A Generative Perspective on MRFs in Low-Level Vision

Uwe Schmidt* Qi Gao* Stefan Roth
Department of Computer Science, TU Darmstadt

Abstract

Markov random fields (MRFs) are popular and generic probabilistic models of prior knowledge in low-level vision. Yet their generative properties are rarely examined, while application-specific models and non-probabilistic learning are gaining increased attention. In this paper we revisit the generative aspects of MRFs, and analyze the quality of common image priors in a fully application-neutral setting. Enabled by a general class of MRFs with flexible potentials and an efficient Gibbs sampler, we find that common models do not capture the statistics of natural images well. We show how to remedy this by exploiting the efficient sampler for learning better generative MRFs based on flexible potentials. We perform image restoration with these models by computing the Bayesian minimum mean squared error estimate (MMSE) using sampling. This addresses a number of shortcomings that have limited generative MRFs so far, and leads to substantially improved performance over maximum a-posteriori (MAP) estimation. We demonstrate that combining our learned generative models with sampling-based MMSE estimation yields excellent application results that can compete with recent discriminative methods.

1. Introduction and Related Work

Markov random fields (MRFs) provide a sound probabilistic framework for modeling and integrating prior knowledge of images and scenes, and have found widespread use across low-level vision, *e.g.*, in image restoration [8, 20, 31], super-resolution [25], stereo [3], optical flow [14], *etc*. While the study of natural image and scene statistics [23] provides key motivations for the use of MRFs as well as particular modeling choices, such as the shape of the potential functions, it is rarely evaluated how well these statistical properties are captured by MRF models. More than 10 years ago, Zhu and Mumford [31] advocated the use of sampling for evaluating their image prior and computed derivative histograms from model samples. Ever since, the generative properties of MRFs have been largely ignored. Instead, model evaluation usually happens in the context of a particular application, *e.g.*, image denoising in case of image priors [20]. This stems from the fact that computing marginal distributions and other probabilistic properties of MRFs is difficult. Sampling is largely the only choice, but widely applicable algorithms, such as standard Gibbs sampling [31], are cumbersome and inefficient.

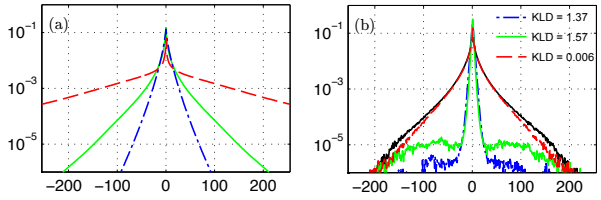


Figure 1. **Pairwise MRF potentials and derivative marginals:** *(a)* Generalized Laplacian [25] (blue, dash-dotted), fit of the marginals [22] (green, solid), and proposed flexible potential (red, dashed). *(b)* Derivative histogram of samples from corresponding MRFs, and statistics of natural images (black, solid). Typical models [22, 25] lead to an incorrect representation of image statistics.

Recent years have seen a trend to move away from a strict probabilistic interpretation of MRFs in low-level vision. On the one hand, this is due to the prevalence of maximum a-posteriori (MAP) estimation (*e.g.*, [3]), which is not affected by an arbitrary rescaling of the model's energy. On the other hand, even though probabilistically trained generative models of low-level vision [20, 27, 31] have been successful, they have frequently required ad-hoc modifications to perform well in practice. Moreover, due to the difficulties of learning generative models, non-probabilistic discriminative methods have gained momentum, including max-margin [24] and loss-function based training [1, 14, 21]. While the performance of such application-specific models is highly desirable, they lack the statistical interpretability and versatility of generative MRFs.

In this paper we revisit and explore the *generative, probabilistic* aspects of MRFs in low-level vision, and demonstrate that a rigorous probabilistic interpretation and good generative properties can go hand-in-hand with excellent application performance. Even though many of our discussions remain general, we focus on image priors and image restoration as concrete examples. We rely on a general class of MRFs proposed by [27], whose clique potentials use Gaussian scale mixtures (GSMs) [19] to model the responses to a bank of linear filters. This model encompasses

*The first two authors contributed equally to this work.

common pairwise MRFs [11, 13, 25] and Fields of Experts (FoEs) [20], high-order models with large cliques. We take advantage of the fact that MRFs with GSM potentials can be efficiently sampled using an auxiliary-variable Gibbs sampler [7, 12, 26]. With this toolset we analyze the generative properties of image priors using sampling. This includes marginals of *model features, multi-scale derivatives, and random filters*. As a quantitative measure, we propose the *marginal KL-divergence* of the respective marginals. In contrast to [31], we use a more efficient sampler, and analyze popular recent pairwise and high-order MRFs with regards to their probabilistic properties. Surprisingly, we find that these models are quite poor generative models (*e.g.*, Fig. 1), which apparently contradicts their good application performance.

To understand this, we go beyond previous work and exploit the auxiliary-variable Gibbs sampler to learn pairwise and high-order MRFs using contrastive divergence [9], and analyze their generative properties. In contrast to typical MRF models with simple parametric potentials, we rely on more flexible GSMs that admit a wide range of possible shapes. Unlike [27], we do not fit the GSM potentials to the marginals ahead of time, but let the learning algorithm determine their shape. For both pairwise MRFs and FoEs, we find *heavier-tailed potentials* than have been considered before, and demonstrate their ability to capture the statistics of natural images. To our knowledge this provides the first analysis of which potential shapes are crucial for capturing natural image statistics in pairwise and high-order MRFs with learned filters. Despite significantly improved generative properties, at a first glance our models perform surprisingly poorly in an image denoising application.

In the last part of this paper we show that we need to move away from MAP estimation to take full advantage of generative MRFs. A number of recent theoretical and empirical results [13, 17, 29] have already pointed to deficiencies of MAP estimation. In the context of image denoising, we show that there is only a modest correlation between the generative quality of the image prior and the resulting denoising performance, which may also explain the prevalence of hand-tweaked models. To address these issues, we use the auxiliary-variable Gibbs sampler to infer the posterior mean, or *Bayesian minimum mean squared error estimate* (MMSE) in image denoising and inpainting. Contrary to common belief, we show that using sampling for posterior inference of MRFs in image restoration is both feasible and practical. Moreover, we demonstrate that the MMSE estimate not only substantially outperforms MAP, but also avoids several of its problems. Firstly, our approach no longer requires ad-hoc modifications (*cf*. [20]), but achieves state-of-the-art image restoration results in a *purely generative setting*, as compared to other random field models in the pixel domain. Secondly, using MMSE we find the generative quality of the model to be highly correlated with the quality of the denoising result. Finally, we observe that the MMSE avoids the inherent bias of MAP estimates toward δ-like marginals [29]. In contrast to [29], our approach does not require a modification of the proven MRF framework.

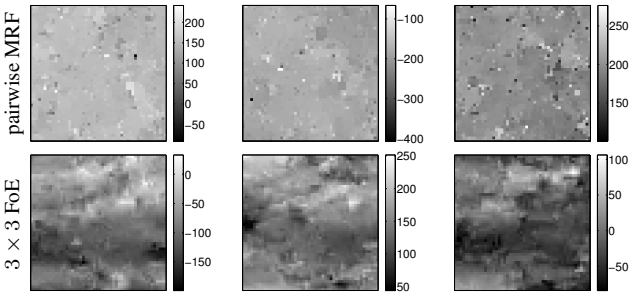


Figure 2. **Rapid mixing:** Three subsequent samples (l. to r.) from our learned models after reaching the equilibrium distribution.

Other Related Work. Even though MRFs provide a generative framework of natural images and scenes, their generative properties have been studied only rarely since [31], which is in contrast to local statistical models for which this is commonplace [19]. Weiss and Freeman [27] derive a likelihood bound that can be used to evaluate GSM-based high-order MRF models. Lyu and Simoncelli [15] analyze the marginals of their MRF model of wavelet coefficients by sampling, and find them to be non-Gaussian, but not as heavy-tailed as those of real image data. Levi [12] analyzes the marginals of Fields of Experts using sampling, and finds them to lack the heavy-tailed properties of natural images. Here, we extend this analysis to a wider range of different MRFs and show how to remedy this problem.

A number of authors have relied on sampling the posterior distribution for inference in MRFs, starting with Geman and Geman [8]. Fox and Nicholls [5] sample the posterior of binary MRFs using perfect sampling, and find the MMSE estimate to lead to more robust results. Barbu and Zhu [2] develop an efficient Swendsen-Wang sampling scheme and apply it to segmentation and stereo, but remain focused on MAP estimation in an annealing framework. Kim *et al*. [10] use population-based Markov chain Monte Carlo (MCMC) methods for stereo, but also remain in MAP setting. Works by Geman and Yang [7] and Levi [12] are most closely related to ours, as they both use efficient posterior sampling for image restoration with MRFs. Performance is limited in both cases, however, since [12] only uses a single posterior sample, and [7] employs annealing-based MAP estimation.

While there is little work on using MMSE estimation with MRFs aside from [5], MMSE estimation is more popular with local statistical models [19] and sparse representations [30]. Portilla and Simoncelli [19], for example, use the MMSE for their GSM-based model of wavelet coefficients. In contrast to the MRFs considered here, the local nature of these models makes the MMSE much easier to apply.

2. Flexible MRF Model and Efficient Sampling

For our analysis of MRFs in low-level vision, we focus on image priors to ease exposition. Yet, we expect many of the results to generalize to other models of scenes. Rather than proposing a new prior, we rely on Fields of Experts (FoEs) [20], high-order MRFs whose clique potentials model the responses to a bank of linear filters $\mathbf{J}_i$. The probability density of an image $\mathbf{x}$ under the FoE is written as

$$p(\mathbf{x};\boldsymbol{\Theta}) = \frac{1}{Z(\boldsymbol{\Theta})} e^{-\epsilon\|\mathbf{x}\|^2/2} \prod_{c\in\mathcal{C}} \prod_{i=1}^{N} \phi\big(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_{(c)};\boldsymbol{\alpha}_i\big), \quad (1)$$

where $\mathcal{C}$ are the maximal cliques, $\mathbf{x}_{(c)}$ are the pixels of clique c, ϕ is an expert function, $\boldsymbol{\alpha}_i$ are the expert parameters for $\mathbf{J}_i$, and $Z(\boldsymbol{\Theta})$ is the partition function that depends on all parameters $\boldsymbol{\Theta} = \{\mathbf{J}_i, \boldsymbol{\alpha}_i | i = 1,\dots,N\}$. The very broad Gaussian factor $e^{-\epsilon\|\mathbf{x}\|^2/2}$ with $\epsilon = 10^{-8}$ guarantees the model to be normalizable even if the experts do not fully constrain the image space [27]. Following [27], we use flexible Gaussian scale mixtures (GSMs) [19] as experts[1]:

$$\phi(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_{(c)};\boldsymbol{\alpha}_i) = \sum_{j=1}^{J} \alpha_{ij} \cdot \mathcal{N}(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_{(c)}; 0, \sigma_i^2/s_j). \quad (2)$$

α_{ij} are the (normalized) weights of the Gaussian component with scale s_j and base variance σ_i^2. GSMs have the advantage that they allow a wide variety of heavy-tailed potentials to be represented, including Student-t [20] and generalized Laplacians [25], and are yet computationally relatively easy to deal with. We use a fixed base variance and a wide range of 15 scales $s = \exp(0, \pm1, \dots, \pm5, \pm7, \pm9)$ to support a broad range of shapes.

Apart from a variety of FoE-based models [20, 21, 27], this general class of MRFs subsumes popular pairwise MRF models as well, *e.g.*, [11, 13, 25]. For the pairwise case we define a single fixed filter $\mathbf{J}_1 = [1, -1]^{\mathrm{T}}$ and let the maximal cliques $\mathcal{C}$ be all pairs of horizontal and vertical neighbors.

2.1. Auxiliary-variable Gibbs sampler

In order to provide a practical way of analyzing the generative properties of MRF priors through samples, an efficient sampling procedure is required. Since direct sampling is usually not feasible, Markov chain Monte Carlo (MCMC) methods have to be used. Single-site Gibbs samplers [8, 31] are very inefficient, as they need many iterations to reach the equilibrium distribution. Other Metropolis-based samplers, such as hybrid Monte Carlo [20], are sufficient for small images, but exhibit slow mixing for larger ones as sample dynamics have to be very small.

Here, we take a different route and exploit that our potentials use Gaussian scale mixtures. In the context of Products of Experts [9], Welling *et al.* [28] showed that it is beneficial to retain the scales of the GSM as an explicit hidden random vector $\mathbf{z} \in \{1,\dots,J\}^N$, one scale for each expert. Similar to a regular mixture model, one can define a joint distribution $p(\mathbf{x},\mathbf{z};\boldsymbol{\Theta})$ of $\mathbf{x}$ and the auxiliary mixture coefficients $\mathbf{z}$ such that $\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\Theta}) = p(\mathbf{x};\boldsymbol{\Theta})$. [28] showed that this allows defining a rapidly mixing auxiliary-variable Gibbs sampler that alternates between sampling $\mathbf{z}^{(t+1)} \sim p(\mathbf{z}|\mathbf{x}^{(t)};\boldsymbol{\Theta})$ and $\mathbf{x}^{(t+1)} \sim p(\mathbf{x}|\mathbf{z}^{(t+1)};\boldsymbol{\Theta})$, where t denotes the current iteration. If one only cares about obtaining samples of $\mathbf{x}$, the $\mathbf{z}$s can later be discarded. Similar ideas have been pioneered by Geman and Yang [7] in the context of MRFs. Levi and Weiss [12, 26] showed that this general framework can also be applied to MRFs with arbitrary Gaussian mixture potentials where $\mathbf{z} \in \{1,\dots,J\}^{N\times|\mathcal{C}|}$. From Eqs. (1) and (2), we obtain the following conditionals

$$p(z_{ic}|\mathbf{x};\boldsymbol{\Theta}) \propto \alpha_{iz_{ic}} \cdot \mathcal{N}(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_{(c)}; 0, \sigma_i^2/s_{z_{ic}}) \quad (3)$$

$$p(\mathbf{x}|\mathbf{z};\boldsymbol{\Theta}) \propto \mathcal{N}\left(\mathbf{x}; \mathbf{0}, \left(\epsilon\mathbf{I} + \sum_{i=1}^{N} \mathbf{W}_i\mathbf{Z}_i\mathbf{W}_i^{\mathrm{T}}\right)^{-1}\right), \quad (4)$$

where $\mathbf{W}_i$ are filter matrices that correspond to a convolution of the image with filter $\mathbf{J}_i$, and $\mathbf{Z}_i = \mathrm{diag}\{s_{z_{ic}}/\sigma_i^2\}$ are diagonal matrices with entries for each expert and clique. Sampling the scales according to Eq. (3) is straightforward, since their discrete distributions are conditionally independent given the image. Since the conditional distribution of the image given the scales in Eq. (4) is Gaussian, it can also be sampled without too much trouble. Difficulties arise from the fact that the (inverse) covariance matrix is huge with large images, which prevents an explicit Cholesky decomposition as in [28]. Levi and Weiss [12, 26] showed that this can be avoided by rewriting the covariance as

$$\boldsymbol{\Sigma} = \left(\epsilon\mathbf{I} + \sum_{i=1}^{N} \mathbf{W}_i\mathbf{Z}_i\mathbf{W}_i^{\mathrm{T}}\right)^{-1} = \left(\mathbf{W}\mathbf{Z}\mathbf{W}^{\mathrm{T}}\right)^{-1} \quad (5)$$

and obtaining a sample $\mathbf{x}$ by solving a least-squares problem

$$\mathbf{W}\mathbf{Z}\mathbf{W}^{\mathrm{T}}\mathbf{x} = \mathbf{W}\sqrt{\mathbf{Z}}\mathbf{y}, \qquad \mathbf{y} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \quad (6)$$

where the vector $\mathbf{y}$ is sampled from a unit normal; solving this sparse linear system of equations is much more efficient than a Cholesky decomposition. The advantage over single-site Gibbs samplers [31] or patch Gibbs samplers is that the whole image vector can be sampled at once, which leads to an efficient sampling procedure with rapid mixing (see Fig. 2) and fast convergence to the equilibrium distribution.

One problem frequently encountered with Markov random fields is that boundary pixels are less constrained than pixels in the image interior, which can lead to artifacts [18]. Boundary pixels are overlapped by fewer cliques in the MRF, and tend to take on extreme values when sampling the model. Since we found that this can affect learning

[1] Note that we sometimes use the terms potential and expert interchangeably, depending on the context.

and the analysis of the model through sampling, we follow Norouzi *et al*. [18] to keep a small number of pixels around the boundary, $\mathbf{x}^b$, fixed and conditionally sample the interior $\mathbf{x}^i$ according to $p(\mathbf{x}^i|\mathbf{x}^b, \mathbf{z}; \boldsymbol{\Theta})$. Since $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\Theta})$ is Gaussian, the required conditional distribution is easy to derive. All derivations can be found in the *supplemental material*.

2.2. Convergence analysis

Whenever MCMC methods are used to compute expected values and marginals, only fair samples after converging to the equilibrium distribution should be used to estimate the quantities of interest. While the auxiliary-variable Gibbs sampler mixes rapidly (see Fig. 2), a more rigorous procedure for monitoring convergence is still desirable. We use the popular approach by Gelman and Rubin [6], which relies on running several Markov chains in parallel and initializing them at different over-dispersed starting points. By computing the within-sequence variance W and the between-sequence variance B of a scalar estimand (here, the model energy), one can monitor convergence by estimating the potential scale reduction

$$\hat{R} = \sqrt{\big((n-1)W + B\big)/\big(nW\big)}, \tag{7}$$

where n is the number of iterations per chain. If $\hat{R}$ is near 1, we can regard the sampler to have approximately converged, since the chains have "forgotten" about their initialization. For computing $\hat{R}$ the first half of the samples is always conservatively discarded. We refer to [6] for details.

3. Generative Properties of Common MRFs

Evaluating the quality of MRF priors is a long-standing issue, as computing the likelihood of the model is intractable, and likelihood bounds [27] can only provide limited insight. Consequently, the performance in a certain application context is often measured instead [20, 25]; only providing a rather indirect measurement of how good the prior model is. In this paper we propose to revive and extend the methodology of Zhu and Mumford [31], which takes advantage of the generative nature of image priors by drawing samples from the model and evaluating its quality through the samples. The central advantage is that this provides a fully application-neutral way of evaluating MRFs.

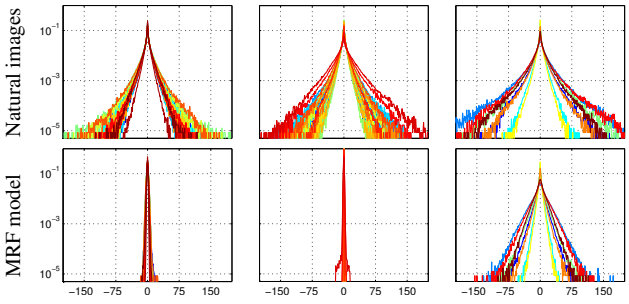
In order to exploit the efficient Gibbs sampler for an analysis of common MRF models, we convert them into the required form, if needed. For this we fit the flexible GSM potential from Eq. (2) to the target potential by minimizing their KL-divergence through simple nonlinear optimization of the weights α_{ij}. We achieve very good fits through a wide range of different potential shapes (KLD < 0.0005).

To evaluate the baseline statistics of natural images, we use a validation set of 3000 randomly cropped 32×32 non-overlapping patches from the test images of the Berkeley image segmentation dataset [16], and convert it to grayscale. The properties of the MRF models, on the other hand, are obtained by randomly sampling 3000 images of size 50×50. To avoid boundary artifacts, we condition on fixed image boundaries from a separate set of 3000 image patches. The fixed boundaries are $m - 1$ pixels wide/high, where m is the maximum extent of the largest clique; thus every interior pixel is constrained by equally many cliques. To avoid the effects of the boundary creeping into the analysis, we only collect sample statistics from 32×32 pixels in the middle. To draw a single sample from the model distribution, we set up three chains and assess convergence as described. We use three over-dispersed starting points: the interior of the boundary image, a median-filtered version, and a noisy version with Gaussian noise ($\sigma = 15$) added.

Pairwise MRFs. We first analyze the generative properties of pairwise MRF models, which remain popular until today due to their simplicity. The study of natural image statistics has widely found marginal histograms of image derivatives to exhibit a sharp peak at 0 and heavy tails (see Fig. 1), which motivates the use of heavy-tailed potentials with shapes similar to the empirical derivative statistics [11, 13, 25]. It has also become common to fit potential functions directly to the derivative histogram [22, 27]. This is at least unsatisfactory, since there is no direct correspondence between potential functions and marginals in MRFs.

Do these potential functions actually allow capturing the derivative statistics of natural images? We first consider generalized Laplacians ($\phi(y) = \exp(-\beta|y|^\gamma)$, typically $\gamma < 1$), which have been popular in the literature [13, 25] (here, $\beta = 0.5, \gamma = 0.7$). We also consider GSM potentials that have been directly fitted to the empirical marginals, similar to [22, 27]. As Fig. 1 shows, neither potential allows pairwise MRFs to capture the derivative statistics of real images. The model marginals are much too tightly peaked and the tails have an incorrect shape. Other potentials such as truncated quadratics exhibit similar issues. Evaluating other model properties appears pointless, since not even the statistics of the *model features* (*i.e*., derivatives) are captured. This seems surprising, however, given how widely used such models are. Since pairwise MRFs can be interpreted as maximum entropy models that capture first derivatives (*cf*. [31]), the potential shape can be the only culprit.

High-order MRFs. Since pairwise MRFs are quite restricted as they (at best) model the statistics of first image derivatives, high-order MRF models have become increasingly popular. While the early FRAME model [31] was found to exhibit heavy-tailed derivative marginals, only modest levels of image restoration performance have been achieved. The more recent Field of Experts (FoE) and variants [20, 21, 27] differs through its parametric expert functions and learned filters, and has shown to be among the



(a) [20], KLD = 2.19 (b) [27], KLD = 5.26 (c) Ours, KLD = 0.08

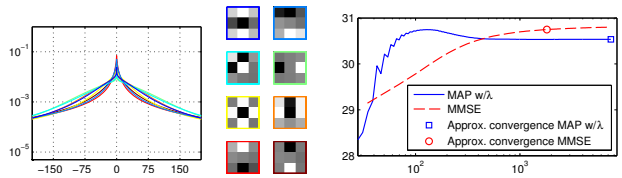
Figure 3. **Filter statistics of natural images and filter marginals of FoE models** (filters are normalized for ease of display)**:** The models of *(a)* Roth and Black [20], and *(b)* Weiss and Freeman [27] show poor generative properties. The 3×3 FoE learned here *(c)* matches the statistical properties of natural images much better.

best-performing image priors.

We analyze their generative properties by looking at the marginal distributions of filter responses. Since different models have different features (*i.e.*, filters), we evaluate each model w.r.t. its learned bank of filters. We consider both the original FoE with Student-t experts [20] and the GSM-based FoE model of [27]. The study of natural images has found that even arbitrary zero-mean filters have heavy-tailed statistics, which also holds for the learned filters (see Fig. 3, top). We confirm and extend the findings of Levi [12], that the original FoE model [20] does not capture the filter statistics (Fig. 3, bottom). The model marginals are much too peaky for all filters, and exhibit a high marginal KL-divergence. Beyond this, we find that the model of Weiss and Freeman [27] shows similarly unsatisfactory results. This is again surprising, given how well FoEs perform in real applications. Since FoEs can also be interpreted as maximum entropy models [31] that constrain the statistics of the bank of learned filters, this means either the parametric form of the experts or the learning procedure is at fault.

4. Learning Better Generative MRFs

To better understand the generative deficiencies of popular MRFs, we learn and analyze alternative MRFs. We rely on the flexible GSM-based FoE models from Sec. 2 and train them using the auxiliary-variable Gibbs sampler and contrastive divergence (CD) [9], an efficient alternative to maximum likelihood that does not require expensive equilibrium samples. We learn the GSM weights α_{ij} as well as the filters $\mathbf{J}_i$, except where mentioned. We ensure positive weights by updating their log, and enforce $\sum_j \alpha_{ij} = 1$. The remaining details of the learning procedure are similar to [20]: We use a fixed learning rate, exponential smoothing of the gradient, zero-mean filters, and stochastic gradient descent with mini-batches of 20 images. Our training set consists of 5000 grayscale 50×50 image patches, which were randomly cropped from the training images of [16]. To



(a) 3×3 FoE with 8 experts and filters (b) Time (sec.) vs. PSNR (dB)

Figure 4. *(a)* **Learned experts and filters** and *(b)* **efficiency** of CG-based MAP w/λ and sampling-based MMSE denoising.

avoid artifacts at the image boundaries, we use conditional sampling during learning as suggested by [18]. We found that this conditional learning procedure avoids overfitting on the boundary pixels, yet is more efficient than avoiding boundary effects by simply training on larger images.

Learned models. We trained (1) a pairwise MRF with fixed horizontal and vertical derivative filters and a single GSM potential, and (2) an FoE with 3×3 cliques and 8 GSM experts including filters. For the pairwise MRF we used 15 MCMC iterations per CD step, as this yielded slightly better results than 1-step CD. For the FoE we used 1-step CD for efficiency. Fig. 1(a) shows the learned pairwise potential, which is *significantly heavier-tailed than the marginal derivative statistics* and looks similar to a Student-t distribution [11]. In contrast to the popular pairwise MRFs from above, it correctly captures the marginal derivative statistics of natural images (Fig. 1(b), marginal KLD $= 0.006$). Because of the maximum entropy interpretation of MRFs, this potential shape is *optimal* for generative pairwise MRF image models. As far as we are aware, this is the first time that such an optimal pairwise potential has been reported ([31] only presented optimal potentials for high-order MRFs).

In case of the 3×3 FoE model, we find *very broad experts with a small, narrow peak* (see Fig. 4(a)). Their almost δ-like shape is in contrast to the experts used before [20, 27]. Fig. 3(c) shows that these learned experts lead to a much better match between the filter statistics of natural images and the filter marginals of the learned model. But despite the significant improvement, the filter statistics are not perfectly captured yet. Hence, further research needs to go into finding even better parametric potentials for high-order MRFs. We can conclude that the Student-t experts of [20] were not heavy-tailed enough, and that fitting experts to marginal statistics [27] is not appropriate. Instead, flexible expert functions and full learning of the experts are necessary to achieve good generative properties.

Analysis. To fully comprehend the modeling power of MRF priors, it is instructive to go beyond the model's features. A characteristic property of natural images is that even arbitrary zero-mean filters have heavy-tailed marginal statistics, as can be seen in Fig. 5(a) for filters of varying size. Another important property of natural images is the scale invariance of their derivative statistics [23]

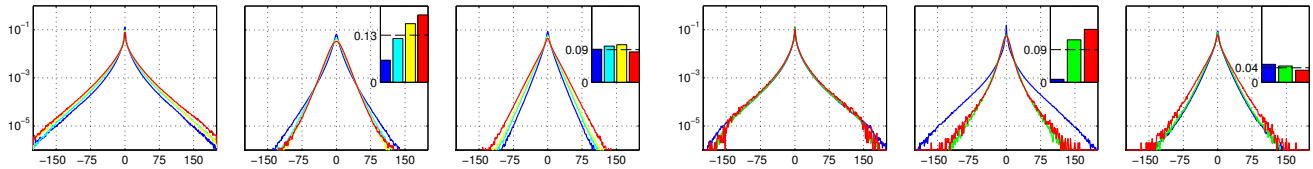


Figure 5. **Random filter statistics and scale-invariant derivative statistics:** *(a–c)* Average marginals of 8 random zero-mean unit-norm filters of various sizes (3×3 blue, 5×5 cyan, 7×7 yellow, 9×9 orange). *(d–f)* Derivative statistics at three spatial scales (1–blue, 2–green, 4–red; 1 refers to the original scale). The bar charts show the marginal KL-divergence of each feature. The learned pairwise MRF only captures short-range and small-scale statistics well. Our high-order FoE also models long-range and large-scale statistics. *Best in color.*

(*cf*. Fig. 5(d)). We propose to analyze generative models regarding these properties and to quantitatively measure the KL-divergence between the marginal statistics of images and the marginals of the model computed via sampling. Note that [31] only analyzed the derivative statistics and did not perform quantitative measurements. From Figs. 5(b) and (e), we can see that the learned pairwise MRF captures the statistics of small random 3×3 filters and derivatives at the smallest scale well. The model marginals of larger random filters and large-scale derivatives, however, tend toward being Gaussian. The learned high-order FoE on the other hand (Figs. 5(c) and (f)), captures the characteristics of natural images across a much wider range of random filter sizes and derivative scales, which clearly demonstrates the improved modeling power. This also becomes apparent by visually comparing samples from both models (Fig. 2). But since the statistics are still not perfectly captured, there is clearly room for improved models in the future.

5. Image Restoration

To analyze the impact of our improved generative models on real-world applications, we evaluate them in the context of image denoising and image inpainting. As is common in denoising, we assume i. i. d. Gaussian noise with known standard deviation, and evaluate the peak signal-to-noise ratio (PSNR). We test on two different test sets from the Berkeley segmentation dataset [16]: A set of 10 images used by [11], and a set of 68 images used by [1, 20, 21]. In case of image inpainting, we rely on a user-defined mask and fill in the missing pixels using the prior alone (*cf*. [20]).

MAP estimation. As a baseline, we restore images through maximum a-posteriori (MAP) estimation, as is usual in the literature. We maximize $p(\mathbf{x}|\mathbf{y};\boldsymbol{\Theta}) \propto p(\mathbf{y}|\mathbf{x})\cdot p(\mathbf{x};\boldsymbol{\Theta})^{\lambda}$ w.r.t. $\mathbf{x}$ using conjugate gradients (CG), where $p(\mathbf{y}|\mathbf{x})$ is the application specific likelihood, and λ is an optional regularization weight, which has frequently been employed to obtain good application performance (*e.g*., [20]). We compare our learned models against pairwise MRFs with three potential functions (standard and generalized Laplacian [25], and a marginal fit as in [22]) as well as two Fields-of-Experts models [20, 27]. Table 1 shows that despite their good generative properties, our learned models perform rather poorly when using MAP estimation and no regularization weight. With an optional regularization weight λ (optimized on the test set), we can achieve improved results, but still do not outperform previous models. Moreover, such a regularization weight deteriorates the generative properties (at least for our models). Why do good generative models not perform well?

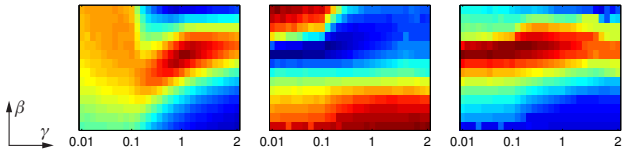
In unfortunately little known work, Nikolova [17] showed this to be an intrinsic problem of MAP estimation. To better understand this, we analyze the denoising performance of pairwise MRFs with a wide range of potentials from the family of generalized Laplacians $\phi(y;\beta,\gamma) = \exp\left(-\beta|y^2+\epsilon|^{\frac{\gamma}{2}}\right)$, where β controls the width of the potential, γ controls the heavy-tailedness, and the small $\epsilon > 0$ ensures differentiability. Moreover, we measure the generative quality of the model through the KL-divergence between the image derivative statistics and the model marginals. From Fig. 6 we make two important observations about MAP: First, the best performance is obtained from a convex potential ($\gamma = 1.0$, *i.e*., Laplacian). Second, there is only a moderate correlation between the generative quality of the model and denoising performance. Not only does this confirm the results of [17], it also offers an explanation why better generative models have not been used in the literature: *They simply performed poorly in the context of MAP.*

5.1. Alternative approach: MMSE estimation

To circumvent these problems, we propose to perform image restoration with MRFs by computing the *Bayesian minimum mean squared error estimate* (MMSE)

$$\hat{\mathbf{x}} = \arg\min_{\tilde{\mathbf{x}}} \int ||\tilde{\mathbf{x}} - \mathbf{x}||^2 p(\mathbf{x}|\mathbf{y};\boldsymbol{\Theta})\,\mathrm{d}\mathbf{x} = E[\mathbf{x}|\mathbf{y}], \quad (8)$$

which is equal to the mean of the posterior distribution and generally differs from the maximum in case of non-Gaussian posteriors, as are used here. Contrary to MAP, the MMSE estimate exploits the uncertainty of the model, but was long regarded as being impractical due to the difficulty of taking expectations over entire images (*cf*. [17]). To make this practical, we extend the efficient auxiliary-variable Gibbs sampler to the posterior. In case of removing Gaussian noise, we alternate between sampling the hidden



(a) PSNR of MAP (b) KL-divergence (c) PSNR of MMSE

Figure 6. **Correlation between generative properties and denoising performance:** Pairwise MRF with generalized Laplacian potential and different parameters. *(a)* MAP denoising with CG ("Lena", 128×128 pixels, σ=20, red – high PSNR); $\text{PSNR}_{\max} = 28.07$dB. *(b)* KL-divergence between derivative statistics of images and model marginals (blue – low KLD). *(c)* MMSE denoising with sampling; $\text{PSNR}_{\max} = 28.26$dB. While the correlation between KLD and PSNR of MAP is low (NCC $= -0.43$), the KLD and PSNR of MMSE are highly correlated (NCC $= -0.84$).

scales according to Eq. (3) and sampling the image according to

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}; \boldsymbol{\Theta}) = \mathcal{N}\big(\mathbf{x}; \widetilde{\boldsymbol{\Sigma}}\mathbf{y}/\sigma^2, \widetilde{\boldsymbol{\Sigma}}\big), \tag{9}$$

where σ^2 is the noise variance, $\widetilde{\boldsymbol{\Sigma}} = \big(\mathbf{I}/\sigma^2 + \boldsymbol{\Sigma}^{-1}\big)^{-1}$, and $\boldsymbol{\Sigma}$ is defined as in Eq. (5). MMSE estimation for image inpainting is possible through conditional sampling.

We compute the MMSE estimate by running 4 parallel Markov chains from different starting points (noisy image and smoothed versions from median, Wiener, and Gauss filtering), which allows to assess convergence using the potential scale reduction. After discarding the burn-in samples, we average all subsequent samples until the average images from the 4 samplers are sufficiently close to one another (< 1 grayvalue difference on average). The final image is obtained by averaging the samples from all samplers[2]. Figs. 7 and 8 show example results for inpainting and denoising. Fig. 4(b) shows the evolution of the PSNR over time (and thus no. of samples) for one of the images. We find that computing the MMSE using sampling is practical, despite our simple MATLAB implementation, and much superior to using only a single sample [12]. It is also easy to accelerate through parallel programming due to multiple samplers. Moreover, while MAP-based denoising using CG achieves a high PSNR early on, the performance at the (local) optimum of the posterior is often worse. The MMSE does not exhibit such a problem for our models.

[2]Note that the *learned models* and *sampling code* for model analysis and image restoration are available on the authors' webpages.

Table 1. Denoising results (avg. PSNR) for 10 test images [11].

Model	MAP		MAP w/λ		MMSE	
	σ=10	σ=20	σ=10	σ=20	σ=10	σ=20
pairw. (marg. fit [22])	28.35	23.96	30.98	26.92	29.70	24.72
pairw. (g. Lapl. [25])	27.35	22.97	31.54	27.59	28.64	23.92
pairwise (Laplacian)	29.36	24.27	**31.91**	**28.11**	30.34	25.47
pairwise (**ours**)	**30.27**	**26.48**	30.41	26.55	**32.09**	**28.32**
5×5 FoE from [20]	27.92	23.81	**32.63**	**28.92**	29.38	24.95
15×15 FoE from [27]	22.51	20.45	32.27	28.47	23.22	21.47
3×3 FoE (**ours**)	**30.33**	**25.15**	32.19	27.98	**32.85**	**28.91**

Figure 7. *(Left)* **Derivative statistics** of 10 denoised test images (blue, dotted – MAP; red, dashed – MMSE) and of corresponding clean originals (black, solid), σ=10, 20. *(Right)* **Sampling-based inpainting result** (MMSE) from our learned pairwise MRF.

Experiments. Table 1 compares MMSE estimation against MAP estimation, the latter with and without a regularization weight. We find that the MMSE outperforms MAP estimation, and when applied to good generative models, such as our learned ones, it even stays ahead of MAP with an optimal regularization weight. Note that this is despite MMSE estimation operating in a purely generative setting with *no regularization weight required*.

More extensive experiments on 68 test images [20, 21] confirm these findings. Table 2 shows that using MMSE-based denoising even our learned pairwise MRF outperforms the FoE of [20] using MAP, despite their much larger 5×5 cliques and noise-adaptive regularization weight. With MMSE denoising using posterior sampling, our learned 3×3 FoE model not only further improves the performance, but even outperforms the results of [21]. This is remarkable since their discriminative approach explicitly maximizes the denoising performance of MAP estimates, and furthermore uses larger cliques and more experts. In consequence, MMSE estimation enables *application-neutral generative MRFs* to be competitive with MAP-based denoising-specific discriminative MRFs (see *supplemental material* for additional results). We also compared to two MMSE-related state-of-the-art methods outside the random field literature: non-local means [4] (with tuned parameters), and the wavelet-based BLS-GSM [19]; we clearly outperform the former, and can compete with the latter despite not being limited to denoising.

Beyond improved quantitative results, MMSE-based image restoration has two other important advantages: First, MAP solutions have often been found to be piecewise constant with staircasing, which results in incorrect statistics of the output image (*cf*. [29] and Fig. 7). Woodford *et al*. [29] developed models that explicitly enforce certain statistical properties of the MAP estimate, but needed to abolish

Table 2. Denoising results for 68 test images [20, 21] ($\sigma = 25$).

Model	Learning	Inference	avg. PSNR
5×5 FoE from [20]	CD (generative)	MAP w/λ	27.44dB
5×5 FoE from [21]	discriminative	MAP	27.86dB
pairwise (**ours**)	CD (generative)	MMSE	27.54dB
3×3 FoE (**ours**)	CD (generative)	MMSE	27.95dB
Non-local means [4]	–	(MMSE)	27.50dB
BLS-GSM [19]	–	MMSE	**28.02**dB

(a) Original image (b) Noisy image (σ=25), 20.34dB (c) Learned pairwise MRF, MMSE, 26.12dB (d) 5×5 FoE from [20], MAP w/λ, 25.36dB (e) 5×5 discrimin. FoE from [21], MAP, 26.19dB (f) Learned 3×3 FoE, MMSE, 26.27dB

Figure 8. **Image denoising example** (cropped): Excellent restoration performance (PSNR) from generative models and MMSE estimation.

the well-understood MRF framework and had to rely on a rather complex inference procedure. From a practical point of view, Fig. 7 shows that we do not need to replace MRFs, but that replacing MAP with MMSE estimation is *already sufficient* to achieving the desired statistics of the output image and circumventing this long-standing problem. Second, Fig. 6 shows that the denoising performance of MMSE is highly correlated with the generative quality of the model, which in contrast to MAP suggests that better generative models are likely to improve application results without requiring any ad-hoc modifications.

6. Summary and Conclusions

Based on an efficient framework for analyzing the quality of MRF models using sampling, we found common image priors to exhibit poor generative properties. We demonstrated that this can be remedied with learned, flexible potentials. Moreover, we showed that MMSE estimation addresses a number of shortcomings of the prevalent MAP framework, and enables us to obtain excellent application performance from generative application-neutral models, that can even compete with recent, specialized discriminative approaches. We hope that our results will stimulate a renewed interest in generative models for low-level vision.

Acknowledgements: We are very grateful to Yair Weiss for sharing his ideas on the efficient Gibbs sampler that enabled this work. We would also like to thank Arjan Kuijper and Michael Goesele for discussions; and Kegan Samuel and Marshall Tappen for making detailed results of their paper available to us.

References

[1] A. Barbu. Learning real-time MRF inference for image denoising. *CVPR 2009*.

[2] A. Barbu and S.-C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *PAMI*, 27(8):1239–1253, 2005.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[4] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. *CVPR 2005*.

[5] C. Fox and G. K. Nicholls. Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. Technical report, Department of Mathematics, Auckland University, Auckland, New Zealand, 2000.

[6] A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[7] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE TIP*, 4(7):932–946, 1995.

[8] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *PAMI*, 6:721–741, 1984.

[9] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002.

[10] W. Kim, J. Park, and K. M. Lee. Stereo matching using population-based MCMC. *IJCV*, 83(2):195–209, 2009.

[11] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order Markov random fields. *ECCV 2006*.

[12] E. Levi. Using natural image priors – Maximizing or sampling? Master's thesis, The Hebrew University of Jerusalem, 2009.

[13] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. *CVPR 2009*.

[14] Y. Li and D. P. Huttenlocher. Learning for optical flow using stochastic optimization. *ECCV 2008*.

[15] S. Lyu and E. P. Simoncelli. Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures. *PAMI*, 31(4):693–706, 2009.

[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV 2001*.

[17] M. Nikolova. Model distortions in Bayesian MAP reconstruction. *AIMS J. on Inverse Problems and Imaging*, 1(2):399–422, 2007.

[18] M. Norouzi, M. Ranjbar, and G. Mori. Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning. *CVPR*, 2009.

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

[20] S. Roth and M. J. Black. Fields of experts. *IJCV*, 82(2):205–229, 2009.

[21] K. G. G. Samuel and M. F. Tappen. Learning optimized MAP estimates in continuously-valued MRF models. *CVPR 2009*.

[22] H. Scharr, M. J. Black, and H. W. Haussecker. Image statistics and anisotropic diffusion. *ICCV 2003*.

[23] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *J. Math. Imaging Vision*, 18(1):17–33, 2003.

[24] M. Szummer, P. Kohli, and D. Hoiem. Learning CRFs using graph cuts. *ECCV 2008*.

[25] M. F. Tappen, B. C. Russell, and W. T. Freeman. Exploiting the sparse derivative prior for super-resolution and image demosaicing. *Int. Workshop SCTV*, 2003.

[26] Y. Weiss. Personal communication, 2005.

[27] Y. Weiss and W. T. Freeman. What makes a good model of natural images? *CVPR 2007*.

[28] M. Welling, G. E. Hinton, and S. Osindero. Learning sparse topographic representations with products of Student-t distributions. *NIPS*2002*.

[29] O. J. Woodford, C. Rother, and V. Kolmogorov. A global perspective on MAP inference for low-level vision. *ICCV 2009*.

[30] I. Yavneh and M. Elad. MMSE approximation for denoising using several sparse representations. *4th World Conf. of the IASC*, 2008.

[31] S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *PAMI*, 19(11):1236–1250, 1997.