

M A D



NYU

COURANT INSTITUTE OF  
MATHEMATICAL SCIENCES

# MATHEMATICS OF DEEP LEARNING

---

JOAN BRUNA , CIMS + CDS, NYU, SPRING'18

*Lecture 3: The Scattering Transform and beyond*

# LECTURE 3 OVERVIEW

---

- The Scattering Transform.
- Stability Properties of Scattering Transforms
- Rigid Scattering
- From Scattering to CNNs.

# AVERAGE AND UNIQUENESS

---

- The only linear, translation-invariant operator is the average:

# AVERAGE AND UNIQUENESS

---

- The only linear, translation-invariant operator is the average:

$$\begin{aligned} \forall v, \Phi(x) = \Phi(\varphi_v x) \implies \Phi(x) &= \frac{1}{|G|} \int \Phi(\varphi_v x) dv \\ \implies \Phi(x) &= \Phi\left(\frac{1}{|G|} \int \varphi_v x dv\right) = \Phi\left(\frac{1}{|G|} \int x(u) du\right). \end{aligned}$$

- And a similar argument can be used locally.

# FROM AVERAGES TO WAVELETS

---

- Low-pass information is insufficient:

The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and



# FROM AVERAGES TO WAVELETS

---

- Low-pass information is insufficient:

The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and



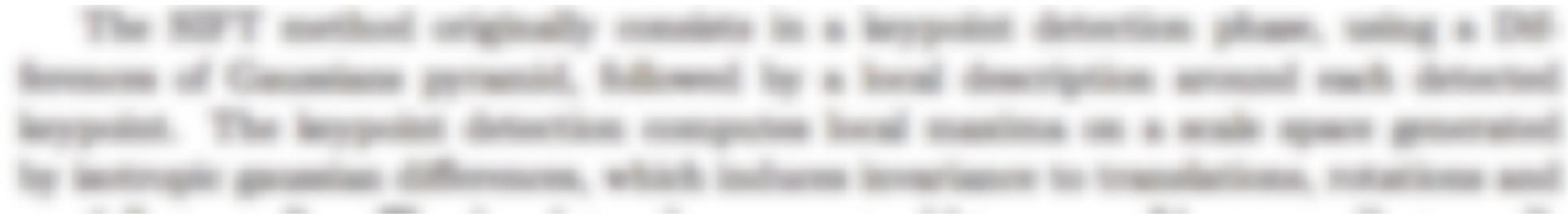
- Thus, we must capture high-frequency.
- These new measurements must involve a non-linearity.

# FROM AVERAGES TO WAVELETS

---

## ➤ Low-pass information is insufficient:

The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and



- Thus, we must capture high-frequency.
- These new measurements must involve a non-linearity.
- We want them to preserve stability to deformations.
- And we want them to preserve inter-class variability.

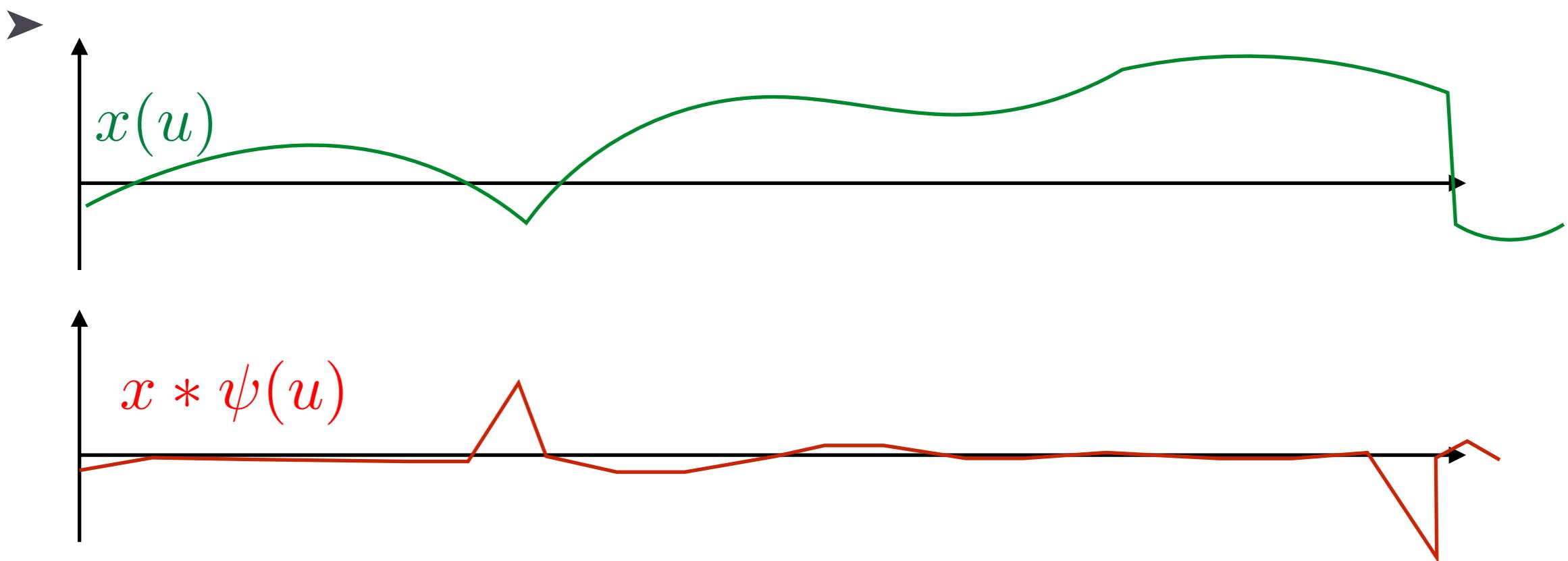
# WAVELETS

---

- $\psi$ : bandpass (ie oscillating) signal, well localized in space and frequency.
- At least one vanishing moment:  $\int \psi(u)du = 0$

(we say that  $\psi$  has  $k$  vanishing moments if  $\int \psi(u)u^l du = 0$  for  $l < k$ )

If  $x(u)$  is piece-wise smooth, then  $x * \psi(u)$  is mostly zero



# WAVELETS

---

- The local average  $x * \phi$  is a “blurry” version of  $x$ , whereas
- $x * \psi$  carries the details lost by the blurring.

# WAVELETS

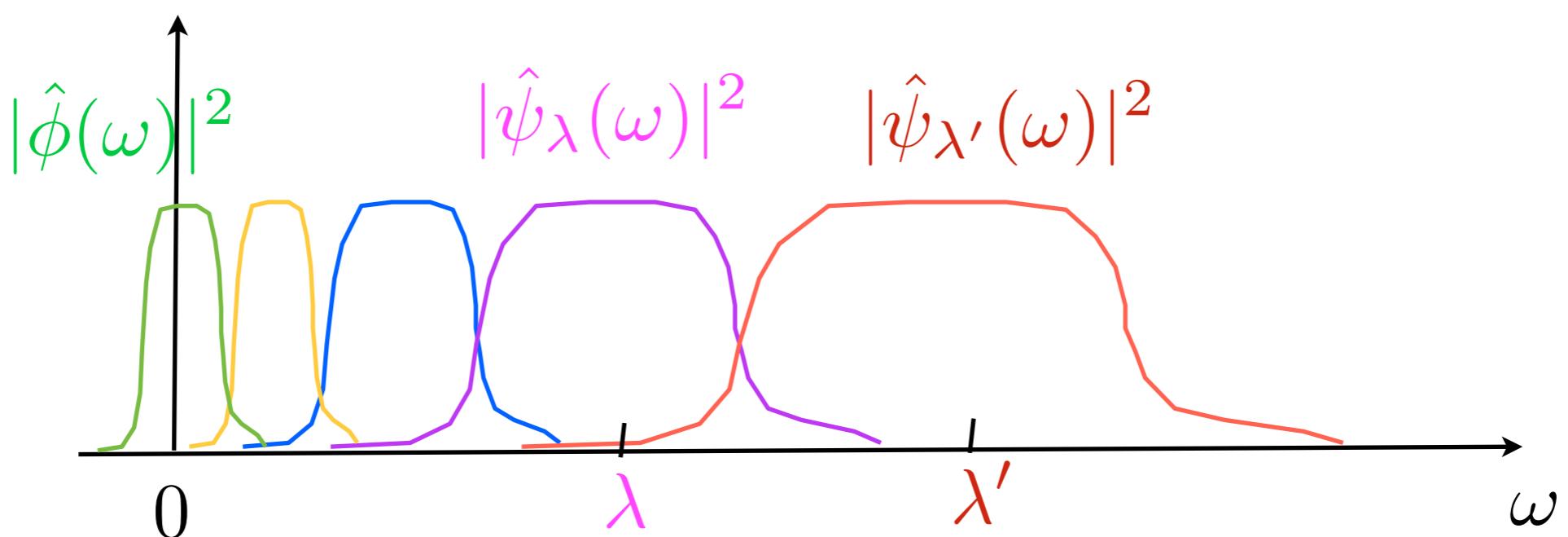
---

- The local average  $x * \phi$  is a “blurry” version of  $x$ , whereas
- $x * \psi$  carries the details lost by the blurring.
- The details are relative to a given resolution. How to obtain a decomposition that captures details at *all* resolutions?

# WAVELETS

---

- The local average  $x * \phi$  is a blurry version of  $x$ , whereas
  - $x * \psi$  carries the details lost by the blurring.
  - The details are relative to a given resolution. How to obtain a decomposition that captures details at *all* resolutions?
- 
- Dilated wavelets:  $\hat{\psi}_j(u) = 2^{-j}\psi(2^{-j}u)$ ,  $j \in \mathbb{Z}$

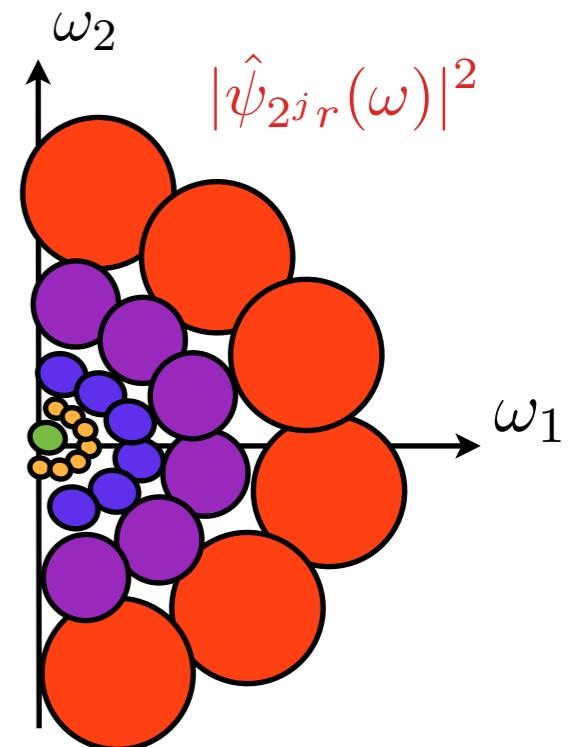


# LITTLEWOOD-PALEY WAVELET FILTER BANKS

---

- For images, dilated and rotated wavelets:

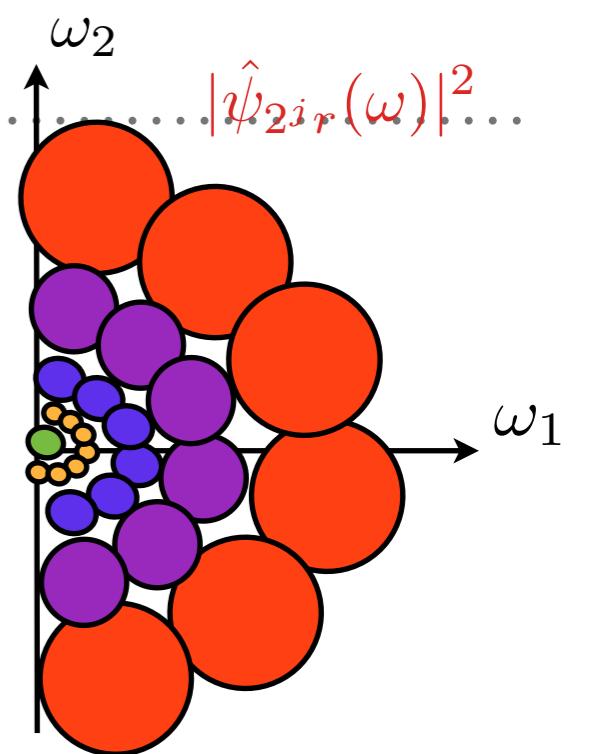
$$\psi_\lambda(u) = 2^{-j/2} \psi(2^{-j} r u) , \text{ with } \lambda = 2^j r$$



# LITTLEWOOD-PALEY WAVELET FILTER BANKS

- For images, dilated and rotated wavelets:

$$\psi_\lambda(u) = 2^{-j/2} \psi(2^{-j}ru) , \text{ with } \lambda = 2^j r$$



$$Wx = \{x \star \phi(u), x \star \psi_\lambda(u)\}_{\lambda \in \Lambda}$$

$$x \star \psi(u) = \int x(v) \psi(u - v) dv .$$

**Theorem** (Littlewood-Paley): If there exists  $\delta > 0$  such that

$$\forall \omega > 0 , 1 - \delta \leq |\hat{\phi}(\omega)|^2 + \frac{1}{2} \sum_{\lambda} |\hat{\psi}(\lambda^{-1}\omega)|^2 \leq 1 ,$$

then  $\forall x \in L^2 , (1 - \delta) \|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2$  .

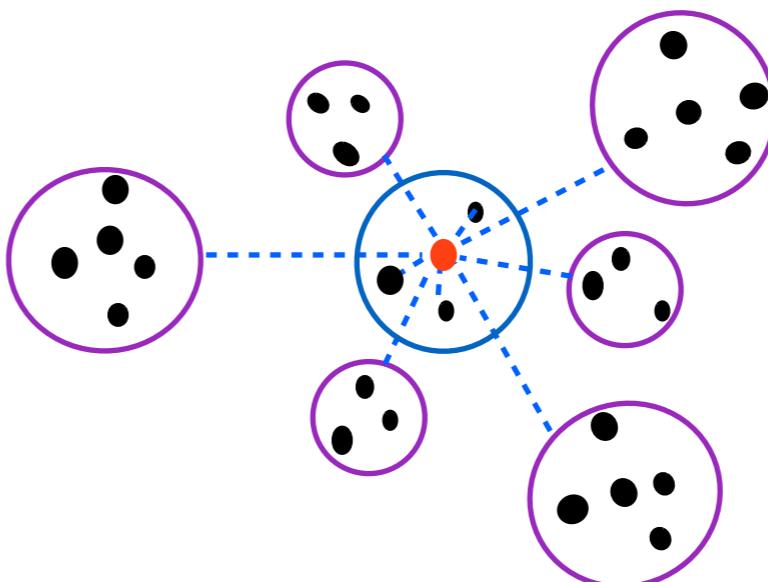
# WAVELETS AND SCALE SEPARATION

---

Interactions of  $d$  variables of  $X$ : pixels, particules, agents...

- Markov: each variable interacts only with its neighbours and ignores large scale interactions.

Factorisation  
into multiscale  
interactions



Multiscale regroupments reduce the number of interactions  
from  $d$  to  $O(\log d)$

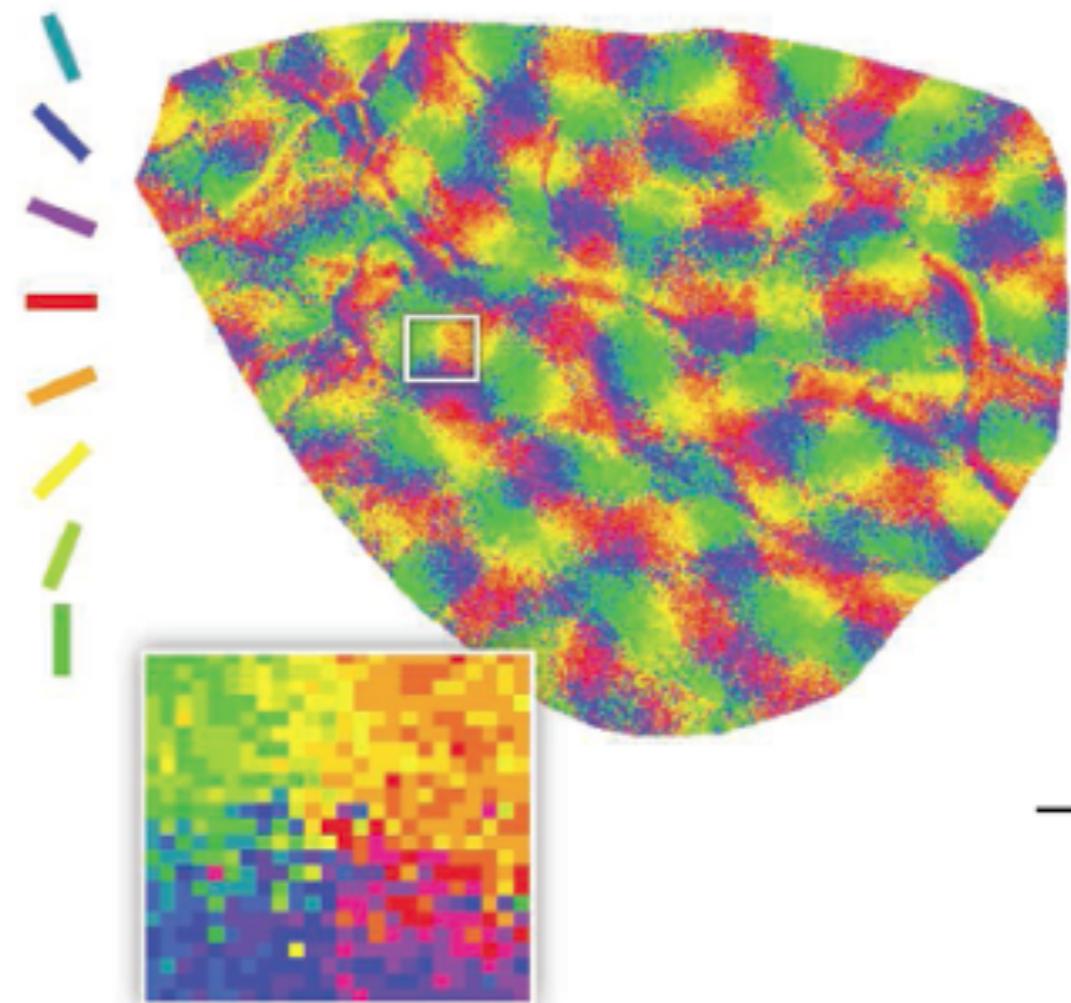
⇒ wavelet transforms and deep convolutional networks

See also the Fast Multipole Method [Greengard and Rokhlin]

# WAVELETS IN VISION

---

- *V1* Model of Simple and Complex cells: First layer of processing is selective in orientation, scale and position.



- cells are organized in *pinwheels*. (more on that later).

# WAVELETS AND DEFORMATIONS

---

- We saw before that a blurring kernel is nearly invariant to deformations:

**Proposition:** The local averaging  $\Phi(x) = x * \phi_J$  satisfies  
 $\forall \|x\| = 1 \in L^2, \tau, \|\Phi(x) - \Phi(\varphi_\tau x)\| \leq C\|\tau\|.$

- What about the wavelet operator  $\Phi(x) = \{x * \psi_\lambda\}_\lambda$ ?

# WAVELETS AND DEFORMATIONS

---

- We saw before that a blurring kernel is nearly invariant to deformations:

**Proposition:** The local averaging  $\Phi(x) = x * \phi_J$  satisfies  
 $\forall \|x\| = 1 \in L^2, \tau, \|\Phi(x) - \Phi(\varphi_\tau x)\| \leq C\|\tau\|.$

- What about the wavelet operator  $\Phi(x) = \{x * \psi_\lambda\}_\lambda$ ?
  - We don't have local invariance, but we have a form of local equivariance:

**Proposition [Mallat]:** For each  $\delta > 0$  there exists  $C > 0$  such that for all  $J$  and all  $\tau \in C^2$  with  $\|\nabla \tau\|_\infty \leq 1 - \delta$  we have

$$\|W_J \varphi_\tau - \varphi_\tau W_J\| \leq C(J\|\nabla \tau\|_\infty + \|H\tau\|_\infty).$$

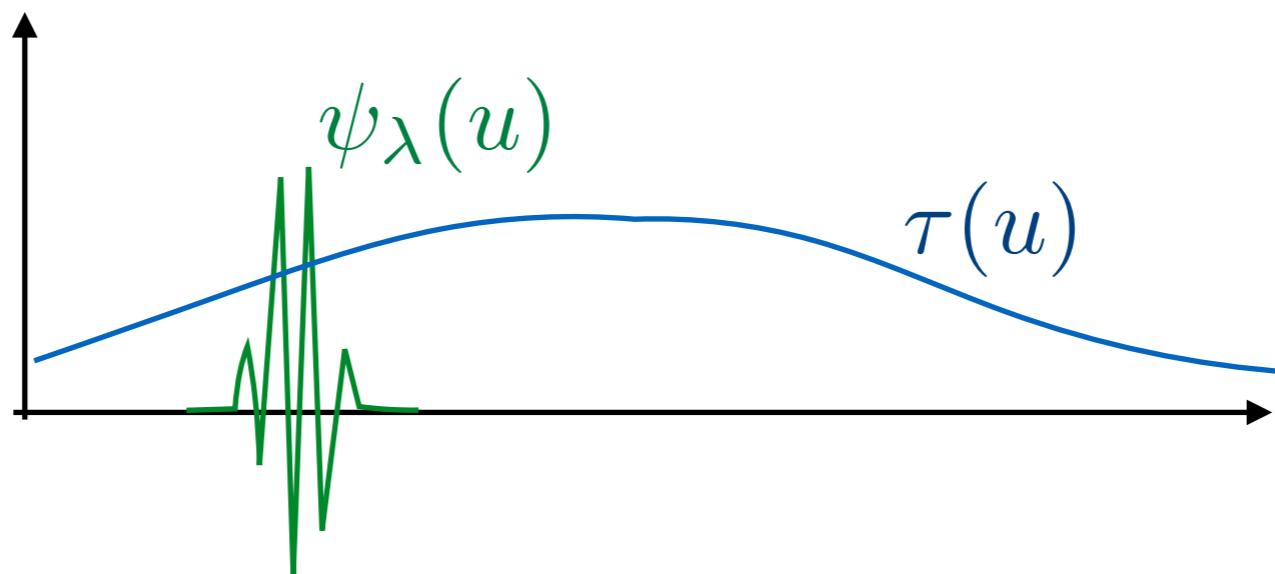
( $H\tau$ : Hessian of  $\tau$ )

# WAVELETS AND DEFORMATIONS

---

► Qualitative idea behind this result:

Each  $\psi_\lambda$  only “sees” the part of the deformation  $\tau$  that intersects its support.

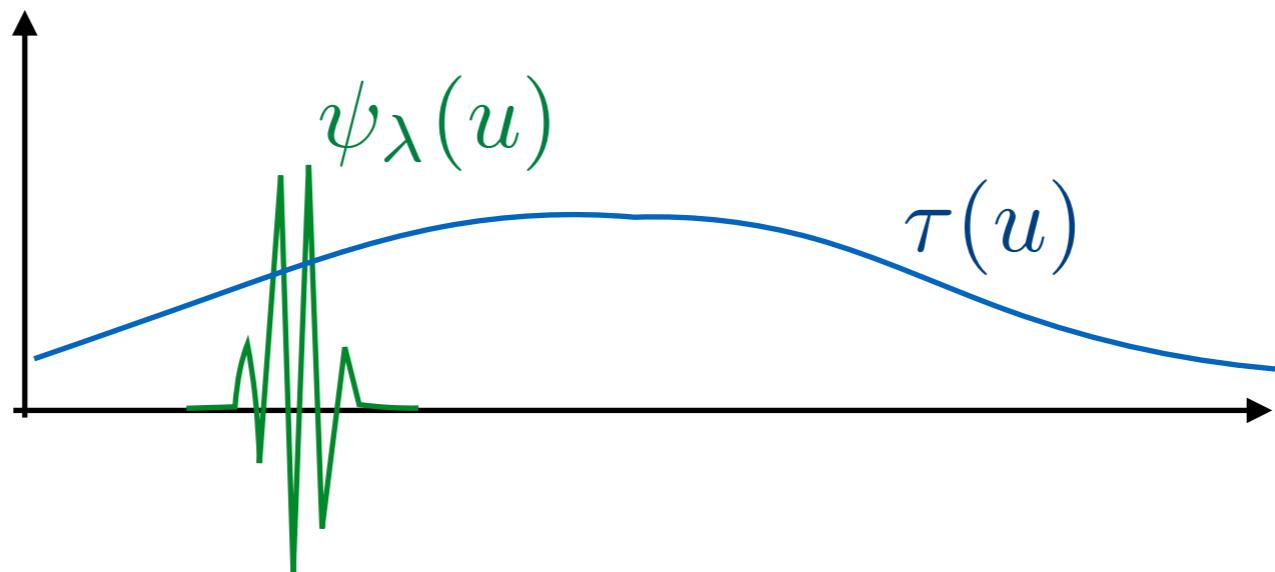


# WAVELETS AND DEFORMATIONS

---

► Qualitative idea behind this result:

Each  $\psi_\lambda$  only “sees” the part of the deformation  $\tau$  that intersects its support.



For small scales,  $\psi_\lambda$  has small support, and for  $u, v$  within that support, because  $\tau$  is smooth,  $|\tau(v) - \tau(u)| \sim 2^{-j} |\nabla \tau|_\infty$ .

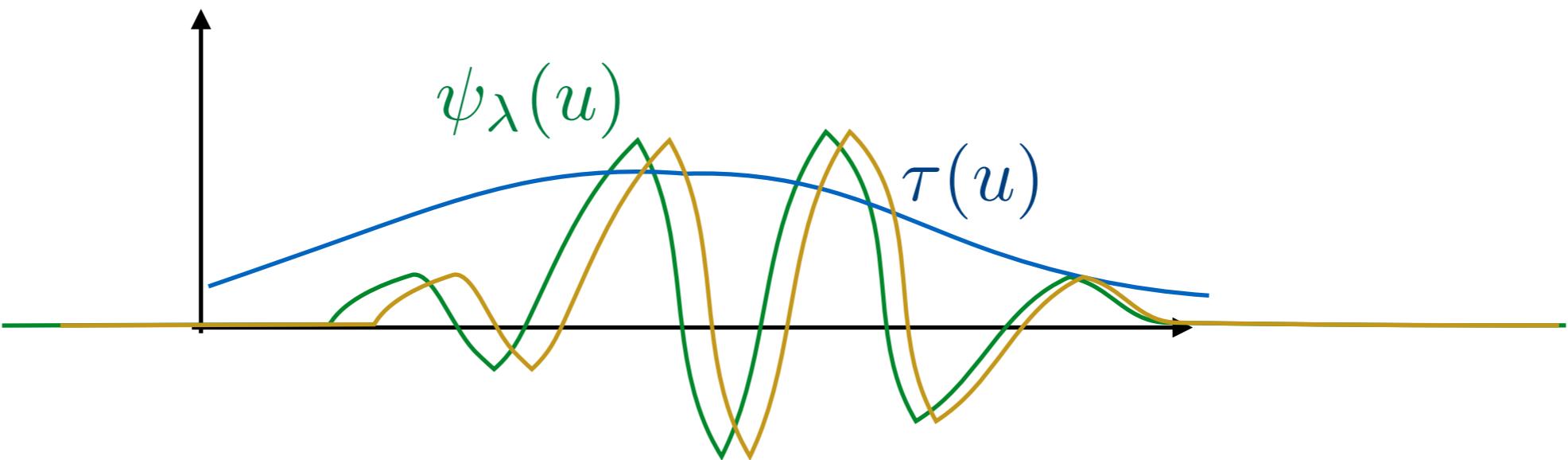
Thus  $|(\varphi_\tau x) * \psi_\lambda(u) - x * \psi_\lambda(u - \tau(u))| \sim |\nabla \tau|_\infty$ .

# WAVELETS AND DEFORMATIONS

---

► Qualitative idea behind this result:

Each  $\psi_\lambda$  only “sees” the part of the deformation  $\tau$  that intersects its support.

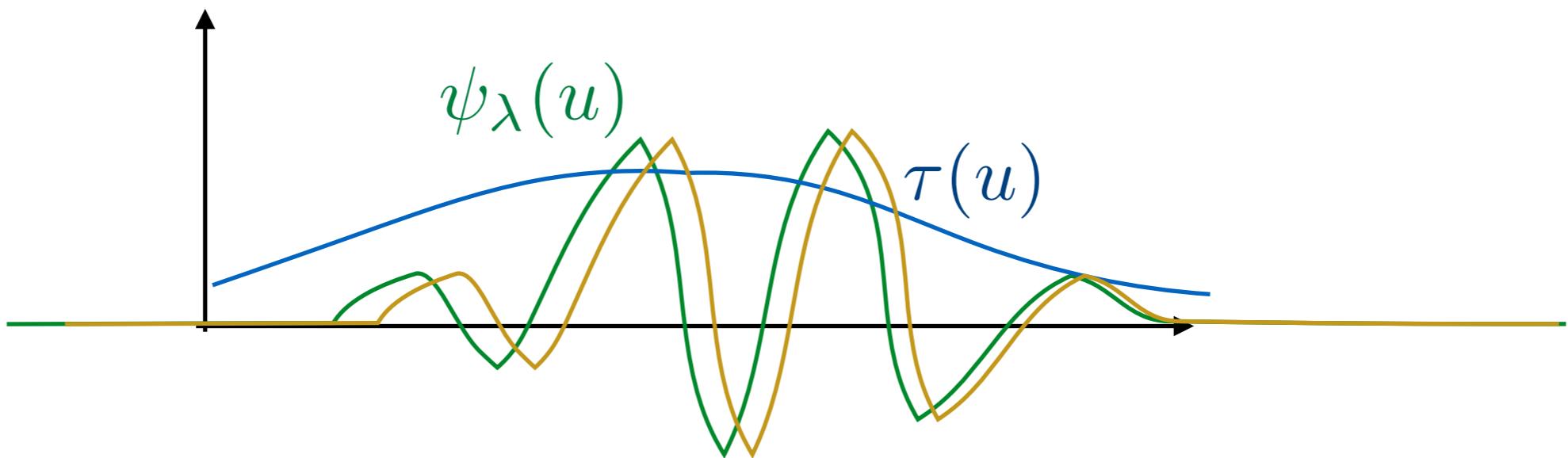


# WAVELETS AND DEFORMATIONS

---

► Qualitative idea behind this result:

Each  $\psi_\lambda$  only “sees” the part of the deformation  $\tau$  that intersects its support.



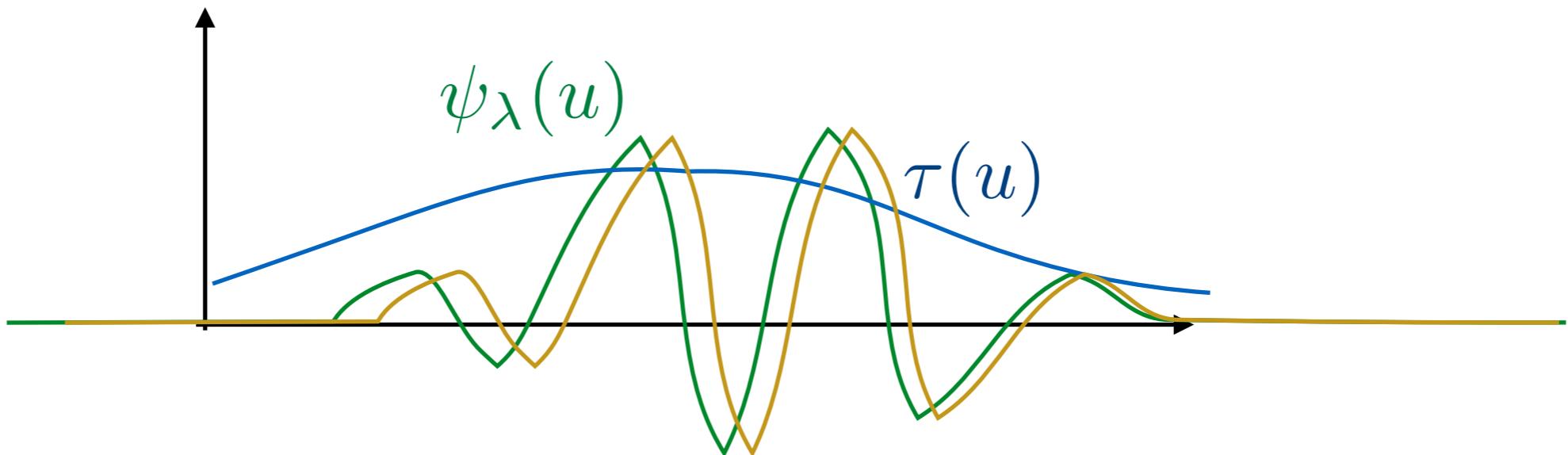
For large scales,  $\psi_\lambda$  is itself smooth, thus  
 $|\varphi_\tau(x * \psi_\lambda) - (\varphi_\tau x) * \psi_\lambda| \sim \|\nabla \tau\|_\infty$ .

# WAVELETS AND DEFORMATIONS

---

► Qualitative idea behind this result:

Each  $\psi_\lambda$  only “sees” the part of the deformation  $\tau$  that intersects its support.



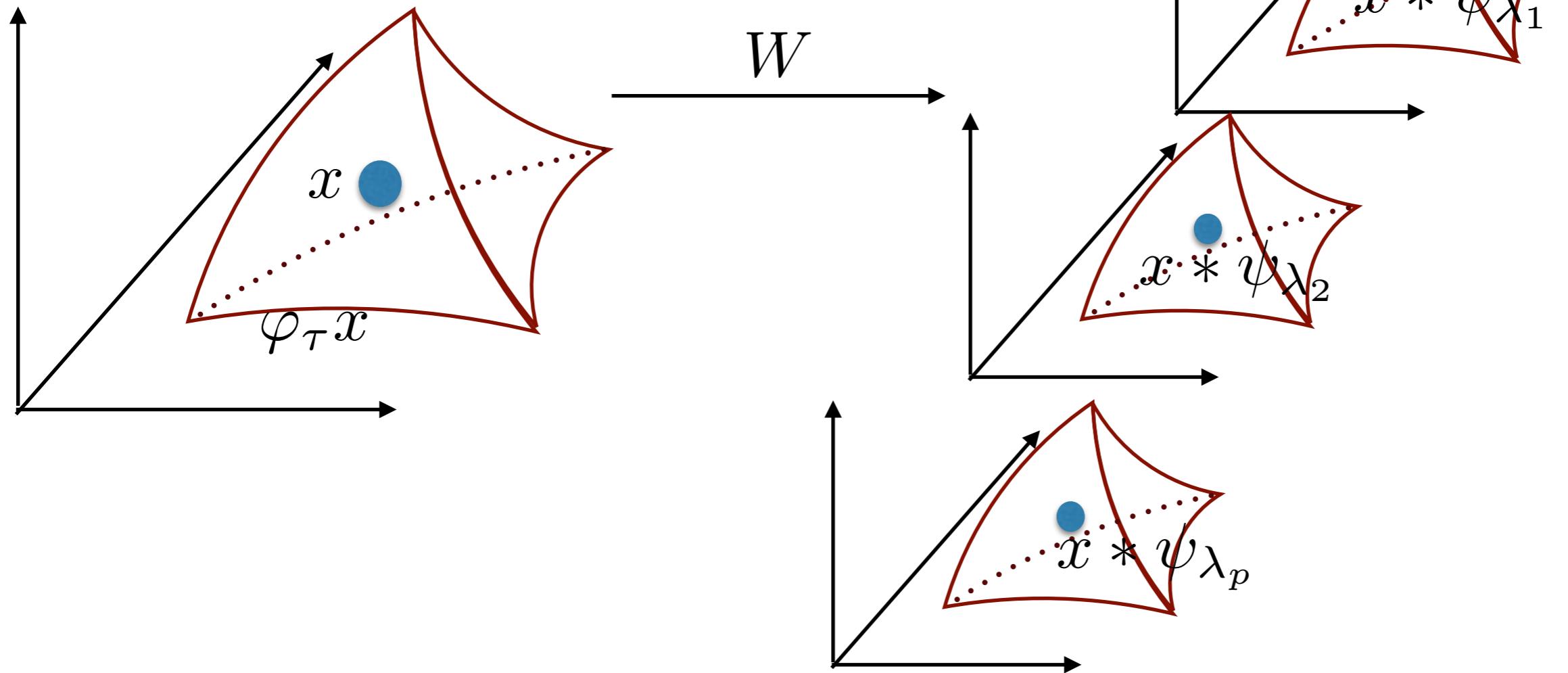
For large scales,  $\psi_\lambda$  is itself smooth, thus  
 $|\varphi_\tau(x * \psi_\lambda) - (\varphi_\tau x) * \psi_\lambda| \sim \|\nabla \tau\|_\infty$ .

And, most importantly, wavelet separates scales  
(so errors do not accumulate)

# WAVELETS AND NON-LINEARITIES

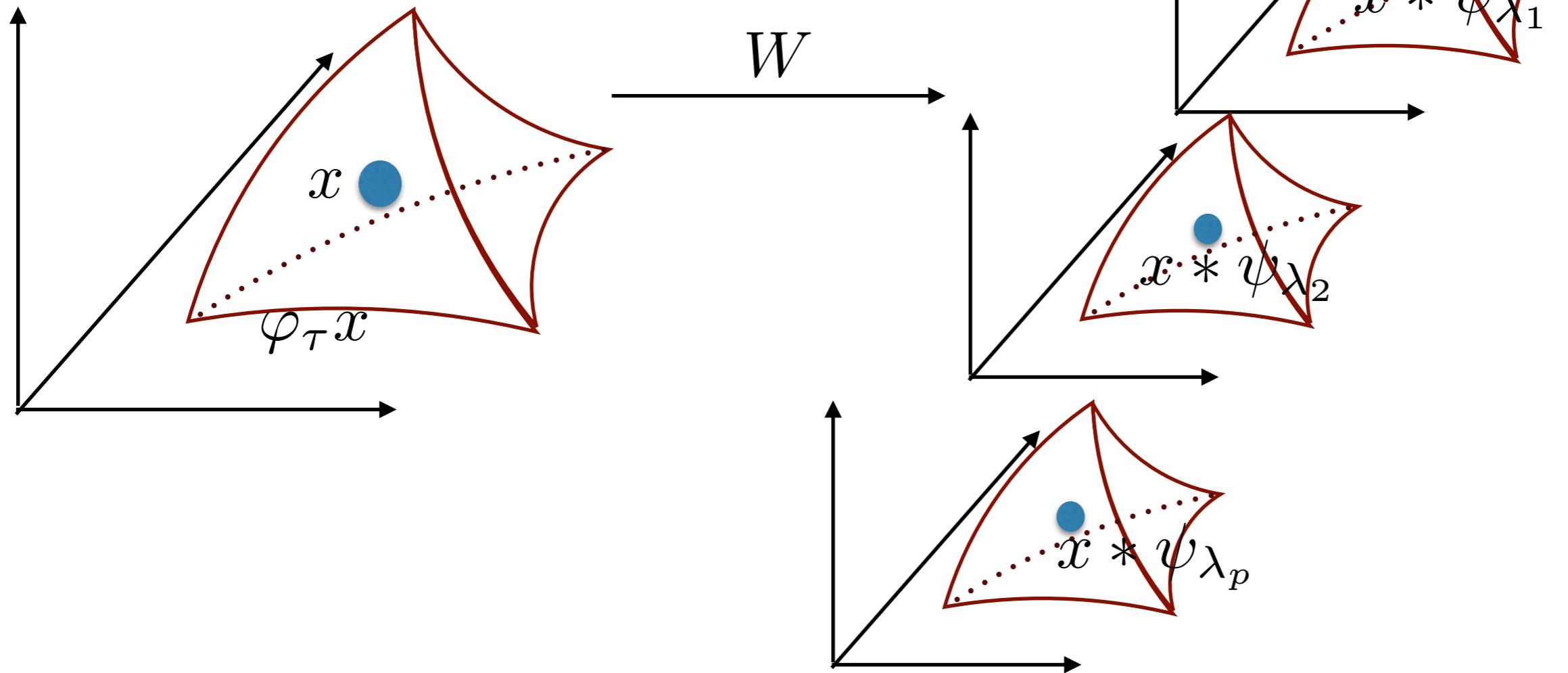
---

- The commutation property says that deformations in the input are approximately mapped to deformations in the wavelet domain:



# WAVELETS AND NON-LINEARITIES

- The commutation property says that deformations in the input are approximately mapped to deformations in the wavelet domain:



- We want to extract again stable measurements: *need non-linear operator.*

# CHARACTERIZATION OF STABLE NON-LINEARITIES

---

► Preserve additive stability:

$$\|Mx - Mx'\| \leq \|x - x'\| . \quad M \text{ non-expansive} .$$

# CHARACTERIZATION OF STABLE NON-LINEARITIES

---

- Preserve additive stability:

$$\|Mx - Mx'\| \leq \|x - x'\| . \quad M \text{ non-expansive} .$$

- Preserve geometric stability: It is sufficient to commute with diffeomorphisms.

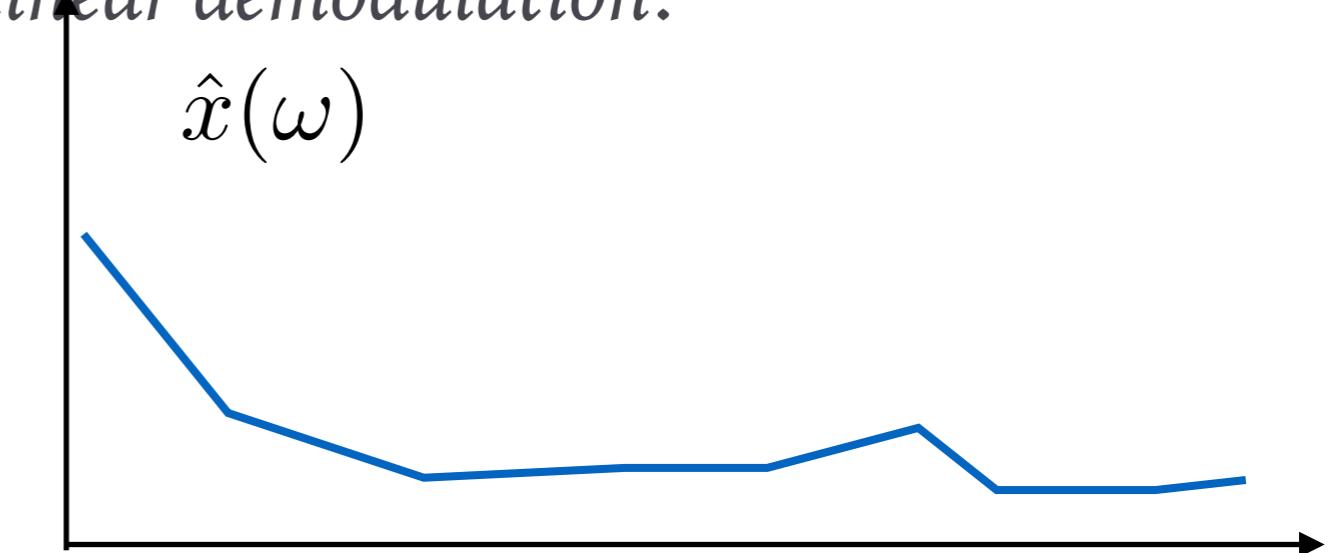
**Theorem:** If  $M$  is non-expansive operator in  $L^2$  such that  $\varphi_\tau M = M\varphi_\tau$  for all  $\tau$ , then  $M$  is point-wise:

$$Mx(u) = \rho(x(u)) .$$

# UNDERSTANDING THE EFFECT OF NONLINEARITIES

---

► Rectifiers thus perform a *non-linear demodulation*:

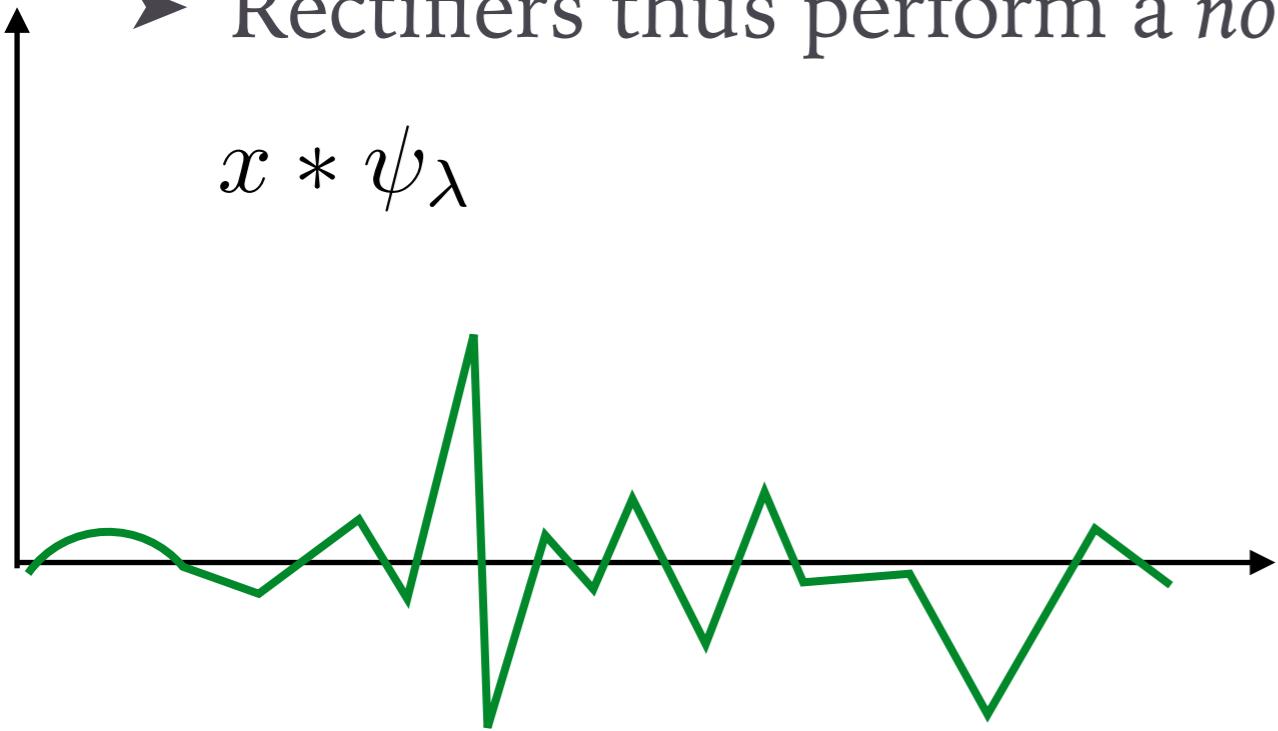


# UNDERSTANDING THE EFFECT OF NONLINEARITIES

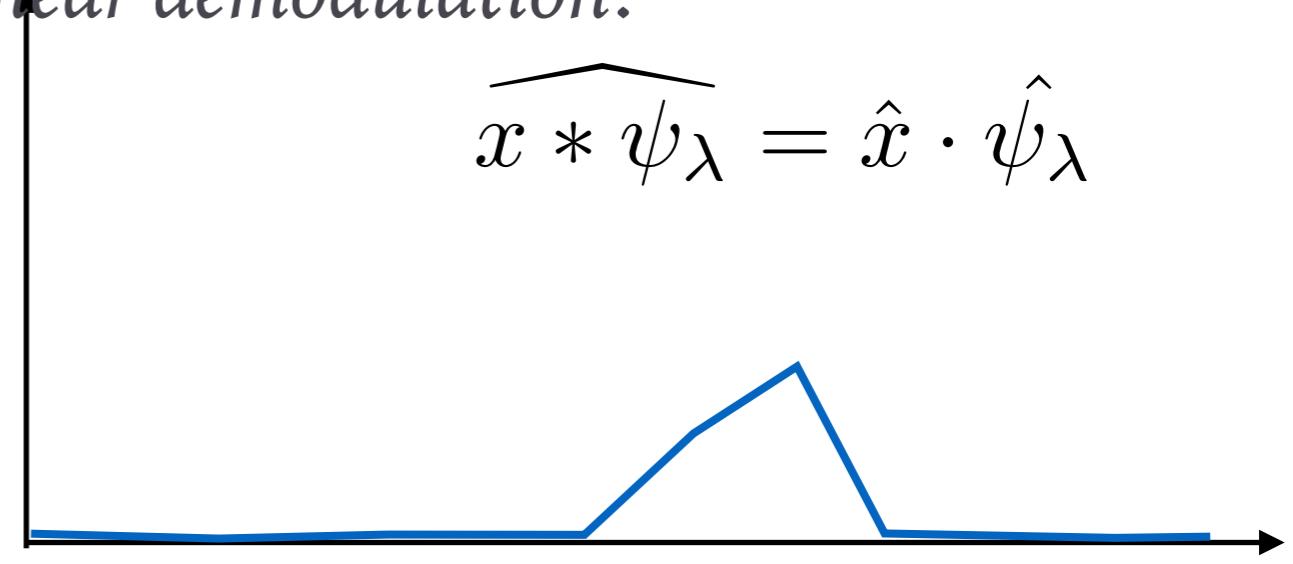
---

- Rectifiers thus perform a *non-linear demodulation*:

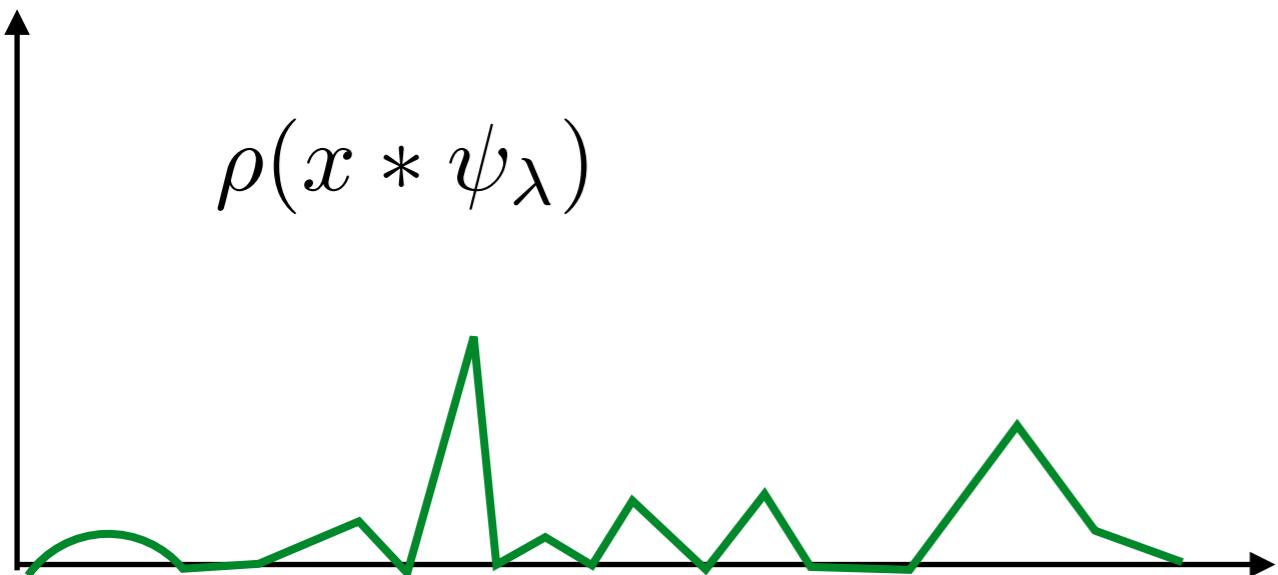
$$x * \psi_\lambda$$



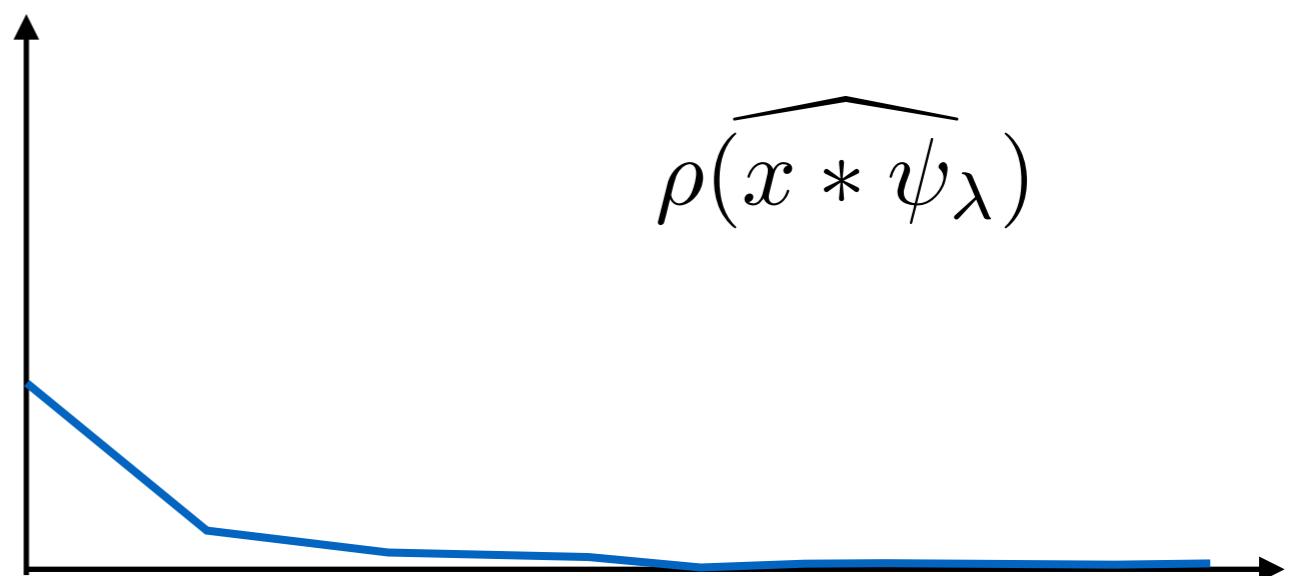
$$\widehat{x * \psi_\lambda} = \hat{x} \cdot \hat{\psi}_\lambda$$



$$\rho(x * \psi_\lambda)$$



$$\rho(\widehat{x * \psi_\lambda})$$



sometimes called the envelope

# CHOICE OF POINTWISE NONLINEARITY

---

- Full rectification  $\rho(z) = |z|$  preserves energy:
  - When the wavelet is complex, it produces smoother envelopes (thus more stable features).
- Half rectification (ReLU)  $\rho(z) = \max(z, 0)$  captures half the energy, and it also creates *sparsity*.
  - We will see that this is important to perform *detection*.
- Sigmoid nonlinearity  $\rho(z) = (1 + e^{-z})^{-1}$ .
  - It is not homogeneous
  - Saturating regimes are problematic for learning via back propagation in deep models.
- Other: “Leaky” ReLU [MSR’14]: parametrized half-rectifier, ELU , etc.

# SEPARABLE SCATTERING OPERATORS

---

- Local averaging kernel:  $x \star \phi_J$ 
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.

# SEPARABLE SCATTERING OPERATORS

---

- Local averaging kernel:  $x \star \phi_J$ 
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.
- Recover lost information:  $\mathcal{U}_J(x) = \{x \star \phi_J, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J}$  .
  - Point-wise, non-expansive non-linearities: maintain stability.
  - Complex modulus maps energy towards low-frequencies.

# SEPARABLE SCATTERING OPERATORS

---

- Local averaging kernel:  $x \star \phi_J$ 
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.
- Recover lost information:  $\mathcal{U}_J(x) = \{x \star \phi_J, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J}$  .
  - Point-wise, non-expansive non-linearities: maintain stability.
  - Complex modulus maps energy towards low-frequencies.
- Cascade the *recovery* operator:  
$$\mathcal{U}_J^2(x) = \{x \star \phi_J, |x \star \psi_\lambda| \star \phi_J, ||x \star \psi_\lambda| \star \psi_{\lambda'}||\}_{\lambda, \lambda' \in \Lambda_J} .$$

# SEPARABLE SCATTERING OPERATORS

---

- Local averaging kernel:  $x \star \phi_J$ 
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.
- Recover lost information:  $\mathcal{U}_J(x) = \{x \star \phi_J, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J}$ .

- Point-wise, non-expansive non-linearities: maintain stability.
- Complex modulus maps energy towards low-frequencies.

- Cascade the “recovery” operator:

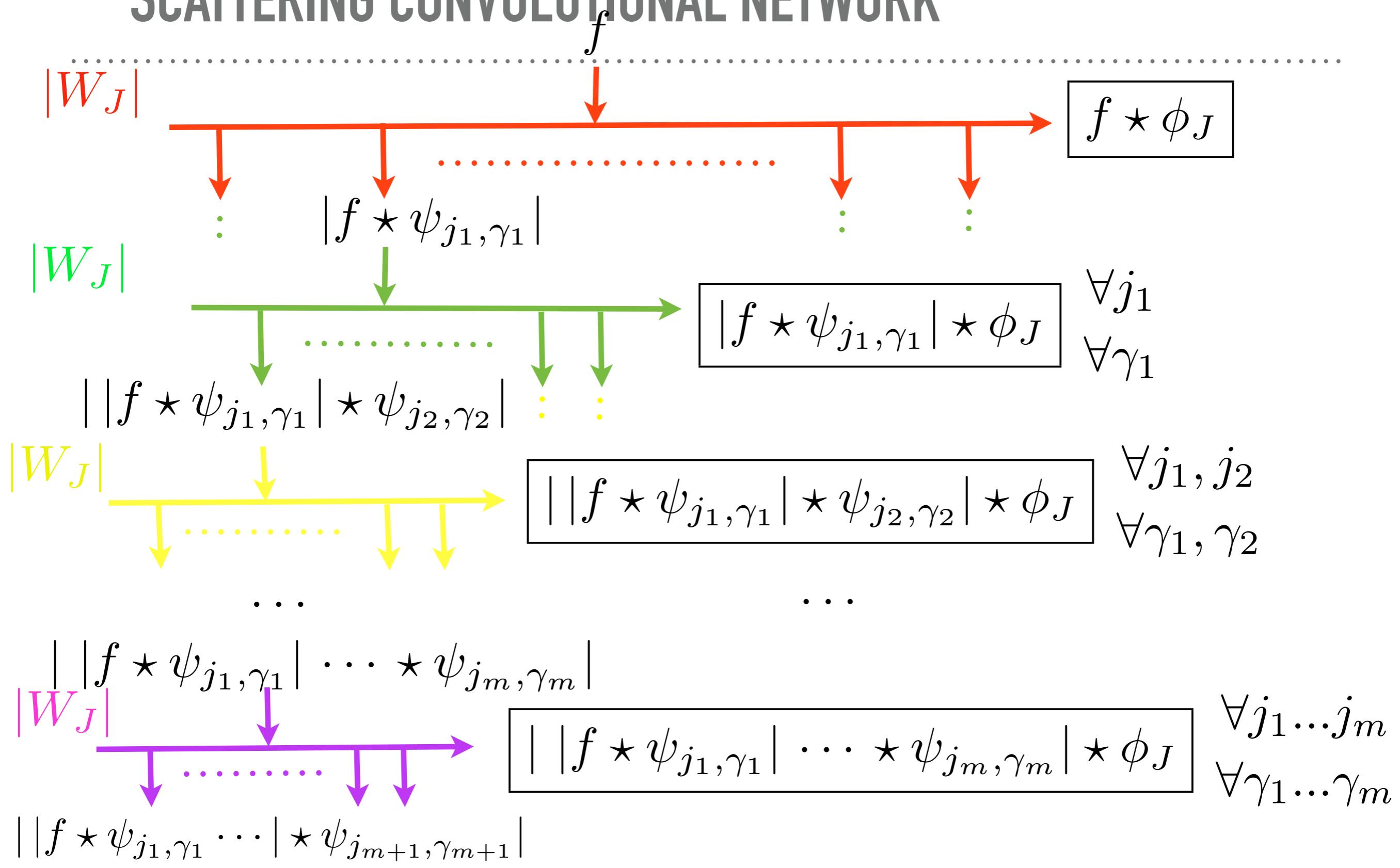
$$\mathcal{U}_J^2(x) = \{x \star \phi_J, |x \star \psi_\lambda| \star \phi_J, ||x \star \psi_\lambda| \star \psi_{\lambda'}||\}_{\lambda, \lambda' \in \Lambda_J}.$$

$p = (\lambda_1, \dots, \lambda_m)$  :

- Scattering coefficient along a path

$$S_J[p]x(u) = |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \dots | \star \psi_{\lambda_m}| \star \phi_J(u).$$

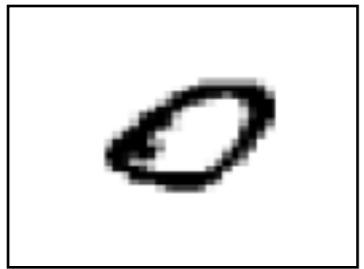
# SCATTERING CONVOLUTIONAL NETWORK



Cascade of contractive operators.

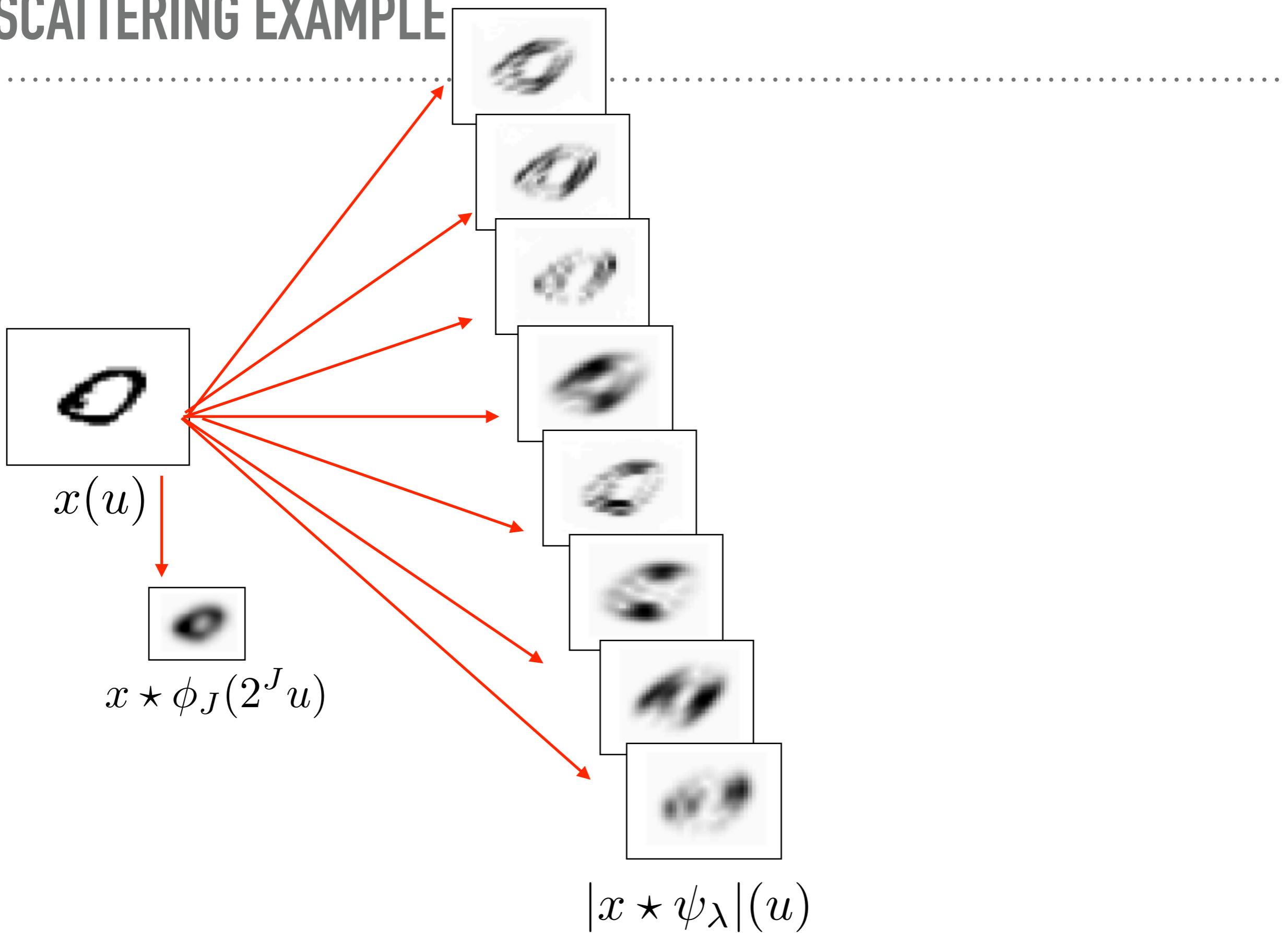
# SCATTERING EXAMPLE

---



$x(u)$

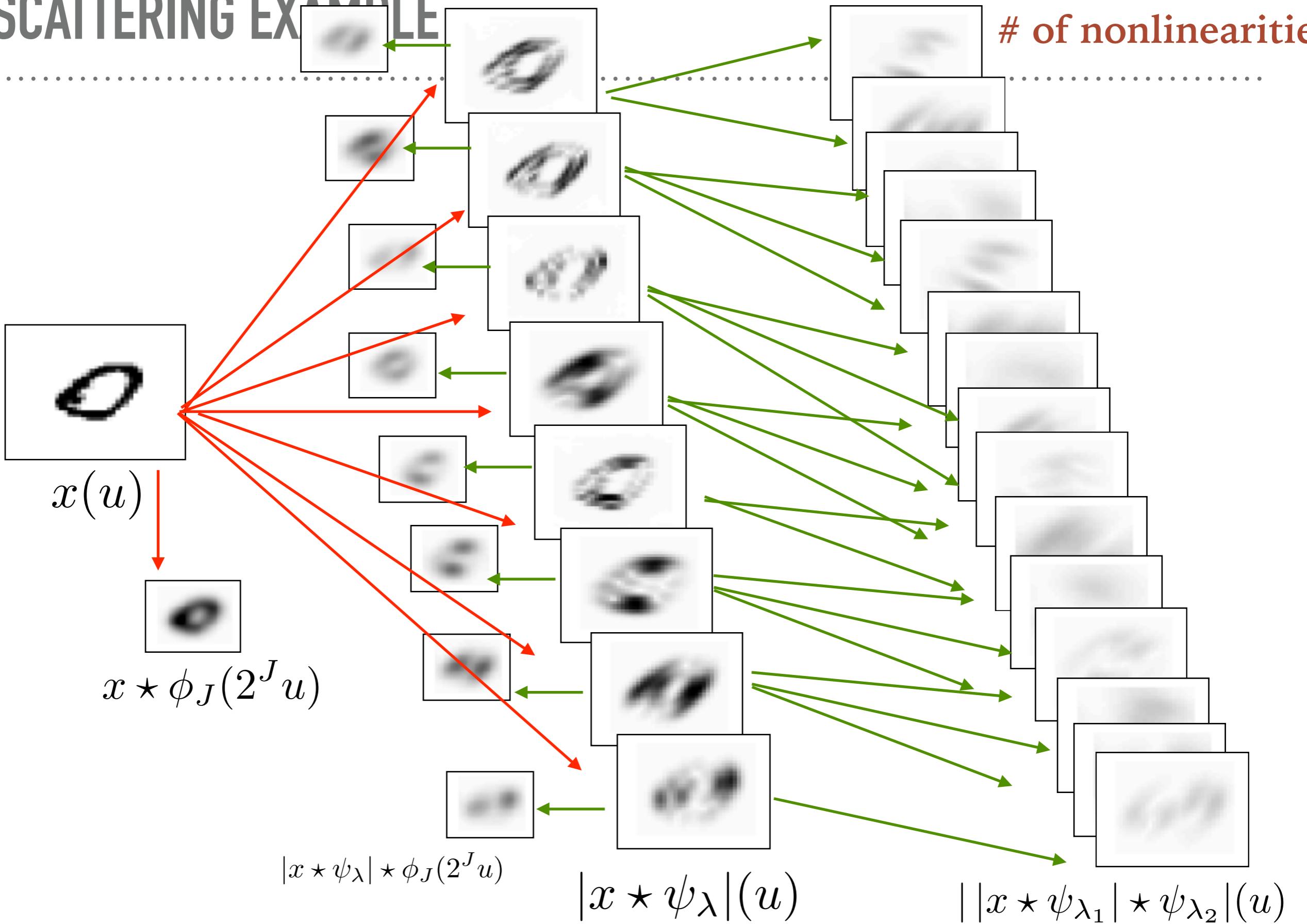
# SCATTERING EXAMPLE



# SCATTERING EXAMPLE

#layers:

# of nonlinearities



# SCATTERING WITH MULTI-RESOLUTION WAVELETS

---

- We have considered a collection  $\psi_{j,\theta}$  of oriented and dilated wavelets, and a translation co-variant wavelet decomposition operator:

$$Wx = \{x \star \phi_J, x \star \psi_{j,\theta}\}$$

- With  $J$  scales and  $L$  orientations, the redundancy is  $(1+JL)$ .

# MULTI-RESOLUTION WAVELETS

---

- At each scale  $j$ , we consider a low-pass *scaling filter*  $h$  and band-pass filters  $g_\theta$ ,  $\theta \in [1, \dots, L]$ .
- Wavelets and the blurring kernel are obtained at each  $j$  by cascading these filters:

$$\phi_j = \phi_{j-1} \star h_j \quad \psi_{j,\theta} = \phi_{j-1} \star g_{j,\theta} .$$

- Decompositions are obtained by cascading fine-to-coarse:

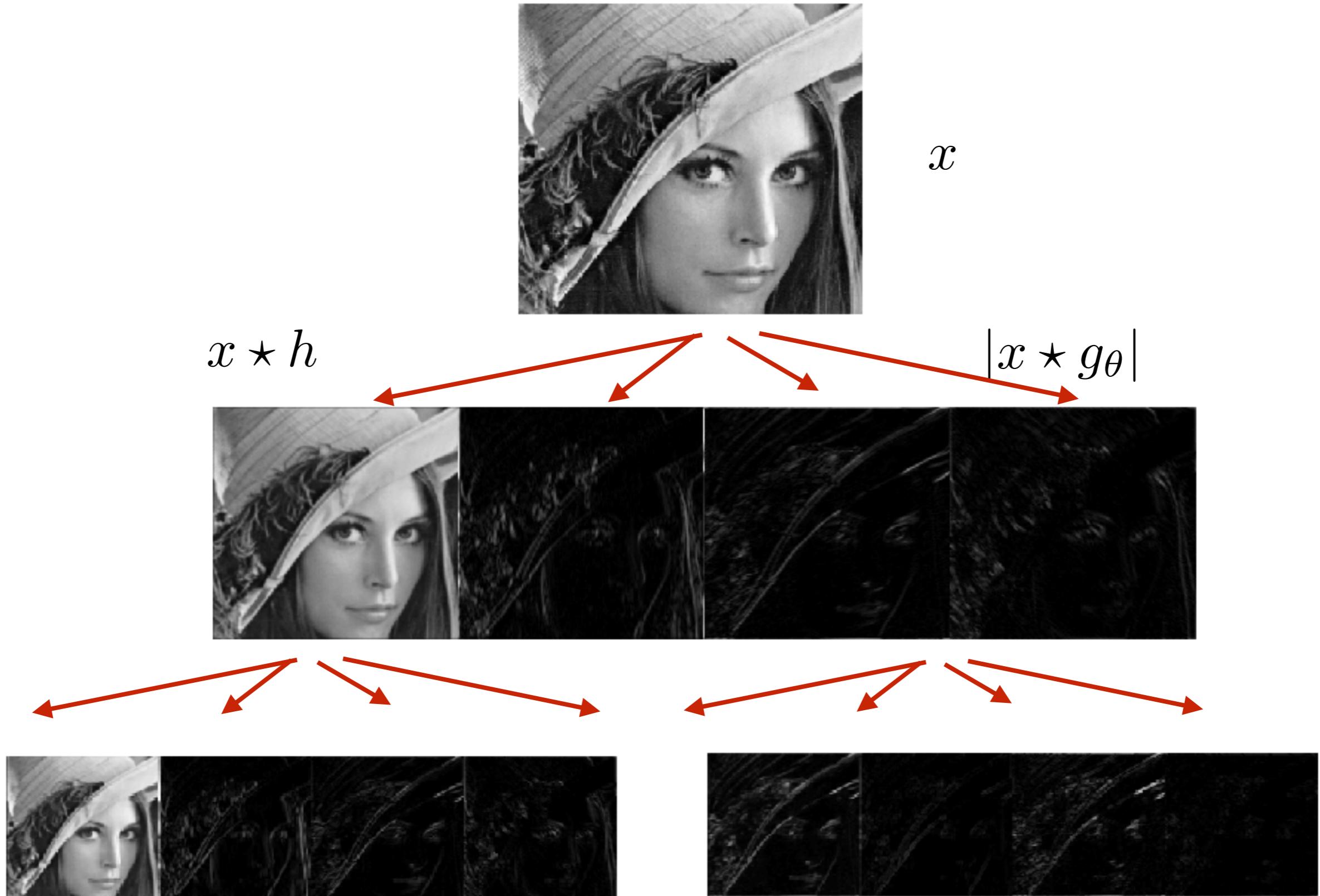
$$x \star \phi_j(u) = (x \star \phi_{j-1}) \star h_j(u) , \quad x \star \psi_{j,\theta}(u) = (x \star \phi_{j-1}) \star g_{j,\theta}(u) .$$

- Downsampling (or “*stride*”) adaptive to signal smoothness:

$$x \star \phi_j(u) = (x \star \phi_{j-1}) \star h(2u) , \quad x \star \psi_{j,\theta}(u) = (x \star \phi_{j-1}) \star g_\theta(2u) .$$

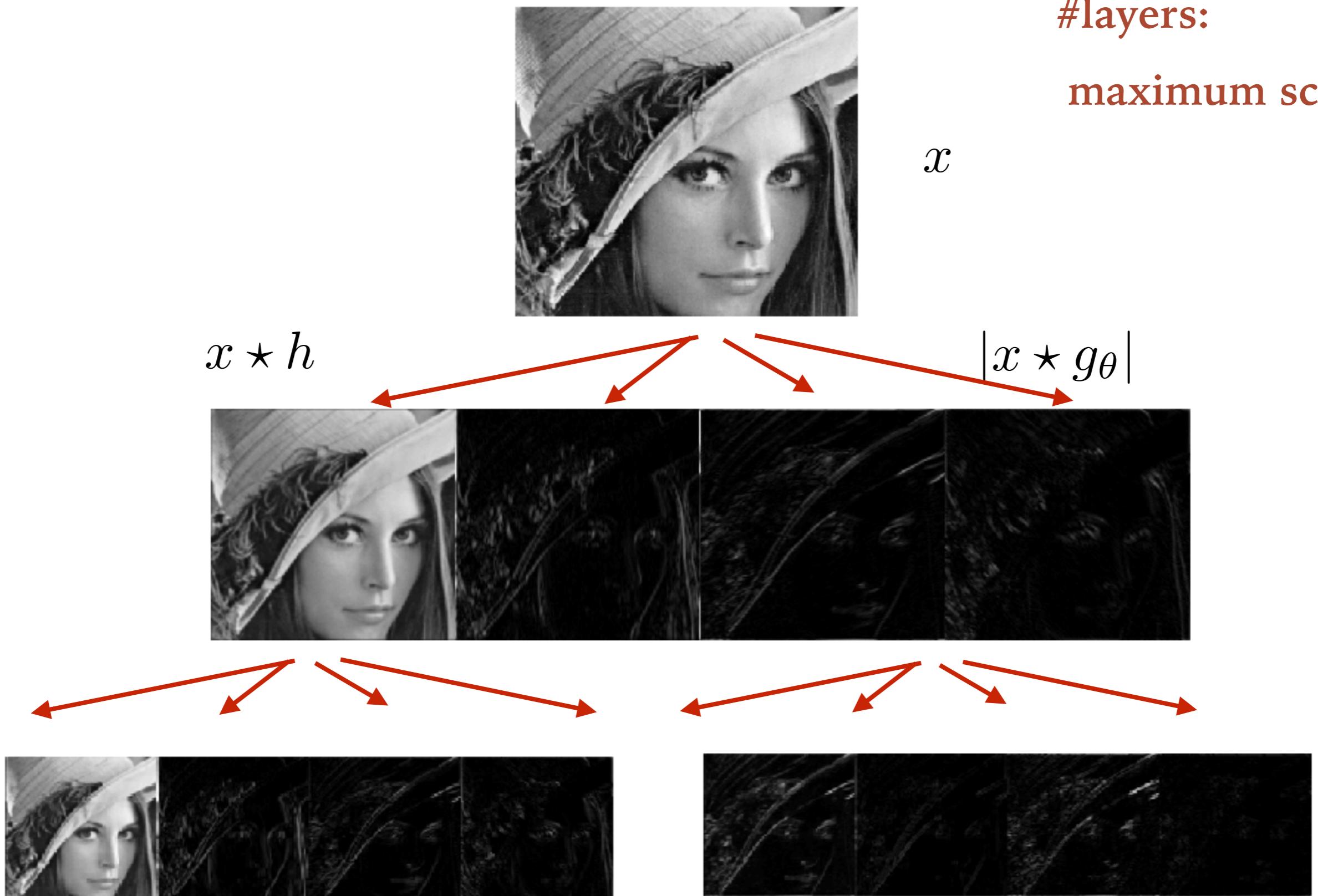
# SCATTERING WITH MULTI-RESOLUTION WAVELETS

---



# SCATTERING WITH MULTI-RESOLUTION WAVELETS

#layers:  
maximum scale



# SCATTERING CONSERVATION OF ENERGY

---

- Additive stability and conservation of energy:

**Theorem (Mallat):** For appropriate wavelets, the scattering representation is contractive,  $\|S_J x - S_J x'\| \leq \|x - x'\|$ , and unitary,  $\|S_J x\| = \|x\|$ .

$$\|S_J x\|^2 = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\|^2$$

# SCATTERING CONSERVATION OF ENERGY

---

**Theorem (Mallat):** For appropriate wavelets, the scattering representation is contractive,  $\|S_Jx - S_Jx'\| \leq \|x - x'\|$ , and unitary,  $\|S_Jx\| = \|x\|$ .

$$\|S_Jx\|^2 = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\|^2$$

- In practice, the transform is limited to a finite number of layers  $m_{max}$ . This result shows residual error converges to 0.
- The result requires complex wavelets (ie, not real).

# INTERPRETATION

---

- Unitary Wavelet decomposition preserves energy:

$$\|x\|^2 = \|x \star \phi_J\|^2 + \sum_{j \leq J, \theta} \|x \star \psi_{j,\theta}\|^2 .$$

- Repeat formula on each output  $|x \star \psi_{j,\theta}|$  :

$$|||x \star \psi_{j,\theta}|||^2 = |||x \star \psi_{j,\theta}| \star \phi_J||^2 + \sum_{j_2 \leq J, \theta_2} |||x \star \psi_{j,\theta}| \star \psi_{j_2,\theta_2}||^2 .$$

$$\|x\|^2 = \|S_J[0]x\|^2 + \sum_{|p|=1} \|S_J[p]x\|^2 + \sum_{|p|=2} |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}||^2$$

$\forall m$

$$\|x\|^2 = \sum_{|p| < m} \|S_J[p]x\|^2 + \sum_{|p|=m} |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} | \dots \psi_{\lambda_m}||^2$$

# INTERPRETATION

---

- Unitary Wavelet decomposition preserves energy:

$$\|x\|^2 = \|x \star \phi_J\|^2 + \sum_{j \leq J, \theta} \|x \star \psi_{j,\theta}\|^2.$$

$$|x \star \psi_{j,\theta}|$$

► Repeat formula on each output

$$\||x \star \psi_{j,\theta}|\|^2 = \||x \star \psi_{j,\theta}| \star \phi_J\|^2 + \sum_{j_2 \leq J, \theta_2} \||x \star \psi_{j,\theta}| \star \psi_{j_2, \theta_2}\|^2.$$

$$\|x\|^2 = \|S_J[0]x\|^2 + \sum_{|p|=1} \|S_J[p]x\|^2 + \sum_{|p|=2} \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|^2$$

$\forall m$

$$\|x\|^2 = \sum_{|p| < m} \|S_J[p]x\|^2 + \sum_{|p|=m} \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} | \dots \psi_{\lambda_m}\|^2$$

# INTERPRETATION

---

- Result amounts to proving that

$$\lim_{m \rightarrow \infty} \sum_{|p|=m, j_i \leq J} \| |x \star \psi_{\lambda_1}| \star \dots | \star \psi_{\lambda_m} \|^2 = 0 .$$

- *Fact:* Every time we apply the (complex) wavelet modulus, we push energy towards the low frequencies.
- Result is obtained by formally proving this fact.

# INTERPRETATION

---

- Result amounts to proving that

$$\lim_{m \rightarrow \infty} \sum_{|p|=m, j_i \leq J} \| |x \star \psi_{\lambda_1} | \star \dots | \star \psi_{\lambda_m} | \|^2 = 0 .$$

- Fact: Every time we apply the (complex) wavelet modulus, we push energy towards the low frequencies.
- Result is obtained by formally showing this fact.
- Extended to general wavelets in [Waldspurger,'17] and to other frames in [Czaja, Li.16].
- Decay is in fact exponential for band-limited signals.

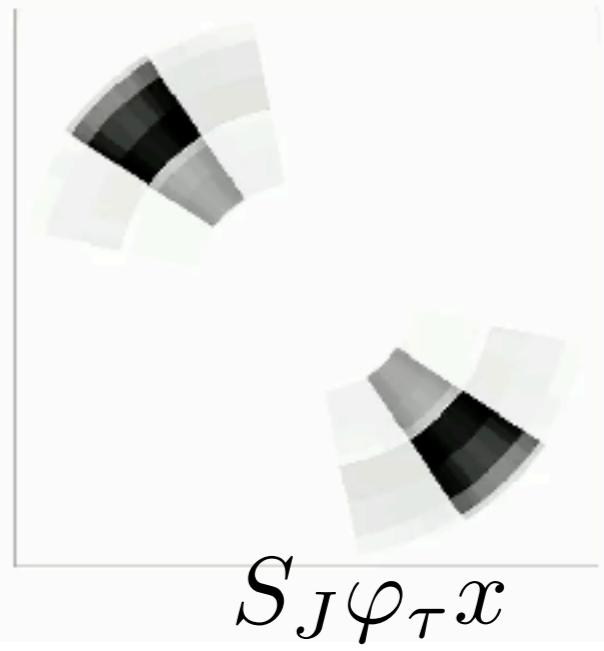
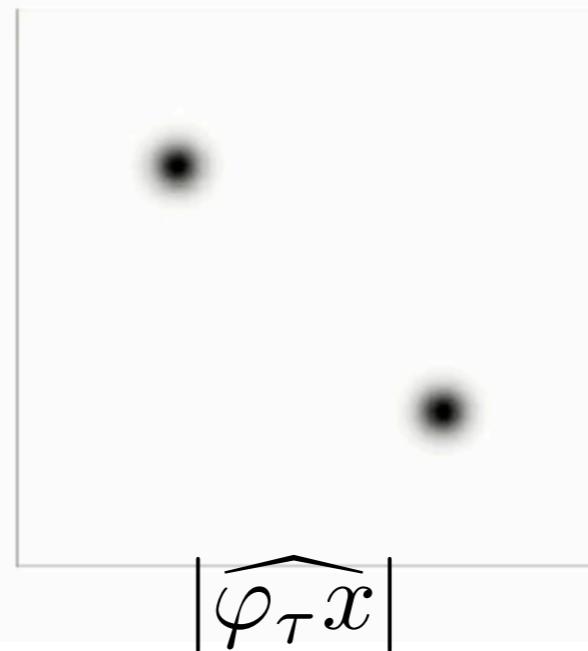
# SCATTERING GEOMETRIC STABILITY

► Geometric Stability:

$$\|S_J x\|^2 = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\|^2$$

**Theorem** (Mallat'10): There exists  $C$  such that for all  $x \in L^2(\mathbb{R}^d)$  and all  $m$ , the  $m$ -th order scattering satisfies

$$\|S_J \varphi_\tau x - S_J x\| \leq Cm\|x\|(2^{-J}|\tau|_\infty + \|\nabla \tau\|_\infty + \|H\tau\|_\infty) .$$



# INTERPRETATION

---

- Denote

$$A_J x = x \star \phi_J \quad W_J x = \{x \star \psi_\lambda\}_\lambda \quad Mx = |x|$$

- We know that

$$\|A_J - A_J \varphi_\tau\| \leq C(2^{-J} |\tau|_\infty + |\nabla \tau|_\infty)$$

$$\|W_J \varphi_\tau - \varphi_\tau W_J\| \leq C(J |\nabla \tau|_\infty + |H \tau|_\infty)$$

$$M \varphi_\tau = \varphi_\tau M$$

( $[A, B] = AB - BA$  : Commutator)

- $S_J = \{A_J, A_J M W_J, A_J M W_J M W_J, \dots\}$

# INTERPRETATION

---

- Each order contributes separately:

$$\|S_J - S_J \varphi_\tau\|^2 = \|A_J - A_J \varphi_\tau\|^2 + \|A_J M W_J - A_J M W_J \varphi_\tau\|^2 + \dots$$

# INTERPRETATION

---

- Each order contributes separately:

$$\|S_J - S_J \varphi_\tau\|^2 = \|A_J - A_J \varphi_\tau\|^2 + \|A_J M W_J - A_J M W_J \varphi_\tau\|^2 + \dots$$

- Let us inspect a generic term:

$$\left\| A_J \underbrace{M W_J M W_J \dots M W_J}_{(U_J = M W_J) \text{ } k \text{ times}} - A_J \underbrace{M W_J M W_J \dots M W_J}_{k \text{ times}} \varphi_\tau \right\|$$

$$\|A_J U_J^k - A_J U_J^k \varphi_\tau\| \leq \|A_J U_J^k - A_J U_J^{k-1} \varphi_\tau U_J\| + \|A_J U_J^{k-1} \varphi_\tau U_J - A_J U_J^k \varphi_\tau\|$$

# INTERPRETATION

---

- Each order contributes separately:

$$\|S_J - S_J \varphi_\tau\|^2 = \|A_J - A_J \varphi_\tau\|^2 + \|A_J M W_J - A_J M W_J \varphi_\tau\|^2 + \dots$$

- Let us inspect a generic term:

$$\left\| A_J \underbrace{M W_J M W_J \dots M W_J}_{k \text{ times}} - A_J \underbrace{M W_J M W_J \dots M W_J}_{k \text{ times}} \varphi_\tau \right\|$$
$$(U_J = M W_J)$$

$$\begin{aligned} & \|A_J U_J^k - A_J U_J^k \varphi_\tau\| \leq \|A_J U_J^k - A_J U_J^{k-1} \varphi_\tau U_J\| + \|A_J U_J^{k-1} \varphi_\tau U_J - A_J U_J^k \varphi_\tau\| \\ & \leq \|A_J U_J^{k-1} - A_J U_J^{k-1} \varphi_\tau\| + \|A_J U_J^{k-1} [\varphi_\tau, U_J]\| \\ & \leq \|A_J U_J^{k-1} - A_J U_J^{k-1} \varphi_\tau\| + \|[\varphi_\tau, U_J]\| \\ & \leq k \|[\varphi_\tau, U_J]\| + \|A_J - A_J \varphi_\tau\| \leq k \|[\varphi_\tau, W_J]\| + \|A_J - A_J \varphi_\tau\| \end{aligned}$$

# DISCRIMINABILITY AND SPARSITY

---

- Typical non-linearities are contractive:

$$\|\rho(x) - \rho(x')\| \leq \|x - x'\|$$

# DISCRIMINABILITY AND SPARSITY

---

- Typical non-linearities are contractive:

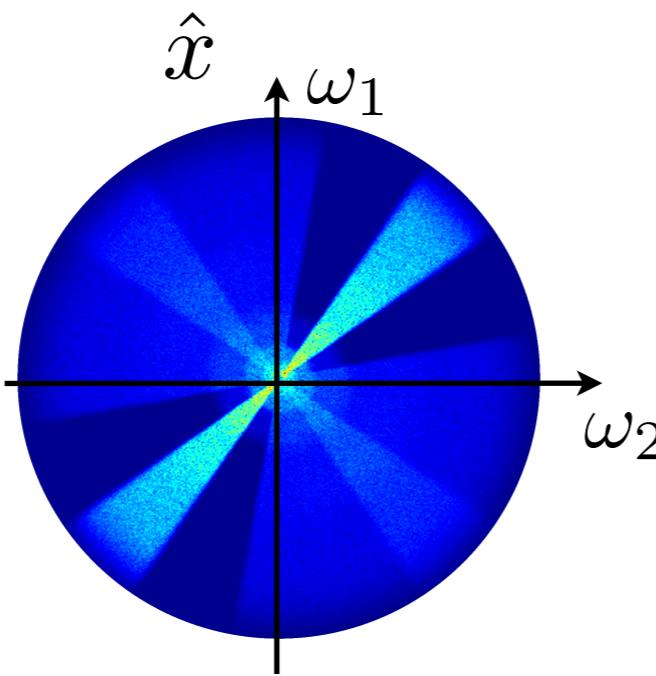
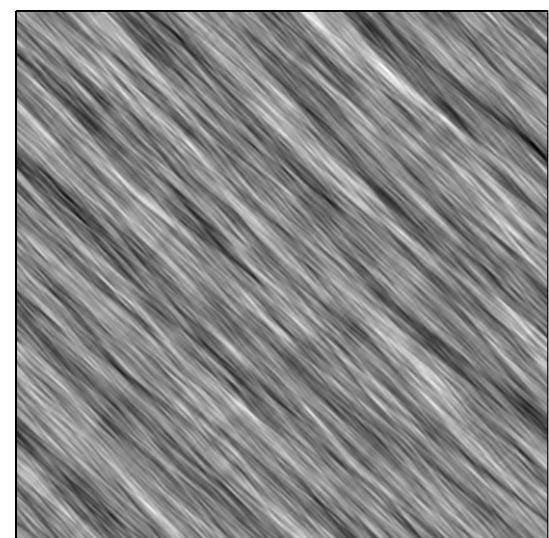
$$\|\rho(x) - \rho(x')\| \leq \|x - x'\|$$

- However, if  $x, x'$  are sparse, this inequality is an equality in most of the signal domain.
- Thus sparsity is a means to control and prevent excessive contraction of different signal classes.

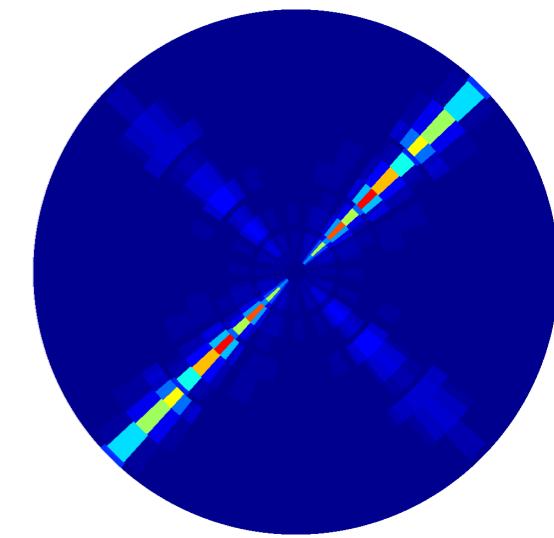
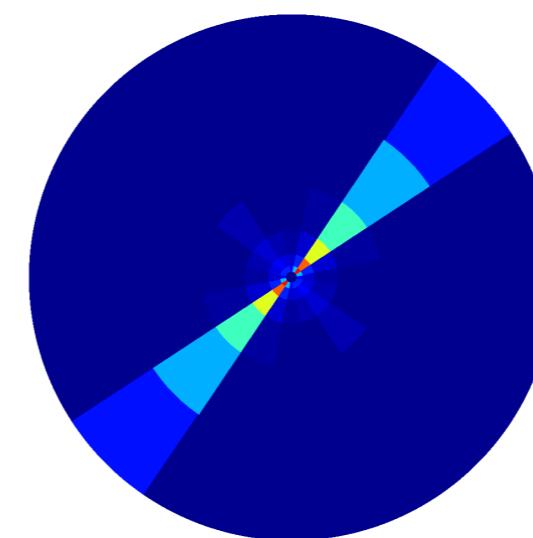
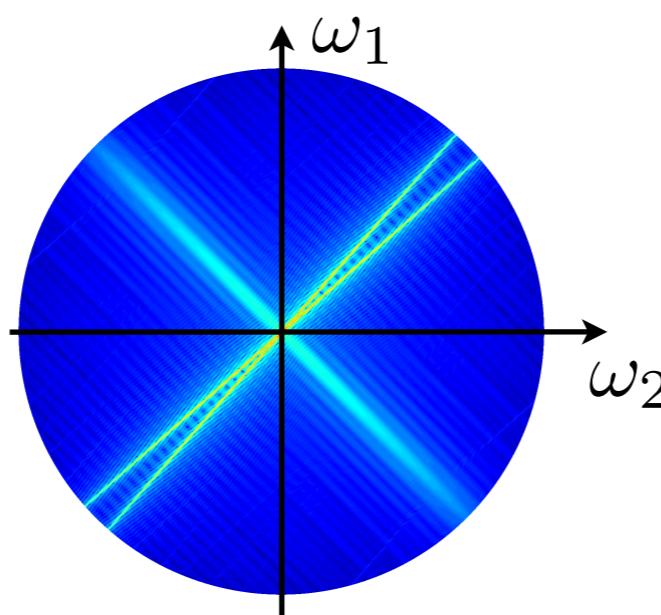
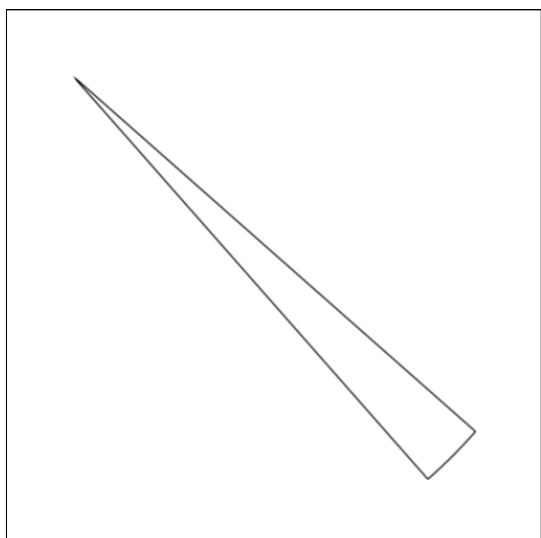
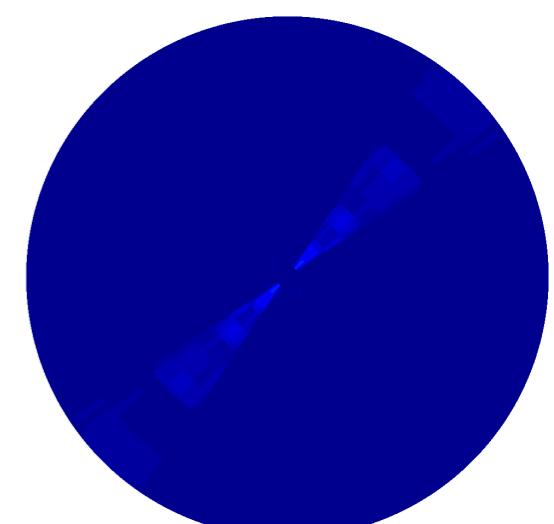
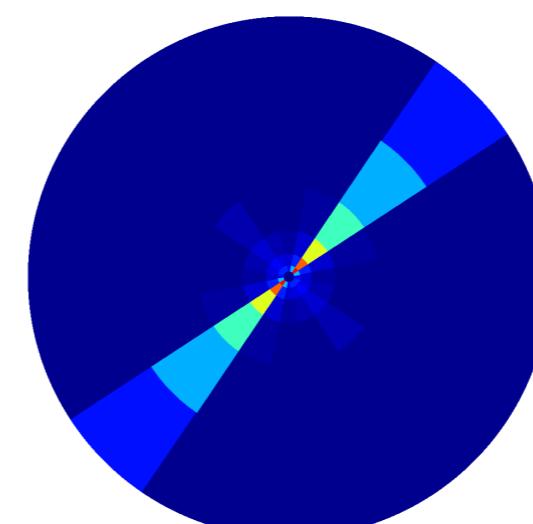
# IMAGE EXAMPLES

Images ..... Fourier ..... Wavelet Scattering .....

$x$



$$|x \star \psi_{\lambda_1}| \star \phi_J \quad ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J$$

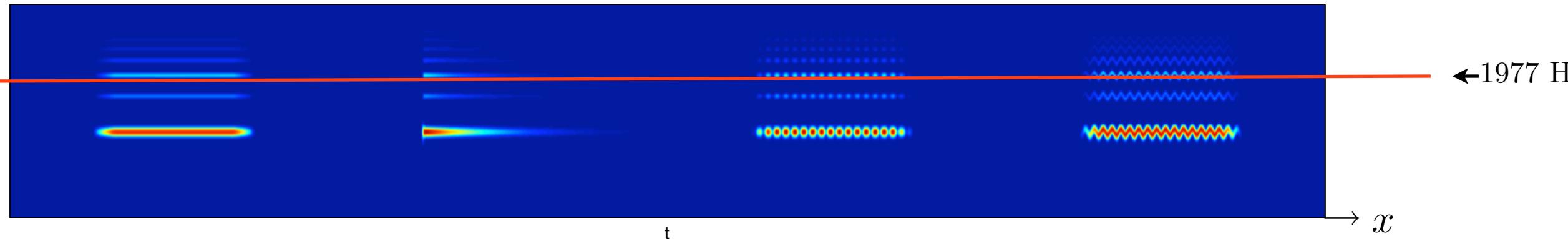


window size = image size

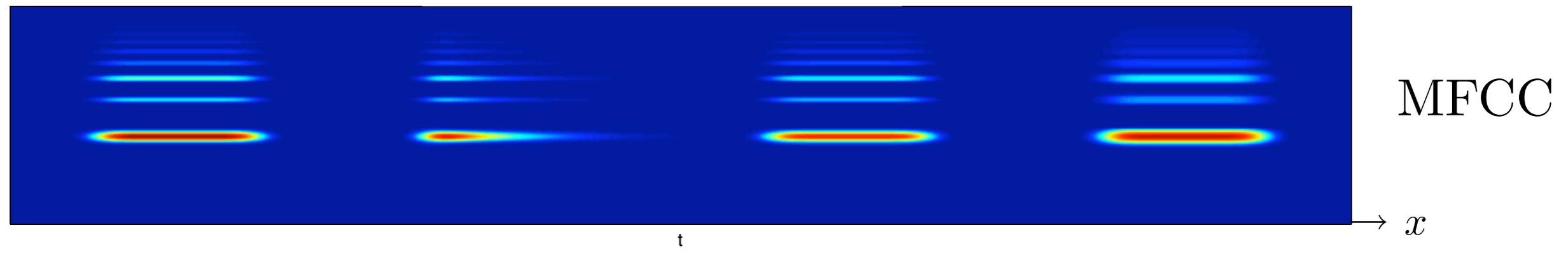
# SOUND EXAMPLES

(courtesy J. Anden)

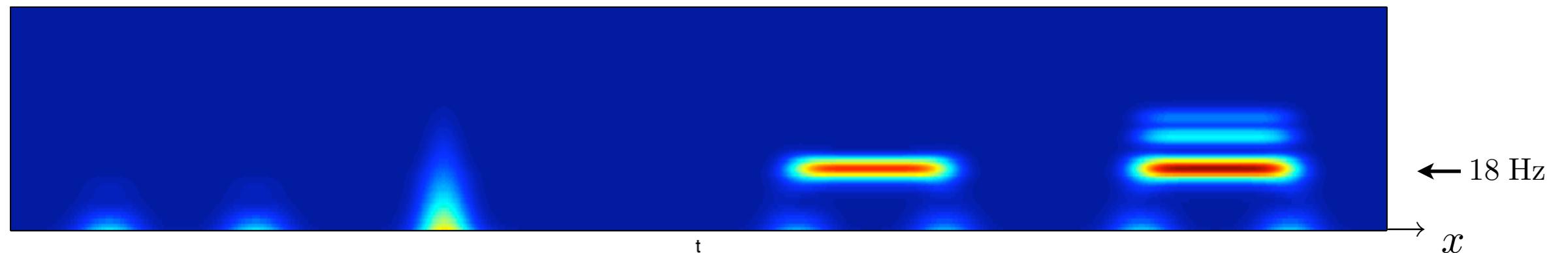
$$\lambda_1 = \log(\omega_1)$$



$$\lambda_1 = \log(\omega_1)$$



$$\lambda_2 = \log(\omega_2)$$



# LIMITATIONS OF SEPARABLE SCATTERING

---

- No feature dimensionality reduction
  - The number of features increases exponentially with depth and polynomially with scale.

# LIMITATIONS OF SEPARABLE SCATTERING

---

- No feature dimensionality reduction
  - The number of features increases exponentially with depth and polynomially with scale.
- We are indirectly assuming that each wavelet band is deformed independently
  - We cannot capture the *joint* deformation structure of feature maps
  - Loss of discriminability.

# LIMITATIONS OF SEPARABLE SCATTERING

---

- No feature dimensionality reduction
  - The number of features increases exponentially with depth and polynomially with scale.
- We are indirectly assuming that each wavelet band is deformed independently
  - We cannot capture the *joint* deformation structure of feature maps
  - Loss of discriminability.
- The deformation model is rigid and non-adaptive
  - We cannot adapt to each class
  - Wavelets are hard to define *a priori* on high-dimensional domains.

# JOINT VERSUS SEPARABLE INVARIANCE

---

- Suppose we simply want stable translation invariance.
- Two-dimensional translation group in a periodic domain:

$$G \cong (\mathbb{R}/([0, N]))^2 = S^1 \times S^1 \cong \mathbb{T}^2$$



# JOINT VERSUS SEPARABLE INVARIANCE

---

- Suppose we simply want stable translation invariance.
- Two-dimensional translation group in a periodic domain:

$$G \cong (\mathbb{R}/([0, N]))^2 = S^1 \times S^1 \cong \mathbb{T}^2$$



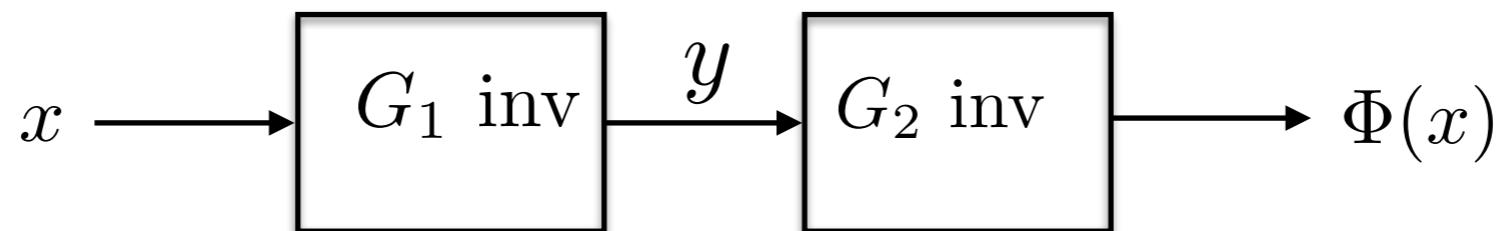
- Each  $S^1$  acts on images along a different coordinate:

$$\varphi_a^1 x(u_1, u_2) = x(u_1 - a, u_2), \quad \varphi_a^2 x(u_1, u_2) = x(u_1, u_2 - a)$$

# JOINT VERSUS SEPARABLE INVARIANCE

---

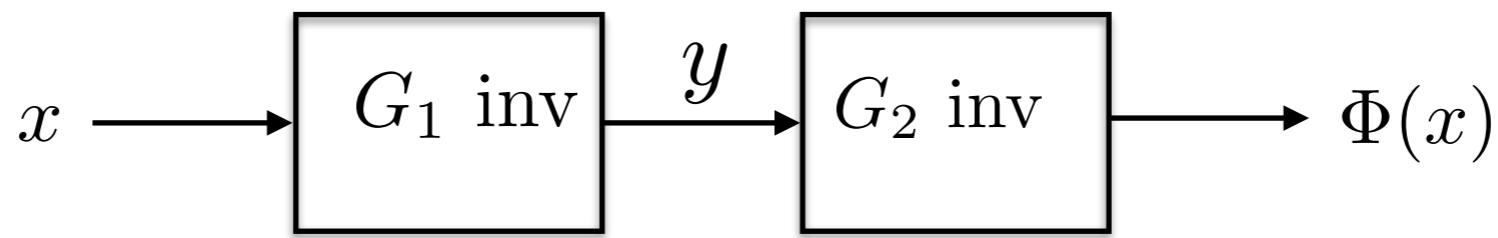
- So we could just consider one-dimensional (stable) translation invariant representations and compose:  
$$G = G_1 \times G_2$$



# JOINT VERSUS SEPARABLE INVARIANCE

---

- So we could just consider one-dimensional (stable) translation invariant representations and compose:  
$$G = G_1 \times G_2$$



- If for each  $u_2$ ,  $x(\cdot, u_2) \mapsto \Phi_1(x)(\cdot, u_2)$  is  $G_1$  invariant then  $\Phi_1(\varphi^1 x) = \Phi_1(x)$  for all  $x$  and  $\varphi^1 \in G_1$
- If for each  $\lambda$ ,  $y(\lambda, \cdot) \mapsto \Phi_2(y)(\lambda, \cdot)$  is  $G_2$  invariant then  $\Phi_2(\varphi^2 y) = \Phi_2(y)$  for all  $y$  and  $\varphi^2 \in G_2$

# JOINT VERSUS SEPARABLE INVARIANCE

---

- Thus, if  $\Phi_1$  is  $G_1$  invariant and  $G_2$  covariant, and  $\Phi_2$  is  $G_2$  invariant, then  $\Phi = \Phi_2 \circ \Phi_1$  satisfies

$$\forall \varphi \in G, \varphi = \varphi^1 \varphi^2, \varphi^i \in G_i$$

$$\Phi(\varphi x) = \Phi_2 \Phi_1(\varphi^1 \varphi^2 x) = \Phi_2 \Phi_1(\varphi^2 x) = \Phi_2 \varphi^2 \Phi_1(x) = \Phi_2 \Phi_1(x) = \Phi(x)$$

# JOINT VERSUS SEPARABLE INVARIANCE

---

- Thus, if  $\Phi_1$  is  $G_1$  invariant and  $G_2$  covariant, and  $\Phi_2$  is  $G_2$  invariant, then  $\Phi = \Phi_2 \circ \Phi_1$  satisfies

$$\forall \varphi \in G, \varphi = \varphi^1 \varphi^2, \varphi^i \in G_i$$

$$\Phi(\varphi x) = \Phi_2 \Phi_1(\varphi^1 \varphi^2 x) = \Phi_2 \Phi_1(\varphi^2 x) = \Phi_2 \varphi^2 \Phi_1(x) = \Phi_2 \Phi_1(x) = \Phi(x)$$

- So we achieve further invariance by composing partial invariances.
- Is there a problem here?

# JOINT VERSUS SEPARABLE INVARIANCE

---

- The factorization does not capture the joint action of  $G_1$  along the domain  $(u_1, u_2)$ .
- We are invariant to *too many* things.



# WAVELET COVARIANTS

---

- If we replace input image by first layer output:

$$\rho(x_0 \star \psi_{j,\theta})(u) = x_1(u, j, \theta)$$

Let  $\tilde{x}_0 = R_\alpha x_0$  be a rotation of  $\alpha$  degrees.

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(R_\alpha u, j, \theta + \alpha)$$

# WAVELET COVARIANTS

---

- If we replace input image by first layer output:

$$\rho(x_0 \star \psi_{j,\theta})(u) = x_1(u, j, \theta)$$

Let  $\tilde{x}_0 = R_\alpha x_0$  be a rotation of  $\alpha$  degrees.

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(R_\alpha u, j, \theta + \alpha)$$

- Similarly, roto-translation acts on  $x_1$  by rotating and translating spatial coordinates and translating orientation coordinates

Let  $\tilde{x}_0 = \varphi_{(v,\alpha)} x_0$  be a roto-translation with parameters  $(v, \alpha)$ .

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(\varphi_v R_\alpha u, j, \theta + \alpha)$$

# WAVELET COVARIANTS

---

- If we replace input image by first layer output:

$$\rho(x_0 \star \psi_{j,\theta})(u) = x_1(u, j, \theta)$$

Let  $\tilde{x}_0 = R_\alpha x_0$  be a rotation of  $\alpha$  degrees.

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(R_\alpha u, j, \theta + \alpha)$$

- Similarly, roto-translation acts on  $x_1$  by rotating and translating spatial coordinates and translating orientation coordinates

Let  $\tilde{x}_0 = \varphi_{(v,\alpha)} x_0$  be a roto-translation with parameters  $(v, \alpha)$ .

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(\varphi_v R_\alpha u, j, \theta + \alpha)$$

- So we can replace convolutions over translation by convolutions over roto-translations.

# GROUP CONVOLUTIONS

---



**Definition:** Let  $G$  be a group equipped with a Haar measure  $d\mu$ , acting on  $\Omega$ , and  $h \in L^1(G)$ . The group convolution  $x \star_G h$  is defined as

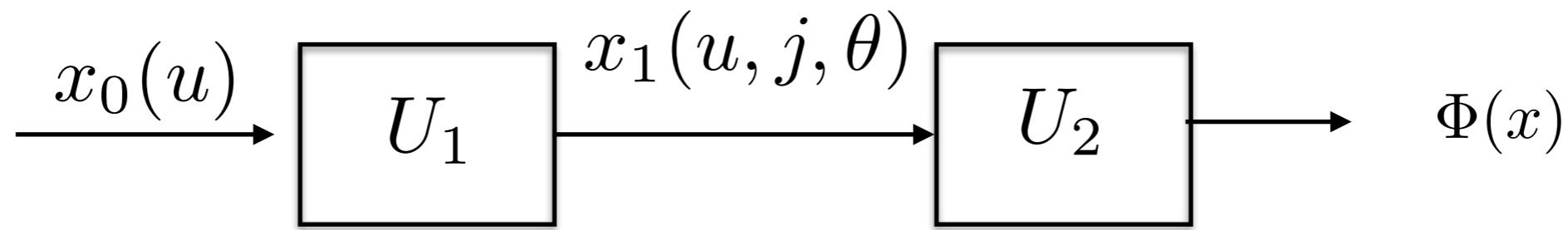
$$x \star_G h(u) = \int_G h(g)x(\varphi_g u)d\mu(g) , \quad x \in L^2(\Omega) .$$

If  $x = x_1(u, j, \theta)$  and  $G$  are roto-translations, these convolutions recombine different orientation channels.

# JOINT SCATTERING

---

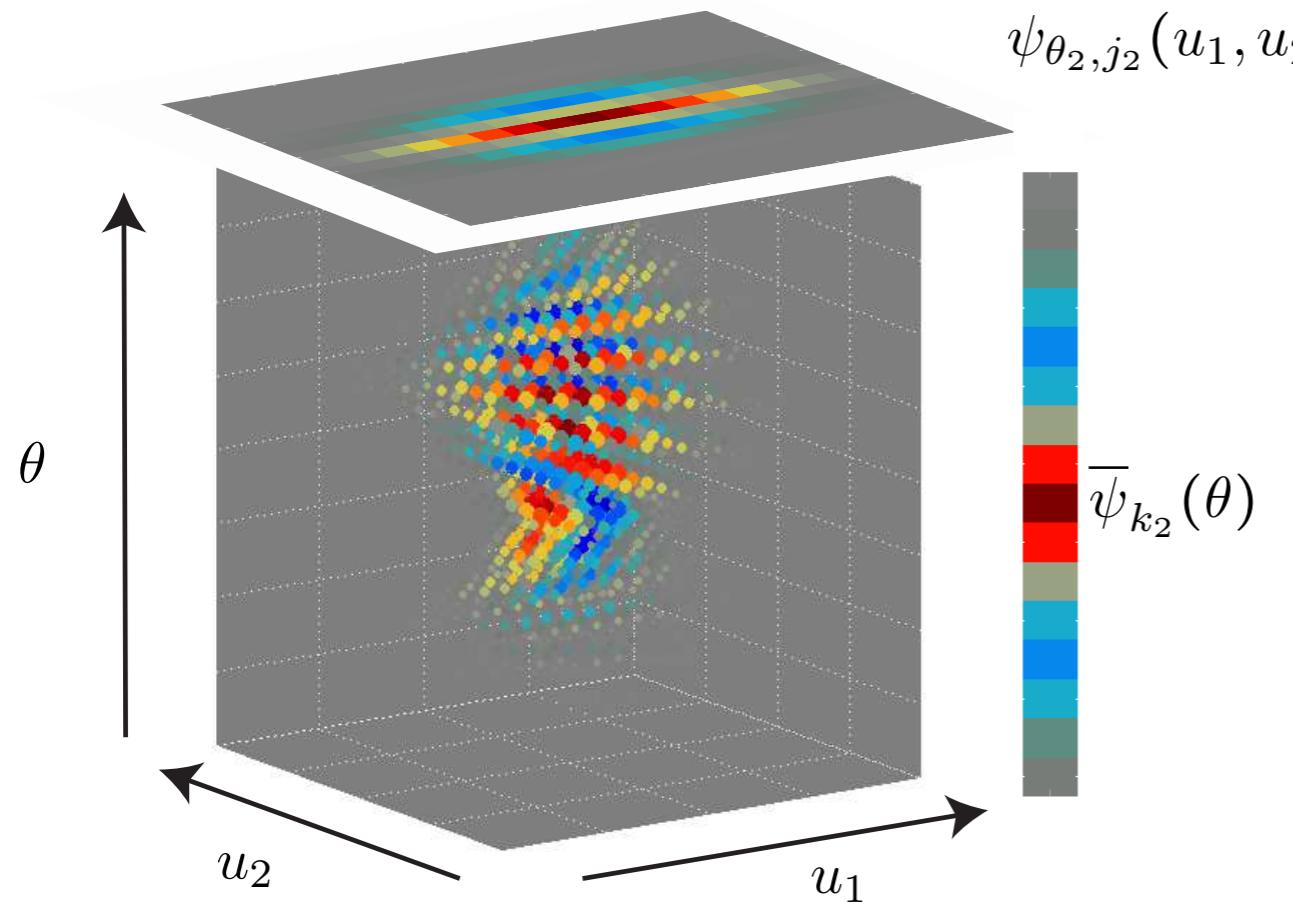
- We start by *lifting* the image with spatial wavelet convolutions: stable and covariant to roto-translations.



- We then adapt the second wavelet operator to its joint variability structure.
- More discriminability.
- Requires defining wavelets on more complicated domains:  
 $\Omega = \mathbb{R}^2 \times S^2$ .

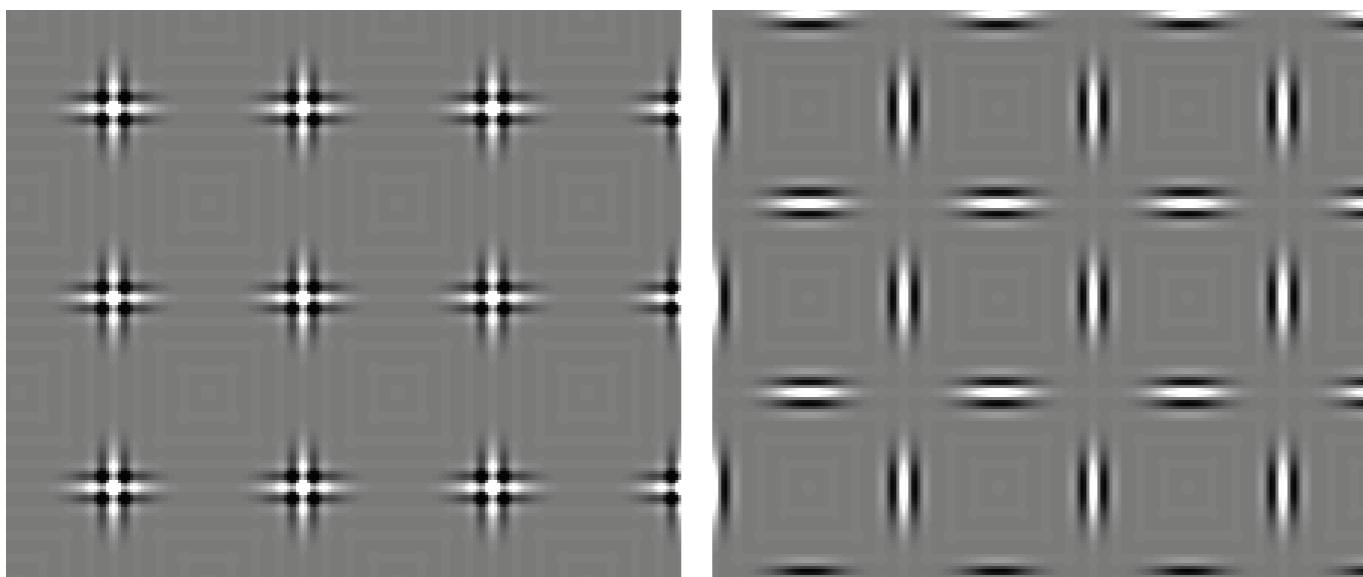
# EXAMPLE: ROTO-TRANSLATION SCATTERING

- [Sifre and Mallat'13]



$\psi_{\theta_2, j_2}(u_1, u_2)$   
second layer wavelets constructed by a  
separable product on spatial and rotational  
wavelets

$$\Psi_\lambda(u, \theta) = \psi_{\lambda_1}(u)\psi_{\lambda_2}(\theta)$$



example of patterns that are  
discriminated by joint scattering but  
not with separable scattering.

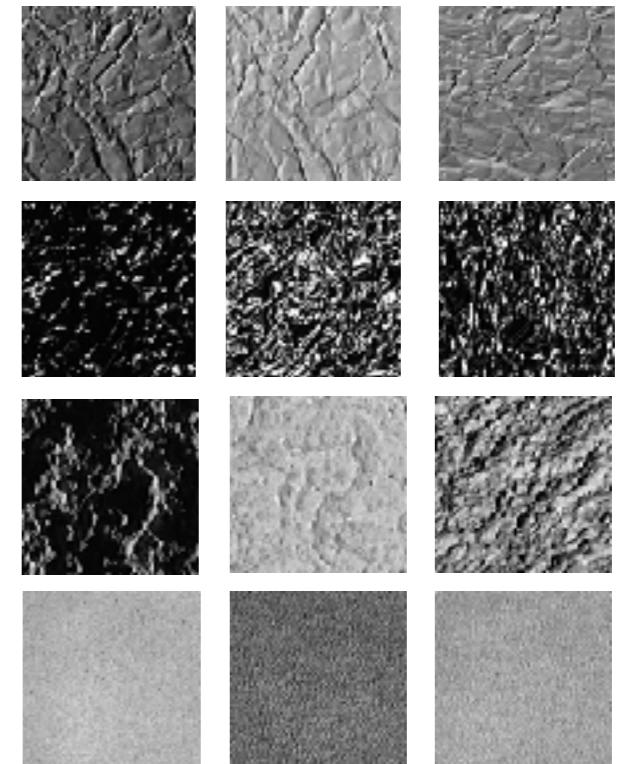
# CLASSIFICATION WITH SCATTERING

---

- State-of-the art on pattern and texture recognition using separable scattering followed by SVM:

- MNIST, USPS [Pami'13]

3 6 8 1 7 9 6 6 9 1  
6 7 5 7 8 6 3 4 8 5  
2 1 7 9 7 1 2 8 4 6  
4 8 1 9 0 1 8 8 9 4

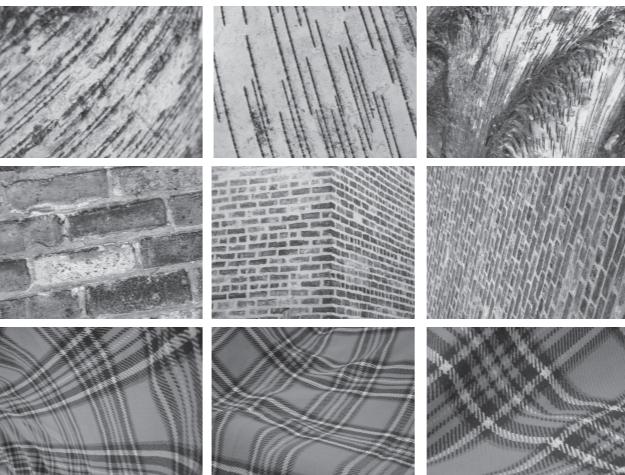


- Texture (CUREt) [Pami'13]

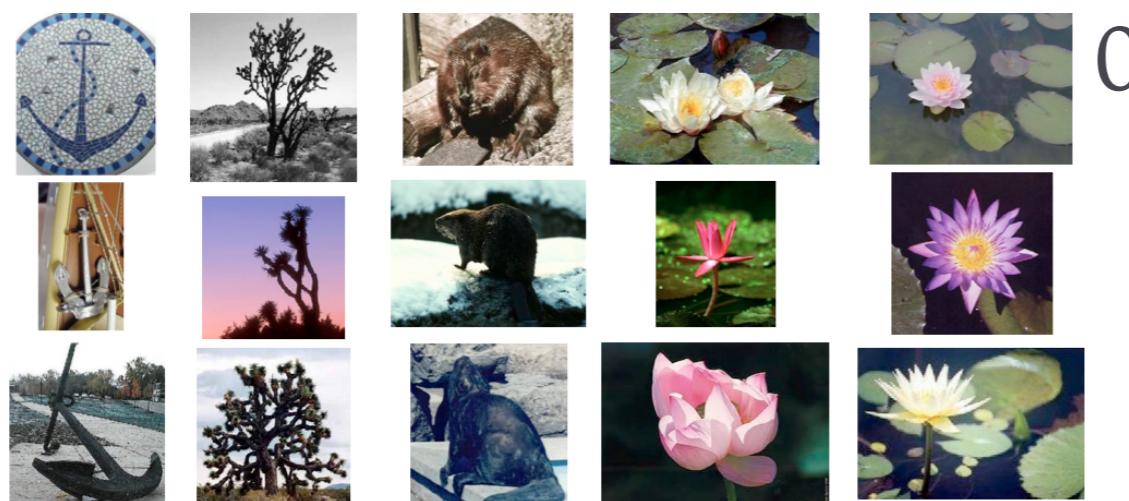
- Music Genre Classification (GTZAN) [IEEE Acoustic '13]

# CLASSIFICATION WITH SCATTERING

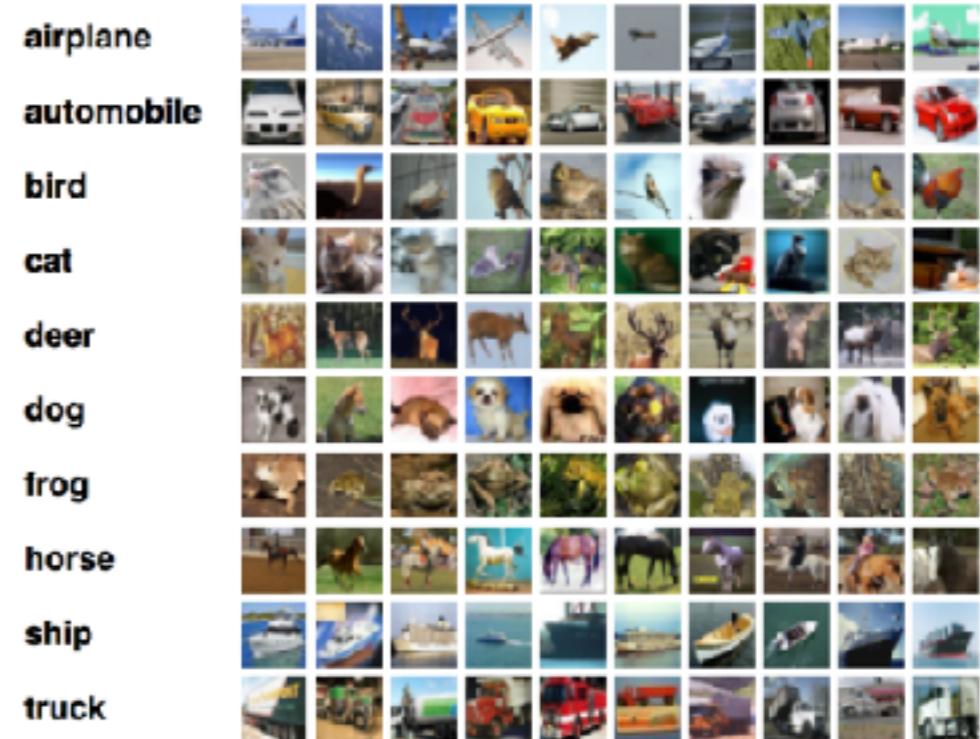
- Joint Scattering Improves Performance:
  - More complicated Texture (KTH,UIUC,UMD)  
[Sifre&Mallat, CVPR'13]



- Small-mid scale Object Recognition (Caltech, CIFAR)  
[Oyallon&Mallat, CVPR'15]



0



# LIMITATIONS OF JOINT SCATTERING

---

- Variability from physical world expressed in the language of transformation groups and deformations
  - However, there are not many possible groups: essentially the affine group and its subgroups.
- As a new wavelet layer is introduced, we create new coordinates, but we do not destroy existing coordinates
  - Hard to scale: dimensionality reduction is needed.
  - Wavelet design complicated beyond roto-translation groups.
- Beyond physics, many deformations are class-specific and not small.
  - Learning filters from data rather than designing them.

# FROM SCATTERING TO CNNS

---

- Given  $x(u, \lambda)$  and a group  $G$  acting on both  $u$  and  $\lambda$ , we defined wavelet convolutions over  $G$  as

$$x \star_G \psi_{\lambda'}(u, \lambda) = \int_v \int_{\alpha} \psi_{\lambda}(R_{-\alpha}(u - v)) x(v, \alpha) dv d\alpha$$

- In discrete coordinates,

$$x \star_G \psi_{\lambda'}(u, \lambda) = \sum_v \sum_{\alpha} \bar{\psi}_{\lambda'}(u - v, \alpha, \lambda) x(v, \alpha)$$

- Which in general is a convolutional tensor.