Multiscale Sparse Models in Convolutional Networks

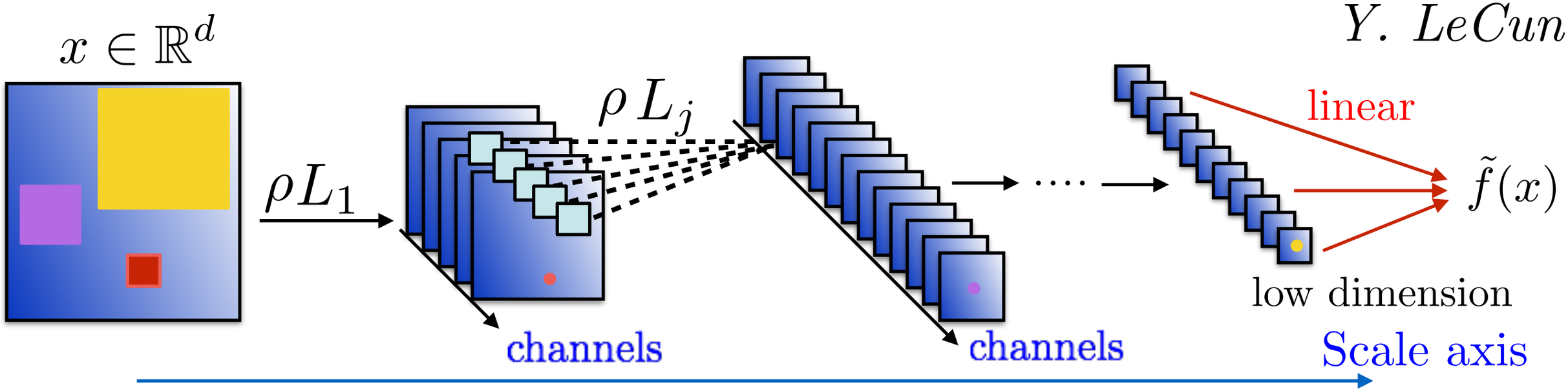
Tomas Anglès, Roberto Leonarduzi,
Stéphane Mallat, Louis Thiry,
John Zarka, Sixin Zhang

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

Deep Convolutional Network

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

Y. LeCun



L_j: spatial convolutions and linear combination of channels

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

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

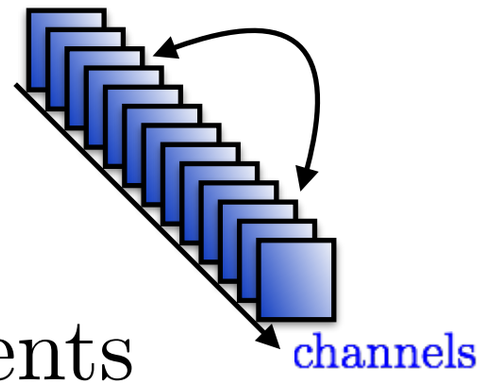
Exceptional results for *images, speech, language, bio-data quantum chemistry regressions, ...*

How does it reduce dimensionality ?

Multiscale , Sparsity, Invariants

Statistical Models from 1 Example

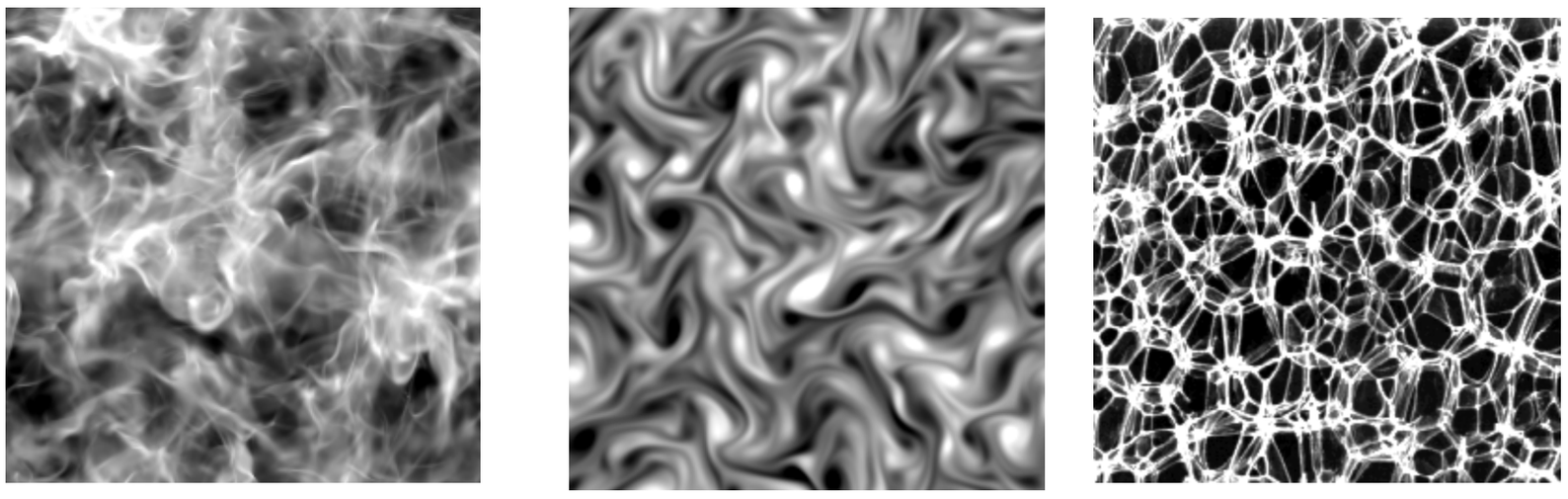
M. Bethdge et. al.



- 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

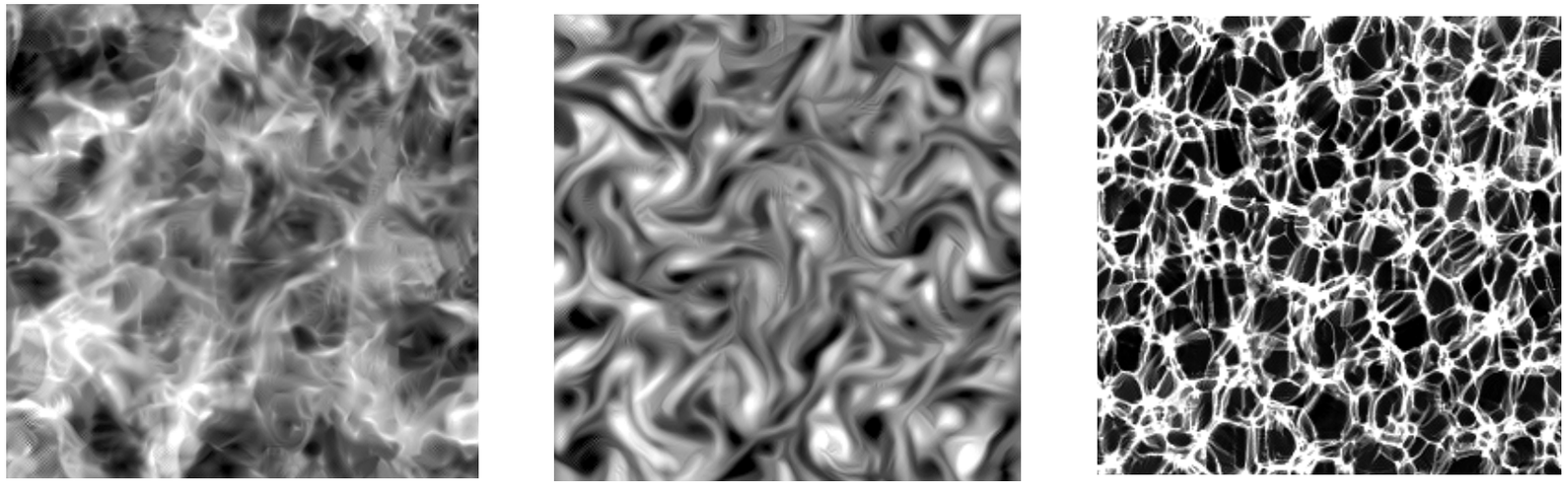
x

$6\,10^4$ pixels



$\tilde{x}$

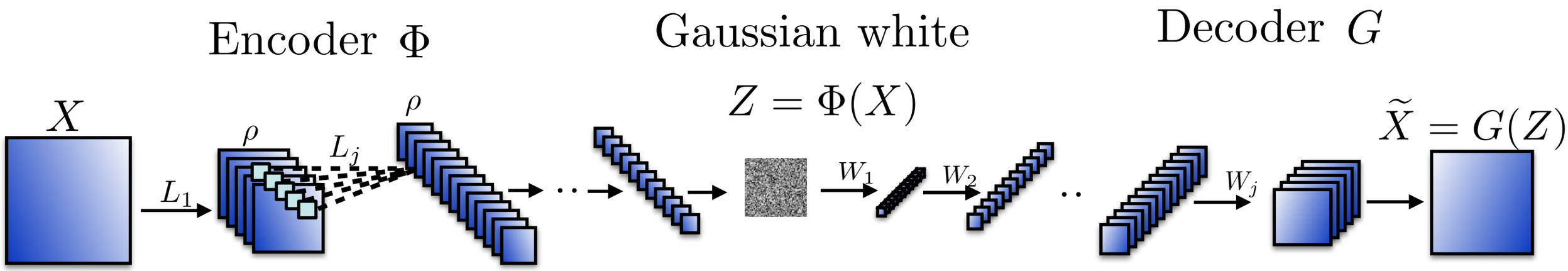
$2\,10^5$ correlations

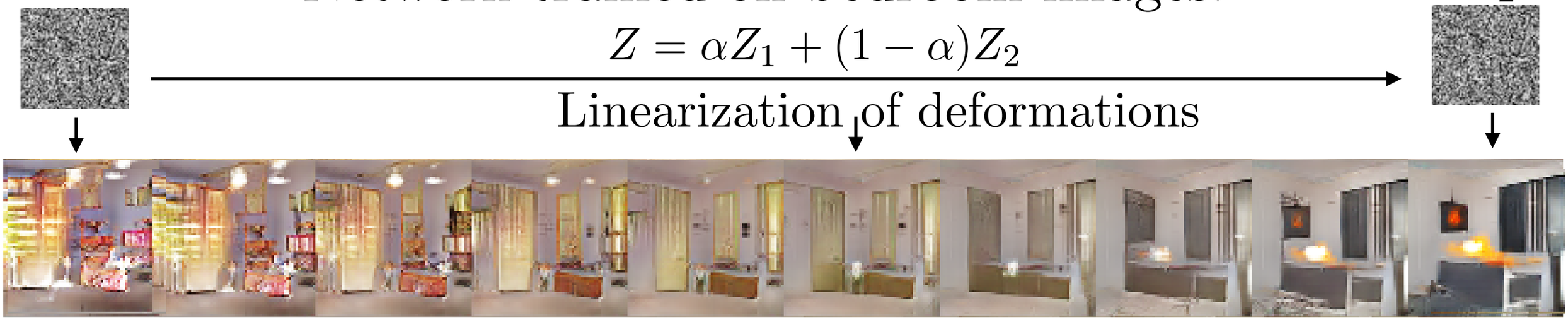


What mathematical interpretation ?

Learned Generative Networks

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





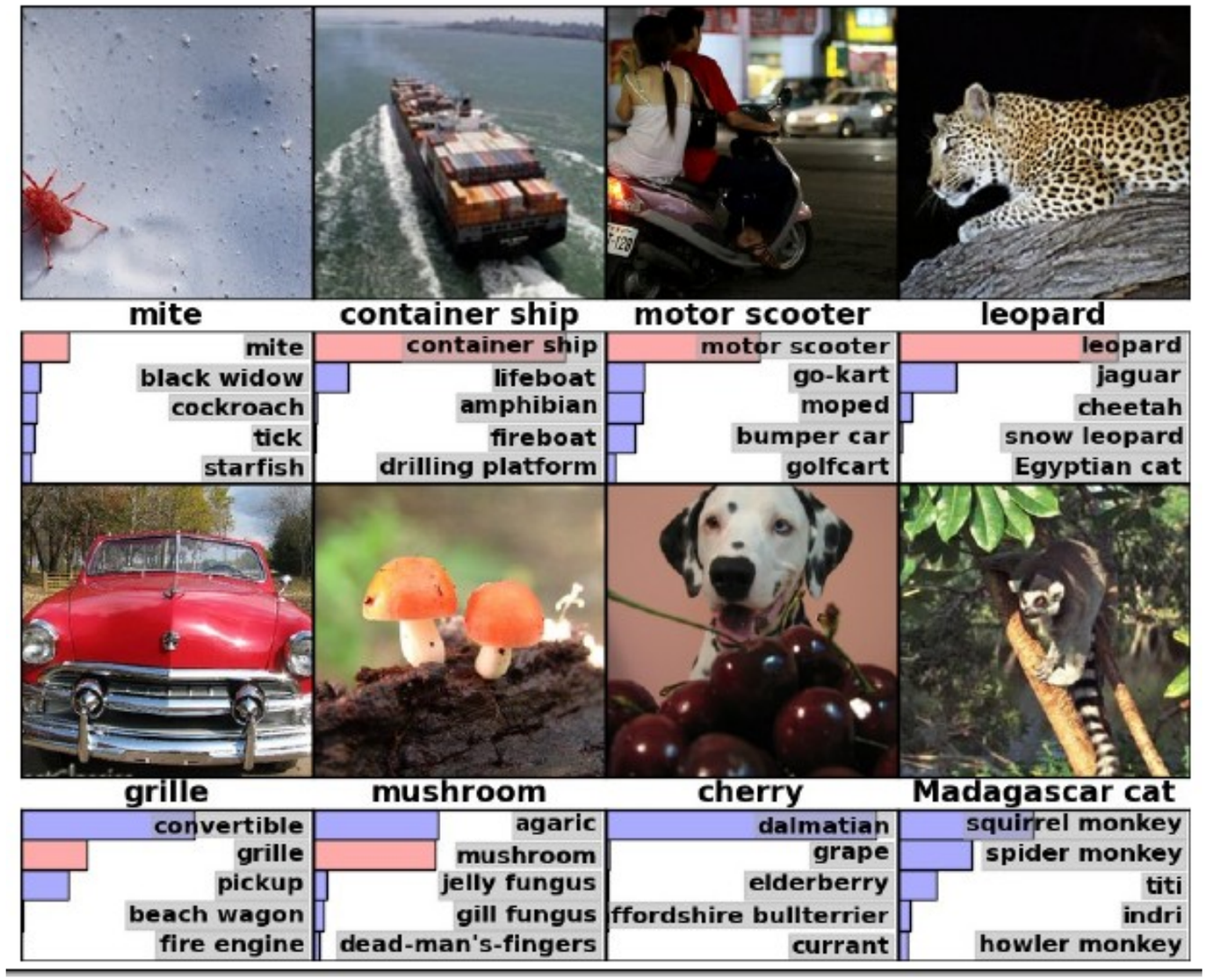
Network trained on faces of celebrities:

$G(Z)$

What mathematical interpretation ?

Image Classification: ImageNet 2012

1000 classes, 1.2 million labeled training images, of 224×224 pixels



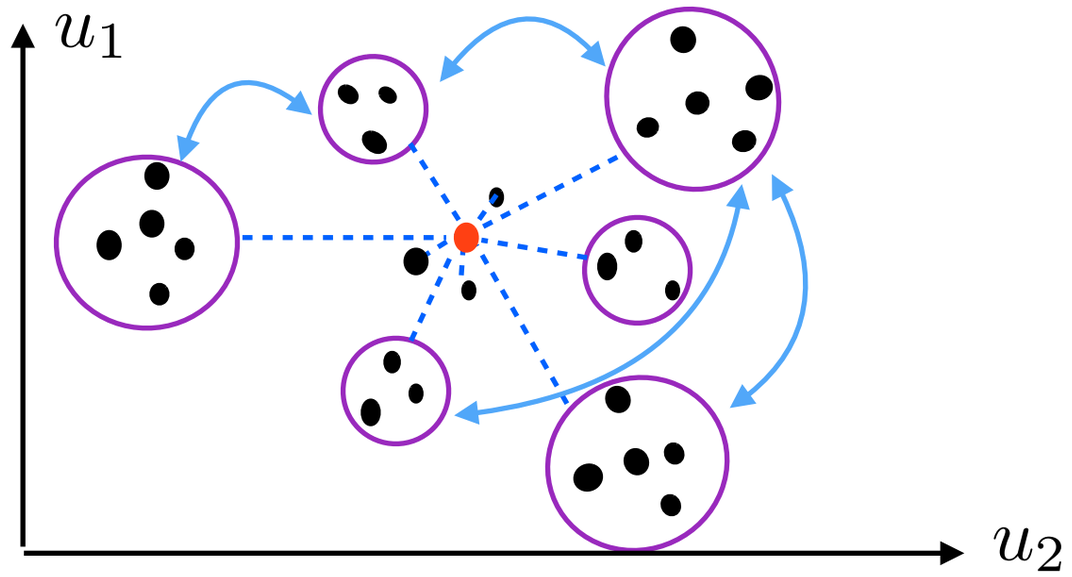
	Alex-Net	ResNet
Top 5 error	20%	10%

Scale Separation and Interactions

- **Dimension reduction:**

Interactions de d bodies represented by $x(u)$: particles, pixels...

Interactions across scales



Multiscale regroupement of interactions of d bodies into interactions of $O(\log d)$ groups.

Scale separation $\Rightarrow$ wavelet transforms.

How to capture scale interactions ?

Critical harmonic analysis problems since 1970's

Overview

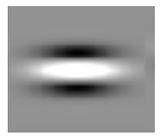
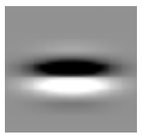
- Scale separation with wavelets and interactions through phase

- Linear scale interaction models:
  - Compressive signal approximations
  - Stochastic models of stationary processes

- Non-linear scale interactions models with sparse dictionaries
  - Generative autoencoders
  - Classification of ImageNet

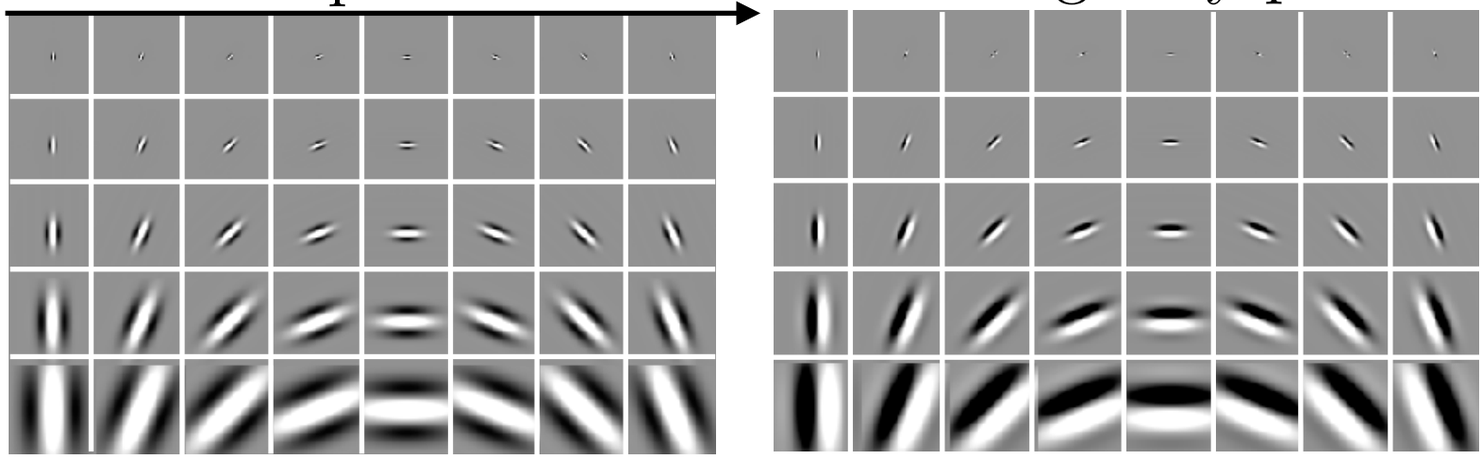
- All these roads go to Convolutional Neural Networks…

Scale separation with Wavelets

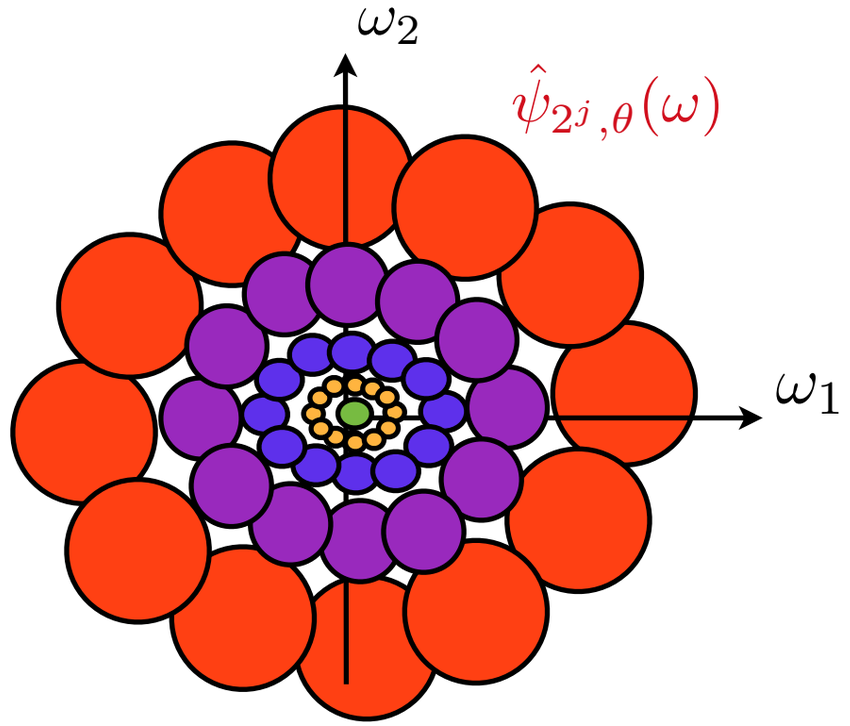
- Wavelet filter $\psi(u)$:  $+ i$  2 phases

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

real parts θ imaginary parts



Fourier



- Wavelet transform: invertible

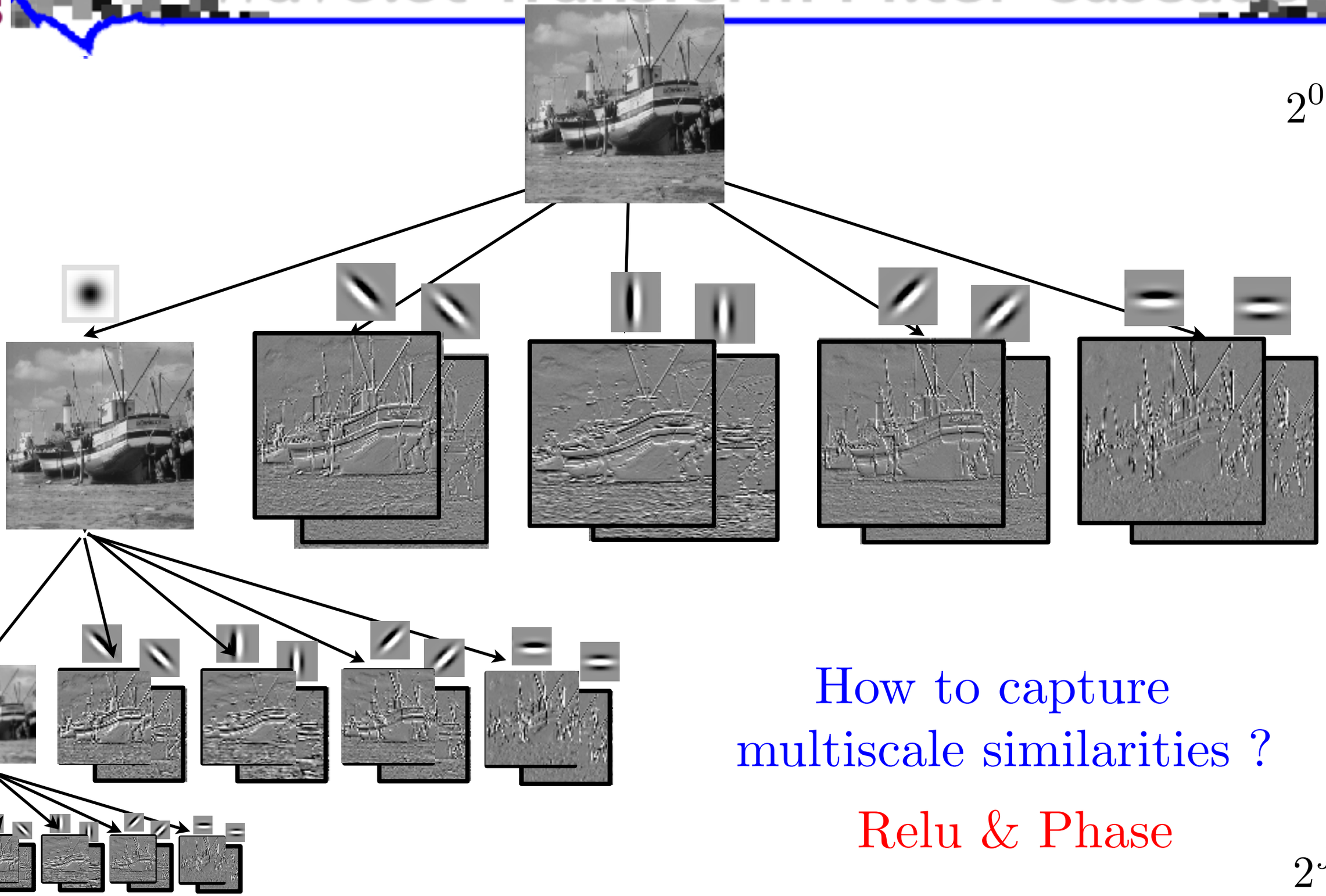
$$Wx = \begin{pmatrix} x \star \phi_{2^J} \\ x \star \psi_\lambda \end{pmatrix}_\lambda$$

$$\widehat{x \star \psi_\lambda}(\omega) = \hat{x}(\omega)\,\hat{\psi}_\lambda(\omega)$$

- Zero-mean and no correlations across scales: problem!

$$\sum_u x \star \psi_\lambda(u)\, x \star \psi^*_{\lambda'}(u) = \sum_\omega |\widehat{x}(\omega)|^2\, \psi_\lambda(\omega)\, \psi_\lambda(\omega)^* \approx 0 \quad \text{if} \quad \lambda \neq \lambda'$$

Wavelet Transform Filter Cascade



How to capture multiscale similarities ?

Relu & Phase

Rectified Wavelet Coefficients

- Multiphase real wavelets: $\psi_{\alpha,\lambda} = \text{Real}(e^{-i\alpha}\,\psi_\lambda)$
- Rectified with $\rho(a) = \max(a, 0)$:

$$Ux = \begin{pmatrix} x \star \phi_{2^J} \\ \rho(x \star \psi_{\alpha,\lambda}) \end{pmatrix}_{\alpha,\lambda} : \text{conv. net. coefficients}$$

- Linearly invertible:

$$\rho(a) + \rho(-a) = a \quad \Rightarrow \quad x = U^{-1}Ux \;\text{ with }\; U^{-1} \text{ linear}$$

- Relu creates non-zero mean and correlations across scales:

$$\sum_u \rho(x \star \psi_{\alpha,\lambda}(u))$$

$$\sum_u \rho(x \star \psi_{\alpha,\lambda}(u))\,\rho(x \star \psi_{\alpha',\lambda'}(u))$$

Linear Rectifiers act on Phase

$$Ux(u,\alpha,\lambda) = \rho(x \star \mathrm{Real}(e^{i\alpha}\,\psi_\lambda)) = \rho(\mathrm{Real}(e^{i\alpha}\,x \star \psi_\lambda))$$

$$x \star \psi_\lambda = |x \star \psi_\lambda|\, e^{i\varphi(x\star\psi_\lambda)}$$

Homogeneous: $\rho(\alpha a) = \alpha\,\rho(a)$ if $\alpha > 0$

$$Ux(u,\alpha,\lambda) = |x \star \psi_\lambda|\,\rho(\cos(\alpha + \varphi(x \star \psi_\lambda))$$

Phase harmonics $\forall z = |z|e^{i\varphi(z)} \in \mathbb{C}\ ,\ [z]^k \triangleq |z|\, e^{ik\varphi(z)}$

A Relu computes phase harmonics:

Theorem : Fourier transform along the phase α:

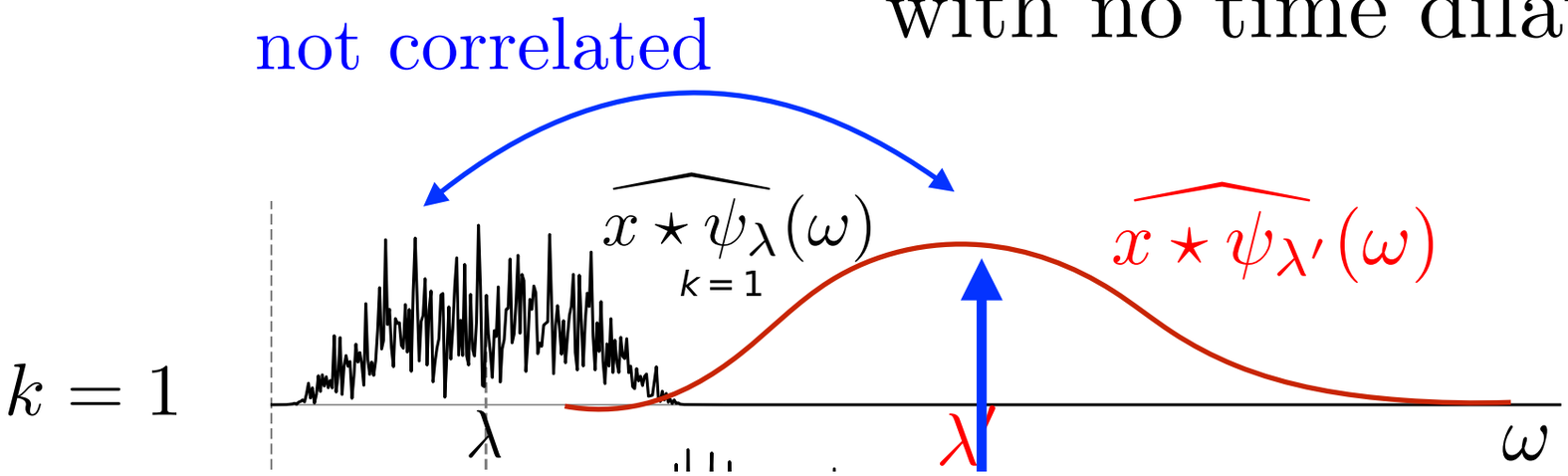
$$\widehat{U}x(u,k,\lambda) = \hat{\gamma}(k)\,|x \star \psi_\lambda(u)|\, e^{ik\,\varphi(x\star\psi_l\,a(u))}$$

with $\gamma(\alpha) = \rho(\cos\alpha)$ for any homogeneous non-linearity ρ.

Frequency Transpositions

Phase harmonics: $[x \star \psi_\lambda]^k = |x \star \psi_\lambda(u)|\, e^{i\,k\,\varphi(x \star \psi_\lambda(u))}$

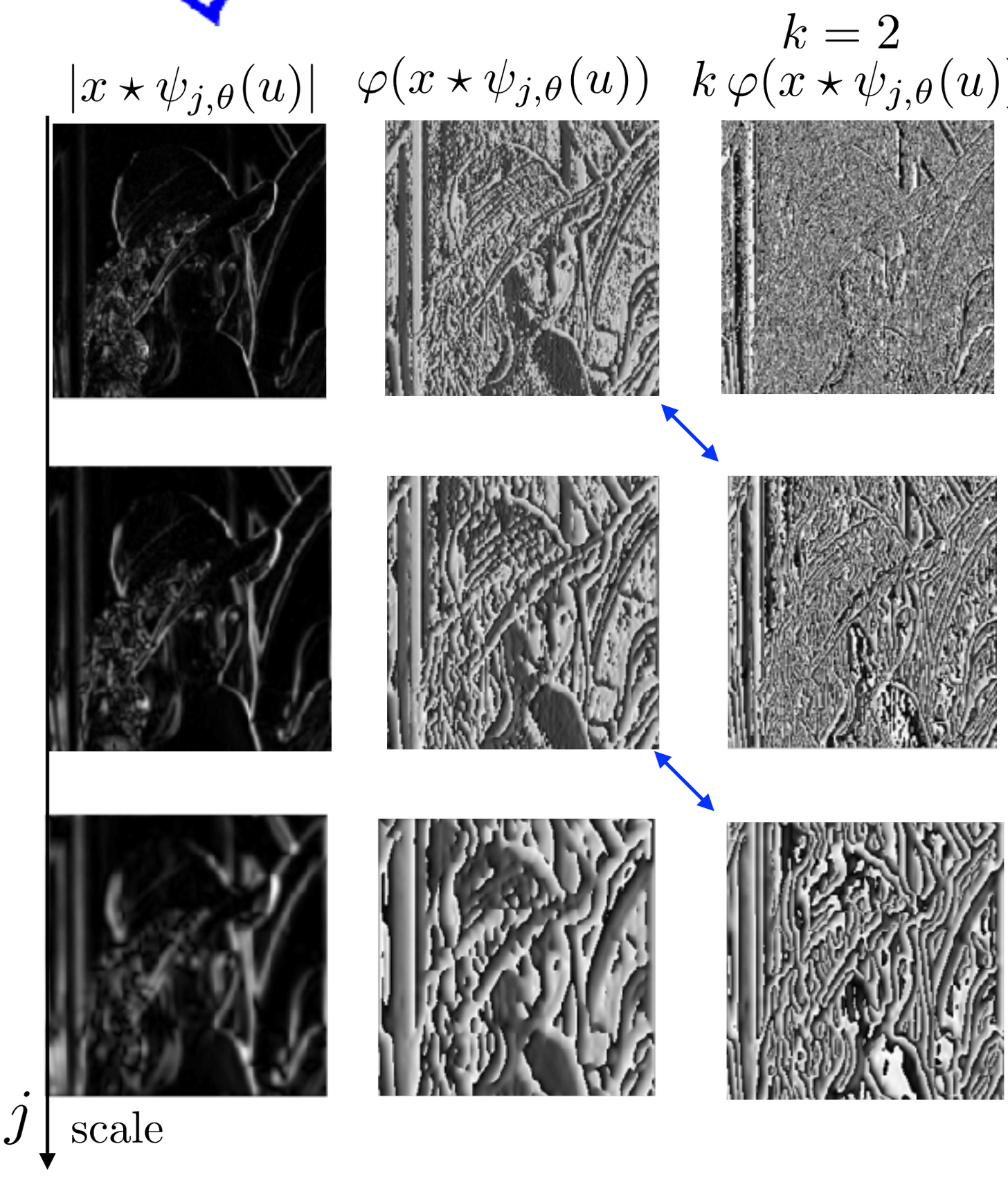
Performs a non-linear frequency dilation / transposition with no time dilation

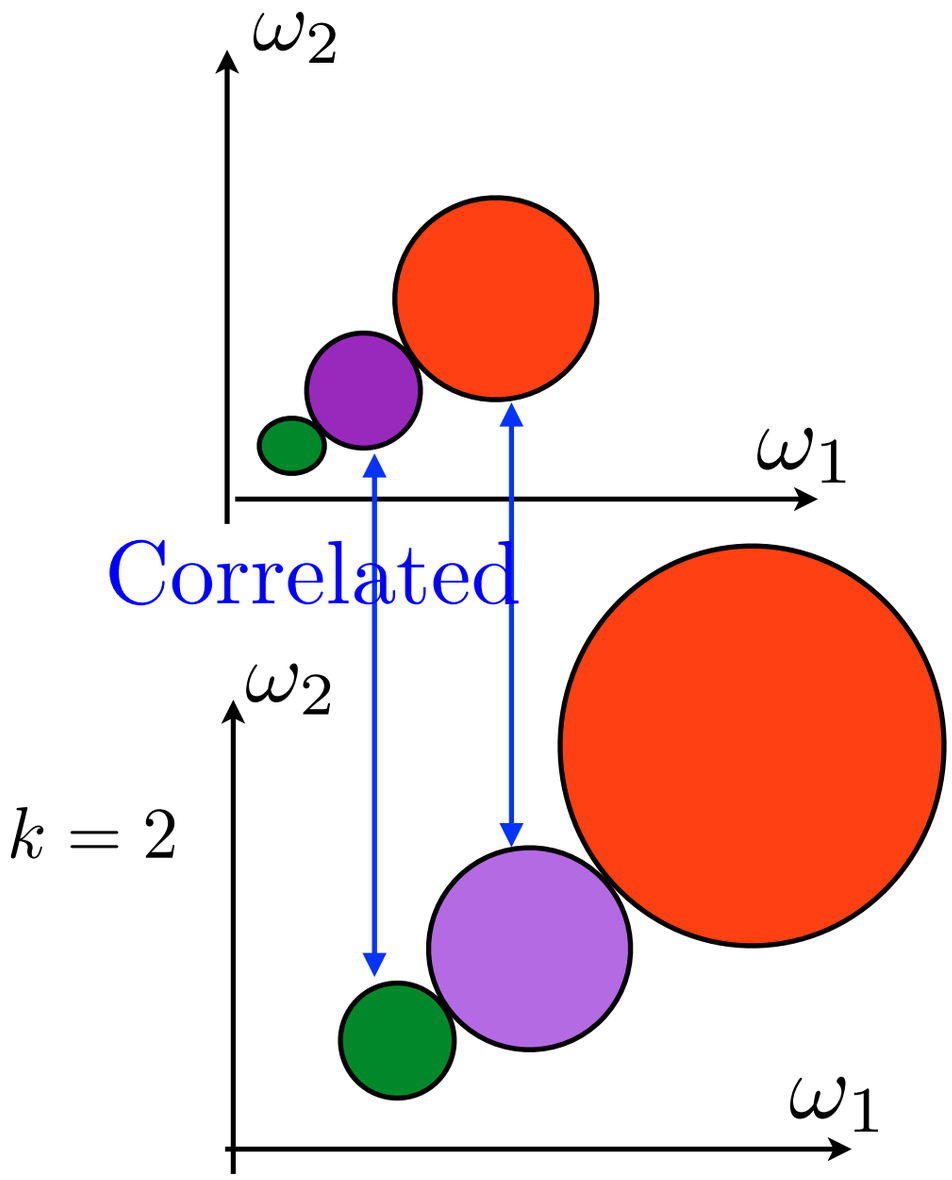


Phase Harmonics

Correlated if $k\lambda \approx \lambda'$

Scale Transposition with Harmonics





Phase harmonics:
Frequency transpositions

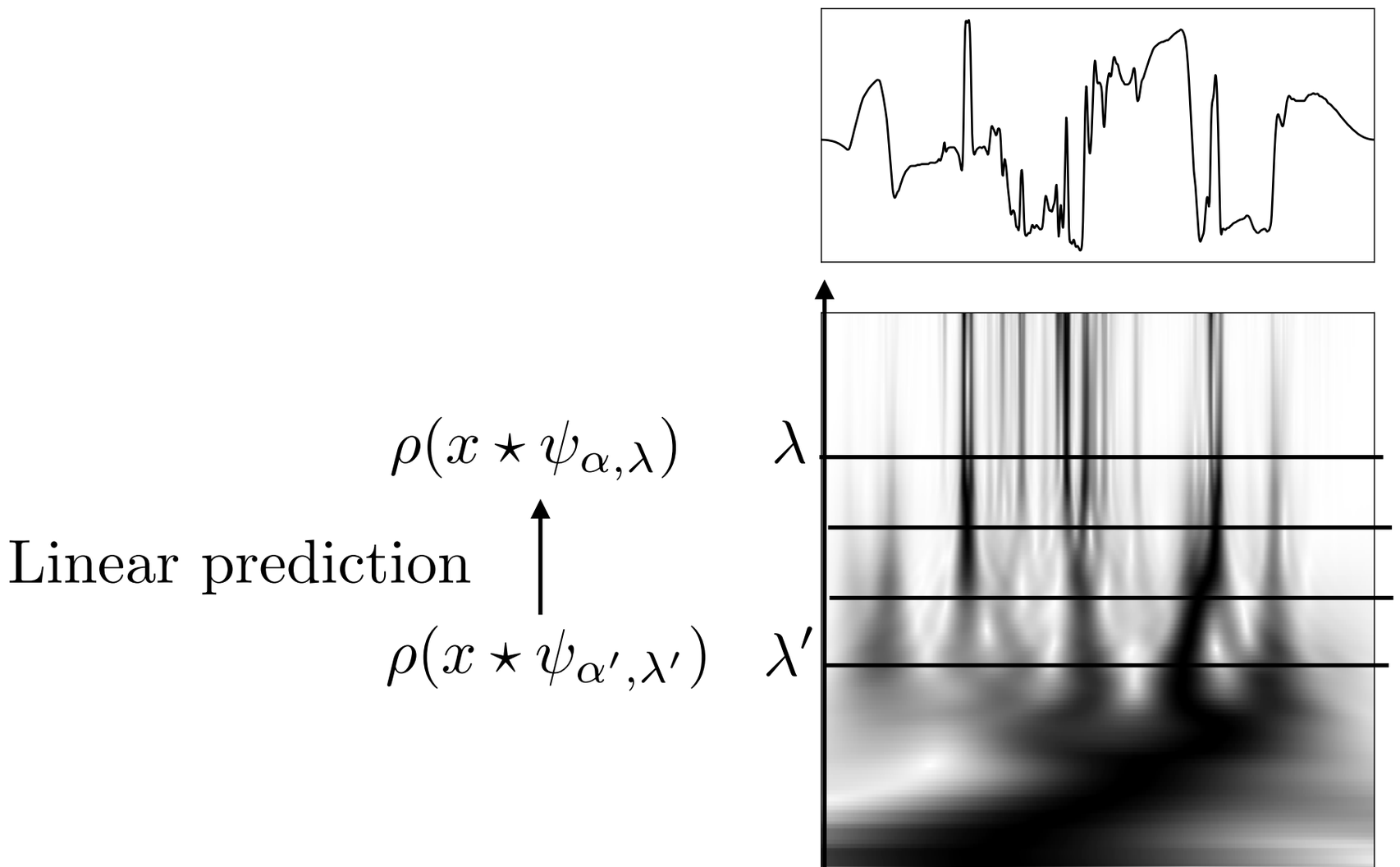
Linear Prediction Across Scales/Freq.

- Relu mean and correlations: invariant to translations

$$M(\alpha, \lambda) = d^{-1} \sum_u \rho(x \star \psi_{\alpha,\lambda}(u))$$

$$C(\alpha, \lambda, \alpha', \lambda') = d^{-1} \sum_u \rho(x \star \psi_{\alpha,\lambda}(u))\, \rho(x \star \psi_{\alpha',\lambda'}(u))$$

Define linear autoregressive model from low to high frequencies:

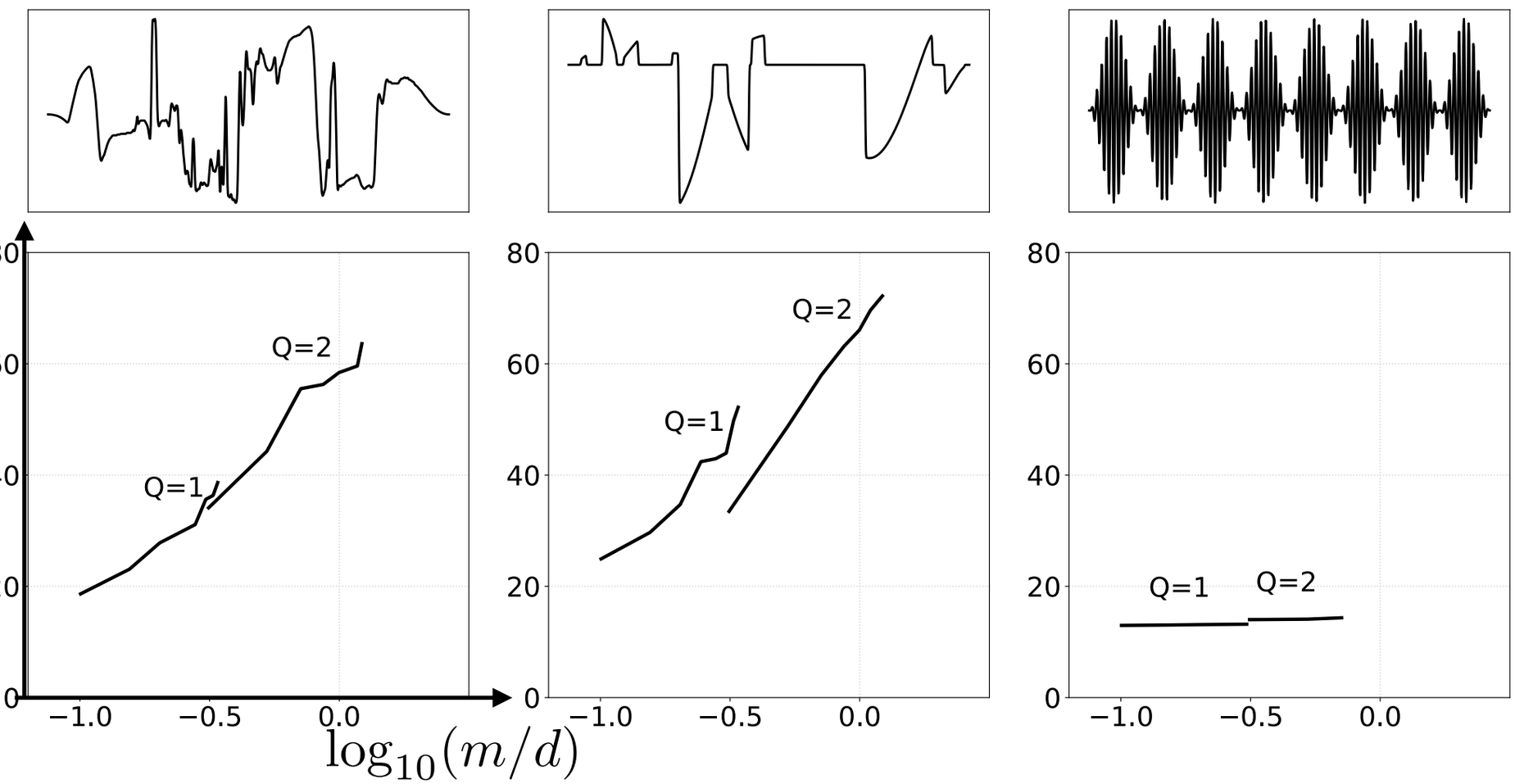


Compressive Reconstructions

Gaspar Rochette, Sixin Zhang

- If $x \star \psi_\lambda$ is sparse then x is recovered from $m \ll d$ phase harmonic means Mx and covariances Cx:

$$\tilde{x} = \arg\min_y \|Cx - Cy + (Mx - My)\,(Mx - My)^*\|^2$$

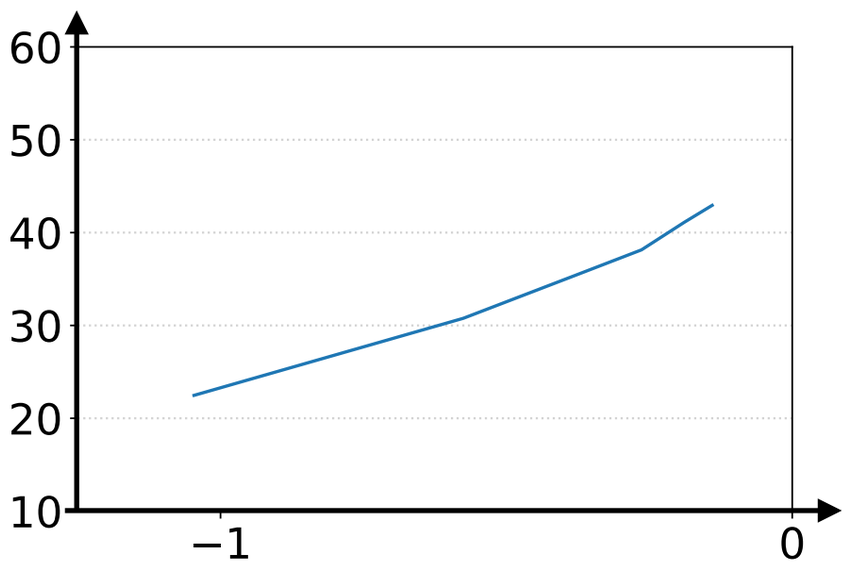


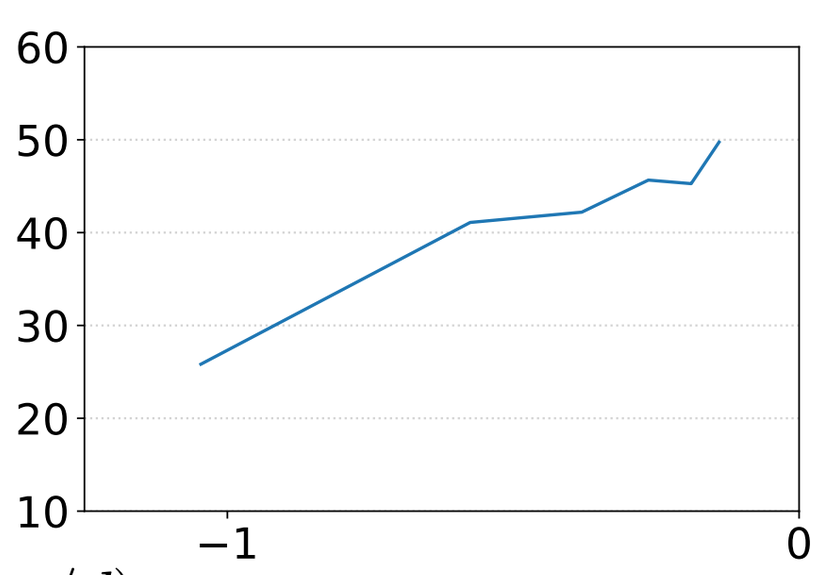
Approximation rate optimal for total variation signals:

$$\|x - \tilde{x}\| \sim m^{-2}$$

Compressive Reconstructions

PSNR (db)





$\log_{10}(m/d)$

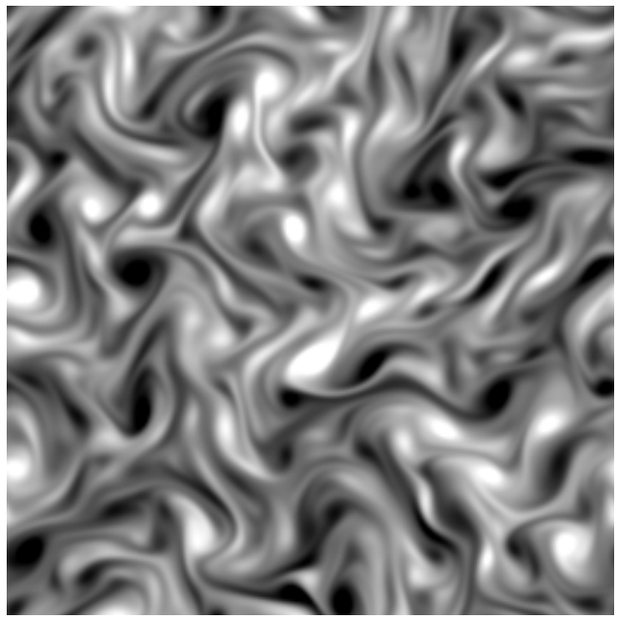
Approximation rate optimal for total variation signals:

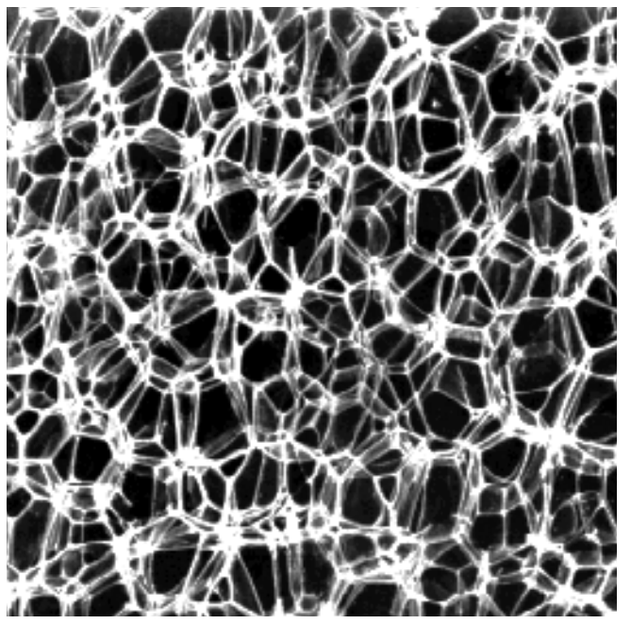
$$\|x - \tilde{x}\| \sim m^{-1}$$

Gaussian Models of Stationary Proc.

What stochastic models for turbulence ?

$$d = 6\,10^4$$



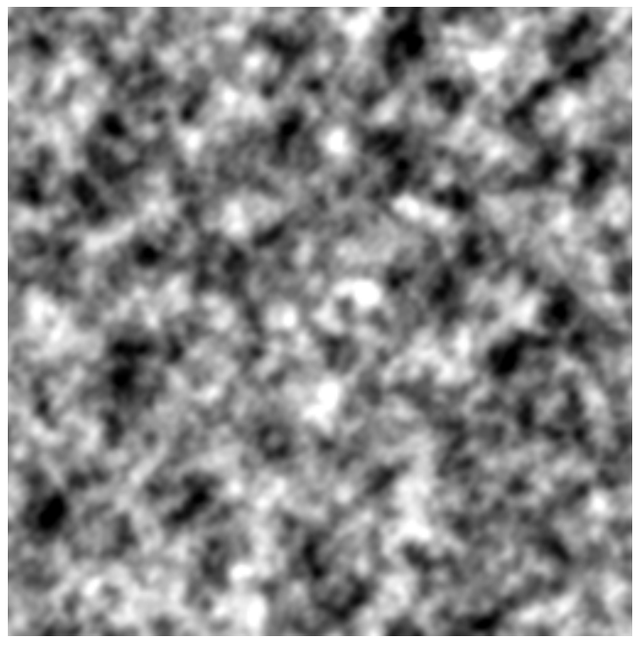


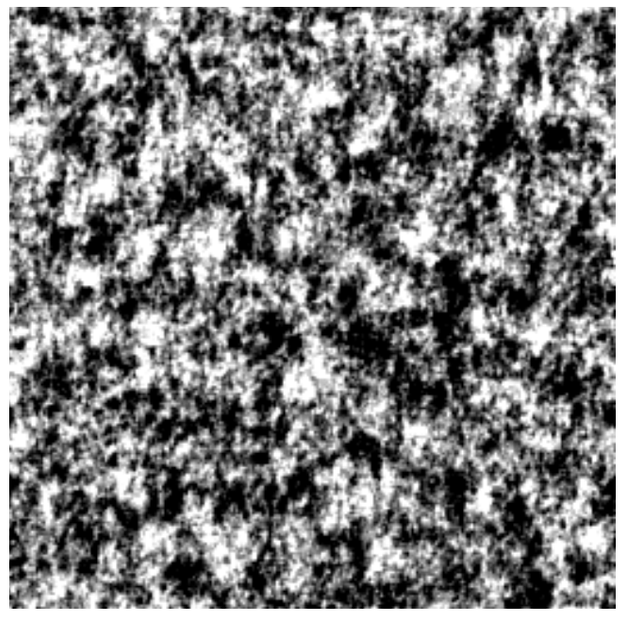
Kolmogorov model:

From d empirical moments:
$d^{-1} \sum_u (x(u)\, x(u-\tau))$

x

$\tilde{x}$





Gaussian model with
same power spectrum

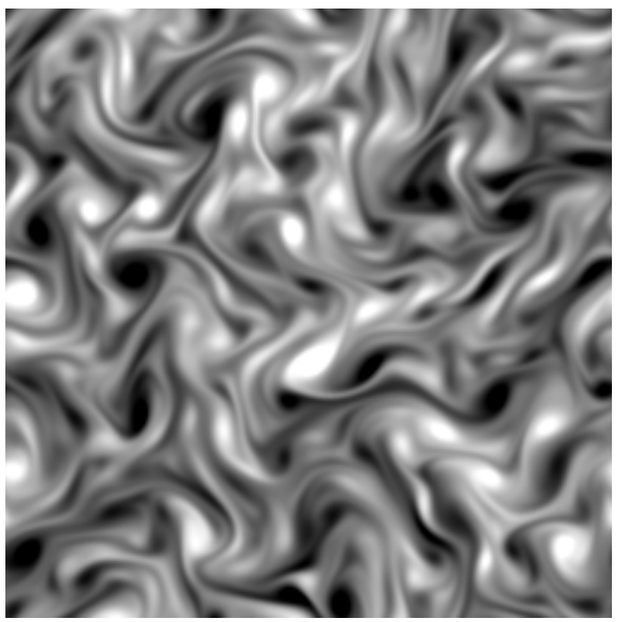
No correlation is captured across scales and frequencies.
Random phases.

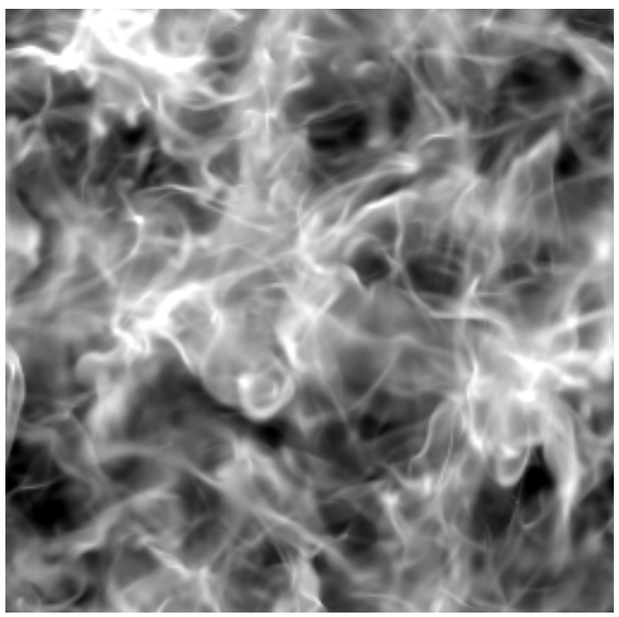
How to capture non-Gaussianity and long range interactions ?

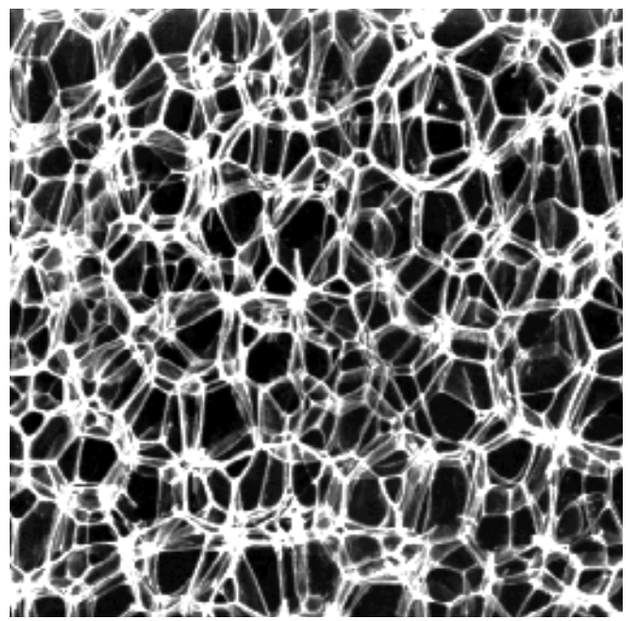
Models of Stationary Processes

Sixin Zhang

x







If ergodic then empirical moments converge:

$$d^{-1}\sum_u \rho(x \star \psi_{\alpha,\lambda}(u)) \xrightarrow[d\to\infty]{} \mathbb{E}\Big(\rho(x \star \psi_{\alpha,\lambda})\Big)$$

$$d^{-1}\sum_u \rho(x \star \psi_{\alpha,\lambda}(u))\,\rho(x \star \psi_{\alpha',\lambda'}(u)) \xrightarrow[d\to\infty]{} \mathbb{E}\Big(\rho(x \star \psi_{\alpha,\lambda})\,\rho(x \star \psi_{\alpha',\lambda'})\Big)$$

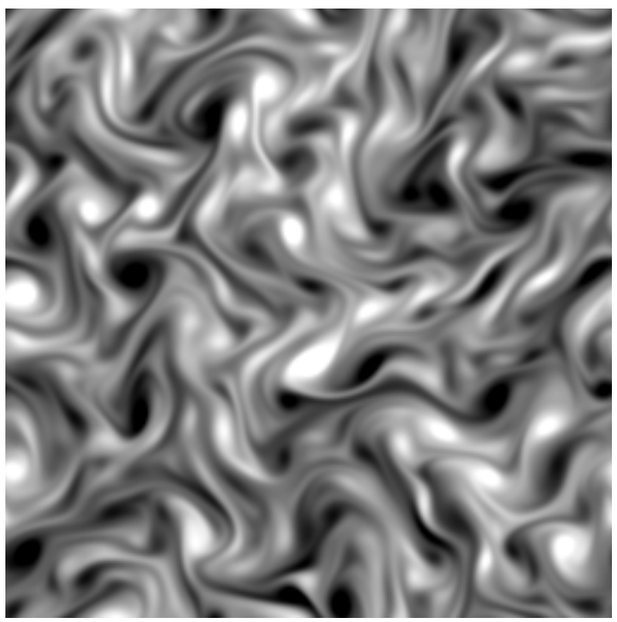
- Stationary processes conditioned by translation invariant moments

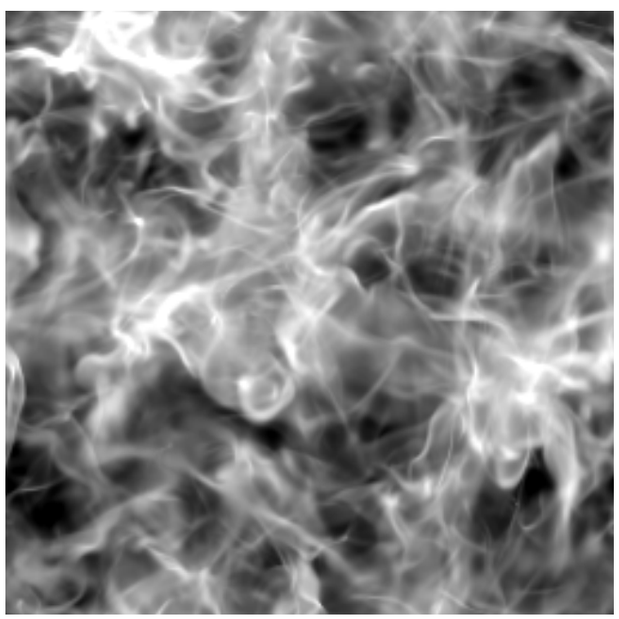
Ergodic Stationary Processes

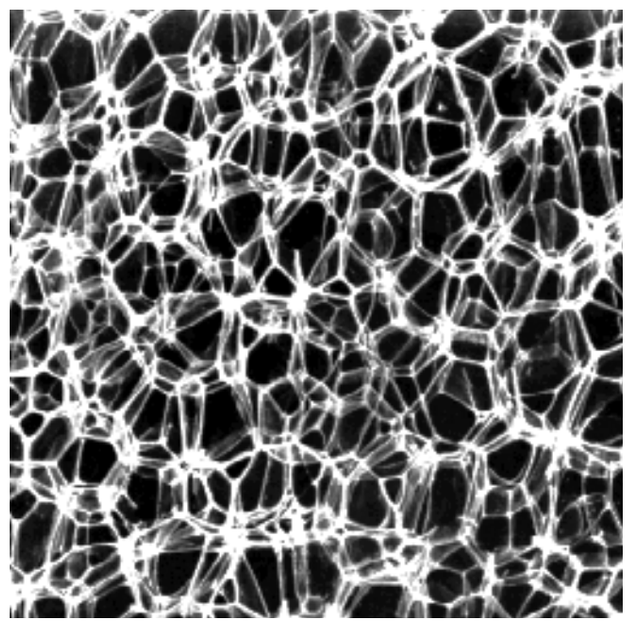
Sixin Zhang

$d = 6\,10^4$

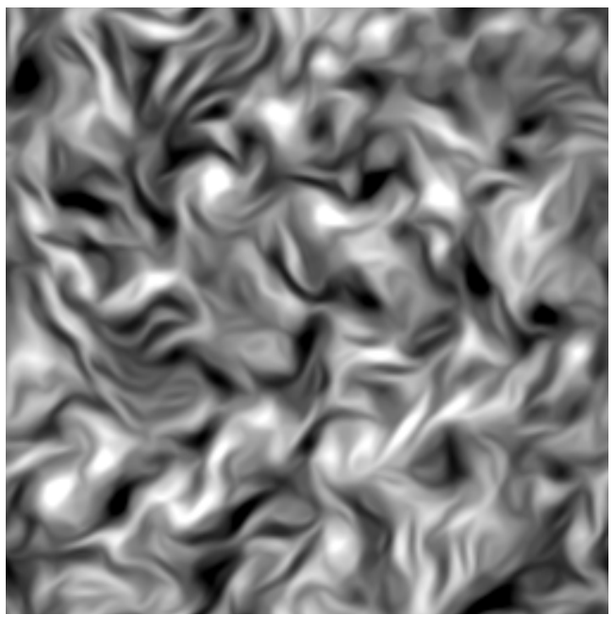
x

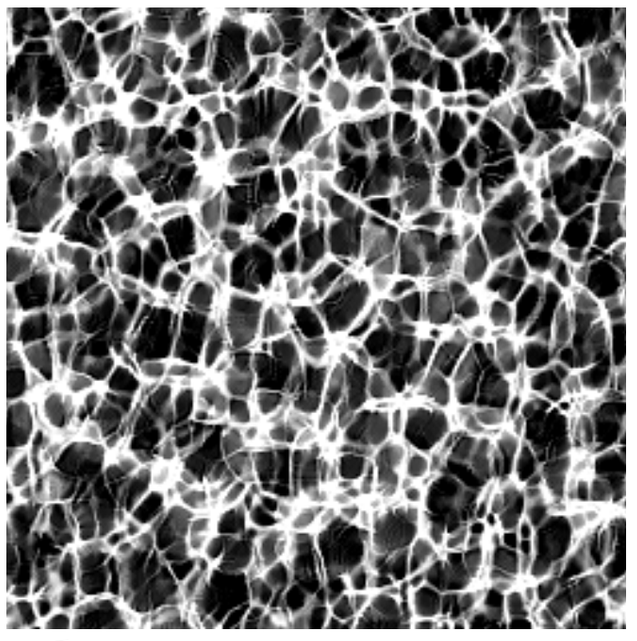






$\tilde{x}$





$m = 3\,10^3$

number
of moments

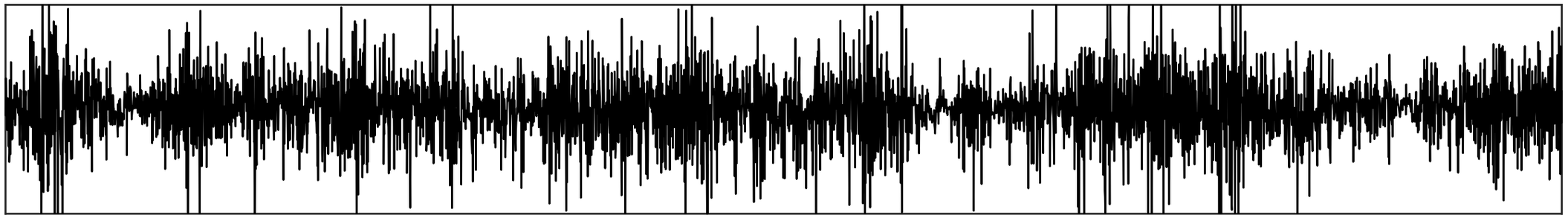
Phase coherence is captured
Same quality as with learned Deep networks
with much less moments

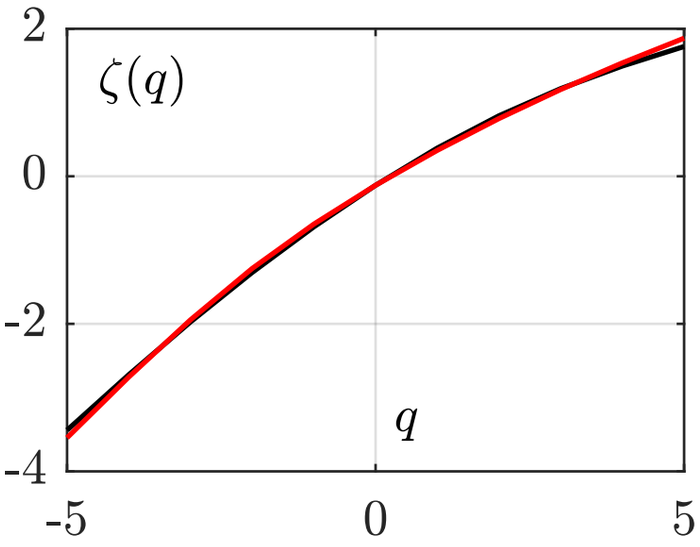
Multifractal Models

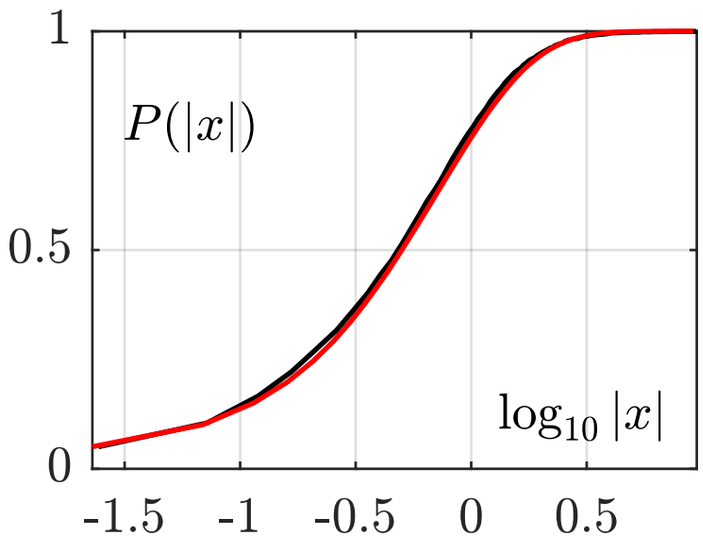
without high order moments

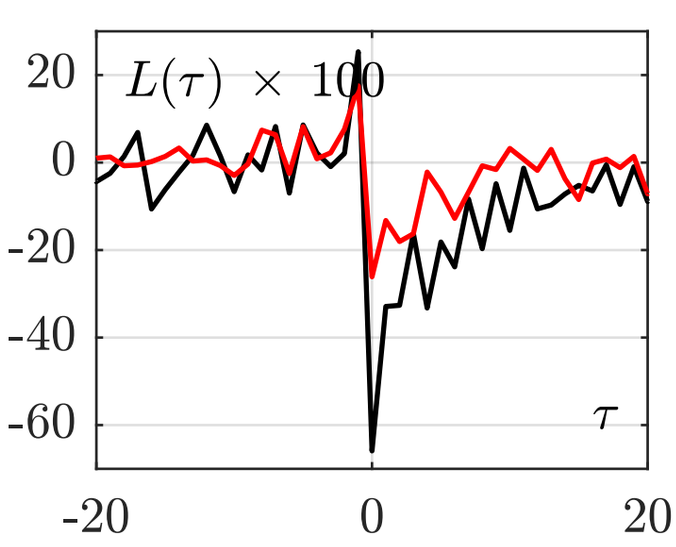
Roberto Leonarduzi

Financial S & P 500 returns:





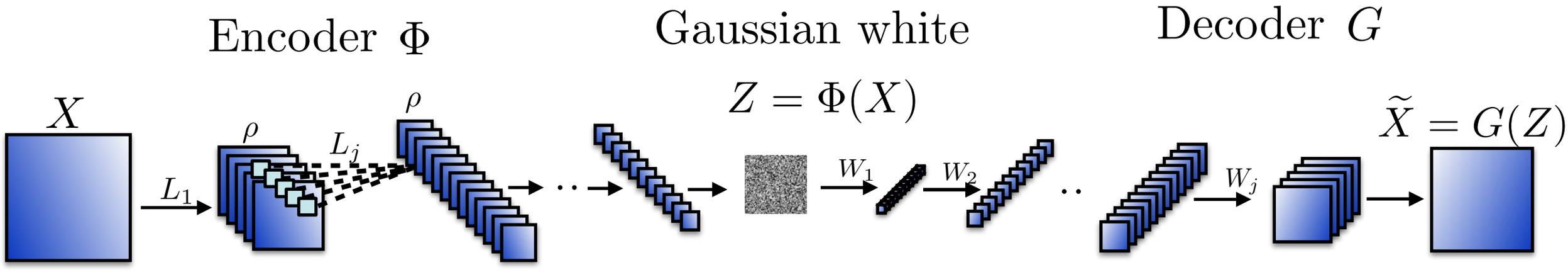


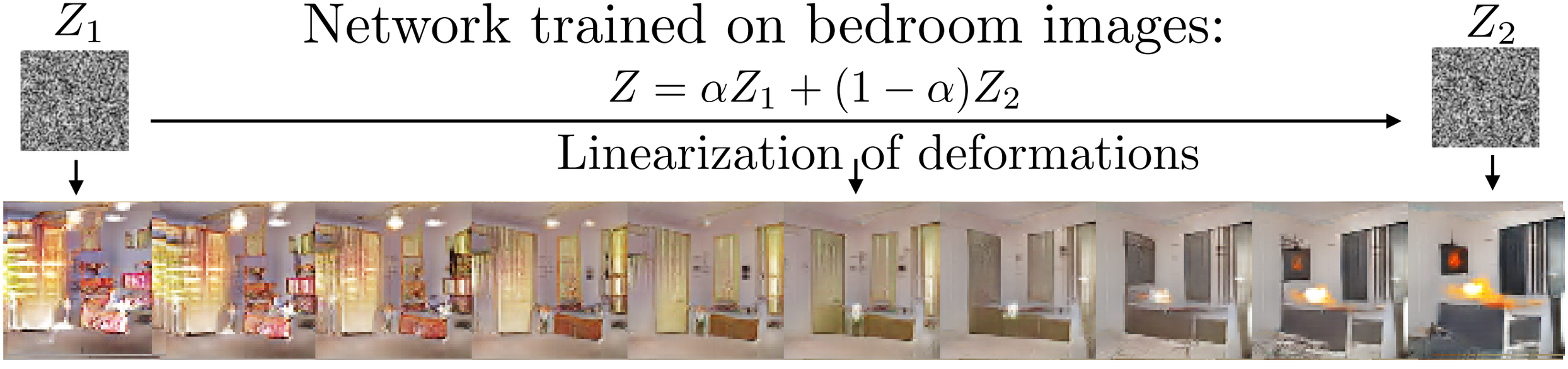


- **Multifractal properties:** $E[\,|X \star \psi|^q\,] \sim 2^{j\zeta(q)}$ reproduce high-order moments
- **Probability distribution:** $P(|x|)$
- **Leverage correlation:** $L(\tau) = E\left[\,|X(t+\tau)|^2 X(t)\right]$: time asymmetry

Learned Generative Networks

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





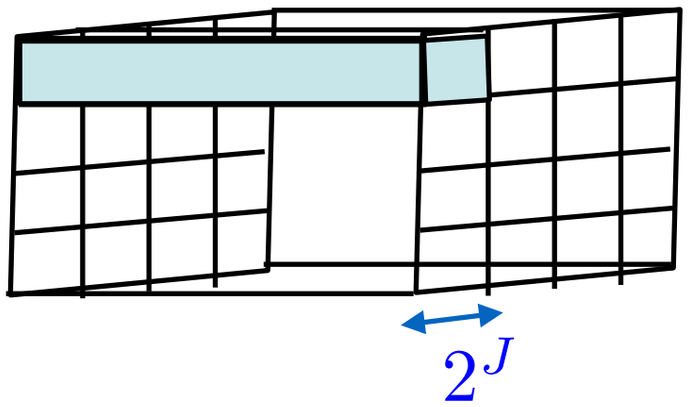
- Encoder Lipschitz continuous to actions of deformations

How to build such auto encoders ?

Averaged Rectified Wavelets

Spatial averaging at a large scale 2^J:

$$Ux \star \phi_J = \begin{pmatrix} x \star \phi_{2^J}(2^J n) \\ \rho(x \star \psi_{\alpha,\lambda}) \star \phi_J(2^J n) \end{pmatrix}_{\alpha,\lambda}$$



$Ux \star \phi_J$

Scale separation and spatial averaging with ϕ_J:

- Gaussianization
- Linearize small deformations

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

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

Multiscale Autoencoder

Tomas Angles

- **Encoder:** convolutional network

$$x \xrightarrow{d = 10^4} Ux \star \phi_J \xrightarrow{Id - Pr} Ix \xrightarrow[\text{Innovation}]{L} \underset{\text{White noise} \approx \text{Gaussian}}{Z} \quad d' = 10^2$$

- Innovations: prediction errors are decorrelated across scales
- Spatial decorrelation and dimension reduction

- **Generator:** sparse deconvolution

$$Z \xrightarrow[\text{pseudo-inverse}]{L^{-1}} Ix + \epsilon \xrightarrow{(Id - Pr)^{-1}} Ux \star \phi_J + \epsilon' \xrightarrow[\text{CNN}]{\textit{Deconvolution}} \widetilde{U}x \xrightarrow{U^{-1}} \widetilde{x}$$

Non-linear Dictionary model

Progressive Sparse Deconvolution

Tomas Angles

- Progressive sparse deconvolution of $x \star \phi_j$ for j decreasing.

- Learns a dictionary D_j where $Ux \star \phi_j$ is sparse

the CNN computes a sparse code α so that:

$$Ux \star \phi_j + \epsilon' = D_j \, \alpha$$

The CNN is learned jointly with D_j

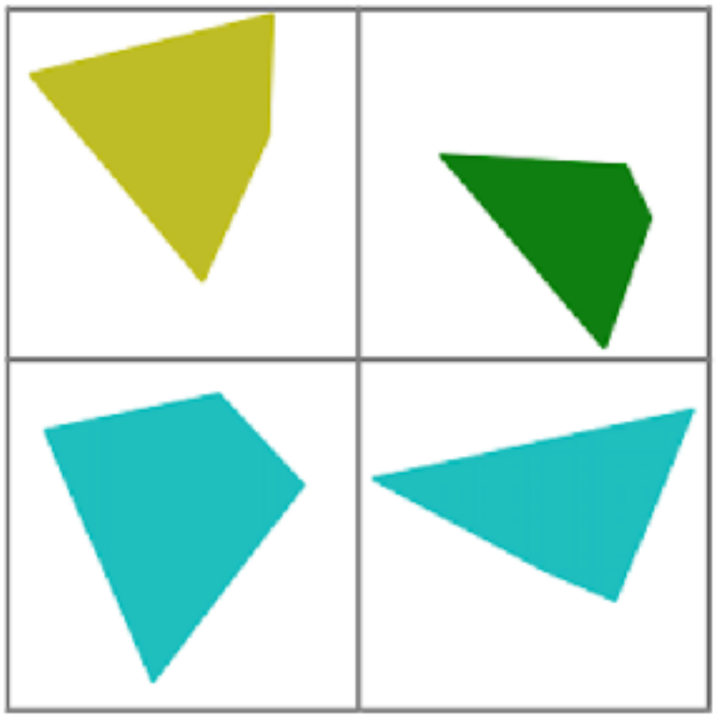
by minimising the average error $\|\epsilon'\|^2$ over a data basis.

What sparse code is computed by the CNN ?

Could it be an $\mathbf{l^1}$ sparse code ?

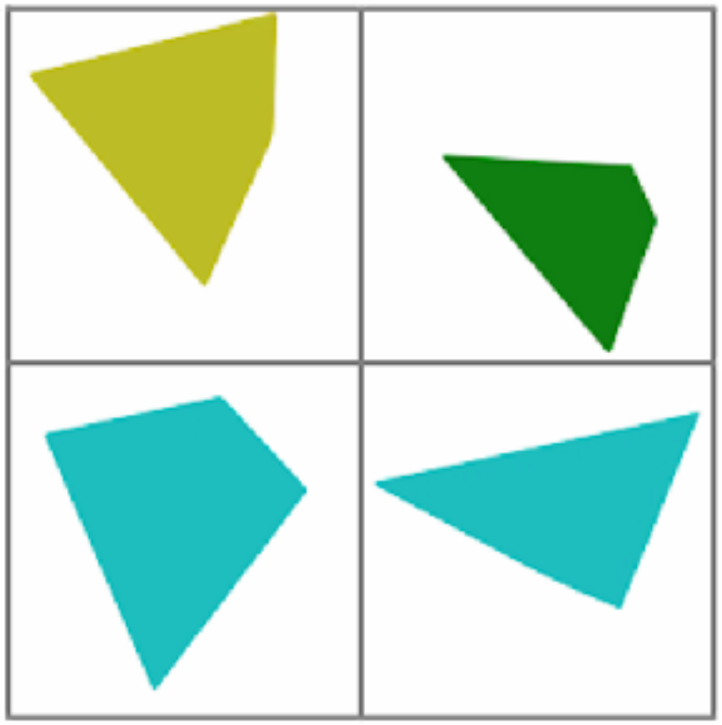
Training Reconstruction

Training x_i
Polygones

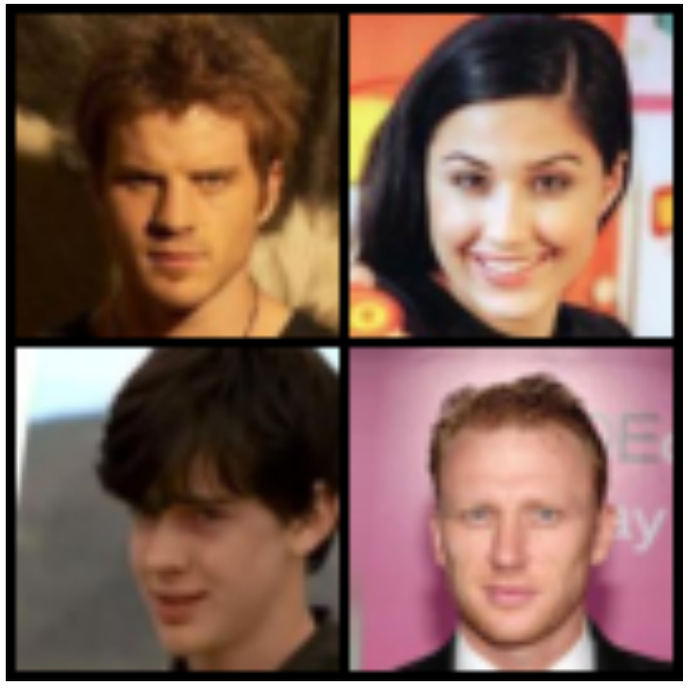


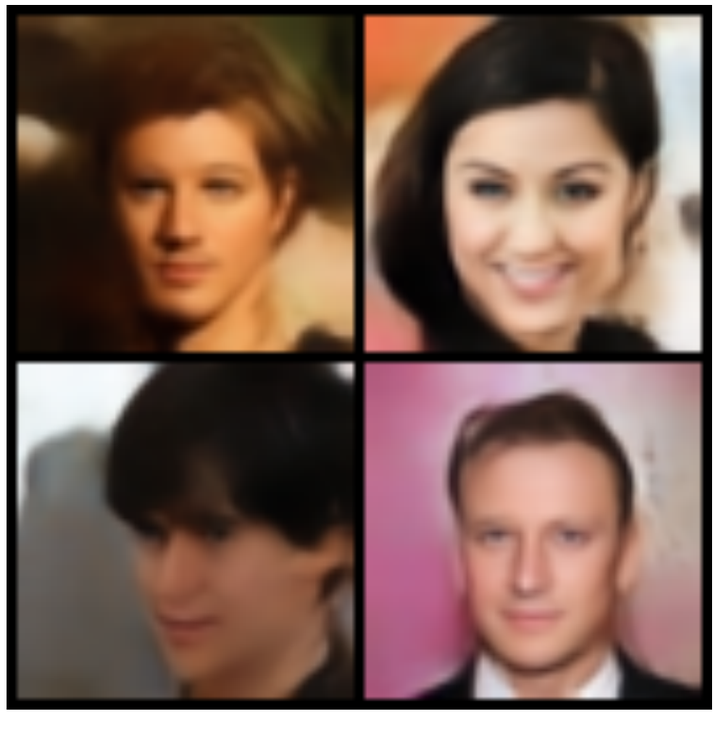
$2^J = 16$

$G(S_J(x_i))$ *Tomas Angles*



Celebrities Data Basis

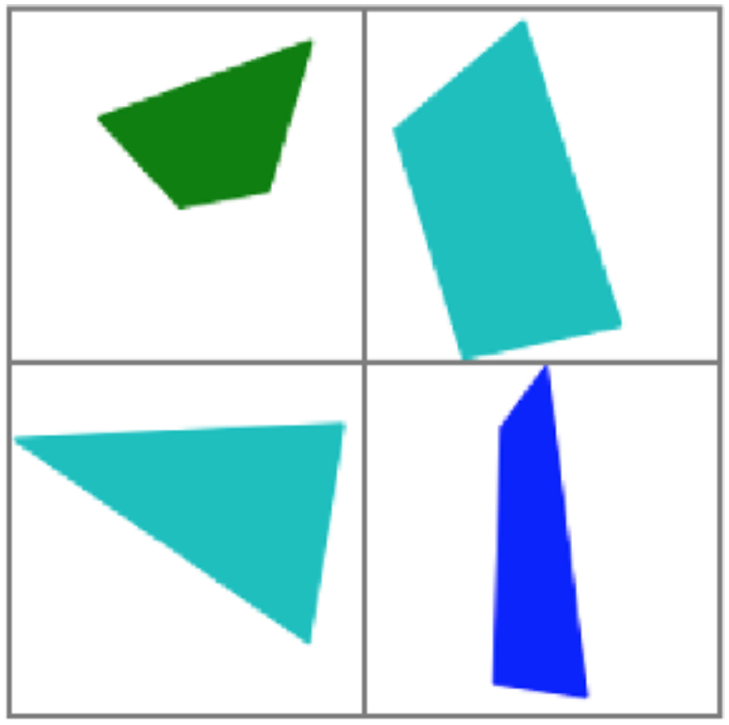




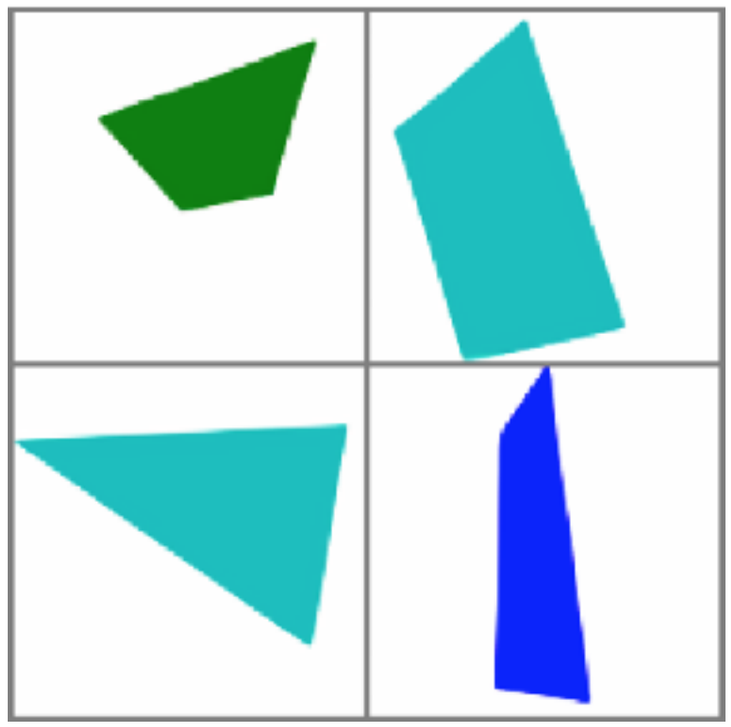
Testing Reconstruction

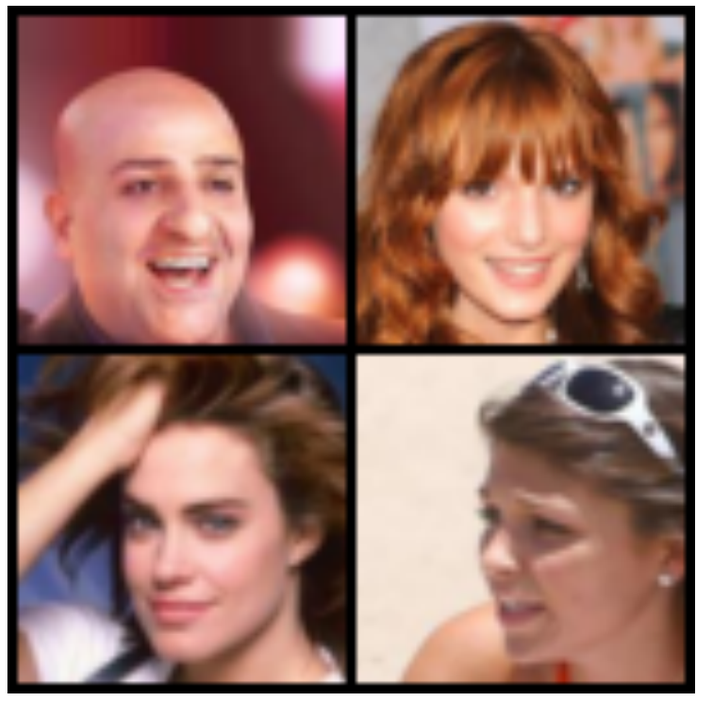
Tomas Angles

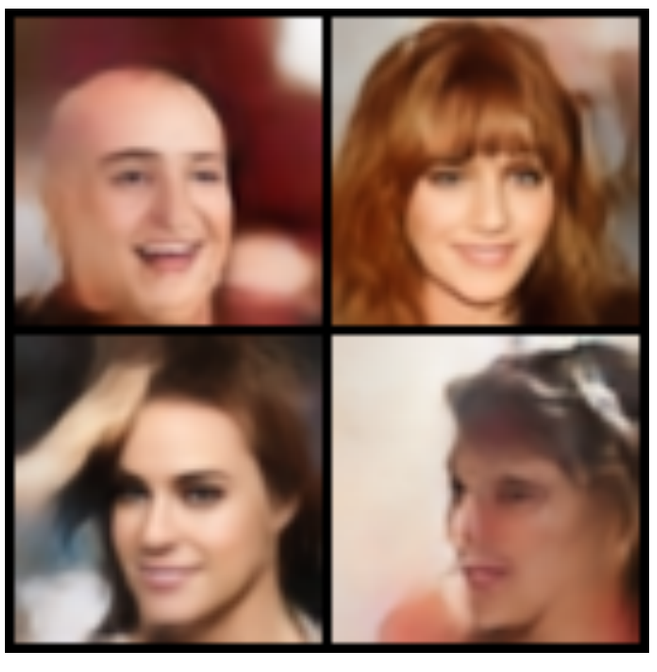
Testing
x_t



$G(S_J(x_t))$





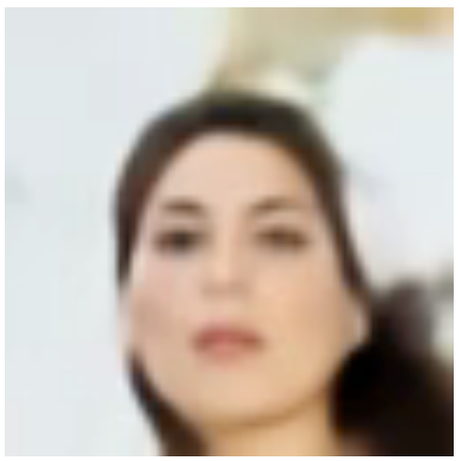


Generative Interpolations

Tomás Angles

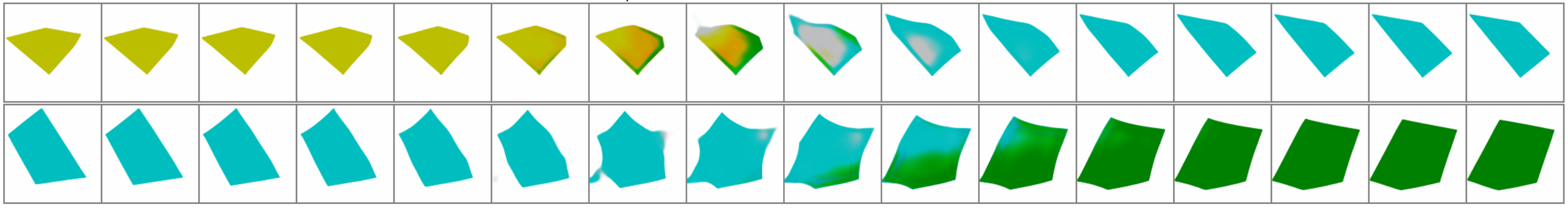
Polygons

Celebrities



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

Z_1 $\downarrow G$ — Z $\downarrow G$ — Z_2 $\downarrow G$



Random Sampling

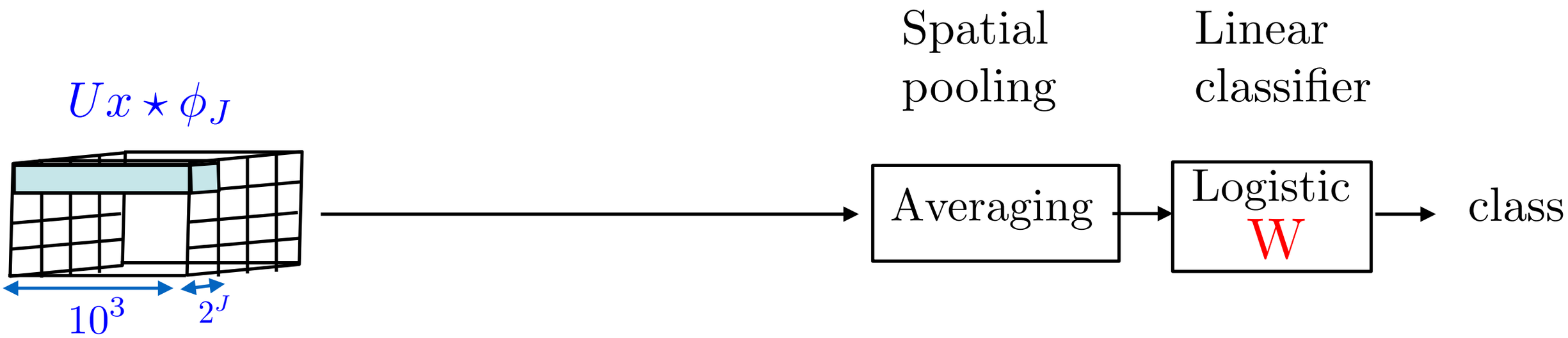
Tomas Angles

- Images synthesised from a Gaussian white noise

Classification by Dictionary Learning

Louis Thiry, John Zarka

1000 classes, 1.2 million labeled training images, of 224×224 pixels



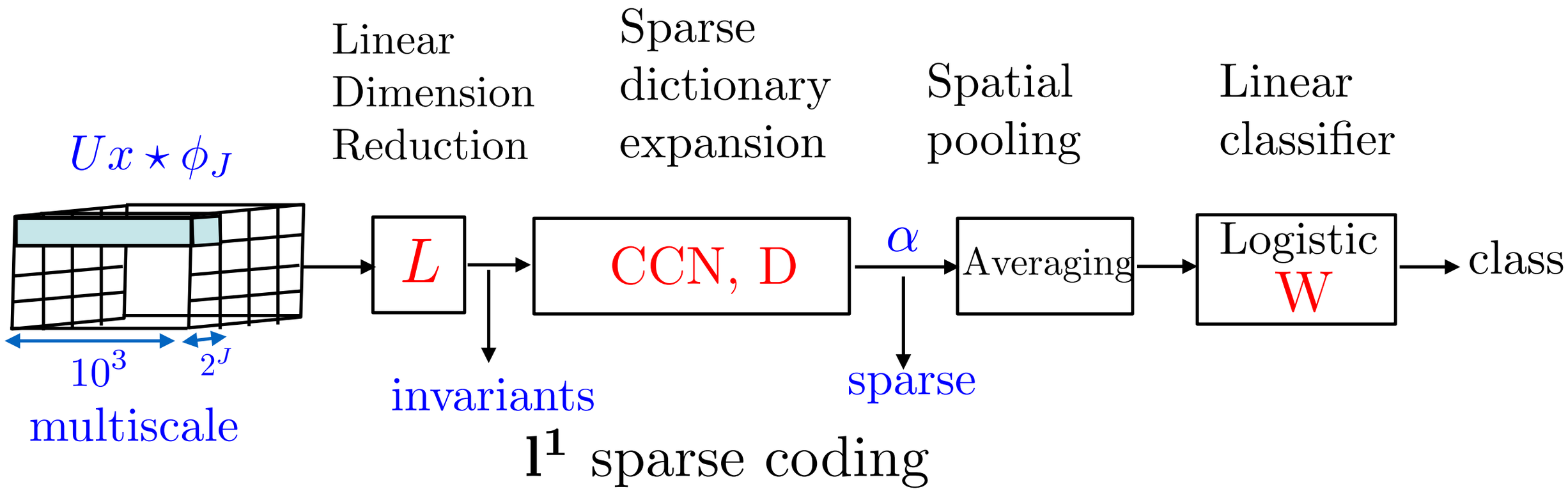
Phase Harmonic

	Alex-Net	Wavelets
Top 5 error	20%	70%

Classification by Dictionary Learning

Louis Thiry, John Zarka

1000 classes, 1.2 million labeled training images, of 224×224 pixels



l1 Sparse Coding: LISTA

- $\mathbf{l^1}$ sparse coefficients in a convolutional dictionary D

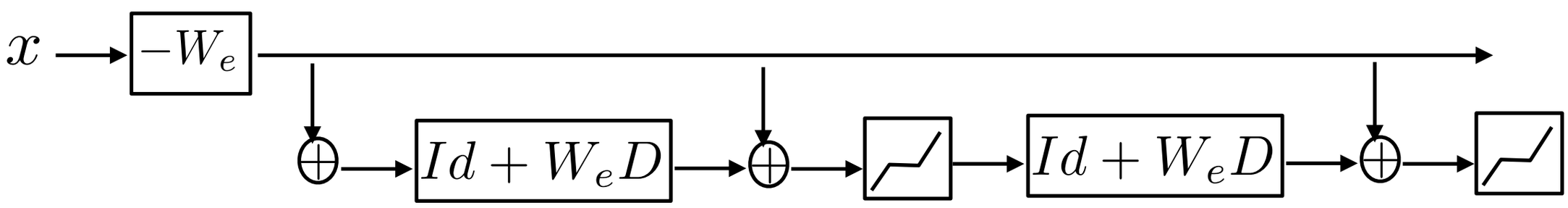
$$\widetilde{\alpha} = \arg\min_z \|x - D\alpha\|^2 + \gamma \|\alpha\|_1$$

Gregor & LeCun

- With a deep neural network implemented with D and W_e:

$$\alpha_{k+1} = \text{soft-thresh}(\alpha_k + W_e(D\alpha_k - x))$$

LISTA: CNN with soft-threshold non-linearity

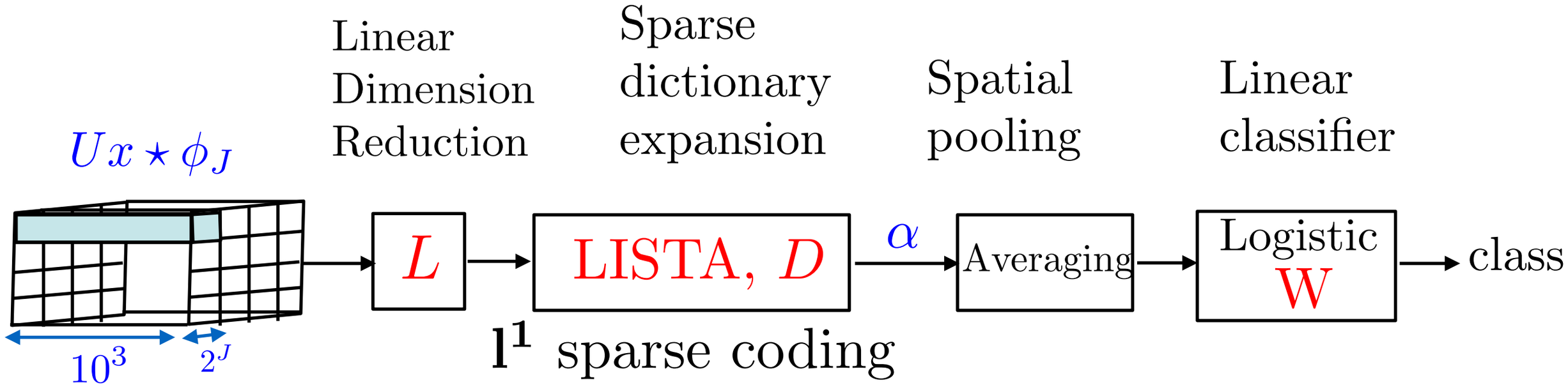


Can be used to learn the dictionary D *Gyries et. al*

Classification by Dictionary Learning

Louis Thiry, John Zarka

1000 classes, 1.2 million labeled training images, of 224×224 pixels



CNN architecture: convolutions, Relu and soft-thresholdings

end-to-end optimisation of L, D, W

D: sparse informative pattern across scales

	Alex-Net	Wavelets	Phase Harmonic Wavelets + Sparse
Top 5 error	20%	70%	30%

Multiscale Approximations

- A Relu on multiscale wavelet filters can produce scale interactions: creates phase harmonics

- Autoregressive models over multiscale phase harmonics approximate sparse signals and large classes of non-Gaussian and long range interaction processes

- Non-linear models based on sparse dictionaries may reproduce some CNN results for generation and classification

- Still need functional analysis models and approximation theorems with decay rates.