Hierarchical Dirichlet Process Hidden Markov Trees for Multiscale Image Analysis

Jyri Kivinen

International Computer Science Institute, Berkeley, USA

Helsinki University of Technology, Espoo, Finland

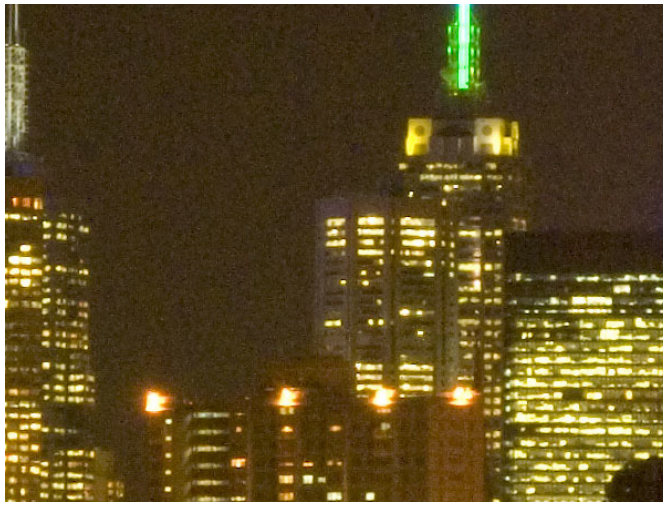
The University of Edinburgh, UK

Joint work with

Erik Sudderth, Brown University

Michael Jordan, UC Berkeley

Low-level Image Analysis



Noise Removal

Deblurring

Inpainting & Restoration

Goals:

- Accurately model the statistics of *natural images*
- Exploit the availability of large digital *image collections*
  - Suggests use of data-driven, *nonparametric* models

Natural Scene Categorization

Coast **Forest** **Open Country** **Street** **Tall Building**

Goals:

- Visually *recognize* natural scene categories
- Accurately model the statistics of *natural scenes*
- Learn *global* statistical scene models

Outline

Multiscale Models for Natural Images

- Nonparametric Hidden Markov Trees (HDP-HMTs)
- Learning with Monte Carlo methods
- Truncated representations for efficient learning from large datasets

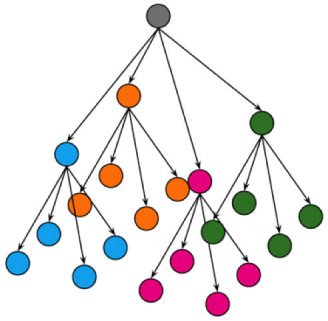
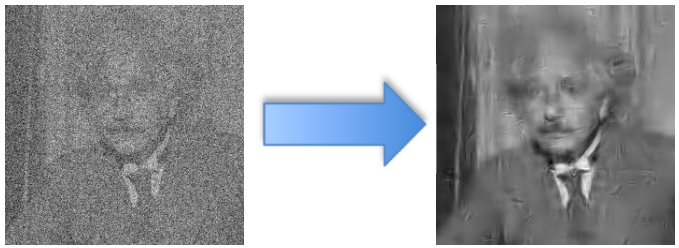


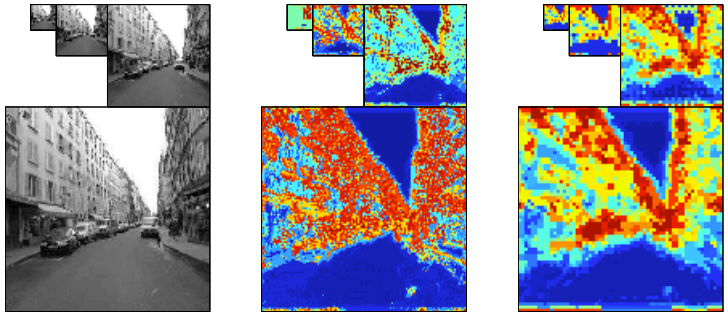
Image Denoising

- Transfer natural image statistics for making robust predictions

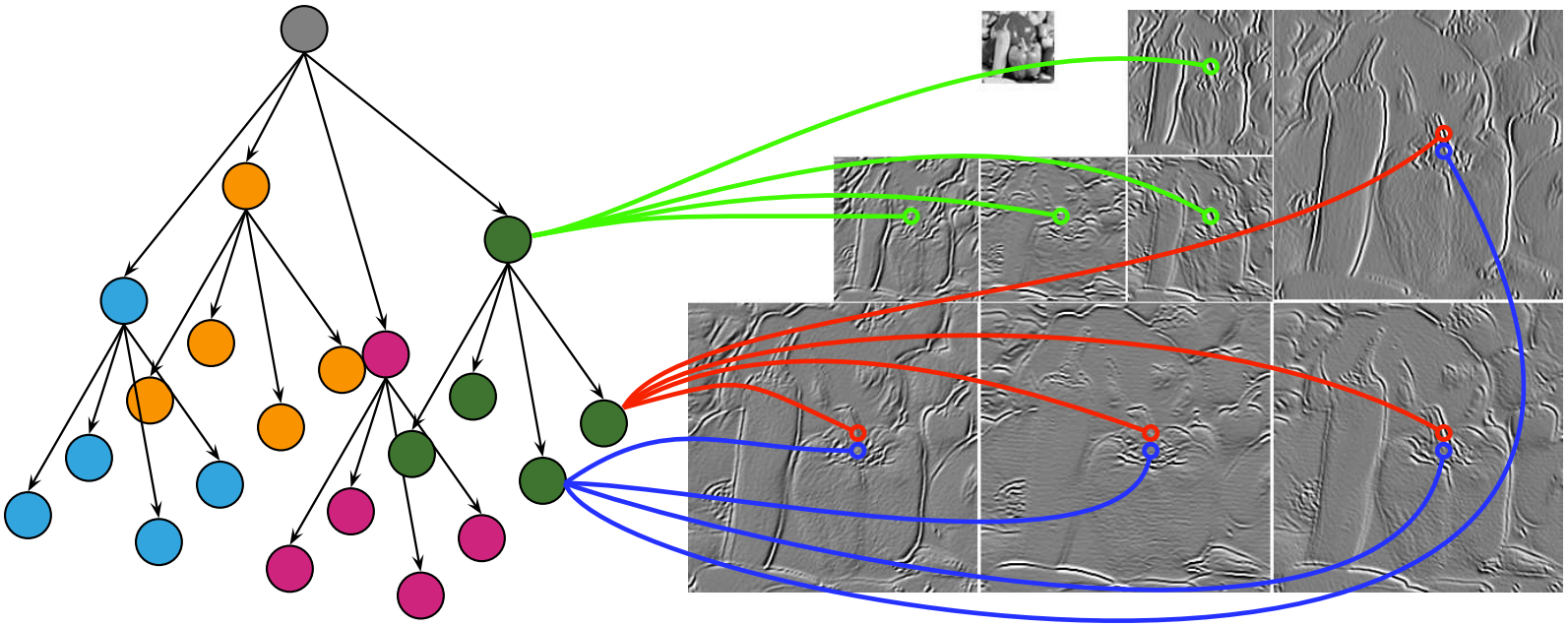


Natural Scene Analysis

- Global, data-driven scene models via HDP-HMT
- Fast categorization via Belief Propagation methods



Multiscale Models for Natural Images

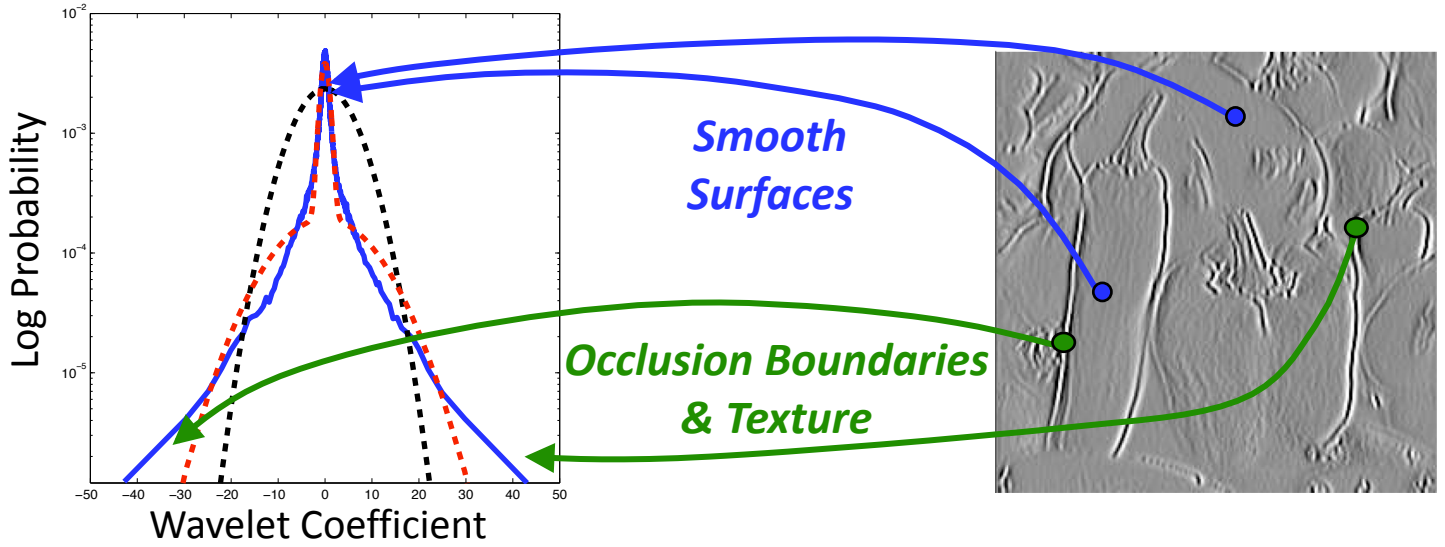


Goal: Accurately model the statistics of natural images

Approach:

- Capture multiscale dependencies using a *tree* of latent variables
- Automatically adapt to data complexity via nonparametric, *Dirichlet process* priors

Mixture Models for Heavy-Tailed Wavelet Marginals



- Extreme coefficient values resultant from edges and texture occur more frequently than with a Gaussian
- Gaussian scale mixtures provide good matches for the highly kurtotic, heavy tailed distributions

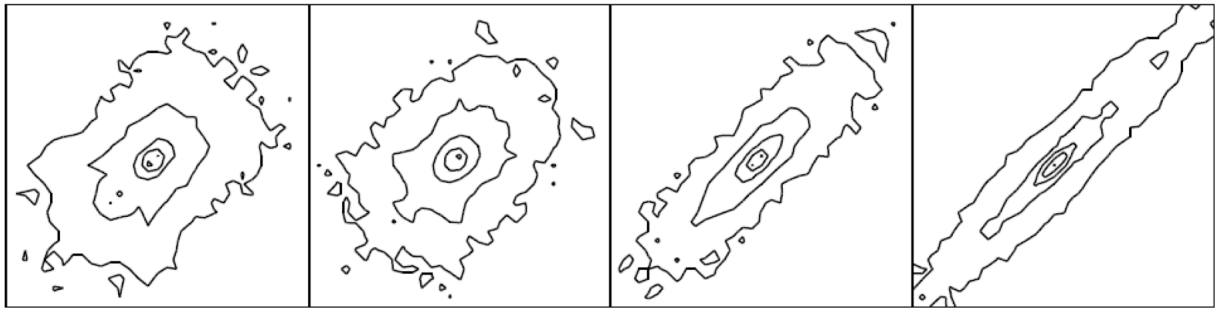
$$x_{ti} = v_{ti} u_{ti} \;;\; v_{ti} \geq 0 \;,\; u_{ti} \sim \mathcal{N}(0, \Lambda)$$

- Discrete mixtures easier to work with, reasonable denoising results even with binary mixtures:

$$x_{ti} \sim \pi \mathcal{N}(0, \Lambda_0) + (1 - \pi)\mathcal{N}(0, \Lambda_1)$$

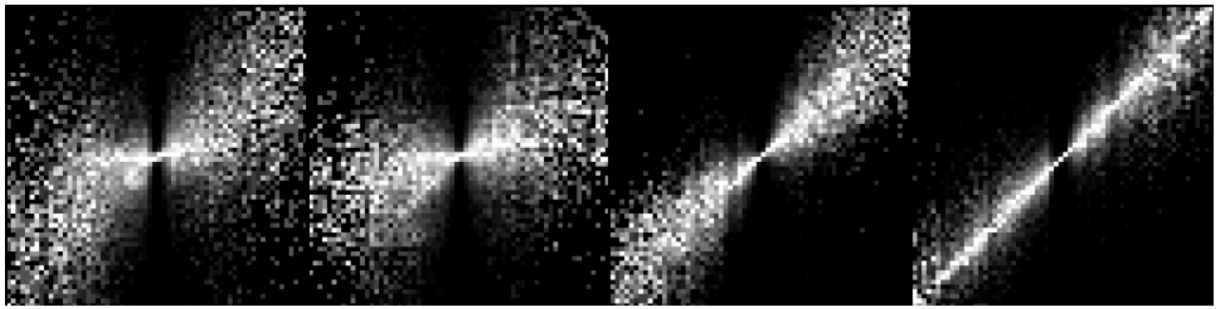
Joint Statistics of Wavelet coefficients

Pairwise Joint Histograms:

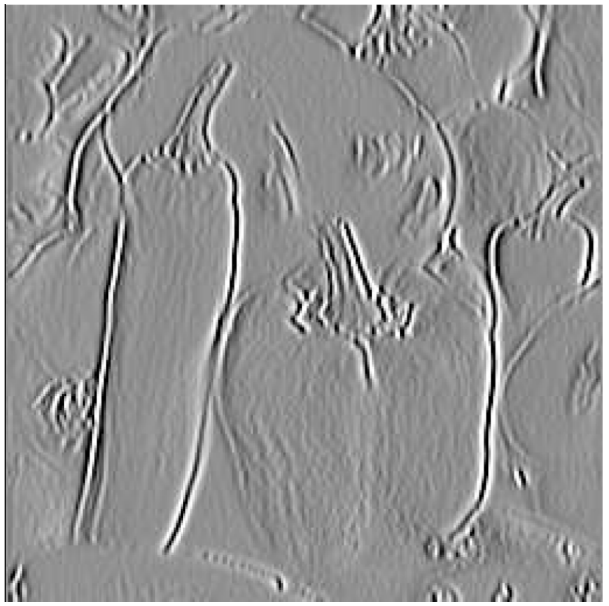


Orientation *Scale* *Vertical* *Horizontal*

Pairwise Conditional Histograms:



Orientation *Scale* *Vertical* *Horizontal*

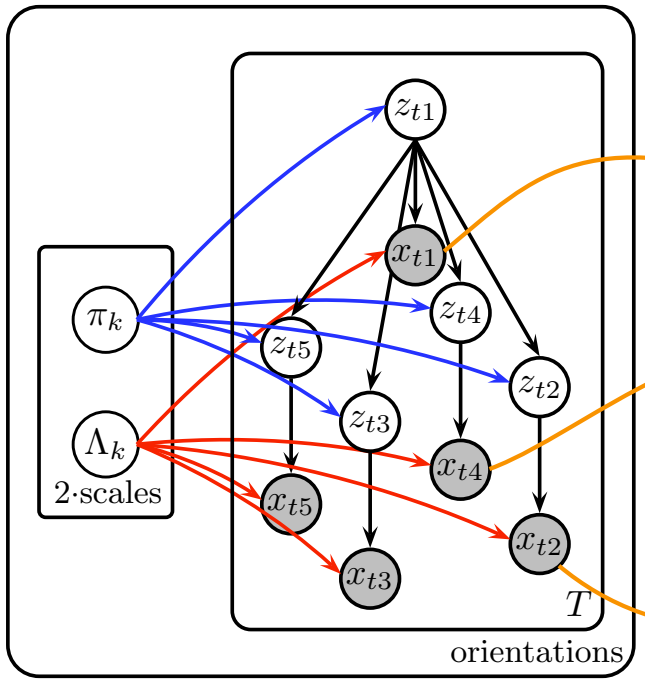


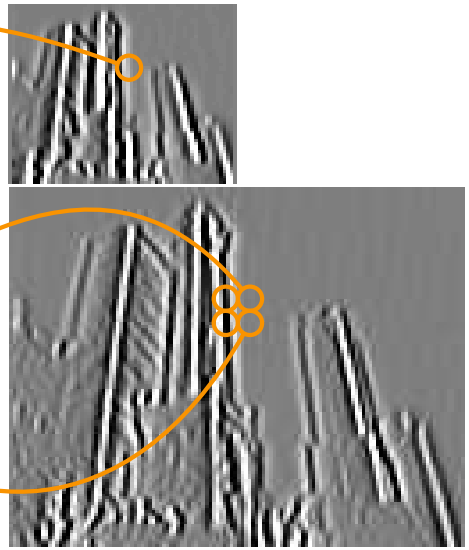
Large magnitude wavelet coefficients…

- *Persist* across multiple scales
- *Cluster* at adjacent spatial locations

Binary Hidden Markov Trees

Crouse, Nowak, & Baraniuk, 1998





$\pi_k \rightarrow$ state *transition* distributions

$z_{ti} \sim \pi_{z_{\mathrm{Pa}(ti)}}$

$\Lambda_k \rightarrow$ state-specific *emission* covariances

$x_{ti} \sim \mathcal{N}(0, \Lambda_{z_{ti}})$

$z_{ti} \rightarrow$ hidden *state* or cluster assignment

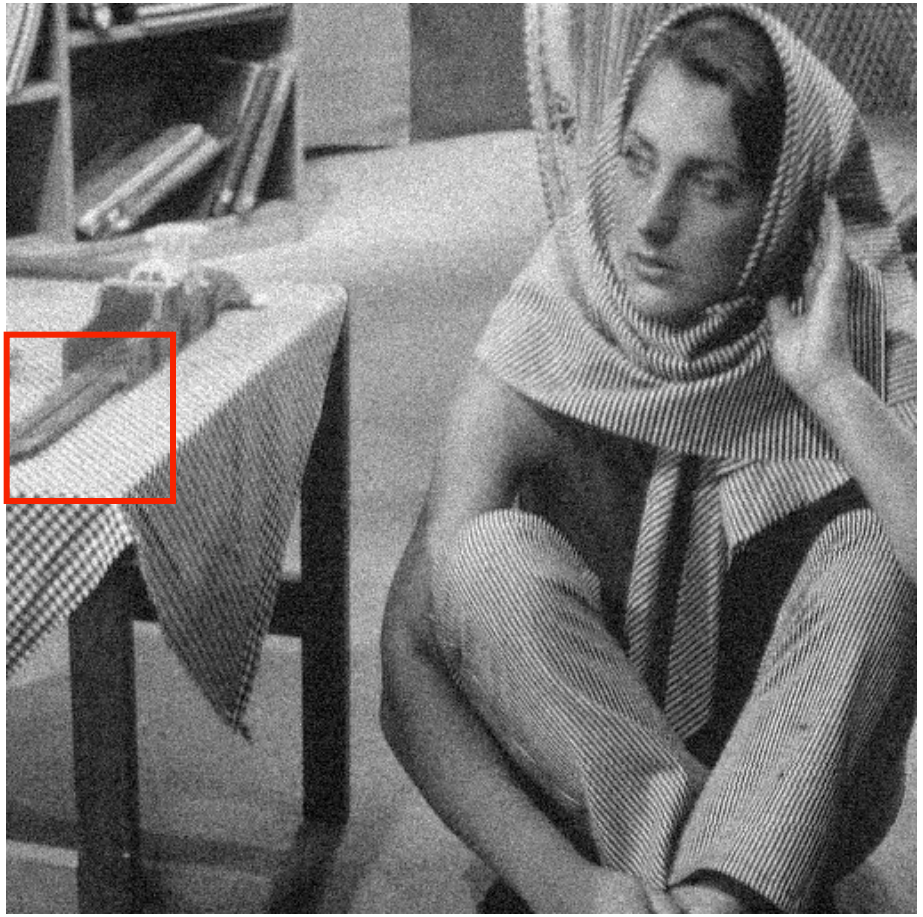
$z_{ti} \in \{0, 1\}$

$x_{ti} \rightarrow$ *observed* wavelet coefficient

- Wavelet coefficients marginally distributed as mixtures of two Gaussians
- Markov dependencies between hidden states capture persistence of image contours across locations and scales
- Models each scale and orientation independently

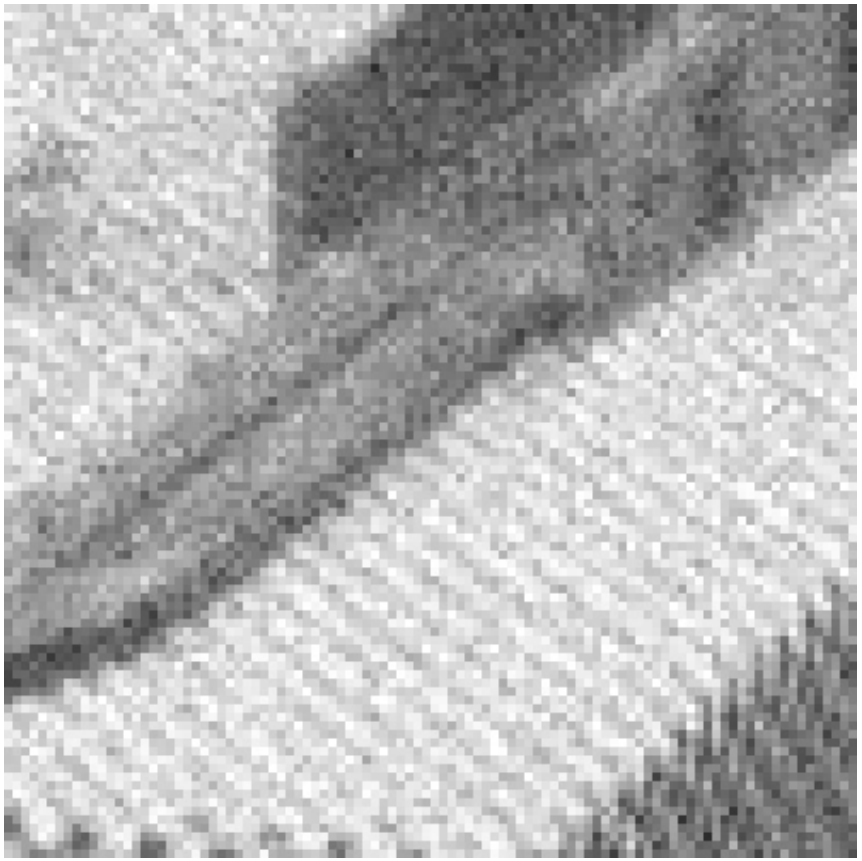
Validation : Image Denoising

Original

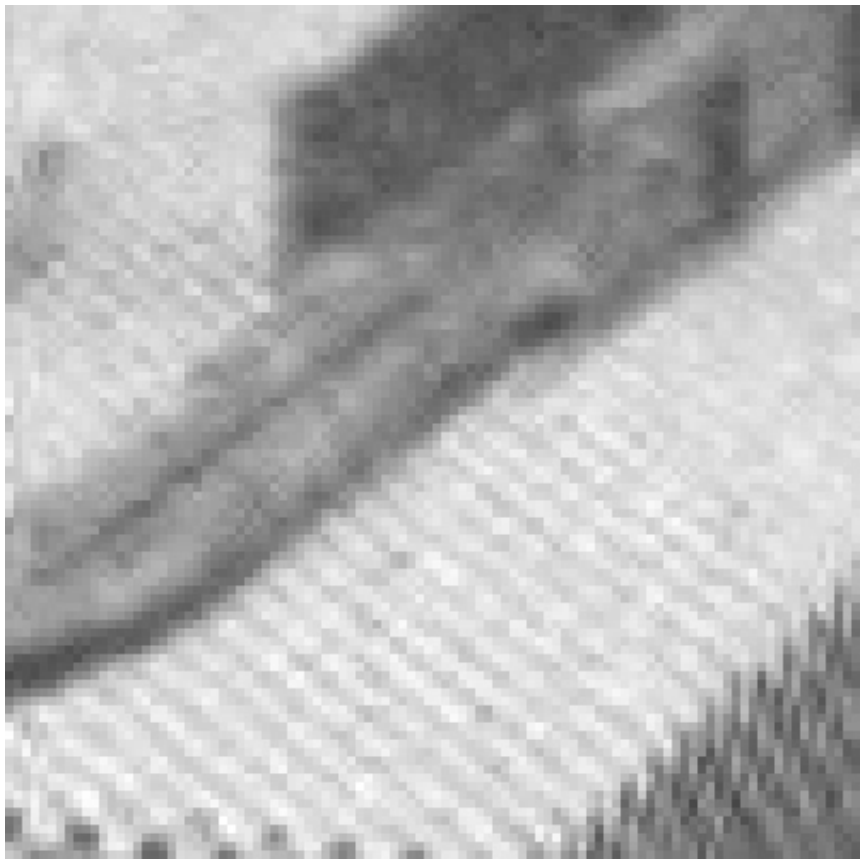


Corrupted by Additive White Gaussian Noise (PSNR = 24.61 dB)

Denoising with Binary HMTs



Noisy Input



Denoised (EM algorithm)

- Is two states per scale sufficient? How many is enough?
- Should states be shared the same way for all images, or for all wavelet decompositions?

Dirichlet Process Mixtures

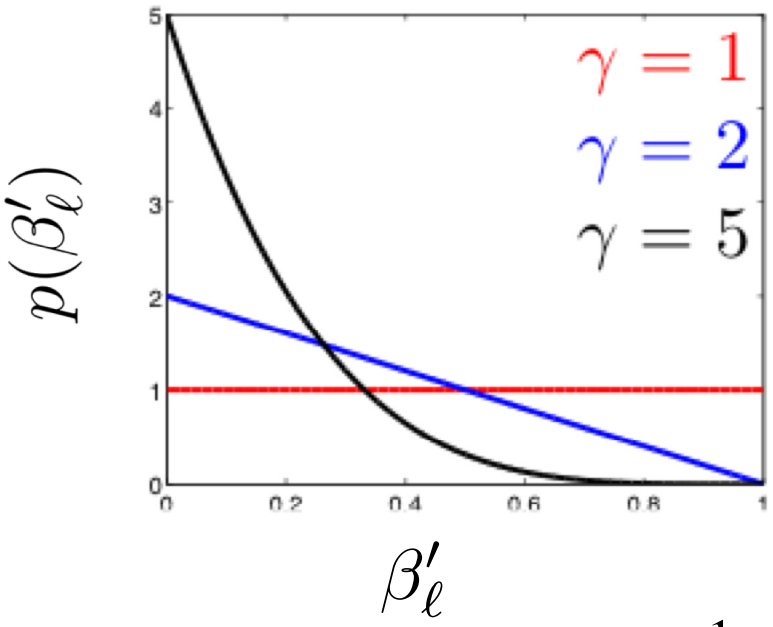
$$p(x_{ti} \mid \beta, \Lambda_1, \Lambda_2, \ldots) = \sum_{k=1}^{\infty} \beta_k \mathcal{N}(x_{ti}; 0, \Lambda_k)$$

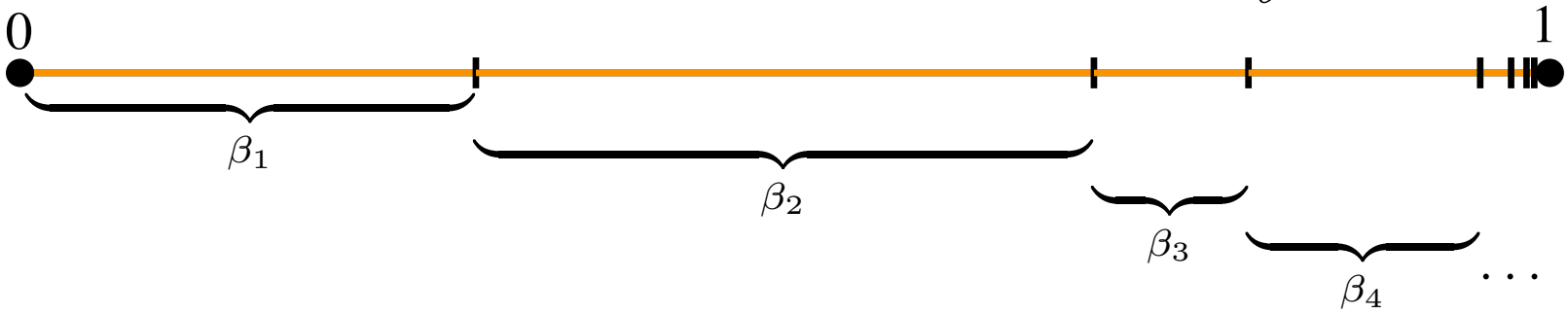
Stick-breaking prior for mixture weights controls complexity:

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell)$$

$$\beta'_\ell \sim \text{Beta}(1, \gamma)$$

$\gamma \longrightarrow$ Concentration parameter





Why the Dirichlet Process ?

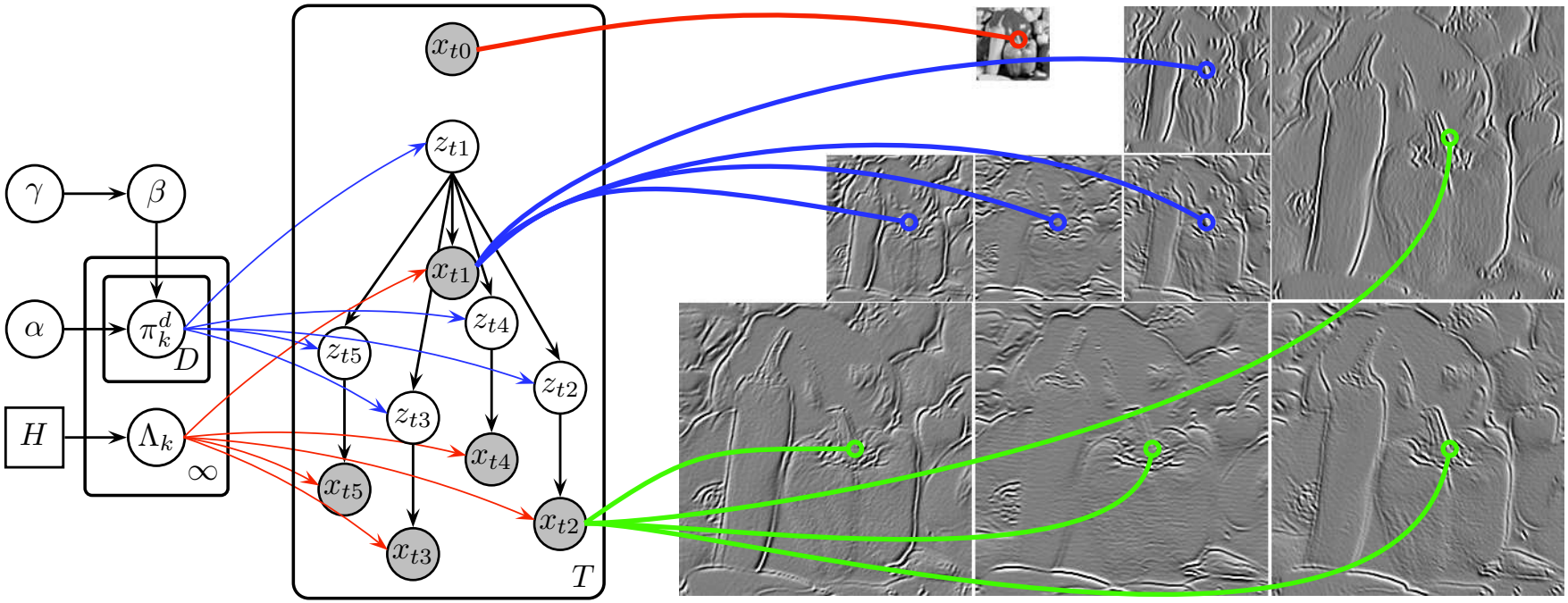
$$p(x) = \sum_{k=1}^{\infty} \beta_k f(x \mid \Lambda_k)$$

$$\beta \sim \text{Stick}(\gamma)$$

$$\Lambda_k \sim H$$

- Basis for *nonparametric* models whose complexity grows as additional data is observed
  - Makes simple predictions given few observations
  - Low-weight clusters capture details of very large datasets
- Attractive *asymptotic guarantees*
  - Posterior consistency of DP mixture density estimators
  - Convergence to finite mixture parameters of any order
- Leads to simple, effective *computational methods*
  - Growing literature on Monte Carlo and variational methods
  - Integrated, efficient handling of models of varying orders

Hierarchical Dirichlet Process Hidden Markov Trees



z_{ti} → indexes *infinite* set of hidden states

$$z_{ti} \in \{1, 2, 3, \ldots\}$$

π_k → infinite set of state *transition* distributions

$$z_{ti} \sim \pi_{z_{\mathrm{Pa(ti)}}}^{d_{ti}}$$

x_{ti} → observed *vector* of wavelet coefficients

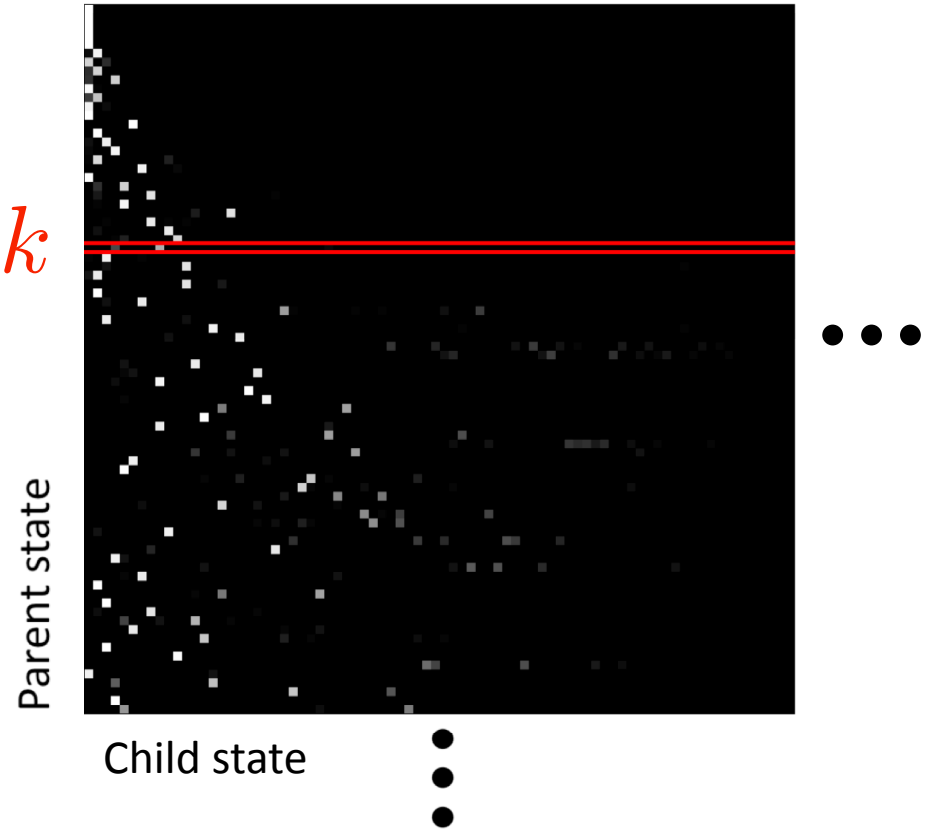
Λ_k → state-specific *emission* covariances

$$x_{ti} \sim \mathcal{N}(0, \Lambda_{z_{ti}})$$
$$\Lambda_k \sim H$$

Why a Hierarchical DP ? *(Teh et. al. 2004)*

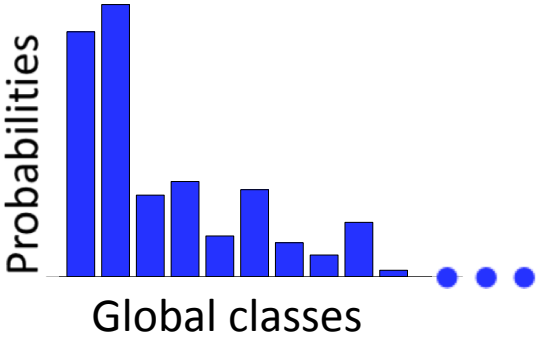
- Hierarchical DP prior allows us to learn a potentially infinite set of *appearance patterns* from natural images
- Hierarchical coupling ensures, with high probability, that a common set of *child* states are reachable from each *parent*

$$\pi_k^{d_{ti}}(\ell) = \Pr\left[z_{ti} = \ell \,|\, z_{\mathrm{Pa}(ti)}\right]$$



$$\beta \sim \mathrm{Stick}(\gamma)$$

Average state frequencies



$$\pi_k^d \sim \mathrm{DP}(\alpha, \beta)$$

Transition distributions

$$\mathbb{E}\left[\pi_k^d\right] = \beta$$

$\alpha \longrightarrow$ *Sparsity & variability of transition distributions*

Learning HDP-HMT Models with a Collapsed Gibbs Sampler

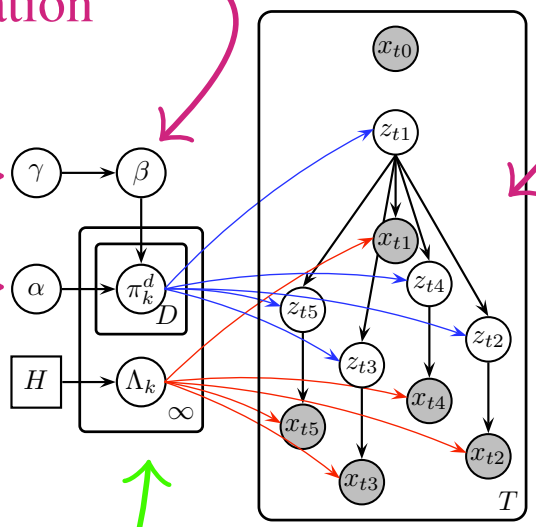
Sample global mixture proportions β

Sample hidden assigments z:

Consider
- states assigned to existing clusters
- a potential new state

(state cardinality determination)

Sample concentration parameters γ, α



Marginalize state-specific parameters $\{\pi, \theta\}$

Rao-Blackwellization improves the efficiency and accuracy of MCMC methods

Learning with a Truncated Gibbs Sampler

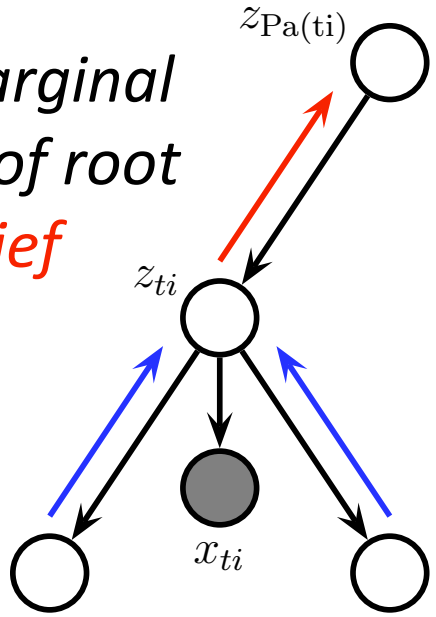
- Weak limit approximation uses high probability *upper bounds* on the number of states underlying a finite dataset:

$$\beta = (\beta_1, \ldots, \beta_K) \sim \mathrm{Dir}(\gamma/\mathrm{K}, \ldots, \gamma/\mathrm{K})$$

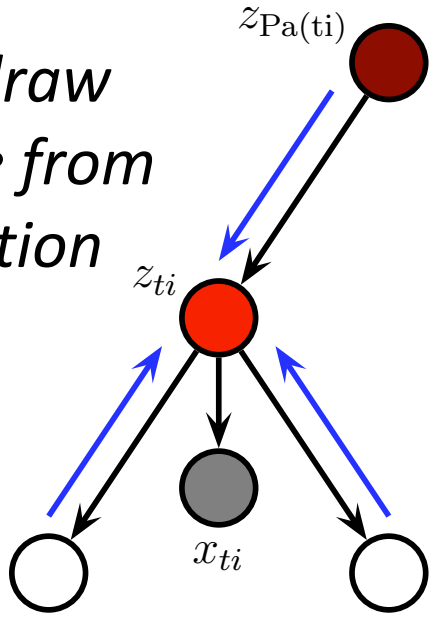
$$\pi_t = (\pi_{t1}, \ldots, \pi_{tK}) \sim \mathrm{Dir}(\alpha\beta_1, \ldots, \alpha\beta_\mathrm{K})$$

- Predictions from *truncated* model converge in distribution to HDP as $K \to \infty$, and allow efficient *blocked sampling*:

Compute marginal distribution of root state via belief propagation

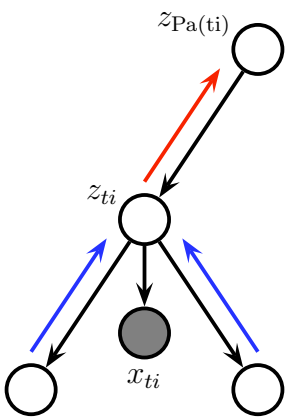


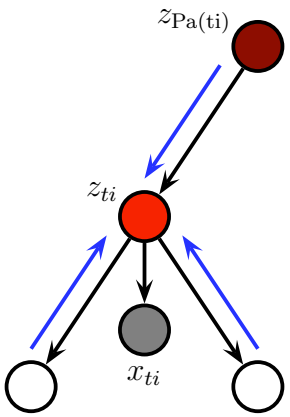
Recursively draw exact sample from joint distribution of all states



Learning with a Truncated Gibbs Sampler

Sample hidden state assignments jointly using belief propagation:

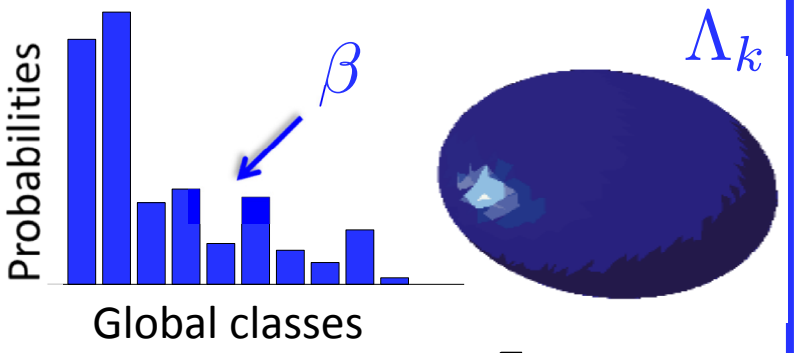




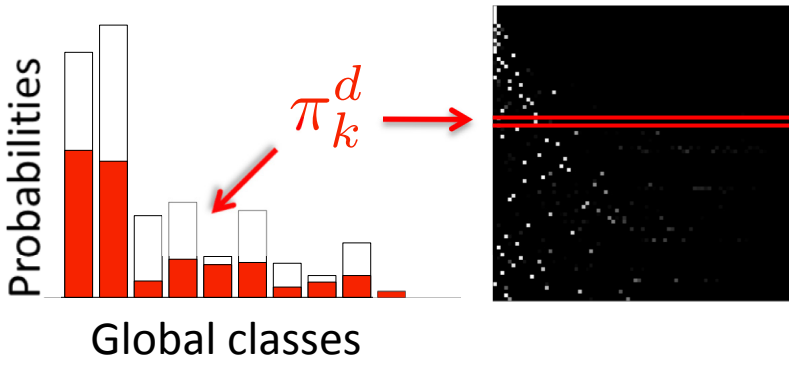
Compute messages from leaves to root

Sample assignments from root to leaves

Sample parameters:

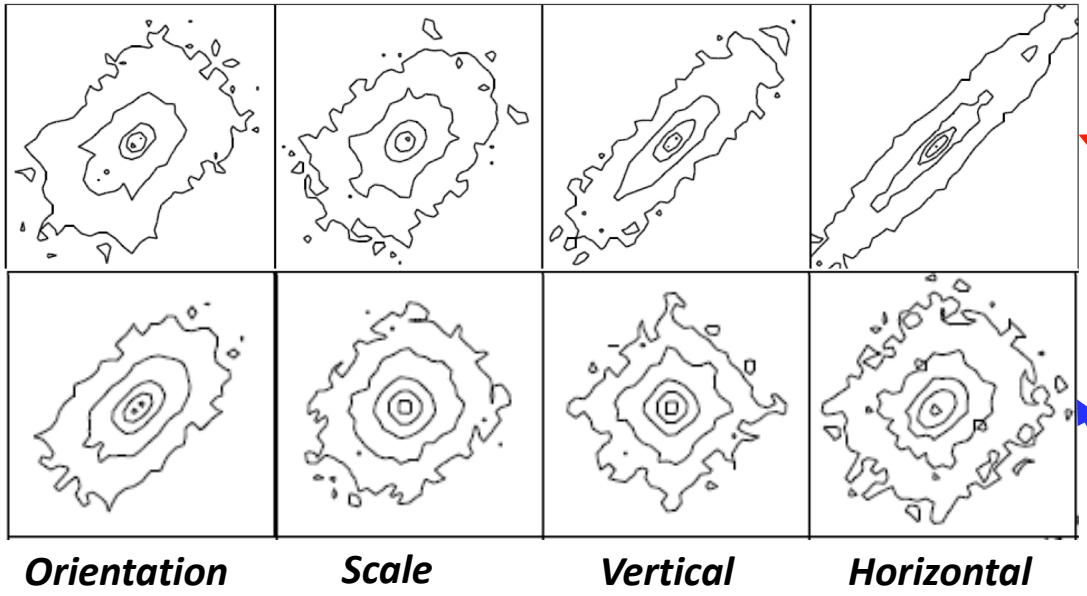


Feature distributions

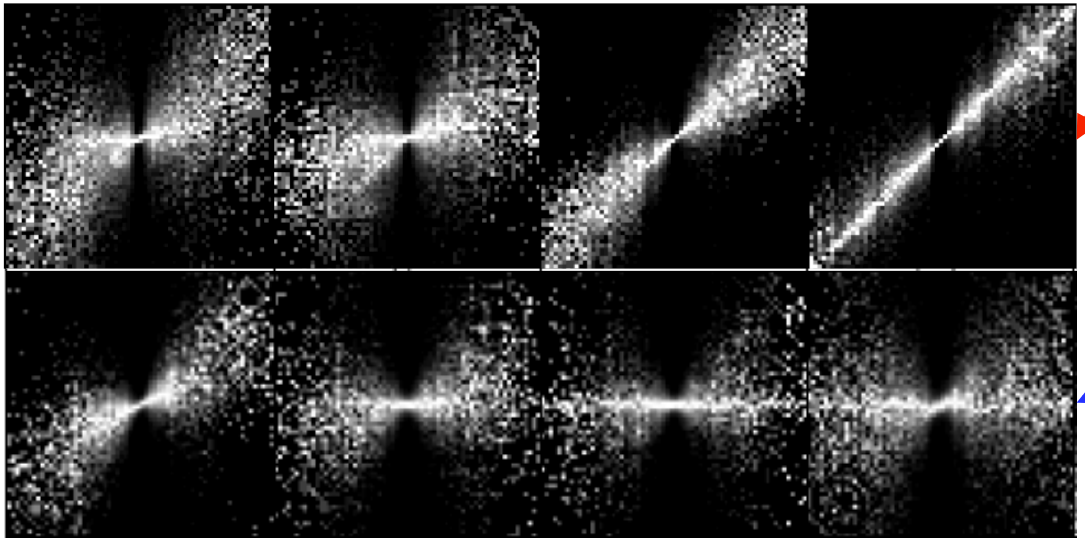


Transition probabilities

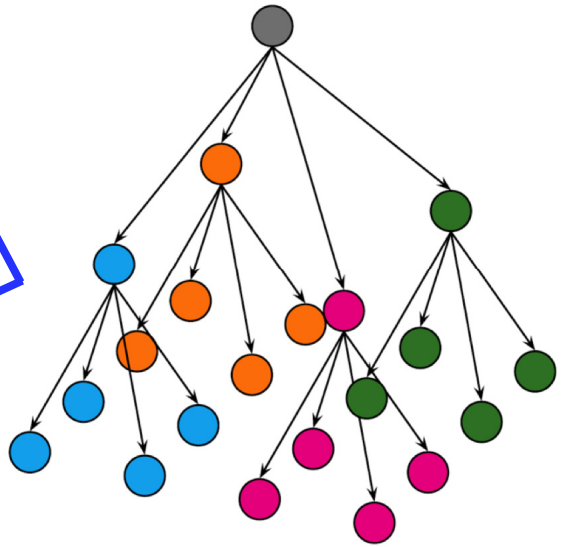
Pairwise Wavelet Histograms



Image



Model



Outline

Multiscale Models for Natural Images

- Nonparametric Hidden Markov Trees (HDP-HMTs)
- Learning with Monte Carlo methods
- Truncated representations for efficient learning from large datasets

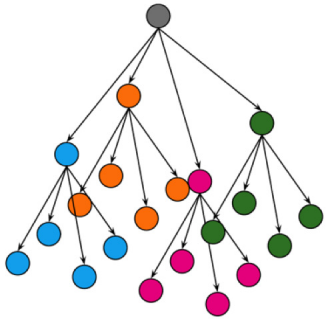
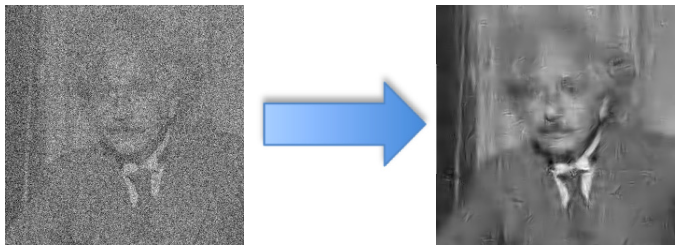


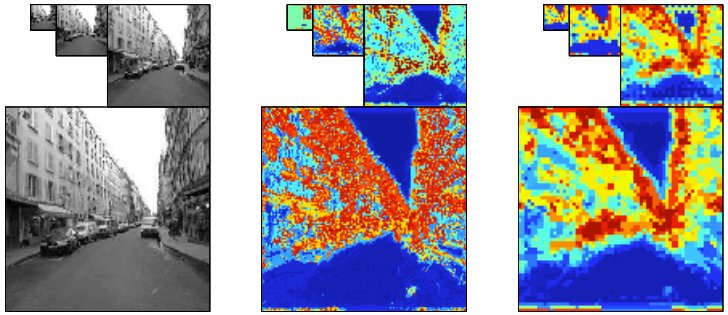
Image Denoising

- Transfer natural image statistics for making robust predictions



Natural Scene Analysis

- Global, data-driven scene models via HDP-HMT
- Fast categorization via Belief Propagation methods



Denoising Images in Wavelet-Domain

Noisy

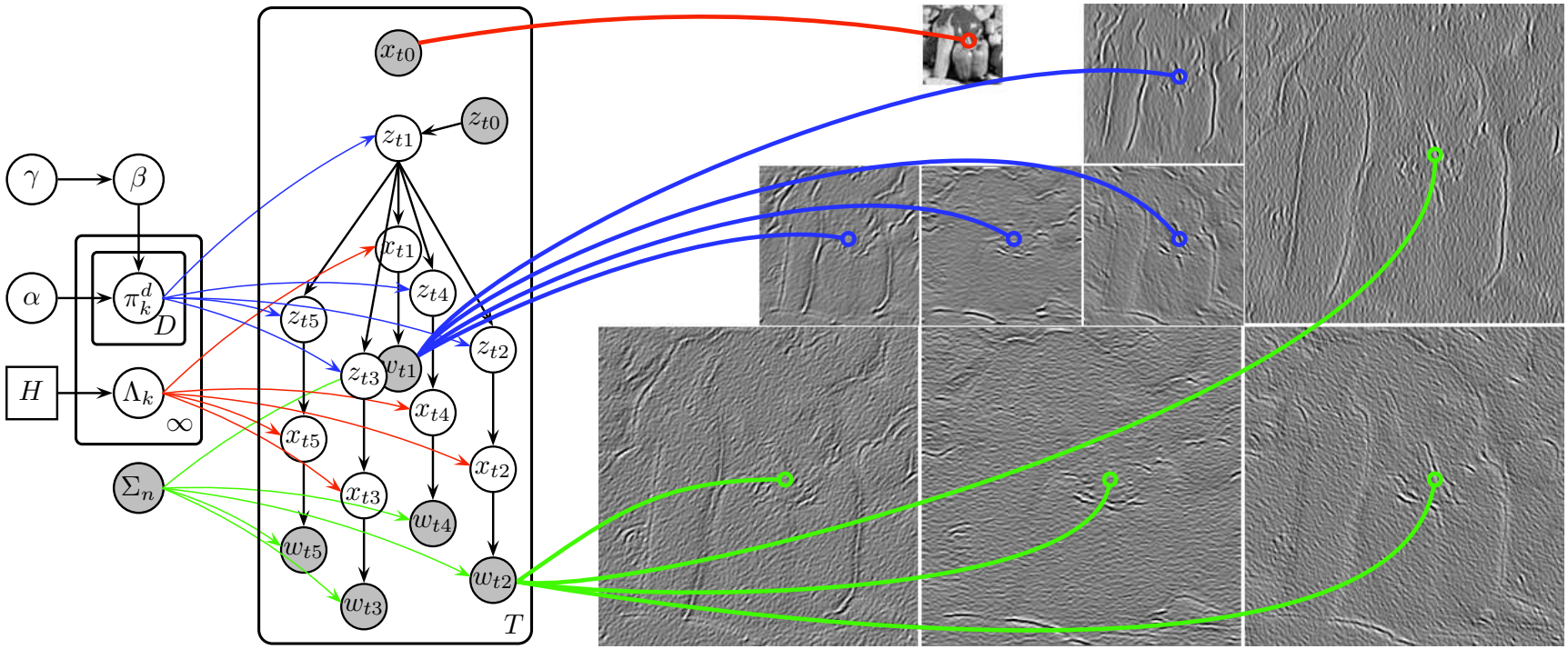
HDP-HMT (Emp. Bayes)

HDP-HMT (Transfer)

Apply learned statistics to denoise images

- Using a global model increases robustness
- Exploit availability of large image databases to develop efficient *transfer* denoising algorithms
- Improve performance by reusing statistics of clean images

HDP-HMT for noisy data



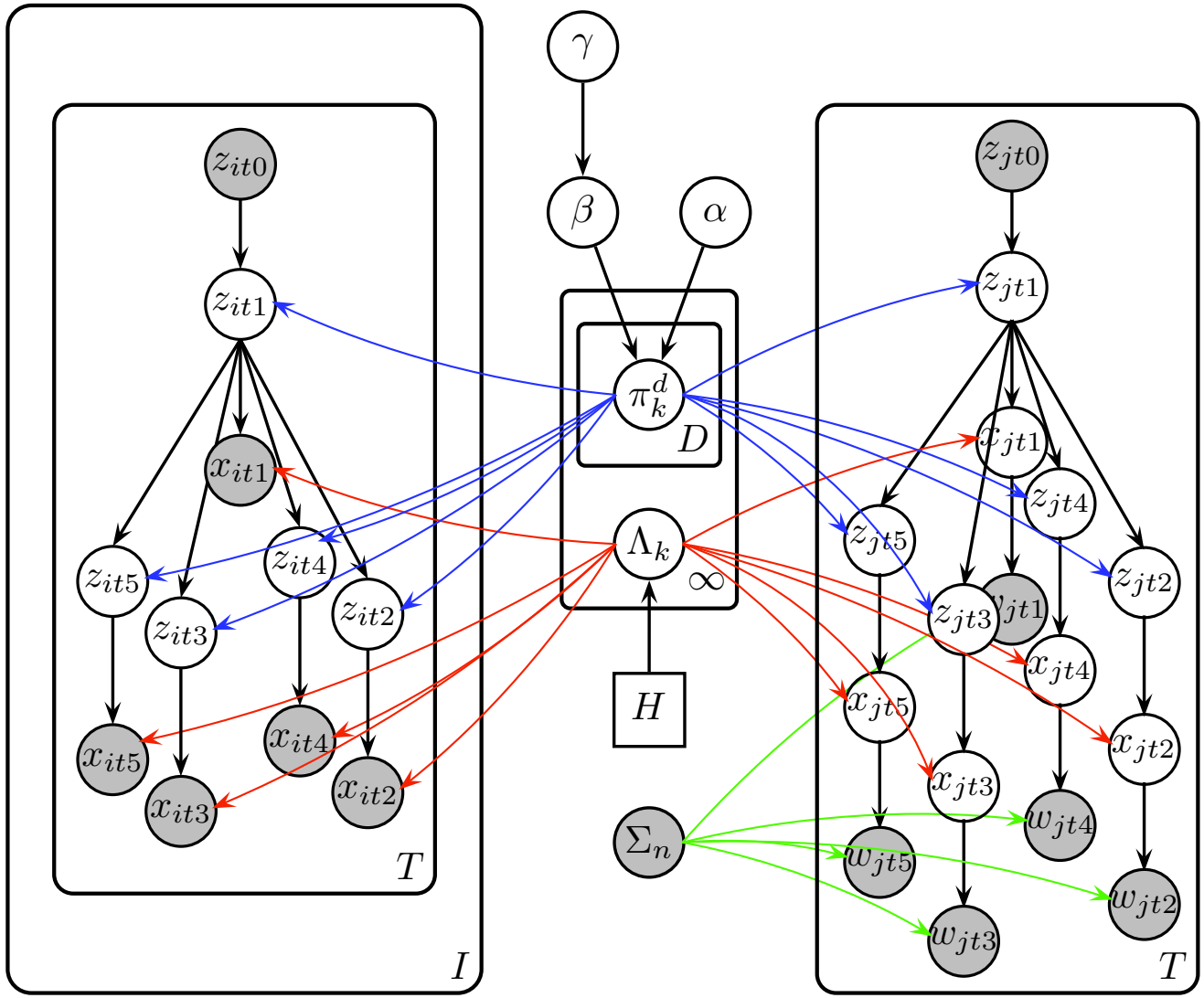
x_{ti} → unobserved vector of *clean* wavelet coefficients

Σ_n → noise variance

w_{ti} → observed vector of *noisy* wavelet coefficients

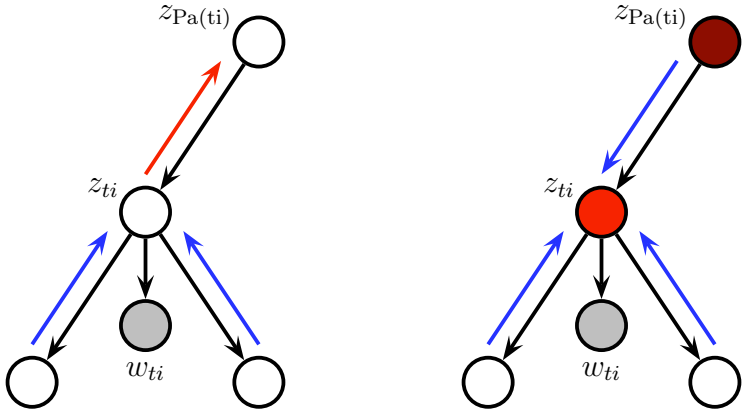
$$w_{ti} \sim \mathcal{N}(x_{ti}, \Sigma_n)$$

… and for clean data as well



Learning via Gibbs Sampling

Sample hidden states jointly using BP:



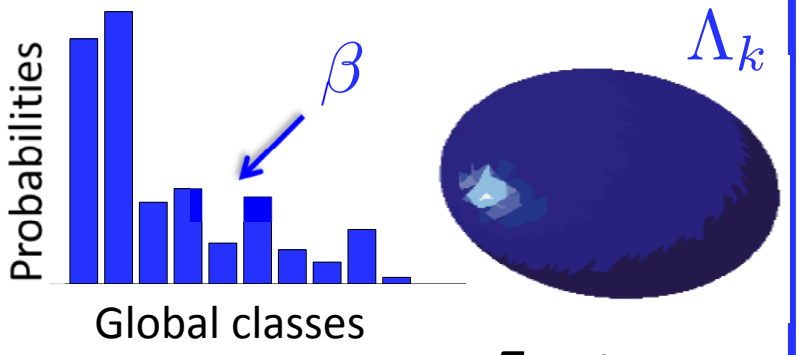
Compute msgs from leaves to root

Sample states from root to leaves

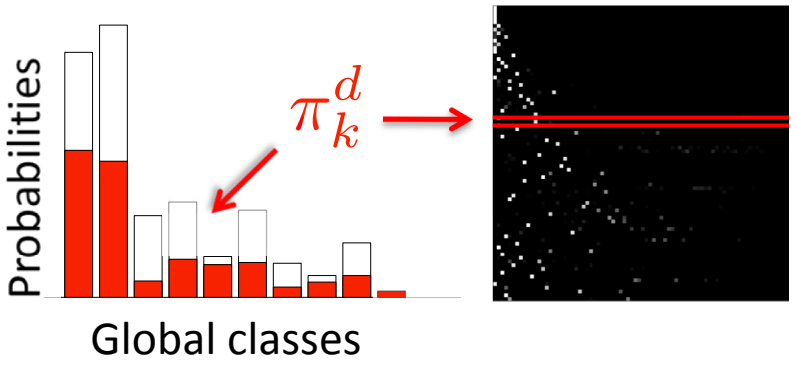
For noisy images, sample clean wavelet coefficients:

$$x_{ti} \sim \mathcal{N}(\mu, \Sigma)$$
$$\mu = \left(\Lambda_{z_{ti}}^{-1} + \Sigma_n^{-1}\right)^{-1} \Sigma_n^{-1} w_{ti}$$
$$\Sigma = \left(\Lambda_{z_{ti}}^{-1} + \Sigma_n^{-1}\right)^{-1}$$

Sample parameters:

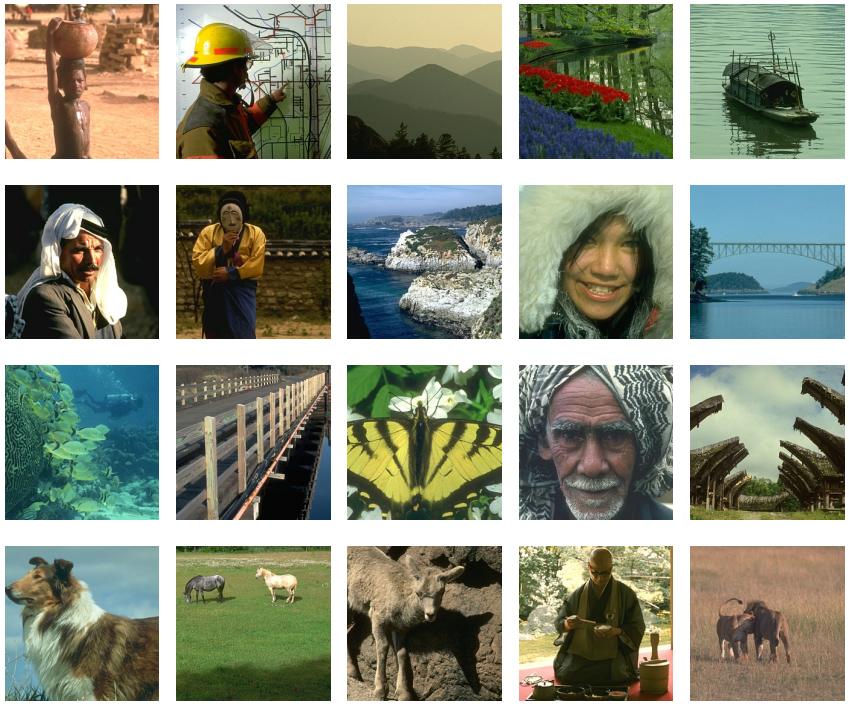


Feature distributions

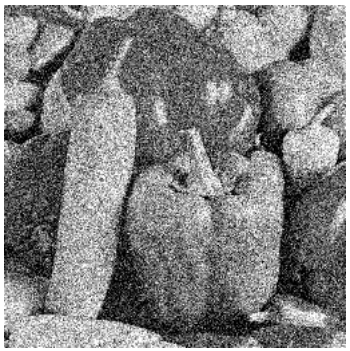


Transition probabilities

Estimating Clean Images



Empirical Bayesian approach estimates model parameters from the noisy image



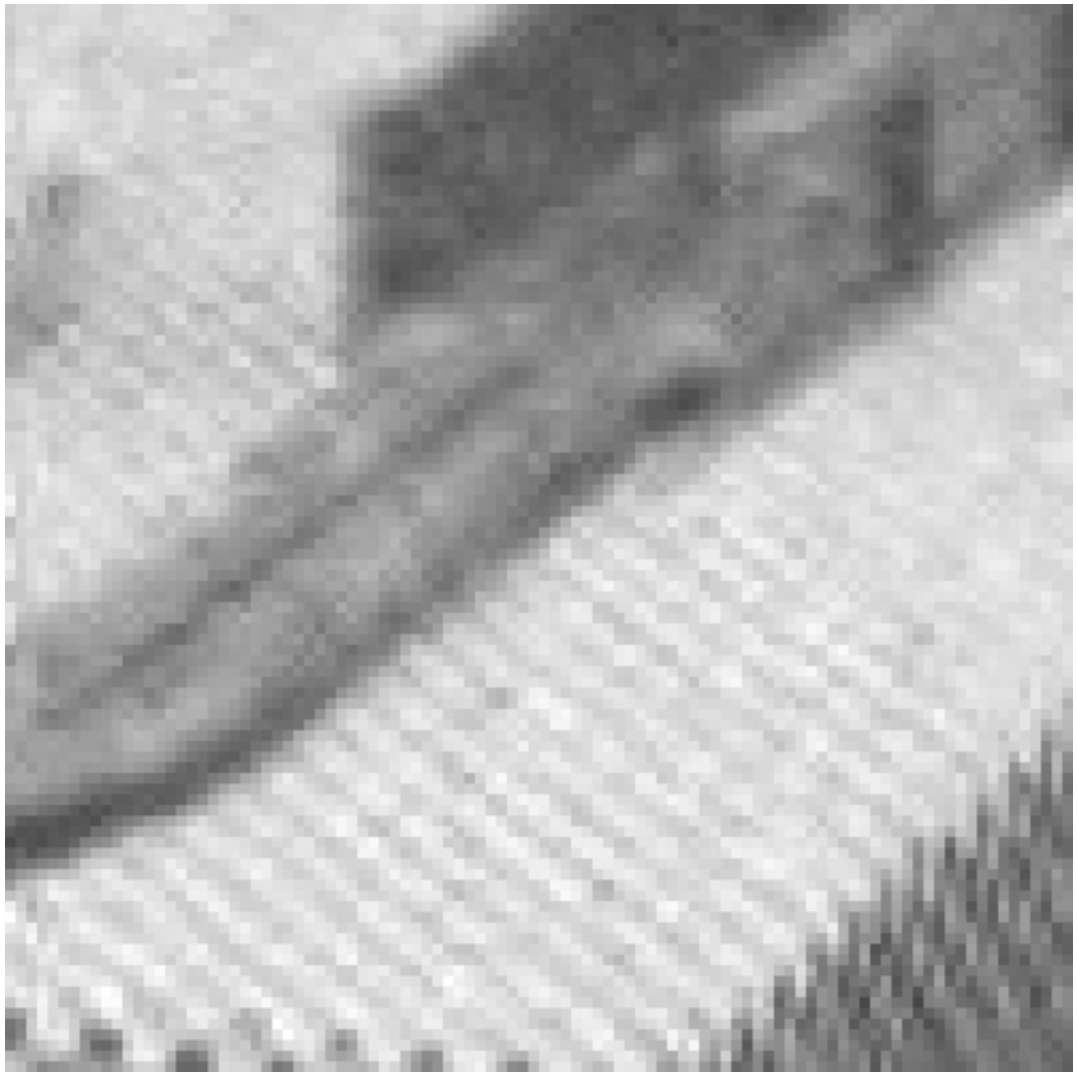
Transfer denoising approach reuses multiscale hidden state patterns of clean images for making robust predictions

$$\mathbb{E}\left[x_{ti} \mid \mathbf{w}, \theta^{(s)}\right] = \sum_{k=1}^{K_s} p(z_{ti} = k \mid \mathbf{w}, \theta^{(s)})\, \mathbb{E}\left[x_{ti} \mid w_{ti}, \Lambda_k^{(s)}\right]$$

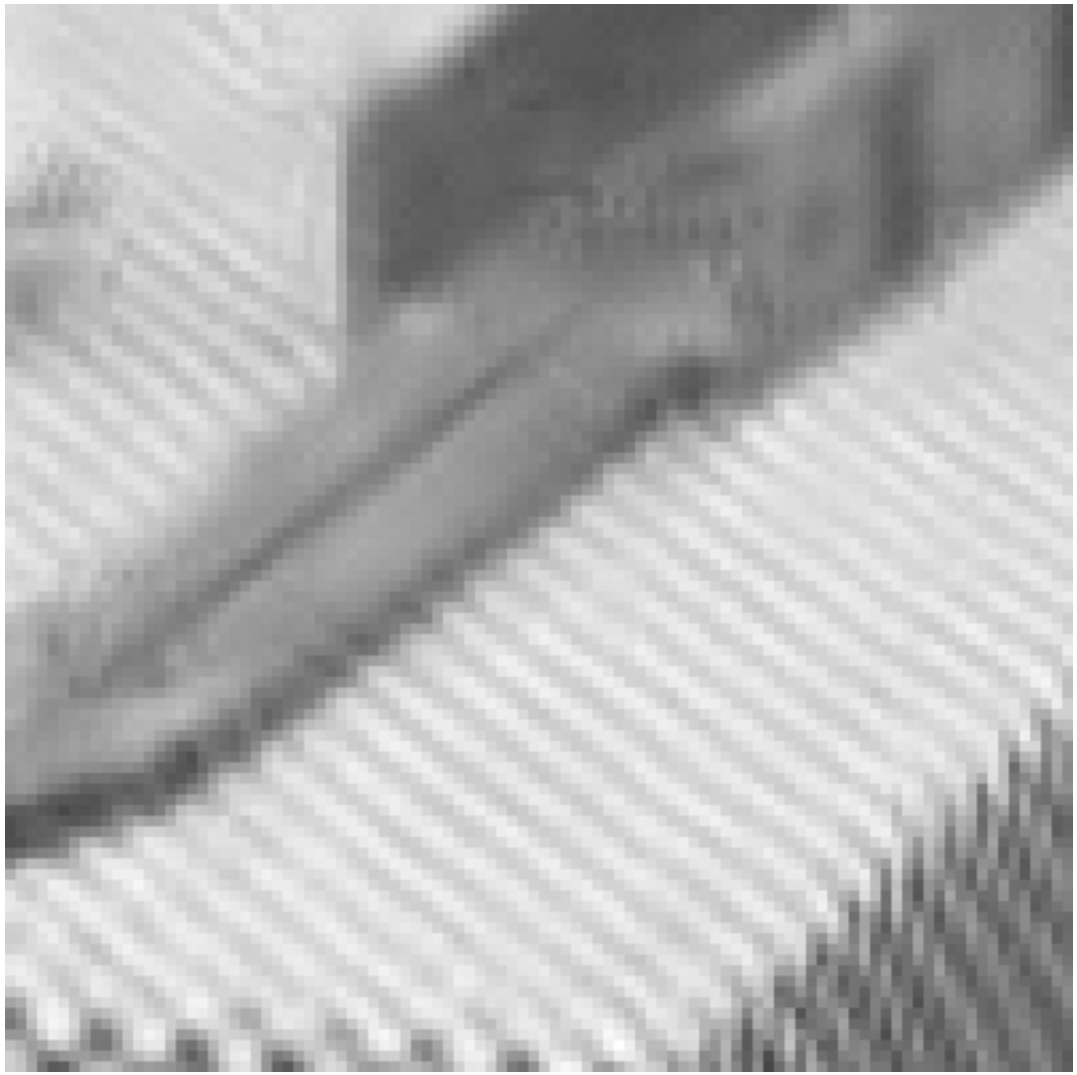
From belief propagation

Linear least-squares smoothing

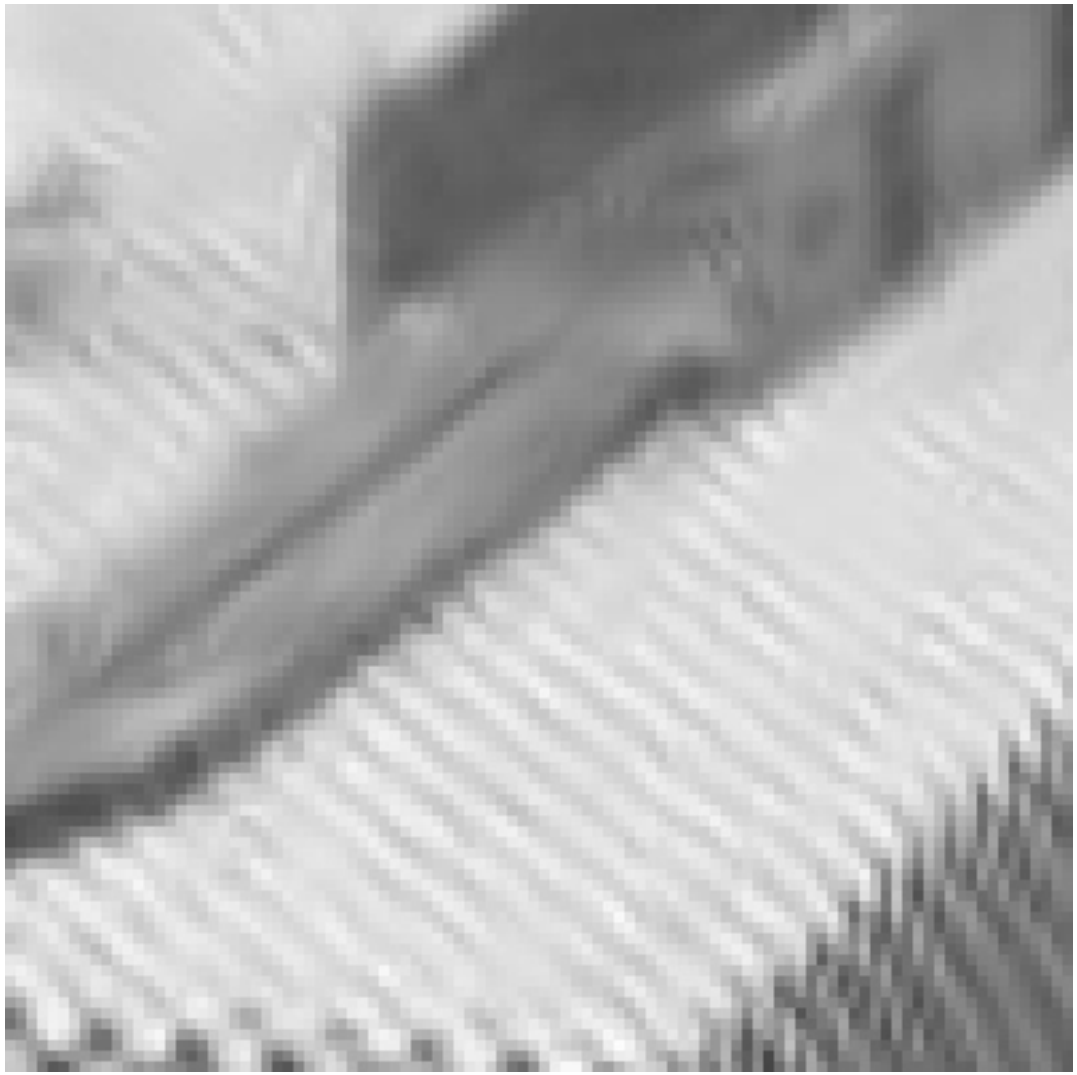
Denoising: Binary HMT



Denoising: HDP-HMT (Emp. Bayes)



Denoising: Local GSM



Portilla, Strela, Wainwright, & Simoncelli, 2003

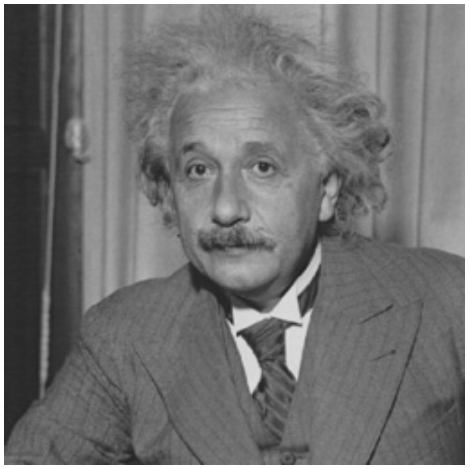
Denoising Einstein

Noisy
10.60 dB, 0.057

HDP-HMT
(Emp. Bayes)
25.64 dB, 0.564

HDP-HMT
(Transfer)
26.80 dB, 0.664

Original



BLS-GSM
26.38 dB, 0.647

BM3D
26.49 dB, 0.659

Denoising Hill

Noisy
8.12 dB, 0.038

HDP-HMT
(Emp. Bayes)
24.56 dB, 0.540

HDP-HMT
(Transfer)
24.74 dB, 0.568

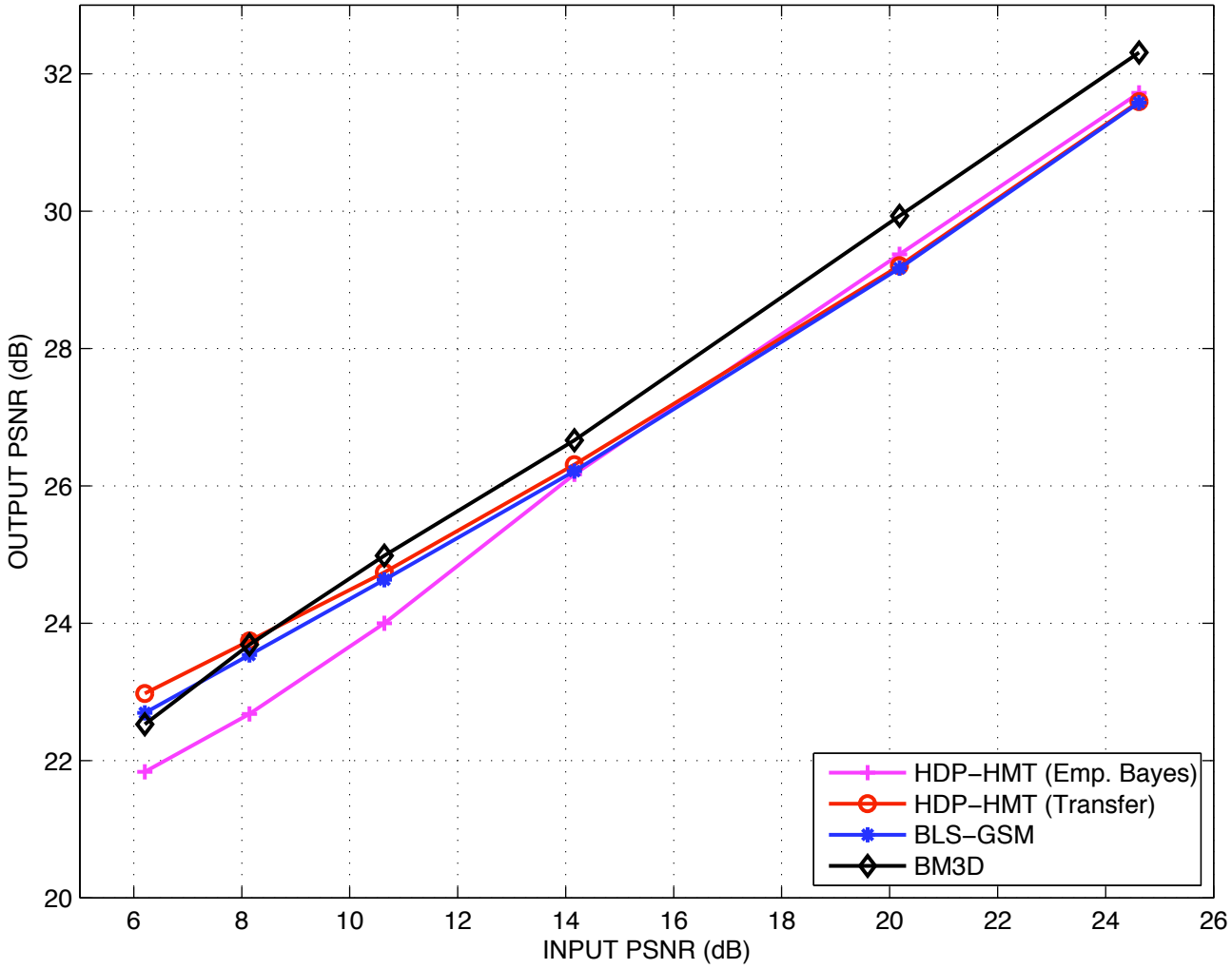
Original

BLS-GSM
24.54 dB, 0.544

BM3D
24.39 dB, 0.548

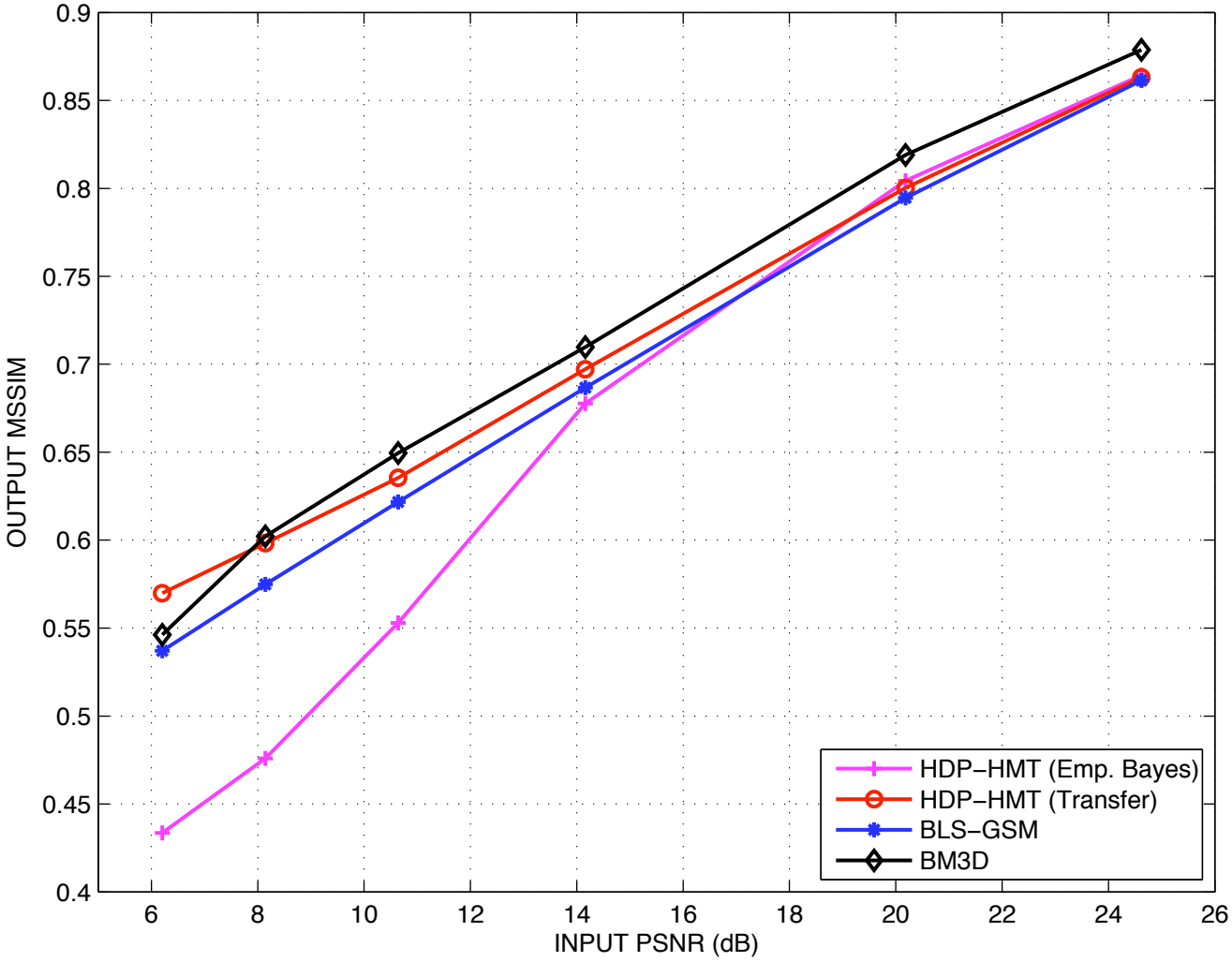
Average Denoising Performance

Peak signal-to-noise ratio (PSNR)



Average Denoising Performance

Mean structural similarity index (SSIM)



Natural Scene Denoising

Noisy
8.14 dB, 0.033

HDP-HMT
(Emp. Bayes)
24.24 dB, 0.519

HDP-HMT
(Transfer)
26.50 dB, 0.794

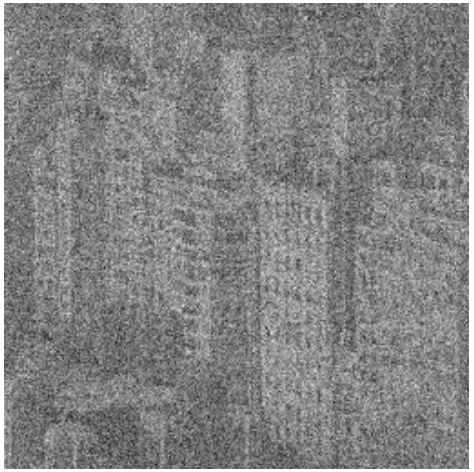
Original

BLS-GSM
25.59 dB, 0.726

BM3D
25.74 dB, 0.751

Natural Scene Denoising

Noisy
8.14 dB, 0.177



HDP-HMT
(Emp. Bayes)
18.55 dB, 0.484

HDP-HMT
(Transfer)
18.77 dB, 0.486

Original

BLS-GSM
18.59 dB, 0.454

BM3D
18.65 dB, 0.470

Natural Scene Denoising

HDP-HMT (Transfer)
23.28 dB, 0.653

BLS-GSM
23.14 dB, 0.617

BM3D
23.23 dB, 0.651

Outline

Multiscale Models for Natural Images

- Nonparametric Hidden Markov Trees (HDP-HMTs)
- Learning with Monte Carlo methods
- Truncated representations for efficient learning from large datasets

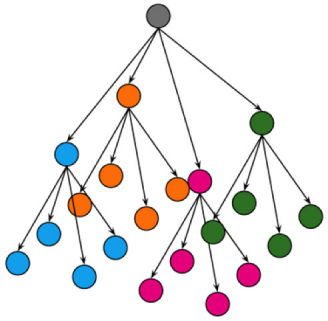
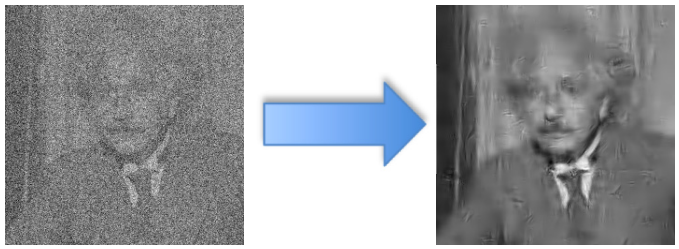


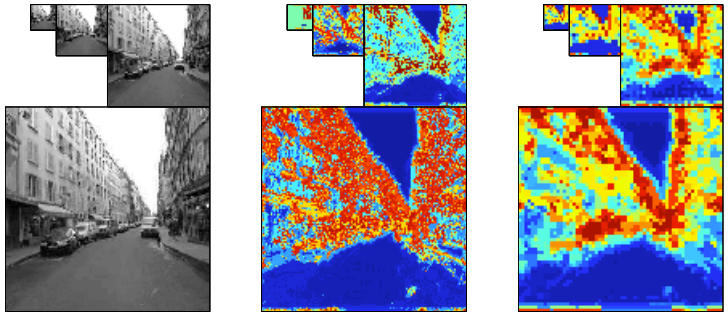
Image Denoising

- Transfer natural image statistics for making robust predictions

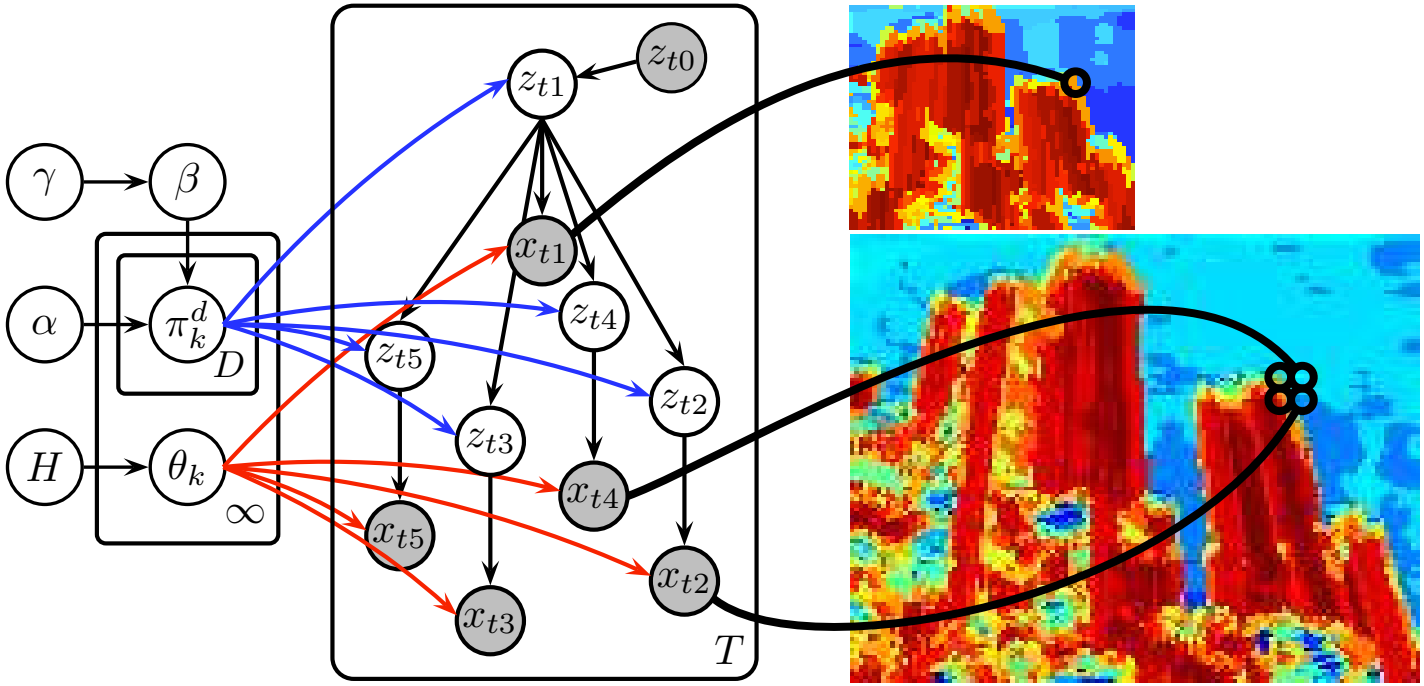


Natural Scene Analysis

- Global, data-driven scene models via HDP-HMT
- Fast categorization via Belief Propagation methods



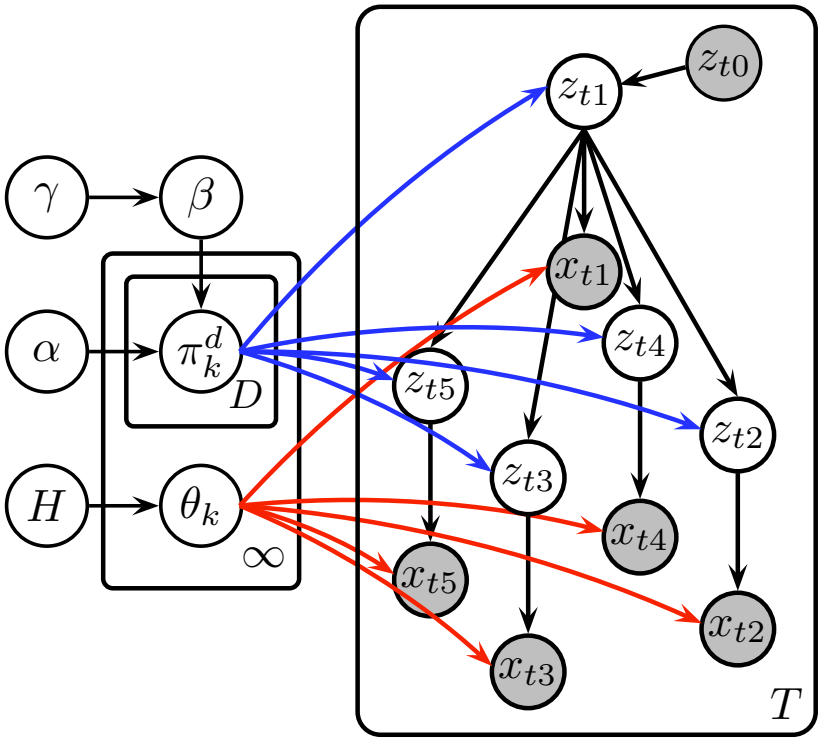
HDP-HMT Scene Model



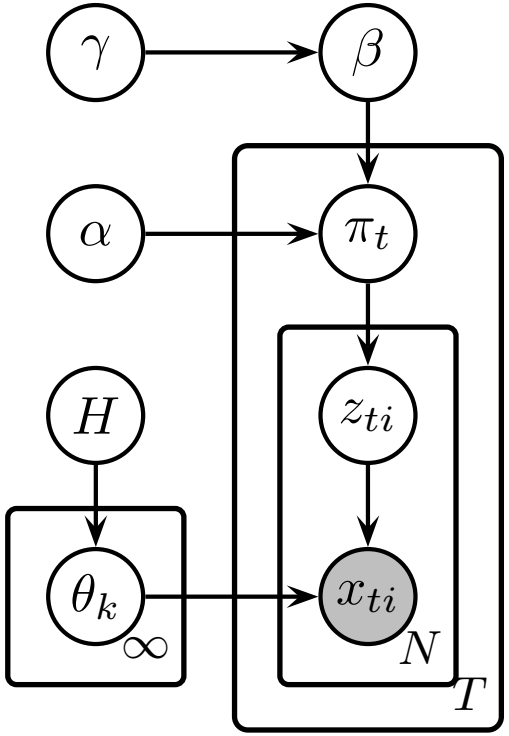
- Hidden states z_{ti} generate vectors of clean wavelet coefficients x_{ti} at multiple orientations or SIFT-descriptors

… versus baseline HDP-BOF

HDP-HMT

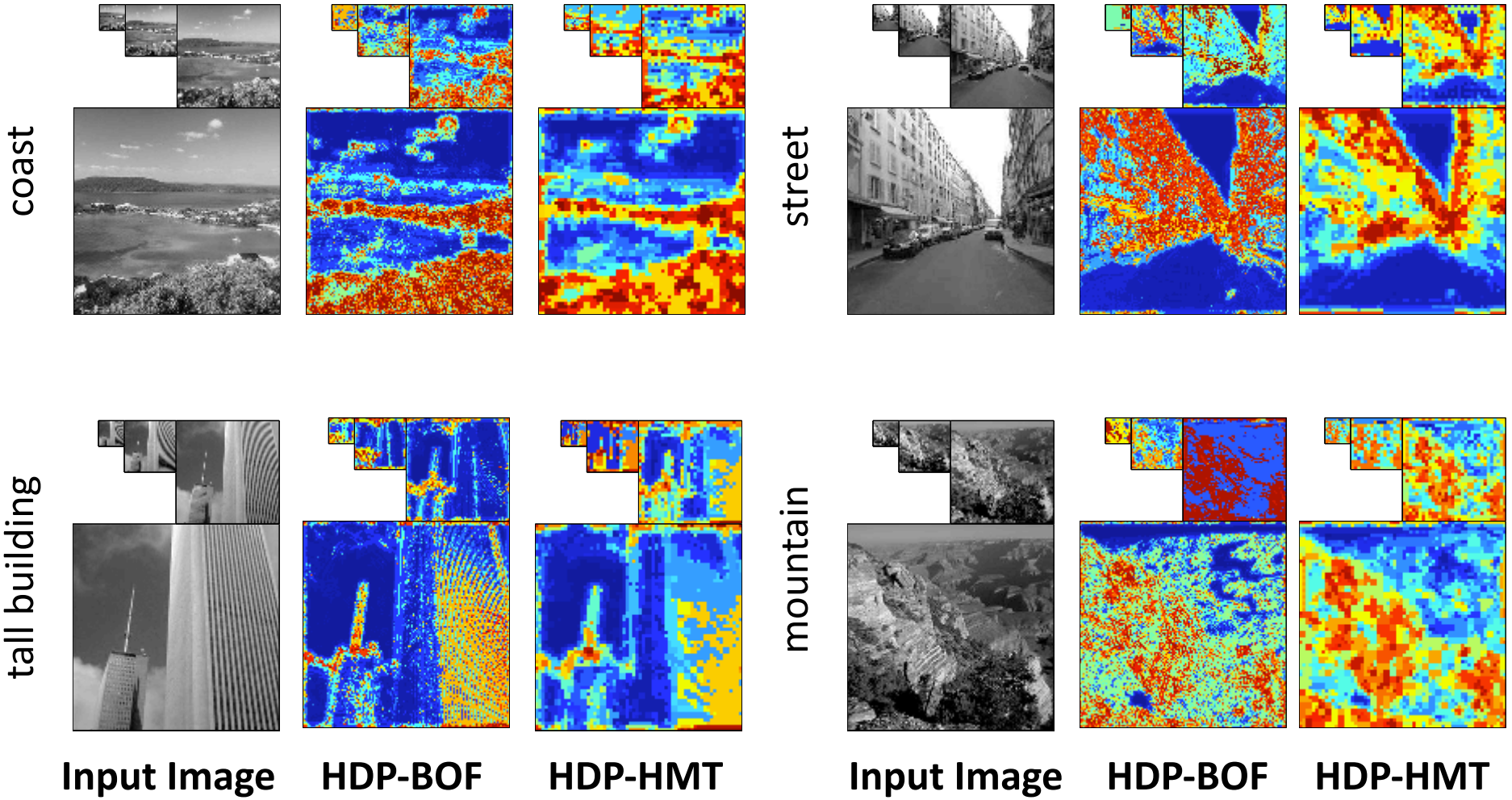


HDP-BOF



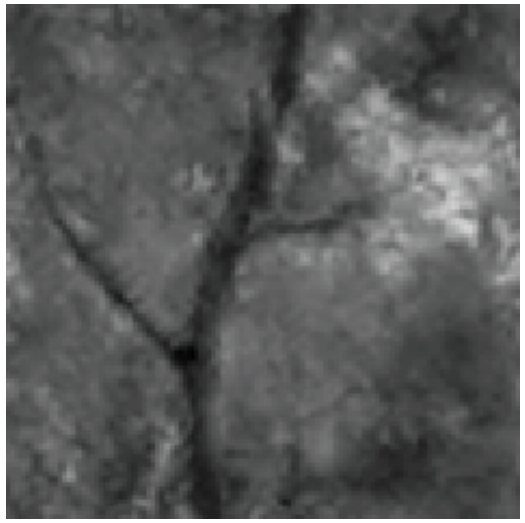
- A nonparametric Bayesian extension of the LDA-based model for scene categorization by Fei-Fei and Perona (2005), which ignores spatial dependencies in the appearance of locally extracted image features
- The HDP-HMT further extends this by incorporating dependencies in feature appearance across location and scale

MAP assignments (sp5)

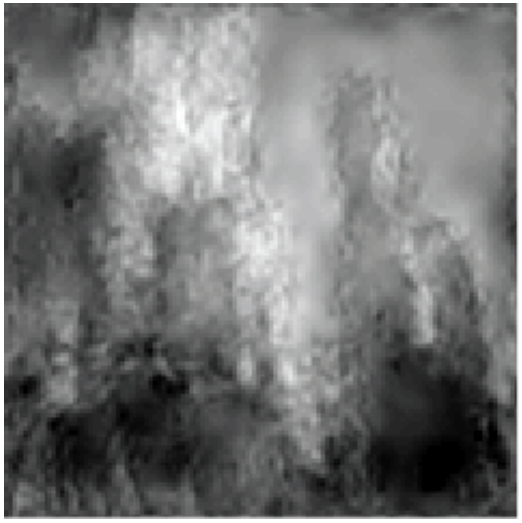


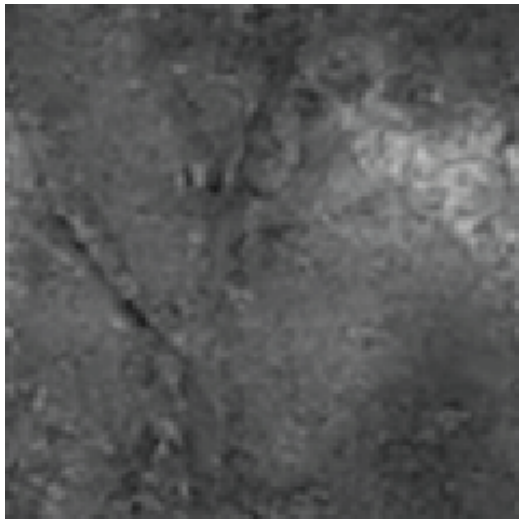
Samples given MAP states

Input Image



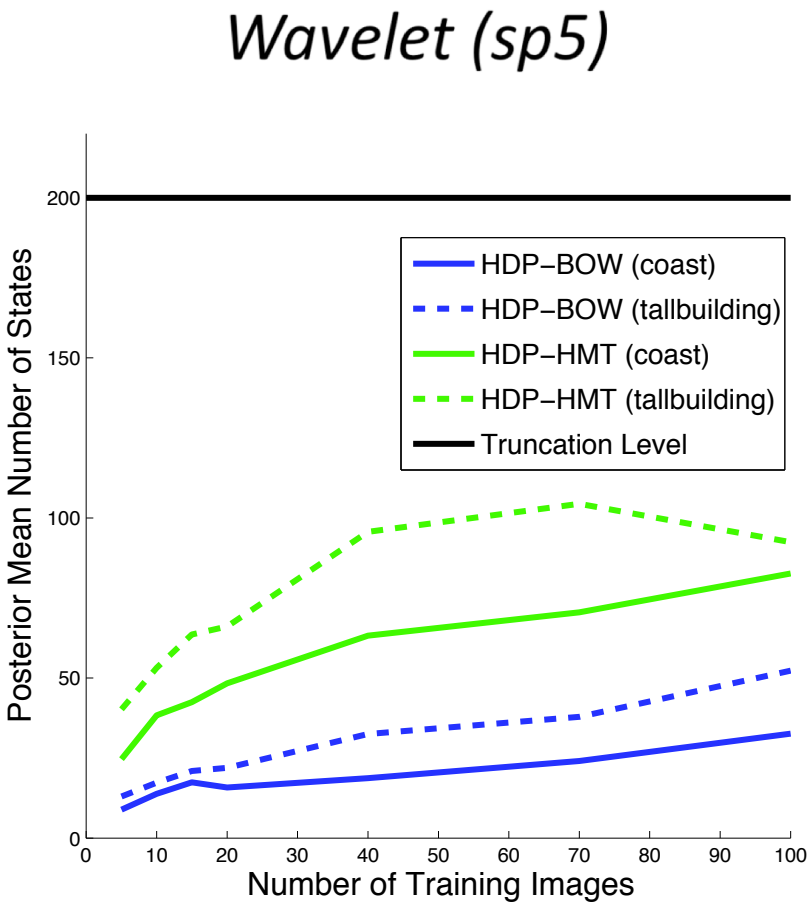
HDP Hidden Markov Tree

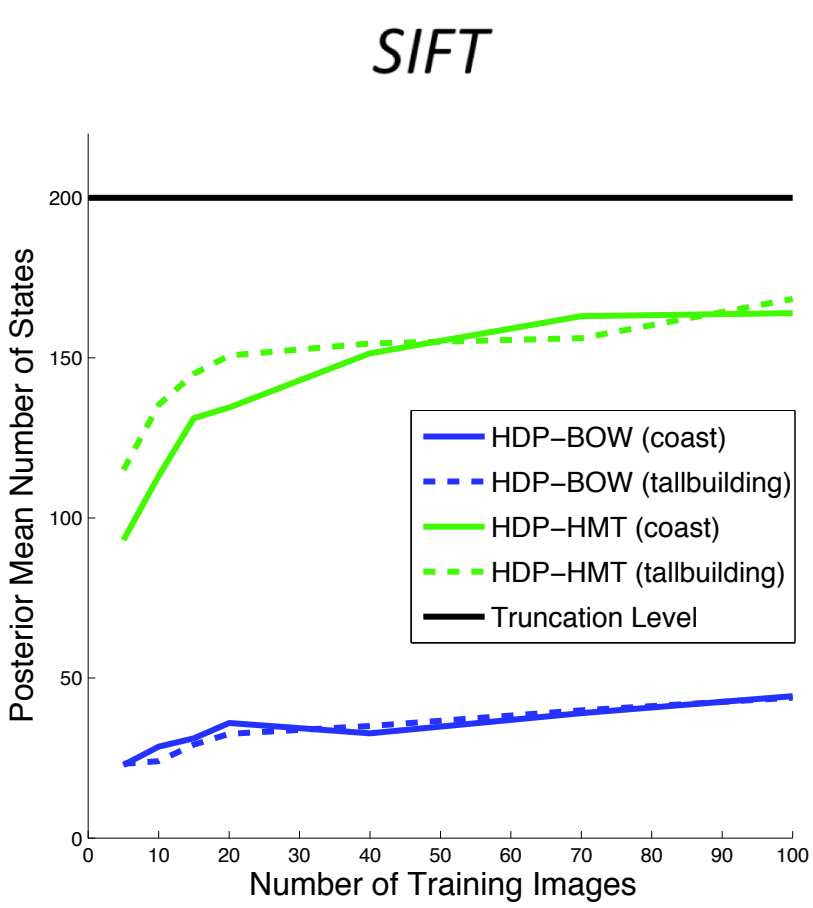




HDP Bag of Features

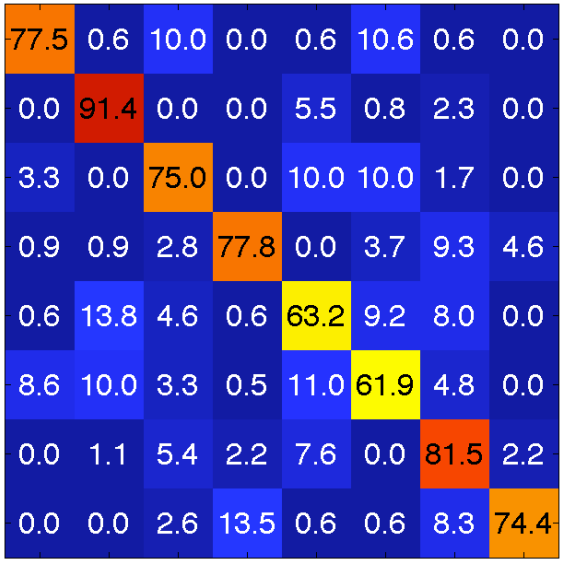
Number of States





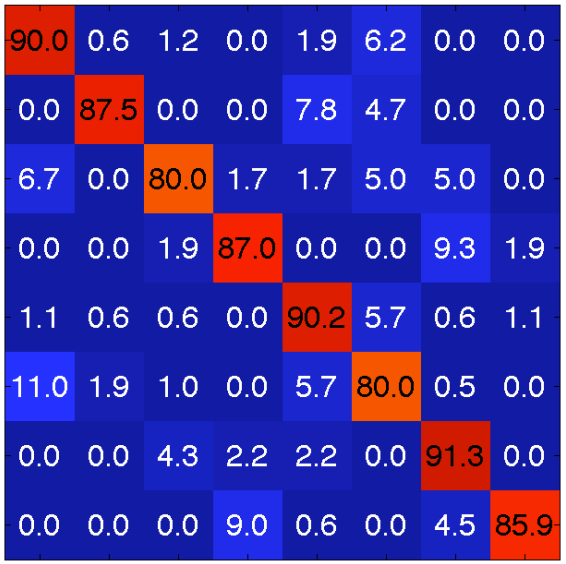
Categorizing Natural Scenes

Wavelet (sfp7)

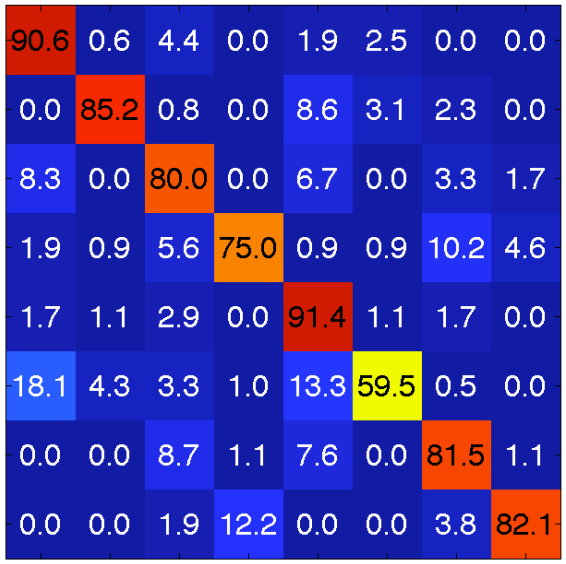


HDP-BOF [75.3 %]

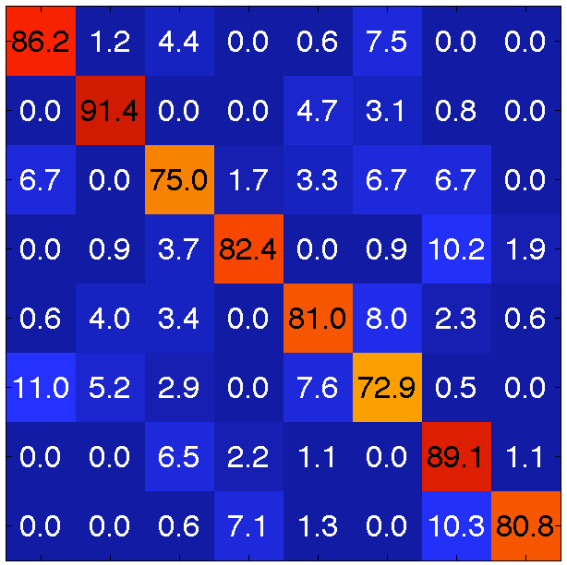
SIFT



HDP-BOF [82.4 %]



HDP-HMT [80.7 %]



HDP-HMT [86.5 %]

Average Categorization Performance

	Wavelet (sfp7)		SIFT	
Man-made	82.9	85.4	86.4	89.7
Natural	78.6	83.5	85.7	87.7
Eight	75.3	80.7	82.4	86.5
Thirteen			75.9	81.8
Fifteen			69.7	77.1
	HDP-BOW	HDP-HMT	HDP-BOW	HDP-HMT

Summary and Conclusions

- Presented a hierarchical nonparametric Bayesian model for multiscale datasets with complex non-local dependencies
- Presented MCMC methods for learning HDP-HMT parameters from clean and noisy images
- Truncated representations of the DP to allow efficient blocked sampling algorithms and learning from large datasets
- The HDP-HMT captures complex natural image structures and leads to effective learning algorithms for denoising and categorization
- Robust restoration with natural image statistics transfer

Pairwise Statistics of Wavelets