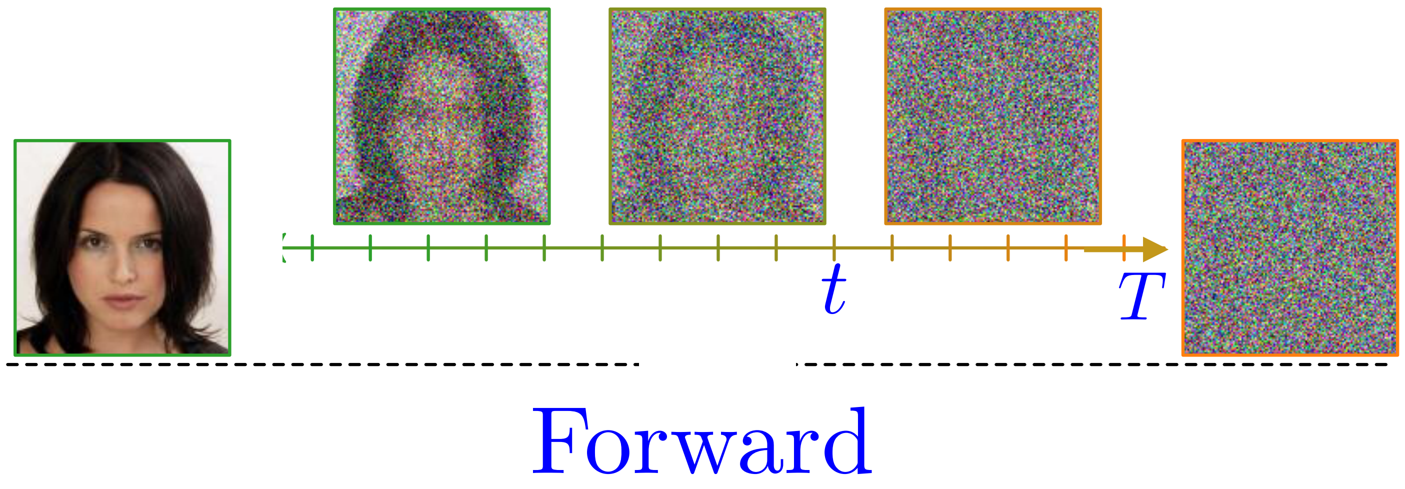


- The inverse transport is learned from data. What transport ?
  - AI score diffusion generation (2020): along noise variance
  - Physics Wilson renormalisation group (1970): along scales

Score Diffusion Generation

Yang Song et. al.

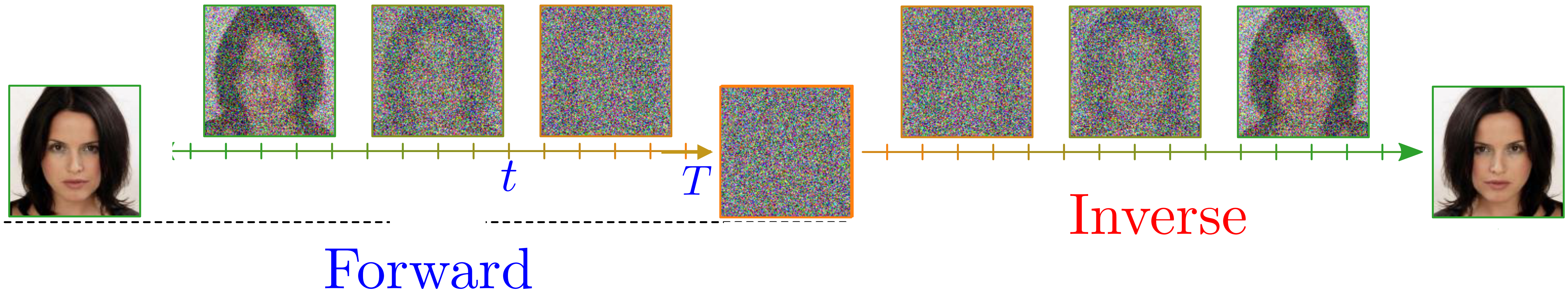
- Forward diffusion: add noise with Ornstein-Uhlenbeck equation



Score Diffusion Generation

Yang Song et. al.

- Forward diffusion: add noise with Ornstein-Uhlenbeck equation



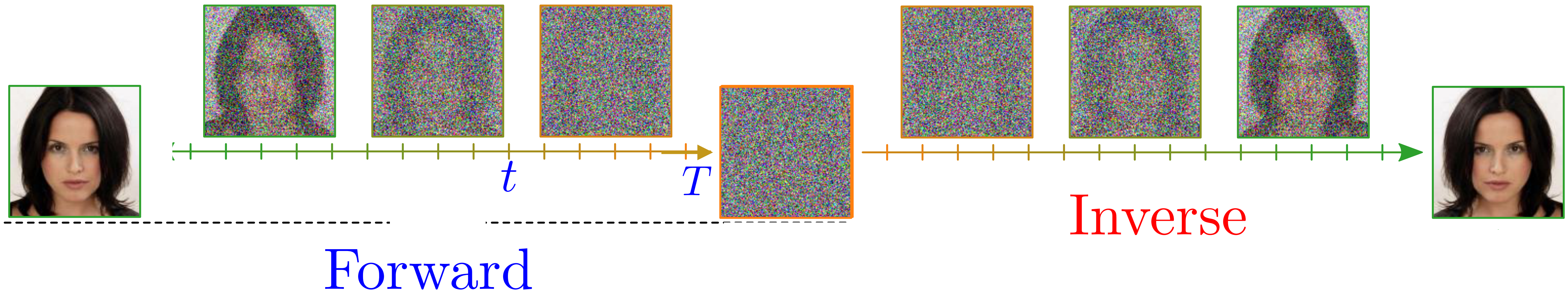
- The diffusion is inverted with a damped-Langevin equation:

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

Score Diffusion Generation

Yang Song et. al.

- Forward diffusion: add noise with Ornstein-Uhlenbeck equation



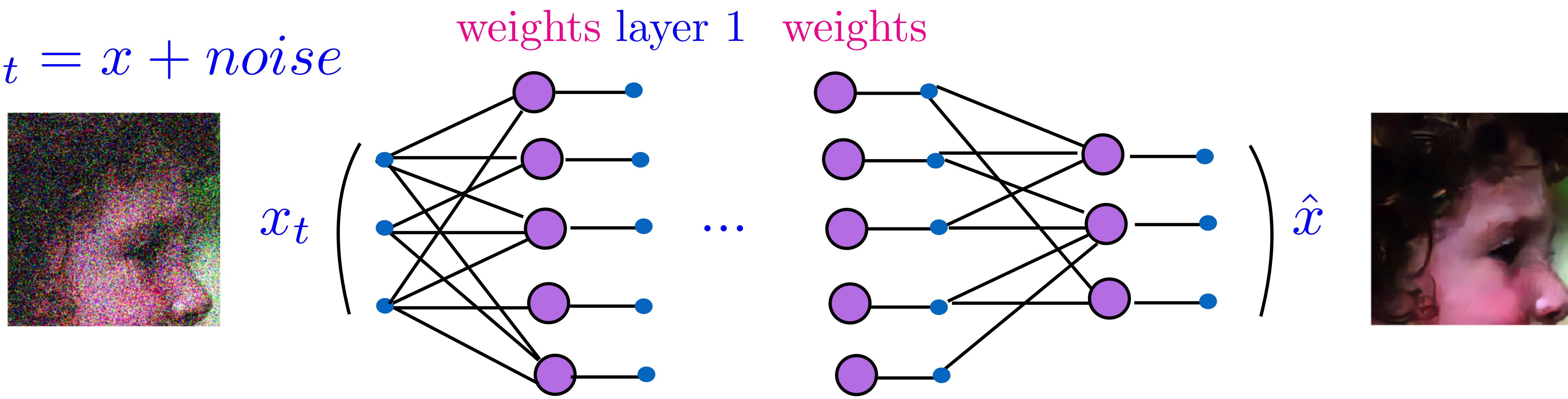
- The diffusion is inverted with a damped-Langevin equation:

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

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

Score Estimation by Denoising

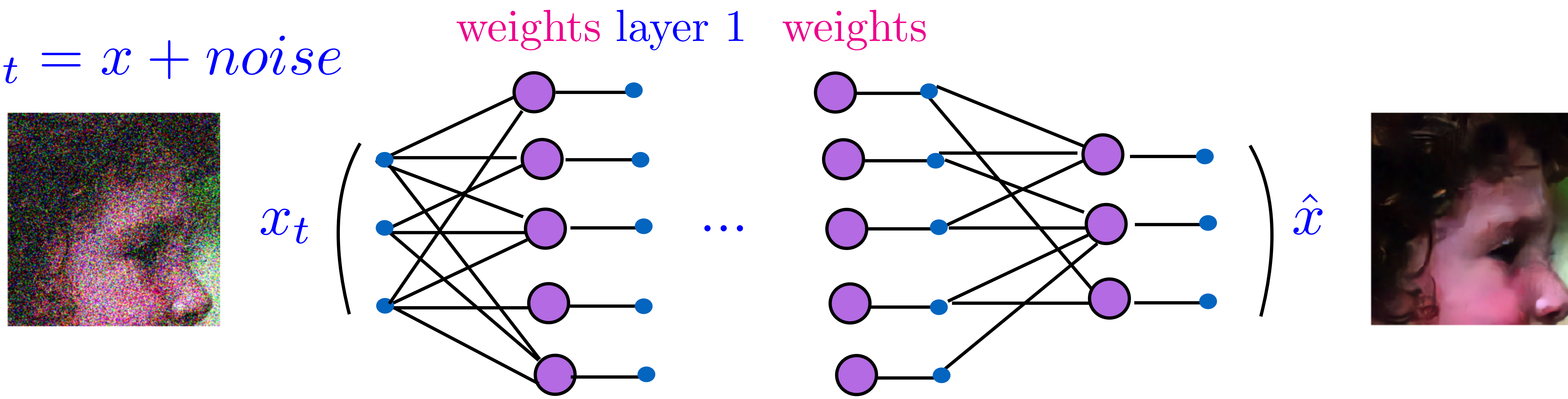
$x_t = x + noise$



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

Score Estimation by Denoising

$x_t = x + noise$



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

Tweetie, Robbins, Myasawa formula for the optimal $\hat{x}$:

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

Does it really work ? 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 ?

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

N=1

Closest in S_1

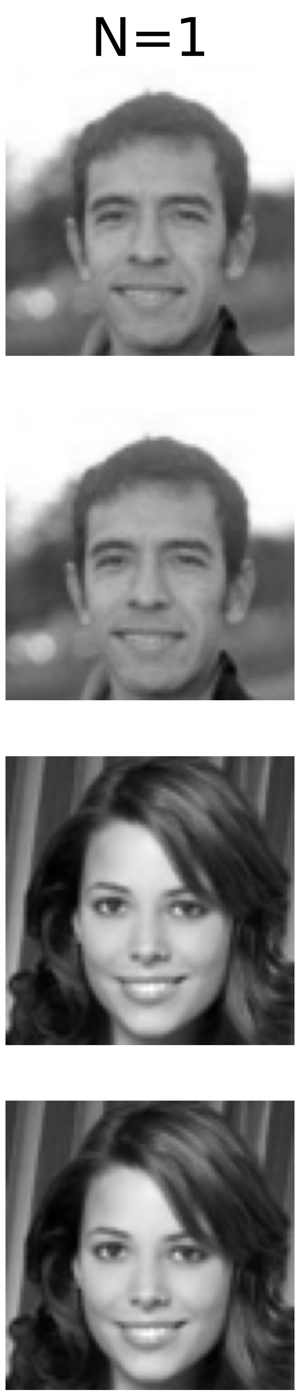
Synthesized from S_1

Synthesized from S_2

Closest in S_2

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



Closest in S_1

Synthesized from S_1

Synthesized from S_2

Closest in S_2

Generalises or Memorises ?

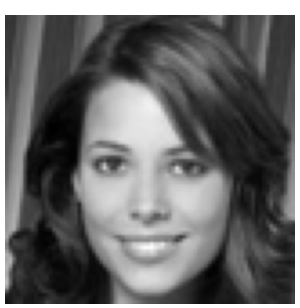
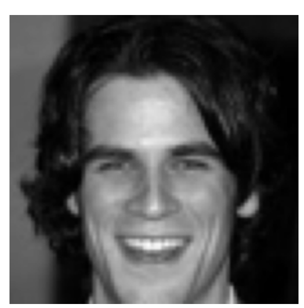
Images reconstructed from the same noise with 2 scores estimated from 2 different train sets S_1 and S_2 of N images of 80×80 pixels

N=1 N=10

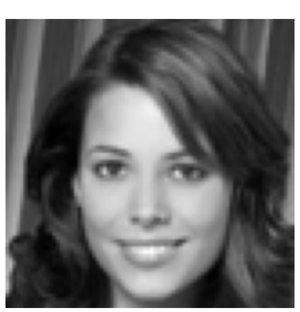
Closest in S_1

Synthesized from S_1


Synthesized from S_2

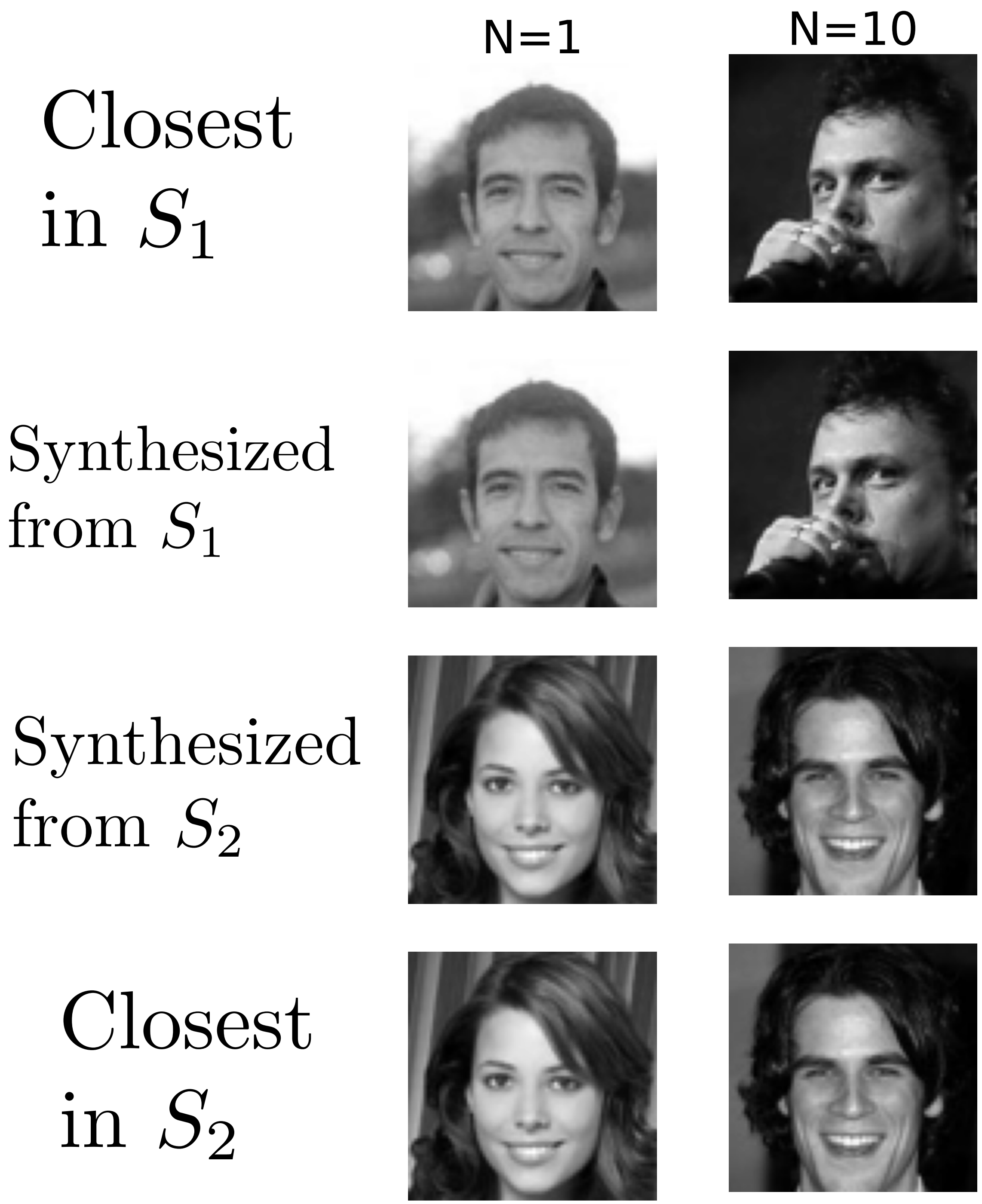



Closest in S_2



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

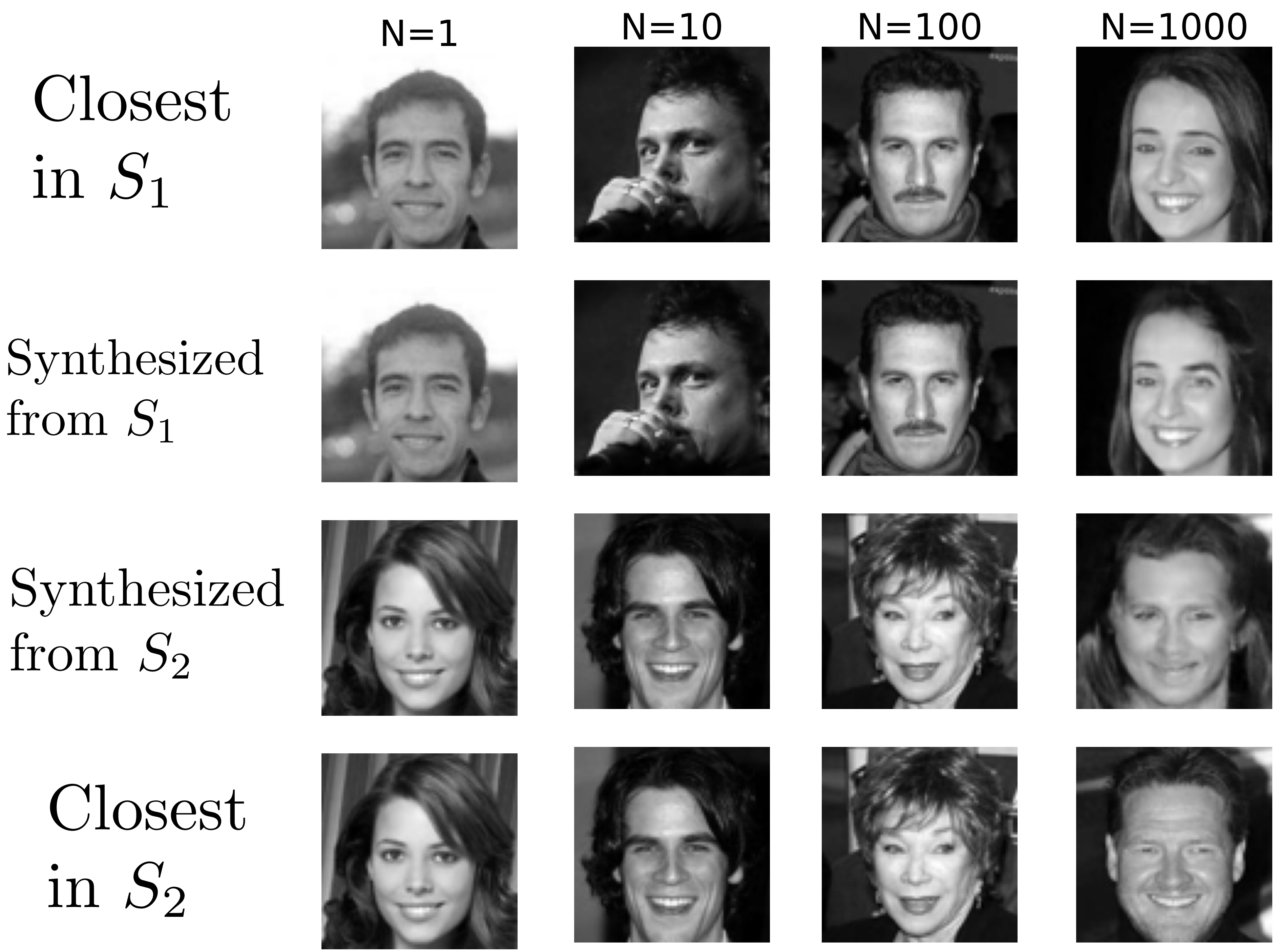


Generalises or Memorises ?

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

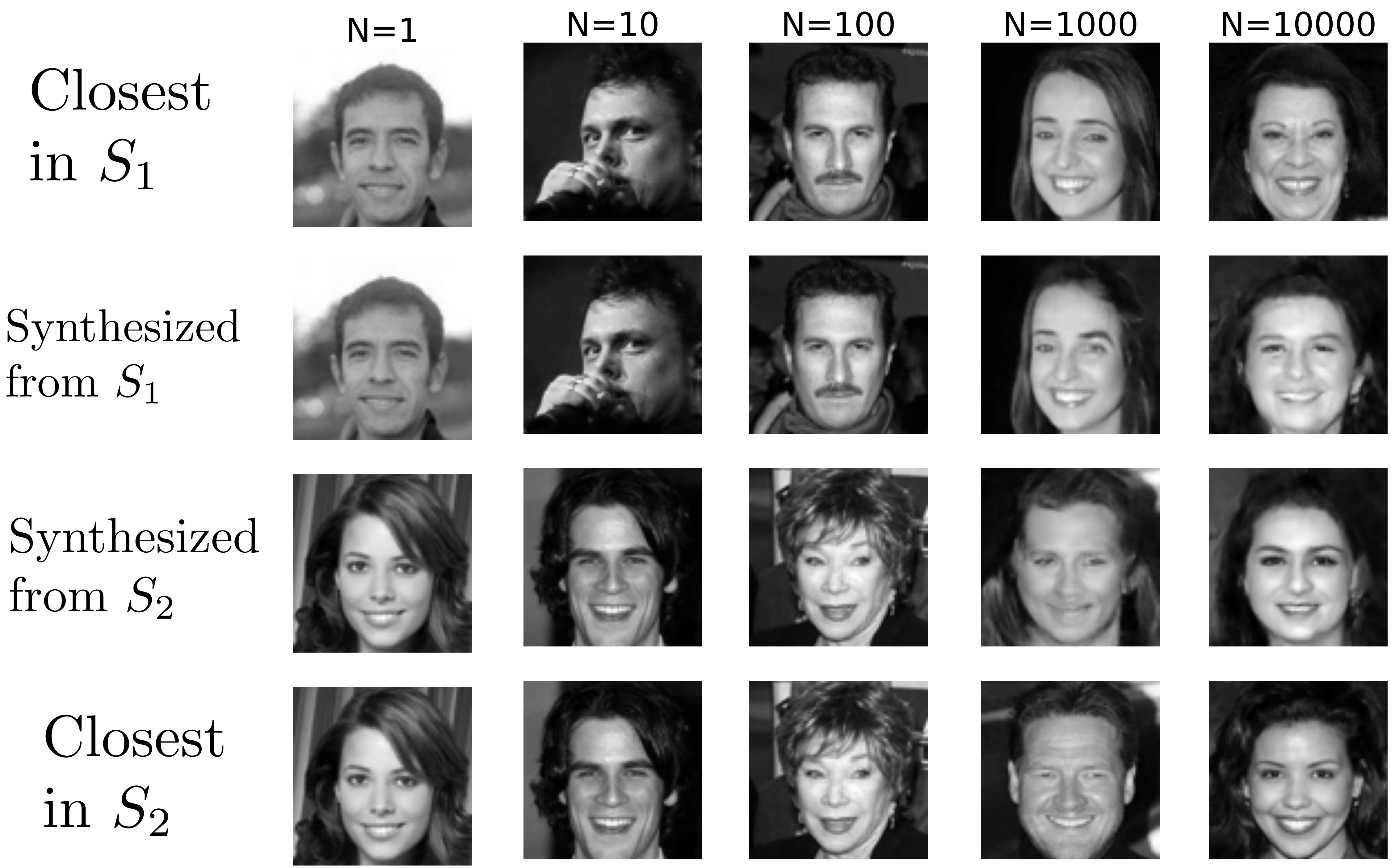
Generalises or Memorises ?

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 or Memorises ?

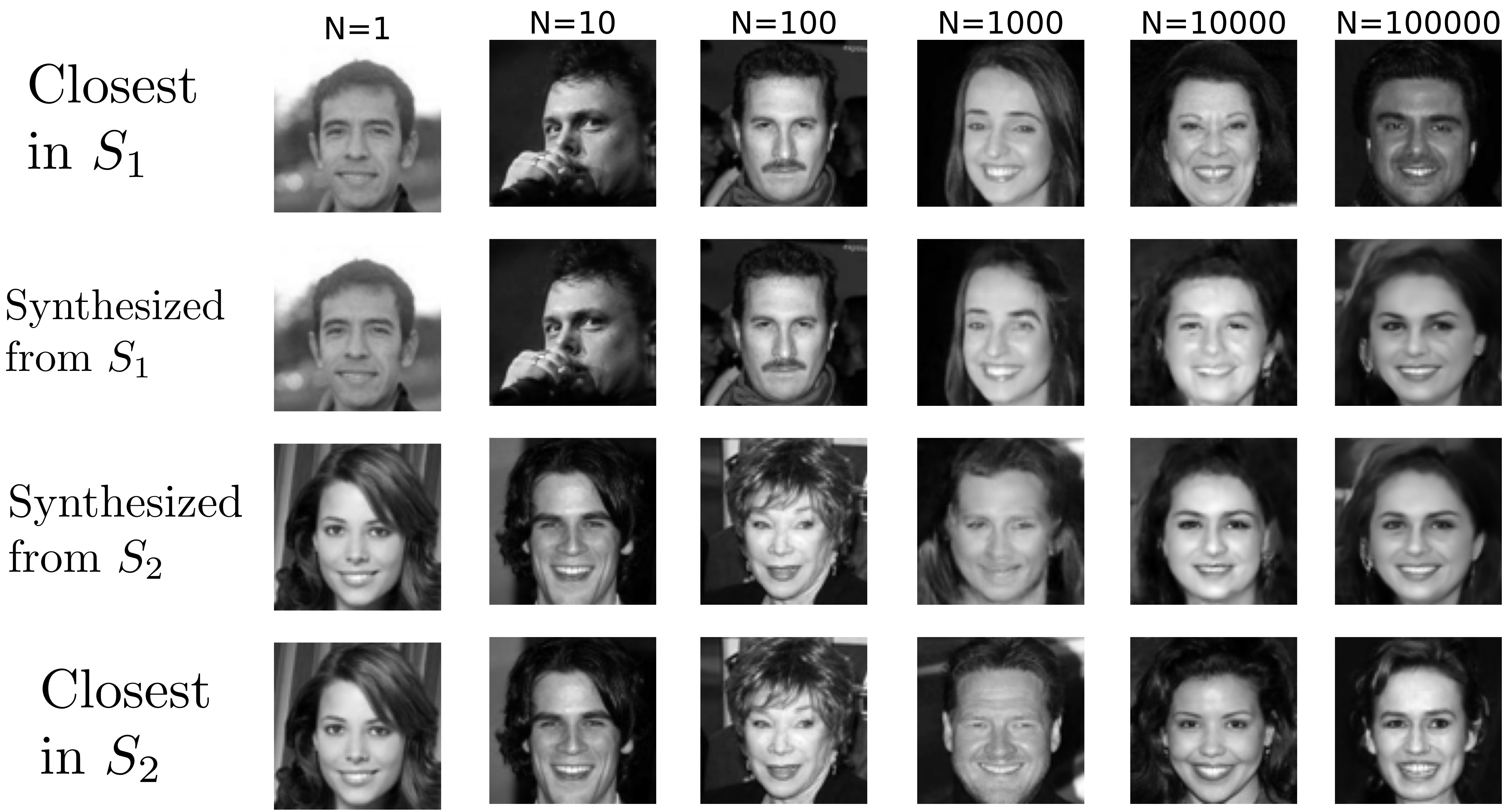
Images reconstructed from the same noise with 2 scores estimated from 2 different train sets S_1 and S_2 of N images of 80×80 pixels



Generalises 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

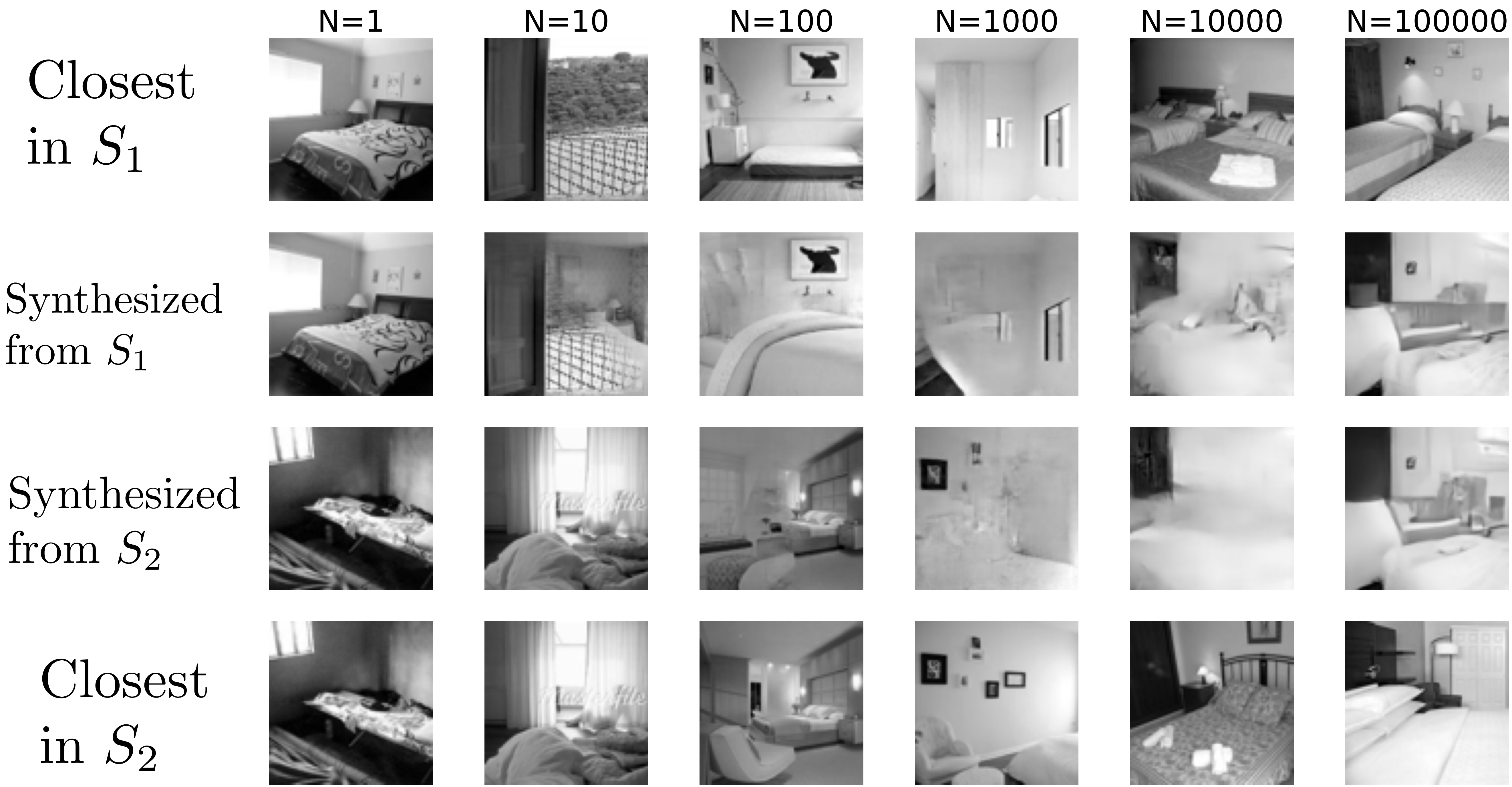
Synthesized from S_2

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

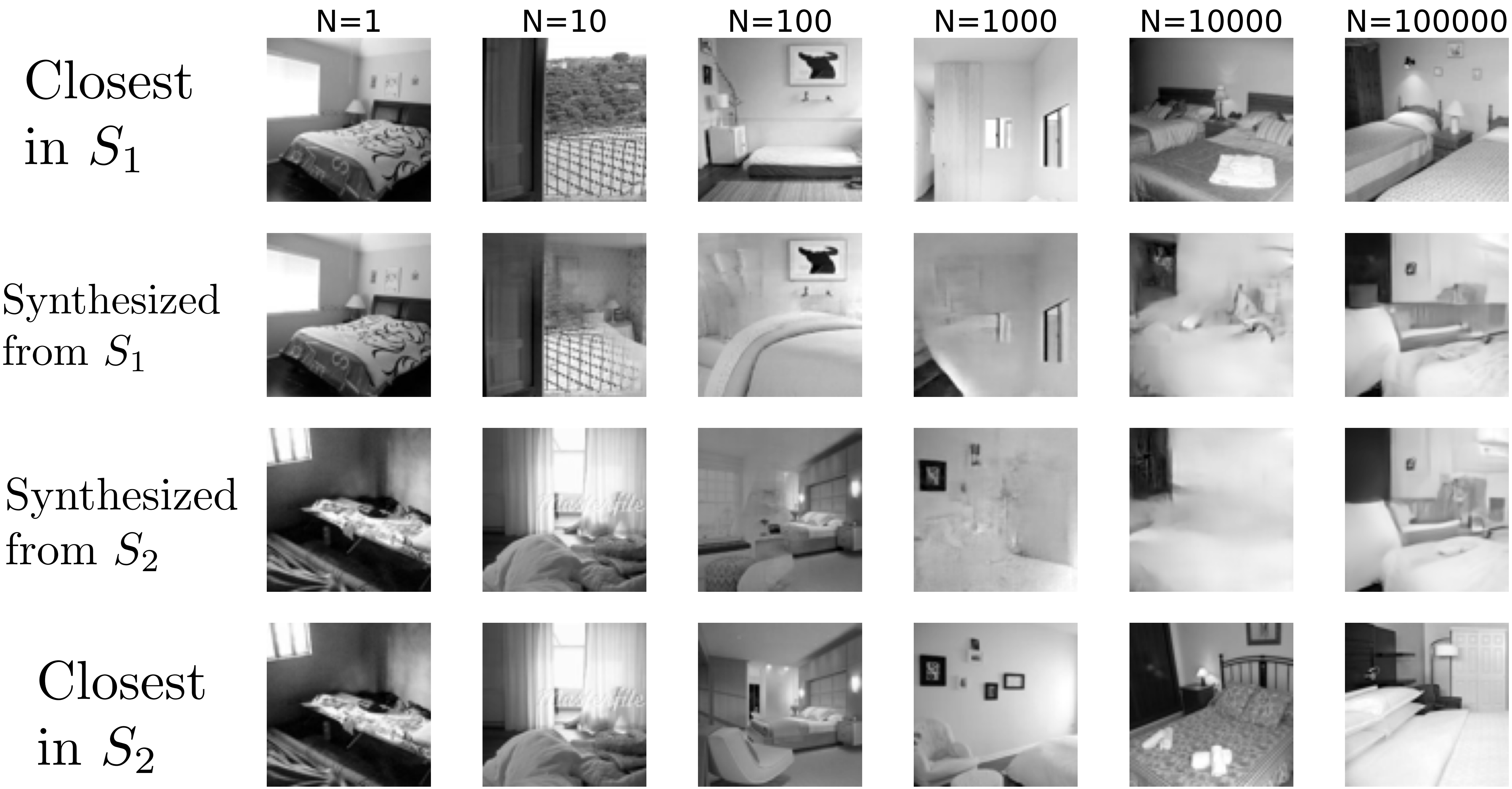
Generalises!



Generalisation Test: Memorise ?

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

Generalises!



2. High Dimensional Models

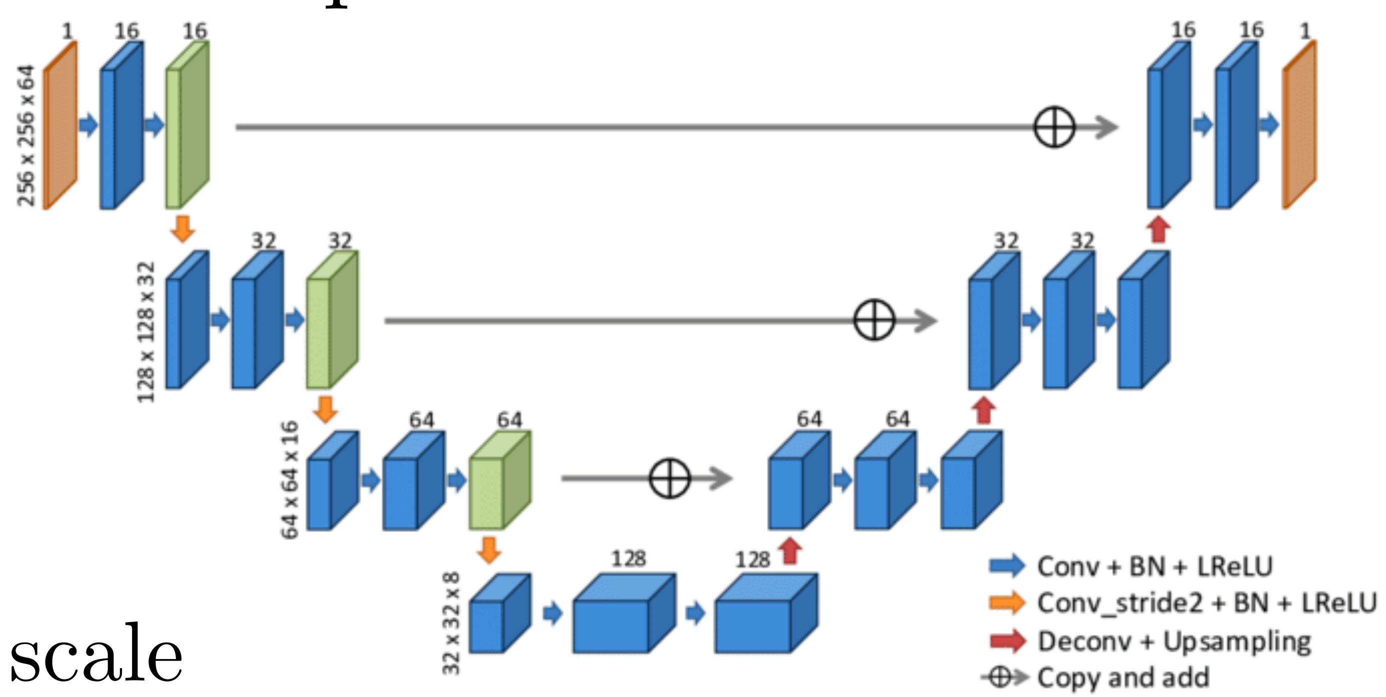
- Score diffusion generalises with enough training examples
- Generalisation depends upon the number of network parameters
- Circumvents the curse of dimensionality: how ?

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 ?

Can we build accurate models with fewer examples ?

Open the black box



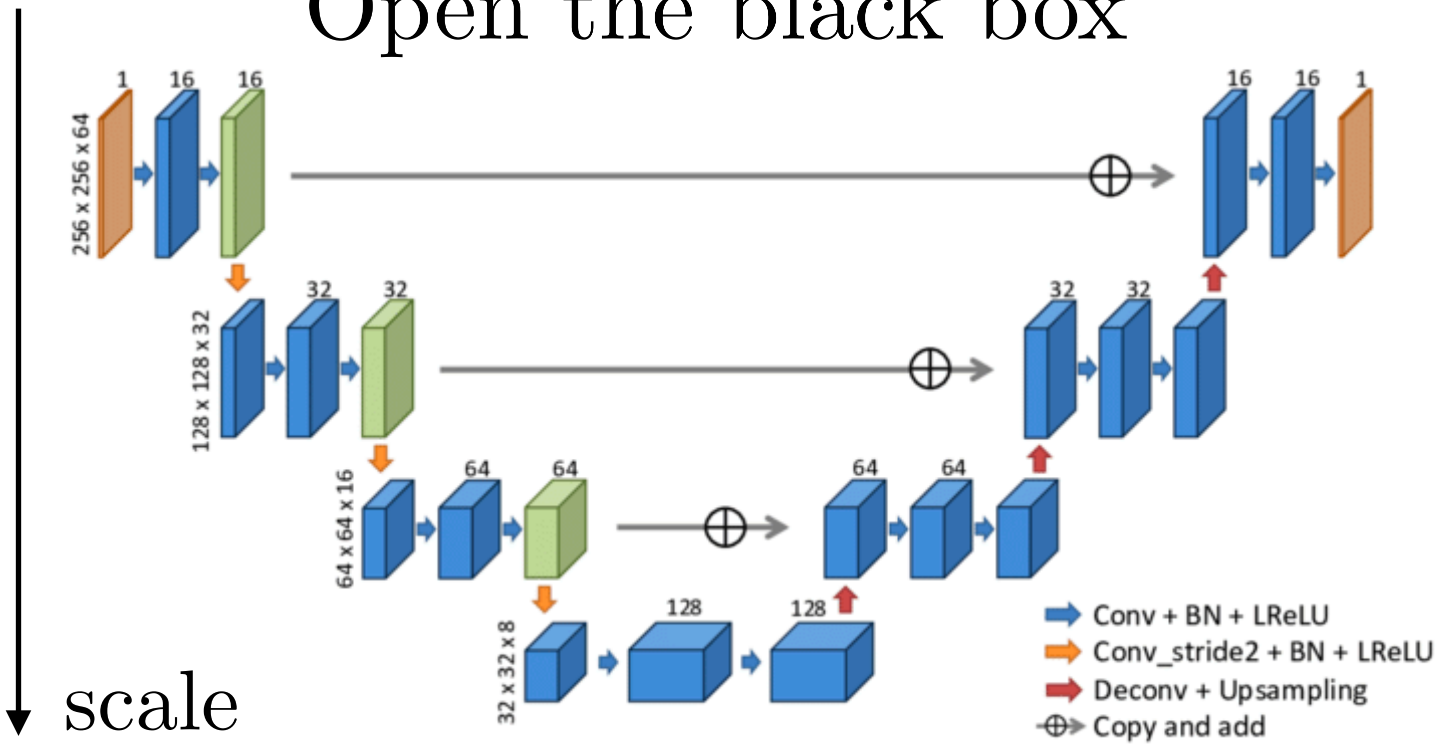
scale

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 ?

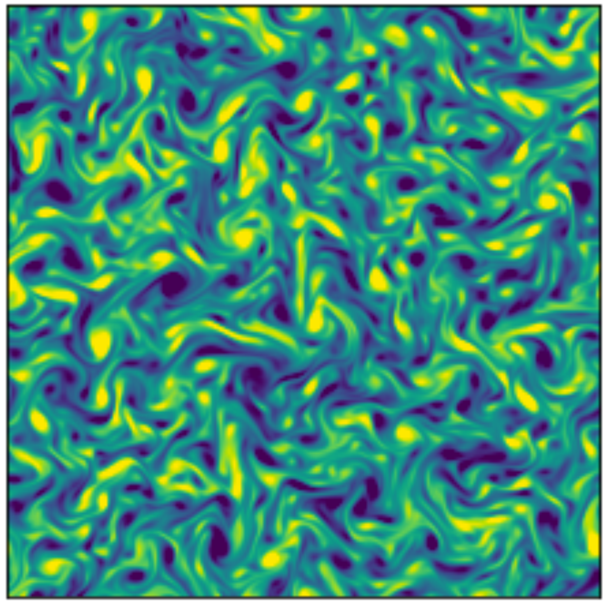
Can we build accurate models with fewer examples ?

Open the black box



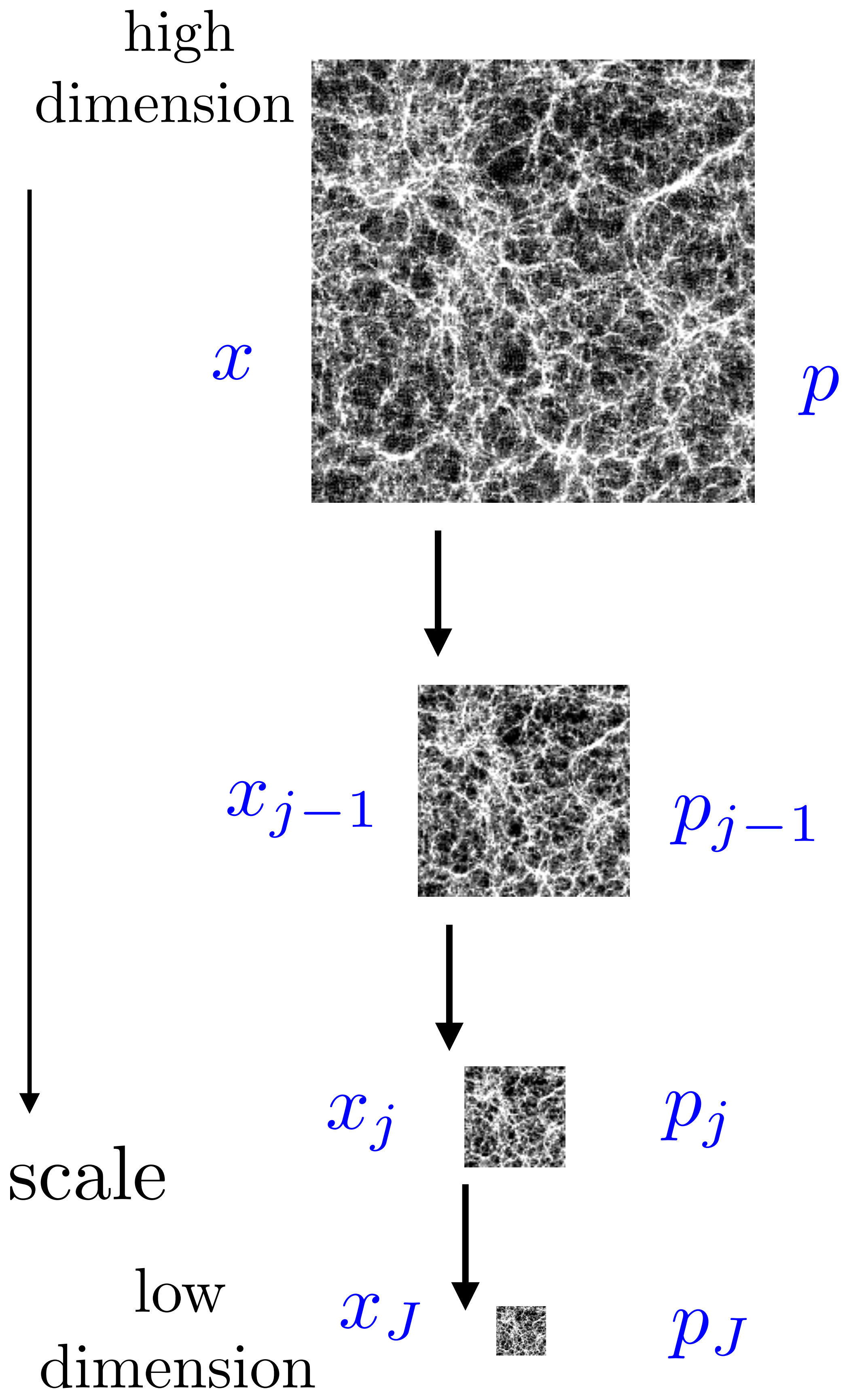
How to capture an image geometry ?

Can we model physical turbulences ?



Renormalisation Group : Hierachy

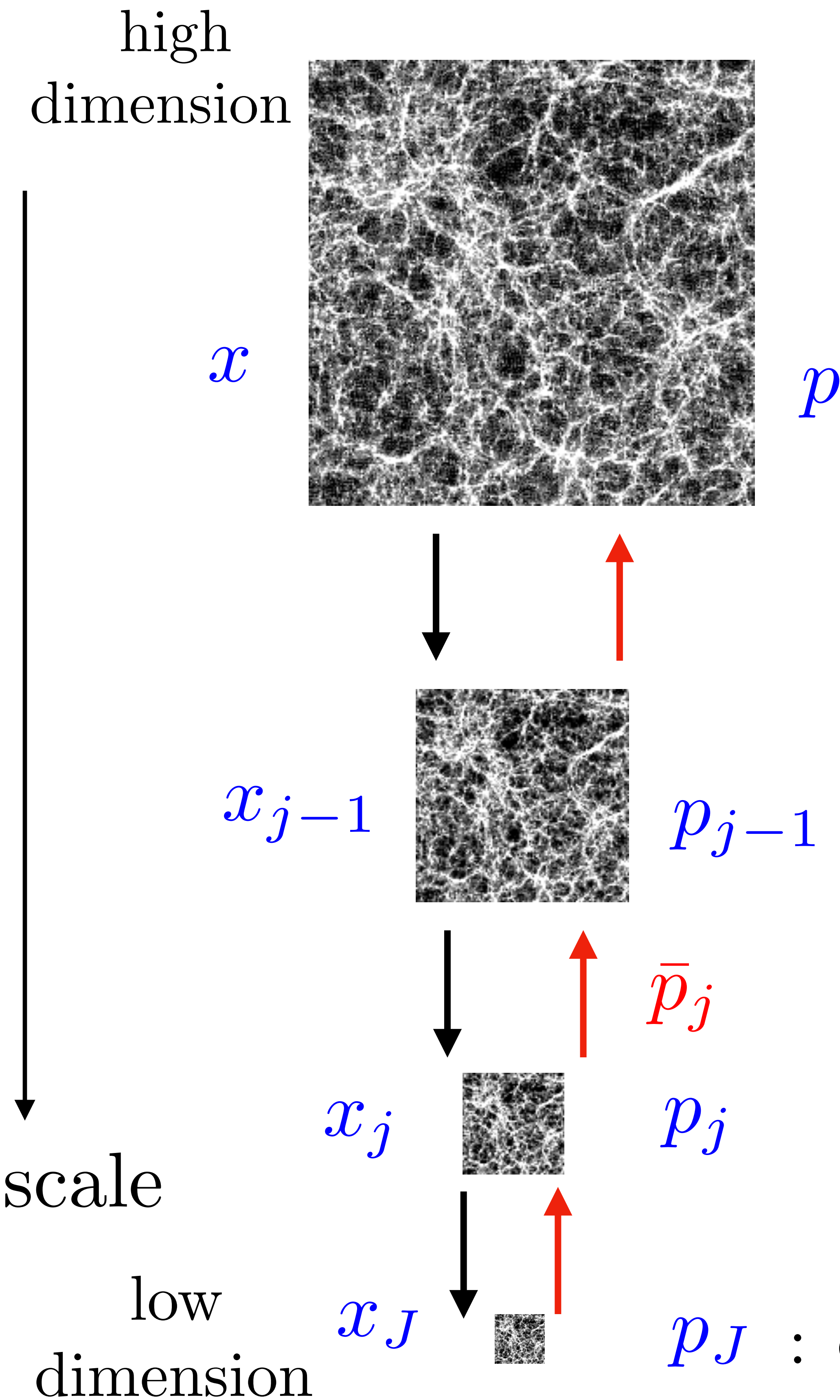
Kadanoff, Wilson 1970



Probability transport across scales

Renormalisation Group : Hierachy

Kadanoff, Wilson 1970



p_J : easy to estimate and sample

Probability transport across scales

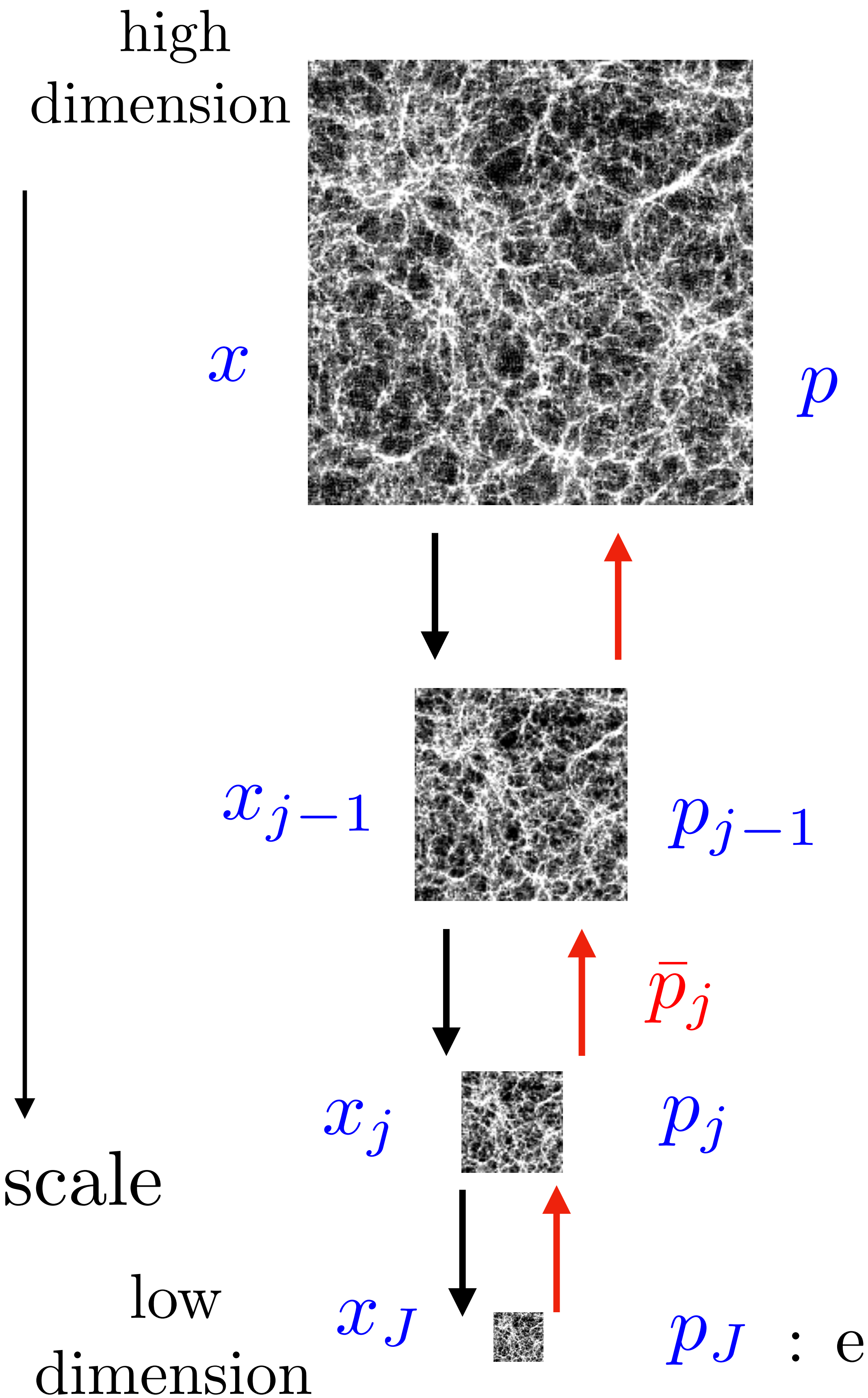
Inverse Markov chain

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

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

Renormalisation Group : Hierachy

Kadanoff, Wilson 1970



Probability transport across scales

Inverse Markov chain

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

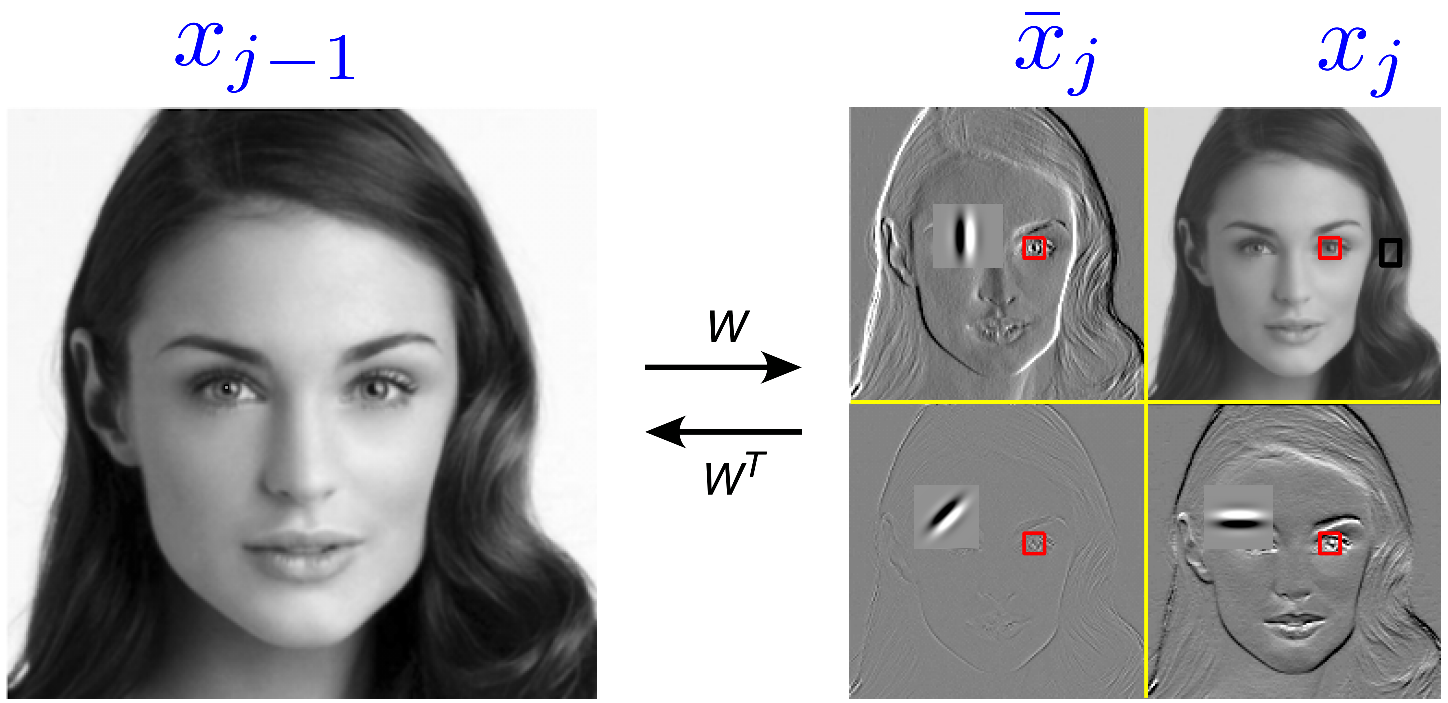
G. Biroli, E. Lempereur
T. Marchand, M. Ozawa, S. M.

Wilson:
Easier to estimate $\bar{p}_j(x_{j-1}|x_j)$
than directly $p_{j-1}(x_{j-1})$

p_J : easy to estimate and sample

Transition Probabilities Across Scales

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



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

Local conditional dependencies over wavelet coefficients.

Hierarchic Sampling

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

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

Hierarchic Sampling

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

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

x_J

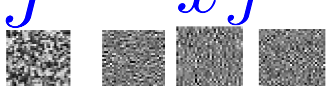
sample $p_J(x_J)$

Hierarchic Sampling

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

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

x_J $\overline{x}_J$



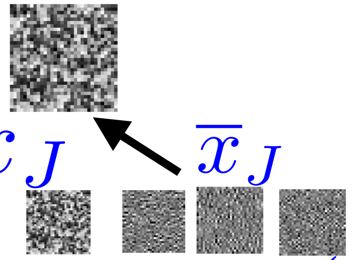
sample $\bar{p}_J(\overline{x}_J | x_J)$

sample $p_J(x_J)$

Hierarchic Sampling

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

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



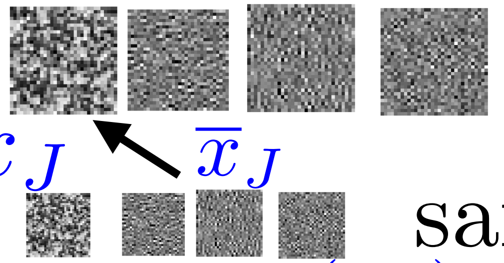
sample $\bar{p}_J(\overline{x}_J | x_J)$

sample $p_J(x_J)$

Hierarchic Sampling

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

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



x_J $\bar{x}_J$

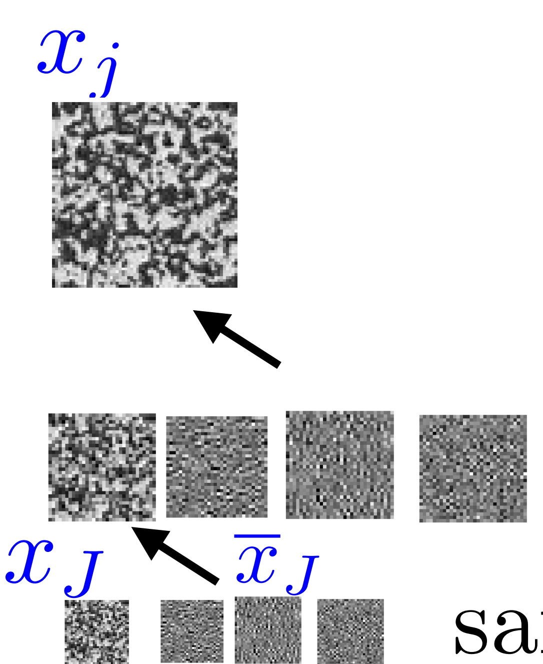
sample $\bar{p}_J(\bar{x}_J | x_J)$

sample $p_J(x_J)$

Hierarchic Sampling

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

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



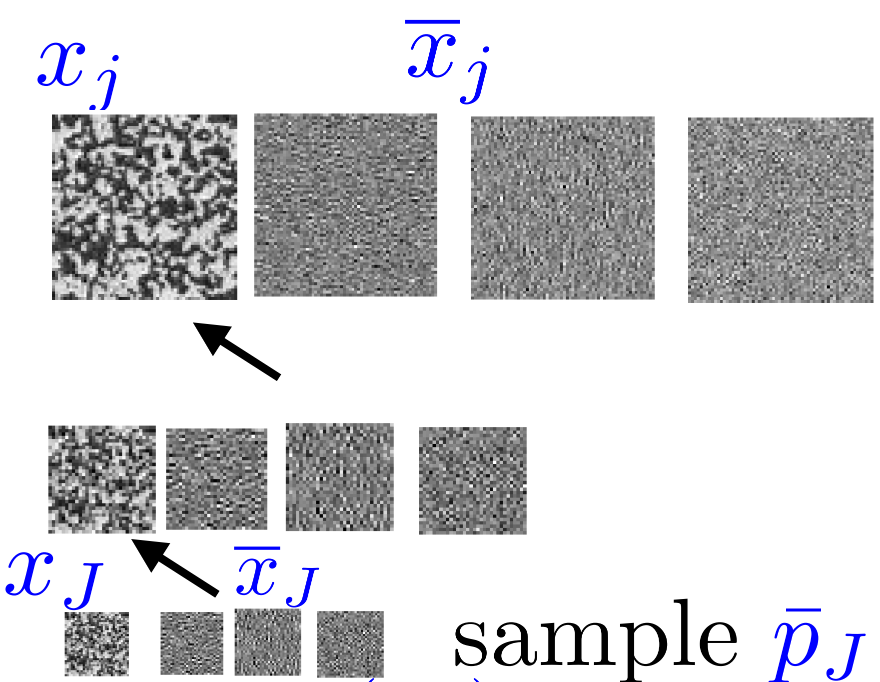
sample $\bar{p}_J(\overline{x}_J | x_J)$

sample $p_J(x_J)$

Hierarchic Sampling

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

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



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

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

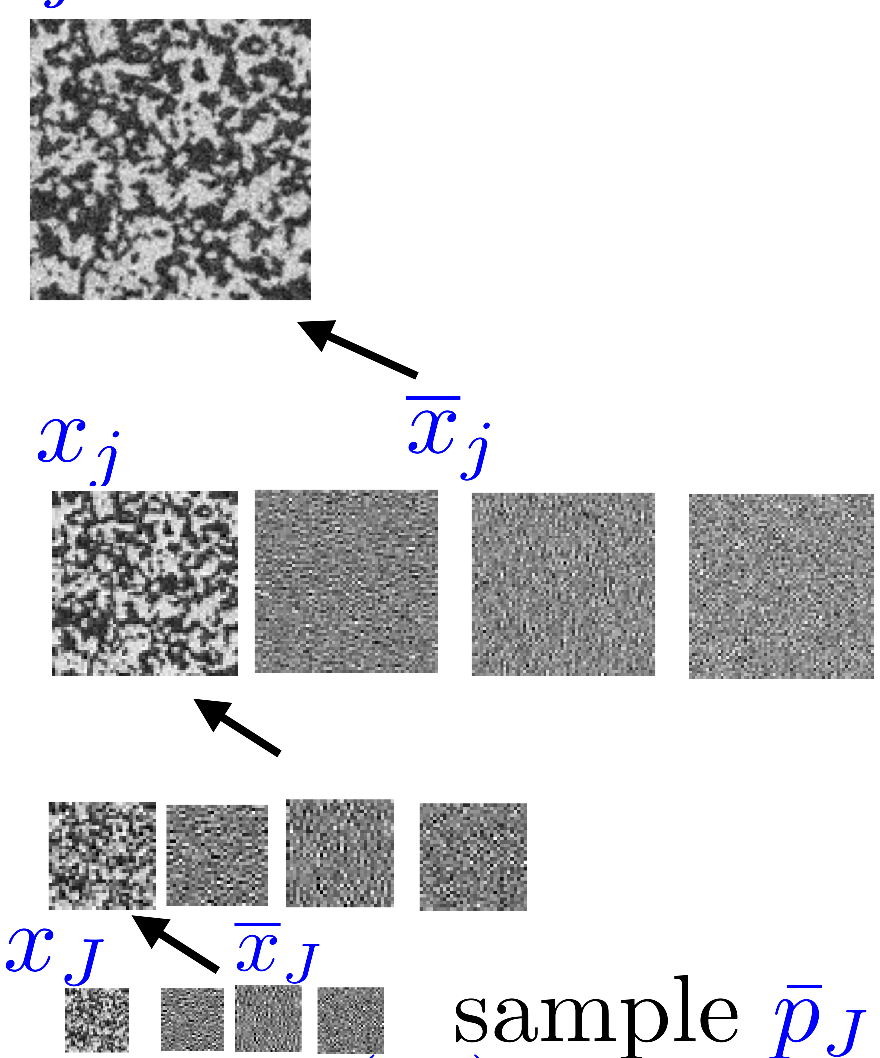
sample $p_J(x_J)$

Hierarchic Sampling

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

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

x_{j-1}



x_j $\overline{x}_j$

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

x_J $\overline{x}_J$

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

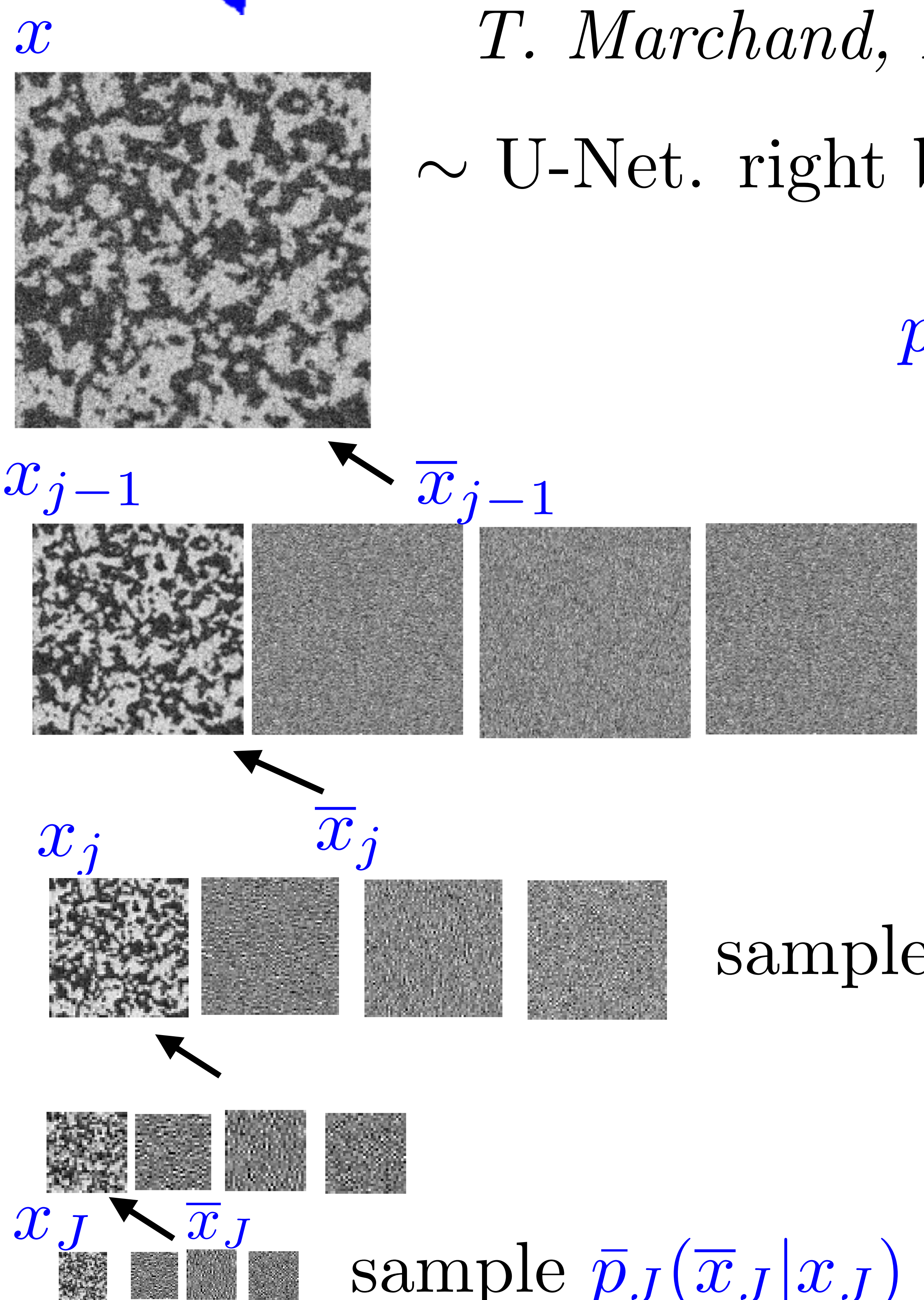
sample $p_J(x_J)$

Hierarchic Sampling

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

$\sim$ U-Net. right branch

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



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

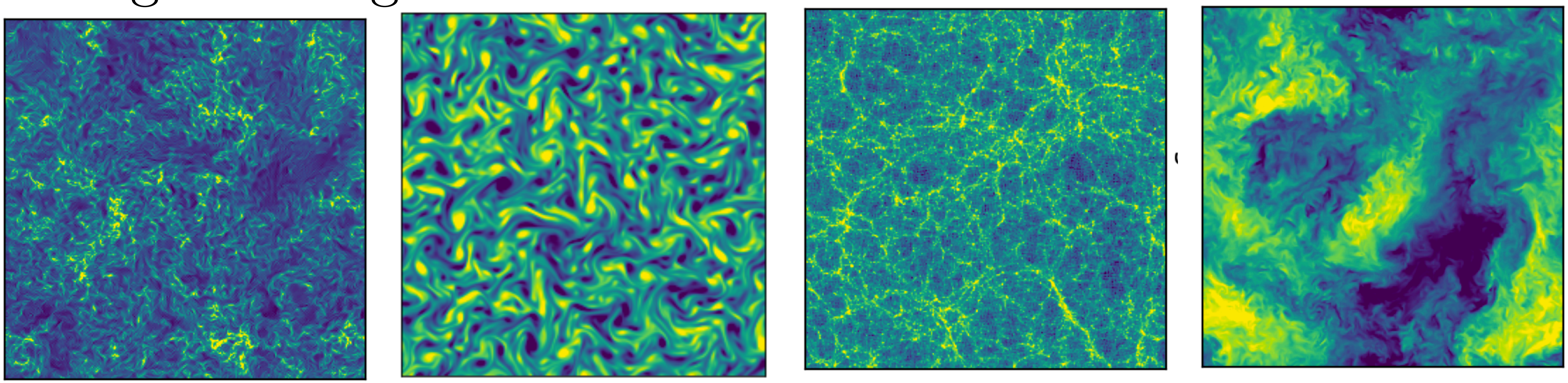
sample $\bar{p}_J(\overline{x}_J | x_J)$

sample $p_J(x_J)$

Generation from Scattering Models

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

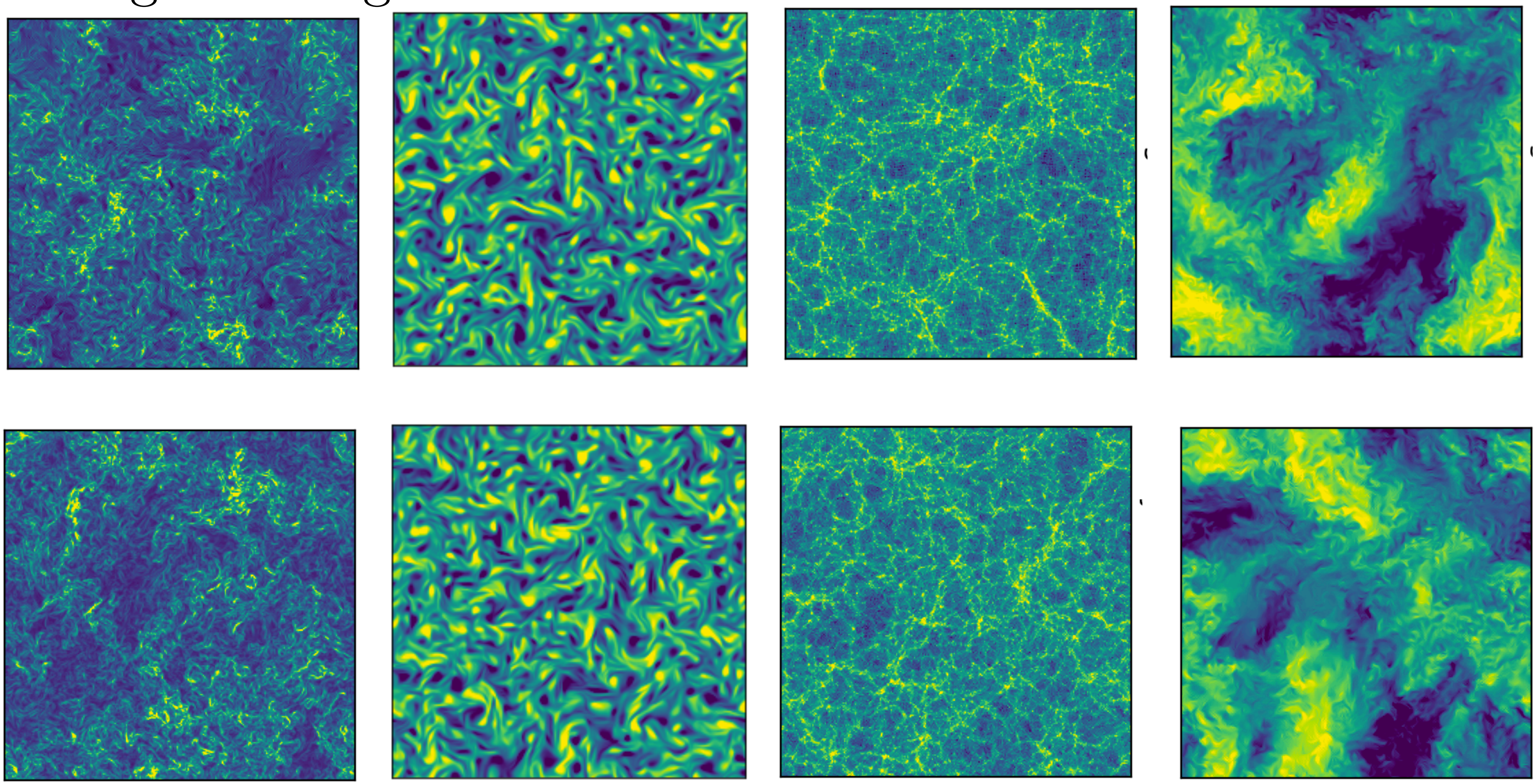
Original images of dimension $d = 5\,10^4$



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)

Conclusion

Conclusion

- Neural network score generation do generalise: they do not just memorise if the data set is large enough: very large…

- Hierarchical organisations reduce the curse of dimensionality

Conclusion

- Neural network score generation do generalise: they do not just memorise if the data set is large enough: very large…
- Hierarchical organisations reduce the curse of dimensionality

Conclusion

- Neural network score generation do generalise: they do not just memorise if the data set is large enough: very large…

- Hierarchical organisations reduce the curse of dimensionality

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