Generalization in diffusion models arises from geometry-adaptive harmonic representations

Zahra Kadkhodaie

September 2024

Florentin Guth

Eero Simoncelli

Stéphane Mallat

ENS
ÉCOLE NORMALE SUPÉRIEURE

NYU Center for Data Science

FLATIRON INSTITUTE
Center for Computational Neuroscience

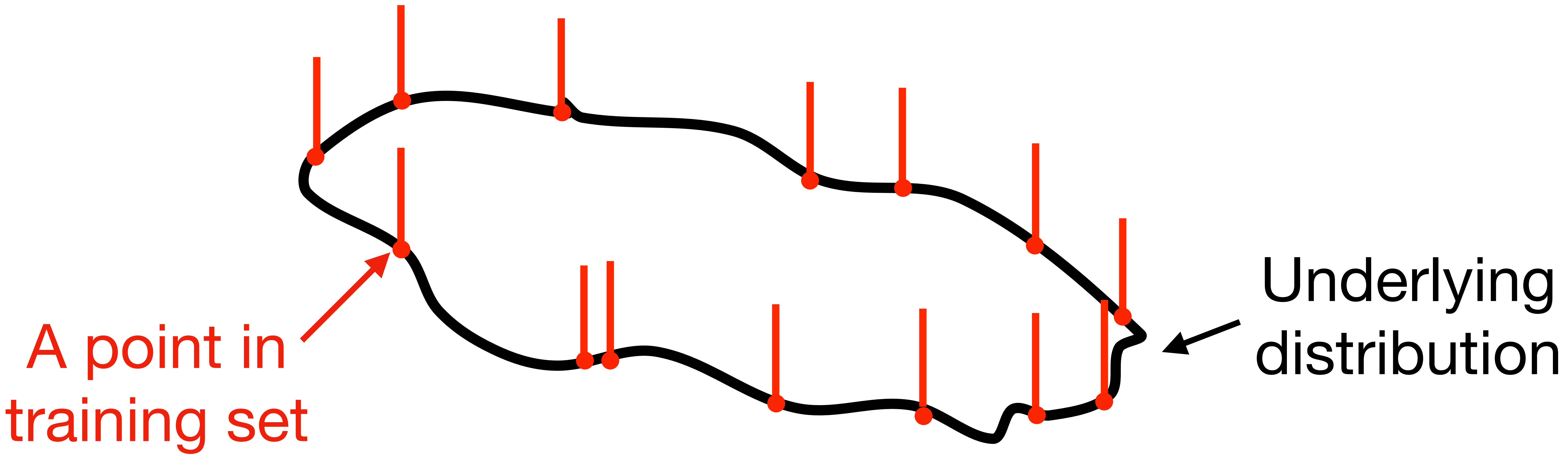
Diffusion models embed densities

- Sample from a **learned** density [Song & Ermon 2019; Ho et al 2020]
- How is this possible, given the "Curse of dimensionality" ?!

generated by a diffusion model

[Ho et al, 2022]

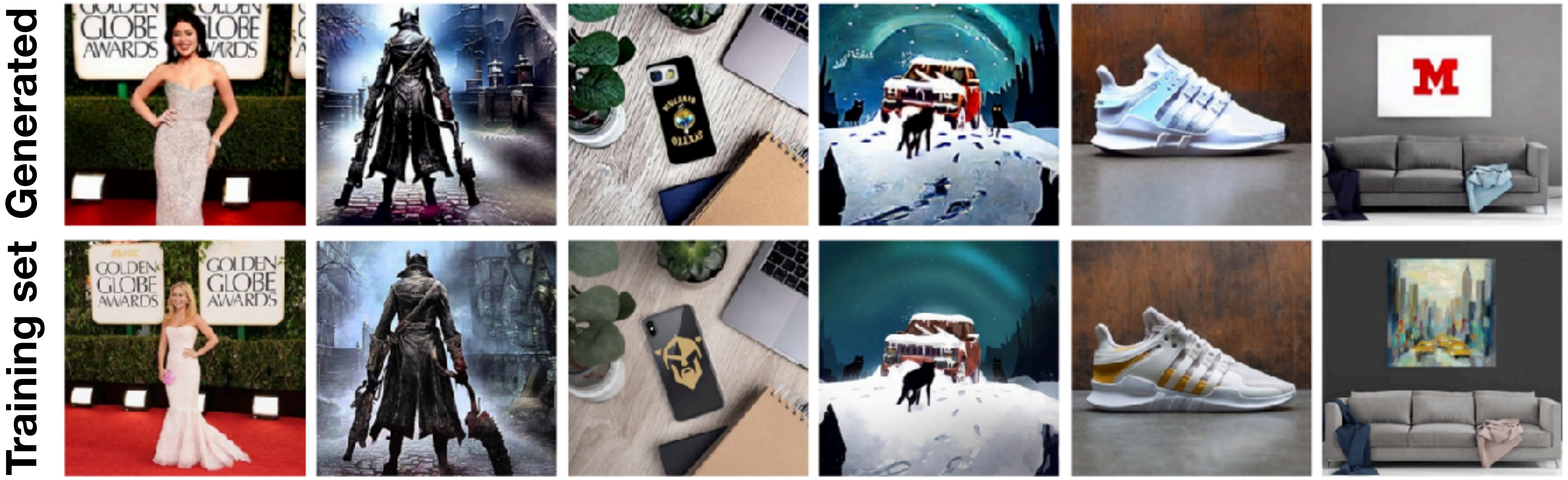
Memorization vs. generalization



A collection of delta functions

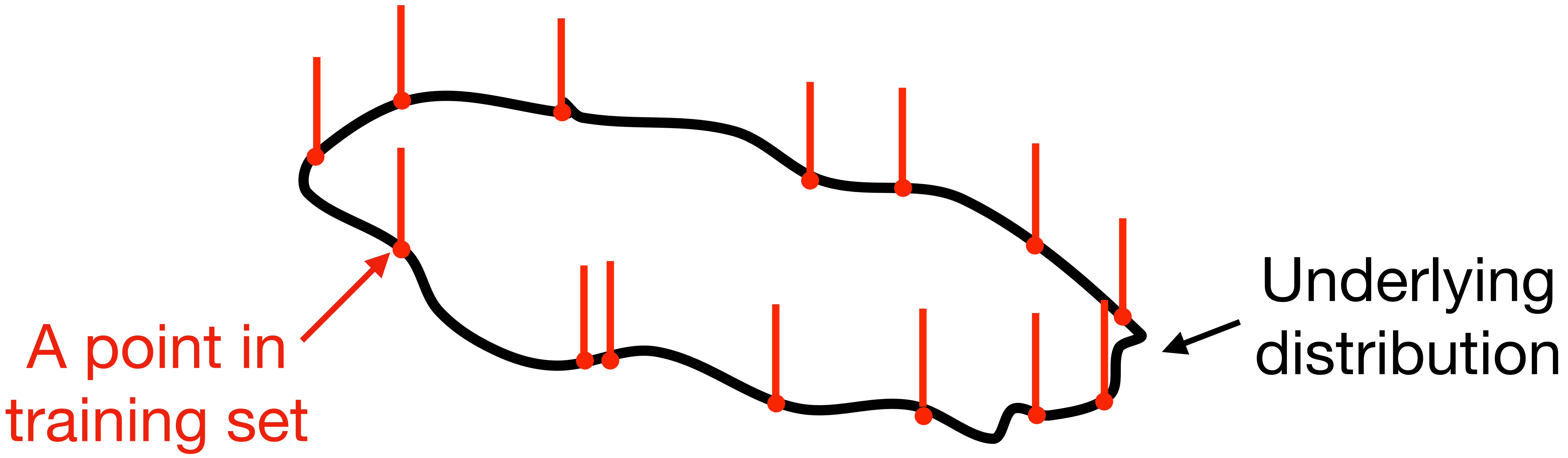
From training set | **Generated sample**

[Carlini et al, 2023]

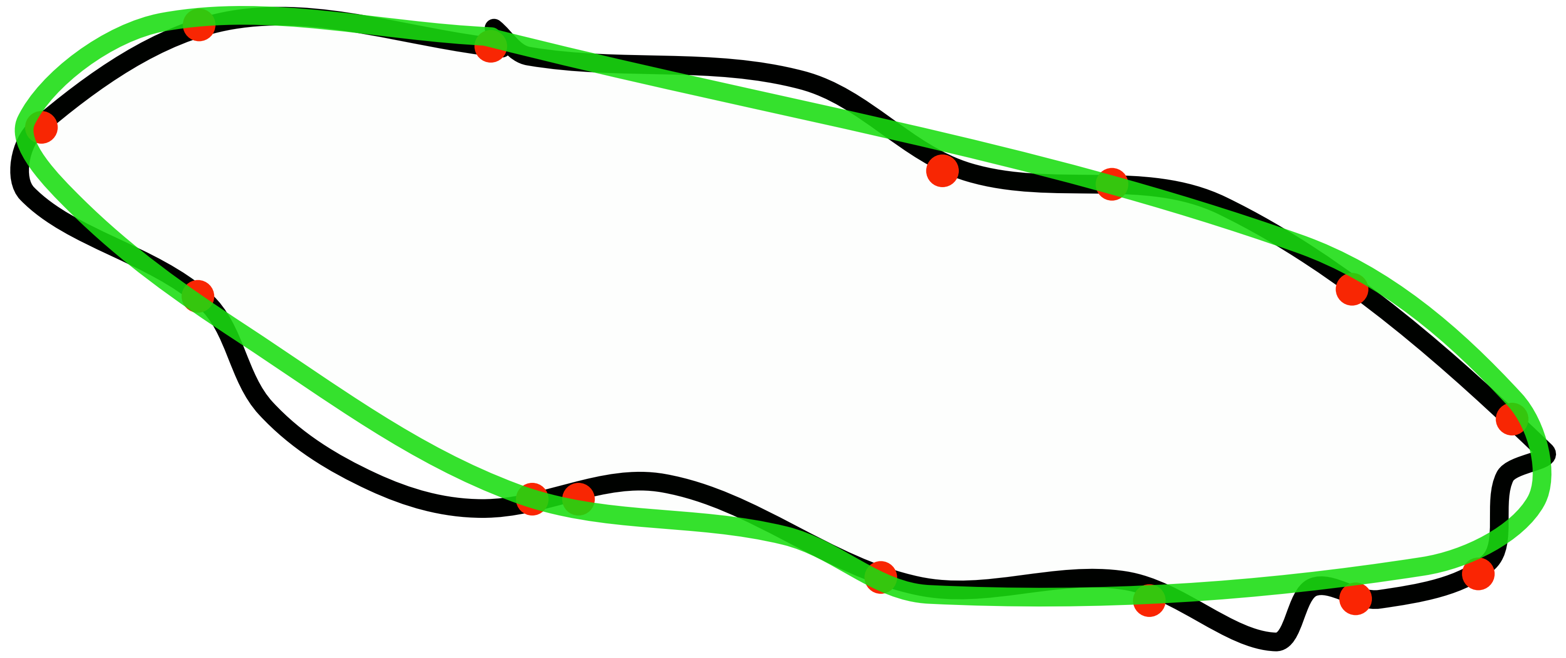


[Somepalli et al, 2023]

Memorization vs. generalization

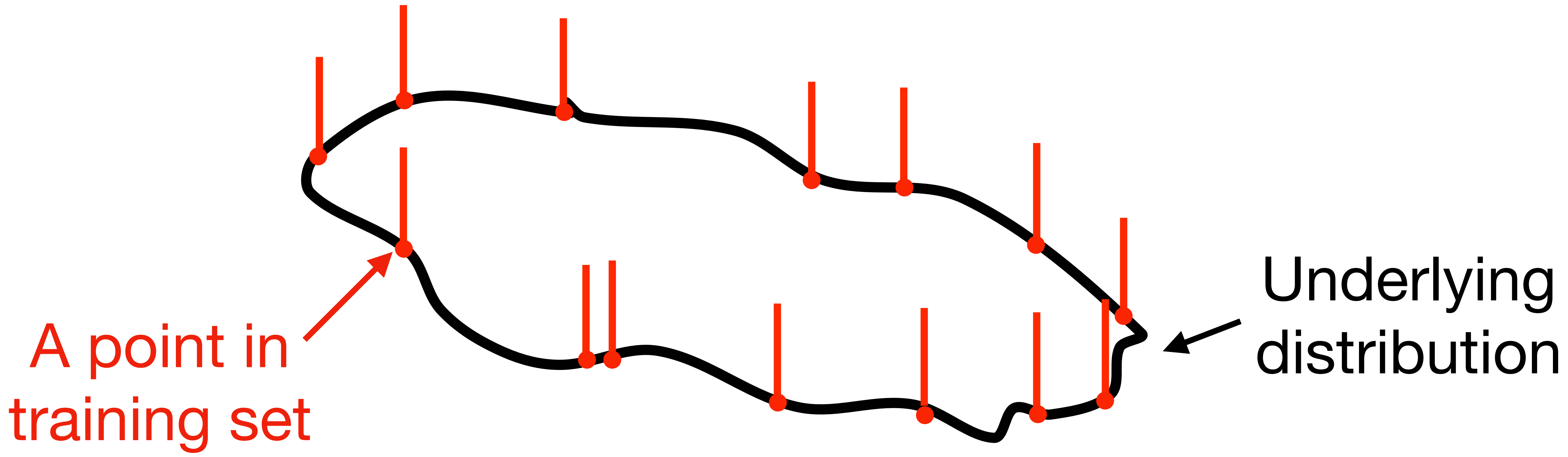


A collection of delta functions

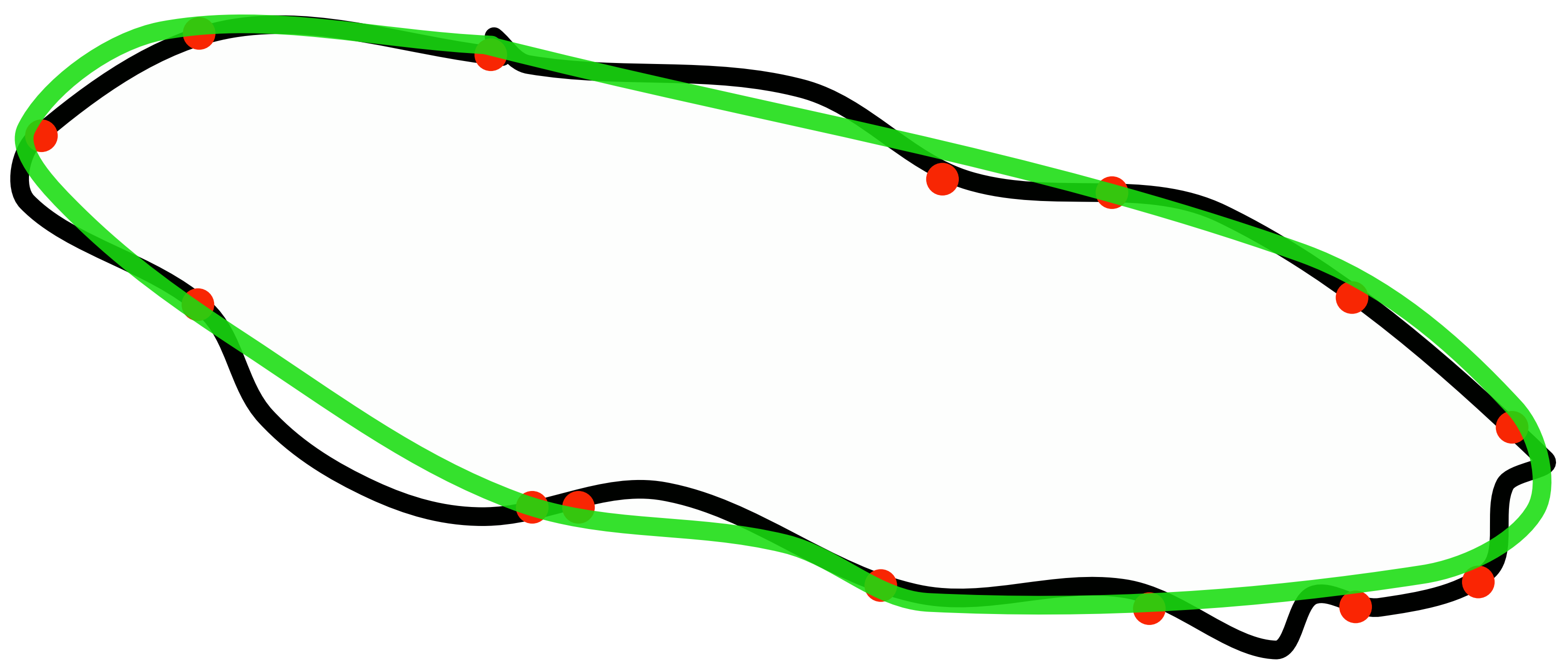


A continuous model of the underlying distribution

Memorization vs. generalization



A collection of delta functions



A continuous model of the underlying distribution

1. Can diffusion models generalize?

2. If so, how?

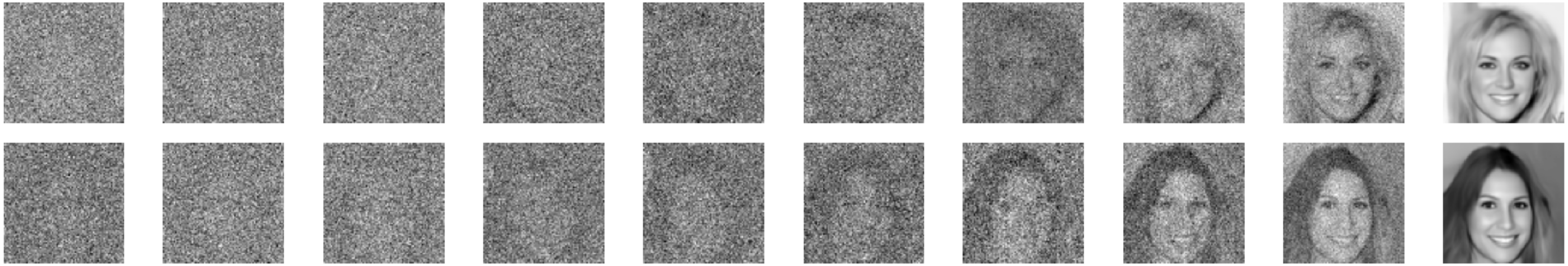
Diffusion models and denoising

noisy image $y = x + z$

$z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

$\hat{f}(y)$ denoised image

Denoiser is applied iteratively and partially



[Kadkhodaie & Simoncelli arXiv2020, NeurIPS2021]

Diffusion models and denoising

noisy image $y = x + z$

$z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

Deep Neural Network

$\hat{f}\,(y)$ denoised image

$$\textit{Minimize}\ \mathbb{E}\left[\|x - f(y)\|^2\right]$$

[Tweedie, via Robbins, 1956;
Miyasawa, 1961]

$$f^\star(y) = \mathop{\mathbb{E}}_{x}[x|y] = \int x p^\star(x|y)\mathrm{d}x \ = \boxed{f^\star(y) = y + \sigma^2 \nabla \log p^\star_\sigma(y)}$$

optimal denoiser

mean of the posterior distribution

Diffusion models and denoising

noisy image $y = x + z$

$z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

$\hat{f}(y)$ denoised image

$$\textit{Minimize} \;\; \mathbb{E}\left[\| x - f(y) \|^2 \right]$$

[Tweedie, via Robbins, 1956;
Miyasawa, 1961]

$$f^\star(y) = \mathop{\mathbb{E}}_x [x|y] = \int x p^\star(x|y) \mathrm{d}x \;=\; \boxed{f^\star(y) = y + \sigma^2 \nabla \log p^\star_\sigma(y)}$$

optimal denoiser

mean of the posterior distribution

$$p^\star_\sigma(y) = \int p(y|x)\, p^\star(x)\, \mathrm{d}x = \int g_\sigma(y - x)\, p^\star(x)\, \mathrm{d}x,$$

$p_\sigma(y)$ is a blurred (diffused) version of $p(x)$

Diffusion models and denoising

noisy image $y = x + z$

$z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

$\hat{f}\,(y)$ denoised image

[Tweedie, via Robbins, 1956;
Miyasawa, 1961]

$$\boxed{f^\star(y) = y + \sigma^2 \nabla \log p^\star_\sigma(y)}$$ **optimal denoiser**

$$\hat{f}\,(y) = y + \sigma^2 \nabla \log \hat{p}_\sigma(y)$$ **learned denoiser**

Coarse-to-fine **gradient ascent**

$$\nabla_y \log \hat{p}_\sigma(y) \approx (\hat{f}\,(y) - y)/\sigma^2$$

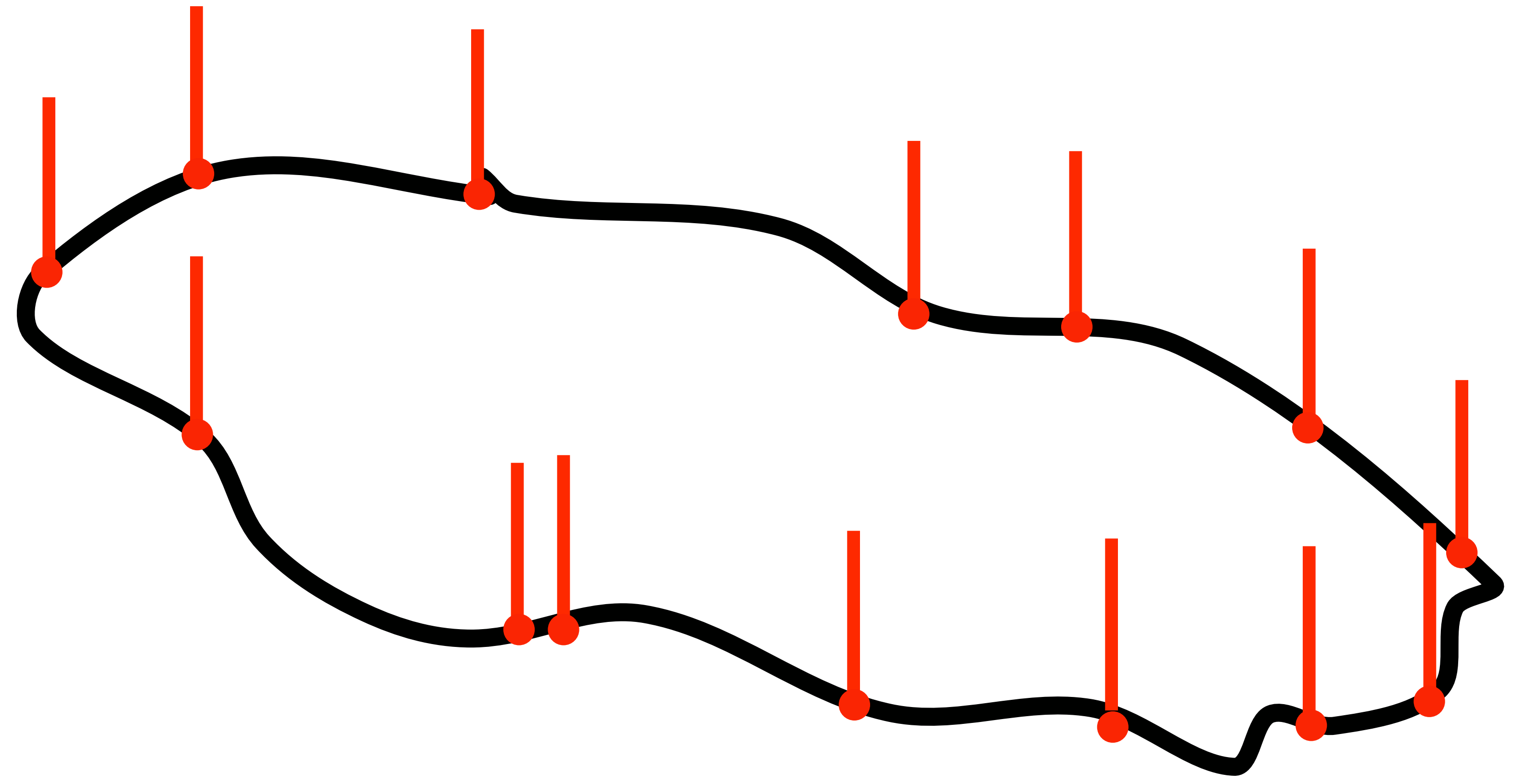
Diffusion models and denoising

$$\boxed{f^\star(y) = y + \sigma^2 \nabla \log p^\star_\sigma(y)}$$ optimal denoiser

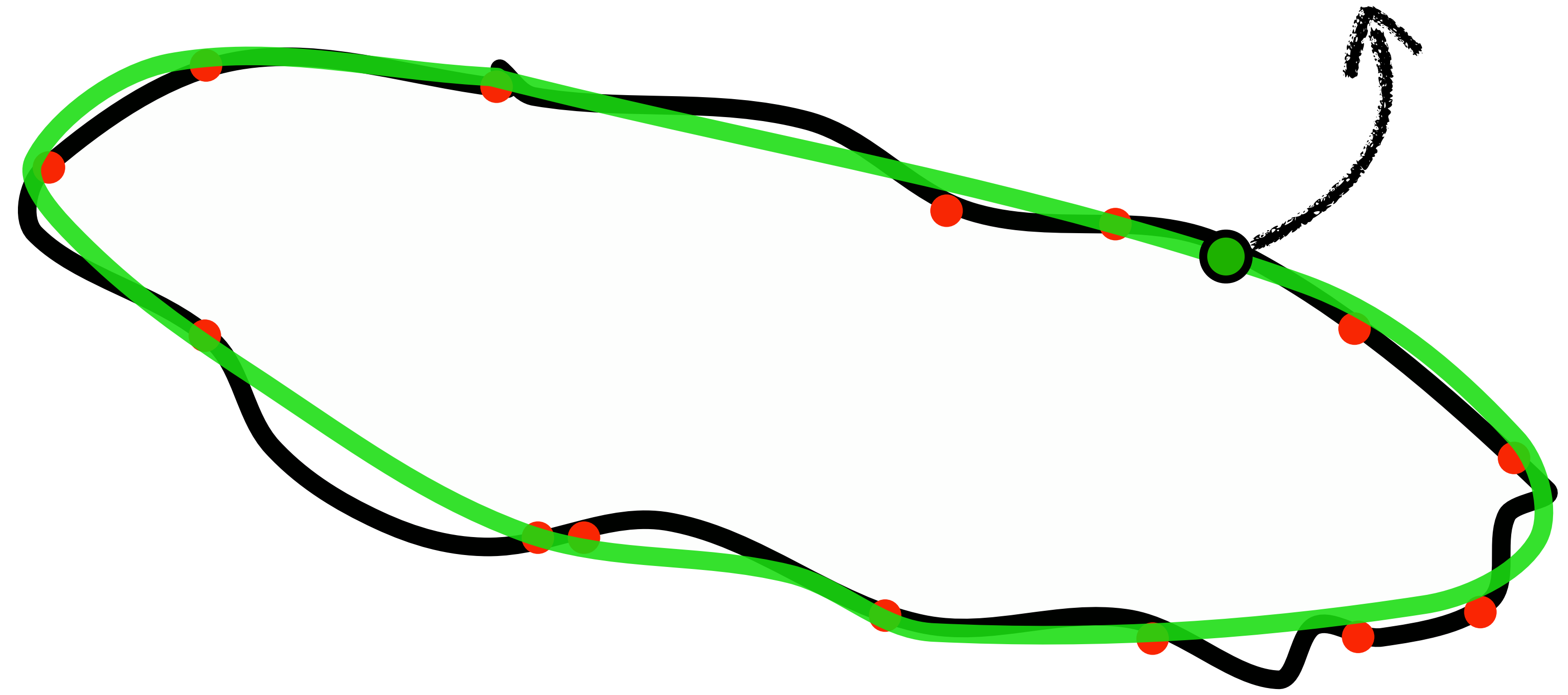
$$\hat{f}(y) = y + \sigma^2 \nabla \log \hat{p}_\sigma(y)$$ **learned denoiser**

Samples from this

Diffusion models and denoising



$$\hat{p}_\sigma(y) \overset{?}{\approx} p^\star_\sigma(y)$$



$$\boxed{f^\star(y) = y + \sigma^2 \nabla \log p^\star_\sigma(y)}$$ **optimal denoiser**

$$\hat{f}(y) = y + \sigma^2 \nabla \log \hat{p}_\sigma(y)$$ **learned denoiser**

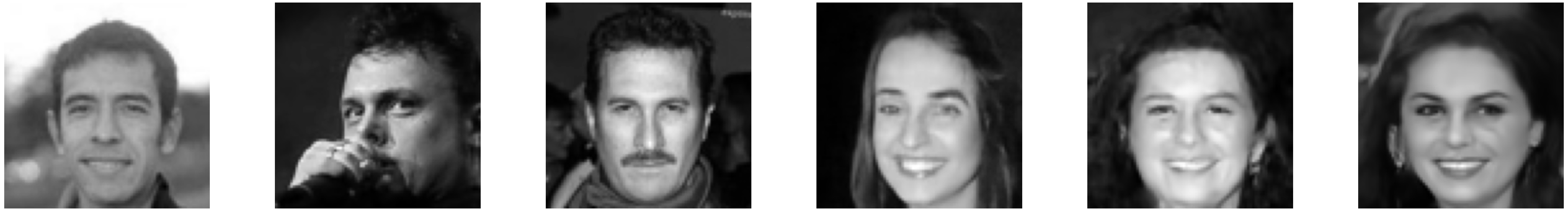
↓

Samples from this

Transition from memorization to generalization

Training set size: 1 10 100 1,000 10,000 100,000

Samples, model trained on set A:



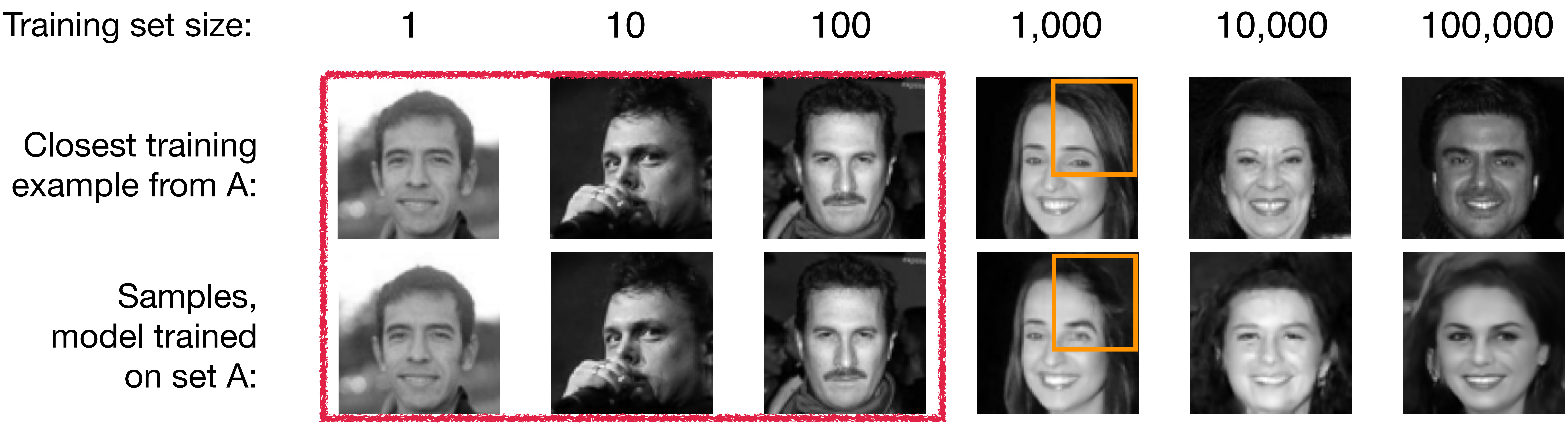
Transition from memorization to generalization

Training set size: 1 | 10 | 100 | 1,000 | 10,000 | 100,000

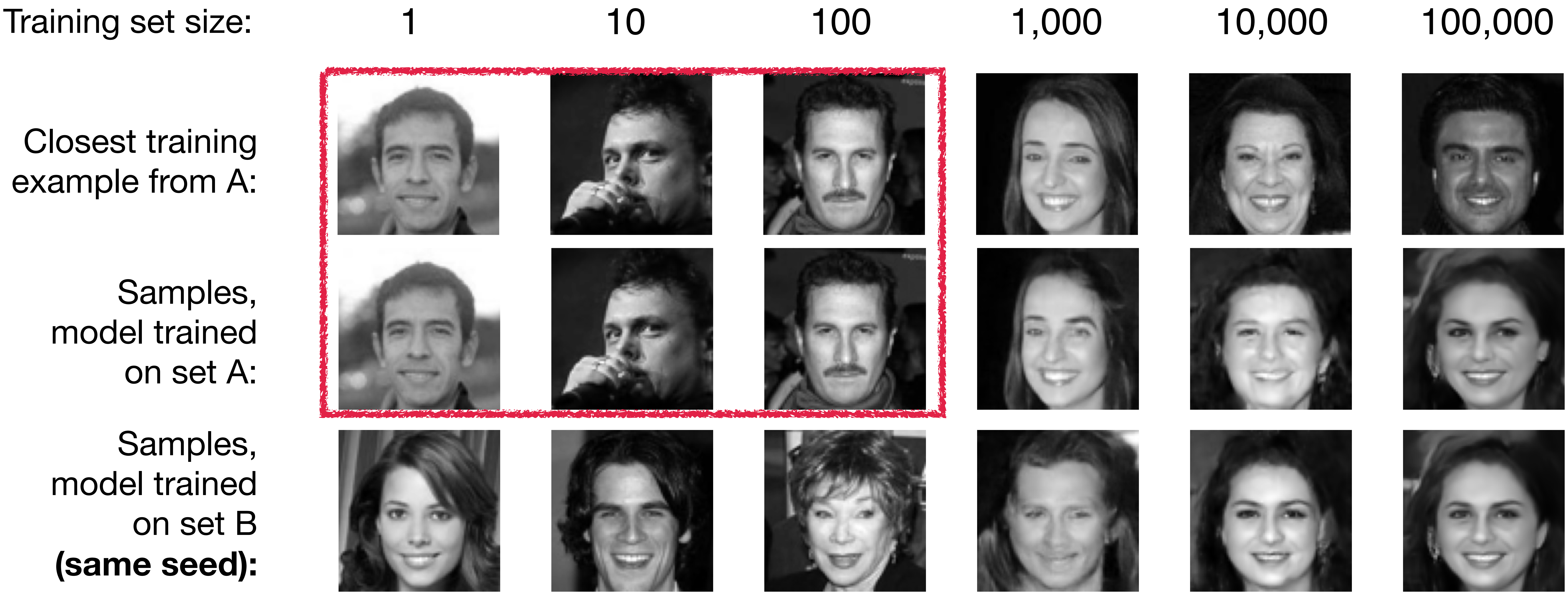
Closest training example from A:

Samples, model trained on set A:

Transition from memorization to generalization

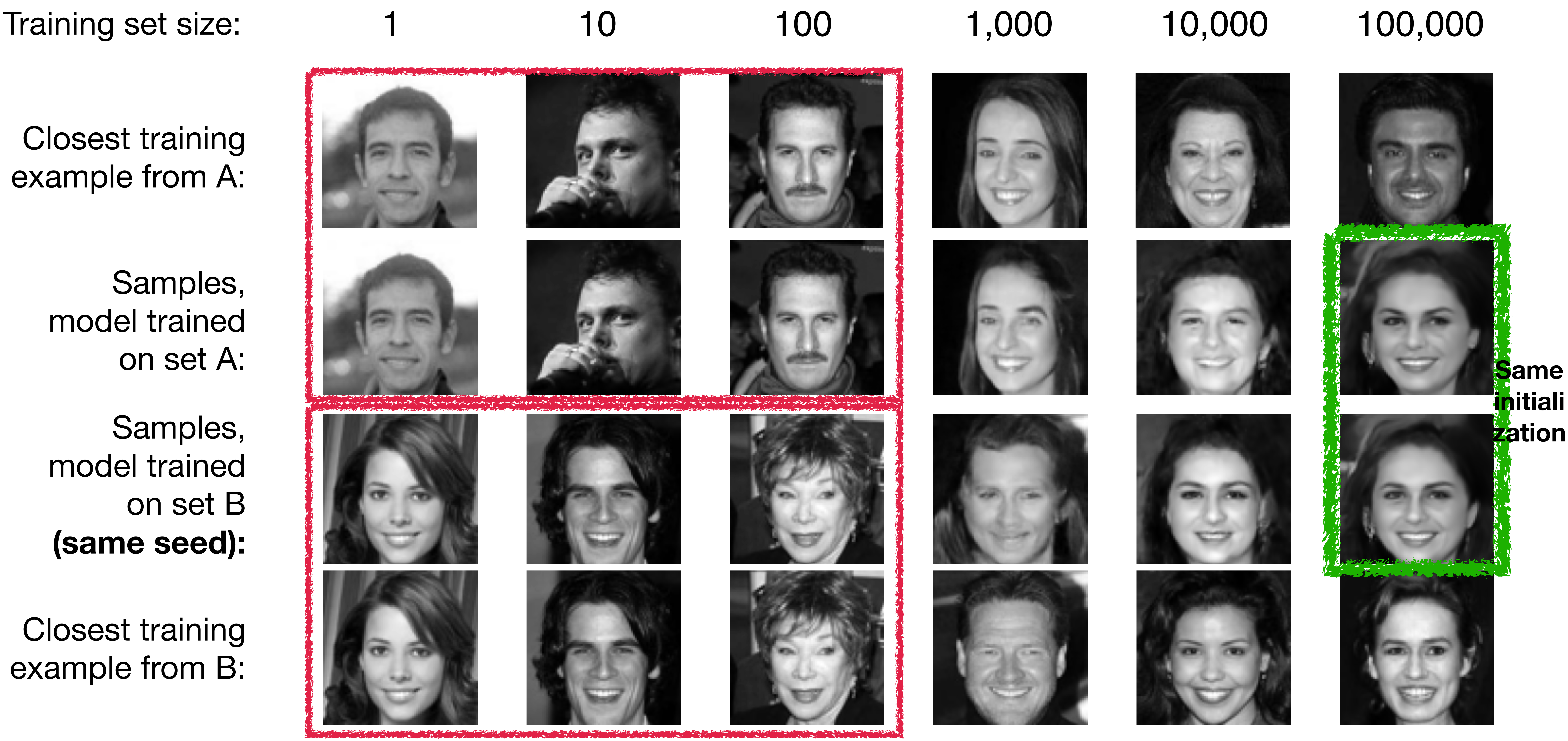


Transition from memorization to generalization

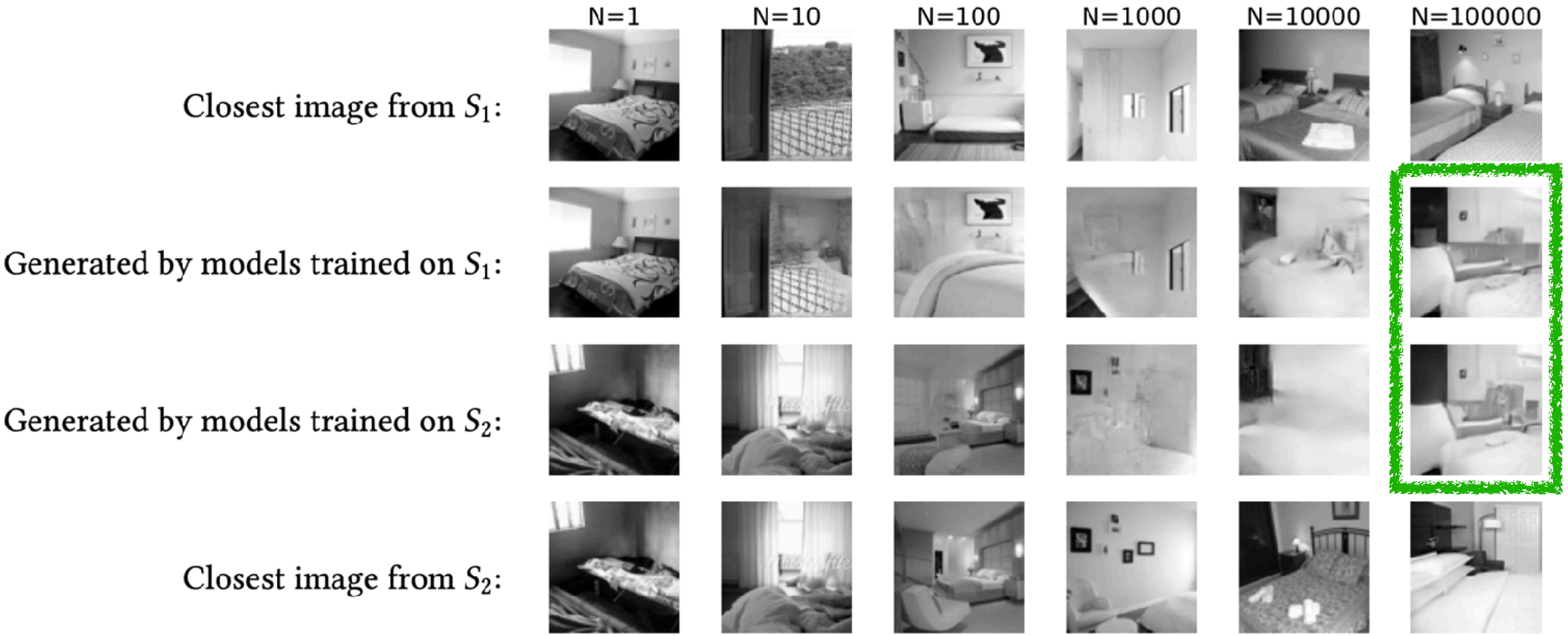


Transition from memorization to generalization

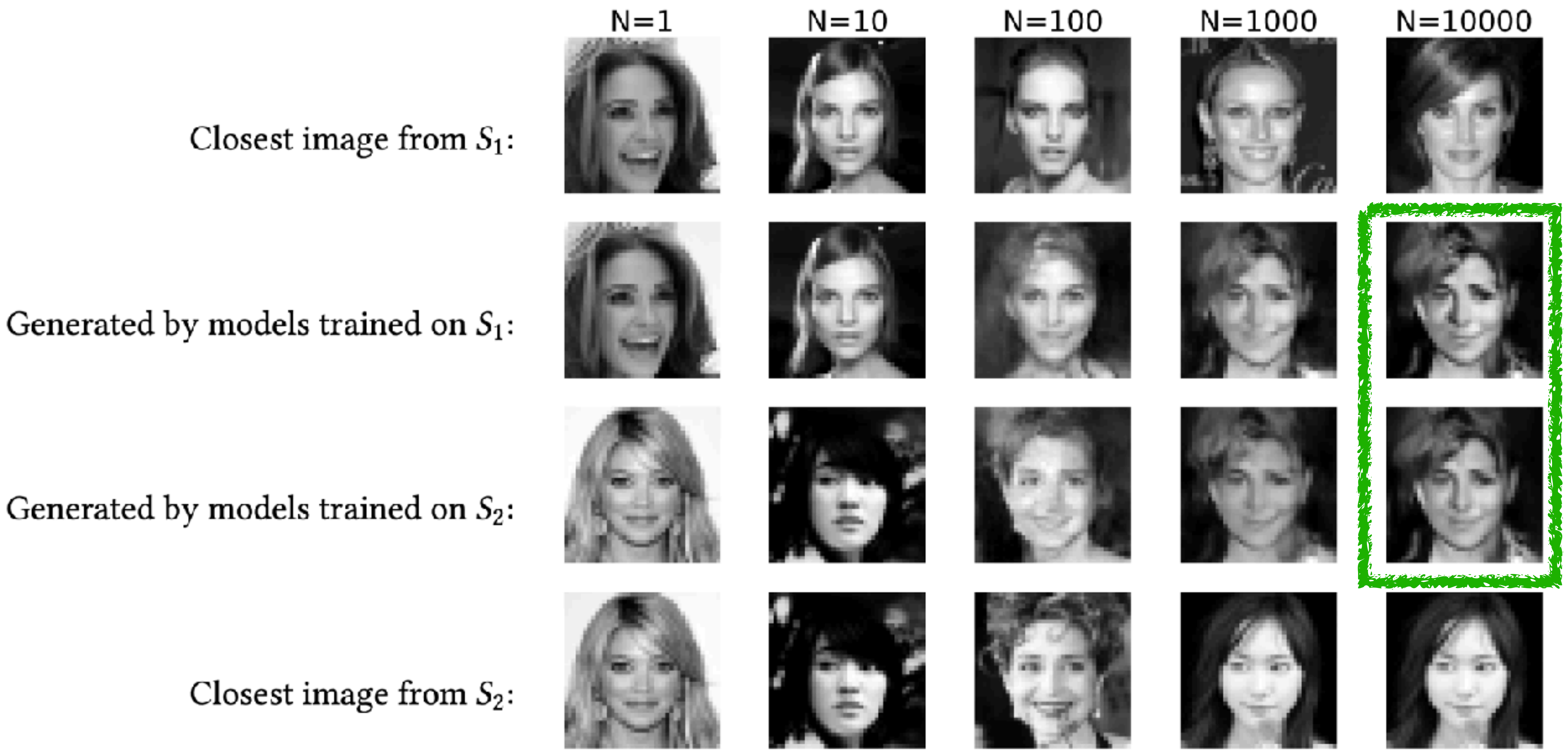
Transition from memorization to generalization



Strong generalization in LSUN bedroom dataset



Strong generalization in BF-CNN architecture



How do diffusion models generalize?

What are inductive biases of the denoiser?

Denoising as shrinkage in a basis

Classical framework for denoising:

1. Transform the noisy image to a basis where noise and signal are separable

2. Suppress the noise (shrinkage)

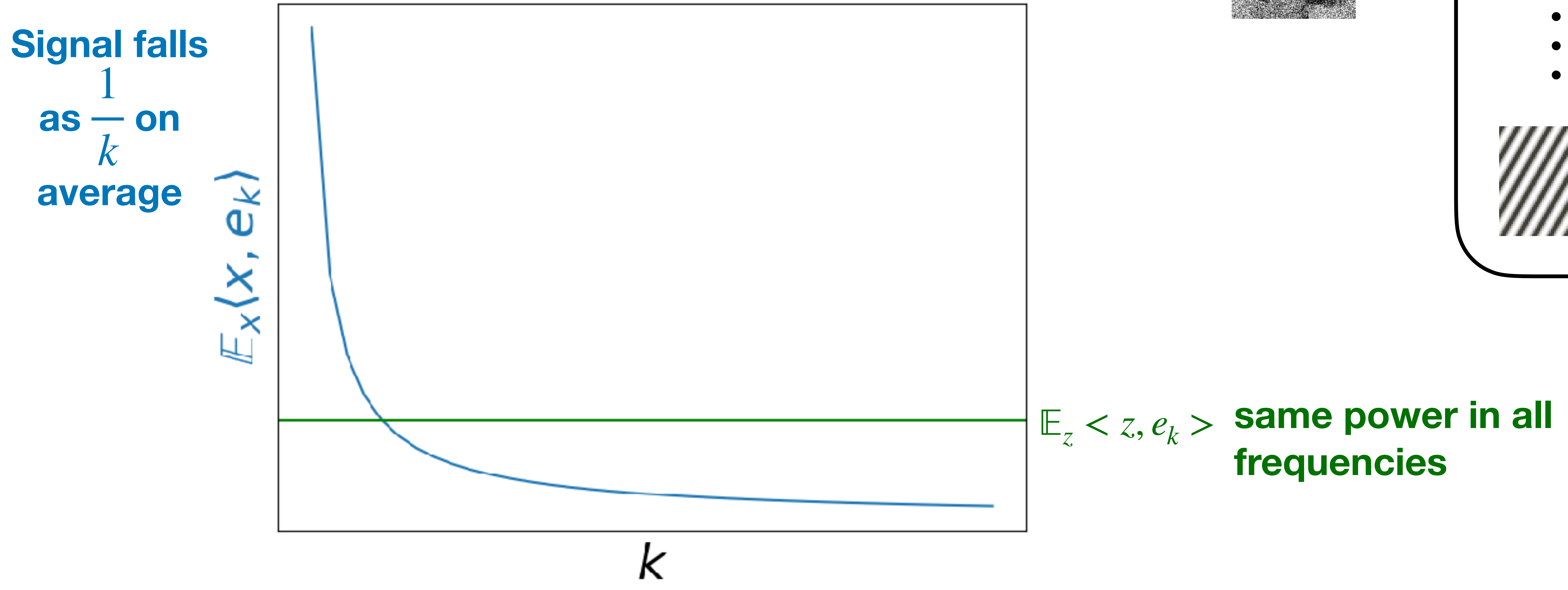
3. Transform back to pixel domain

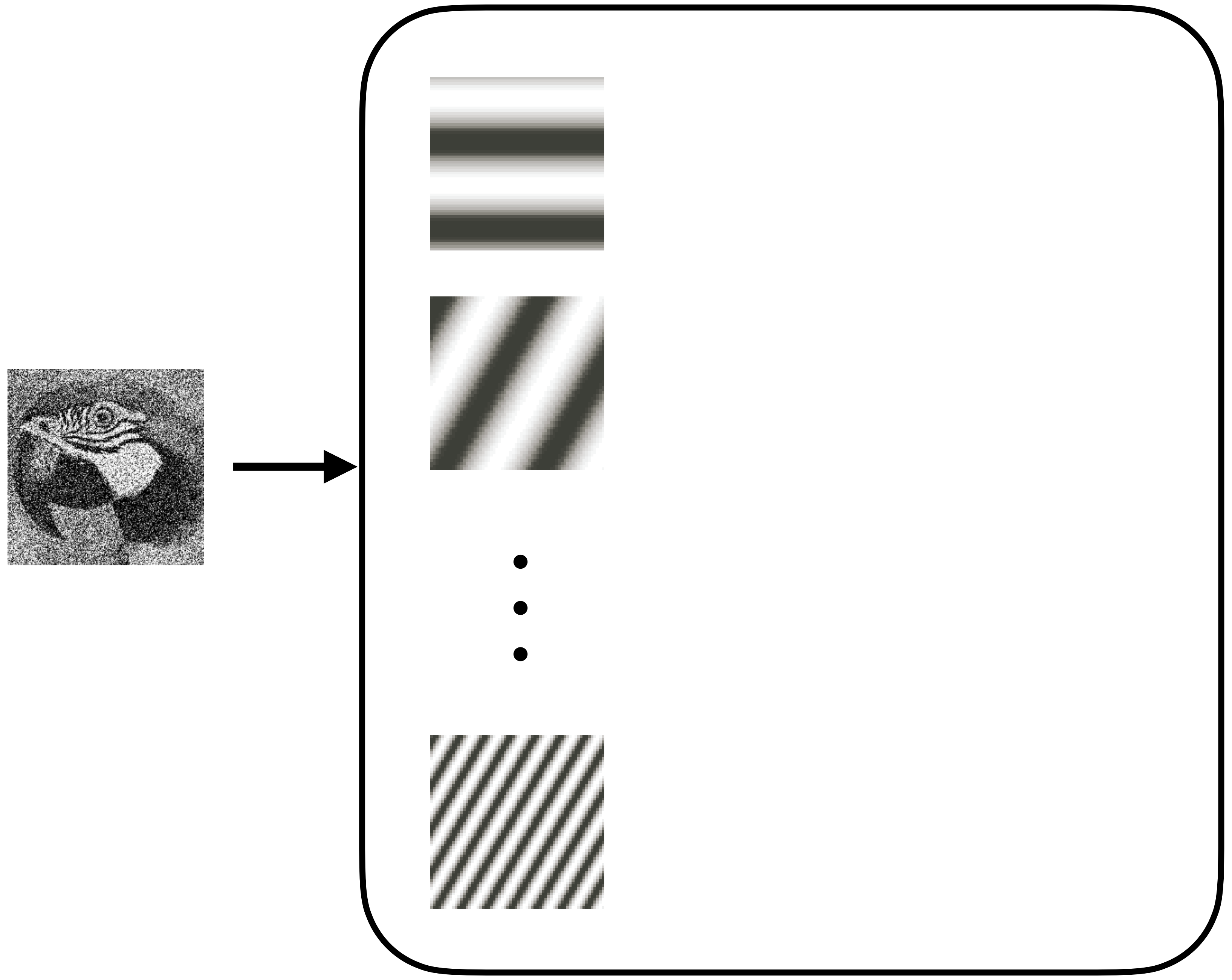
Denoising as shrinkage in a basis

Fixed basis, fixed shrinkage

$$f(y) = \sum_k \lambda_k \langle y, e_k \rangle e_k$$

Fourier basis





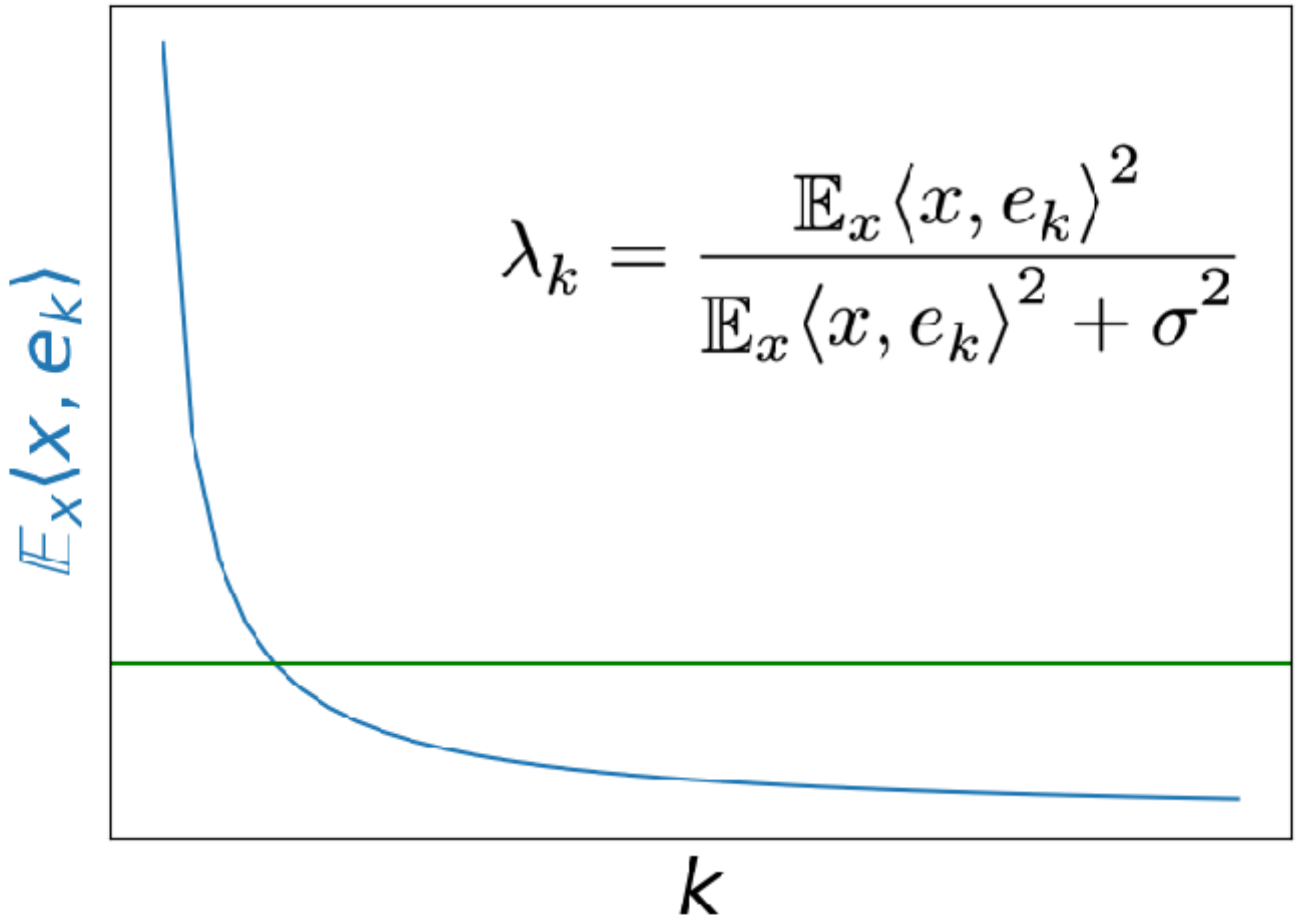
Denoising as shrinkage in a basis

Fixed basis, fixed shrinkage

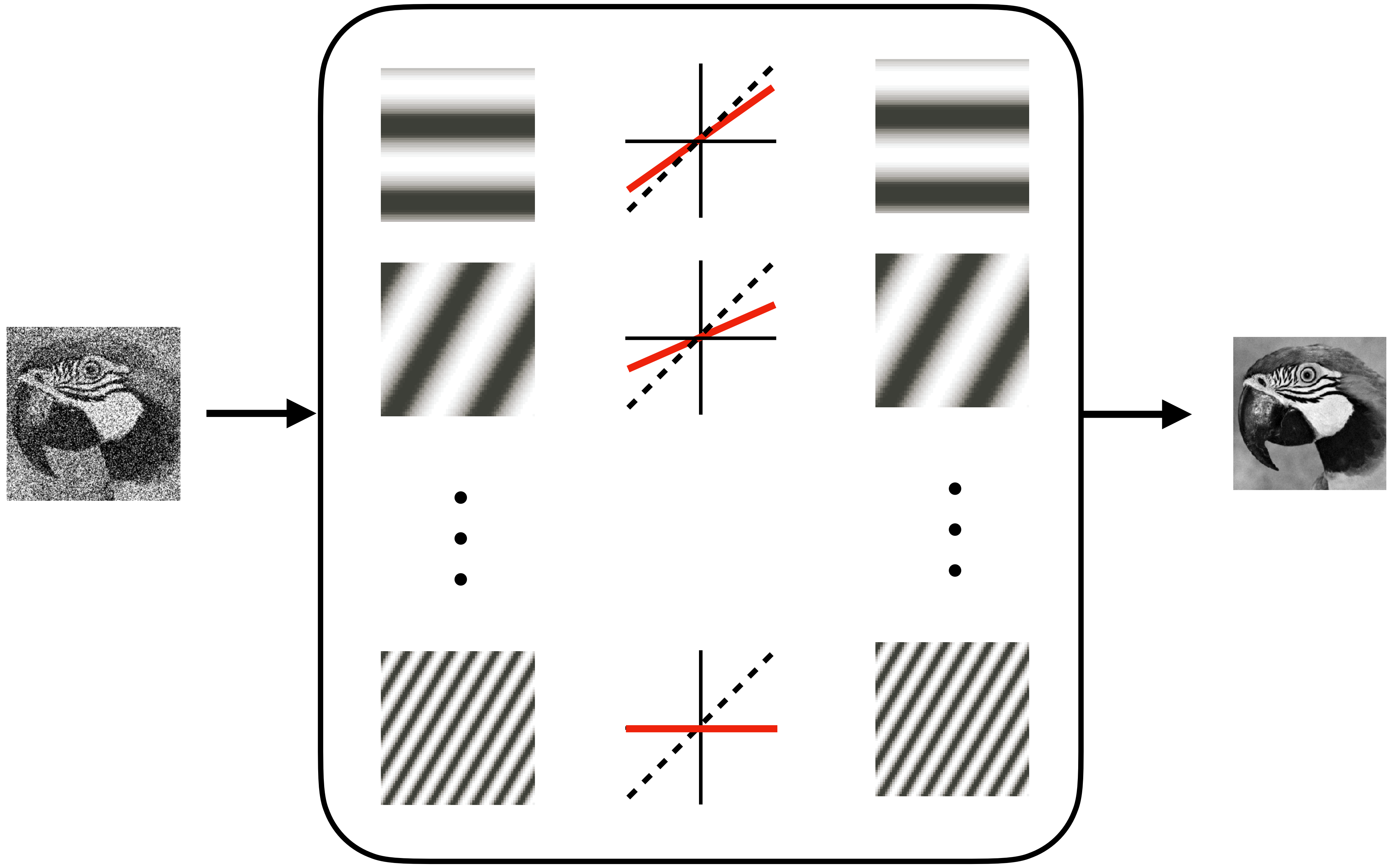
$$f(y) = \sum_k \lambda_k \langle y, e_k \rangle e_k$$

Fourier basis

Signal falls as $\frac{1}{k}$ on average



$\mathbb{E}_z < z, e_k >$ **same power in all frequencies**



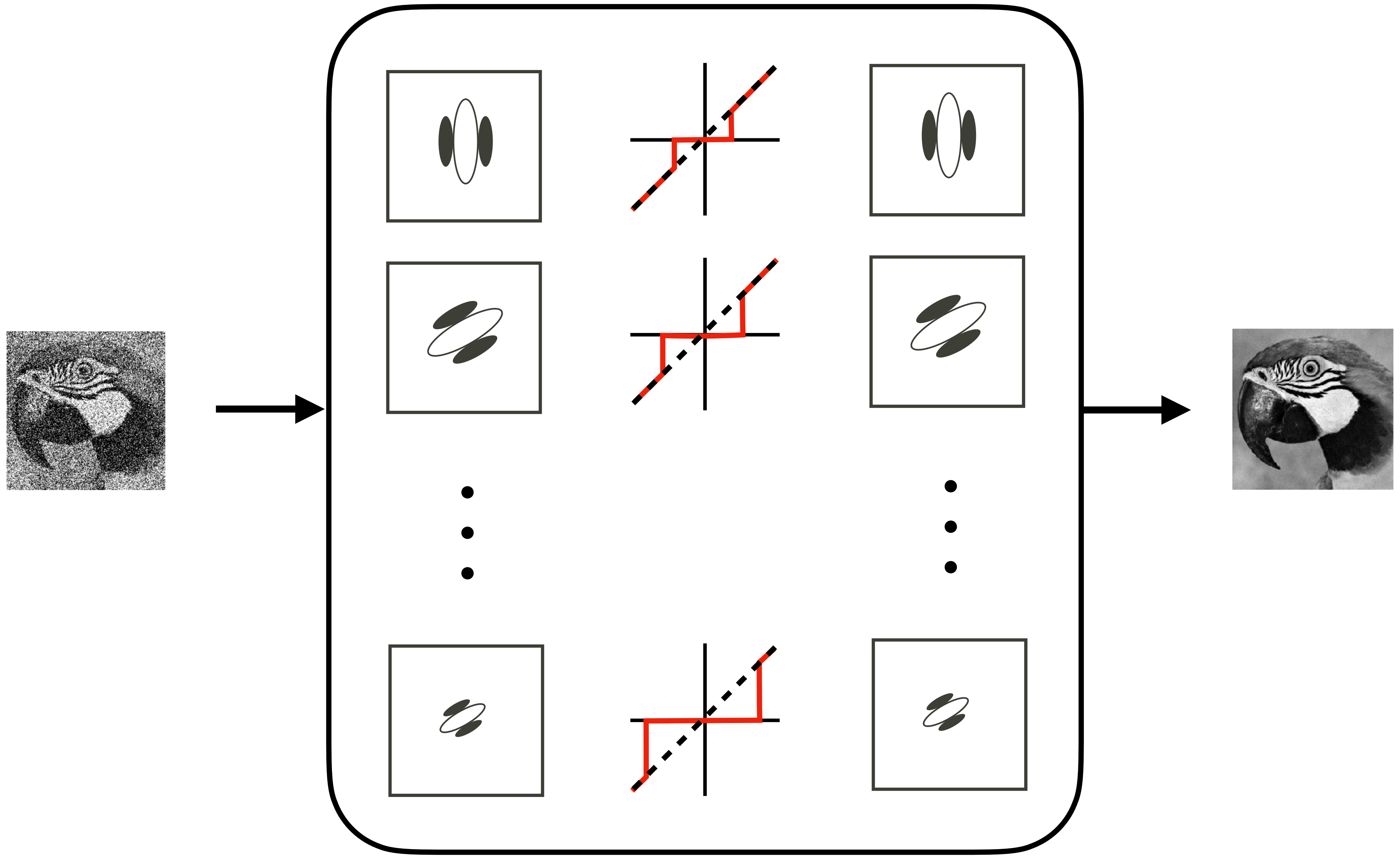
"Wiener filter" (1942)

Denoising as shrinkage in a basis

Fixed basis, adaptive shrinkage

$$f(y) = \sum_k \lambda_k(y) \langle y, e_k \rangle e_k$$

Wavelet basis



“wavelet thresholding”

[Donoho & Johnstone 94]

☑ Coefficients fall faster in wavelet basis. More compact representation of signal.
Easier separation between noise and signal with sparse signal

Denoising as shrinkage in a basis

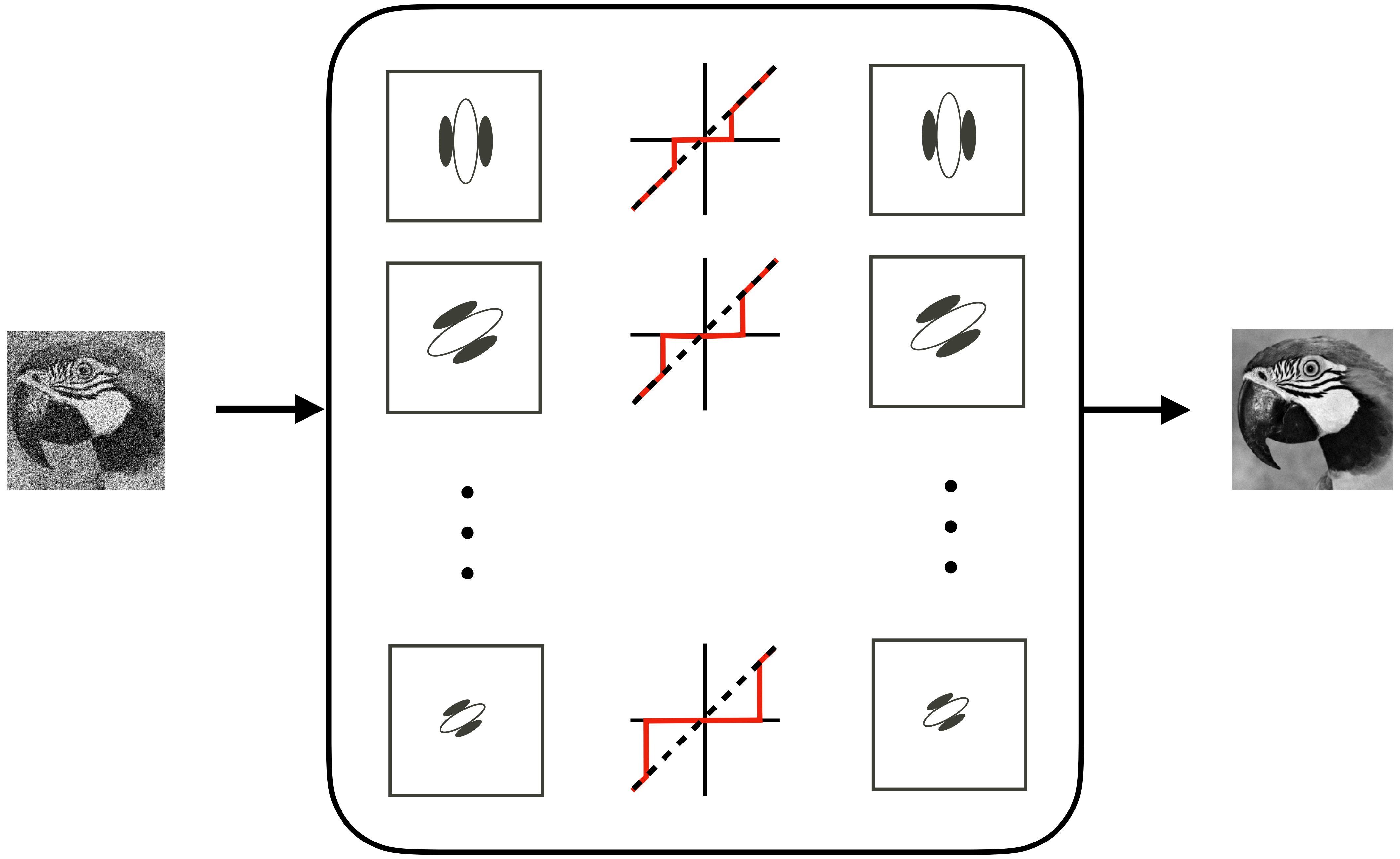
Fixed basis, adaptive shrinkage

$$f(y) = \sum_k \boxed{\lambda_k(y)} \langle y, e_k \rangle \, e_k$$

Wavelet basis

Adaptive thresholding

$$\lambda_k(y) = \begin{cases} 1 & |\langle y, e_k \rangle| > \alpha\sigma \\ 0 & \text{Otherwise} \end{cases}$$



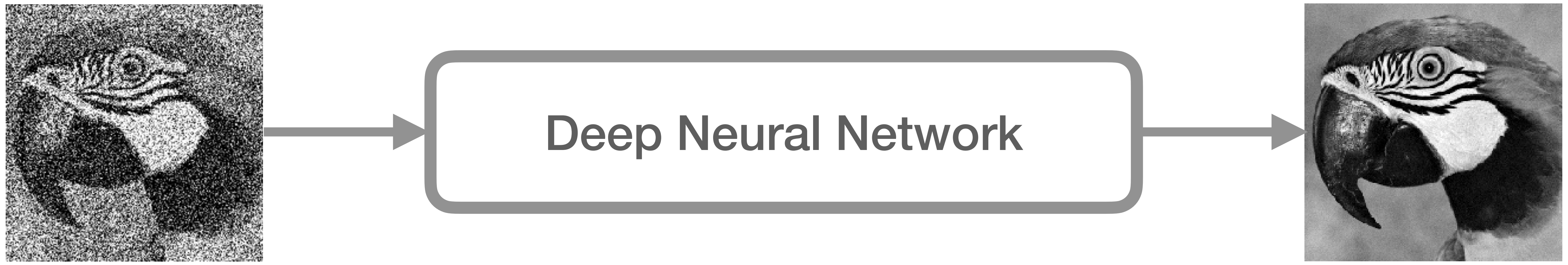
"wavelet thresholding"

[Donoho & Johnstone 94]

☑ Coefficients fall faster in wavelet basis. More compact representation of signal.
Easier separation between noise and signal with sparse signal

Denoising as shrinkage in a basis

Adaptive basis, adaptive shrinkage



Locally linear function:

$$\hat{f}(y) = \underbrace{\nabla \hat{f}(y)}\, y$$

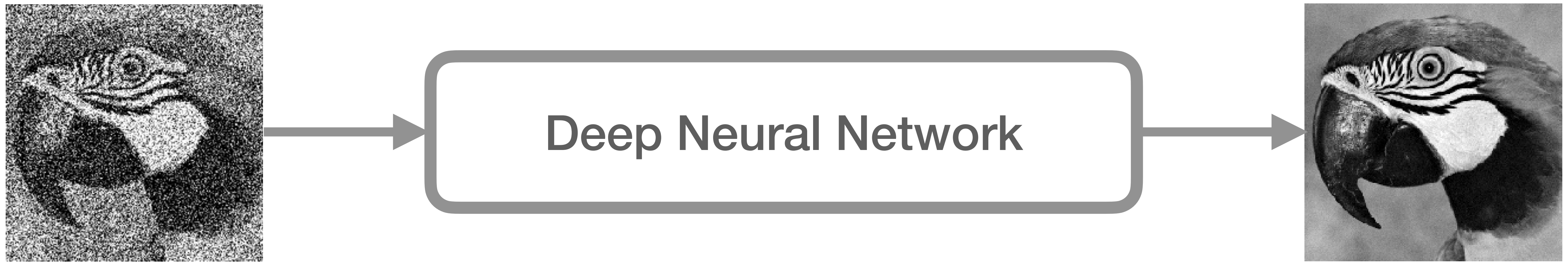
Jacobian w.r.t. Input y

Nearly symmetric

[Mohan*, Kadkhodaie*, Simoncelli, Fernandez-Granda, ICLR 2020]

Denoising as shrinkage in a basis

Adaptive basis, adaptive shrinkage



Eigen decomposition of Jacobian

Locally linear function:

$$\hat{f}(y) = \nabla \hat{f}(y)\, y = \sum_k \lambda_k(y) \, \langle y, e_k(y) \rangle \, e_k(y)$$

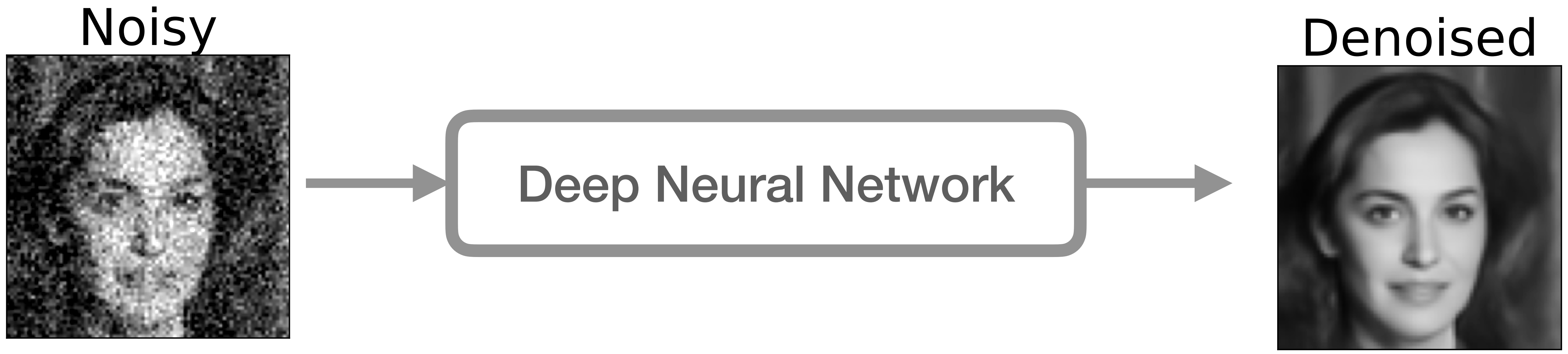
Shrinkage factors

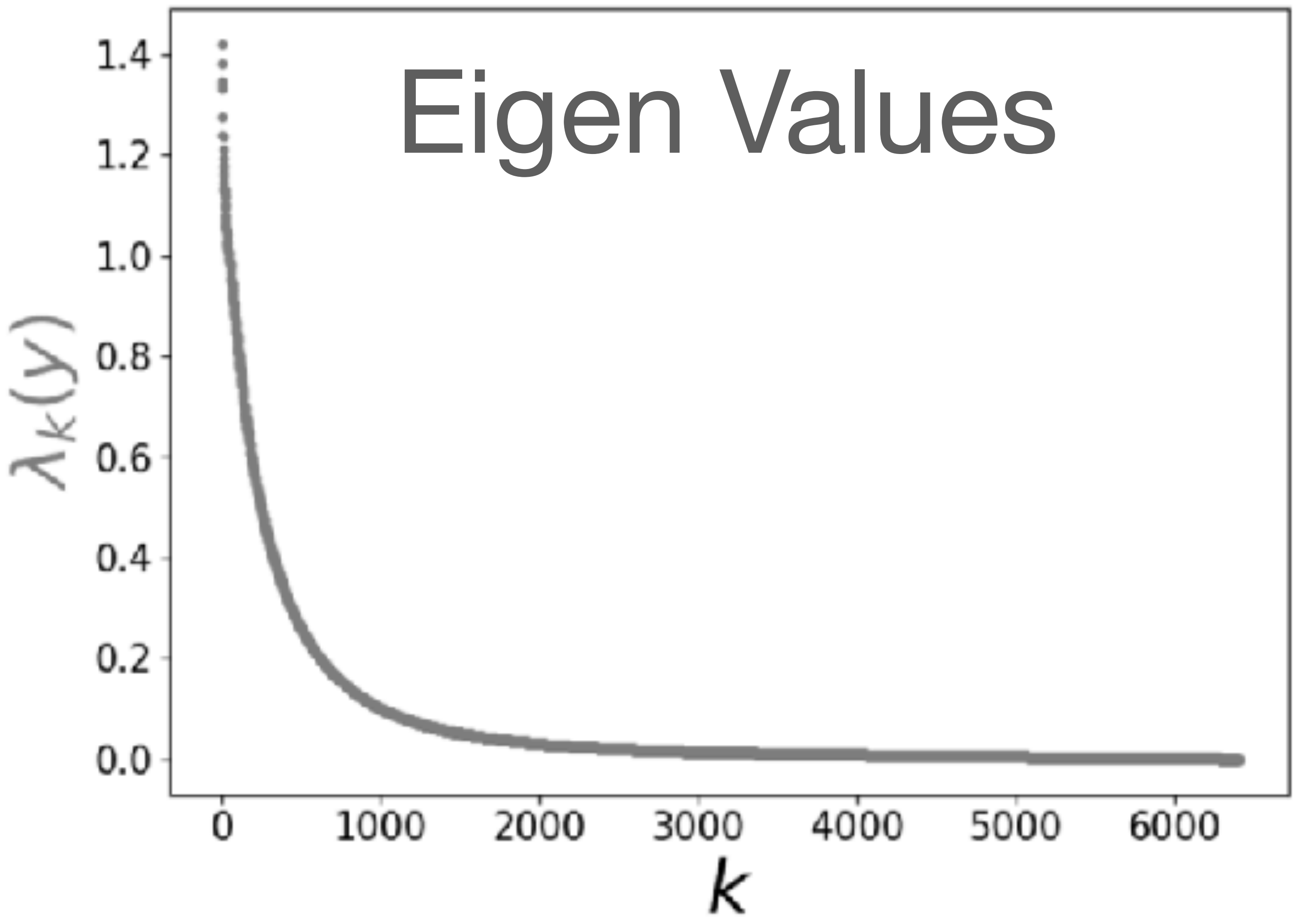
Eigen basis

[Mohan*, Kadkhodaie*, Simoncelli, Fernandez-Granda, ICLR 2020]

Denoising as shrinkage in an adaptive basis

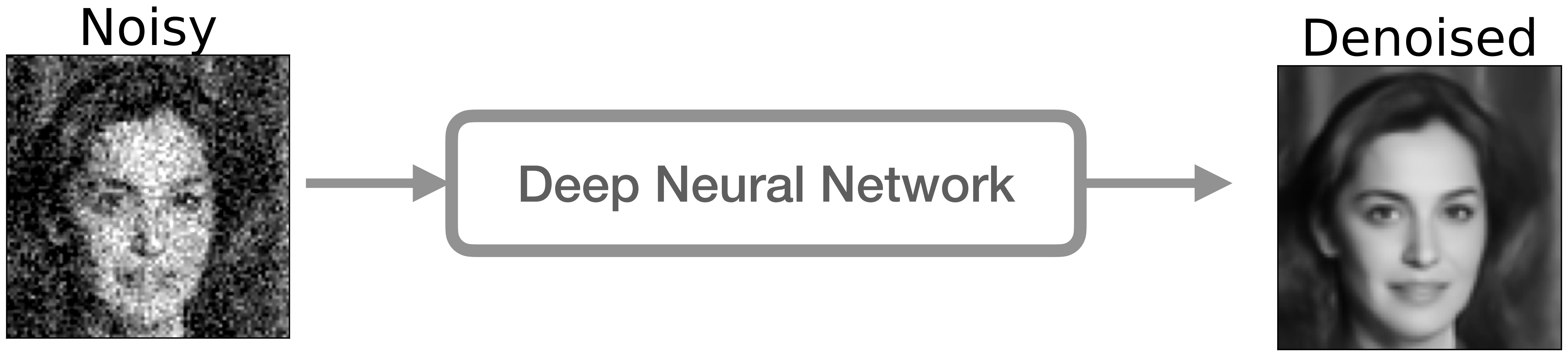
Adaptive basis, adaptive shrinkage

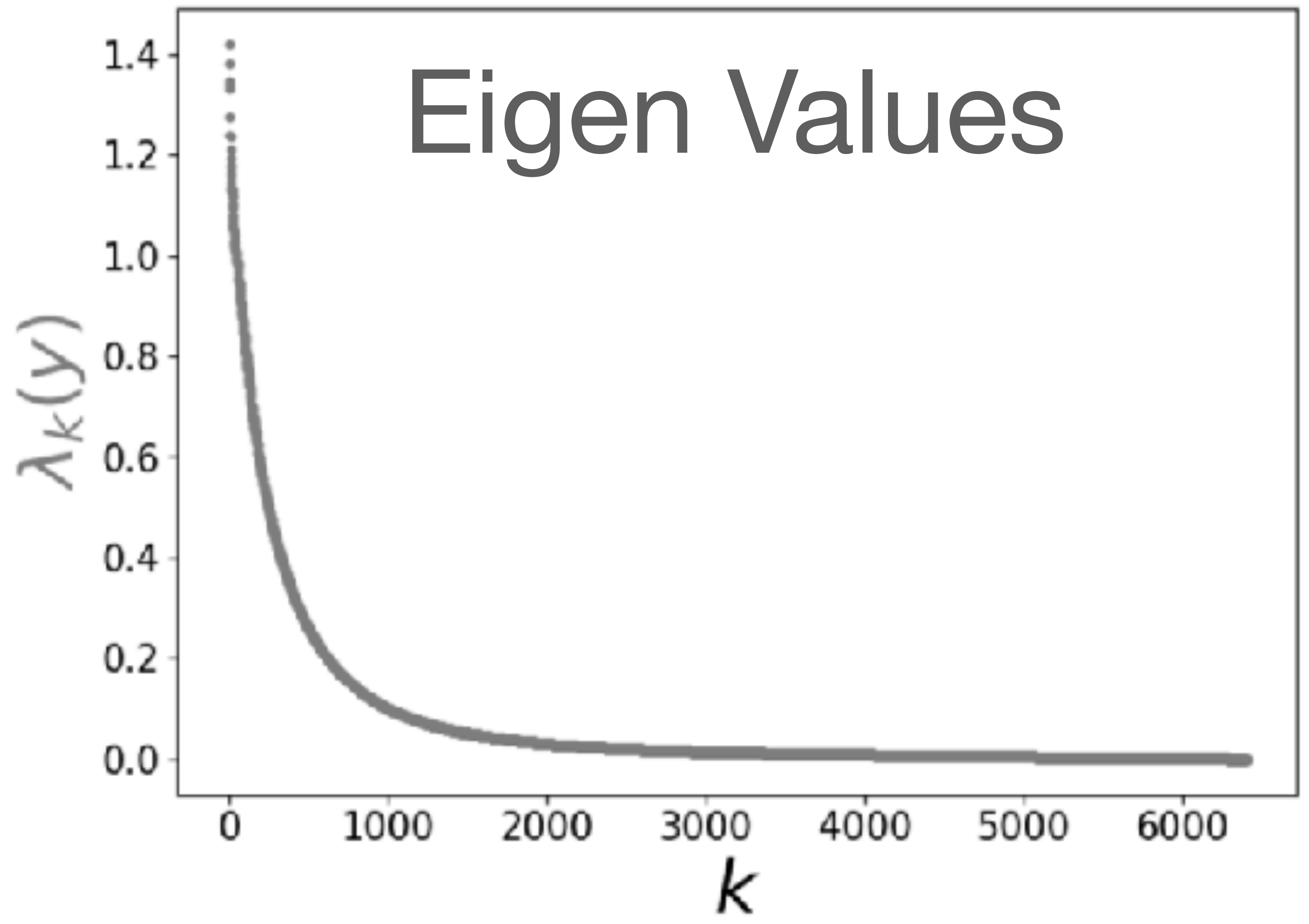




Denoising as shrinkage in an adaptive basis

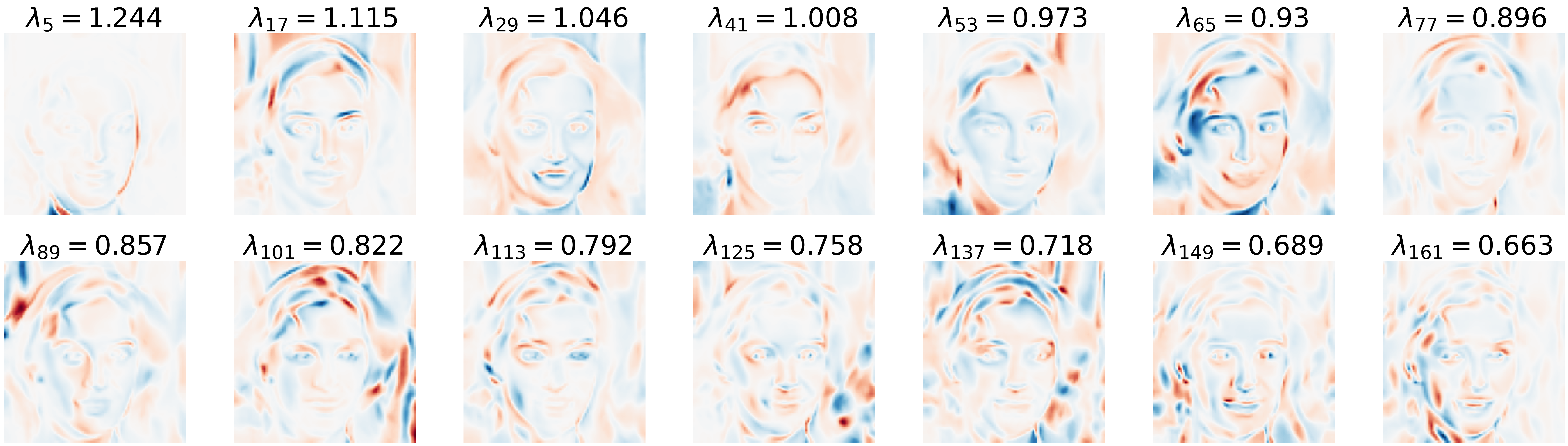
Geometry Adaptive Harmonic Basis (GAHBs)





Some top Eigenvectors

1. Adaptive
2. oscillatory



Denoising as shrinkage in an adaptive basis

Geometry Adaptive Harmonic Basis (GAHBs)

hypothesis:

DNN denoisers have inductive biases towards learning GAHBs

Denoising as shrinkage in an adaptive basis

Geometry Adaptive Harmonic Basis (GAHBs)

hypothesis:

DNN denoisers have inductive biases towards learning GAHBs

How to test this?

Synthetic images

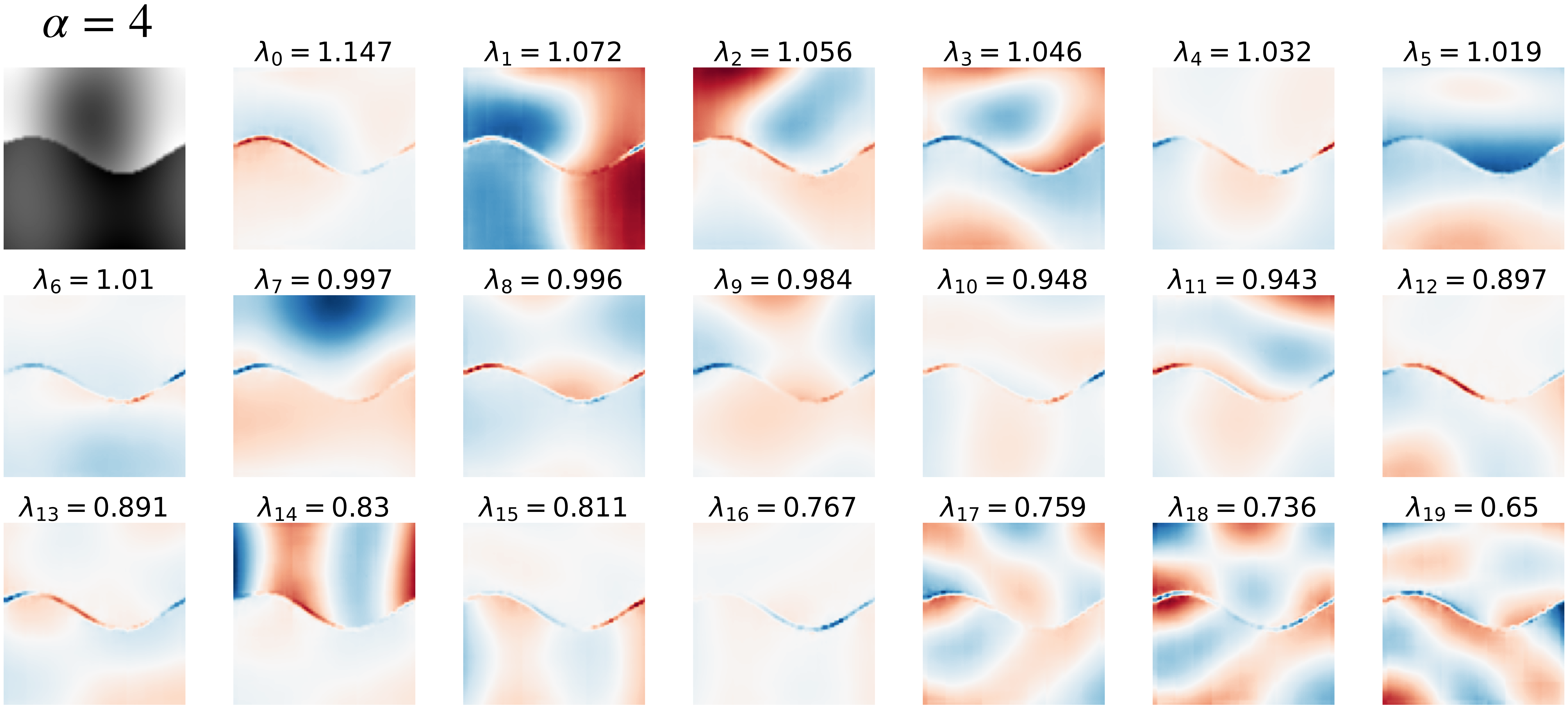
Geometric C^α images

$\alpha = 1$ $\alpha = 2$ $\alpha = 3$ $\alpha = 4$ $\alpha = 5$

GAHBs are optimal for denoising these

[Korostelev & Tsybakov, 1993; Donoho, 1999; Peyré & Mallat, 2008]

Geometric C^α images



geometry adaptive harmonic basis

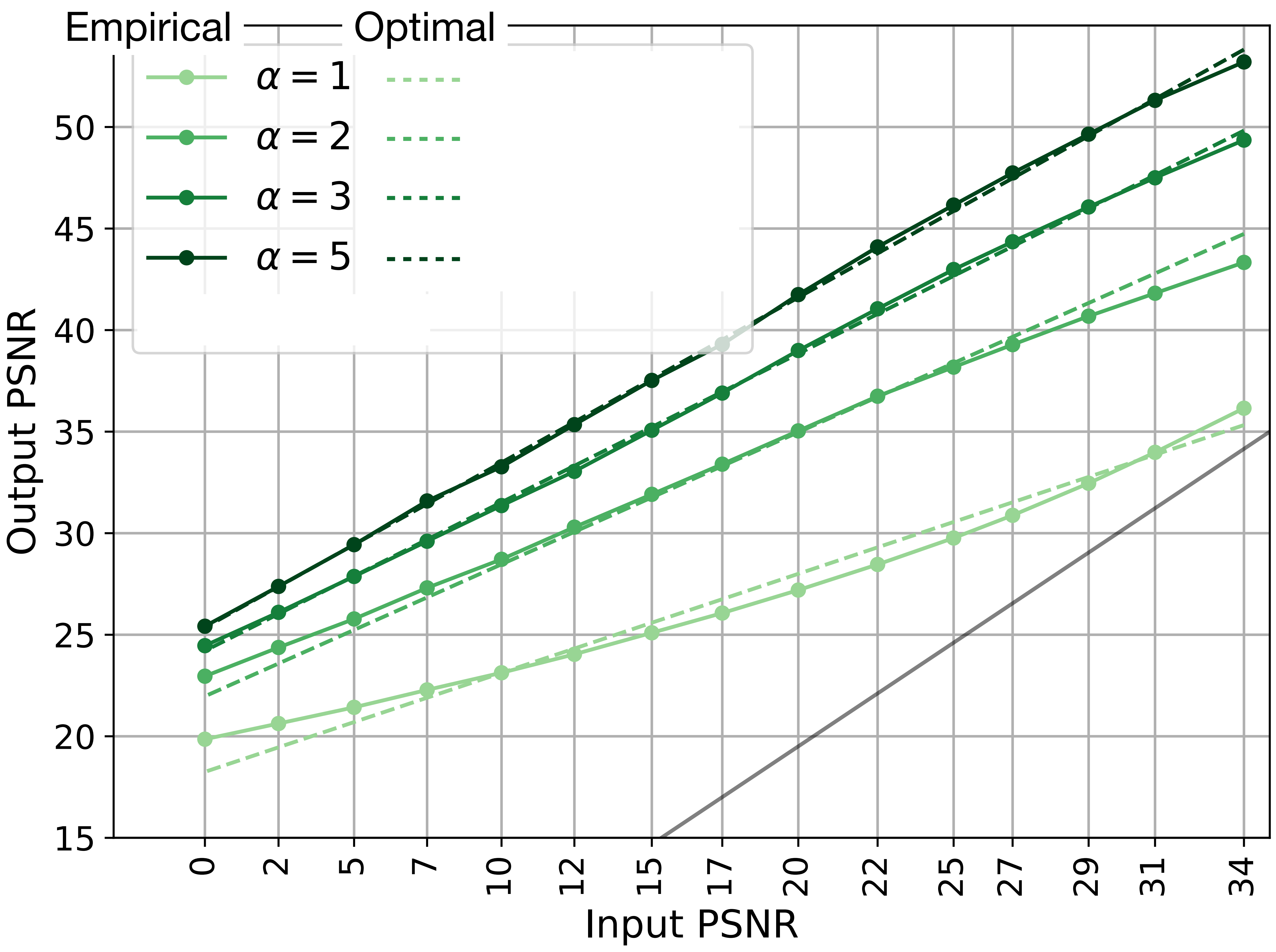
Geometric C^α images

Optimal denoiser on C^α images has slope $\dfrac{\alpha}{\alpha+1}$.

[Korostelev & Tsybakov, 1993]

[Peyré & Mallat, 2008]

Denoising performance



Geometric C^α images

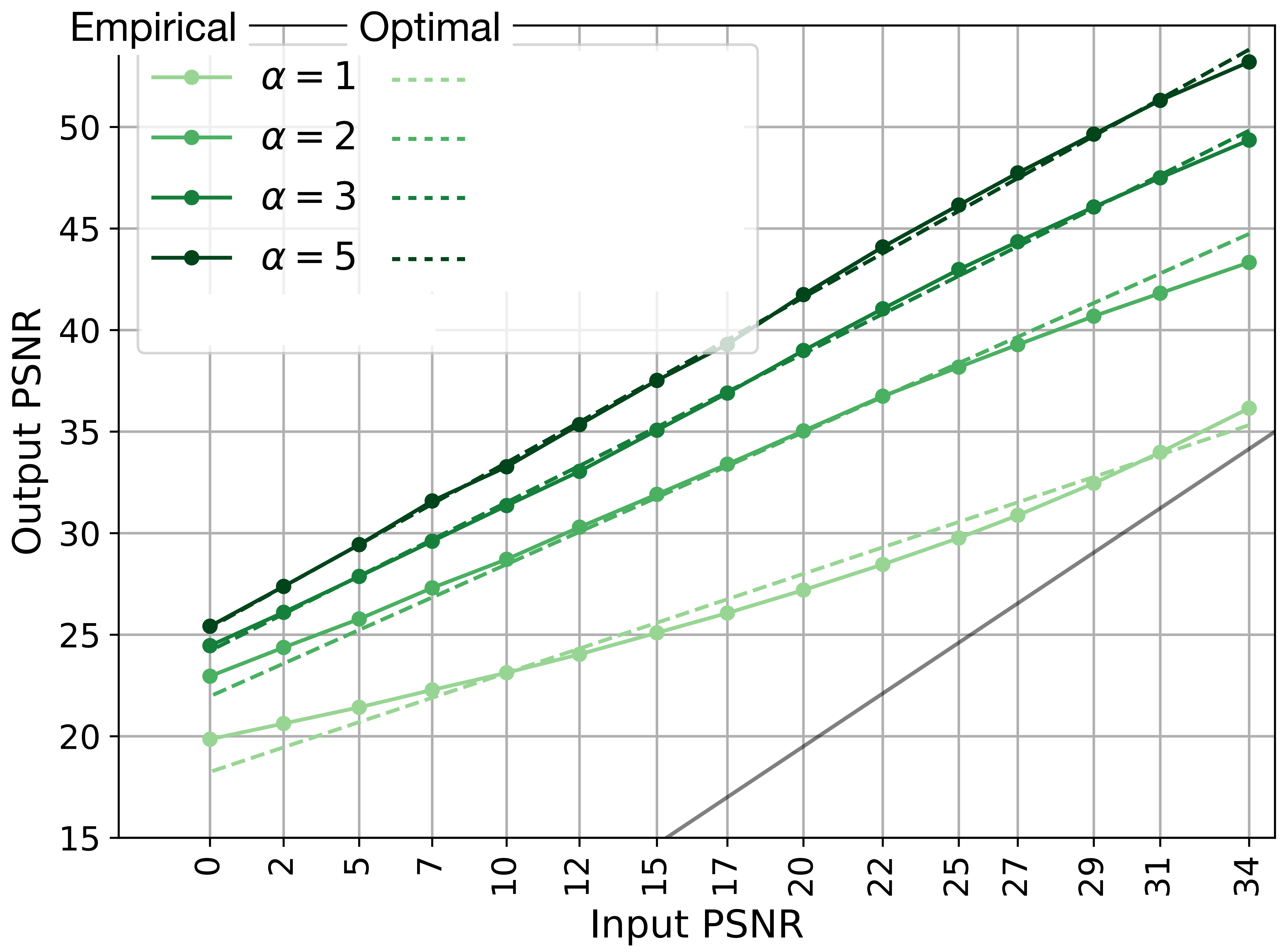
Optimal denoiser on C^α images has slope $\dfrac{\alpha}{\alpha+1}$.

[Korostelev & Tsybakov, 1993]

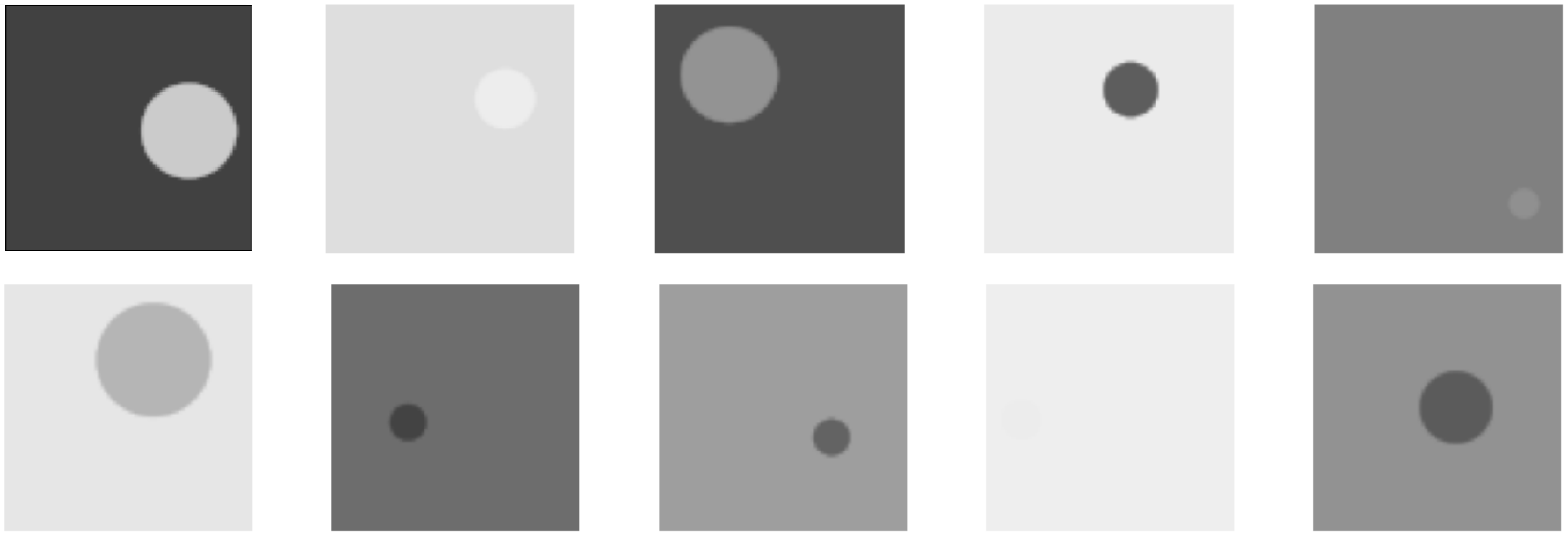
[Peyré & Mallat, 2008]

✔ **Deep nets learn GAHB for denoising when it's optimal**

Denoising performance



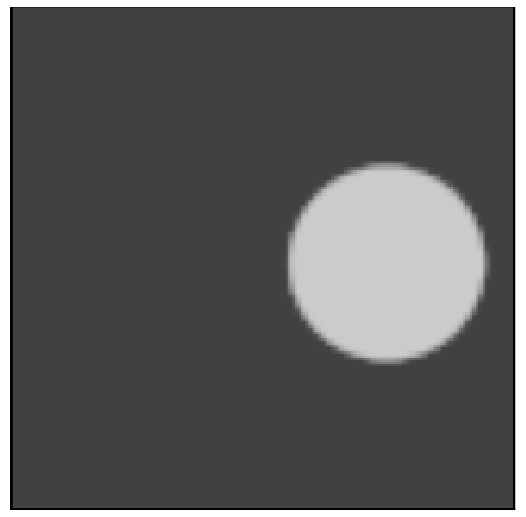
Manifold of disks

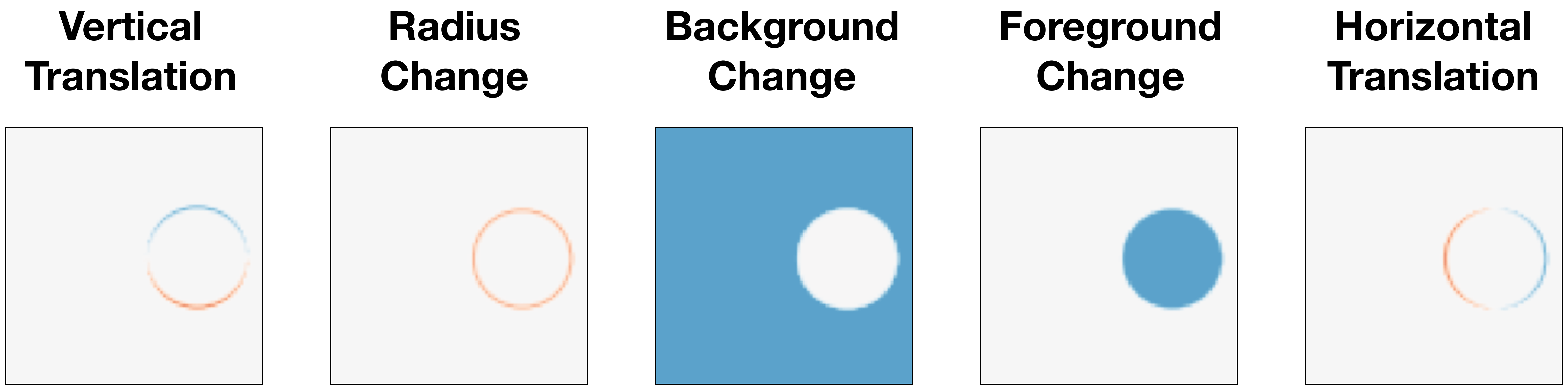


five-dimensional *curved* manifold

1. Vertical position
2. Horizontal position
3. Radius/size
4. Foreground intensity
5. Background intensity

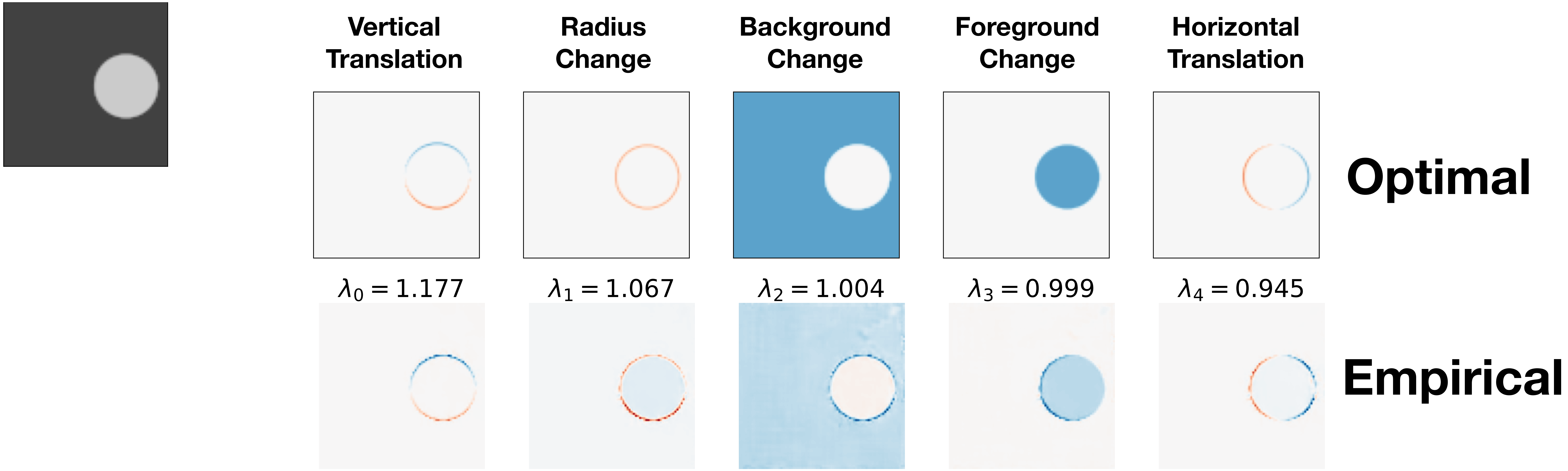
Manifold of disks



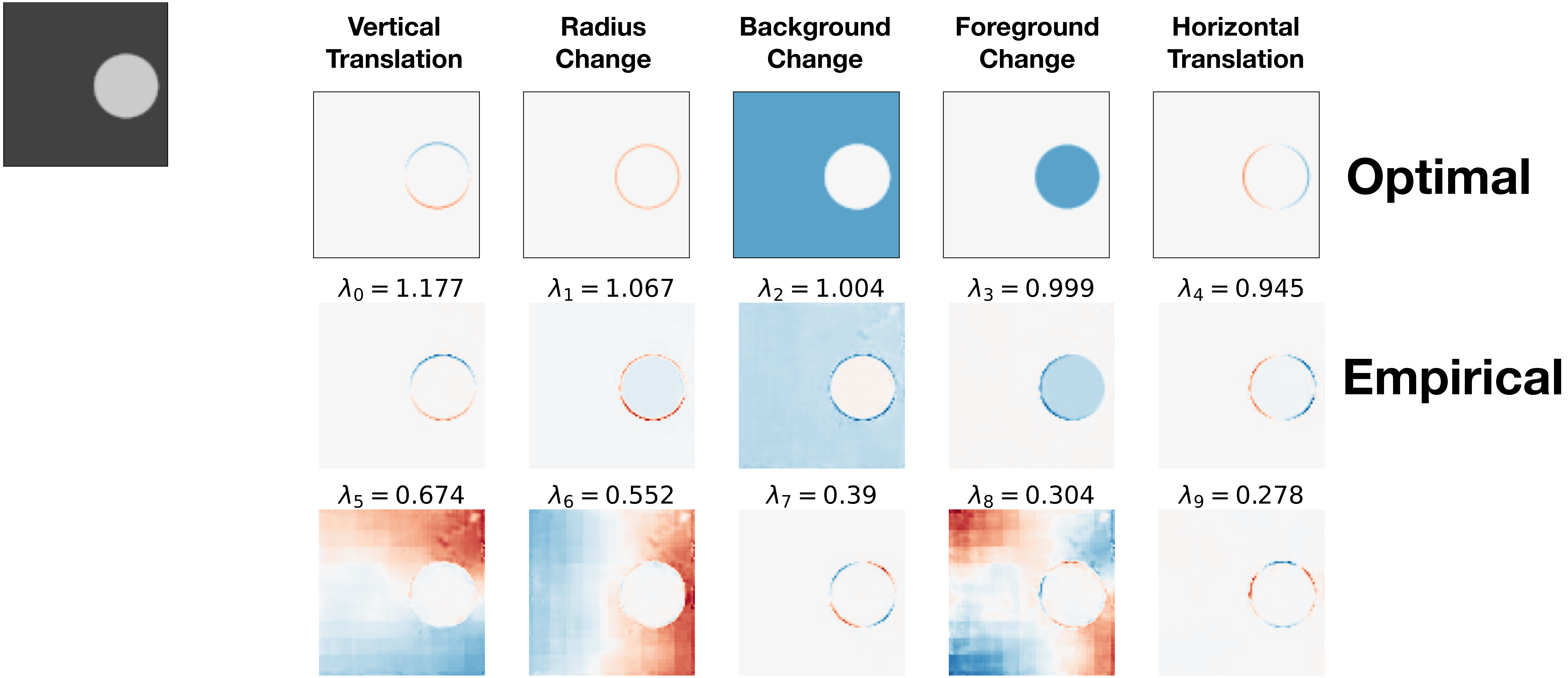


Optimal

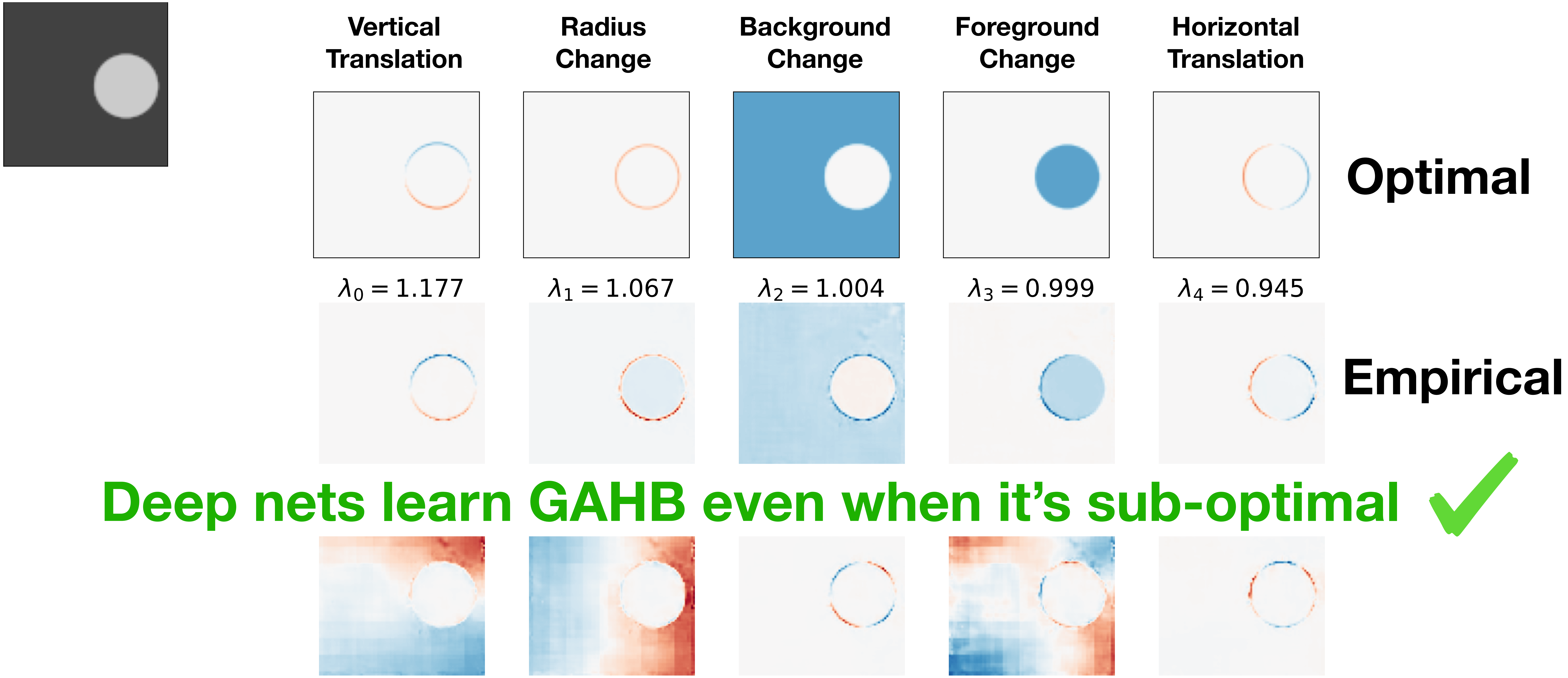
Manifold of disks



Manifold of disks



Manifold of disks



Interim summary

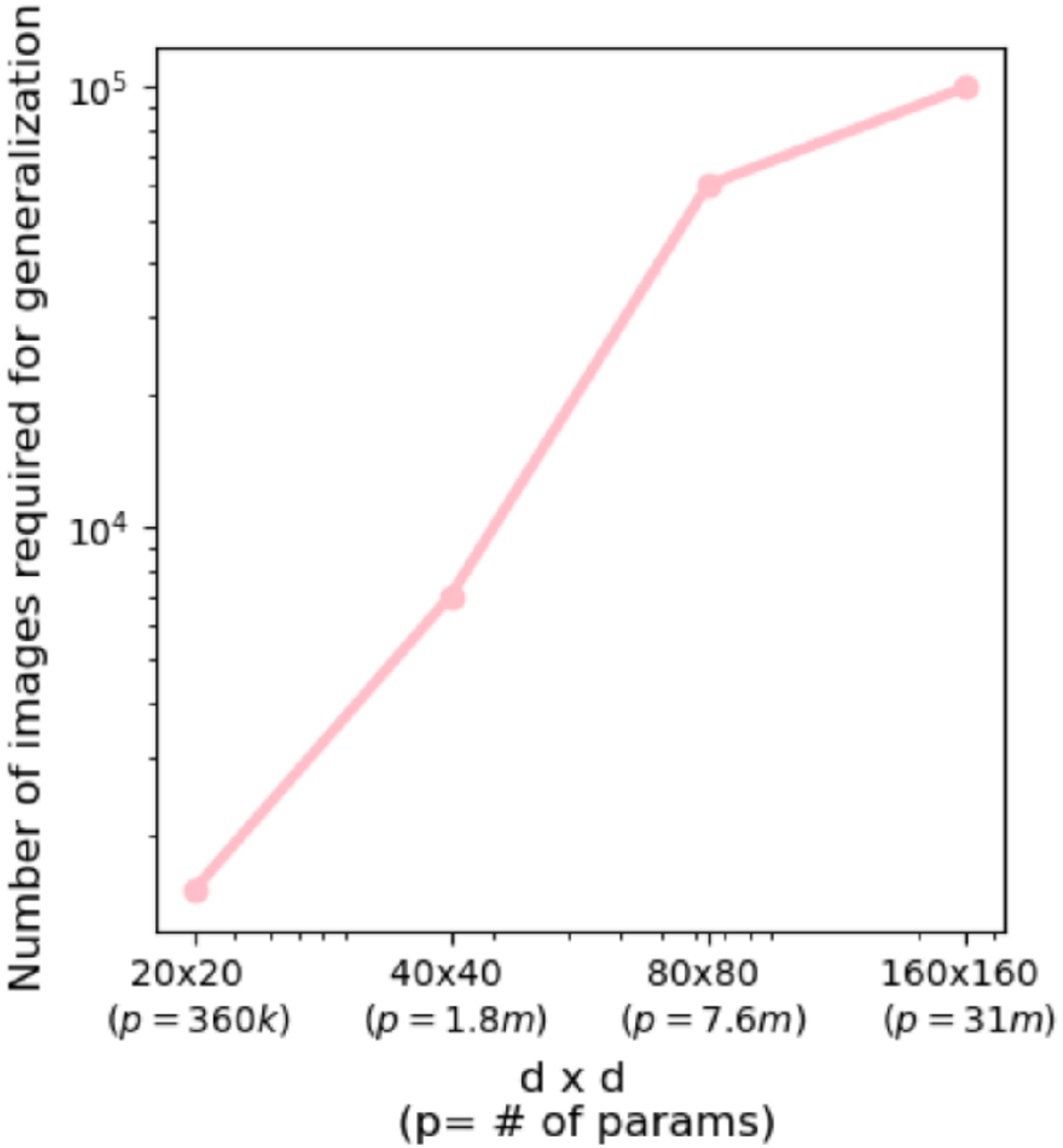
- Diffusion models can transition from **memorization** to **generalization** with large enough training set size
- Generalization is **strong:** two denoisers trained on non-overlapping training sets converge to nearly the same function
- Generalization due to an **inductive bias** corresponding to shrinkage in a Geometry Adaptive Harmonic Basis (GAHB)

Kadkhodaie Z, Guth F, Simoncelli EP, Mallat S.
"*Generalization in diffusion models arises from geometry-adaptive harmonic representation*".ICLR 2024.

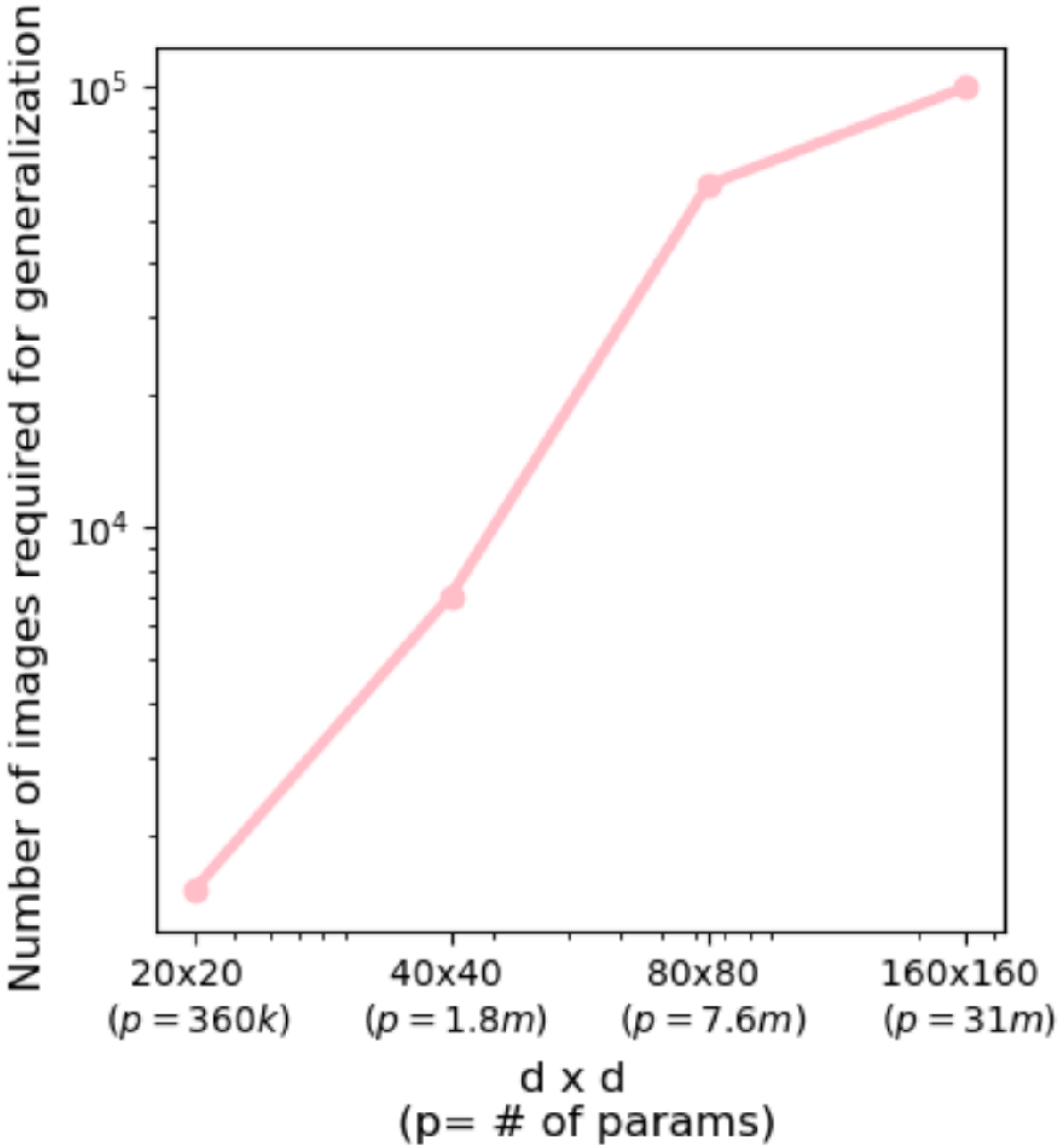
- We are in good shape with learning densities
- Can we reduce the size of training set required for generalization?

- Image resolution
- Network size
- Complexity of image dataset

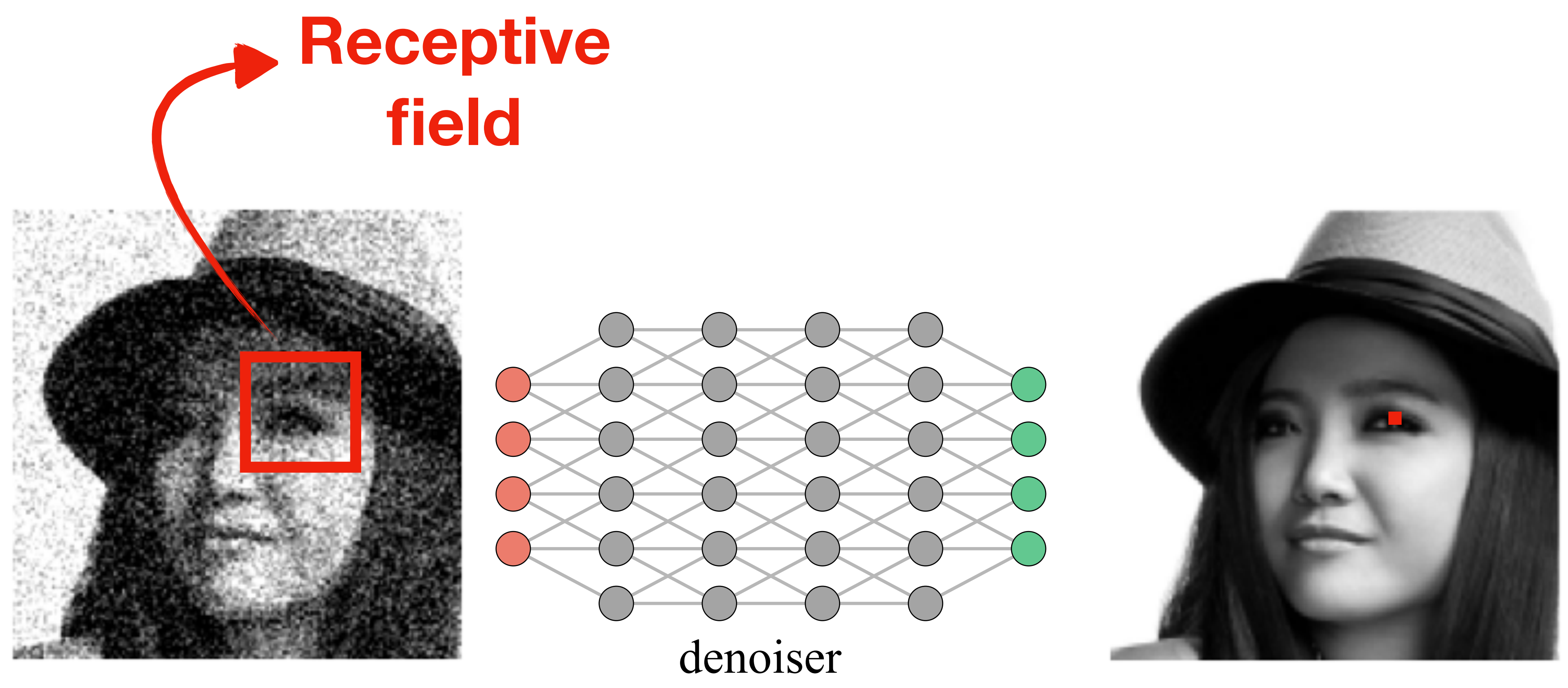
- Image resolution
- Network size
- Complexity of image dataset



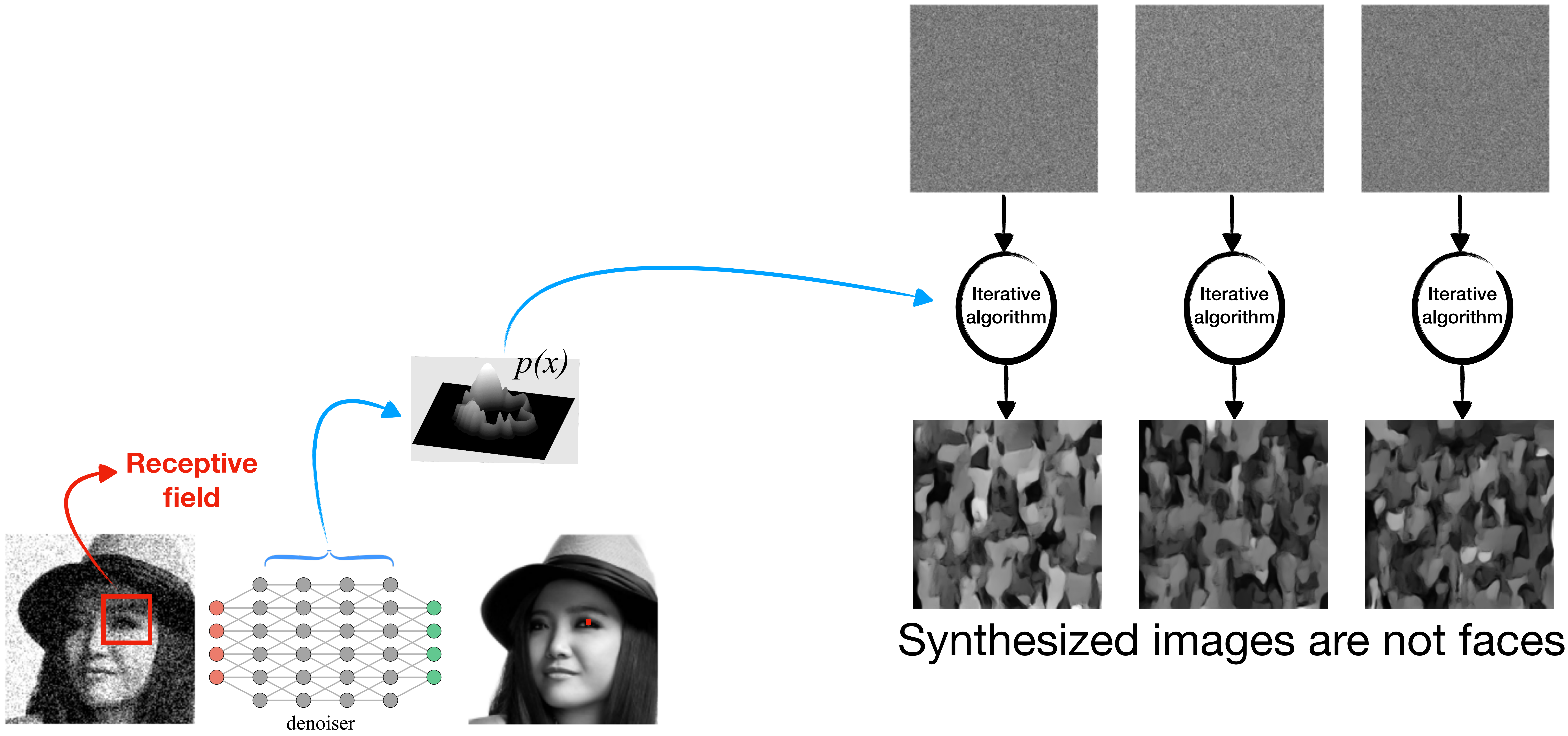
Hundreds of millions or
billions of parameters
with **global receptive fields**



Network receptive field



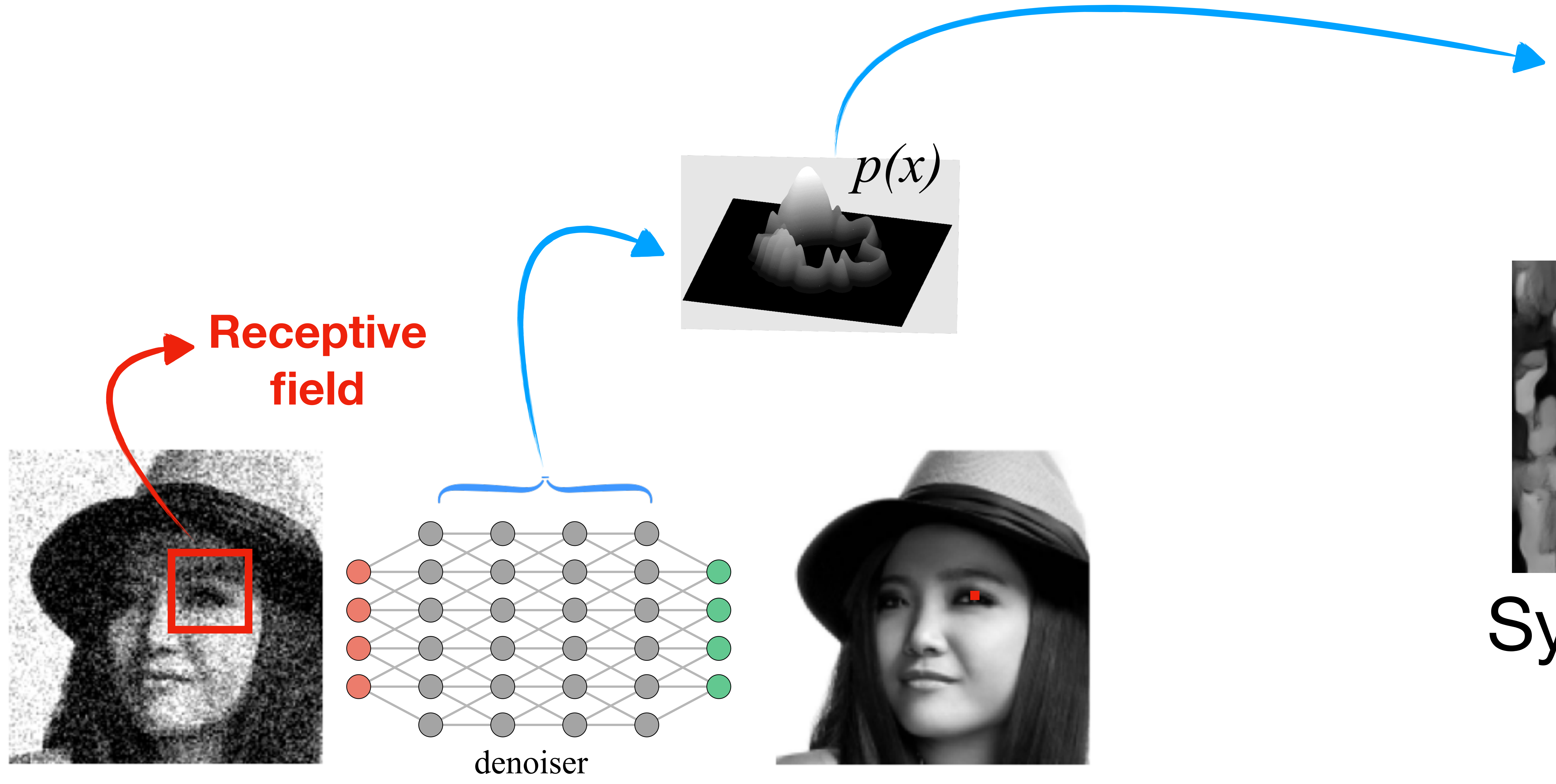
Synthesis fails without global receptive fields!

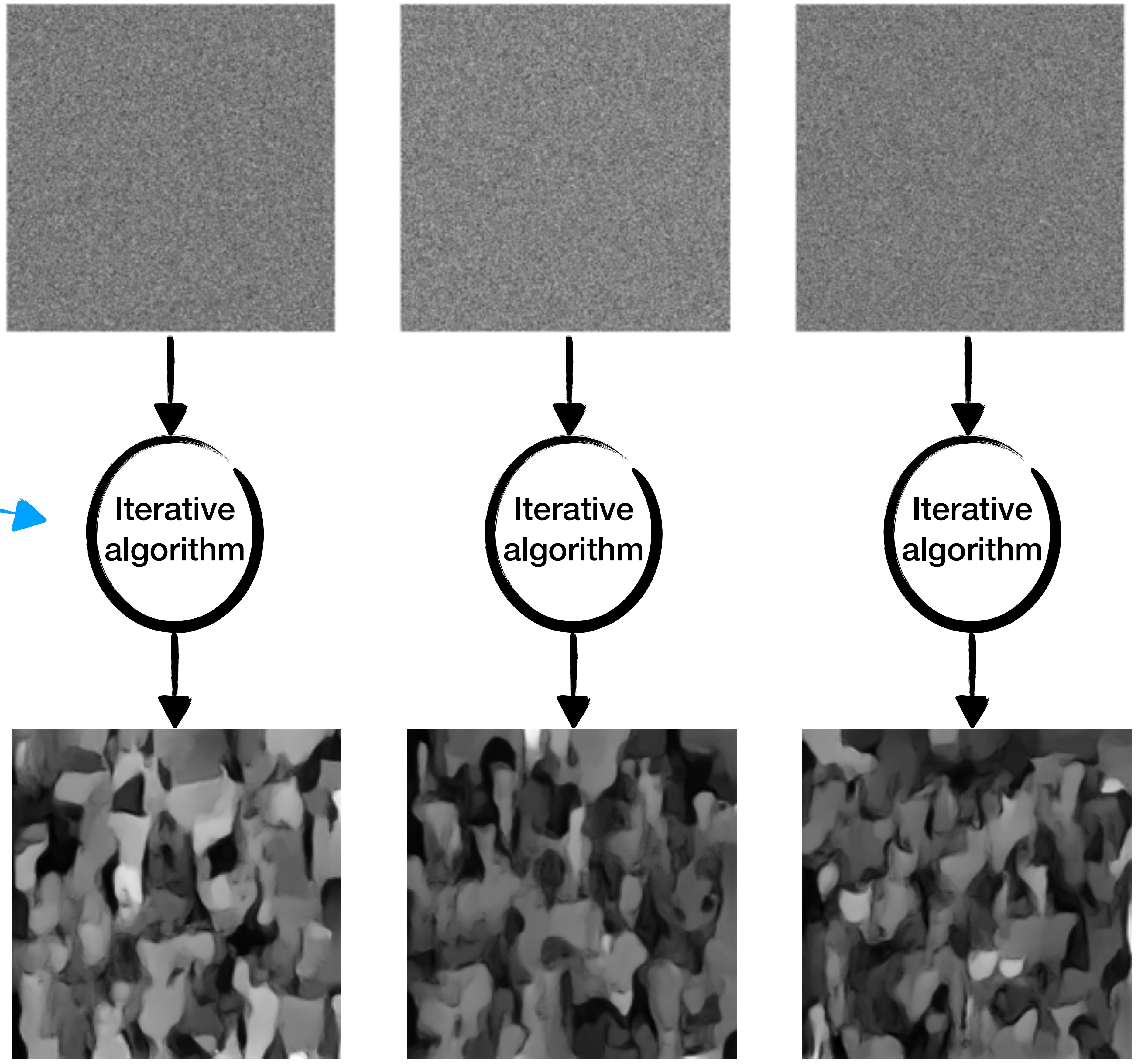


Synthesized images are not faces

Synthesis fails without global receptive fields!

local and translation invariant RF $\Rightarrow$ learn density is an MRF



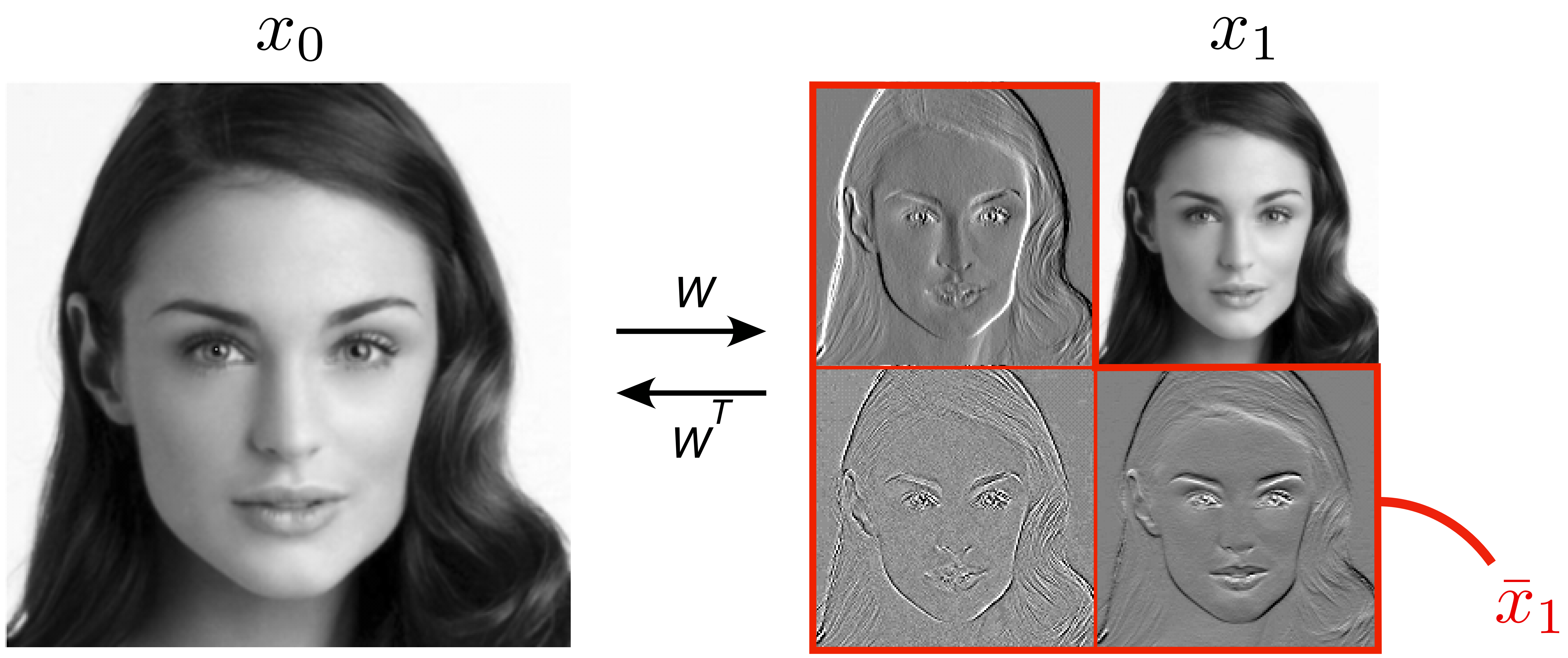


Synthesized images are not faces

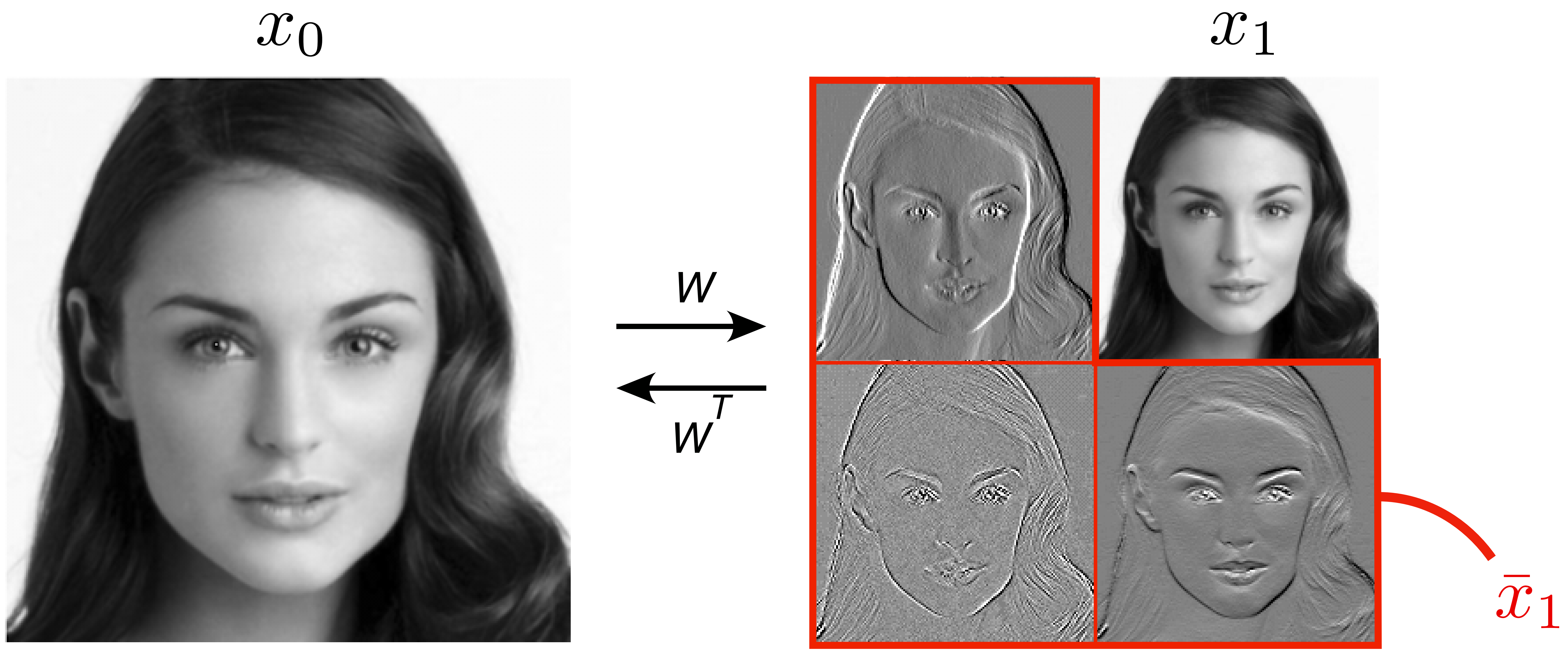
Do bigger images require bigger models?

x_0 x_1

Wavelet decomposition

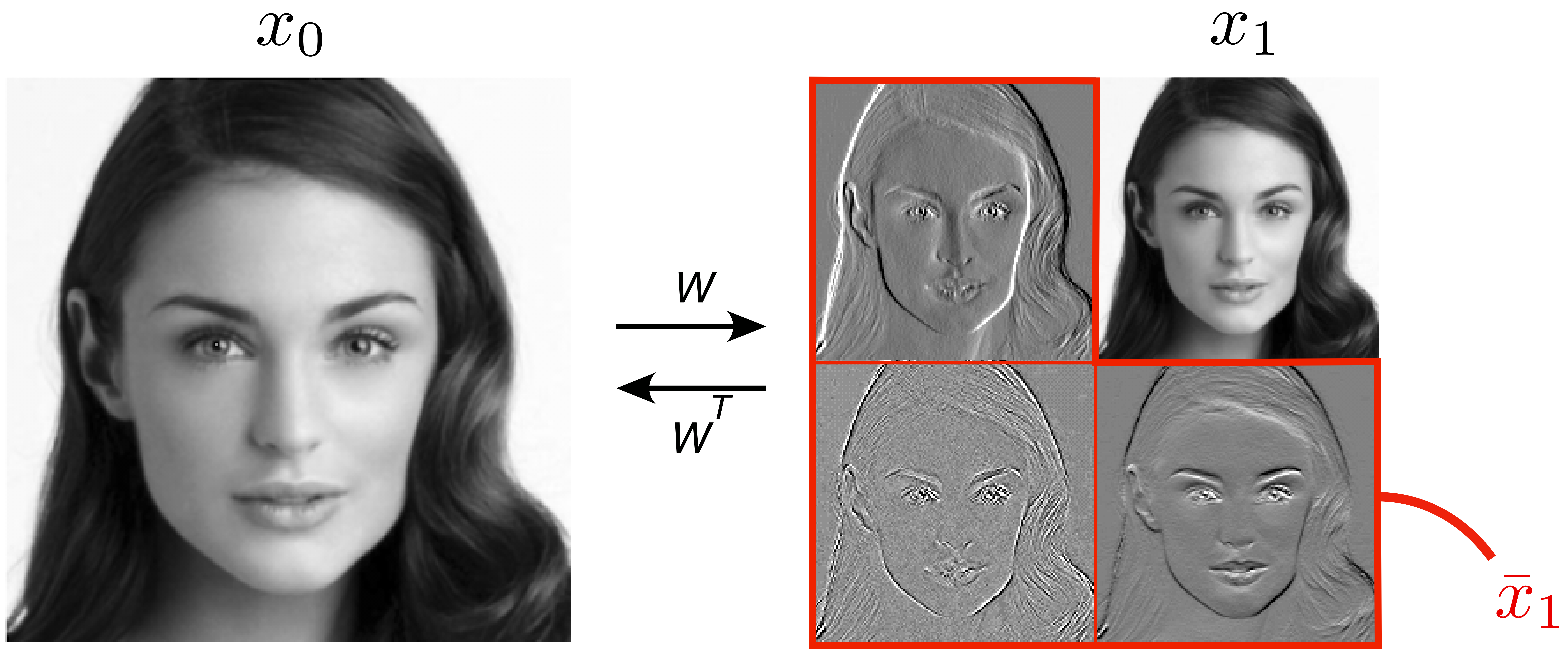


Wavelet decomposition



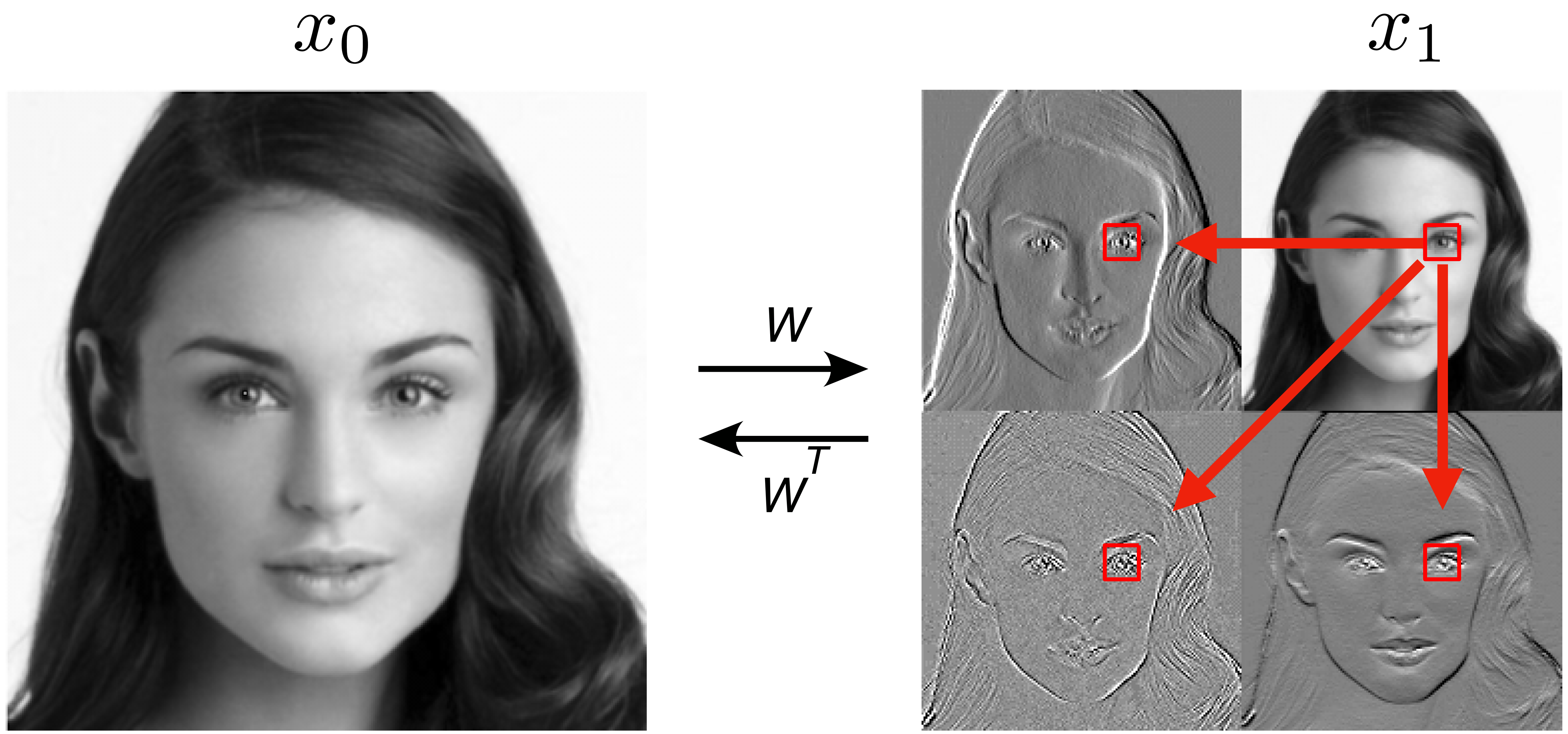
$$p(x_0) = p(x_1, \bar{x}_1)$$

Factorization of the prior



$$p(x_0) = p(x_1)\, p(\bar{x}_1 | x_1)$$

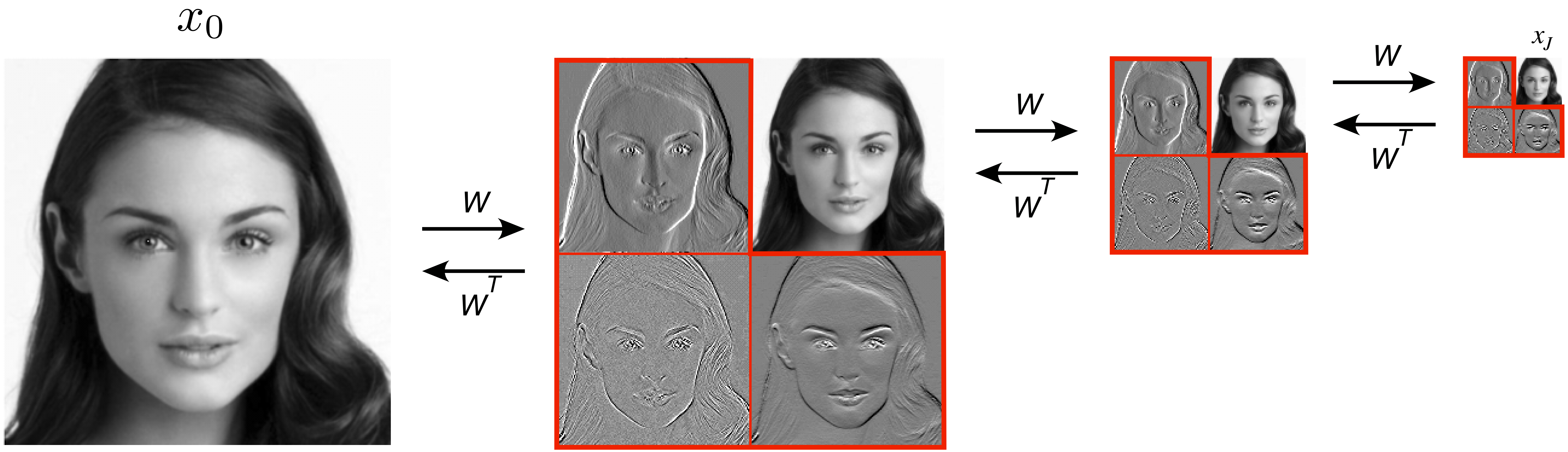
Conditional densities are local



$$p(x_0) = p(x_1)\, p(\bar{x}_1 | x_1)$$

Global **Local**

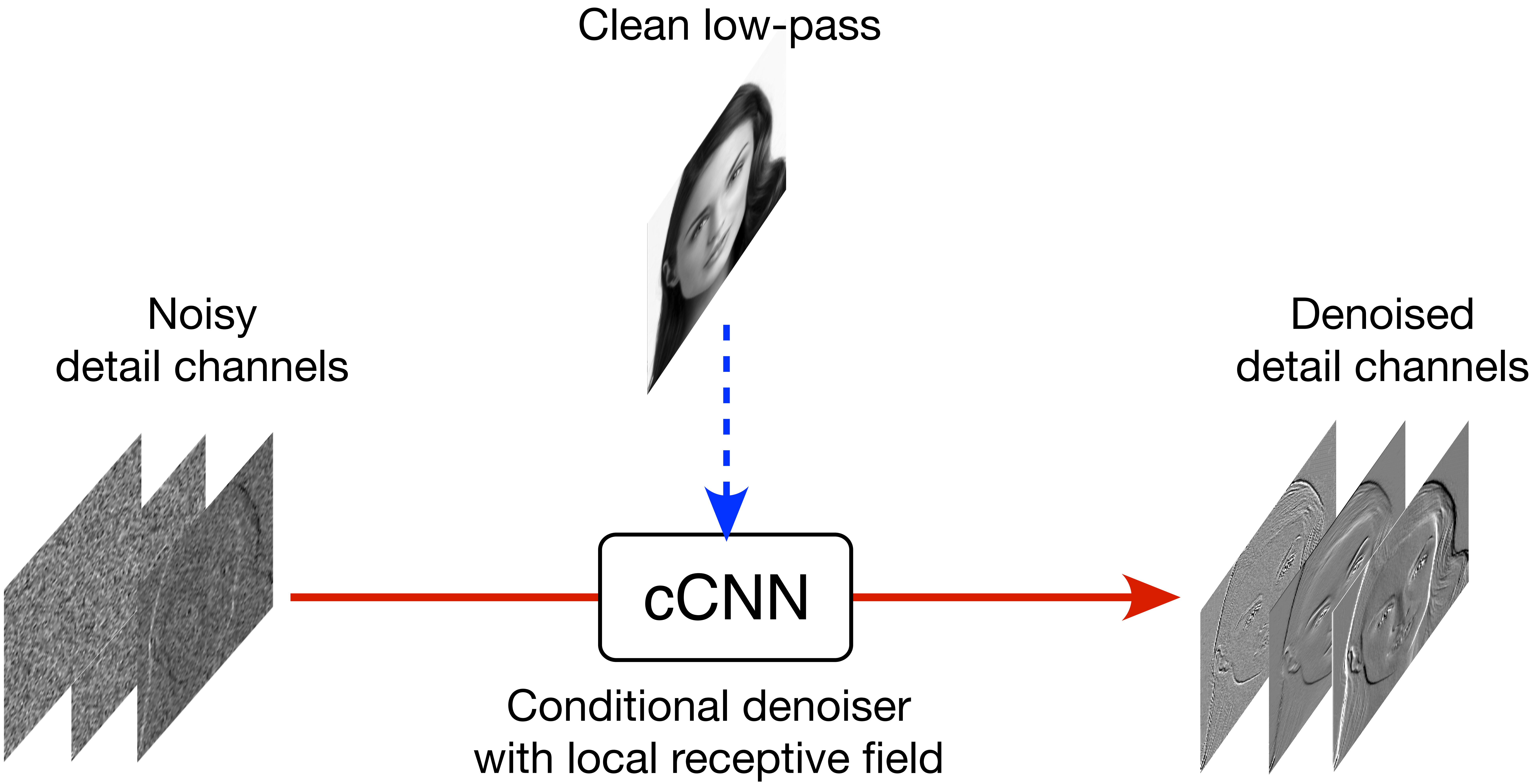
Multi-scale wavelet representation



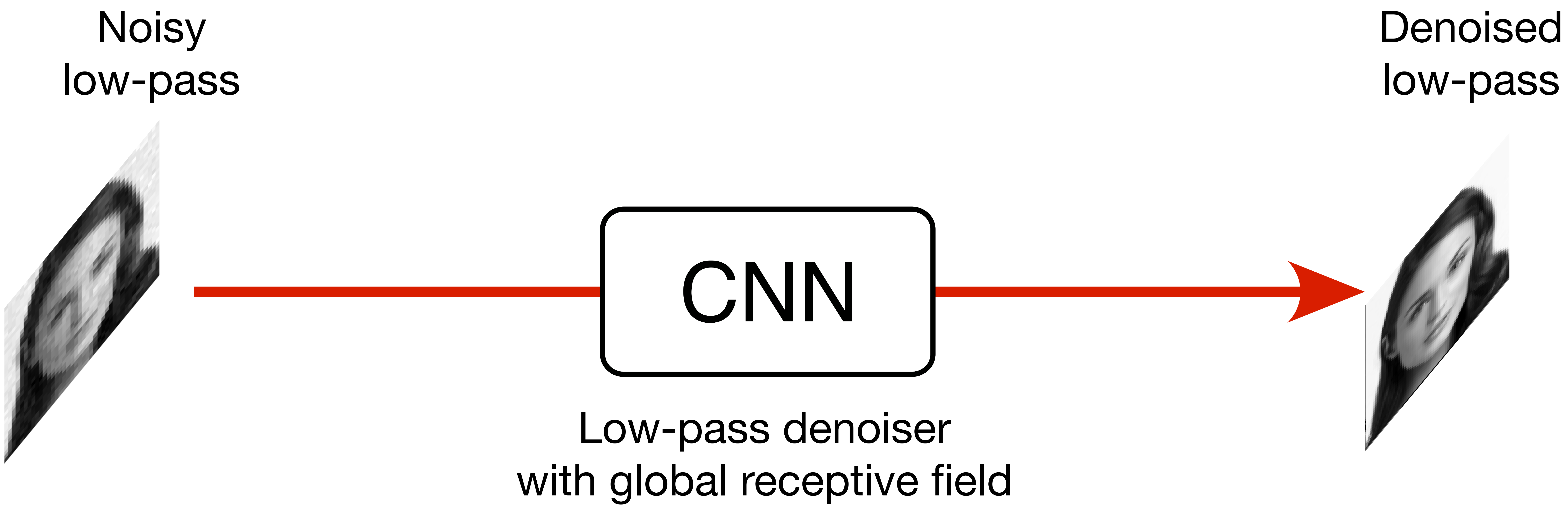
$$p(x_0) = p(x_J) \prod_{j=1}^{J} p(\bar{x}_j | x_j)$$

Global **Local**

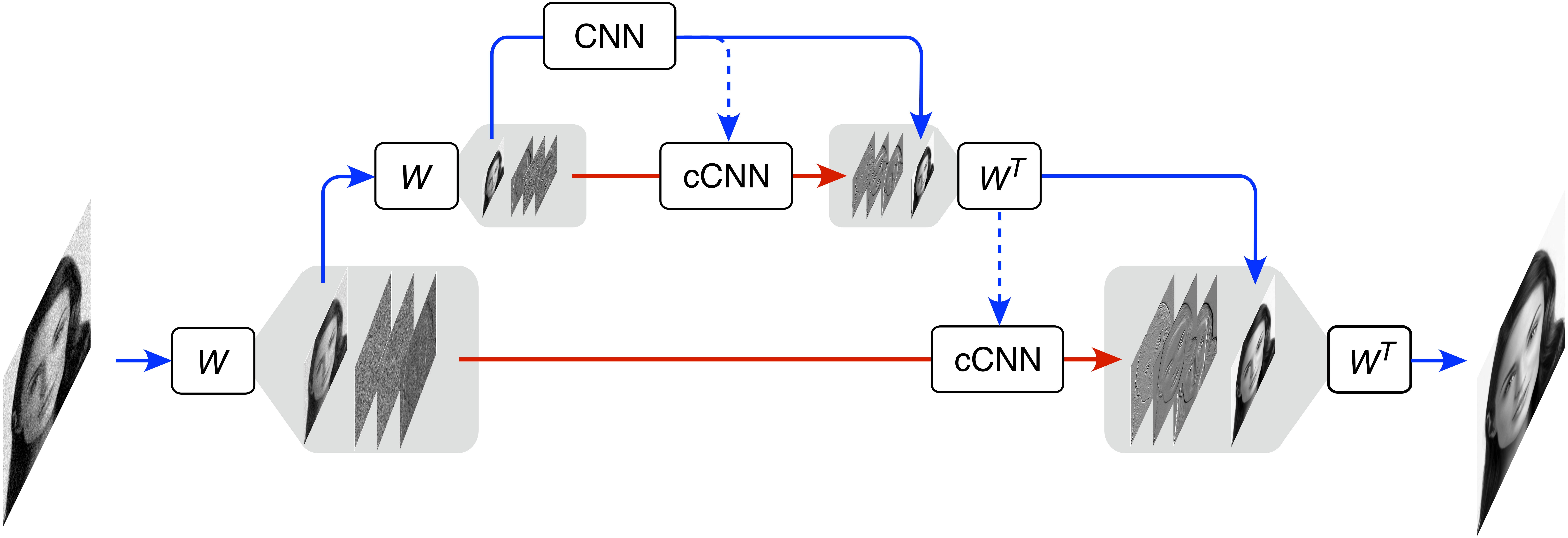
Conditional denoisers



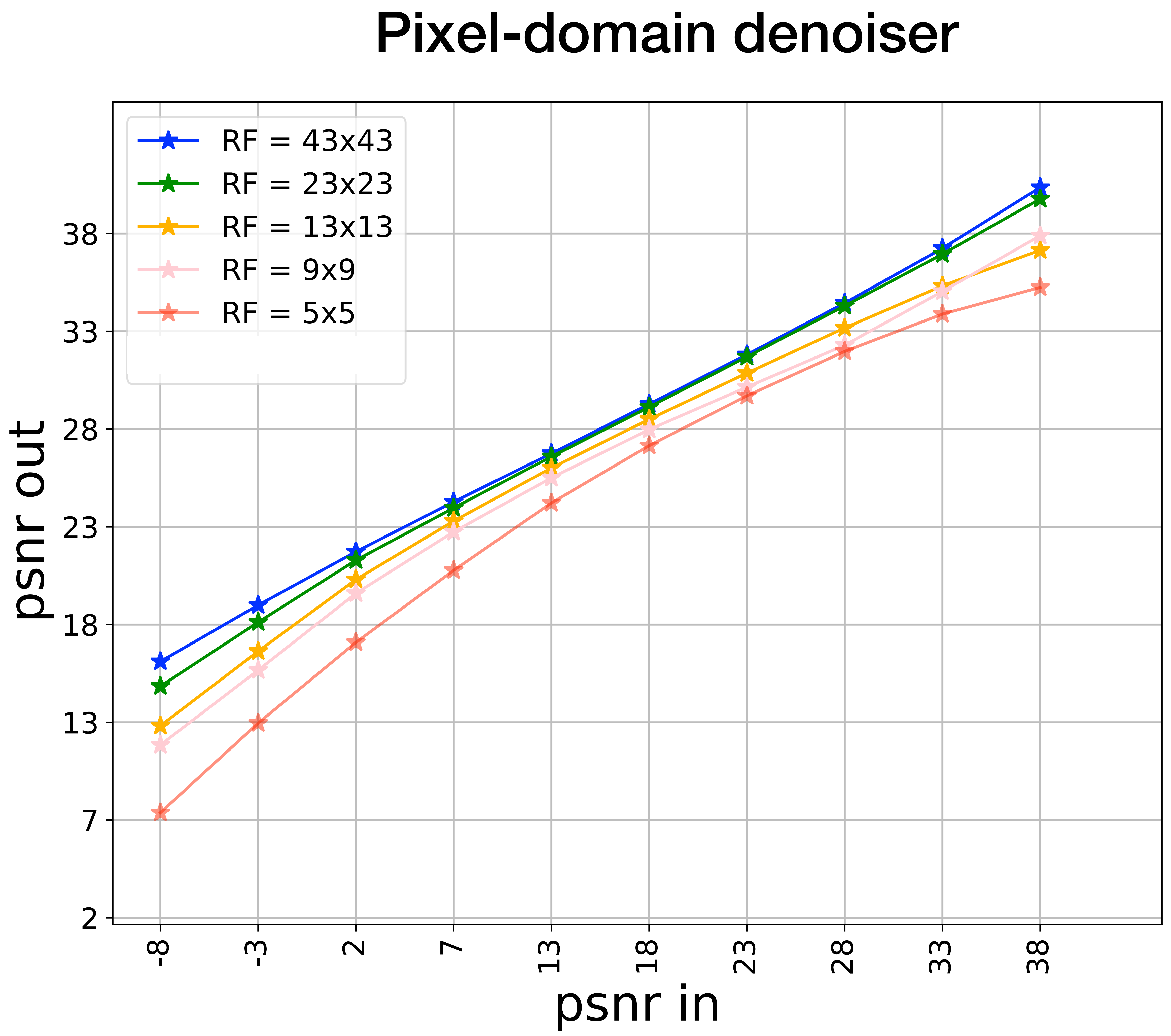
Low pass denoiser



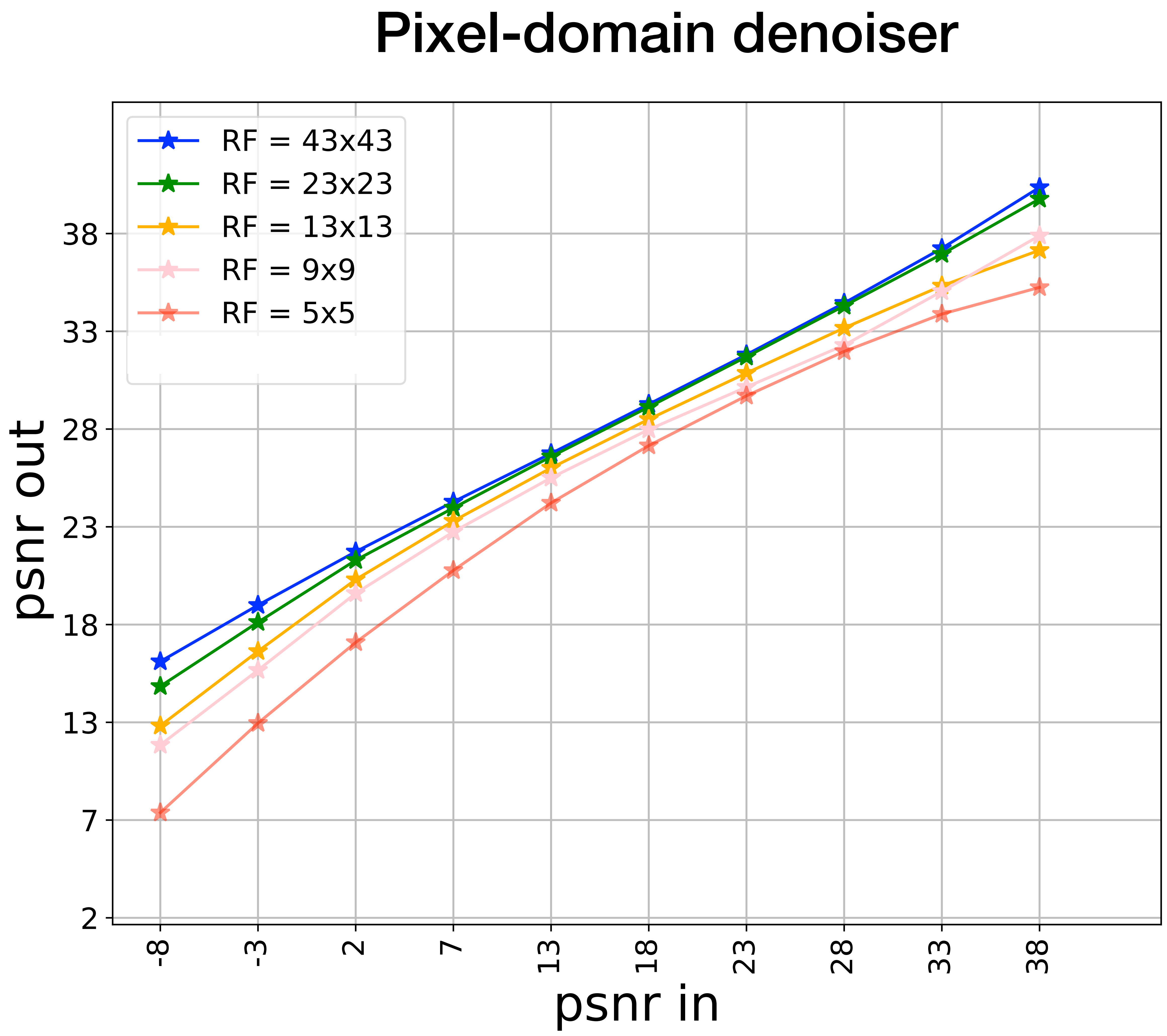
Multi-scale wavelet conditional denoiser



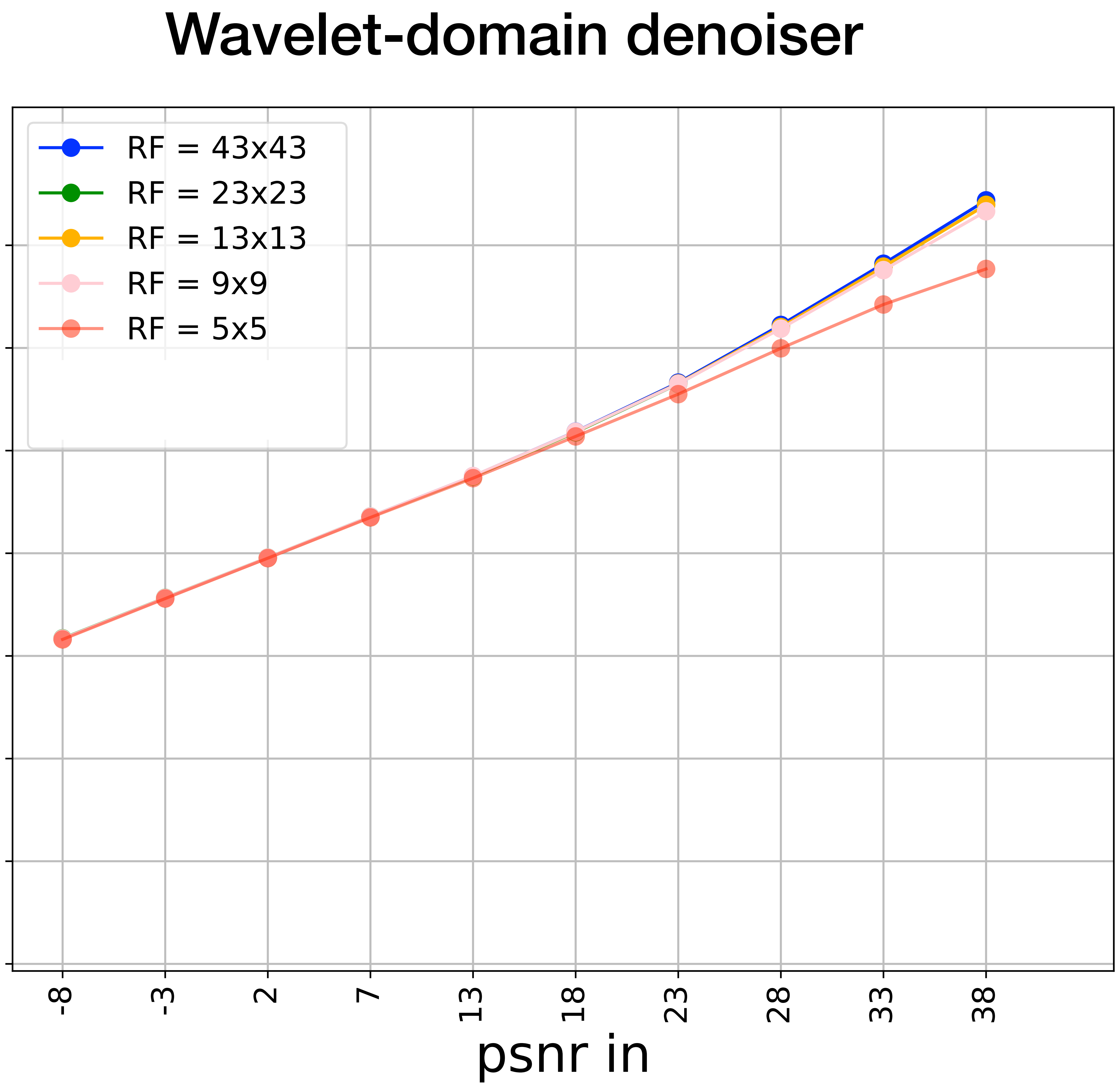
How local?


Pixel-domain denoiser
RF = 43x43
RF = 23x23
RF = 13x13
RF = 9x9
RF = 5x5
psnr out
psnr in
38
33
28
23
18
13
7
2
-8
-3
2
7
13
18
23
28
33
38

How local?

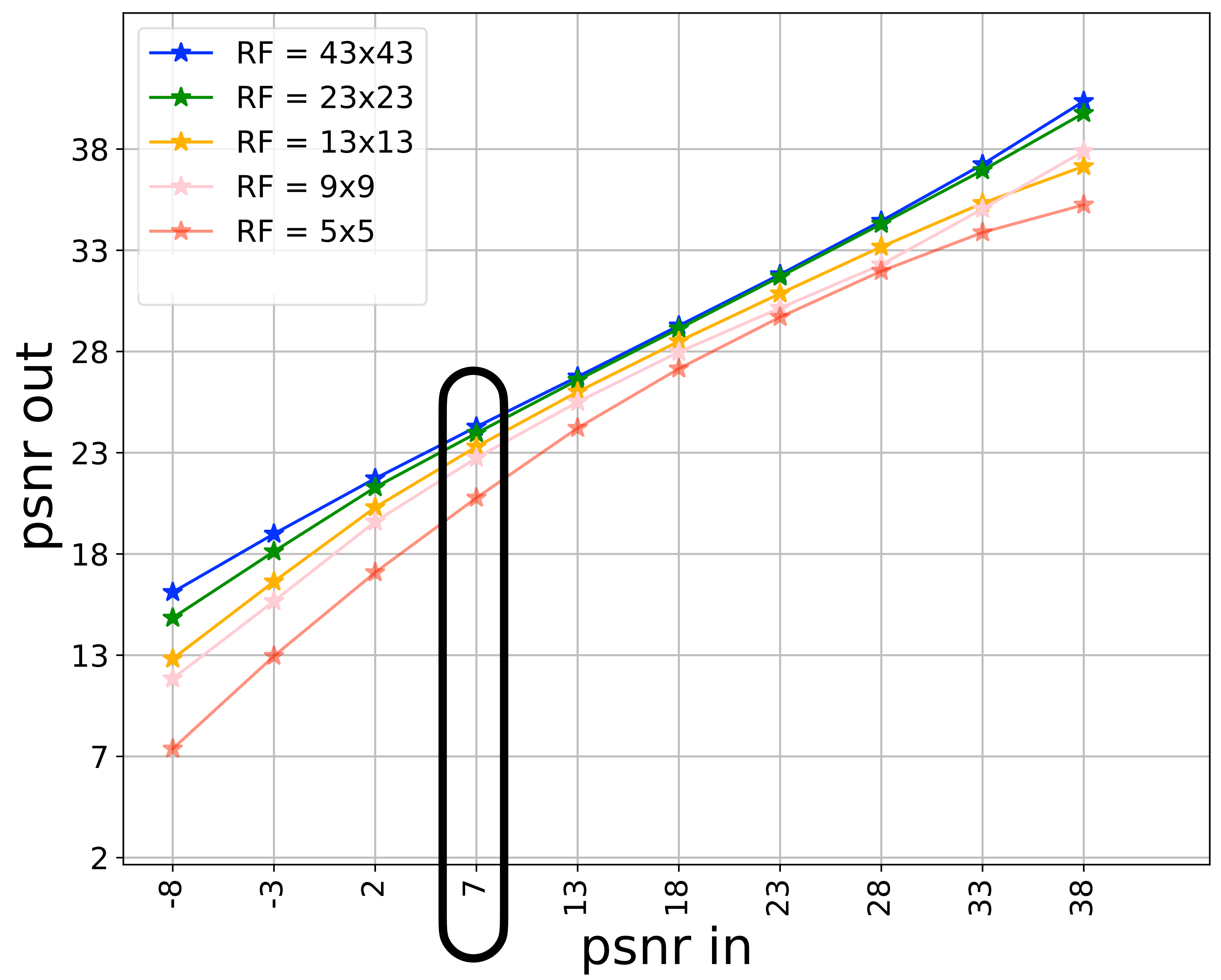


Pixel-domain denoiser
RF = 43x43
RF = 23x23
RF = 13x13
RF = 9x9
RF = 5x5
psnr out
psnr in
2
7
13
18
23
28
33
38
-8
-3
2
7
13
18
23
28
33
38



Wavelet-domain denoiser
RF = 43x43
RF = 23x23
RF = 13x13
RF = 9x9
RF = 5x5
psnr in
-8
-3
2
7
13
18
23
28
33
38

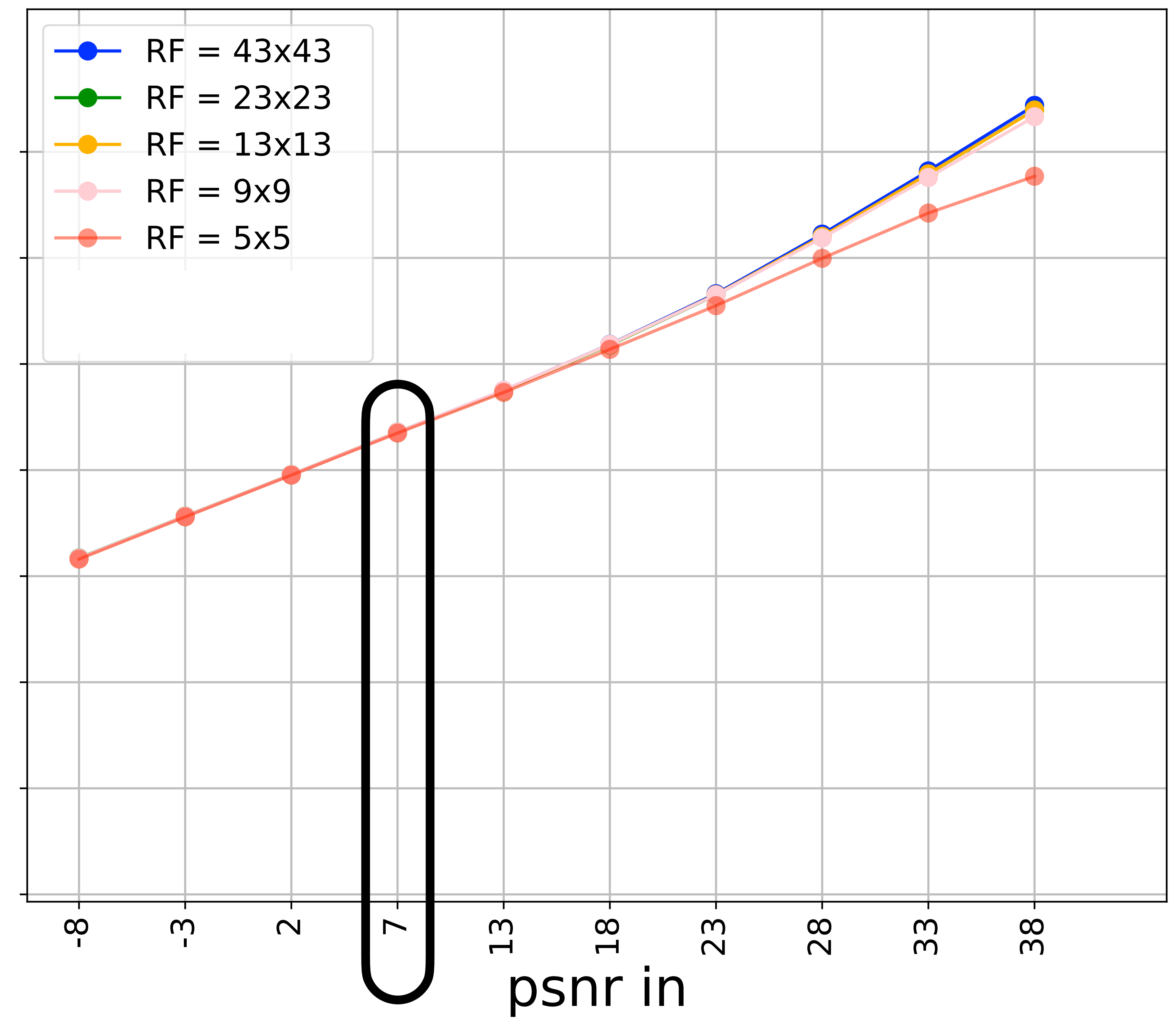
Pixel-domain denoiser



psnr in = 7

Image size: 320x320

Wavelet-domain denoiser



RF = 43x43

RF = 23x23

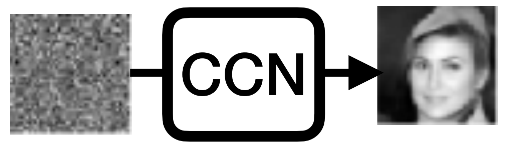
RF = 9x9

RF = 43x43

RF = 23x23

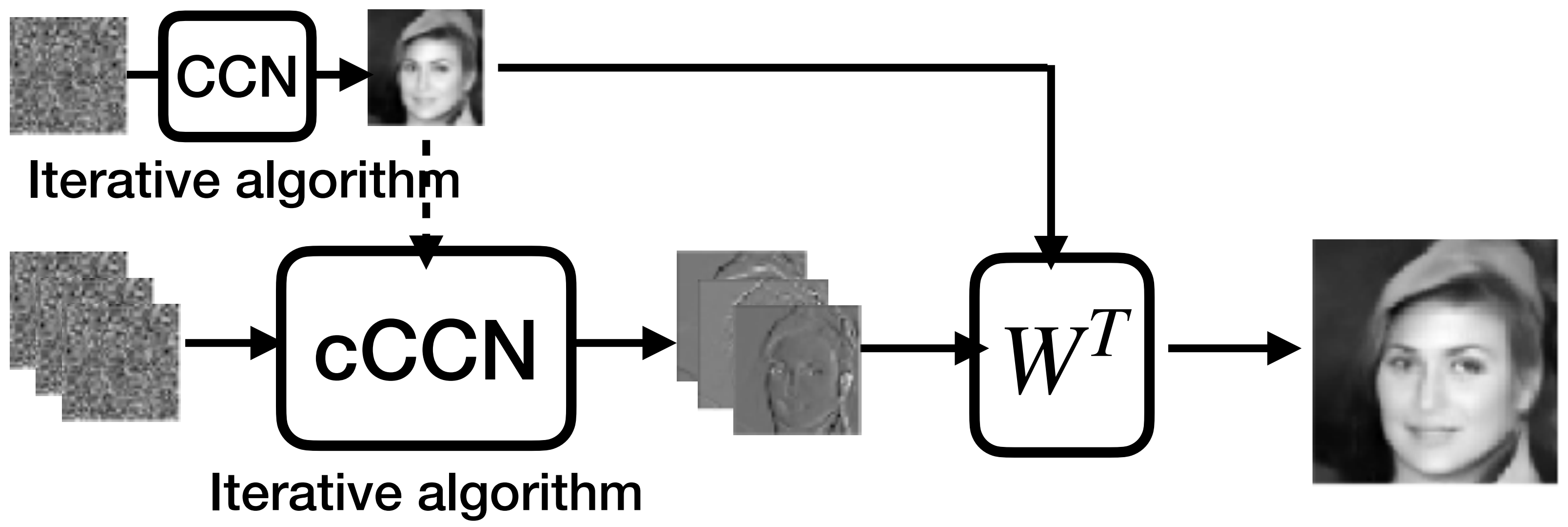
RF = 9x9

Synthesis example

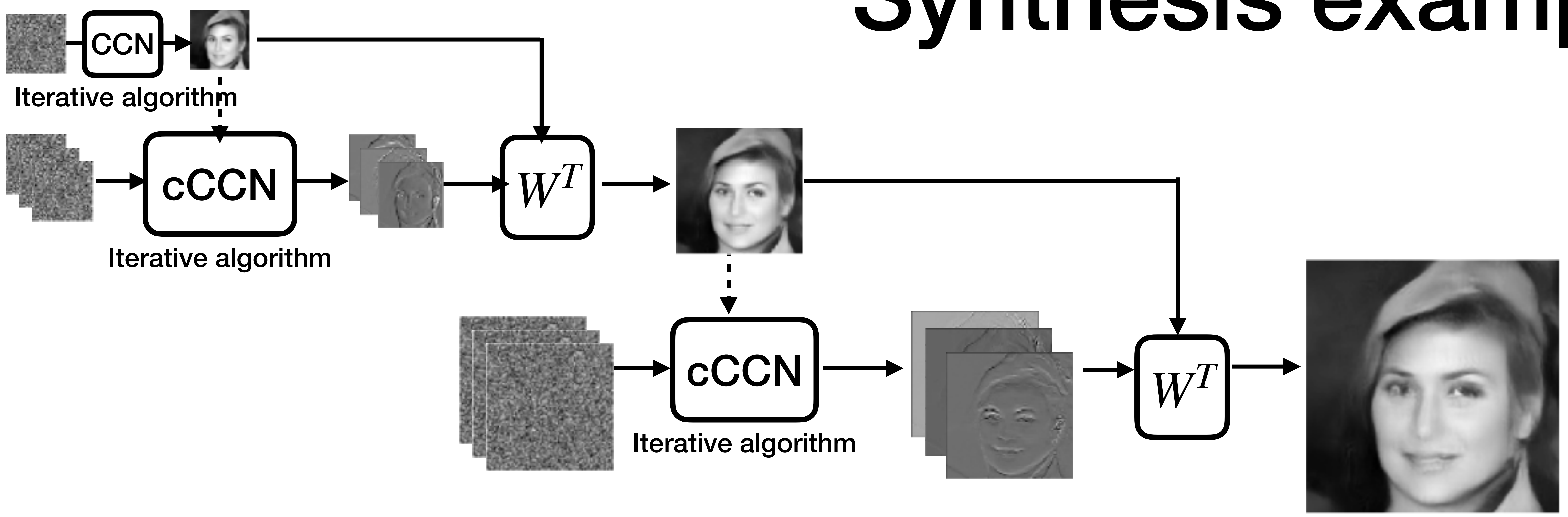


Iterative algorithm

Synthesis example



Synthesis example



Synthesis example

Total number of parameters ~ 1 million

Synthesis example

To sum up:

- We can model probability of large images with small networks.
- The global structure is captured by a global prior over a small low-pass image.
- Details can be modeled using local (Markov) conditional probability distributions in the wavelet domain.

Thank you!

Kadkhodaie, Guth, Simoncelli, Mallat, “*Generalization in diffusion models arises from geometry-adaptive harmonic representation”.* ICLR 2024

Kadkhodaie, Guth, Mallat, & Simoncelli, “*Learning multi-scale local conditional probability models of images.*” ICLR 2023