Deep Unsupervised Learning from Maximum Entropy to Deep Generative Networks

Stéphane Mallat

École Normale Supérieure
Collège de France
www.di.ens.fr/data

Deep Convolution Networks

- Spectacular supervised learning : prediction of y given data x . Classification, regression: *images, speech, natural languagee, bio-data, go*… but black boxes.

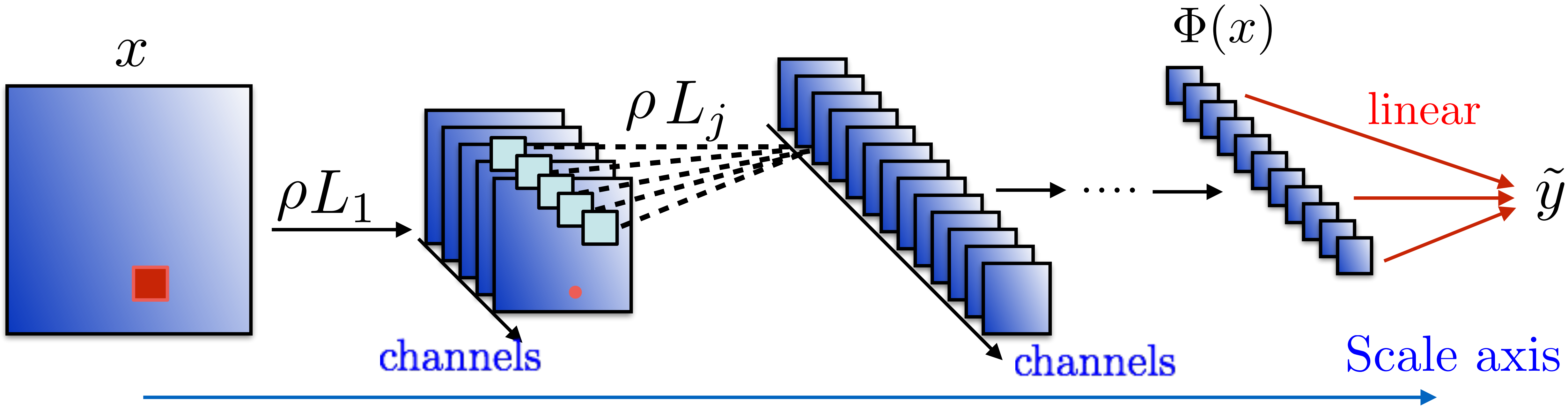
- Spectacular unsupervised learning: data models x. Generation of *textures, complex images, speech, music*…

- Good results for inverse problems and denoising: improvements of *1db* relatively to state of the art *(Unser et. al.)*.

- **Opening the black box**: powerful statistical tools.

Supervised Deep Learning

- Deep convolutional neural network to predict $y = f(x)$:



L_j: spatial convolutions and linear combination of channels

$\rho(a) = \max(a, 0)$: Relu

Supervised learning of L_j from n examples $\{x_i, y_i\}_{i \leq n}$

Exceptional results for *images, speech, language, bio-data...*

Transfer learning of $\Phi(x)$ to classify over different data bases.

Open questions:

- Why is such a filter-bank architecture effective ?
- Need to learn $\Phi(x)$ or prior information could be enough ?

Unsupervised Learning

- Estimation $\tilde{p}(x)$ of a probability density $p(x)$ for $x \in \mathbb{R}^d$ given n realizations $\{x_i\}_{i \leq n}$ of a random vector X.

- Generation of a typical realisation by sampling $\tilde{p}(x)$

- Models for all statistical applications

- If $p(x)$ is locally regular: Lipschitz

$$\mathbb{E}(\|p - \tilde{p}\|_{\mathcal{H}}) \leq \epsilon \quad \Rightarrow \quad n \geq C\,\epsilon^{-d}$$

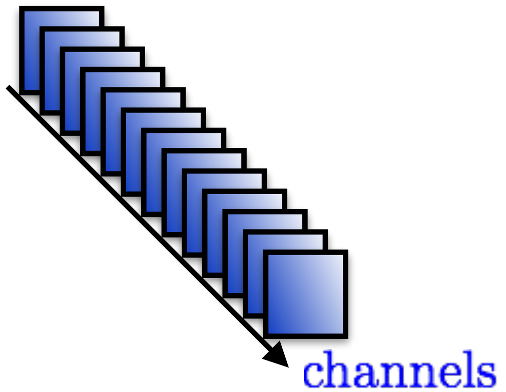
Curse of dimensionality

Problem: Find regularity properties which can break the curse of dimensionality.

Turbulence $x(u)$
$d = 10^6$

Deep Net. Models from 1 Example

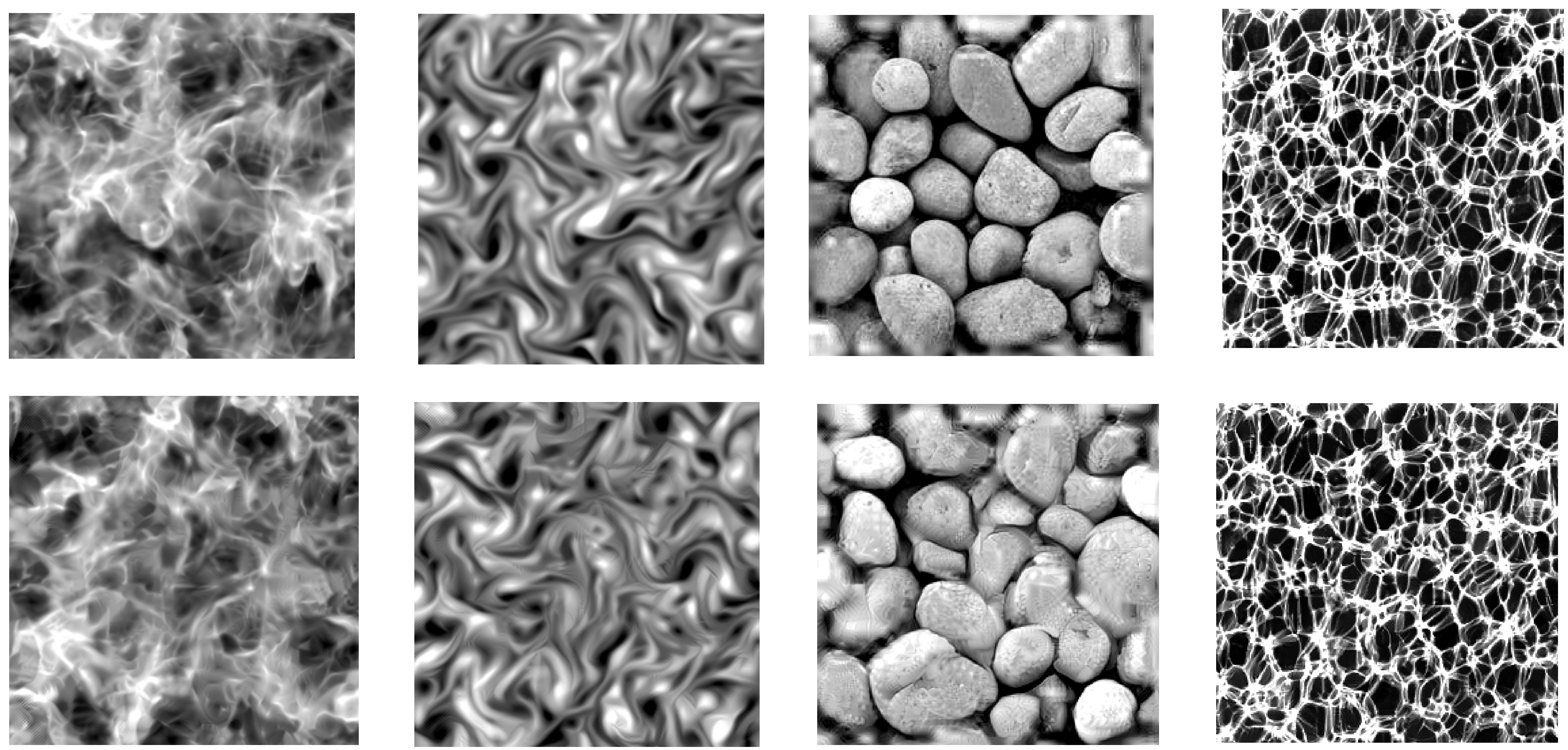
M. Bethdge et. al.



- Supervised network training (ex: on ImageNet)
- For 1 realisation x of X, compute each layer
- Compute correlation statistics of network coefficients
- Synthesize $\tilde{x}$ having similar statistics

x
$6\,10^4$ pixels

$\tilde{x}$
$2\,10^5$ correlations

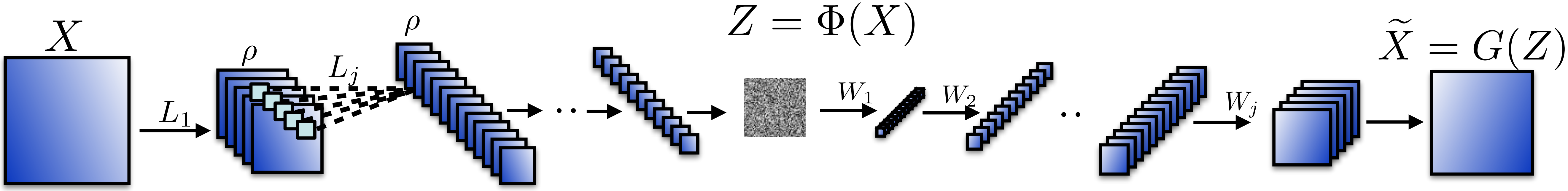


What mathematical interpretation ?

Learned Generative Networks

- Variational autoencoder: trained on n examples $\{x_i\}_{i \leq n}$

Encoder Φ Gaussian white Decoder G



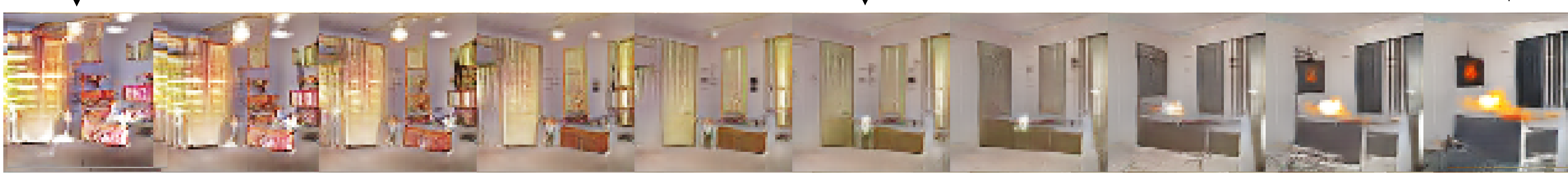
Z_1

Network trained on bedroom images:

$Z = \alpha Z_1 + (1 - \alpha) Z_2$

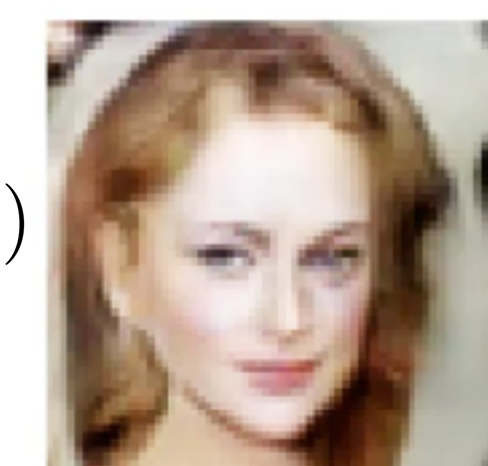
Linearization of deformations

Z_2



Network trained on faces of celebrities:

$G(Z)$



What mathematical interpretation ?

Maximum Entropy with Moments *Jaynes*

Approximation of $p(x)$ conditioned on K moments $\mathbb{E}_p(\phi_k(x))$ by $\tilde{p}$ which maximizes the entropy $H_{\tilde{p}} = -\int \tilde{p}(x) \, \log \tilde{p}(x) \, dx$

Theorem [*Gibbs distributions*] If $\tilde{p}(x)$ satisfies

$$\forall k \leq K \quad , \quad \mathbb{E}_{\tilde{p}}(\phi_k(x)) = \int_{\mathbb{R}^N} \phi_k(x) \; \tilde{p}(x) \, dx = \mathbb{E}_p(\phi_k(x))$$

and maximizes $H_{\tilde{p}} = -\int \tilde{p}(x) \, \log \tilde{p}(x) \, dx$ then

$$\tilde{p}(x) = \mathcal{Z}^{-1} \exp\Big(-\sum_{k=1}^{K} \beta_k \, \phi_k(x) \Big) \; .$$

How to choose the ϕ_k ?

Can we avoid computing the β_k ?

Which moments ?

- **Linearization:**

$$-\log \tilde{p}(x) = \log \mathcal{Z} + \sum_{k=1}^{K} \beta_k \, \phi_k(x) \approx -\log p(x)$$

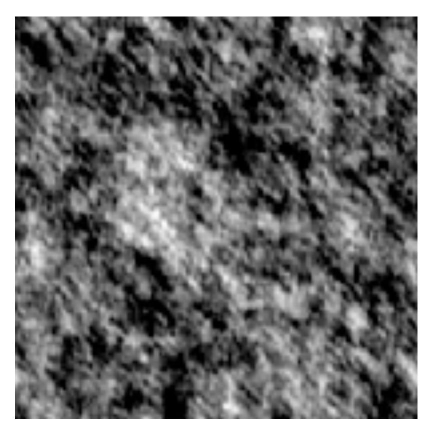
$\phi_k(x)$ specified from "priors" on the regularity of $p(x)$

- Stationarity: $p(x)$ invariant to translations
  obtained with $\phi_k(x)$ also invariant

  Other priors: regularity to deformations, ...

- Quadratic: $\phi_k(x) = \sum_u x(u)\, x(u - \tau_k)$ $\Rightarrow$ $\tilde{p}(x)$ is Gaussian

Gaussian



Kolmogorov

- Higher order moments: large variance, sensitive to outliers.
  Failure!

Prior: Scale Separation

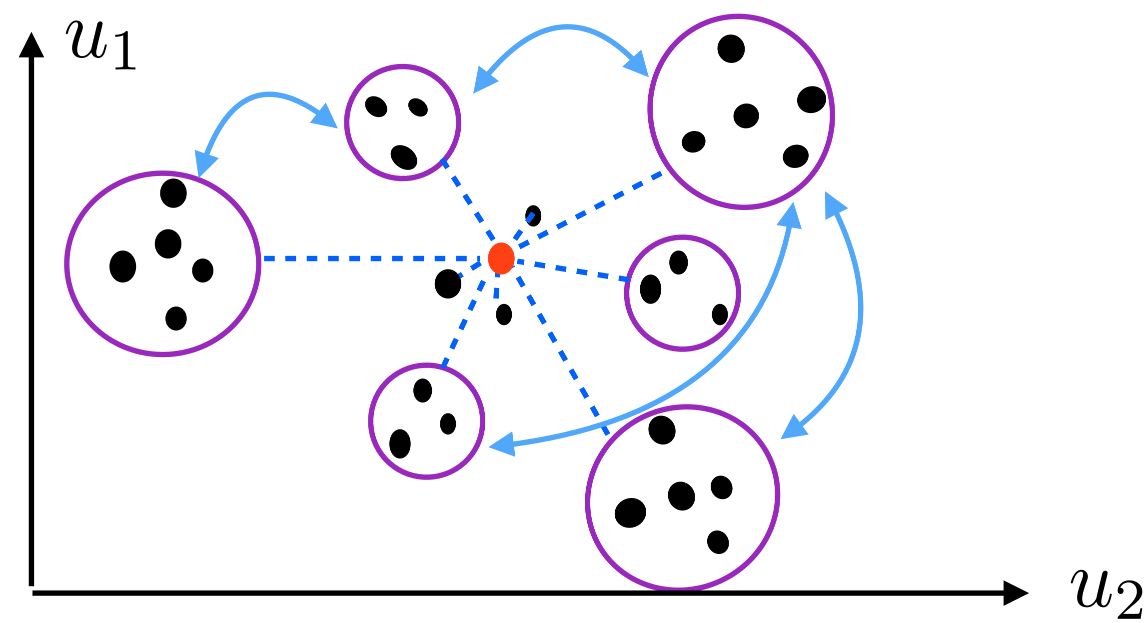
- Architecture of complexity: hierarchical *Herbert Simons*

Prior: Scale Separation

- Architecture of complexity: hierarchical *Herbert Simons*

Interactions de d variables $x(u)$: pixels, particules, agents...

Interactions across scales



Multiscale regroupement of interactions of d variables

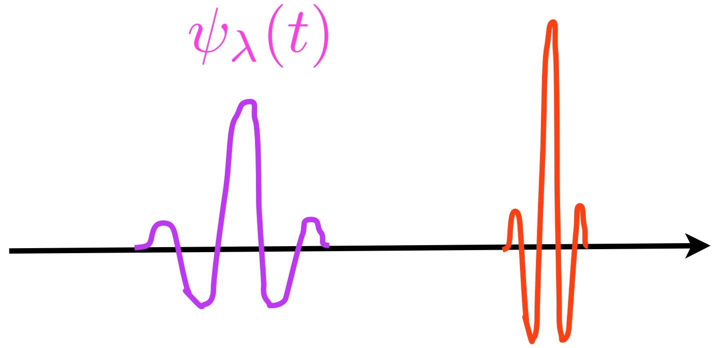
into interactions of $O(\log d)$ groups of variables,

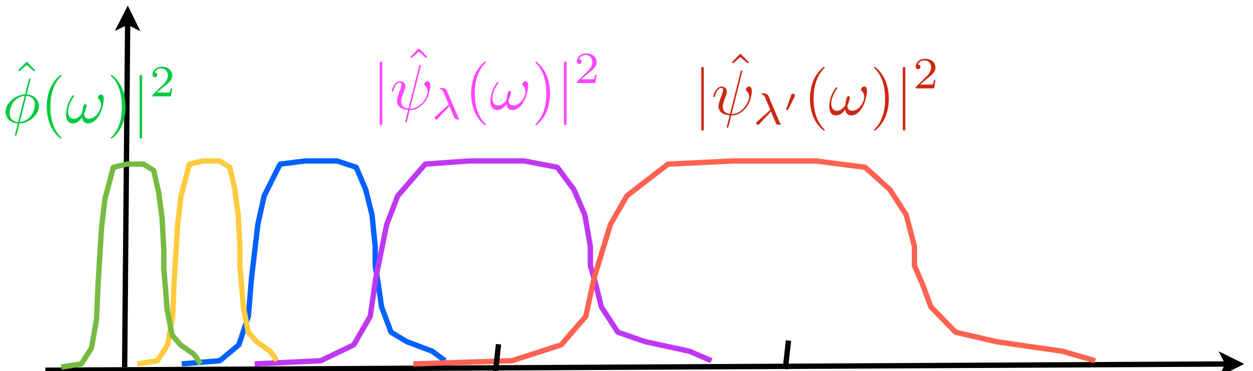
Scale separation $\Rightarrow$ wavelet transforms, filter banks

A path to Deep Nets.

Multiscale Wavelet Transform

- Dilated wavelets: $\psi_\lambda(t) = 2^{-j/Q}\,\psi(2^{-j/Q}t)$ with $\lambda = 2^{-j/Q}$





Q-constant band-pass filters $\hat{\psi}_\lambda$

$$x \star \psi_\lambda(t) = \int x(u)\,\psi_\lambda(t-u)\,du \;\Rightarrow\; \widehat{x \star \psi_\lambda}(\omega) = \widehat{x}(\omega)\,\widehat{\psi}_\lambda(\omega)$$

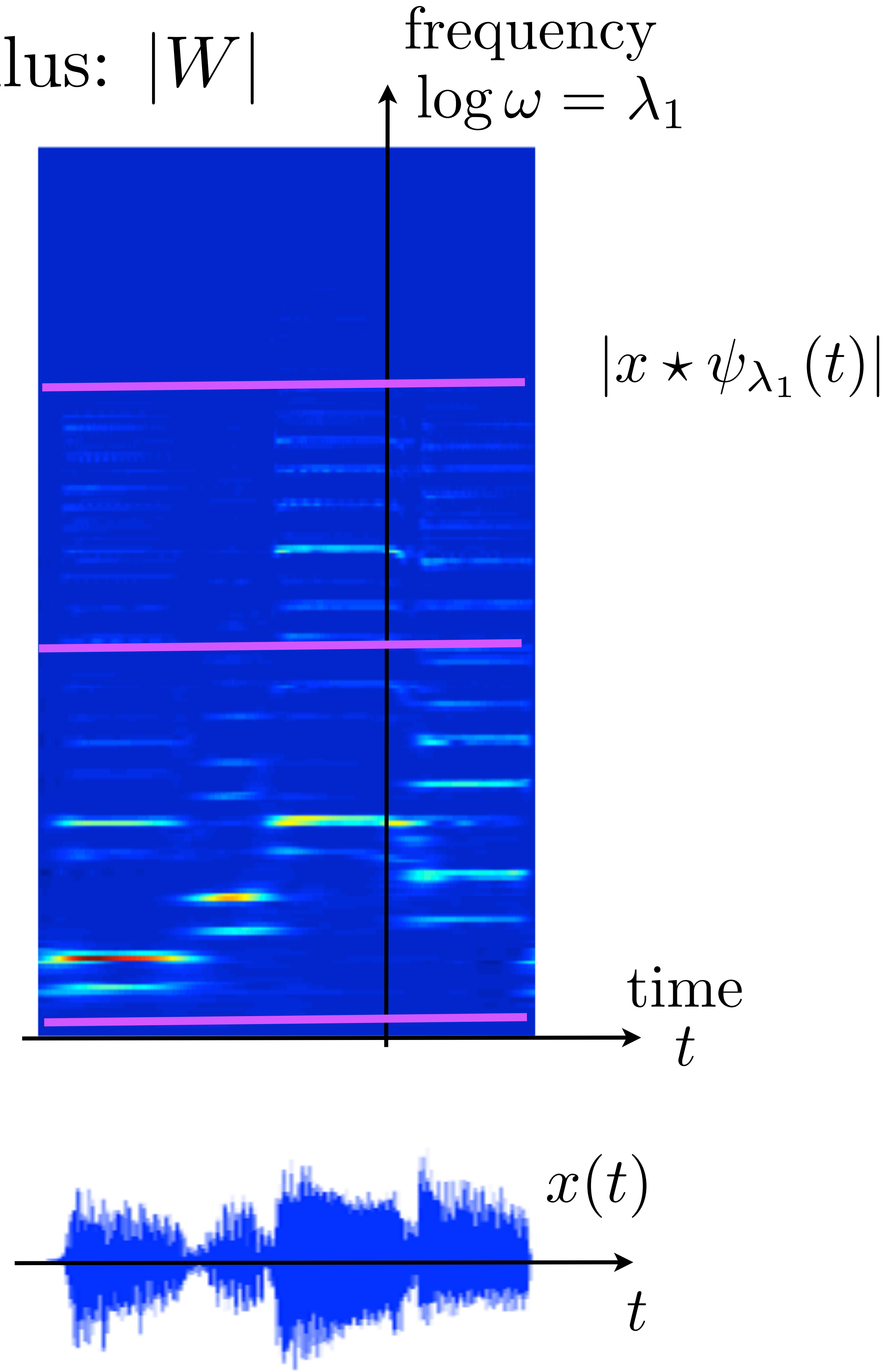
- Wavelet transform: $Wx = \begin{pmatrix} x \star \phi_{2^J}(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{\lambda \leq 2^J}$: average ; : higher frequencies

Preserves norm: $\|Wx\|^2 = \|x\|^2$.

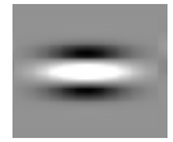
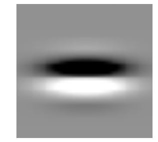
Wavelets are stable to deformations

Wavelet Spectrogram

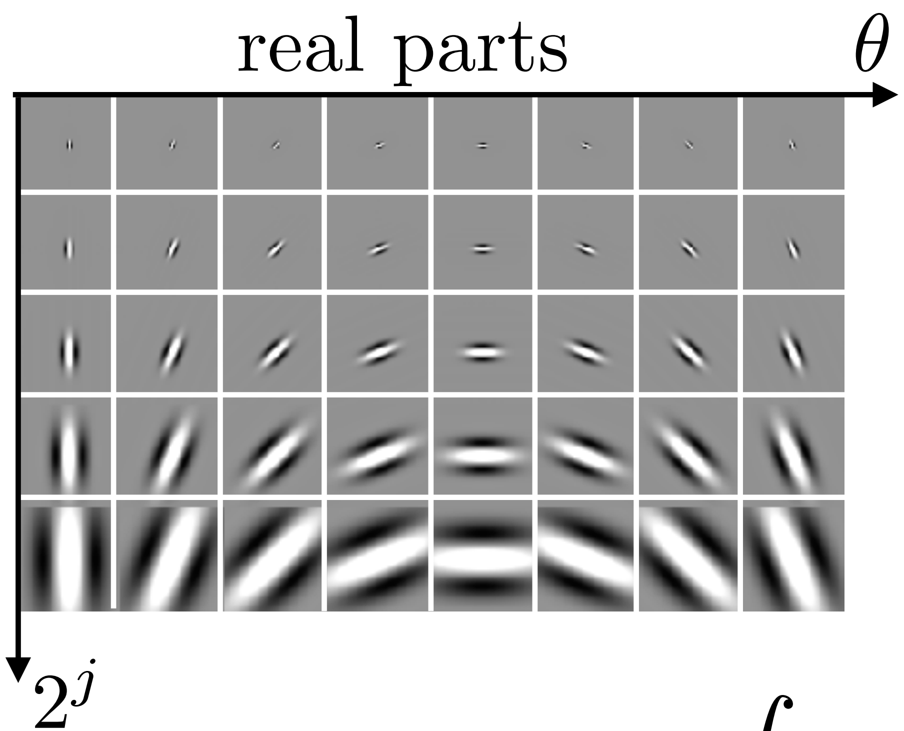
Wavelet transform modulus: $|W|$

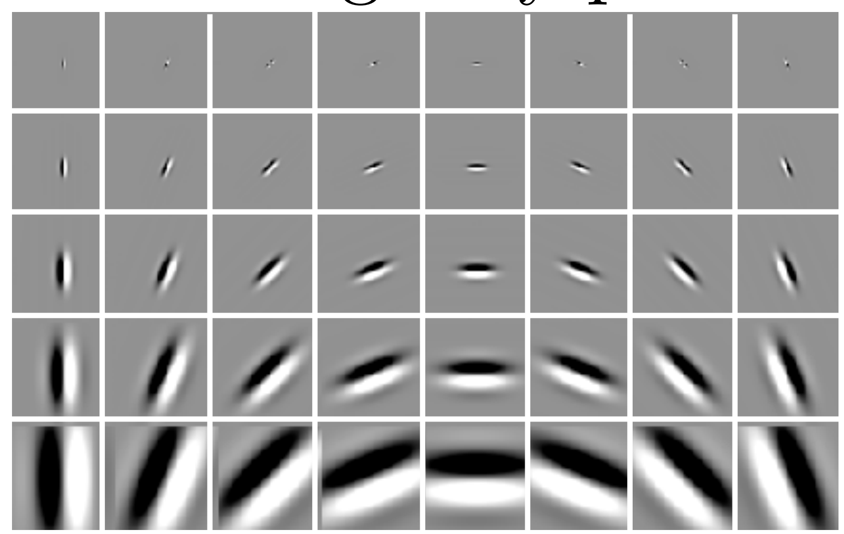


Scale separation with Wavelets

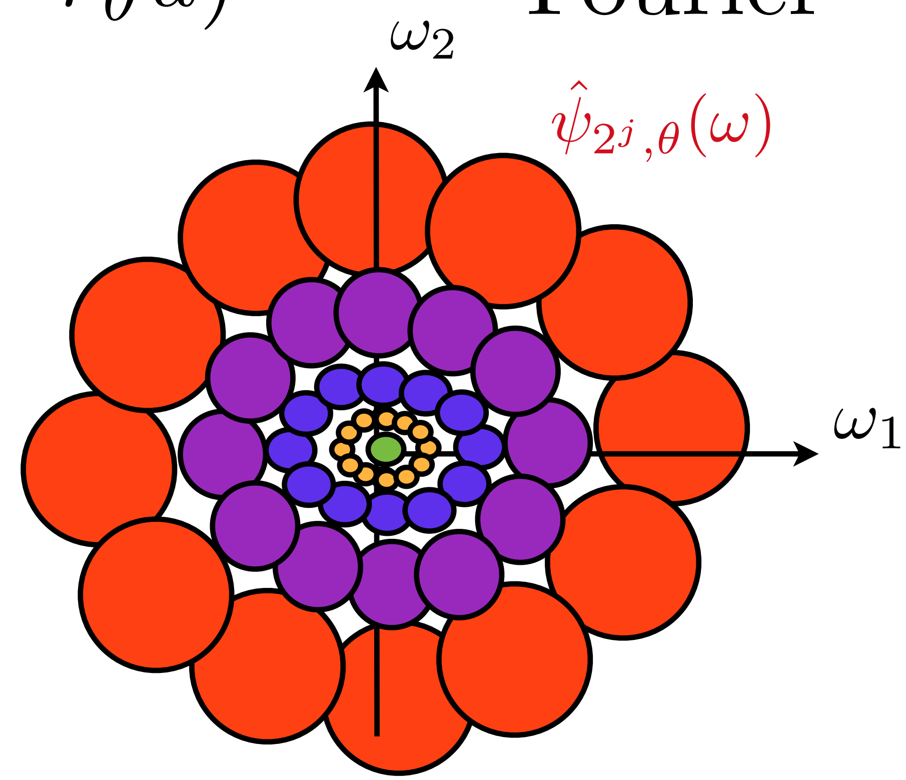
- Wavelet filter $\psi(u)$:  $+\, i$ 

rotated and dilated: $\psi_{2^j,\theta}(u) = 2^{-j}\,\psi(2^{-j} r_\theta u)$





Fourier



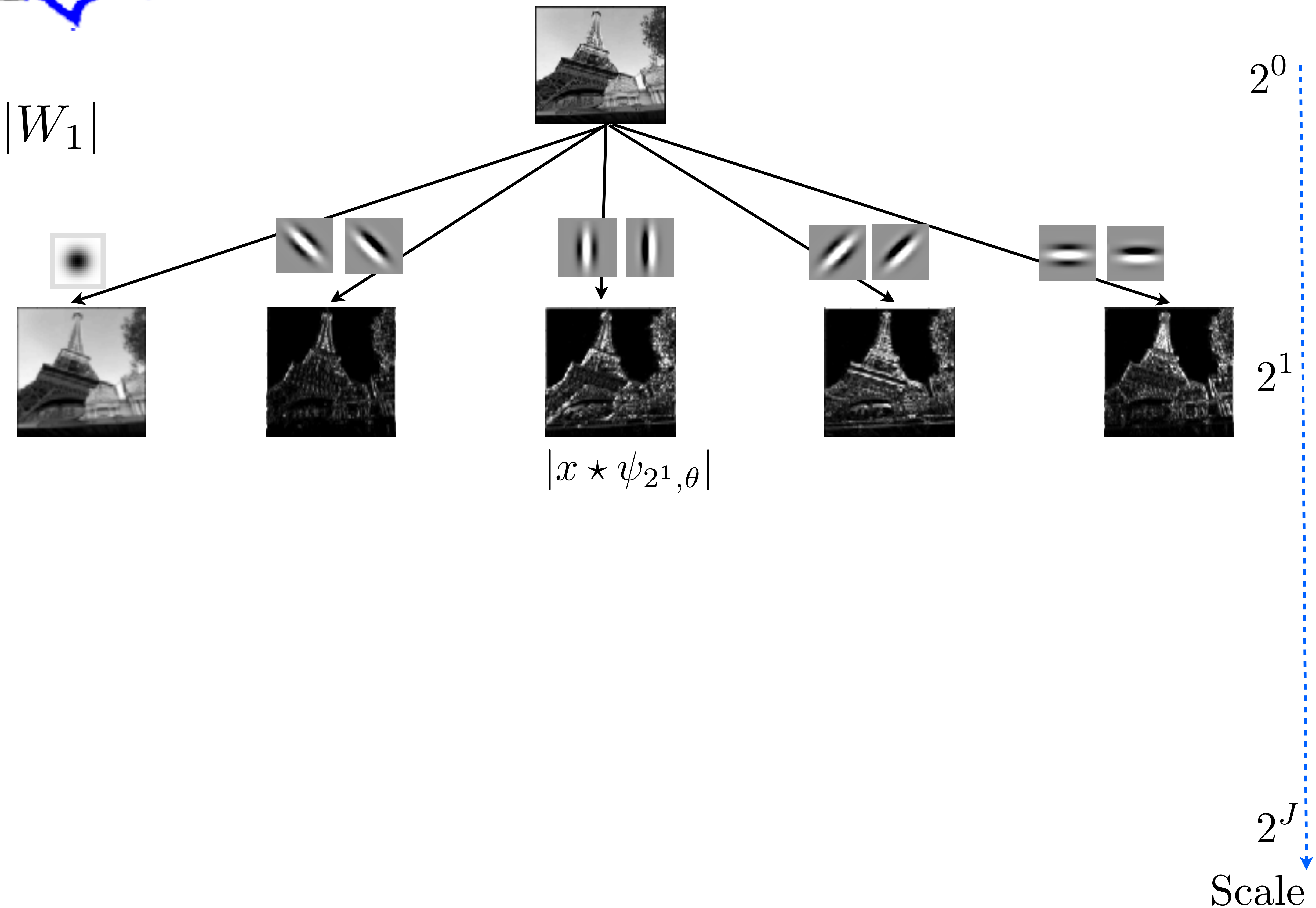
$$x \star \psi_{2^j,\theta}(u) = \int x(v)\,\psi_{2^j,\theta}(u-v)\,dv \quad \Rightarrow \quad \hat{x}(\omega)\,\hat{\psi}_{2^j,\theta}(\omega)$$

- Wavelet transform: $Wx = \begin{pmatrix} x \star \phi_{2^J}(u) \\ x \star \psi_{2^j,\theta}(u) \end{pmatrix}_{j \le J, \theta}$: average ; : higher frequencies

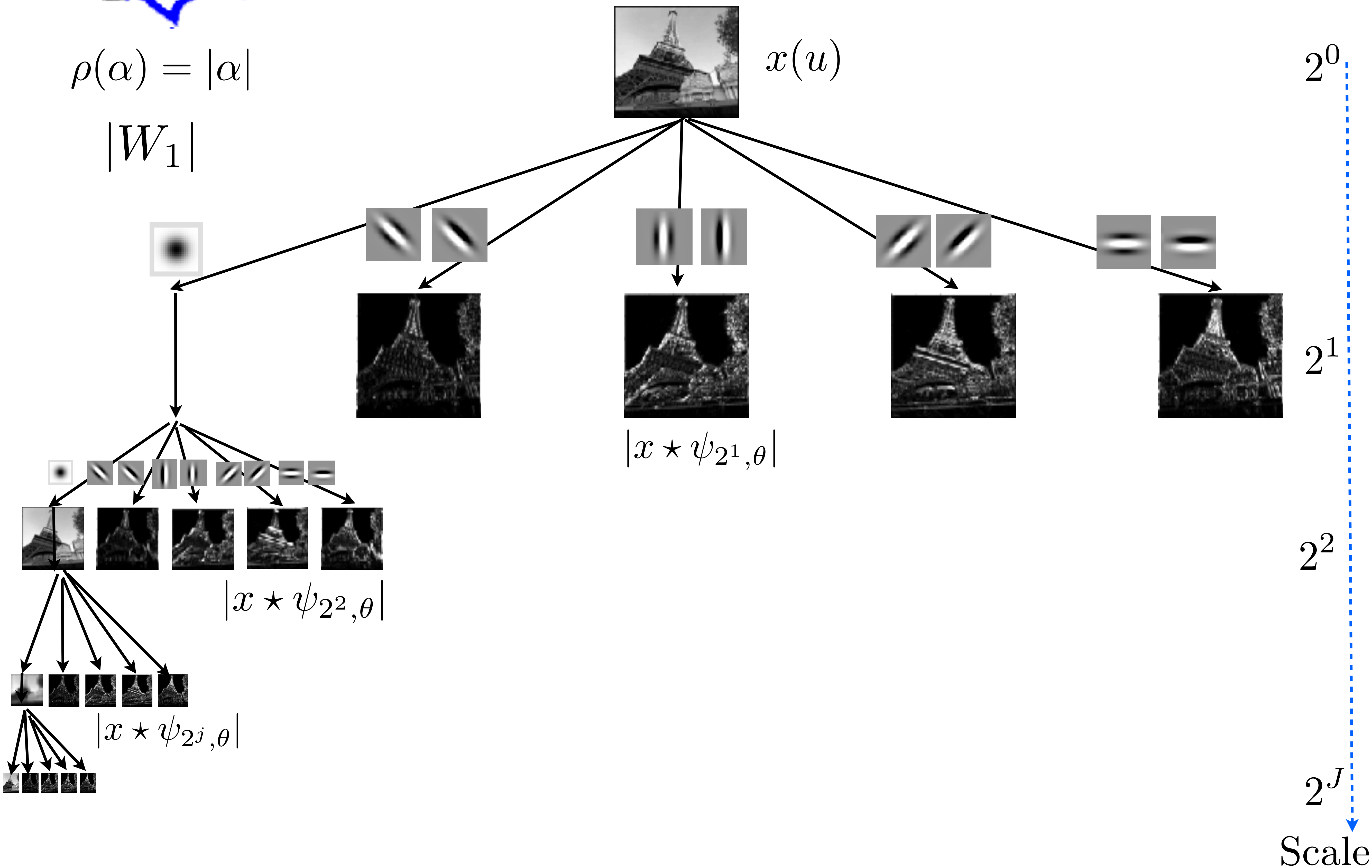
Preserves norm: $\|Wx\|^2 = \|x\|^2$.

Wavelets are stable to deformations

Fast Wavelet Filter Bank



Wavelet Filter Bank

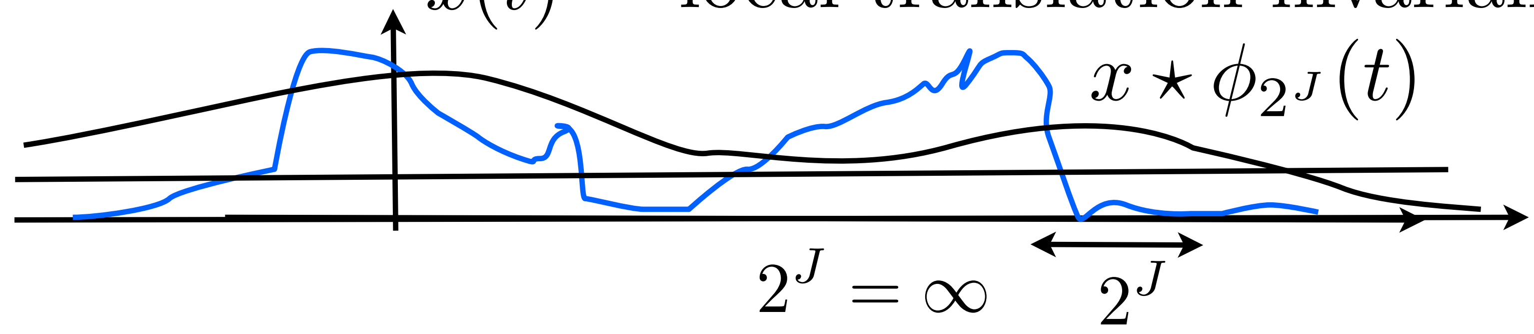


Wavelet Translation Invariance

First wavelet transform

$$W_1 x = \begin{pmatrix} x \star \phi_{2^J} \\ x \star \psi_{\lambda_1} \end{pmatrix}_{\lambda_1}$$

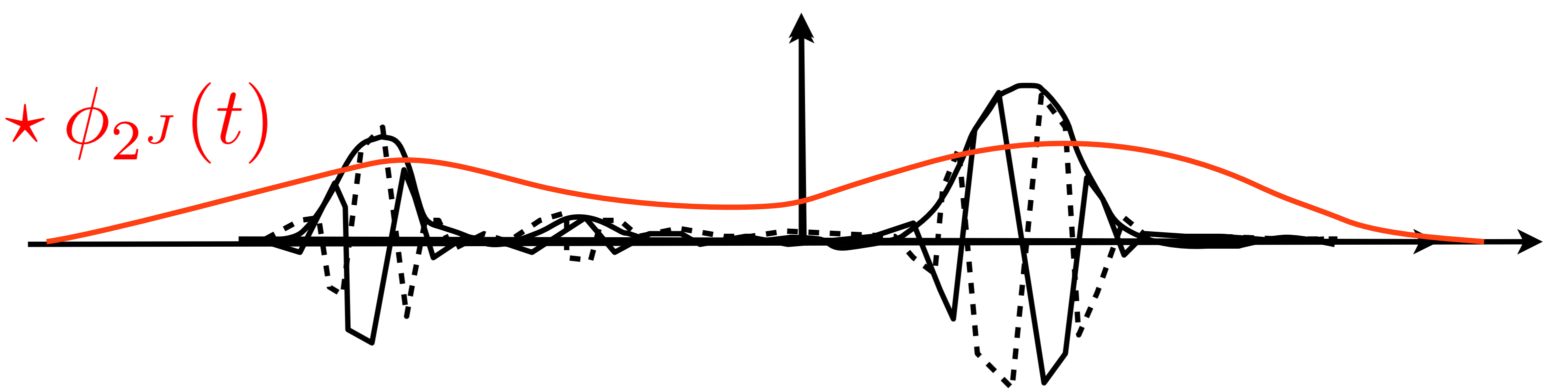
$$|W_1| x = \begin{pmatrix} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \end{pmatrix}_{\lambda_1}$$



Lost high frequencies: $x \star \psi_{\lambda_1}(t)$

Eliminate the phase: $|x \star \psi_{\lambda_1}(t)|$ non-linearity

Invariant: $|x \star \psi_{\lambda_1}| \star \phi_{2^J}(t)$

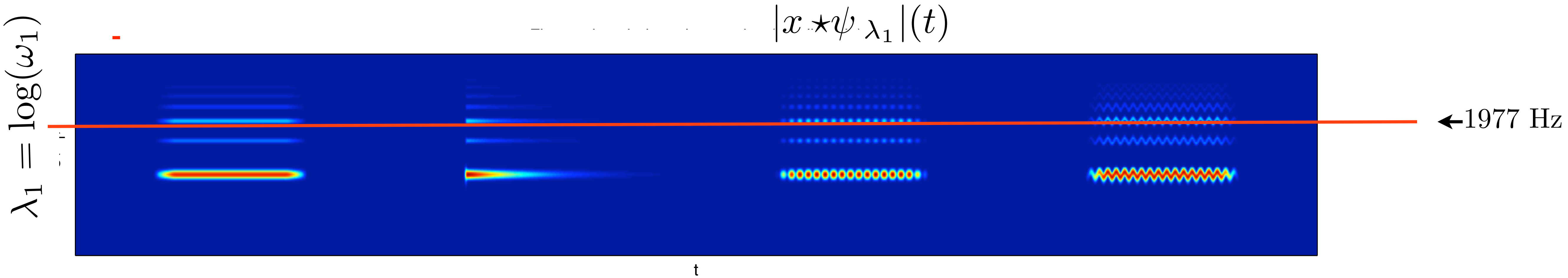


Need to recover lost high frequencies: $|x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)$

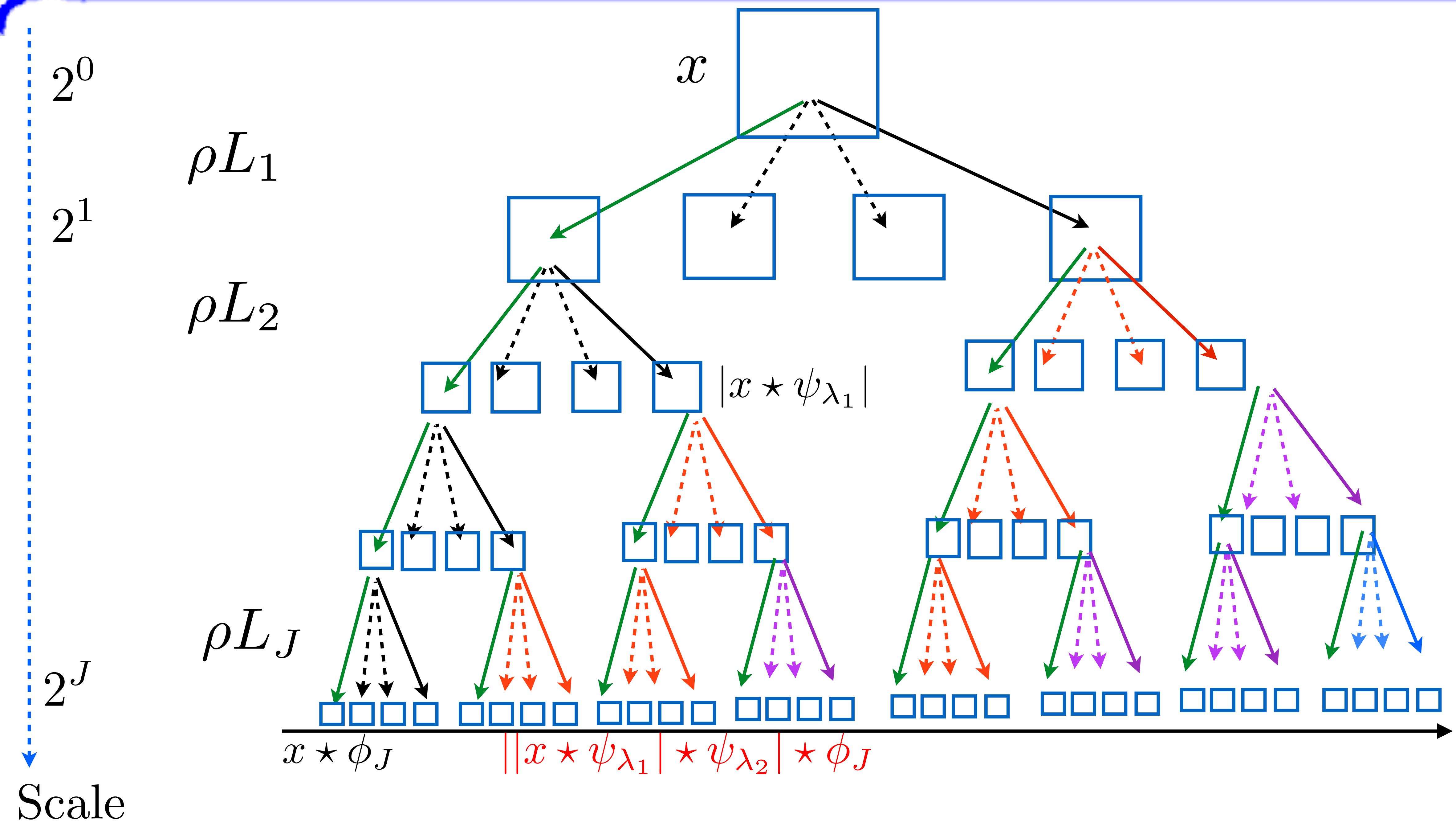
$$\Rightarrow \text{ wavelet transform: } |W_2|\ |x \star \psi_{\lambda_1}| = \begin{pmatrix} |x \star \psi_{\lambda_1}| \star \phi_{2^J}(t) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)| \end{pmatrix}_{\lambda_2}$$

Amplitude Modulation

Harmonic sound: $x(t) = a(t)\, e \star h(t)$ with varying $a(t)$



Wavelet Scattering Network



$$S_J = \rho W_1 \;\; \rho W_2 \;\; \cdots \;\; \rho W_J$$

$$\rho(\alpha) = |\alpha| \qquad S_J x = \left\{ |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \star ...| \star \psi_{\lambda_m}| \star \phi_J \right\}_{\lambda_k}$$

Convolutional tree: no combination along channels

Scattering Network Properties

$$S_J x = \begin{pmatrix} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \star \phi_{2^J} \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J} \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J} \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots} = \dots |W_3|\,|W_2|\,|W_1|\,x$$

$$\|W_k x\| = \|x\| \quad \Rightarrow \quad \||W_k x| - |W_k x'|\| \le \|x - x'\|$$

Theorem: *For appropriate wavelets, a scattering is*

contractive $\|S_J x - S_J y\| \le \|x - y\|$ ($\mathbf{L^2}$ *stability*)

translations invariance and deformation stability:

if $D_\tau x(u) = x(u - \tau(u))$ *then*

$$\lim_{J \to \infty} \|S_J D_\tau x - S_J x\| \le C\, \|\nabla \tau\|_\infty\, \|x\|$$

Unsupervised Learning

Estimate the distribution $p(x)$ of a stationary X

Scattering transform of $X(u)$ up to order 2:

$$S_J(X) = \begin{pmatrix} X \star \phi_J \\ |X \star \psi_{\lambda_1}| \star \phi_J \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J \end{pmatrix}_{\lambda_1, \lambda_2}$$

$$\Big\downarrow \; 2^J \to \infty$$

Concentration
with ergodicity/decorrelation conditions

$$\mathbb{E}(S(X)) = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(|X \star \psi_{\lambda_1}|) \\ \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \end{pmatrix}_{\lambda_1, \lambda_2}$$

concentration towards
: scattering moments.
low-order

How to avoid computing Lagrange multipliers of max entropy ?

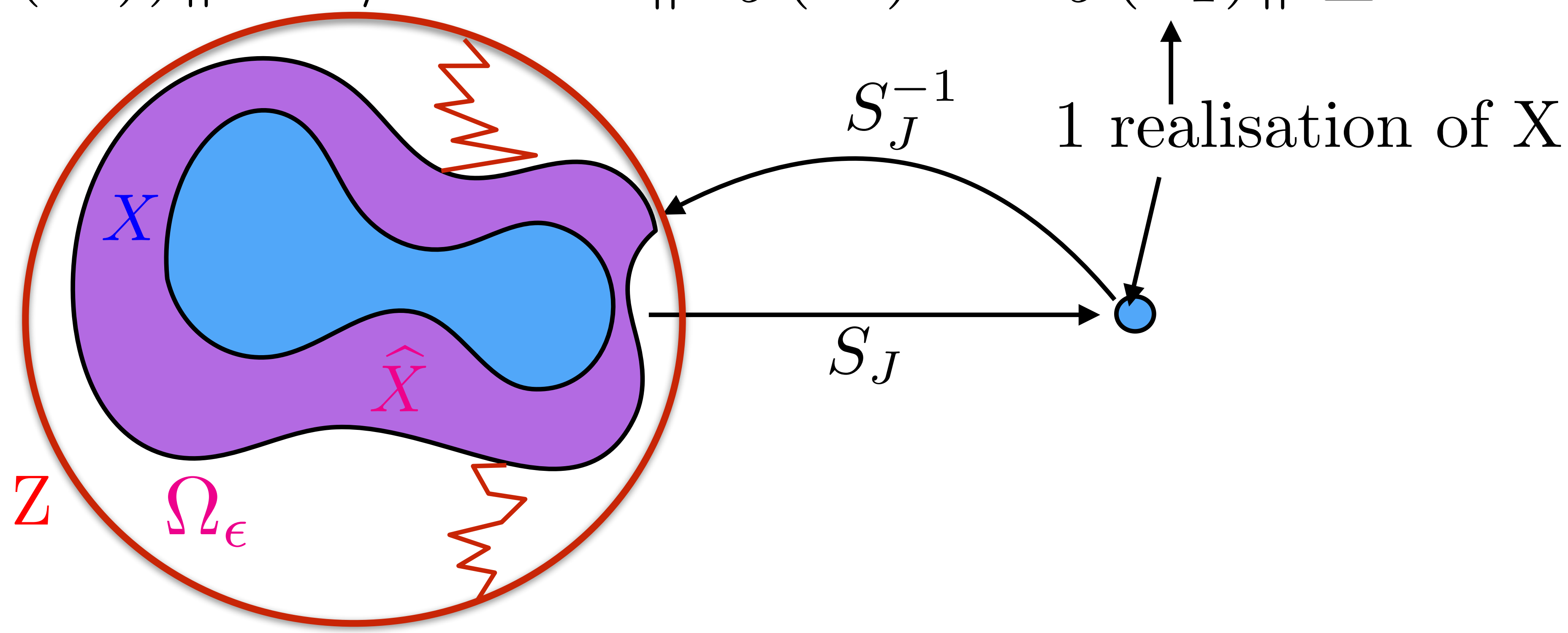
Microcanonical Models: Stat. Phys.

Joan Bruna

Maximum scale 2^J = signal size

concentration: with high probability

$$\|S_J(X) - \mathbb{E}(S(X))\| \leq \epsilon/2 \;\; \Rightarrow \;\; \|S_J(X) - S_J(x_1)\| \leq \epsilon$$



$\Omega_\epsilon = \{x \; : \; \|S_J(x) - S_J(x_1)\| \leq \epsilon\}$: microcanonial ensemble.

$\hat{p}(x)$: uniform density over the microcanoniccal ensemble

Max entropy inversion of S_J: micro canonical model $\hat{\hat{X}}$

- Sampling: initialize x with Gaussian white noise Z

  Minimize $\|S_J(x) - S_J(x_1)\|^2$ by stochastic gradient descent

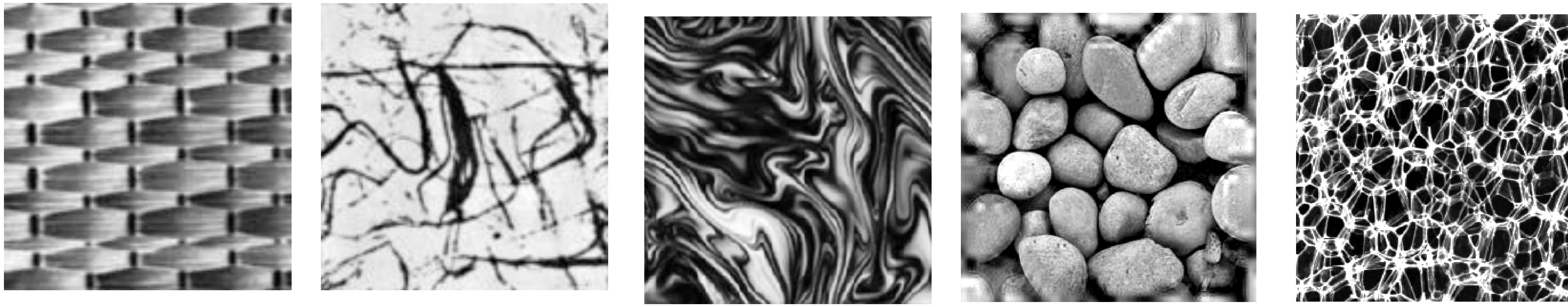
Texture Reconstructions

Joan Bruna

Texture of d pixels

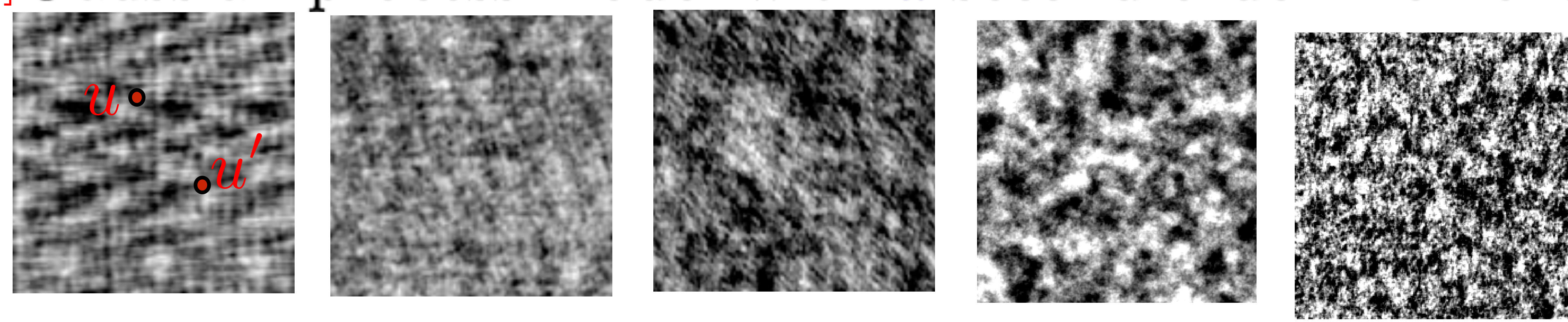
Turbulence 2D

$d = 6\,10^4$



$\mathbb{E}[X(u)\,X(u')]$ Gaussian process model with d second order moments

$d' = 6\,10^4$



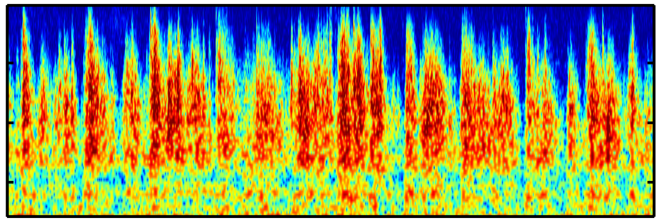
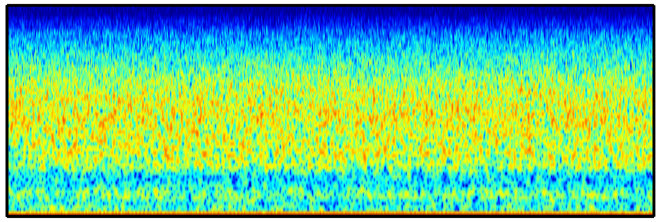
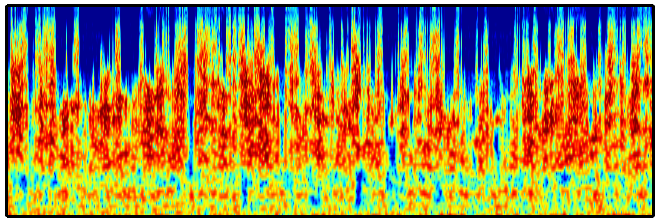
From $O(\log^2 d)$ 2nd order scattering coefficients

$d' = 6\,10^2$

What is missing ?

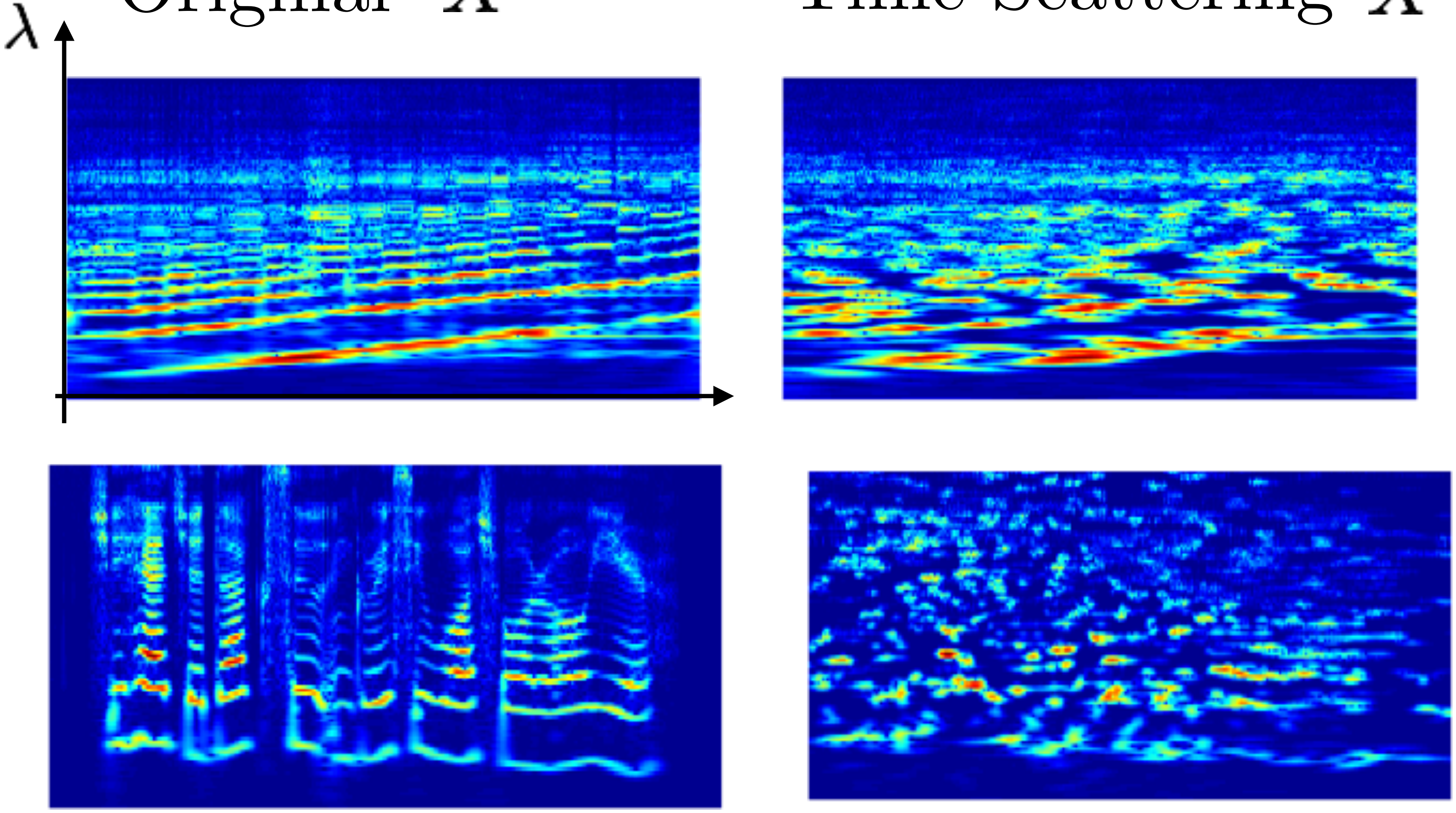
Representation of Audio Textures

Joan Bruna

	Original	Gaussian in time	Scattering Order 2
Paper			
Cocktail Party			

Failures of Audio Synthesis

Original X Time Scattering $\tilde{X}$

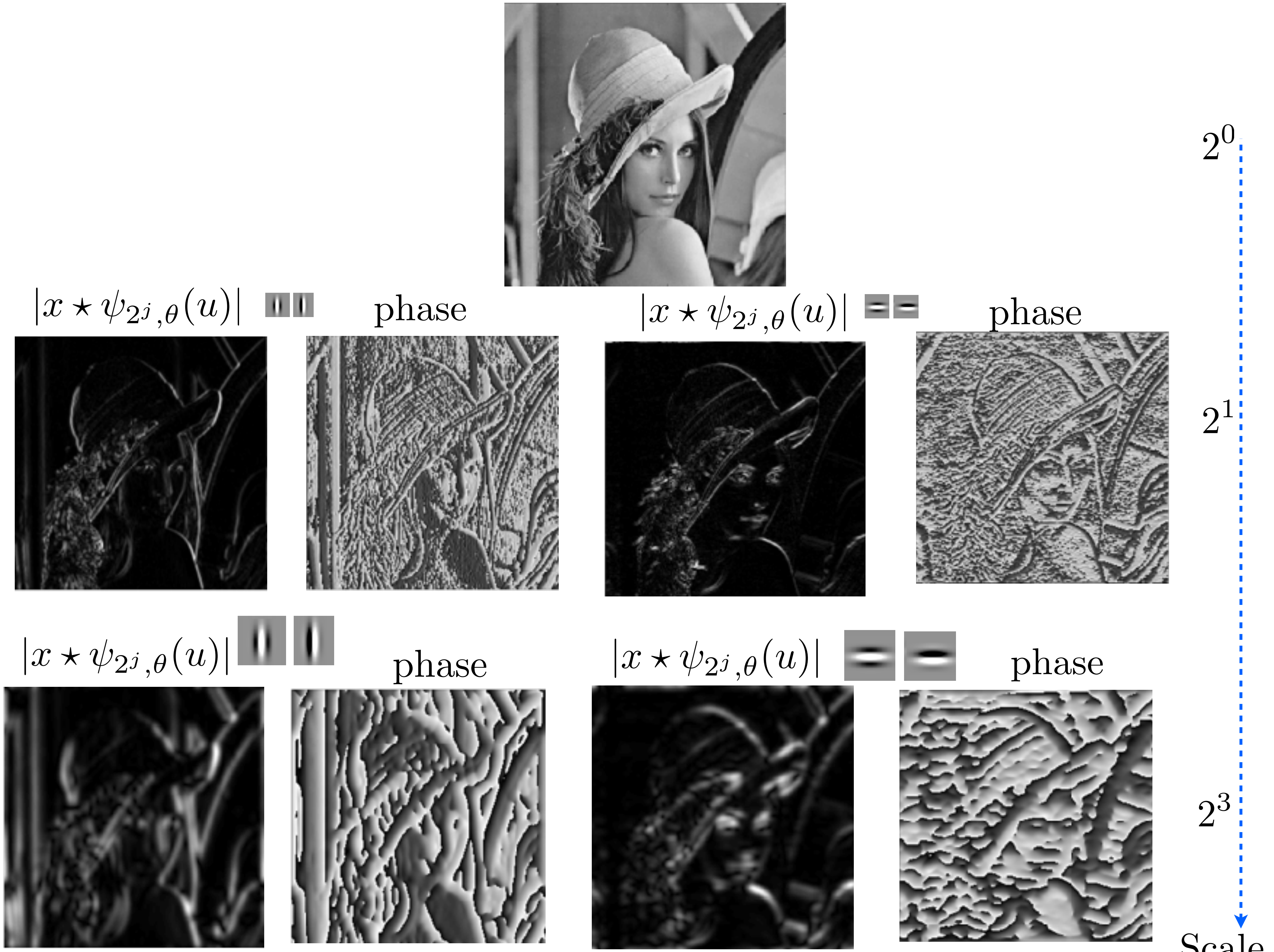


Typical of $\tilde{X}$ is not typical of X

- Missing frequency connections $\Rightarrow$ misalignments

How to connect frequencies ?

Using Complex Phase



Lines of Constant Phase

- Lines of constant phase specify the geometry: edge detection

Phase Harmonics

- Filters with a phase shift α *Sixin Zhang*

$$\psi_{j,\theta,\alpha} = \mathrm{Real}(e^{i\alpha}\,\psi_{j,\theta})$$

Rectification:

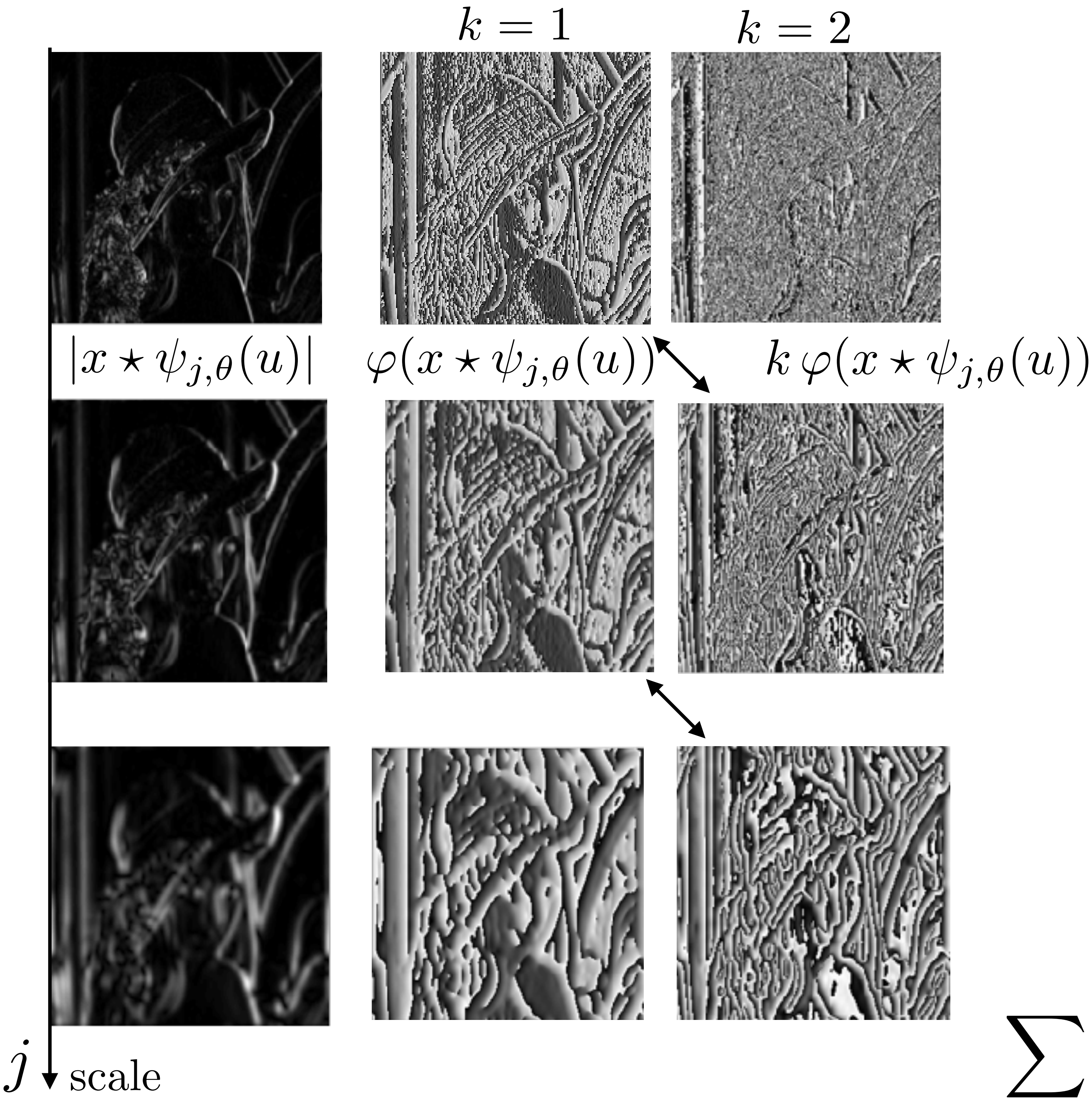
$$x(u,j,\theta,\alpha) = \rho\Big(x \star \psi_{j,\theta,\alpha}(u)\Big) \ \text{ with } \rho(a) = \max(a,0)$$

Theorem : Fourier transform along the phase α:

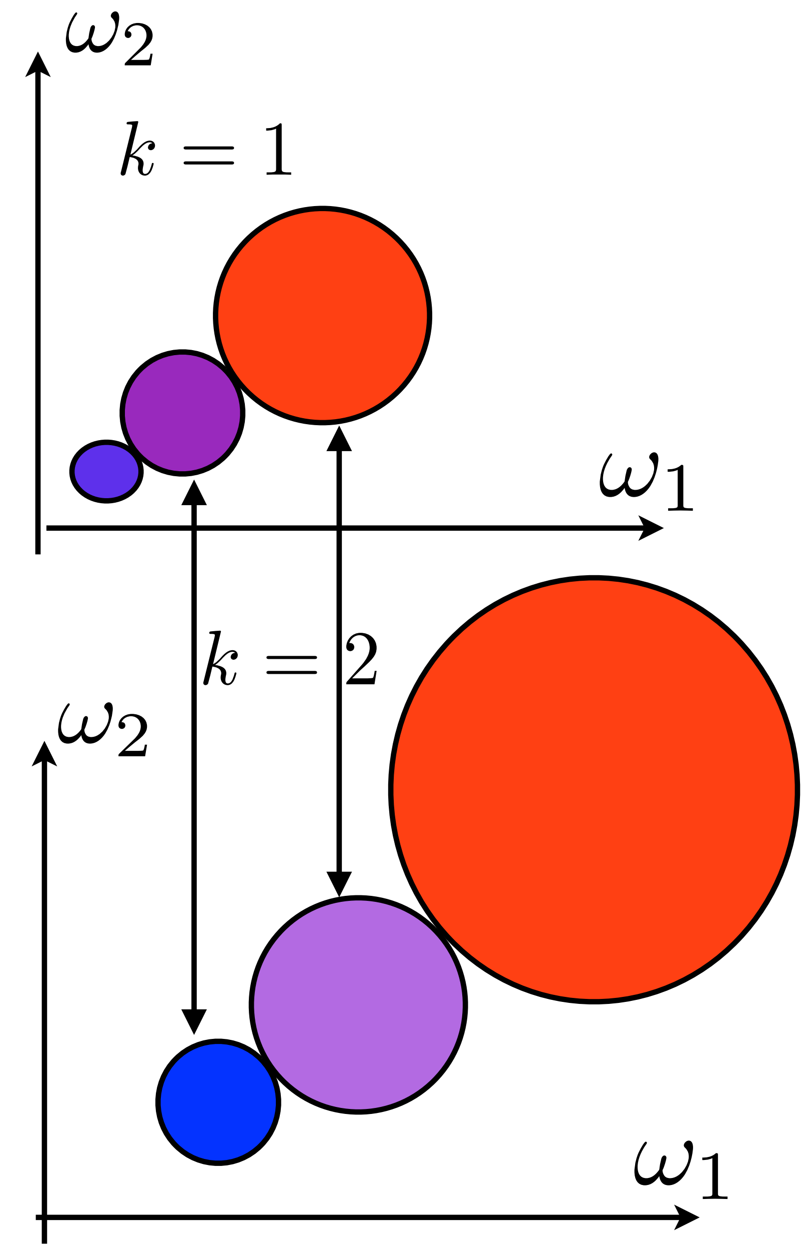
$$\widehat{x}(u,j,\theta,k) = c_k\,|x \star \psi_{j,\theta}(u)|\,e^{i\,k\,\varphi(x\star\psi_{j,\theta}(u))}$$

Linear combination across network channels creates harmonics of the phase $\Rightarrow$ connections across scales/frequencies.

Scale Connections with Harmonics



Frequency domain



Correlations:

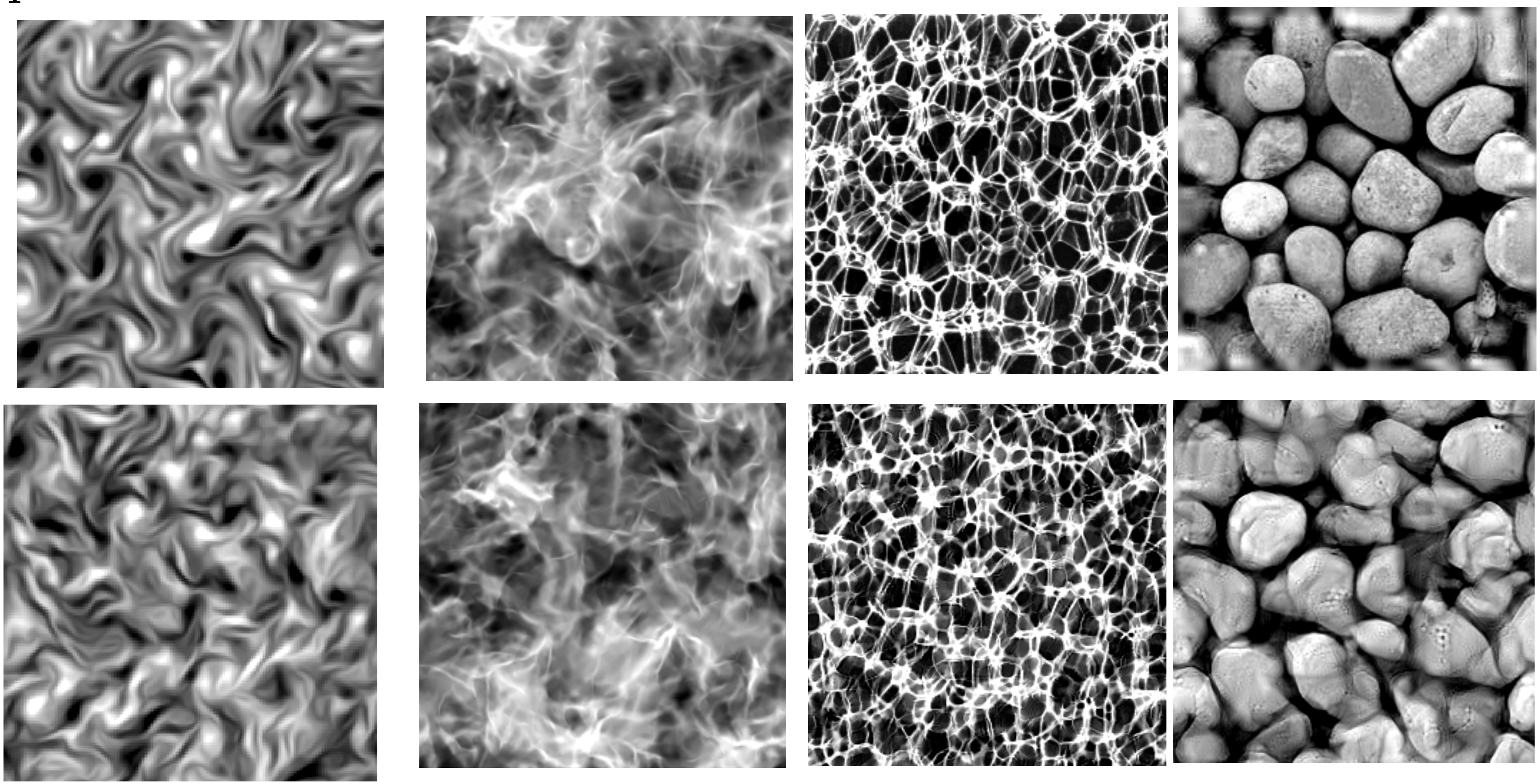
$$\sum_u \hat{x}(u, j, \theta, k)\, \hat{x}(u, j+1, 2k)$$

Synthesis with Scale Connections

$6\,10^4$ pixels

x

$\tilde{x}$



$3\,10^3$ correlations

Same quality as with learned Deep networks
with much less moments

II- Adversarial Network Generation

- What if X is not stationary ?

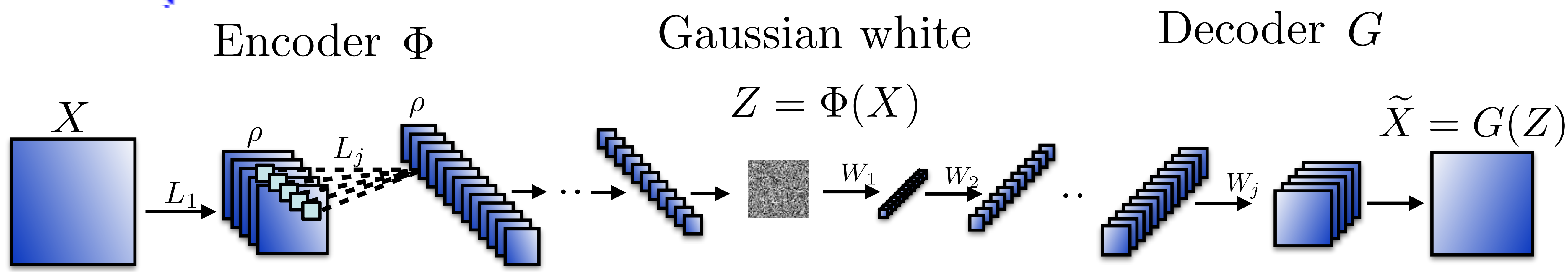
  Estimate $\widehat{X}$ from many realisations $\{x_i\}_i$ of X

- Spectacular results with a jungle of convolutional networks: GAN's, Autoencoders, Recurrent Neural Nets, WaveNets...

- Complex training with dual networks and little theory.

- Can we simplify algorithms and relate it to existing maths ?

Variational Autoencoders



- The encoder Φ must produce a nearly white noise $Z = \Phi(X)$

  variational cost: $KL(p_\Phi(z/x)||\mathcal{N}(0, Id))$

  Problem: distance estimation is untractable *Arora et. al.*

- The decoder G must nearly restore X: inverse problem

  by minimizing $\mathbb{E}(\|X - G(\Phi(X))\|^2)$

Generation as an Inverse Problem

- From prior on $p(x)$, define Φ with $\Phi(X)$ nearly Gaussian.
  Avoids the intractable step.

- Encode by whitening with a linear operator L: $\widehat{Z} = L\,\Phi(X)$.

$$\underbrace{X \longrightarrow \boxed{\Phi} \longrightarrow \boxed{L} \longrightarrow}_{Encoder} \widehat{Z} \underbrace{\longrightarrow \boxed{L^{-1}} \longrightarrow \boxed{G} \longrightarrow \widehat{X}}_{Decoder}$$

- The generator should invert Φ on X: $G(\Phi(X)) \approx X$.

  Regularized inverse over deep network operators $\mathcal{G}$:

  Does not maximize entropy

Gaussianization from Prior

- $p(x)$ is locally or globally invariant to translations of x
  nearly invariant to small deformations
  has sparse typical realisations with wavelets

$$S_J(X) = \left(\begin{array}{c} X \star \phi_J \\ |X \star \psi_{\lambda_1}| \star \phi_J \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J \end{array} \right)_{\lambda_1, \lambda_2}$$

- Averaging by ϕ_J Gaussianizes: central limit theorem
  when $2^J \to \infty$

Scattering Inverse Network

Tomás Angles

- Encoder: whitens $S_J X$ with a linear operator L

$$X \longrightarrow \boxed{S_J} \longrightarrow \boxed{L} \longrightarrow \widehat{Z}$$

: white nearly Gaussian

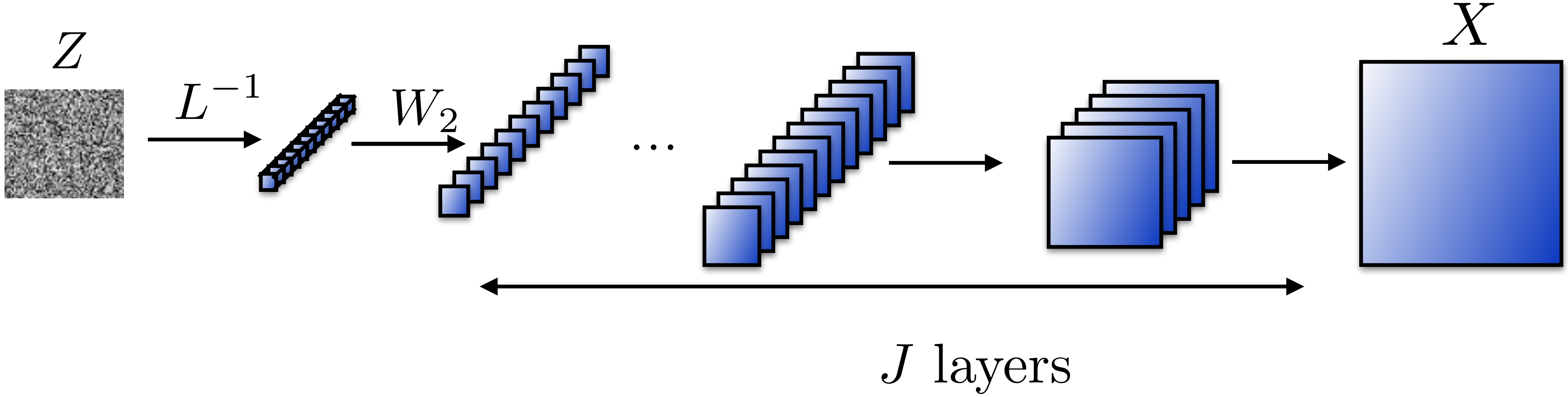
- Decoder:

$$Z \sim \mathcal{N}(\mu, Id) \longrightarrow \boxed{L^{-1}} \longrightarrow \boxed{G} \longrightarrow \widehat{X}$$

Regularized inversion of S_J with a deep net G minimising:

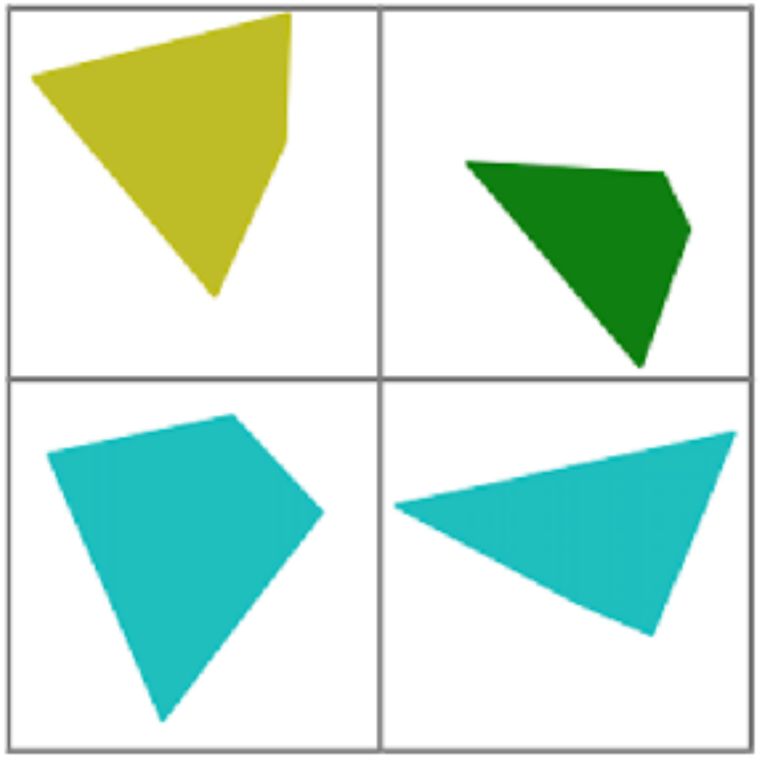
$$G = \min_{\hat{G} \in \mathcal{G}} \sum_i \Big(\|\hat{G}(S_J(x_i)) - x_i\| \Big)$$

regularized inverse scattering

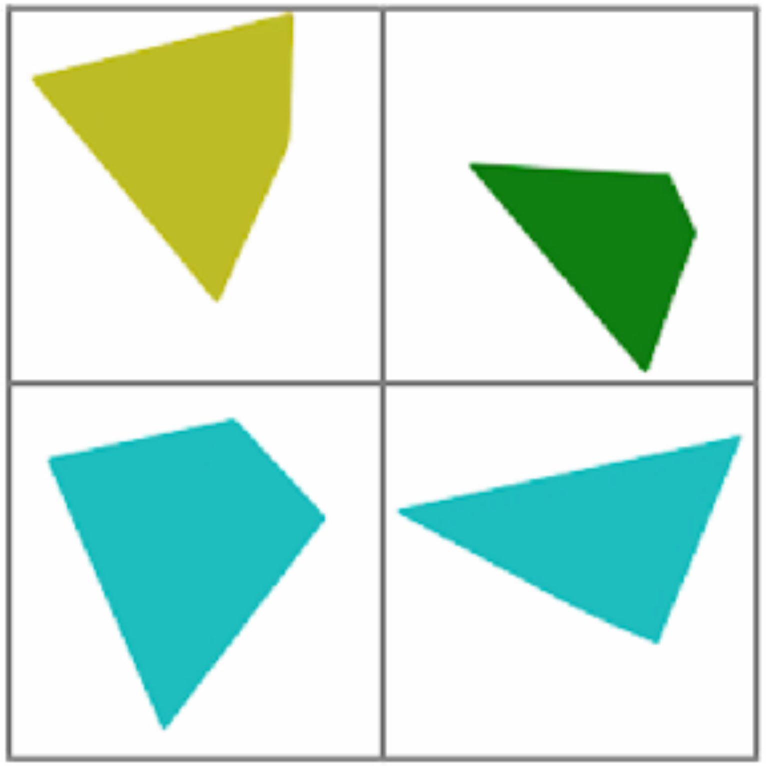


Training Reconstruction

Training x_i
Polygones



$G(S_J(x_i))$

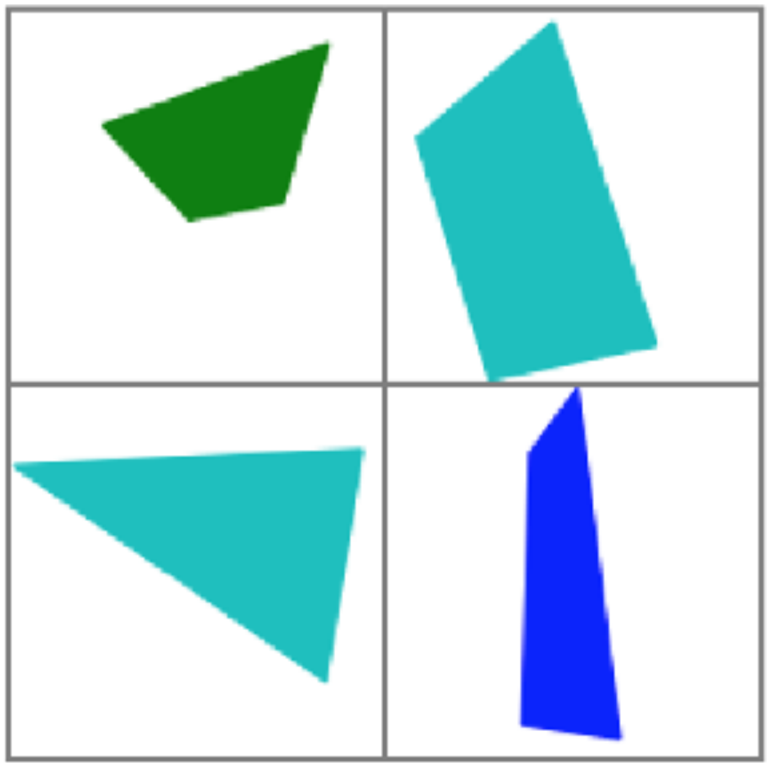


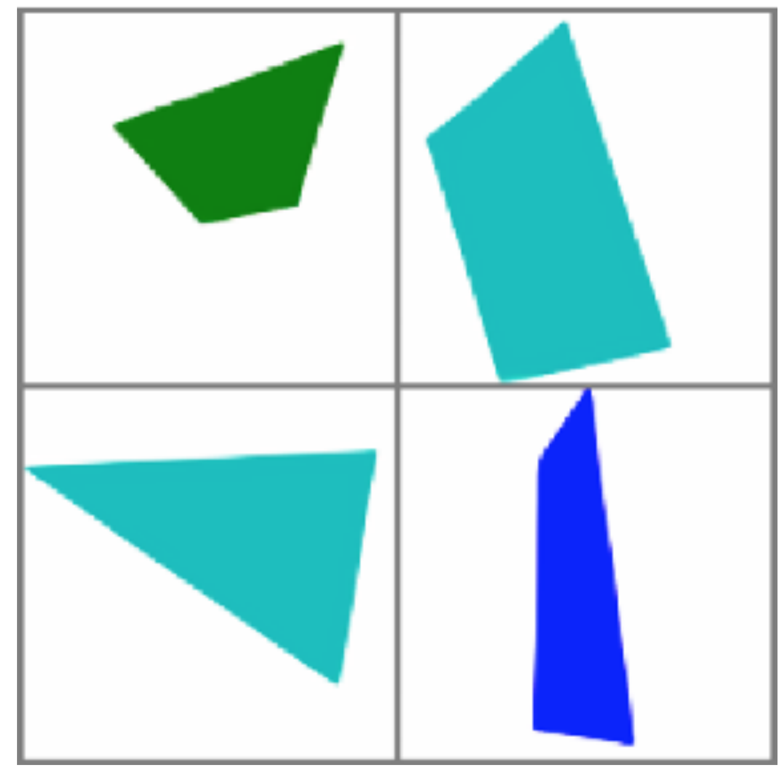
Celebrities Data Basis

Testing Reconstruction

Testing
x_t

$G(S_J(x_t))$



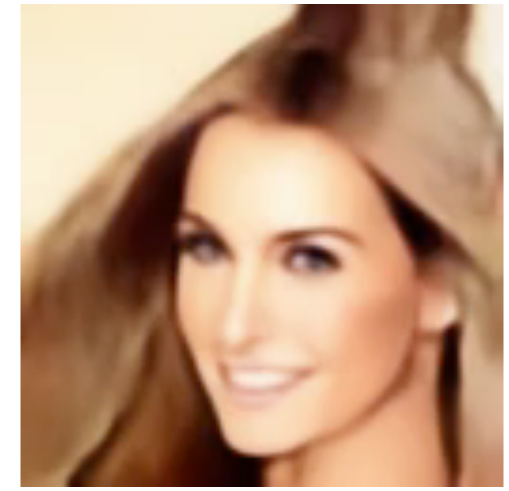


Generative Interpolations

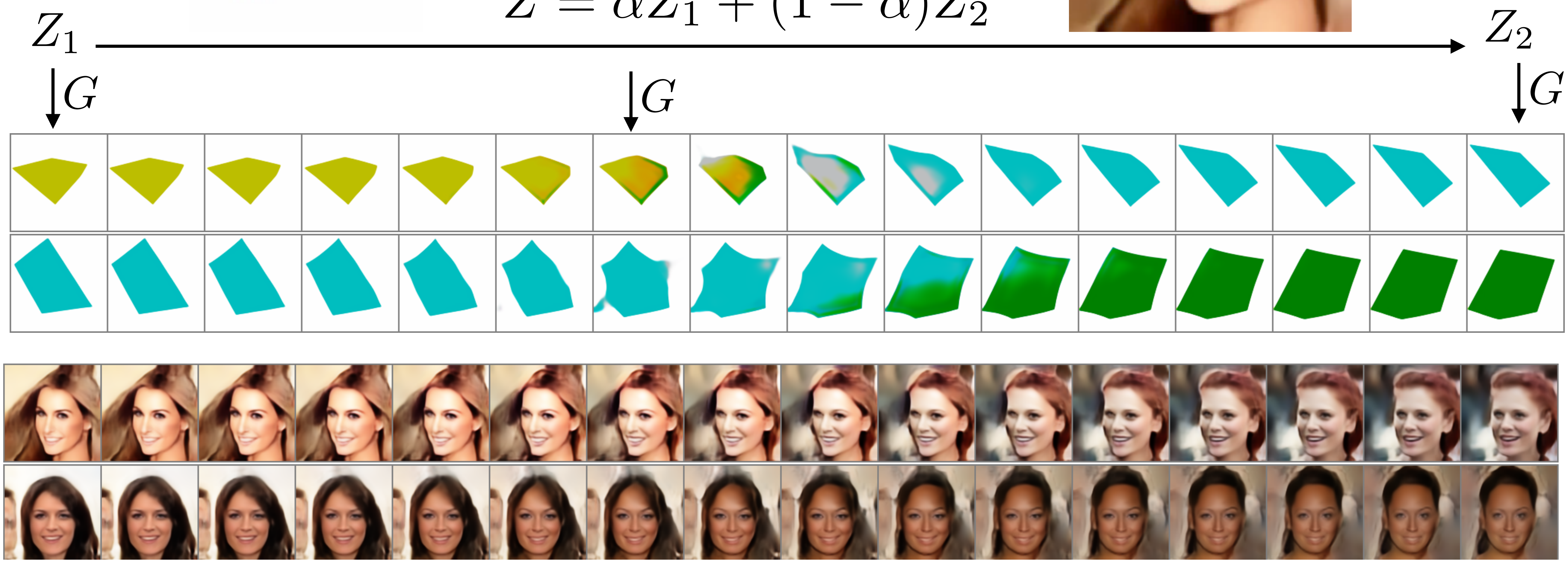
Tomás Angles

Polygons

Celebrities



$$Z = \alpha Z_1 + (1-\alpha) Z_2$$



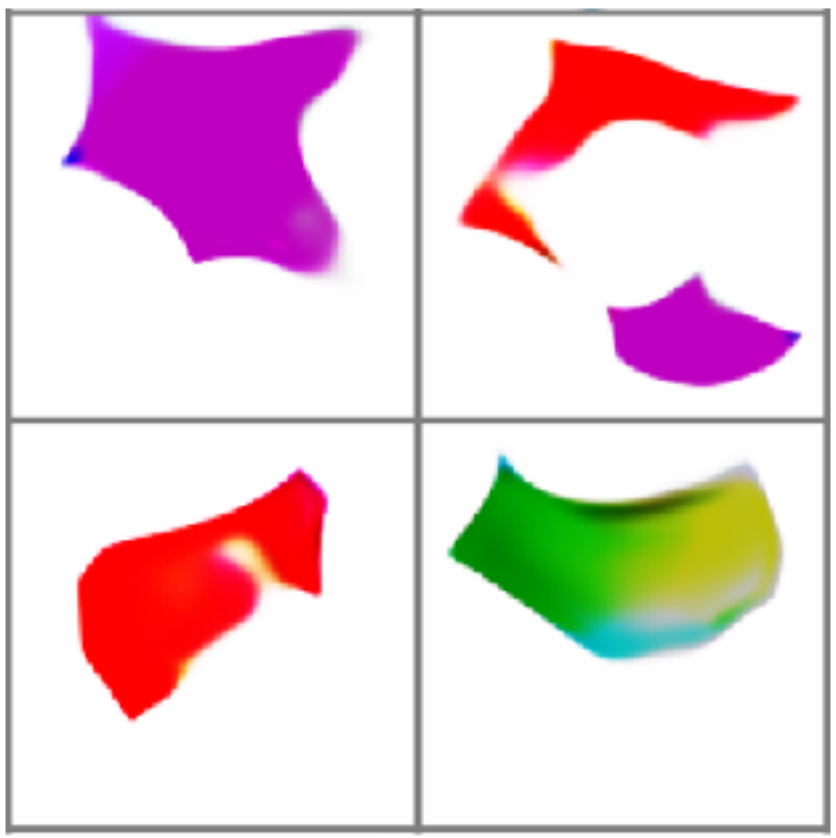
Generative Scattering Networks

Tomás Angles

Images synthesised from Gaussian white noise Z:

$$Z \sim \mathcal{N}(\mu, Id) \longrightarrow \boxed{L^{-1}} \longrightarrow \boxed{G} \longrightarrow \widehat{X}$$

G: regularized inversion of S_J



Networks regularize with some form of "memory storage".
Sparse activations for images from data basis.

More Complex Bases (bedrooms)

Memory can saturate if data basis is too complex:

Loss of resolution or loss of structures (mode dropping)

Training images

Generative Adversarial Nets.

Variational Autoencoders

Reconstructed from Noise

Generative Adversarial Networks

T. Karras, T. Aila, S. Laine, J. Lehtinen

Generated from Hollywood celebrities data basis

Generative adversarial networks do not reduce quality but "forget" images (mode dropping).

Conclusion

- Deep neural network architectures are providing a new statistical tools beyond high order moments.
- Scale separation and interactions through filters/wavelets.
- Distributed memory storage: not understood as most properties…
- Opening the black box: a beautiful statistical and information processing problem!