From Patches to Images: A Nonparametric Generative Model

Geng Ji [1] Michael C. Hughes [2] Erik B. Sudderth [1 3]

Abstract

We propose a hierarchical generative model that captures the self-similar structure of image regions as well as how this structure is shared across image collections. Our model is based on a novel, variational interpretation of the popular expected patch log-likelihood (EPLL) method as a model for randomly positioned grids of image patches. While previous EPLL methods modeled image patches with finite Gaussian mixtures, we use nonparametric Dirichlet process (DP) mixtures to create models whose complexity grows as additional images are observed. An extension based on the hierarchical DP then captures repetitive and self-similar structure via image-specific variations in cluster frequencies. We derive a structured variational inference algorithm that adaptively creates new patch clusters to more accurately model novel image textures. Our denoising performance on standard benchmarks is superior to EPLL and comparable to the state-of-the-art, and we provide novel statistical justifications for common image processing heuristics. We also show accurate image inpainting results.

1. Introduction

Models of the statistical structure of natural images play a key role in computer vision and image processing (Srivastava et al., 2003). Due to the high dimensionality of the images captured by modern cameras, a rich research literature instead models the statistics of small image patches. For example, the K-SVD method (Elad & Aharon, 2006) generalizes K-means clustering to learn a dictionary for sparse coding of image patches. The state-of-the-art *learned simultaneous sparse coding* (LSSC, Mairal et al. (2009)) and *block matching and 3D filtering* (BM3D, Dabov et al. (2008)) methods integrate clustering, dictionary learning, and denoising to extract information directly from a single corrupted image. Alternatively, the accurate *expected patch log-likelihood* (EPLL, Zoran & Weiss (2011)) method maximizes the log-likelihood of overlapping image patches under a finite Gaussian mixture model learned from uncorrupted natural images.

We show that with minor modifications, the objective function underlying EPLL is equivalent to a variational log-likelihood bound for a novel generative model of whole images. Our model coherently captures overlapping image patches via a randomly positioned spatial grid. By deriving a rigorous variational bound, we then develop improved nonparametric models of natural image statistics using the *hierarchical Dirichlet process* (HDP, Teh et al. (2006)). In particular, DP mixtures allow an appropriate model complexity to be inferred from data, while the hierarchical DP captures the patch self-similarities and repetitions that are ubiquitous in natural images (Jégou et al., 2009). Unlike previous whole-image generative models such as *fields of experts* (FoE, Roth & Black (2005)), which uses a single set of Markov random field parameters to model all images, our HDP model learns image-specific clusters to accurately model distinctive textures. Coupled with a scalable structured variational inference algorithm, we improve on the excellent denoising accuracy of the LSSC and BM3D algorithms, while providing a Bayesian nonparametric model with a broader range of potential applications.

2. Expected Patch Log-likelihood

Our approach is derived from models of small (8×8 pixel) patches of a large natural image x. Let P_i be a binary indicator matrix that extracts the $G = 8^2$ pixels $P_i x \in \mathbb{R}^G$ in patch i. To reduce sensitivity to lighting variations, a *contrast normalizing* transform is applied to remove the mean (or "DC component") of the pixel intensities in each patch:

$$v_i = P_i x - \tfrac{1}{G}\mathbf{1}^T P_i x = B P_i x, \qquad (1)$$

for a "zero-centering" matrix B. Zoran & Weiss (2012) show that a finite mixture of K zero-mean Gaussians,

$$p(v_i) = \textstyle\sum_{k=1}^{K} \pi_k \mathrm{Norm}(v_i \mid 0, \Lambda_k^{-1}), \qquad (2)$$

is superior to many classic image models in terms of predictive likelihood and patch denoising performance.

[1] Brown University, Providence, RI, USA. [2] Harvard University, Cambridge, MA, USA. [3] University of California, Irvine, CA, USA. Correspondence to: Geng Ji <gji@cs.brown.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

The widely-used EPLL image restoration framework measures the quality of a reconstruction by the expected patch log-likelihood, “assuming a patch location in the image is chosen uniformly at random” (Zoran & Weiss, 2011). Given a corrupted image y, EPLL estimates a clean image x by minimizing the objective:

$$\min_{x} \frac{\lambda}{2}\|x-y\|^2 - \textstyle\sum_i \log p(BP_i x). \qquad (3)$$

Here, the sum ranges over all *overlapping*, completely visible (uncropped) image patches. The constant λ is determined by the noise level of the corrupted image y.

Direct optimization of Eq. (3) is challenging, so inspired by *half quadratic splitting* (Geman & Yang, 1995), the EPLL objective can be reformulated as follows:

$$\min_{x,\bar{v}} \frac{\lambda}{2}\|x-y\|^2 + \sum_i \frac{\kappa}{2}\|P_i x - \bar{v}_i\|^2 - \log p(B\bar{v}_i). \quad (4)$$

Each patch i is allocated an auxiliary variable $\bar{v}_i$, which (unlike the v_i variable in Eq. (1)) includes an estimate of the mean patch intensity. This augmented objective leads to closed-form coordinate descent updates.

Gating. Assign each patch i to some cluster z_i:

$$z_i = \arg\max_k \; \pi_k \operatorname{Norm}\big(BP_i x \mid 0, \Lambda_k^{-1} + \kappa I\big). \qquad (5)$$

Filtering. Given an approximate clean image x and cluster assignments z, denoise patches via least squares:

$$\bar{v}_i = \Big(I + \kappa^{-1} B^T \Lambda_{z_i} B\Big)^{-1} P_i x. \qquad (6)$$

Mixing. Given a fixed set of auxiliary patches $\bar{v}$ and the noisy image y, a denoised image x is estimated as

$$x = \Big(\lambda I + \kappa \sum_i P_i^T P_i\Big)^{-1} \Big(\lambda y + \kappa \sum_i P_i^T \bar{v}_i\Big). \qquad (7)$$

Annealing. Optimal solutions of Eq. (4) approach those of the EPLL objective in Eq. (3) as $\kappa \to \infty$. EPLL denoising algorithms slowly increase κ via an annealing schedule that must be tuned for best performance.

Justification? Empirically, the intuitive EPLL objective is much more effective than baselines which use only a subset of non-overlapping patches, or average independently denoised patches (Zoran & Weiss, 2011). But why should we optimize the expected *log*-likelihood, instead of the expected likelihood or another function of patch-specific likelihoods? And how can the EPLL heuristic be generalized to capture more complex statistics of natural images? This paper answers these questions by linking EPLL to a rigorous, nonparametric generative model of whole images.

3. Mixture Models for Grids of Image Patches

We now develop the HDP-Grid generative model summarized in Fig. 1, which uses randomly placed patch grids to formalize the EPLL objective, and hierarchical DP mixtures to capture image patch self-similarity.

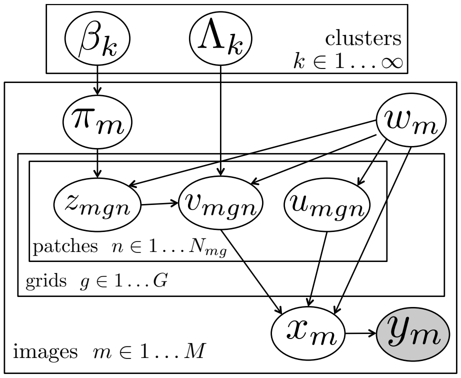


Figure 1. Directed graphical model for our HDP-Grid model of M natural images. Clean image x_m is generated via a randomly placed grid w_m of patches v_m generated by a hierarchical Gaussian mixture model. We observe corrupted images y_m.

3.1. Hierarchical Dirichlet Process Mixtures

The *hierarchical Dirichlet process* (HDP, Teh et al. (2006)) is a Bayesian nonparametric prior used to cluster groups of related data; we model natural images as groups of patches. The HDP shares visual structure, such as patches of grass or bricks, by sharing a common set of clusters (called *topics* in applications to text data) across images. In addition, the HDP models image-specific variability by allowing each image to use this shared set of clusters with unique frequencies; grass might be abundant in one image but absent in another. Via the HDP, we can learn the proper number of hidden clusters from data, and discover new clusters as we collect new images with novel visual textures.

The HDP uses a stick-breaking construction to generate a corpus-wide vector $\pi_0 = [\pi_{01}, \pi_{02}, \ldots, \pi_{0k}, \ldots]$ of frequencies for a countably infinite set of visual clusters:

$$\beta_k \sim \operatorname{Beta}(1, \gamma), \quad \pi_{0k}(\beta) \triangleq \beta_k \textstyle\prod_{\ell=1}^{k-1}(1-\beta_\ell). \qquad (8)$$

The HDP allocates each image m its own cluster frequencies π_m, where the vector π_0 determines the mean of a DP prior on the frequencies of shared clusters:

$$\pi_m \sim \operatorname{DP}(\alpha\pi_0), \qquad \mathbb{E}[\pi_{mk}] = \pi_{0k}. \qquad (9)$$

When the concentration parameter $\alpha < 1$, we capture the “burstiness” and self-similarity of natural image regions (Jégou et al., 2009) by placing most probability mass in π_m on a sparse subset of global clusters.

3.2. Image Generation via Random Grids

We sample pixels in image m via a randomly placed grid of patches. When each patch has G pixels, Fig. 2 shows there are exactly G grid alignments for an image of arbitrary size. The alignment $w_m \in \{1, \ldots, G\}$ has a uniform prior:

$$w_m \sim \operatorname{Cat}(1/G, \ldots, 1/G). \qquad (10)$$

Modeling multiple overlapping grids is crucial to capture real image statistics. As the true grid alignment for each image is uncertain, posterior inference will favor images

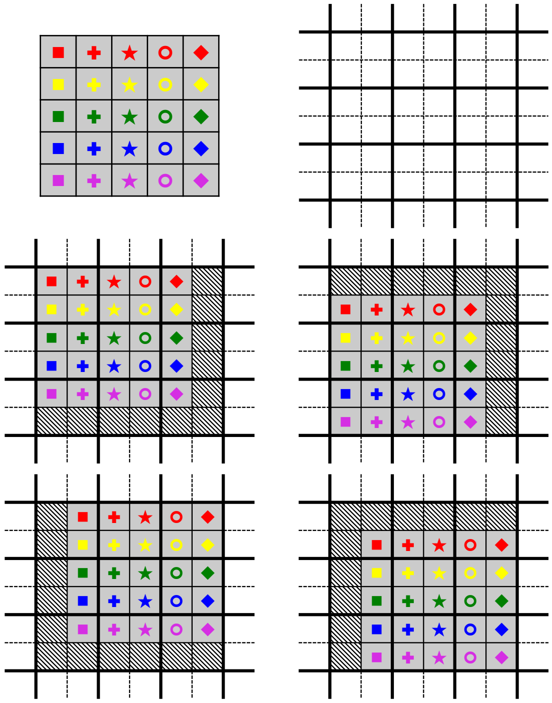


Figure 2. Generation of a complete image via a randomly positioned grid of non-overlapping patches. *Top left:* A 5 × 5 pixel image, where each pixel is identified by a distinct colored symbol. *Top right:* An infinite 2D grid of pixels, divided into 2 × 2 patches. *Bottom:* The four possible ways a 5 × 5 image may be generated from 2 × 2 patches. Shaded pixels are clipped by the image boundary (see Sec. 3.4).

that are likely under *all* possible w_m. Models based on a single, fixed grid produce severe artifacts at patch boundaries, as shown in Fig. 2 of Zoran & Weiss (2011).

3.3. Patch Generation via Gaussian Mixtures

Gaussian mixtures provide excellent density models for natural image patches (Zoran & Weiss, 2012). We associate clusters with zero-mean, full-covariance Gaussian distributions on patches with G pixels. We parameterize cluster k by a precision (inverse covariance) matrix $\Lambda_k \sim \text{Wish}(\nu, W)$, whose conjugate Wishart prior has ν degrees of freedom and scale matrix W. Given that $w_m = g$, each of the N_{mg} patches v_{mgn} in grid g is sampled from an infinite mixture with image-specific cluster frequencies:

$$p(v_{mgn}|w_m = g) = \sum_{k=1}^{\infty} \pi_{mk} \text{Norm}(v_{mgn}|0, \Lambda_k^{-1}). \quad (11)$$

Let $z_{mgn} \mid w_m = g \sim \text{Cat}(\pi_m)$ denote the cluster that generates patch n. To account for the contrast normalization of Eq. (1), the intensities in patch n are shifted by an independent, scalar "DC offset" u_{mgn}:

$$p(u_{mgn} \mid w_m = g) = \text{Norm}(u_{mgn} \mid r, s^2). \quad (12)$$

Finally, if $w_m \neq g$ so that grid g is unobserved, we sample $(z_{mgn}, v_{mgn}, u_{mgn})$ from some reference distribution independent of the HDP mixture model parameters.

3.4. From Patches to Corrupted Images

Given patches v_{mg} with offsets u_{mg} generated via grid $w_m = g$, we sample a whole "clean image" x_m as

$$\text{Norm}\Big(x_m \mid \textstyle\sum_{n=1}^{N_{mg}} P_{mgn}^T \bar{v}_{mgn}, \delta^2 I\Big), \quad (13)$$

where $\bar{v}_{mgn} \triangleq C_{mgn} v_{mgn} + u_{mgn}$. Binary indicator matrices P_{mgn}, as in Sec. 2, stitch together patches in the chosen grid g. Image x_m is then generated by adding independent Gaussian noise with small variance δ^2. Most patches in the chosen grid will be fully observed in x_m, but as illustrated in Fig. 2, some may be clipped by the image boundary. Indicator matrices C_{mgn} are defined so $C_{mgn} v_{mgn} + u_{mgn}$ is a vector containing the observed pixels from patch n.

For image restoration tasks, the observed image y_m is a corrupted version of some clean image x_m that we would like to estimate. Models of natural image statistics are commonly validated on the problem of image denoising, where x_m is polluted by additive white Gaussian noise:

$$p(y_m \mid x_m) = \text{Norm}(y_m \mid x_m, \sigma^2 I). \quad (14)$$

The variance $\sigma^2 \gg \delta^2$ indicates the noise level. We also validate our model on image inpainting problems (Bertalmio et al., 2000), where some pixels are observed without noise but others are completely missing. By replacing Eq. (14) with other linear likelihood models, our novel generative model for natural images may be easily applied to other tasks including image deblurring (Zoran & Weiss, 2011), image super resolution (Yang & Huang, 2010), and color image demosaicing (Mairal et al., 2009).

4. Variational Inference

We now develop scalable learning algorithms for our nonparametric, grid-based image model. We first examine a baseline *DP Grid* model in which the same cluster frequencies π_0 are shared by all images. Our full *HDP Grid* model then learns image-specific cluster frequencies π_m, and instantiates new clusters to model unique visual textures.

4.1. DP Grid: Variational Inference

Our goal is to infer the DP Grid model parameters that best explain observed images which may be clean (x_m) or corrupted by noise (y_m). The DP Grid model uses the same cluster probabilities π_0, generated from stick-breaking weights β as in Eq. (8), for all images.

Learning from clean images. Given a training set $\mathcal{D}$ of *uncorrupted* images $x_1, \dots x_M$, we estimate the posterior distribution $p(\beta, \Lambda, w, \Psi^{\text{patch}} \mid x)$ for our global mixture model parameters β and Λ, grid assignment indicators w_m, and patch-level latent variables $\Psi_m^{\text{patch}} = \{u_m, v_m, z_m\}$.

Exact posterior inference is intractable, so we instead find an approximate posterior $q(\cdot) = q(\beta, \Lambda, w, \Psi^{\text{patch}})$ minimizing the KL divergence (Wainwright & Jordan, 2008) from the true posterior $p(\cdot|x)$. Equivalently, our variational method maximizes the following objective $\mathcal{L}$:

$$\max_{q \in \mathcal{Q}} \mathcal{L}(q, x) = \max_{q \in \mathcal{Q}} \mathbb{E}_q \left[\log \frac{p(x, \cdot)}{q(\cdot)} \right] \leq \log p(x). \quad (15)$$

We constrain the solution of our optimization to come from a tractable family of *structured* mean-field distributions $\mathcal{Q}$, parameterized by free parameters. Unlike naïve mean-field methods which assume complete posterior independence, our structured mean-field approximation is more accurate and includes dependencies between some latent variables:

$$q(\cdot) = \prod_{k=1}^{\infty} q(\Lambda_k) q(\beta_k) \cdot \prod_{m=1}^{M} q(w_m) q(\Psi_m^{\text{patch}} | w_m).$$

As in Hughes & Sudderth (2013), this approximate posterior family contains *infinitely* many clusters, just like the true posterior. Rather than applying a fixed truncation to the stick-breaking prior (Blei & Jordan, 2006), we dynamically truncate the patch assignment distributions $q(z)$ to only use the first K clusters to explain the M observed images. Clusters with indices $k > K$ then have factors $q(\Lambda_k)$ set to the prior, and need not be explicitly represented.

Global mixture model. The global cluster weights β and precision matrices Λ have standard exponential family forms (free parameters are marked by hats):

$$q(\Lambda_k) = \text{Wish}\big(\hat{\nu}_k, \hat{W}_k\big), q(\beta_k) = \text{Beta}\big(\hat{\rho}_k \hat{\omega}_k, (1 - \hat{\rho}_k)\hat{\omega}_k\big).$$

Here $\hat{\rho}_k = \mathbb{E}_q[\beta_k]$, and $\hat{\omega}_k$ controls the variance of $q(\beta_k)$.

Image-specific alignment. For natural images, all grid alignments are typically of similar quality, so we fix a *uniform* alignment posterior $q(w_m) = \text{Cat}\big(\frac{1}{G}, \ldots, \frac{1}{G}\big)$. This simplifies many updates while still avoiding artifacts that would arise from a single, non-overlapping patch grid.

Patch-specific factors. The patch-specific variables Ψ^{patch} have *structured* posteriors, conditioned on the value of the grid indicator w_m for the current image:

$$q(z_{mgn} \mid w_m = g) = \text{Categorical}\big(\hat{r}_{mgn1}, \ldots, \hat{r}_{mgnK}\big),$$
$$q(u_{mgn} \mid w_m = g) = \text{Norm}\big(\hat{u}_{mgn}, \hat{\phi}^u_{mgn}\big),$$
$$q(v_{mgn} \mid w_m = g, z_{mgn} = k) = \text{Norm}\big(\hat{v}_{mgnk}, \hat{\phi}^v_{mgnk}\big).$$

Below, we let $\mathbb{E}_q[\cdot]$ denote the *conditional* expectation with respect to the variational distribution q, given w_m.

Learning. Given clean images x, we perform coodinate ascent on the objective $\mathcal{L}$, alternatively updating one factor among $q(\beta)q(\Lambda)q(w)q(\Psi^{\text{patch}})$. Most updates have closed forms due to the exponential families defining $\mathcal{Q}$ (see supplement). As one intuitive example, consider the update for the cluster precision matrix posterior $q(\Lambda_k | \hat{\nu}_k, \hat{W}_k)$:

$$\hat{\nu}_k = \nu + \frac{1}{G} N_k, \quad N_k = \sum_{m=1}^{M} \sum_{g=1}^{G} \sum_{n=1}^{N_{mg}} \hat{r}_{mgnk}, \quad (16)$$

$$\hat{W}_k = W + \frac{1}{G} \underbrace{\sum_{m=1}^{M} \sum_{g=1}^{G} \sum_{n=1}^{N_{mg}} \mathbb{E}_q \left[\mathbb{1}_k(z_{mgn}) v_{mgn} v_{mgn}^T \right]}_{S_k}.$$

Statistic $N_k(\hat{r})$ counts patches assigned to cluster k, while $S_k(\hat{r}, \hat{v}, \hat{\phi}^v)$ aggregates second moments. These updates follow the standard form of prior parameter plus expected sufficient statistic, except the statistics are averaged (*not* simply added) across the G grid alignments.

4.2. Image Denoising and Connections to EPLL

Given a corrupted image y_m, we seek to compute the posterior $p(x_m \mid y_m, \mathcal{D})$, where we condition on the training set $\mathcal{D}$. Our variational posterior family Q now includes an additional factor for the unobserved, "clean" image x_m:

$$q(x_m) = \text{Norm}\big(x_m \mid \hat{x}_m, \hat{\phi}^x_m\big). \quad (17)$$

The variational inference objective becomes

$$\max_{q \in \mathcal{Q}} \mathbb{E}_q \left[\log \frac{p(\mathcal{D}, y_m, x_m, \cdot)}{q(x_m, \cdot)} \right] \leq \log p(y_m, \mathcal{D}), \quad (18)$$

and the coordinate ascent update for $q(x_m)$ equals

$$\hat{x}_m = \hat{\phi}^x_m \left(\frac{y_m}{\sigma^2} + \frac{h_m}{\delta^2} \right), \quad \hat{\phi}^x_m = \frac{\delta^2 \sigma^2}{\delta^2 + \sigma^2} I. \quad (19)$$

The updated covariance is diagonal, improving computational efficiency. The mean depends on the average image vector across all patches in all grids, denoted by h_m:

$$h_m \triangleq \frac{1}{G} \sum_{g=1}^{G} \sum_{n=1}^{N_{mg}} P^T_{mgn} (C_{mgn} \mathbb{E}_q[v_{mgn}] + \hat{u}_{mgn}). \quad (20)$$

Note that the update for $\hat{x}_m$ in Eq. (19) is similar to the EPLL update in Eq. (7), except that some terms involving projection matrices become constants because we account for partially observed patches. Modeling partial patches is necessary to produce a valid likelihood bound in Eq. (18).

In fact, as we show below all three terms in the EPLL objective in Eq. (4) are very similar to our proposed minimization objective function $-\mathcal{L}$, up to a scale factor of G. Of course, a key difference is that our objective seeks full posteriors rather than point estimates, and enables the HDP model of multiple images detailed in Sec. 4.3.

EPLL Term 1. When we set $\lambda \triangleq \frac{G}{\sigma^2}$, the first term of the EPLL objective in Eq. (4) becomes

$$G \cdot \tfrac{1}{2\sigma^2} (x - y)^T (x - y). \quad (21)$$

Similarly, suppressing the subscript m denoting the image for simplicity, $\mathbb{E}_q[-\log p(y|x)]$ in our $-\mathcal{L}$ simplifies as

$$\tfrac{1}{2\sigma^2} \mathbb{E}_q[(x - y)^T (x - y)]. \quad (22)$$

EPLL Term 2. Taking the second term in Eq. (4) and substituting $\kappa = 1/\delta^2$, we have:

$$\frac{1}{2\delta^2}\sum_i (P_i x - \bar{v}_i)^T (P_i x - \bar{v}_i). \tag{23}$$

The corresponding term $\mathbb{E}_q[-\log p(x|w,u,v)]$ in our objective $-\mathcal{L}$ can be written similarly up to a scaling by G:

$$\frac{1}{G}\frac{1}{2\delta^2}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\mathbb{E}_q\Big[(P_{gn}x-\bar{v}_{gn})^T(P_{gn}x-\bar{v}_{gn})\Big]. \tag{24}$$

EPLL Term 3. The third EPLL term assumes zero-centered patches $B\bar{v}_i$ are drawn from Gaussian mixtures:

$$-\sum_i \log p(B\bar{v}_i \mid \pi_0, \Lambda). \tag{25}$$

Similarly, in our minimization objective $-\mathcal{L}$ we draw v_{gn} from a DP mixture model. Explicitly including the cluster assignment z_{gn}, $\mathbb{E}_q[-\log p(v,z|w)]$ equals

$$-\frac{1}{G}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\mathbb{E}_q[\log p(v_{gn}, z_{gn} \mid \pi_0, \Lambda)]. \tag{26}$$

EPLL is similar, but maximizes assignments (Eq. (5)) rather than computing posterior assignment probabilities.

4.3. HDP Grid: Variational Inference

Image-specific frequencies. The DP model above, and the parametric EPLL objective it generalizes, assume the same cluster frequency vector π_0 for each image m. Our HDP Grid model allows image-specific frequencies π_m to be learned from data, via the hierarchical regularization of the HDP prior (Teh et al., 2006). Our approximate posterior family $\mathcal{Q}$ now has the following HDP-specific factors:

$$q(\beta) = \prod_{k=1}^{\infty} \mathrm{Beta}\left(\beta_k \mid \hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k\right), \tag{27}$$

$$q([\pi_{m1}\ldots\pi_{mK}\ \pi_{m>K}]) = \mathrm{Dir}(\hat{\theta}_{m1}\ldots\hat{\theta}_{mK}, \hat{\theta}_{m>K}).$$

This approximate posterior represents infinitely many clusters via a finite partition of π_m into $K+1$ terms: one for each of the K active clusters, and a remainder term at index $>K$ that aggregates the mass of all inactive clusters. The free parameter $\hat{\theta}_m$ is also a vector of size $K+1$ whose last entry represents all inactive clusters. We follow Hughes et al. (2015) to obtain a closed-form update for $\hat{\theta}_m$, and gradient-based updates for $\hat{\rho}, \hat{\omega}$; see the supplement for details. We highlight that the $\hat{\theta}_m$ update naturally includes a $\frac{1}{G}$ rescaling of count sufficient statistics as in Eq. (16). Other factors remain unchanged from the DP Grid model.

Image-specific clusters. Due to the heavy-tailed distribution of natural images (Ruderman, 1997), even with large training sets, test images may still contain unique textural patterns like the striped scarf in the Barbara image in Fig. 3. Fortunately, our Bayesian nonparametric HDP Grid model provides a coherent way to capture such patterns by appending K' novel, image-specific clusters to the original K clusters learned from training images. These novel clusters lead to more accurate posterior approximations $q \in \mathcal{Q}$ that better optimize our objective $\mathcal{L}$.

We initialize inference by creating $K' = 100$ image-specific clusters with the `k-means++` algorithm (Arthur & Vassilvitskii, 2007), which minimizes the cost function

$$\mathcal{J}(z', \Lambda') = \sum_i \sum_{k=1}^{K'} \mathbb{1}_k(z'_i)\mathrm{D}(\tilde{v}_i\tilde{v}_i^T, \Lambda'_k), \tag{28}$$

where the first sum is over the set of fully-observed patches within the image. The function D is the Bregman divergence associated with our zero-mean Gaussian likelihood (Banerjee et al., 2005), and $\tilde{v}_i = BP_i y$ is a zero-centered patch. We initialize the algorithm by sampling K' diverse patches in a distance-biased fashion, and refine with 50 iterations of coordinate descent updates of z' and Λ'.

Then we expand the variational posterior $q(\Lambda)$ into $K + K'$ clusters. The first K indices are kept the same as training, and the last K' indices are set via Eq. (16) using sufficient statistics N', S' derived from hard assignments z':

$$N'_{k'} \leftarrow \sum_i \mathbb{1}_{k'}(z'_i),\ S'_{k'} \leftarrow \left[\sum_i \mathbb{1}_{k'}(z'_i)\tilde{v}_i\tilde{v}_i^T - N_{k'}\sigma^2 I\right]_+.$$

Here, following Portilla et al. (2003) and Kivinen et al. (2007), $S'_{k'}$ estimates the *clean* data statistic $S_{k'}$ by subtracting the expected noise covariance. The $[\cdot]_+$ operator thresholds any negative eigenvalues to zero.

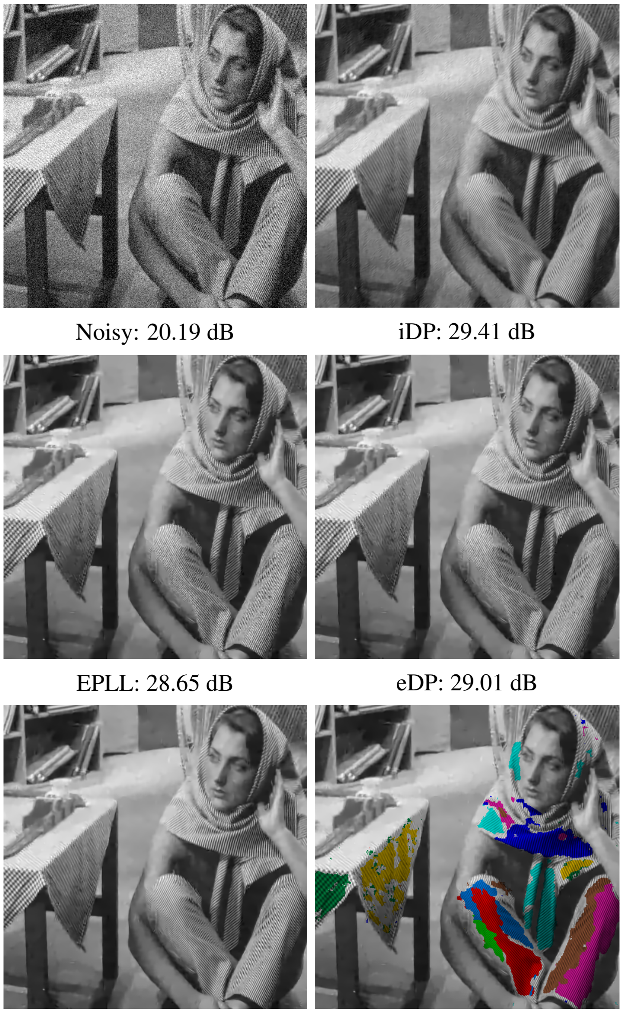
Similarly, the other global variational factor $q(\beta)$ is also expanded to $K + K'$ clusters via sufficient statistics N' and counts of cluster usage from training data. Given $\{\beta, \Lambda\}_{k=1}^{K+K'}$, each factor in q may then be updated in turn to maximize the variational objective $\mathcal{L}$ (see supplement).

Finally, while we initialize K' to a large number to avoid local optima, this may lead to extraneous clusters. We thus delete new clusters that our sparsity-biased variational updates do not assign to any patch. In the Barbara image in Fig. 3, this leaves 9 image-specific clusters. Deletion improves model interpretability and algorithm speed, because costs scale linearly with the number of instantiated clusters.

5. Experiments

Following EPLL, we train our HDP-Grid model using 400 clean training and validation images from the Berkeley segmentation dataset (BSDS, Martin et al. (2001)). We fix $\delta = 0.5/255$ to account for the quantization of image intensities to 8-bit integers. Observed DC offsets u provide maximum likelihood estimates of the mean r and variance s^2 in Eq. (12). Similarly, we compute empirical covariance matrices for patches in the same image segments to estimate hyperparameters W and ν in Eq. (16). Using variational learning algorithms that adapt the number of clusters to the observed data (Hughes & Sudderth, 2013), we discover $K = 449$ clusters for the DP-Grid model, which we use to initialize our HDP model. We set our annealing schedule for κ to match that used by the public EPLL code.

Image denoising methods are often divided into two types (Zontak & Irani, 2011): *external* methods (like



Noisy: 20.19 dB

iDP: 29.41 dB

EPLL: 28.65 dB

eDP: 29.01 dB

HDP: **30.15** dB

HDP: new clusters

Figure 3. For an image with noise level $\sigma = 25$, the HDP improves denoising performance by leveraging both internal clusters (e.g., scarf and tablecloth) and external clusters (e.g., floor and table legs). The bottom right image colors the pixels assigned to each of 9 internal HDP clusters. **Best viewed electronically.**

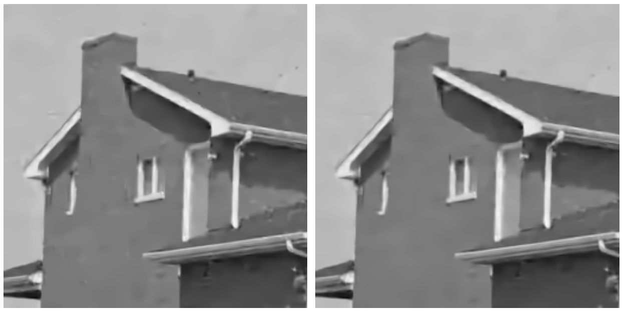
EPLL) that learn all parameters from a training database of clean images, and *internal* methods that denoise patches using other patches of the single noisy image. For example, the K-SVD (Elad & Aharon, 2006) has an external variant that uses a dictionary learned from clean images, and an internal variant that learns its dictionary from the noisy image. A major contribution of our paper is to show that the hierarchical DP leads to a principled hybrid of internal and external methods, in which cues from clean and noisy images are automatically combined in an adaptive way.

5.1. Image Denoising

We test our algorithm on 12 "classic" images used in many previous denoising papers (Mairal et al., 2009; Zoran & Weiss, 2011), as well as the 68 BSDS test images used by (Roth & Black, 2005; Zoran & Weiss, 2011). We evaluate the denoising performance by the *peak signal-to-noise ratio* (PSNR), a logarithmic transform of the *mean squared error* (MSE) between images with normalized intensities,

$$\text{PSNR} \triangleq -20 \log_{10} \text{MSE}. \tag{29}$$

We also evaluate the *structural similarity index* (SSIM, Wang et al. (2004)), which quantifies image quality degradation via changes in structure, luminance, and contrast.



eDP: 32.47 dB

HDP: **32.65** dB

Figure 4. By capturing self-similar patches in the "house" image, our HDP model reduces artifacts in smooth regions such as the sky, roof, and walls. Input noise level $\sigma = 25$ (20.21 dB).

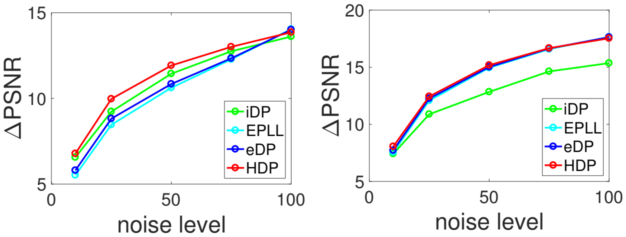


Figure 5. Denoising performance of grid-based models on the Barbara image of Fig. 3 (left) and the house image of Fig. 4 (right), as a function of the noise standard deviation. For both images and all noise levels, the HDP model is superior to baselines that solely use external (eDP) or internal (iDP) training, in terms of PSNR improvement relative to the noisy input image. When the image is extremely noisy ($\sigma = 100$), internal clusters are of poor quality, and the HDP and eDP models are comparable.

Internal vs. external clusters. In result figures, we use *eDP* to refer to our DP-Grid model trained solely on external clean images and *HDP* to refer to the HDP-Grid model that also learns novel image-specific clusters. We also train an internal DP-Grid model, referred to as *iDP*, using only information from the noisy test image. The first four columns of Table 1 compare their average denoising performance, where EPLL can be viewed as a simplification of eDP. For all noise levels and datasets, the HDP model has superior performance. As shown in Fig. 6, HDP is more accurate than EPLL and eDP for every single classic-12 image. Also, the consistent gain in performance from EPLL to eDP demonstrates the benefits of Bayesian nonparametric learning of an appropriate model complexity (for EPLL, the number of clusters was arbitrarily fixed at $K = 200$).

Fig. 3 further illustrates the complementary role of internal

Table 1. Average PSNR and SSIM values on benchmark datasets (larger values indicate better denoising). Methods are highlighted if they are indistinguishable with 95% confidence, according to a Wilcoxon signed-rank test on the fraction of images where one method outperforms another. For all noise levels the patch size of BM3D is fixed to 8 × 8 and LSSC is fixed to 9 × 9.

metric	dataset	σ	iDP	EPLL	eDP	HDP	FoE	eKSVD	iKSVD	BM3D	LSSC
PSNR	classic-12	10	33.66	33.68	33.77	**33.99**	33.11	33.45	33.62	33.98	**34.05**
		25	29.02	29.39	29.47	**29.68**	28.32	28.89	29.11	**29.73**	**29.74**
		50	25.44	26.22	**26.28**	**26.42**	24.69	25.44	25.64	**26.55**	**26.43**
	BSDS-68	10	33.10	33.37	33.42	**33.47**	32.69	33.06	33.08	33.26	33.45
		25	28.33	28.72	28.76	**28.82**	27.76	28.28	28.28	28.55	28.70
		50	25.10	25.72	25.75	**25.83**	24.48	25.17	25.17	25.59	25.50
SSIM	classic-12	10	0.9118	0.9136	**0.9143**	**0.9169**	0.8962	0.9084	0.9111	0.9168	**0.9185**
		25	0.8189	0.8286	0.8299	**0.8337**	0.8018	0.8082	0.8131	**0.8357**	**0.8359**
		50	0.6962	**0.7301**	**0.7316**	**0.7366**	0.6885	0.6926	0.6975	**0.7425**	**0.7390**
	BSDS-68	10	0.9119	0.9219	0.9224	**0.9230**	0.8971	0.9128	0.9135	0.9157	0.9206
		25	0.7964	0.8090	0.8103	**0.8131**	0.7804	0.7859	0.7879	0.8010	0.8109
		50	0.6636	0.6870	0.6880	**0.6962**	0.6585	0.6544	0.6539	0.6840	0.6885

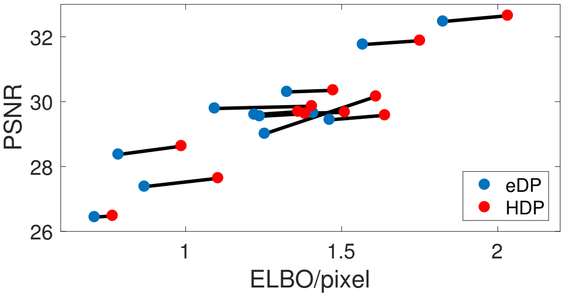


Figure 6. Clean-image evidence lower bound (ELBO) versus output PSNR ($\sigma = 25$) for 12 “classic” images. The horizontal axis plots $\log p(x_{\text{test}}|x_{\text{train}}) \approx \mathcal{L}(x_{\text{test}}, x_{\text{train}}) - \mathcal{L}(x_{\text{train}})$, divided by the number of pixels. Our HDP is uniformly superior to the eDP.

and external clusters for a single test image (“Barbara”). The *internal* iDP perfectly captures some unique textures like the striped clothing, but produces artifacts in smooth background regions. The *external* EPLL and eDP better represent smooth surfaces and contours, which are common in training data, but poorly recover striped textures.

As shown in Fig. 5, while the relative accuracy of the eDP and iDP models varies depending on image statistics, the HDP model adaptively combines external and internal clusters for superior performance at all noise levels. By capturing the expected self-similarity of image patches, the HDP model also reduces artifacts in large regions with regular textures, such as the smoothly shaded areas of Fig. 4.

Computational speed. To denoise a 512 × 512 pixel image on a modern laptop, our Python code for eDP inference with $K = 449$ clusters takes about 12 min. The public EPLL Matlab code (Zoran & Weiss, 2011) with $K = 200$ clusters takes about 5 min. With equal numbers of clusters, the two methods have comparable runtimes. Our open-source Python code is available online at github.com/bnpy/hdp-grid-image-restoration.

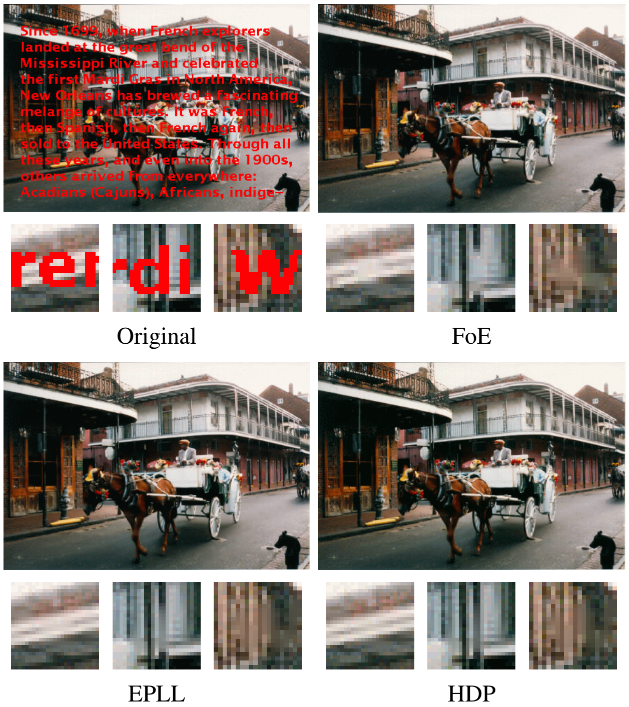

Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-
rer
di
w
Original
FoE
EPLL
HDP

Figure 7. A qualitative comparison of image inpainting algorithms. As illustrated in the three close-up views, the HDP exploits patch self-similarity to better recover fine details.

Learning image-specific clusters for the HDP model is more expensive: our non-optimized Python denoising code currently requires about 30 min. per image. Nearly all of the extra time is spent on the k-means++ initialization of Eq. (28). We expect this can be sped up significantly by coding core routines in C, parallelizing some sub-steps (possibly via GPUs), using fewer internal clusters (100 is often too many), or using faster initialization heuristics.

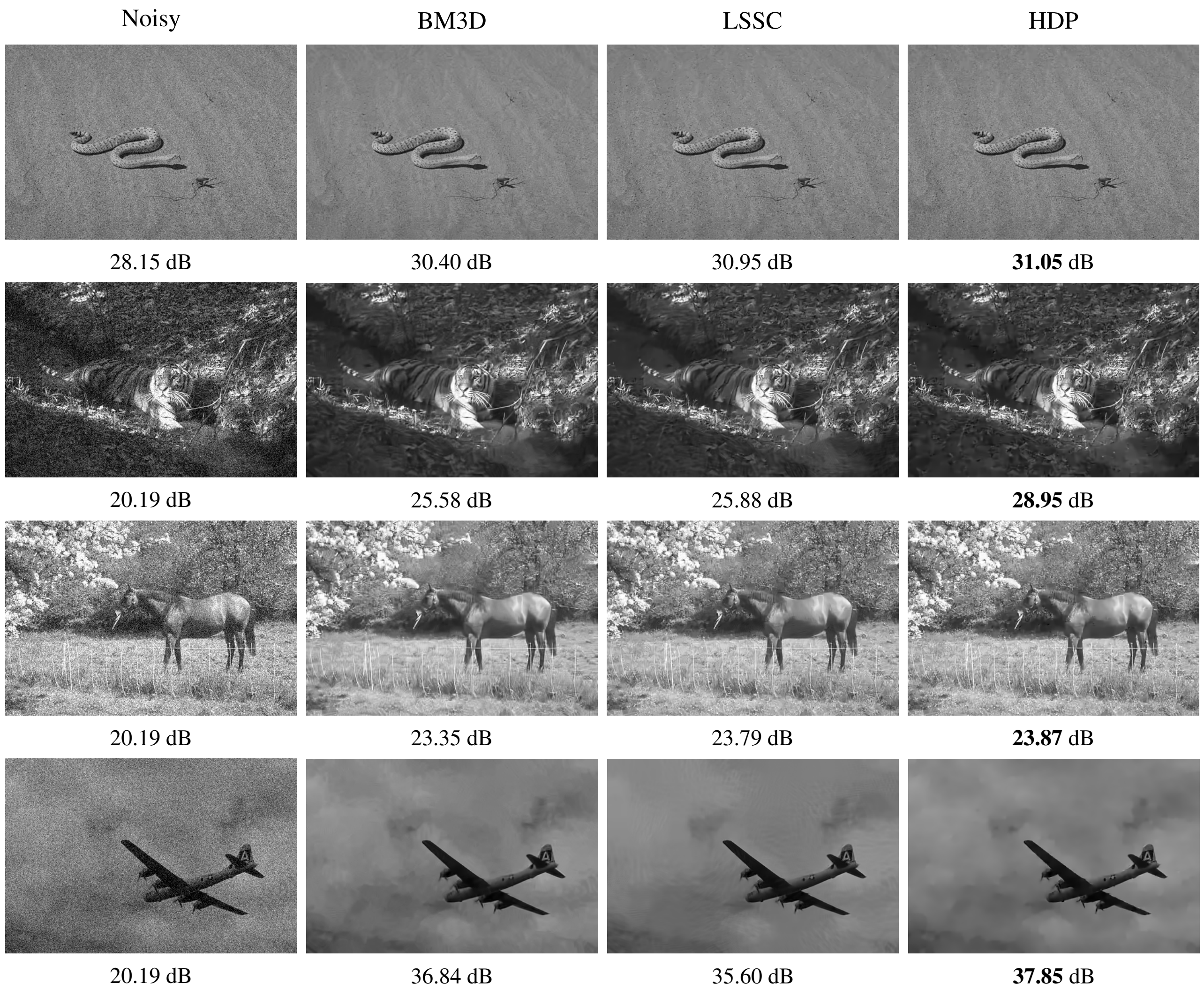


Figure 8. Comparison of image denoising methods on BSDS-68. Unlike our HDP model, the BM3D and LSSC methods learn solely from the noisy image and do not accurately capture some textures such as the sandy ground in *Row 1*, fallen leaves and tiger tail in *Row 2*, trees and grass in *Row 3*, and sky and clouds in *Row 4*. Noise level $\sigma = 10$ in *Row 1*, $\sigma = 25$ elsewhere. **Best viewed electronically.**

Performance. We compare our HDP model to other patch-based denoising methods in Table 1. On classic-12, where many top methods have been hand-tuned to perform well, our model is statistically indistinguishable from the best baselines. On the larger BSDS-68, our performance is superior to the state-of-the-art, showing the value of nonparametric learning from large image collections. See Fig. 8 for examples. At higher noise levels ($\sigma = 50$), LSSC has modestly improved performance (0.2 dB in PSNR) when modeling 12×12 patches (Mairal et al., 2009). HDP models of larger patches are a promising research area.

5.2. Image Inpainting

While many image processing systems are designed for just one problem, our generative model is useful for many tasks. For example, we can "inpaint" occluded image regions (like the red pixels in Fig. 7) by modifying Eq. (14) to let $\sigma^2 \to \infty$ for only those regions and setting $\sigma^2 = 0$ elsewhere. To process color images, we follow the approach of FoE and EPLL and convert to the YCbCr color space before independently inpainting each channel. While ground truth is unavailable for the classic image in Fig. 7, our grid-based HDP produces fewer visual artifacts than baselines.

6. Conclusion

We have developed a coherent Bayesian nonparametric model that, via randomly positioned grids of image patches, provides a novel statistical foundation for the popular EPLL method. We show that HDP mixture models of visual textures can grow in complexity as additional images are observed and capture the self-similarity of natural images. Our HDP-grid image denoising and inpainting algorithms are competitive with the state-of-the-art, and our model is applicable to many other computer vision tasks.

Acknowledgements

This research supported in part by NSF CAREER Award No. IIS-1349774. MCH supported in part by Oracle Labs.

References

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. Image inpainting. In *Computer Graphics and Interactive Techniques*, pp. 417–424, 2000.

Blei, D. M. and Jordan, M. I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1): 121–143, 2006.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Image restoration by sparse 3d transform-domain collaborative filtering. In *Electronic Imaging*, 2008.

Elad, M. and Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

Geman, D. and Yang, C. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.

Hughes, M. C. and Sudderth, E. B. Memoized online variational inference for Dirichlet process mixture models. In *Neural Information Processing Systems*, 2013.

Hughes, M. C., Kim, D. I., and Sudderth, E. B. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2015.

Jégou, H., Douze, M., and Schmid, C. On the burstiness of visual elements. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1169–1176, 2009.

Kivinen, J. J., Sudderth, E. B., and Jordan, M. I. Image denoising with nonparametric hidden Markov trees. In *International Conference on Image Processing*, 2007.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. Non-local sparse models for image restoration. In *International Conference on Computer Vision*, 2009.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference on Computer Vision*, 2001.

Portilla, J., Strela, V., Wainwright, M. J., and Simoncelli, E. P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351, 2003.

Roth, S. and Black, M. J. Fields of experts: A framework for learning image priors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pp. 860–867, 2005.

Ruderman, D. L. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997.

Srivastava, A., Lee, A. B., Simoncelli, E. P., and Zhu, S. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Yang, J. and Huang, T. Image super-resolution: Historical overview and future challenges. *Super-resolution imaging*, pp. 20–34, 2010.

Zontak, M. and Irani, M. Internal statistics of a single natural image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 977–984, 2011.

Zoran, D. and Weiss, Y. From learning models of natural image patches to whole image restoration. In *International Conference on Computer Vision*, 2011.

Zoran, D. and Weiss, Y. Natural images, Gaussian mixtures and dead leaves. In *Neural Information Processing Systems*, 2012.