DISSERTATION

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of

Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Qi Gao, M.Sc.

Born in: Shaanxi, China

Oral examination: 11 July 2023

Filter-Based Probabilistic Markov Random Field Image Priors: Learning, Evaluation, and Image Analysis

Advisor: PD Dr. Karl Rohr

Abstract

Markov random fields (MRF) based on linear filter responses are one of the most popular forms for modeling image priors due to their rigorous probabilistic interpretations and versatility in various applications. In this dissertation, we propose an application-independent method to quantitatively evaluate MRF image priors using model samples. To this end, we developed an efficient auxiliary-variable Gibbs samplers for a general class of MRFs with flexible potentials. We found that the popular pairwise and high-order MRF priors capture image statistics quite roughly and exhibit poor generative properties. We further developed new learning strategies and obtained high-order MRFs that well capture the statistics of the inbuilt features, thus being real maximum-entropy models, and other important statistical properties of natural images, outlining the capabilities of MRFs. We suggest a multi-modal extension of MRF potentials which not only allows to train more expressive priors, but also helps to reveal more insights of MRF variants, based on which we are able to train compact, fully-convolutional restricted Boltzmann machines (RBM) that can model visual repetitive textures even better than more complex and deep models.

The learned high-order MRFs allow us to develop new methods for various real-world image analysis problems. For denoising of natural images and deconvolution of microscopy images, the MRF priors are employed in a pure generative setting. We propose efficient sampling-based methods to infer Bayesian minimum mean squared error (MMSE) estimates, which substantially outperform maximum a-posteriori (MAP) estimates and can compete with state-of-the-art discriminative methods. For non-rigid registration of live cell nuclei in time-lapse microscopy images, we propose a global optical flow-based method. The statistics of noise in fluorescence microscopy images are studied to derive an adaptive weighting scheme for increasing model robustness. High-order MRFs are also employed to train image filters for extracting important features of cell nuclei and the deformation of nuclei are then estimated in the learned feature spaces. The developed method outperforms previous approaches in terms of both registration accuracy and computational efficiency.

Zusammenfassung

Markov Random Fields (MRF) auf der Basis linearer Filterantworten sind eine der beliebtesten Formen zur Modellierung von Bildprioren aufgrund ihrer rigorosen probabilistischen Interpretationen und Vielseitigkeit in verschiedenen Anwendungen. In dieser Dissertation schlagen wir eine anwendungsunabhängige Methode vor, um MRF-Bildprioren quantitativ mithilfe von Stichproben der Modelle zu bewerten. Zu diesem Zweck haben wir effiziente Hilfsvariablen-Gibbs-Sampler für eine allgemeine Klasse von MRFs mit flexiblen Potentialen entwickelt. Wir haben festgestellt, dass die populären Pairwise- und High-Order-MRF Prioren die Bildstatistik sehr grob erfassen und schlechte generative Eigenschaften aufweisen. Wir haben weiterhin neue Lernstrategien entwickelt und High-Order-MRFs erhalten, die die Statistik der eingebauten Merkmale gut erfassen, und damit echte Maximum-Entropie-Modelle und andere wichtige statistische Eigenschaften von Naturbildern darstellen, wodurch die Fähigkeiten von MRFs herausgestellt werden. Wir schlagen eine multimodale Erweiterung der MRF-Potentiale vor, die nicht nur das Training von expressiveren Prioren ermöglicht, sondern auch dazu beiträgt, weitere Einblicke in MRF-Varianten zu gewinnen, auf deren Basis wir kompakte, vollständig konvolutionale Restricted Boltzmann Machines (RBM) trainieren können, die visuelle wiederkehrende Texturen besser modellieren können als komplexere und tiefe Netzwerke.

Die erlernten High-Order-MRFs ermöglichen es uns, neue Methoden für verschiedene Realwelt-Bildanalyseprobleme zu entwickeln. Für die Rauschunterdrückung von natürlichen Bildern und die Dekonvolution von Mikroskopiebildern werden MRF-Prioren in einem rein generativen Rahmen eingesetzt. Wir schlagen effiziente sampling-basierte Methoden vor, um die Bayesschen Minimum-Mean-Squared-Error (MMSE) Schätzungen zu inferieren, die wesentlich besser abschneiden als die Maximum-a-posteriori-Methode (MAP) und mit den State-of-the-Art diskriminativen Methoden konkurrieren können. Für die nichtstarre Registrierung von Lebend-Zellkernen in Zeitraffer-Mikroskopiebildern schlagen wir eine globale optische Fluss-basierte Methode vor. Die Statistik des Rauschens in Fluoreszenz-Mikroskopiebildern wird untersucht, um ein adaptives Gewichtungsverfahren zur Erhöhung der Modellrobustheit abzuleiten. High-Order-MRFs werden auch eingesetzt, um Bildfilter zur Extraktion wichtiger Merkmale von Zellkernen zu trainieren, und die Deformation der Kerne wird dann in den erlernten Merkmalsräumen geschätzt. Die entwickelte Methode übertrifft bisherige Ansätze in Bezug auf Registrierungsgenauigkeit und Recheneffizienz.

Acknowledgment

First and foremost, I would like to express my deepest appreciation to my advisor, *PD Dr. Karl Rohr*, for his invaluable guidance and support. His extensive knowledge and expertise in cross-disciplinary collaboration, scientific writing, and research methodology were essential in shaping my research and preparing me for the future. His unwavering encouragement and insightful feedback helped me navigate through various research challenges.

I would also like to extend my sincere gratitude to my former advisor *Prof. Stefan Roth* for his exceptional guidance and help. His rigorous mathematical training and expert advice provided me with a strong foundation for tackling complex research problems. His mentorship has left a lasting impact on my research and career.

My special thanks go to *Uwe Schmidt* for extensive discussions, and to *Prof. M. Cristina Cardoso* and *Vadim Chagin* for valuable input from another field. Their diverse expertise, constructive feedback, and willingness to share their knowledge were instrumental in advancing my research and broadening my perspective. I am very grateful to my colleagues in the Biomedical Computer Vision (BMCV) group for their kind help, interesting discussions, and successful collaboration. *Marco Tektonidis* generously shared his data and results with me for experimental comparison. *Simon Eck* conducted additional experiments to demonstrate the effectiveness of my method. *Thomas Wollmann* helped me get started quickly with software development on the Galaxy platform. *Ran Li* and *Roman Spilger* offered me many fruitful discussions on object tracking methods. *Yu Qiang* always shared his interested ideas and findings with me. Moreover, I would like to thank *Sabrina Wetzel* and *Christine Herrmann* for their meticulous administrative support.

I am deeply grateful to my family and friends for their priceless support and encouragement throughout this journey.

Finally, I would like to acknowledge the support provided by various organizations, including the DFG through the Cluster of Excellence CellNetworks, the SPP 2202, and the SFB 1129; the Math-Clinic Core Facility at Heidelberg University; and the BMBF through projects de.NBI (HD-HuB) and CancerTelSys.

Publications

- **Qi Gao**, Vadim Chagin, M. Cristina Cardoso and Karl Rohr, "Quantifying newly appearing replication foci in cell nuclei based on 3D non-rigid registration," *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI 2022)*, Kolkata, India, March 28-31, 2022

- Dan Wang, **Qi Gao**, Ina Schaefer, Handan Moerz, Ulrich Hoheisel, Karl Rohr, Wolfgang Greffrath, and Rolf-Detlef Treede, "TRPM3-mediated dynamic mitochondrial activity in nerve growth factor-induced latent sensitization of chronic low back pain," *PAIN*, Vol. 163(11): e1115-e1128, 2022

- Sheng Liu, Veronica Bonalume, **Qi Gao**, Jeremy T.-C. Chen, Karl Rohr, Jing Hu, and Richard Carr, "Pre-synaptic GABA_A in NaV1.8^+ primary afferents is required for the development of punctate but not dynamic mechanical allodynia following CFA inflammation," *Cells*, Vol. 11(15): 2390, 2022

- **Qi Gao**, Vadim Chagin, M. Cristina Cardoso and Karl Rohr, "Non-rigid registration of live cell nuclei using global optical flow with elasticity constraints," *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI 2021)*, Nice, France, April 13-16, 2021, pp.1457-1460

- Ran Li, **Qi Gao**, and Karl Rohr, "Multi-object dynamic memory network for cell tracking in time-lapse microscopy images," *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI 2021)*, Nice, France, April 13-16, 2021, pp.1029-1032

- **Qi Gao** and Karl Rohr, "A global method for non-rigid registration of cell nuclei in live cell time-lapse images," *IEEE Transactions on Medical Imaging*, Vol. 38(10): 2259-2270, 2019

- Frank Adolf, Manuel Rhiel, Bernd Hessling, **Qi Gao**, Andrea Hellwig, Julien Bethune, and Felix T. Wieland, "Proteomic profiling of mammalian COPII and COPI vesicles," *Cell Reports*, Vol. 26: 250-265, 2019

- **Qi Gao**, Simon Eck, Jessica Matthias, Inn Chung, Johann Engelhardt, Karsten Rippe, and Karl Rohr, "Bayesian joint super-resolution, deconvolution, and denoising of images with Possion-Gaussian noise," *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, D.C., USA, April 4-7, 2018, pp. 938-942

- **Qi Gao** and Karl Rohr, "Optical flow-based non-rigid registration of cell nuclei: Global model with adaptively weighted regularization," *Proceedings of IEEE*

International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, April 18-21, 2017, pp. 420-423

- **Qi Gao** and Stefan Roth, “Texture synthesis: from convolutional RBMs to efficient deterministic algorithms,” *Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2014. Lecture Notes in Computer Science*, Vol. 8621, pp. 434–443. Springer, Berlin, Heidelberg.

- **Qi Gao** and Stefan Roth, “How well do filter-based MRFs model natural images?” *Pattern Recognition. DAGM/OAGM 2012. Lecture Notes in Computer Science*, Vol. 7476, pp. 62-72. Springer, Berlin, Heidelberg. (*DAGM 2012 Prize*)

- Uwe Schmidt*, **Qi Gao*** and Stefan Roth, “A generative perspective on MRFs in low-level vision,” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, San Francisco, USA, June 13-18, 2010, pp. 1751-1758 (oral presentation; *equal contribution)

Notation

Symbol	**Description**
$x,\ h_{jc},\ \alpha_i,\ \cdots$	Scalars
$\mathbf{x},\ \mathbf{f}_j,\ \mathbf{w}_{jc},\ \cdots$	Vectors
$\mathbf{A},\ \mathbf{\Sigma},\ \mathbf{\Gamma},\ \cdots$	Matrices
$(\cdot)_i$	Elements of vectors or matrices
$\mathbf{f}_j^{\mathrm{T}},\ \mathbf{H}^{\mathrm{T}}$	Transposed vector and matrix
$g(\mathbf{x}),\ \phi(y)$	Scalar-valued functions
$\boldsymbol{\psi}(\mathbf{x})$	Vector-valued function
$\boldsymbol{\psi}'(\mathbf{x})$	Derivatives of vector function $\boldsymbol{\psi}(\mathbf{x})$ (element-wise)
$\frac{\partial f(\mathbf{x})}{\partial x_i},\ \frac{\partial}{\partial x_i} f(\mathbf{x})$	Partial derivative of scalar function $f(\mathbf{x})$
$\nabla_{\mathbf{x}} f(\mathbf{x}),\ \nabla f(\mathbf{x})$	Gradient of scalar function $f(\mathbf{x})$ w.r.t. $\mathbf{x}$
$\mathbf{f}_j * \mathbf{x}$	Convolution of image $\mathbf{x}$ with filter $\mathbf{f}_j$
$p(x),\ p(\mathbf{x})$	Probability density functions (continuous)
$p(\mathbf{x}\|\mathbf{y})$	Conditional probability (density) of $\mathbf{x}$ given $\mathbf{y}$
$p(\mathbf{x};\ \boldsymbol{\omega}_j)$	Probability (density) of $\mathbf{x}$ given parameters $\boldsymbol{\omega}_j$
$\left\langle f(\mathbf{x}) \right\rangle_{p(\mathbf{x})},\ \left\langle f(\mathbf{x}) \right\rangle_p$	Expected value of $f(\mathbf{x})$ w.r.t. probability (density) $p(\mathbf{x})$
$\left\langle f(\mathbf{x}) \right\rangle_{\mathbf{X}}$	Expected value of $f(\mathbf{x})$ w.r.t. data $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots]$
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$	Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$

Contents

Chapter 1

Introduction

1.1 Motivation

Computer vision is an interdisciplinary field that studies how computers can understand and interpret digital images or videos using computational models and algorithms. Its applications range from pixel-level information extraction (low-level vision) to semantic understanding of the image content (high-level vision) using, for example, methods for image segmentation and object tracking.

A typical low-level vision problem is microscopy image deconvolution, which we use to motivate our work in this thesis (*cf*. Fig. 1.1). The task is to recover sharp images of the samples (or scenes) from the blurry images with noise acquired by microscopes, which is very useful for further analysis. Although this problem has been studied for decades, it has remained difficult, even if the noise statistics and the point spread function (PSF) of the microscope are known. The essential reason is that this problem (as many other low-level vision problems) is mathematically ill-posed: The available data is generally insufficient to constraint the solution, the image formation process is too complex to be perfectly modeled, and the noise is generated randomly by the sensor. To deal with these issues in practice, prior knowledge, for example, in the form of a regularizer, is often imposed in the computational model to constraint the solution.

An important and common strategy for building a computational model is to adopt a probabilistic formulation due to some distinct advantages. For example, the modeling choices can be made based on statistical properties of the data, and the most suitable model parameters can be determined using learning algorithms. For the optimization process to compute the solution, advanced statistical inference methods can be employed. In addition, the uncertainty of the solution is quantitatively accessible, for example, using information theoretic measures such as the entropy. In a popular Bayesian approach, the posterior probability distribution of

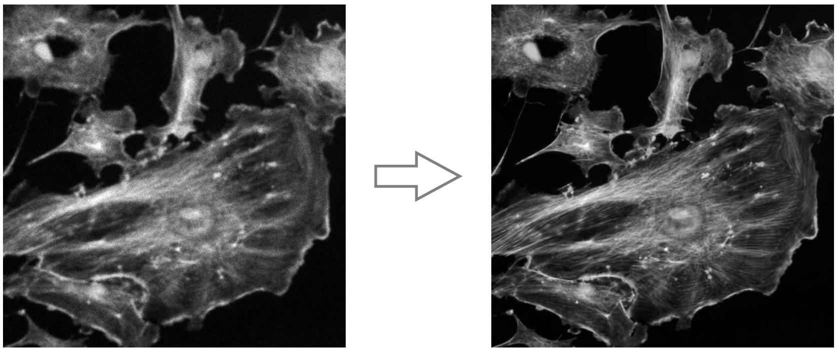


Figure 1.1: Microscopy image deconvolution. (Left) blurry and noisy image acquired by a microscope; (right) recovered "ideal" image.

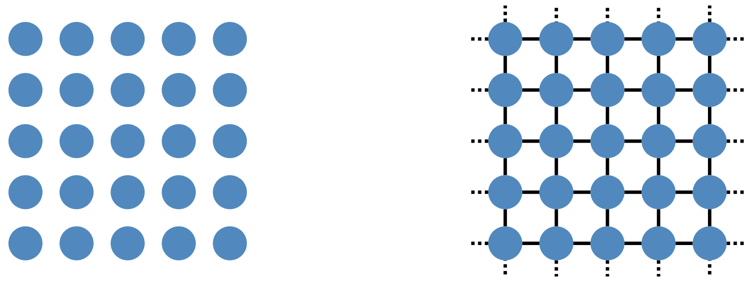
the hidden ideal image given the observed image is defined using Bayes' rule

$$p(\text{ideal} \,|\, \text{observed}) \propto p(\text{observed} \,|\, \text{ideal}) \cdot p(\text{ideal}), \tag{1.1}$$

where the likelihood $p(\text{observed} \,|\, \text{ideal})$ models how the ideal image is degraded during acquisition (or how the observed image is generated in general) and is thus application dependent, and the prior $p(\text{ideal})$ models our belief or prior knowledge about what the ideal state should be and is independent to the observation. Such probabilistic models with separate likelihood and prior terms are called *generative models*. The prior model is application-independent and thus very versatile, and can be used in different applications by changing the likelihood model. In contrast, in a *discriminative model*, the posterior density (e.g., $p(\text{ideal} \,|\, \text{observed})$) is directly modeled. While the discriminative models can generally achieve higher application performances, they have to be trained end-to-end entirely for each application.

Modeling (or training) a probabilistic prior of images is not easy, since an image taken by a common consumer camera today or a modern microscope contains tens of millions of pixels, resulting in a very high dimension of the model if each pixel is considered as a random variable. As a result, priors of small patches were proposed (e.g., ICA-based [1], Products-of-Experts [2], mixture of Gaussians [3]), which could extensively treat the statistics of image patches (due to relatively lower dimensionality) and yield good image restoration results. However, as these models are build on fixed small patches (e.g., 8×8 pixels in [3]), it is generally difficult to adapt them to modeling entire images with arbitrary sizes, which limits their use in applications. Note that, due to the "curse of dimensionality", learning a patch-based prior still requires a large amount of training data and is not easy.

As the pixel grid of an image can be naturally regarded as nodes of a graph (*cf*. Fig. 1.2), undirected graphical models, or more specifically Markov random fields (MRFs), have been used to formulate image priors [4]. MRFs provide a sound probabilistic framework for modeling the entire images and dense scene representations.



(a) Image pixel grid. Nodes represent pixels.

(b) Undirected graph. Nodes are connected by edges.

Figure 1.2: Analogy of pixel grid for images to nodes of undirected graphs.

In general, training a MRF means estimating the potential functions associated with the maximal cliques of the graph (e.g., pairwise connected nodes in Fig. 1.2(b)). The potential functions are often identical for different connected nodes due to the assumption of spatial homogeneity of images, and thus the model training does not require too much data compared to patch-based priors. Different MRF priors have been proposed and found widespread use across low-level vision, ranging from the pairwise models (e.g., [4, 5]) to high-order ones (larger maximal cliques, e.g., [6, 7]), from modeling generic images to modeling visual textures (e.g., [8, 9]), from using hand-crafted filters and/or potential functions (e.g., [10, 6]) to using learned ones (e.g., [7, 11]), and from homogeneous models to content-aware MRFs [12, 13].

A trend in computer vision is to move away from a rigorous probabilistic interpretation of MRF. One reason is the prevalence of non-probabilistic discriminative methods (e.g., loss function-based training [14, 11]). More significantly, deep neural networks have gained momentum. Besides exploiting potential functions, inter-pixel relations can be modeled using deep convolutional networks (CNNs) to learn deep MRFs [15, 16]. More common is directly employing "traditional" MRFs in the deep learning models as a regularizer in the loss function (e.g., optical flow estimation [17, 18], image registration [19, 20, 21], blind image denoising [22], medical image reconstruction [23]). As the learning target of these models coincides with application-specific model evaluation, the performance of such models is highly appealing. In addition, deep generative image models (e.g., [24, 25, 26, 27]) explicitly or implicitly approximate high-dimensional probability distributions of images and can be regarded as image priors. However, the deep learning models lack statistical interpretability and/or versatility of generative MRFs.

Despite that MRF priors are still widely used in image analysis, learning and inference of MRFs, even for the pairwise models, is not easy. Moreover, it is actually not clear whether these MRF models have been properly trained or whether they have right captured important statistics of real images, since they have only been evaluated indirectly in applications.

The potentials and problems of MRFs for modeling image priors motivate our work. The main goal of this dissertation is to explore how well high-order MRFs can model images and how much a good prior can improve the performance of generative models. To this end, we will develop strategies for dealing with the difficulties in learning (especially high-order) MRFs, for application-independent evaluation of the learned prior models, and for efficient probabilistic inference in various computer vision problems.

1.2 Challenges

We use a vector $\mathbf{x} \in \mathbb{R}^N$ to denote all pixels of an image or a dense scene representation. Under a common filter-based MRF framework, where the clique potential is represented as a Product of Experts [28], the probability density of an image is written as

$$p(\mathbf{x};\boldsymbol{\Omega}) = \frac{1}{Z(\boldsymbol{\Omega})} \prod_{c \in \mathcal{C}} \left(\prod_{j} \phi(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; \boldsymbol{\omega}_j) \right), \tag{1.2}$$

where $\mathbf{x}_{(c)}$ are the pixels of maximal clique $c \in \mathcal{C}$, clique potentials use experts[1] $\phi(\cdot)$ with parameters $\boldsymbol{\omega}_j \in \boldsymbol{\Omega}$ to model the responses to linear filters $\mathbf{f}_j$, and $Z(\boldsymbol{\Omega})$ is the partition function. In this section, we use this MRF framework to briefly discuss the challenges.

MRF modeling choices and training

Pairwise MRFs are equivalent to the filter-based MRFs in (1.2) that rely on x- and y-derivative filters (see Sec. 2.2.3 for details), but their small neighborhood with pairwise cliques limit the expressiveness of MRFs. A straightforward extension is to use larger, high-order (derivative) filters that imply larger maximal cliques to define the so-called high-order MRFs (e.g., [8]). In the framework of Fields of Experts (FoE) [7], the filters can even be directly learned from the training data. How to choose or learn the most suitable filters for MRFs is a challenging task.

The choice of potential functions $\phi(\cdot)$ has always been an issue for model performance. Early MRF models use Gaussian potentials (*cf*. Fig. 1.3(b)) due to several reasons, for example, easy inference. However, natural images of common scenes exhibit heavy-tailed marginal distributions, significantly differing from Gaussians (*cf*. Fig. 1.3(a)), which suggests the use of robust or sparse potential functions, e.g., the generalized Laplacian [5, 29] (*cf*. Fig. 1.3(c)) or the Student-t function [7, 30] (*cf*. Fig. 1.3(d)). In [31], the potentials were computed by fitting the image marginals

[1] The terms "potential" and "expert" are sometimes used inter-changeably, depending on the context.

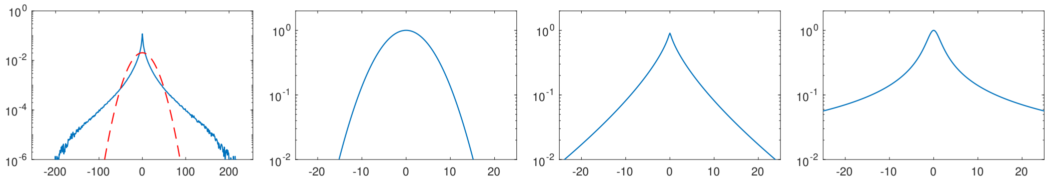
using more flexible Gaussian scale mixtures (GSM) [32]. In fact, many MRFs including high-order ones have been hand-tuned and only roughly capture the statistics of images. Moreover, the most common potential functions in MRFs are uni-modal. While (probably) being sufficient for modeling natural images, they are not able to properly model visual textures due to different statistics [9]. Exceptions are the FRAME models [6, 8] which rely on a discrete non-parametric potential representation (thus allowing for multi-modal potentials) and strict learning procedures. While the FRAME model for texture modeling [8] yields impressive results, the FRAME model for generic natural images [6] does not perform competitively in image restoration tasks.

A special case of MRFs is the restricted Boltzmann machine (RBM) [33] which is a generative stochastic neural network. Its continuous variants are often used for image feature (or filter) learning (e.g., [34, 35]) or modeling visual textures (e.g., [36, 37]). A question is: What are the characteristics of RBMs as image models compared to common filter-based MRFs? In addition, both RBMs and filter-based MRFs can be employed to train filters for image feature extraction in higher-level vision problems. In this case, what is the relation between the two types of learned filters?

Challenges also arise from training MRFs. In the context of training MRF image priors, unsupervised learning of very high-dimensional models should be accomplished. Unfortunately, learning (and probabilistic inference) in many MRF models is NP-hard [38], and the more complex the model is (e.g., high-order MRFs), the more difficult the training will be. For the common maximum likelihood (ML) learning objective, as there is generally no closed form expression for the expectation of model likelihood, exact computation is intractable. Approximate inference, for example, using sampling, must thus be used. Markov chain Monte Carlo (MCMC) approximations are historically most common (e.g., [6]), but very inefficient. Consequently, ML estimation itself was frequently approximated by contrastive divergence (CD) [33], which avoids costly equilibrium samples. Different efficient sampling methods such as Gibbs sampler [4, 33], hybrid Monte Carlo [7] were developed. In addition, deterministic methods including basis rotation [39] and score matching [40] have also been used. These approaches rely on particular potential functions, either fitted after the training process or with limited expressiveness, which constrains their applicability.

Model evaluation

Evaluating the quality of MRF priors is a long-standing issue. Due to the intractable partition function of the MRF, direct computation of the model likelihood is not possible. In [39], likelihood bounds for GSM-based MRFs were derived, but could only provide limited insights. Actually, MRF model evaluation largely happens in



(a) (b) (c) (d)

Figure 1.3: Marginal statistics of natural images and common potential (or expert) functions in MRFs. (a) Histogram of neighboring pixel differences (blue, solid) and its Gaussian fit (red, dashed); (b) Gaussian expert $\phi(y;\sigma)=\exp\left(-\frac{1}{2\sigma^2}y^2\right)$ (here, $\sigma=5$); (c) generalized Laplacian expert $\phi(y;\beta,\gamma)=\exp(-\beta|y|^\gamma)$ (here, $\beta=0.5,\gamma=0.7$); (d) Student-t expert $\phi(y;\alpha)=\left(1+\frac{1}{2}y^2\right)^{-\alpha}$ (here, $\alpha=0.5$). Note that logarithmic (base 10) scale is used for the y-axis.

the context of particular applications, e.g., using image restoration to evaluate image priors [5, 7], which is a rather indirect measurement of how good the prior model is.

Since the statistics of natural images and scenes motivate the modeling choices for MRFs, a central question is how well the statistical properties are captured by these MRF models. This is relatively unclear, although MRFs provide a generative framework of images and scenes. In contrast to local statistical models (e.g., in [32], the local variance of the wavelet coefficients in the neighborhood is modeled), the generative properties of MRFs have been studied only rarely, mainly because computing marginal distributions and other probabilistic properties of MRFs is difficult. In [6], Zhu and Mumford advocated the use of sampling for evaluating their MRF prior and computed derivative histograms from model samples. However, widely applicable sampling algorithms (e.g., standard Gibbs sampling [4, 6], hybrid Monte Carlo [7]) are inefficient, while sampling is almost the only choice. Based on sampling, Lyu and Simoncelli [41] found the marginals of their MRF model of wavelet coefficients to be non-Gaussian, but not as heavy-tailed as those of real image data. Levi [42] also exploited sampling and found the marginals of the Fields of Experts [7] to lack the heavy-tailed properties.

Efficient inference

The highly non-convex energy of common MRF models makes probabilistic inference of the posterior very difficult. In the context of maximum a-posteriori (MAP) estimation, gradient-based methods (e.g., conjugate gradient method in [7]), though being relatively efficient, can typically only find local optima of the objective; graph cuts (e.g., [43, 44]) and belief propagation (e.g., [5, 45]) are good at finding approximate solutions, but are discrete models in contrast to the continuous interpretation of MRFs, and are difficult to apply to models with larger cliques [46]. To achieve best performance in an application, some ad-hoc modifications, e.g., imposing an additional weight on the MRF prior [7], have been made, which contradicts the application-independent setting in generative models. A question is whether this means that the MRF priors are not properly learned.

One of the advantages of probabilistic models is the accessibility of the uncertainty of the solution . The MAP solution is a point estimate and does not reveal any statistical information. Sampling the posterior allows approximating the marginals and the expectation of the posterior (e.g., perfect sampling in [47]), but is generally computationally very expensive. More efficient variational Bayesian methods, in particular, mean field approximation, have been used in multiple applications of MRFs with a certain loss of model accuracy (e.g., [48, 49]).

1.3 Contributions

In this dissertation, we performed an application-independent evaluation of probabilistic Markov random field (MRF) image priors by checking their generative properties and developed new effective learning approaches for training high-order MRF priors that accurately capture key statistics of images and dense scene representations. We also developed new image analysis methods using the learned MRF priors in different real-world problems: denoising of natural images, deconvolution of microscopy images, and non-rigid registration of live cell microscopy images. This demonstrates the versatility and performance competitiveness the learned MRF image priors. The major contributions of this dissertation can be summarized into two categories:

1.3.1 Methods for application-independent evaluation, effective learning, and understanding of MRF image priors

Evaluation of MRF priors using sampling. We developed an efficient auxiliary-variable Gibbs sampler for a general filter-based MRF with potentials represented by flexible Gaussian scale mixtures (GSM, mixture of Gaussians with same mean but different variances) [39]. This allows us to analyze the generative properties of common MRF priors using sampling and to perform an application-independent evaluation by checking the model statistics quantitatively. We surprisingly found that the popular pairwise MRFs [5, 31] and high-order Fields of Experts (FoEs) [7, 39] are relatively poor generative models and can not even capture the statistics of the inbuilt features, i.e. responses to the (learned) model filters. This contradicts the maximum entropy interpretation of filter-based MRFs [6], indicating that the potential functions of these models are chosen or learned inappropriately. The work was published in [50].

Learning MRF priors that capture proper image statistics. We exploited the developed efficient Gibbs sampler to train pairwise MRFs and high-order FoEs with flexible GSM potentials. We found that a standard maximum likelihood learning procedure based on sampling and contrastive divergence (CD) [33] only works well for pairwise MRF. We further proposed strategies to solve several learning issues for the high-order models. Heavier-tailed potentials than have been considered in previous work (e.g., [5, 31, 7, 39]) allow the model to capture the statistics of the inbuilt features correctly; the trained models are thus real maximum entropy models. More importantly and in contrast to previous MRFs, the learned FoEs also quite accurately represent multi-scale derivative statistics, random filter statistics as well as joint feature statistics of natural images. To the best of our knowledge, this is the first time that such close matches between model and natural image statistics have been reported for an MRF image prior. This also demonstrates that FoEs are indeed capable of capturing a large number of key statistical properties of natural images. The work was published in [50, 51].

Understanding the mean and covariance units in MRFs. To improve the model expressiveness, we extended the MRFs with multimodal potentials based on mixtures of GSMs. Such an extension not only enables training MRF priors for specific classes of images (e.g., visual textures), but also helps us to find important insights of MRFs through revisiting the seminal FRAME model [6]. To further understand the importance of "covariance" units that impose regularization in typical MRFs for visual textures, we proposed to learn "mean"-only fully convolutional Gaussian restricted Boltzmann machines (RBMs) for modeling Brodatz textures [52]. Our learned models exhibit several favorable properties of simplicity, efficiency, spatial invariance, and a comparatively small number of structured, more interpretable features; yet they outperform the more complex deep belief networks [37]. The results demonstrate that the "mean" units are actually most important in modeling textures. The work was published in [53].

1.3.2 Image analysis methods based on (learned) MRF priors

Denoising natural images based on sampling. We modeled the denoising problem in a Bayesian framework with the learned high-order MRFs for generic natural images as the prior in a purely generative setting. We extended the efficient auxiliary-variable Gibbs sampler to infer the posterior mean or the Bayesian minimum mean squared error (MMSE) estimate. Experiments show that our approach outperforms previous MRF-based models and can compete with popular denoising methods such as non-local sparse coding [54] and BM3D [55]. Moreover, the MMSE estimate not only substantially outperforms MAP, but also avoids several of its problems (e.g., incorrect statistics of the output images [56]). We demonstrated that a rigorous probabilistic interpretation and good generative properties of MRFs can go hand-in-hand with very good application performance. The work was published in [50, 51].

Joint deconvolution, denoising and super-resolution of microscopy images. We developed a new Bayesian approach that simultaneously performs deconvolution, denoising as well as super-resolution to restore microscopy images with mixed Poisson-Gaussian noise. The model is based on a likelihood modeling the statistics of mixed Poisson-Gaussian noise, and a learned high-order MRF prior. We approximated the likelihood using general mixtures of Gaussians (MoG), which allows a further extension of the developed efficient block Gibbs sampler [50]. The degraded microscopy images are then restored by the MMSE estimate. Experiments using both natural images and fluorescence microscopy images demonstrate that our method can compete with state-of-the-art deconvolution approaches, and the joint computation of deconvolution, denoising and super-resolution is superior to a sequential scheme. Our method is also applied to microscopy images of telomeres acquired via stimulated emission depletion (STED) nanoscopy [57] for distinguishing different experimental conditions. The work was published in [58].

Non-rigid registration of live cell nuclei in time-lapse microscopy images. We introduce a global optical flow-based method to estimate the complex deformations of cell nuclei and perform non-rigid registration. The known properties of the deformation fields are represented by an MRF prior for determining the most suitable regularizer in the model. Based on studying the noise statistics in fluorescence microscopy images, we developed an adaptive weighting scheme to increase model robustness. We further extended the global model by exploiting high-order image features beyond the brightness. As the model formulations of FoEs and convolutional Gaussian RBMs are both consistent with the assumption of high-order feature constancy in the registration model, we learn filter banks with these generative MRF models for extracting high-order features of cell nuclei, thus achieving increased registration accuracy. Using multiple data sets of real 2D and 3D live cell microscopy image sequences as well as synthetic image data, we demonstrated that our proposed approach outperforms the previous methods in terms of both registration accuracy and computational efficiency. The work was published in [59, 60]. In addition, we integrated elasticity properties into the MRF prior to achieve more robust registration performance. This work was published in [61, 62].

1.4 Thesis overview

In Chapter 2, we review basic concepts of Markov random fields and previous important MRF image priors as well as their learning and inference strategies. We also review other generative image models and variational methods. Chapter 3 introduces a method for application-independent evaluation of image priors and new efficient learning approaches to achieve expressive high-order MRFs. We further explore MRFs with multimodal potential functions. Through an analytical comparison between MRFs and restricted Boltzmann machines, we reveal insights of MRFs and introduce an MRF for modeling visual textures. In Chapter 4, we study practical probabilistic inference approaches with the learned high-order image priors and present new methods for denoising of natural images, deconvolution and super-resolution of microscopy images, and non-rigid registration of cell nuclei in time-lapse microscopy images. In Chapter 5, we summarize the findings in this dissertation and discuss the limitations of current work. We also provide some thoughts on future work.

Chapter 2

Background and related work

In this chapter, we review basic concepts of Markov random fields and previous important MRF image priors as well as their learning and inference strategies. We also review other generative image models and variational methods. Related work on methods in particular applications (e.g., image restoration, image registration) is reviewed in Chapter 4.

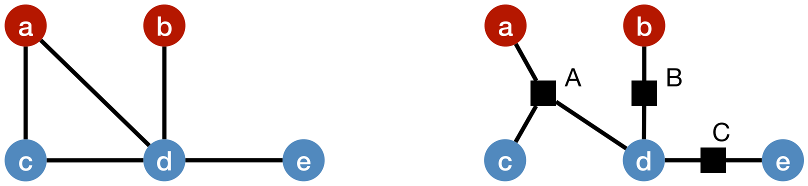
2.1 Graphical models and Markov random fields

A graphical model is a probabilistic model for formalizing the joint probability distribution over a set of random variables. It uses a graph to encode the independences between random variables and to represent how the joint distribution factorizes. The edges of the graph can be either directed or undirected, resulting in directed graphical models and undirected graphical models. The former is also known as Bayesian network and requires the graph to be acyclic (i.e., no directed cycles). But this restriction is generally too strong for modeling pixels in low-level vision. As a result, we will focus on the second class in our work.

Undirected graphical models. In an undirected graphical model, the edges are undirected and the graph may have arbitrary cycles (*cf*. Fig. 2.1(a)). The factorization depends on the cliques of the graph, which are defined as the subsets of nodes with full connection, resulting in the joint probability distribution formulation

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} g_c(\mathbf{x}_{(c)}), \tag{2.1}$$

where $\mathbf{x}_{(c)}$ denote the random variables of the clique $c \in \mathcal{C}$, associated with a potential function $g_c(\cdot) \in \mathbb{R}^+$ for assigning a positive score, and Z is the normalization or partition function (the potential functions g_c do not have to be normalized). This



(a) Undirected graphical model (Markov random field).

(b) Factor graph.

Figure 2.1: Two types of graphical models. Red nodes denote observed random variables, red nodes denote hidden random variables, and black squares are factor nodes.

distribution can be equivalently written in the form of a Gibbs distribution (or Gibbs random field).

While undirected graphical models are able to represent cyclic dependencies and can be seen as a generalization of the directed models, learning and inference of these models are difficult. For example, the connections between potential functions and marginal distributions of cliques are complex; the partition function is generally intractable and the estimation needs expensive computation. In addition, the generalization is not unique, as smaller cliques can be regarded as subsets of larger cliques. To make the factorization unique, maximal cliques are often required.

Markov random fields. A set of random variables $\mathbf{x} = (x_v)_{v \in V}$ and a neighborhood system $\mathcal{N}$ on these variables form a Markov random field (MRF) when the positivity and the Markovianity are both fulfilled [63], i.e., $p(\mathbf{x}) > 0$ and $p(x_v | \mathbf{x}_{V \setminus \{v\}}) = p(x_v | \mathcal{N}(v))$, where $V \setminus \{v\}$ is the set of all variables except v and $\mathcal{N}(v)$ is the set of neighbors of v. In an MRF, the Markovianity also means a variable x_v is conditionally independent of all other variables given the set of its neighbors $\mathcal{N}(v)$. According to the Hammersley–Clifford theorem [64], Markov random fields with a neighborhood system $\mathcal{N}$ are equivalent to Gibbs random fields when the edges follow the same neighborhood system. Therefore, Markov random fields and undirected graphical models are the same class of probability distributions.

Factor graphs. Factor graphs [65] use a bipartite graph with an additional set of nodes $f \in \mathcal{F}$ called factor nodes, to correspond to factors in the factorized joint distribution. Variable nodes are only connected to factor nodes (*cf*. Fig. 2.1(b)), thus the ambiguity of factorization can be removed. The joint probability distribution is given as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{f \in \mathcal{F}} g_f(\mathbf{x}_{(f)}), \tag{2.2}$$

where $\mathbf{x}_{(f)}$ are the random variables connected to factor f, and $g_f(\cdot) \in \mathbb{R}^+$ is the associated potential function. We will see in later sections that the factor graph formalism makes it easier to distinguish and understand different Markov random field image priors studied in this dissertation.

2.2 Markov random field image priors

Many problems in low-level vision are formulated using probabilistic models, particularly Markov random fields (MRFs), to impose prior knowledge [63]. In this section, we review the most prominent MRF image priors as well as related learning and inference approaches.

2.2.1 Pairwise and high-order MRFs

In an MRF image model, the nodes are arranged on a 2D lattice that corresponds to the spatial organization of the image or dense scene representation and each node corresponds to pixel intensities. Factor graphs allow a clear distinction between two types of MRF image priors: (1) pairwise Markov random fields (*cf*. Fig. 2.2(a)), in which factors (or maximal cliques) connect only pairs of nodes, and (2) high-order Markov random fields (*cf*. Fig. 2.2(b)), in which factors (or maximal cliques) connect more than two nodes.

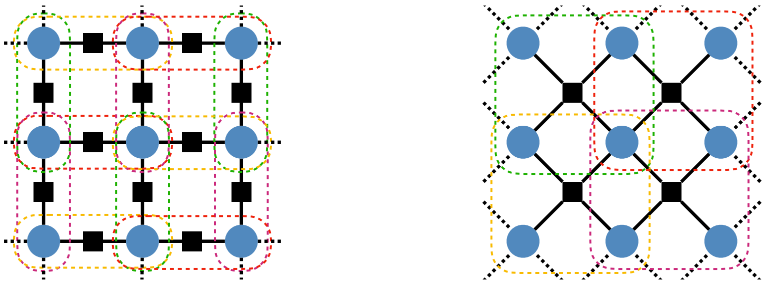
Pairwise MRFs

The probability density of an image $\mathbf{x}$ in a pairwise MRF is generally written as (e.g., [5, 46, 29])

$$p(\mathbf{x}) = \frac{1}{Z_{\text{pw}}} \prod_{i} \prod_{j \in \mathcal{N}(i), j>i} g_{\text{pw}}(x_j - x_i), \tag{2.3}$$

where $\mathcal{N}(i)$ denotes indices of the 4-neighbors of pixel i, and the condition $j > i$ ensures that no pixel pairs are counted twice. The potential g_{pw} is defined as a function of the intensity difference in a pixel pair and usually assumed to be the same at all spatial locations within the image grid, and thus the MRF model is homogeneous and translation-invariant. As can been seen, the pairwise MRF prior is formulated in terms of (the approximation of) first-order image derivatives; though being simple, it poses an important restriction on structures in images and scenes, thus being widely used for modeling prior knowledge of images. Actually, there is an equivalence between pairwise MRF priors and first-order derivative-based regularizers (also called smoothness terms) that are commonly used in variational methods (e.g., [66]; see also Sec. 2.3.3 for more extended review).

Regarding the potential functions, the most common and simplest choices are



(a) Pairwise MRF.

(b) High-order (2×2 cliques) MRF.

Figure 2.2: Typical pairwise and high-order Markov random field image priors illustrated as factor graphs. Colored dashed lines indicate image pixels connected to each factor node.

Gaussians (*cf*. Fig. 1.3(b)). Gaussian potentials (as quadratic regularizers in variatinal models) make the model inference relatively easy, but lead to smooth image discontinuities, which is obviously different than those in generic images. To preserve image discontinuities, robust potentials or regularizers are often chosen (*cf*. Fig. 2.5). The special statistics of image derivatives (*cf*. Fig. 1.3(a)) also motivate the parametric forms of the robust potentials. While the parameters of many pairwise MRF models still remain hand-tuned, there are some methods that learn the parameters from training data. For example, potentials based on generalized Laplacian distributions (*cf*. Fig. 1.3(c)) were trained in [5, 29], and mixture model-based potentials were trained in [31] to fit the image derivative marginals.

High-order MRFs

In pairwise MRFs, the small neighborhood with pairwise cliques limit the expressiveness of the model. A straightforward extension is to use larger neighborhoods (overlapping cliques), yielding a high-order MRF (*cf*. Fig. 2.2(b)). Difficulties for choosing an appropriate potential function arise as the potential needs to be defined over larger cliques. Since the potentials of pairwise MRFs are modeled based on first-order image derivatives, potentials of high-order MRFs can be analogously modeled based on higher-order image derivatives, which can be (approximately) extracted using linear filters. Usually a bank of linear filters are used and the potential function can be written as

$$g(\mathbf{x}_{(c)}) = \prod_j \exp\big(-\rho_j(\mathbf{f}_j^{\mathrm{T}}\mathbf{x}_{(c)})\big), \tag{2.4}$$

where $\mathbf{f}_j$ are linear (high-order derivative) filters with the same size as the cliques, and ρ_j are some (parametric or nonparametric) functions that model the filter responses. Geman and Reynolds [10] modeled gray value images with a high-order MRF based on 5 derivative filters (orders of $1, 2, 3$) of size 3×3 and robust regularization functions. Another important high-order MRF image prior is the FRAME model [8] that can be learned from training data. The potential functions are modeled using a flexible discrete, nonparametric representation. A bank of (much larger) linear filters consisting of Gabor filters of various orientations and scales is first manually chosen as candidates. The leaning procedure includes several rounds of greedily selecting a filter from the candidates using the minimum entropy criterion and adding it to the model, followed by training the model parameters using the maximum entropy criterion. While the FRAME model for visual textures [8] yields impressive results, the FRAME model for natural images [6] does not perform competitively in image restoration applications. As the FRAME model is particularly important in this dissertation, we will revisit and re-evaluate it for modeling generic images with today's learning and inference techniques in Sec. 3.4 and Sec. 4.1.

2.2.2 Fields of Experts

In high-order MRFs, besides the potential functions, the linear filters can also be learned from the training data. Roth and Black proposed the Fields of Experts (FoE) [7], in which the potentials and filters are learned jointly. The potential function in an FoE is formed through a combination of several "expert" functions defined on the filter responses

$$g_{\text{FoE}}(\mathbf{x}_{(c)}) = \prod_j \phi(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; \alpha_j), \tag{2.5}$$

where $\mathbf{f}_j$ are linear filters and $\phi(\cdot)$ are expert functions with parameter α_j. The term "expert" has its origin in the Products of Experts (PoE) framework [28]. The idea behind is that each expert is a simple model describing a particular property of the neighborhood (on a low-dimensional subspace), and the combination of several experts yields an expressive high-dimensional model. Note that the FoE model permits more experts than clique dimensions (over-complete), which allows modeling dependencies between filters.

Natural images and common dense scenes have heavy-tailed marginal distributions. For example, Fig. 1.3(a) shows the marginal log-histograms of x- and y-derivatives, which have a much stronger peak than a Gaussian with the same mean and variance, and very heavy tails that decline slowly away from the mean. Huang [67] showed that natural images exhibit heavy-tailed statistics not only for derivatives and wavelet coefficients, but also for the responses to random, zero-mean linear filters. The distinct heavy-tailed marginal distributions of images motivate the use of heavy-tailed experts. In [7], the experts are modeled using unnormalized Student-t distributions and have the form (*cf*. Fig. 1.3(d))

$$\phi(y; \alpha) = \left(1 + \frac{1}{2}y^2\right)^{-\alpha}, \tag{2.6}$$

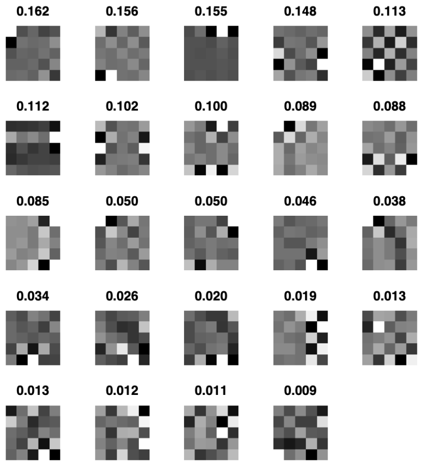


Figure 2.3: 5 × 5 filters obtained by training the Fields of Experts with Student-t experts on natural images [7]. The number above each filter denotes the corresponding expert parameter α_j.

where the parameter $\alpha \in \mathbb{R}^+$ controls the heavy-tailedness of the expert. Note that this continuous, parametric expert representation contrasts with that of the FRAME model [8].

Learning FoEs (i.e., filters $\mathbf{f}_j$ and expert parameters α_j) is very difficult as the model is highly non-convex and the partition function is intractable. The general learning strategy will be reviewed in next section. Fig. 2.3 shows the learned filters with Student-t experts as well as expert parameters on natural images in [7].

Weiss and Freeman [39] extended the FoEs by using Gaussian scale mixtures (GSMs) [32] to represent the experts. GSMs are weighted mixtures of Gaussian components with the same means but different variances (in terms of multiple scales), thus allowing a wider variety of heavy-tailed potentials to be represented, including the Student-t potential in [7] and generalized Laplacians in [5].

While these (learned) heavy-tailed potential (or expert) functions are all uni-modal (except for those of the FRAME models) and can well model the statistics of generic images, Heess *et al.* [9] showed that such potentials are not suitable for modeling textures. They further extended the Student-t expert in (2.6) to comprise both uni-modal and bi-modal cases and trained FoE priors for different visual textures.

2.2.3 Learning

Filter-based MRFs

Above, we have reviewed different pairwise and high-order MRFs (including FoEs). These models can be written in a more general form by

$$p_{\mathrm{MRF}}(\mathbf{x};\mathbf{\Omega}) = \frac{1}{Z(\mathbf{\Omega})}\prod_{c\in\mathcal{C}}\prod_{j}\phi(\mathbf{f}_j^{\mathrm{T}}\mathbf{x}_{(c)};\boldsymbol{\omega}_j), \tag{2.7}$$

where $Z(\mathbf{\Omega})$ is the partition function dependent on model parameters $\mathbf{\Omega}$, $\boldsymbol{\omega}_j$ are parameters for the expert or potential $\phi(\cdot)$ that models responses of filter $\mathbf{f}_j$, and $\mathcal{C}$ denotes all overlapping maximal cliques (overlapping image patches). We can further rewrite the MRF formulation using convolution [7]

$$p_{\mathrm{MRF}}(\mathbf{x};\mathbf{\Omega}) = \frac{1}{Z(\mathbf{\Omega})}\prod_{j}\prod_{i}\phi((\tilde{\mathbf{f}}_j * \mathbf{x})_i;\boldsymbol{\omega}_j). \tag{2.8}$$

Here "$*$" denotes the convolution operation. $\tilde{\mathbf{f}}_j$ is the convolution filter that is the spatially flipped version of filter $\mathbf{f}_j$ (simply by rearranging the coefficients). We use $(\cdot)_i$ to denote the ith pixel in the convolved image.

The pairwise MRF can be subsumed in this class of MRFs in (2.8) easily by defining two derivative filters $\mathbf{f}_1 = [-1, 1]$ and $\mathbf{f}_2 = \mathbf{f}_1^{\mathrm{T}}$ (*cf*. (2.3)).

The expression of MRFs with convolution not only allows efficient implementation, but also makes it easier to derive the learning and inference approaches. In this dissertation we also denote this model class as filter-based MRFs.

Maximum likelihood learning

Learning MRFs means the estimation of model parameters $\mathbf{\Omega}$ (including expert parameters and linear filters) from a set of R training images $\mathbf{X}^0 = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(R)}]$. While there is an extensive literature on learning methods for MRF models, here we focus on the popular maximum likelihood (ML) methods (e.g., used in [2, 30, 7, 9]). By assuming the training images $\mathbf{x}^{(n)}$ are independent, the learning objective is to maximize the log-likelihood

$$\mathcal{L}(\mathbf{X}^0;\mathbf{\Omega}) := \log p_{\mathrm{MRF}}(\mathbf{X}^0;\mathbf{\Omega}) = \sum_{n=1}^{R}\log p_{\mathrm{MRF}}(\mathbf{x}^{(n)};\mathbf{\Omega}). \tag{2.9}$$

As there is generally no closed form solution for $\mathbf{\Omega}$, gradient ascent on the log-likelihood can be performed. The partial derivative of the log-likelihood with respect to a parameter Ω_j is

$$\begin{aligned}\frac{\partial}{\partial\Omega_j}\mathcal{L}(\mathbf{X}^0;\boldsymbol{\Omega}) &= \sum_{n=1}^{R}\frac{\partial}{\partial\Omega_j}\log p(\mathbf{x}^{(n)};\boldsymbol{\Omega})\\ &= \sum_{n=1}^{R}\frac{\partial}{\partial\Omega_j}\Big(- E(\mathbf{x}^{(n)};\boldsymbol{\Omega}) - \log Z(\boldsymbol{\Omega})\Big) \qquad (2.10)\\ &= -\sum_{n=1}^{R}\frac{\partial E(\mathbf{x}^{(n)};\boldsymbol{\Omega})}{\partial\Omega_j} - \sum_{n=1}^{R}\frac{\partial \log Z(\boldsymbol{\Omega})}{\partial\Omega_j}\\ &= -R\cdot\left\langle\frac{\partial E(\mathbf{x};\boldsymbol{\Omega})}{\partial\Omega_j}\right\rangle_{\mathbf{X}^0} - R\cdot\frac{\partial\log Z(\boldsymbol{\Omega})}{\partial\Omega_j}. \qquad (2.11)\end{aligned}$$

E in (2.10) is the unnormalized Gibbs energy according to $p(\mathbf{x};\boldsymbol{\Omega}) = \frac{1}{Z(\boldsymbol{\Omega})}e^{-E(\mathbf{x};\boldsymbol{\Omega})}$ and $\langle\cdot\rangle_{\mathbf{X}^0}$ in (2.11) denotes expectation value (here equivalent to the average) over the training data $\mathbf{X}^0$. The second term in (2.11) can be further simplified as

$$\begin{aligned}\frac{\partial\log Z(\boldsymbol{\Omega})}{\partial\Omega_j} &= \frac{1}{Z(\boldsymbol{\Omega})}\cdot\frac{\partial Z(\boldsymbol{\Omega})}{\partial\Omega_j}\\ &= \frac{1}{Z(\boldsymbol{\Omega})}\cdot\frac{\partial}{\partial\Omega_j}\int e^{-E(\mathbf{x};\boldsymbol{\Omega})}d\mathbf{x}\\ &= \frac{1}{Z(\boldsymbol{\Omega})}\int -\frac{\partial E(\mathbf{x};\boldsymbol{\Omega})}{\partial\Omega_j}e^{-E(\mathbf{x};\boldsymbol{\Omega})}d\mathbf{x}\\ &= -\int\frac{\partial E(\mathbf{x};\boldsymbol{\Omega})}{\partial\Omega_j}p(\mathbf{x};\boldsymbol{\Omega})d\mathbf{x}\\ &= -\left\langle\frac{\partial E(\mathbf{x};\boldsymbol{\Omega})}{\partial\Omega_j}\right\rangle_{p(\mathbf{x};\boldsymbol{\Omega})}. \qquad (2.12)\end{aligned}$$

$\langle\cdot\rangle_p$ denotes the expectation with respect to the model distribution. Thus, the gradient ascent on the log-likelihood leads to the updates of parameters

$$\Omega_j^{(t+1)} = \Omega_j^{(t)} + \eta\left[\left\langle\frac{\partial E}{\partial\Omega_j}\right\rangle_{p(\mathbf{x};\boldsymbol{\Omega}^{(t)})} - \left\langle\frac{\partial E}{\partial\Omega_j}\right\rangle_{\mathbf{X}^0}\right], \qquad (2.13)$$

where η is the learning rate. One conceptual advantage is that this minimizes the Kullback-Leibler (KL) divergence between the model and the data distribution and, in principle, makes the model statistics as close as possible to those of the training data.

Approximate inference using sampling

The equations for computing partial derivatives of the unnormalized Gibbs energy of MRFs in (2.13) can be straightforwardly derived. The expectation over the training data is relatively easy to compute. Various difficulties, however, arise in practice, as there is no closed form expression for the model expectation, and an exact computation is intractable. Approximate inference must thus be used.

The expectation can be approximated using samples $\mathbf{x}_{\mathrm{s}}$ from the MRF model distribution

$$\left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{\mathbf{X}^{\mathrm{s}}} \approx \left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_p, \tag{2.14}$$

where $\mathbf{X}^{\mathrm{s}} = [\mathbf{x}_{\mathrm{s}}^{(1)}, \mathbf{x}_{\mathrm{s}}^{(2)}, \cdots]$ is a set of model samples. To draw samples from MRF models, Markov chain Monte Carlo (MCMC) techniques are commonly used. Despite being powerful, MCMC is often computationally expensive and slow (e.g., in the FRAME model [8]), because the dimensions of entire images are generally very high and MCMC often requires many iterations until the Markov chain converges to the target distribution. One strategy is to make MCMC sampling more efficient using data-driven proposal distributions (e.g., [68]) or domain-specific proposal distributions (e.g., [69]). A second strategy is to perform approximate maximum likelihood, e.g., contrastive divergence (CD) [33]. The samplers are initialized at the training data $\mathbf{X}^0$ and only run for a small, fixed number of MCMC iterations to yield the sample set $\mathbf{X}^{\mathrm{CD}}$, instead of running the Markov chain until convergence. $\mathbf{X}^{\mathrm{CD}}$ are then used as model samples to compute the expectation in (2.14). The interpretation is that samples $\mathbf{X}^{\mathrm{CD}}$ are closer to the (current) model distribution and the change is sufficient to estimate the parameter updates. Although CD was shown to be biased in many cases, the bias is very small in practice [70]. CD has been successfully applied to learning MRF models (e.g., FoEs in [7]) as well as a few related models such as Products of Experts (PoE) [30] and conditional random fields [71, 72].

The development of samplers for performing MCMC sampling is also an important issue. The Gibbs sampler [4] is a classical method commonly used in image modeling, which is performed by repeatedly sampling subsets of the random variables while keeping the other variables fixed. In [4, 8], single-site Gibbs sampling that sample each pixel conditioned on all other pixels was very inefficient, and many MCMC iterations were needed to reach the equilibrium distribution. More efficient block Gibbs samplers was developed for training the restricted Boltzmann machines [33] and the Products of Experts [2], which alternately sample a set of hidden variables given the random variables and vice versa, and thus achieve rapid mixing. For training the FoEs in [7, 9] as well as the PoE in [30], a Metropolis-based sampler, in particular, hybrid Monte Carlo (HMC) [73] was used. The HMC sampler uses the gradient of the log-density to explore the space more effectively and is sufficient for small images, but exhibits slow mixing for larger ones as sample dynamics have to be very small to admit a sufficiently high acceptance ratio.

2.2.4 Inference

Using MRF priors in applications is usually formulated as probabilistic inference of a posterior under a Bayesian framework. Given the observed data (or measurement) $\mathbf{y}$ and the application specific likelihood $p(\mathbf{y}|\mathbf{x})$, the posterior is written as

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}) \cdot p_{\text{MRF}}(\mathbf{x}). \tag{2.15}$$

The inference goal includes finding estimates for $\mathbf{x}$ and computing marginal distributions of the posterior.

The most common estimate for $\mathbf{x}$ is the maximum a-posteriori (MAP) solution. As there is generally no closed form solution for $\mathbf{x}$, a basic strategy is to perform gradient ascent on the log-posterior

$$\mathbf{x} \leftarrow \mathbf{x} + \tau\big(\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) + \nabla_{\mathbf{x}} \log p_{\text{MRF}}(\mathbf{x})\big), \tag{2.16}$$

where τ is the stepsize. According to the MRF formulation in (2.8), the expression for $\nabla_{\mathbf{x}} \log p_{\text{MRF}}(\mathbf{x})$ can be derived as

$$\nabla_{\mathbf{x}} \log p_{\text{MRF}}(\mathbf{x}) = \sum_j \mathbf{f}_j * \boldsymbol{\varphi}'\big(\tilde{\mathbf{f}}_j * \mathbf{x};\ \boldsymbol{\omega}_j\big), \tag{2.17}$$

where $\boldsymbol{\varphi}'(\cdot;\ \boldsymbol{\omega}_j)$ is the vector of derivatives of all log-experts $\log \phi(\cdot_i;\ \boldsymbol{\omega}_j)$. Convolution makes the computation of this term very efficient. When the likelihood is not very complex (e.g., for additive Gaussian noise removal), the inference algorithm is very easy to apply and implement. For example, in [7], conjugate gradient methods [74] were applied for MAP estimation for faster convergence in FoE models. However, since the unnormalized Gibbs energy of MRF models is often highly non-convex, gradient-based method typically only find local optima.

Methods for inference with graphical models are also a common choice in low-level vision, including graph cuts [38, 75, 43, 44] and belief propagation [76, 5, 77, 45]. These algorithms can find very good approximate (sometimes exact) solutions of non-convex optimization problems [43, 44]; but they are discrete models in contrast to the continuous interpretation of MRFs, and are difficult to be applied to models with larger cliques [46].

A number of methods have relied on sampling the posterior distribution to compute approximate MAP solutions. Geman and Yang [78] used posterior sampling for image restoration. Barbu and Zhu [69] developed an efficient Swendsen-Wang sampling scheme and applied it to segmentation and stereo matching. Kim *et al.* [79] used population-based MCMC methods for stereo matching. Moreover, sampling the posterior allows approximating the expectation of the posterior which is also called minimum mean squared error (MMSE) estimate. Fox and Nicholls [47] sampled the

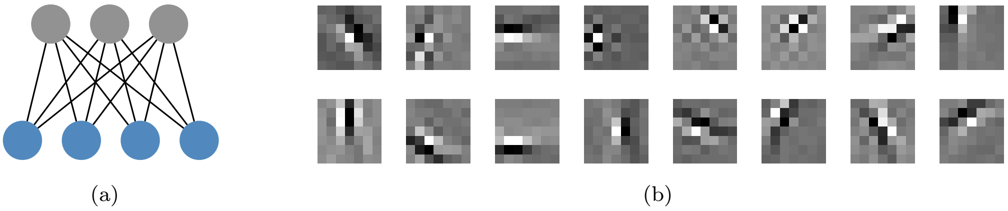


Figure 2.4: Restricted Boltzmann machines and learned filters. (a) Graph representation of a RBM with four visible units (blue) and three hidden units (gray). (b) Filters of 7 × 7 pixels obtained by training a convolutional Gaussian RBM [35] on natural images.

posterior of binary MRFs using perfect sampling, and found the MMSE estimate leading to more robust results.

To directly and more efficiently estimate the MMSE solutions, variational Bayesian methods, in particular, mean field approximation, have been used in multiple applications with high-order MRFs (e.g., [48, 49]), which yielded results with comparable accuracy to sampling methods.

2.3 Other related work

2.3.1 Restricted Boltzmann machines

The restrict Boltzmann machine (RBM) [33] is a generative stochastic neural network and can be represented as a bipartite graph, where there are two groups of nodes (“visible” and “hidden”) with only inter-group connections (*cf*. Fig. 2.4(a)). The RBM is a special case of Markov random fields [80], and its structure allows for efficient training algorithms such as contrastive divergence (CD) [33].

While the standard RBM has binary visible and binary hidden variables, it has been extended to model the density over continuous visible variables (keeping the hidden binary) [81], which makes the RBMs appropriate for modeling images at pixel level. Norouzi *et al.* [35] further applied the continuous RBM convolutionally to all overlapping cliques of a large image to obtain a generative image model. In terms of energy function, such a RBM image model is defined as

$$E_{\mathrm{RBM}}(\mathbf{x},\mathbf{h}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x} - \sum_{c\in\mathcal{C}}\sum_{j} h_{jc}\big(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_{(c)} + b_j\big), \tag{2.18}$$

where $\mathbf{x}$ are visible variables and denote image pixels, $h_{jc} \in \mathbf{h}$ are binary hidden variables, $\mathbf{w}_j$ determine the interaction between pairs of visible and hidden variables, and b_j are the biases. The (joint) distribution of this model can be defined as a Gibbs distribution according to the model energy[1]. As the conditional of $\mathbf{x}$ give the hidden variables $\mathbf{h}$ is a (multi-variate) Gaussian distribution with identity covariance, this continuous model is called convolutional Gaussian RBM (cGRBM).

Note that the cGRBM was not proposed for generic image priors, but to train a bank of filters (which are actually the interaction vectors $\mathbf{w}_j$ in (2.18)) for feature extraction in higher-level vision problems (e.g., object detection [35, 82]). Fig. 2.4(b) shows learned filters from a cGRBM on natural images. Due to the convolution structure in the model (*cf.* (2.8)), the learned cGRBM filters can capture shift-invariant image features.

2.3.2 Filters for image feature extraction

Filtering (or convolution) is an important image processing operation. A large number of computer vision applications rely on image features extracted using filters.

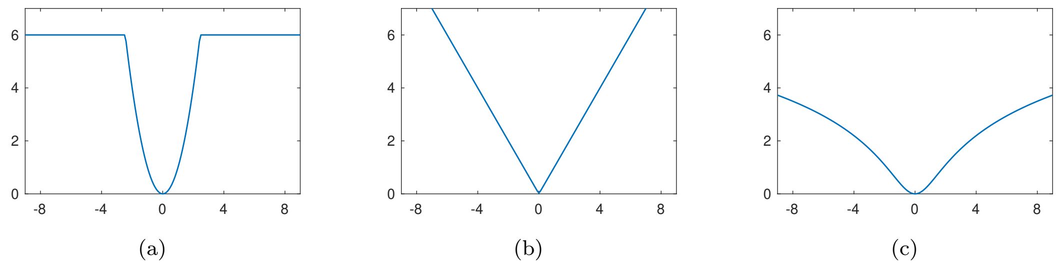
Gabor filters are an popular class of linear filters in image processing designed for frequency analysis and texture feature extraction, which are defined as Gaussian functions modulated by sinusoidal plane waves with different frequencies and directions, thus can be regarded as smooth derivative filters at various scales and orientations. Gabor filters have been used in MRF image models, e.g., the FRAME model [8] for modeling textures.

Using unsupervised feature learning methods, filters can be learned directly from training images. The common methods are patch-based, meaning that the filters are learned from a large number of image patches, and the size of each filter is the same as that of a training patch. These methods include K-means clustering, sparse coding[2] [84, 85], Products of Experts [2], mixtures of Gaussians [3], sparse RBMs [34], sparse auto-encoders [86, 87], and typically (when using whitened training data) yield localized filters, which resemble Gabor filters of various orientations, scales and locations. Note that these methods generally permit more filters than dimensions (number of pixels in a patch). This over-complete feature representation allows dependencies between filters being modeled and consequently is more expressive, resulting in better performance in applications (e.g., image classification) [88].

Filters can also be learned by training high-order generative MRFs (e.g., FoEs or convolutional RBMs, also [11]), which are shift-invariant due to the convolution structure of MRFs. It was shown in [7, 11] that the learned MRF filters are superior to those trained by patch-based models or engineered by hand in the context of image denoising with high-order MRFs. The learned MRF filters have also found widespread use across low-level and high-level vision problems such as denoising [89],

[1] The density for image $\mathbf{x}$ can be easily derived by marginalizing out the hidden variables $\mathbf{h}$.

[2] Principal component analysis (PCA) [83] and independent component analysis (ICA) [1] can be regarded as variations of sparse coding.



(a)

(b)

(c)

Figure 2.5: Common robust regularizers. (a) Truncated quadratic; (b) Charbonnier $\psi(y) = \sqrt{y^2 + \epsilon^2}$; (c) non-convex Lorentzian $\psi(y) = \log(1 + \frac{1}{2}y^2)$.

deconvolution [90], object detection [35], classification [91] to extract important image features.

2.3.3 Regularization in variational methods and deep models

In variational methods, the images or scene representations are regarded as functions of real-valued coordinates (e.g., $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$) and the goal is to determine $f(x, y)$ based on the observations $o(x, y)$ through minimizing the energy functional in the general formulation [92]

$$E(f; o) = \iint D(f, o) + \lambda \cdot \psi(\|\nabla f\|)\, dx\, dy, \tag{2.19}$$

where $D(f, o)$ is the data term that ensures a reasonable relation between the unknown f and the observation o. The second term imposes prior knowledge by penalizing large image gradients using a regularization function ψ, and λ is a weight to balance the two terms. To preserve image discontinuities (e.g., edges) in the results as those that appear in real images, robust regularizers (*cf*. Fig. 2.5) [93, 94, 95] have been used, which less penalize significant image gradients. Variational methods are also called regularization methods, and have continued to be popular since the early work for optical flow estimation [96] and image restoration [97].

Discretization of the variational energy functionals in space allows defining Gibbs random fields, and thus variational methods can be interpreted as Markov random fields [98]. It is easy to find the equivalences of the robust Charbonnier and Lorentzian functions in regularization methods to special cases of generalized Laplacian and Student-t potential functions in MRFs (*cf*. Fig. 1.3(c,d)), respectively. Actually some high-order MRFs (e.g., [11]) were trained using variational methods.

More recently variational methods have also been combined with deep neural networks. For example, non-rigid registration of medical images is a typical complex non-linear optimization problem, for which the deep models could be trained as efficient solvers. Convolutional neural networks (CNN) (in particular, patch-based models [99, 19], models with encoder-decoder structures [100, 20, 21]) for estimating the deformation fields were trained unsupervisedly, where energy functionals with some regularization (e.g., quadratic [99, 20, 19], total-variation (TV) [100], control function [21]) serve as the loss functions.

2.3.4 Deep generative models

Deep generative image models use neural networks with multiple hidden layers to approximate high-dimensional probability distributions of images either explicitly or implicitly. They are trained unsupervisedly and can be regarded as image priors. Below, we review few typical deep generative image models.

In PixelRNN and PixelCNN [24], the tractable probability density of images $p(\mathbf{x})$ is factorized into a product of conditional distributions

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \,|\, x_1, \ldots, x_{i-1}), \tag{2.20}$$

where the image $\mathbf{x} \in \mathbb{R}^N$ is written as a one-dimensional sequence $(x_1, \ldots, x_N)$ and $p(x_i \,|\, x_1, \ldots, x_{i-1})$ is the distribution of the i-th pixel x_i given all previous pixels $(x_1, \ldots, x_{i-1})$. These complex conditional distributions are modeled by a recurrent neural network (RNN) or a convolutional neural network (CNN). Such deep generative models allow explicitly computing the likelihood of images.

Variational Auto-encoders (VAE) [25, 101] define the intractable density of images with latent variables $\mathbf{z}$

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}, \tag{2.21}$$

where θ denotes the model parameters. VAEs are used to generate image samples with a decoder network given the latent variables that can be sampled more efficiently. Training of VAEs is performed by including an encoder network and maximizing the variational lower-bound of the likelihood.

Generative Adversarial Networks (GAN) [26, 27], rather than modeling the image density, take a different strategy to generate image samples from the training distribution. Through a two-player game, GANs simultaneously learn two networks, a generator and a discriminator: The former tries to cheat the latter by generating more realistic images, and the latter try to distinguish between real and generated images. Compared to VAEs, GANs can generate high quality image samples, but are generally more difficult to train due to the minimax optimization objective.

The main aim of these deep generative models is to capture the high-level image structures and generate realistic samples for various purposes such as artwork, super-resolution, and colorization. However, it is hard to use them as a versatile image prior in a variety of image analysis problems (e.g., with a generative setting in a Bayesian framework).

Chapter 3

Learning and evaluation of MRF image priors

Both learning and evaluation of MRF priors for images are challenging problems. In this chapter, we first develop a new efficient auxiliary-variable Gibbs sampler for a general class of MRFs with flexible Gaussian scale mixture (GSM) potentials. This enables us to propose an application-independent method to quantitatively evaluate various common probabilistic MRF priors using model samples. After showing that previous MRF priors only crudely capture image statistics, we further present new strategies for more effectively training MRF priors that can accurately capture key statistical properties of natural images. An extension of the potential functions to the multi-modal case allows to revisit the seminal FRAME model [6] and obtain more in-depth understanding of the learned MRF potential functions. In the last section, we analyze the "mean" and "covariance" units in MRF models and introduce a compact, "mean"-only MRF for efficiently modeling image textures.

The development of the Gibbs sampler and the model evaluation scheme were published in [50]. New MRF learning strategies and results were published in [51]. The extension to multi-modal potentials and revisiting the FRAME model are unpublished work. The analysis and training of MRFs for textures were published in [53].

3.1 Developing an efficient Gibbs sampler

3.1.1 Flexible MRF model

Rather than proposing a new prior, we rely on the filter-based, high-order MRF model Fields of Experts (FoE) [7], whose clique potentials model the responses to a bank of linear filters. The prior probability density of an image $\mathbf{x} \in \mathbb{R}^N$ under the

FoE can be written as

$$p(\mathbf{x};\boldsymbol{\Omega}) = \frac{\mathcal{N}_\epsilon(\mathbf{x})}{Z(\boldsymbol{\Omega})} \prod_{c\in\mathcal{C}} \prod_{j=1}^{J} \phi(\mathbf{f}_j^{\mathrm{T}}\mathbf{x}_{(c)};\boldsymbol{\omega}_j), \tag{3.1}$$

where $\mathbf{x}_{(c)}$ are the pixels of clique $c \in \mathcal{C}$, $\mathbf{f}_j$ are the linear filters and $\phi(\cdot;\boldsymbol{\omega}_j)$ is the respective potential function (or expert, depending on the context) with parameters $\boldsymbol{\omega}_j$. $Z(\boldsymbol{\Omega})$ is the partition function that depends on the model parameters $\boldsymbol{\Omega} = \{\mathbf{f}_j, \boldsymbol{\omega}_j \,|\, j=1,\ldots,J\}$. A very broad unnormalized Gaussian $\mathcal{N}_\epsilon(\mathbf{x}) = e^{-\epsilon\|\mathbf{x}\|^2/2}$ with $\epsilon = 10^{-8}$ is used to guarantee the model to be normalizable even when the potential functions do not fully constrain image pixels [39].

Apart from a variety of FoE models [7, 102, 39], this general class of MRFs in (3.1) subsumes popular pairwise MRF models as well (e.g., [46, 29, 5]). For the pairwise case we define a single fixed filter $\mathbf{f}_1 = [-1, 1]^{\mathrm{T}}$ and let the maximal cliques $\mathcal{C}$ be all pairs of horizontal and vertical neighbors.

Following [39], we use flexible Gaussian scale mixtures (GSM) [32] to represent the potentials. In their finite form they can be written as

$$\phi(\mathbf{f}_j^{\mathrm{T}}\mathbf{x}_{(c)};\boldsymbol{\omega}_j) = \sum_{k=1}^{K} \omega_{jk} \cdot \mathcal{N}(\mathbf{f}_j^{\mathrm{T}}\mathbf{x}_{(c)}; 0, s_k \cdot \sigma_j^2), \tag{3.2}$$

where $\omega_{jk} \geq 0$, $\sum_k \omega_{jk} = 1$ are the mixture weights of the Gaussian components with scales s_k and base variance σ_j^2. GSMs have the advantage that they allow a wide variety of heavy-tailed potentials to be represented, including the Student-t function [7] and generalized Laplacians [5], and are yet computationally relatively easy to deal with.

3.1.2 Auxiliary-variable Gibbs sampler

The partition function $Z(\boldsymbol{\Omega})$ of the MRF priors is intractable, which results in the infeasibility of exact computations of model statistics. As sampling is a distinguished feature of generative models, the statistical properties can be analyzed through samples. In order to make it practical, an efficient sampler, also as a general inference engine for probabilistic models, is required.

Markov chain Monte Carlo (MCMC) is largely the only choice to sample MRFs. Single-site Gibbs samplers [4, 6] are very inefficient, as they need many iterations to reach the equilibrium distribution. Other Metropolis-based samplers, such as hybrid Monte Carlo [7], are sufficient for small images, but exhibit slow mixing for larger ones as sample dynamics have to be very small to admit a sufficiently high acceptance ratio.

Here, we take a different route and exploit that our potentials are represented using Gaussian scale mixtures. In the context of Products of Experts [33], Welling *et al.* [2] showed that it is beneficial to retain the scales of the GSM as an explicit discrete-valued hidden random vector $\mathbf{z} \in \{1, \dots, K\}^J$, one scale for each expert. Similar to a regular mixture model, one can define a joint distribution $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\Omega})$ of $\mathbf{x}$ and the auxiliary mixture coefficients $\mathbf{z}$ such that $\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\Omega}) = p(\mathbf{x}; \boldsymbol{\Omega})$. [2] showed that this allows defining a rapidly mixing auxiliary-variable Gibbs sampler that alternates between sampling $\mathbf{z}^{(t+1)} \sim p(\mathbf{z}|\mathbf{x}^{(t)}; \boldsymbol{\Omega})$ and $\mathbf{x}^{(t+1)} \sim p(\mathbf{x}|\mathbf{z}^{(t+1)}; \boldsymbol{\Omega})$, where t denotes the current iteration. If one only cares about obtaining samples of $\mathbf{x}$, the coefficients $\mathbf{z}$ can later be discarded. Similar ideas have been pioneered by Geman and Yang [78] in the context of MRFs. Levi and Weiss [42, 103] showed that this general framework can also be applied to MRFs with arbitrary Gaussian mixture potentials.

We rewrite the model density in (3.1) and (3.2) as

$$p(\mathbf{x}; \boldsymbol{\Omega}) = \sum_{\mathbf{z}} \frac{\mathcal{N}_\epsilon(\mathbf{x})}{Z(\boldsymbol{\Omega})} \prod_{c \in \mathcal{C}} \prod_{j=1}^{J} p(z_{jc}) \cdot \mathcal{N}(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; 0, s_{z_{jc}} \sigma_j^2), \tag{3.3}$$

where the scales $\mathbf{z} \in \{1, \dots, K\}^{J \times |\mathcal{C}|}$ for each potential function and clique are treated as random variables with $p(z_{jc}) = \omega_{j z_{jc}}$ (i.e., the GSM mixture weights). Instead of marginalizing out the scales, we retain them explicitly and define the joint distribution (*cf*. [2])

$$p(\mathbf{x}, \mathbf{z}; \boldsymbol{\Omega}) = \frac{\mathcal{N}_\epsilon(\mathbf{x})}{Z(\boldsymbol{\Omega})} \prod_{c \in \mathcal{C}} \prod_{j} p(z_{jc}) \cdot \mathcal{N}(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; 0, s_{z_{jc}} \sigma_j^2). \tag{3.4}$$

The conditional distribution $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\Omega})$ can be derived as a multivariate Gaussian

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}; \boldsymbol{\Omega}) &\propto \mathcal{N}_\epsilon(\mathbf{x}) \cdot \prod_{c \in \mathcal{C}} \prod_{j} \exp\left\{ -\frac{(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)})^2}{2 s_{z_{jc}} \sigma_j^2} \right\} \\ &\propto \mathcal{N}\left(\mathbf{x}; \mathbf{0}, \left(\epsilon \mathbf{I} + \sum_{c \in \mathcal{C}} \sum_{j} \frac{\mathbf{w}_{jc} \mathbf{w}_{jc}^{\mathrm{T}}}{s_{z_{jc}} \sigma_j^2} \right)^{-1} \right), \end{aligned} \tag{3.5}$$

where $\mathbf{w}_{jc}$ is defined as $\mathbf{w}_{jc}^{\mathrm{T}} \mathbf{x} = \mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}$ which is the result of applying the filter $\mathbf{f}_j$ to clique c of the image $\mathbf{x}$.

Difficulties of sampling the Gaussian in (3.5) arise from the fact that the (inverse) covariance matrix is huge for large images, which prevents an explicit Cholesky decomposition as in [2]. Levi and Weiss [42, 103] showed that this can be avoided by rewriting the covariance as

$$\mathbf{\Sigma} = \left(\epsilon\mathbf{I} + \sum_{c\in\mathcal{C}}\sum_{j}\frac{\mathbf{w}_{jc}\mathbf{w}_{jc}^{\mathrm{T}}}{s_{z_{jc}}\sigma_j^2}\right)^{-1} = \left(\mathbf{W}\mathbf{\Lambda}\mathbf{W}^{\mathrm{T}}\right)^{-1} \tag{3.6}$$

with $\mathbf{\Lambda}$ being a diagonal matrix and obtaining a sample $\mathbf{x}$ from $p(\mathbf{x}|\mathbf{z};\mathbf{\Omega})$ by solving a least-squares problem

$$\mathbf{W}\mathbf{\Lambda}\mathbf{W}^{\mathrm{T}}\mathbf{x} = \mathbf{W}\sqrt{\mathbf{\Lambda}}\mathbf{y}, \tag{3.7}$$

where the vector $\mathbf{y}$ is sampled from a unit normal $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Solving this sparse linear system of equations is much more efficient than a Cholesky decomposition. The advantage over single-site Gibbs samplers [6] 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 and fast convergence to the equilibrium distribution.

Since the scales $\mathbf{z}$ are conditionally independent given the image, the conditional distribution $p(\mathbf{z}|\mathbf{x};\mathbf{\Omega})$ is given as

$$p(z_{jc}|\mathbf{x};\mathbf{\Omega}) \propto \omega_{jz_{jc}} \cdot \mathcal{N}(\mathbf{f}_j^{\mathrm{T}}\mathbf{x}_{(c)}; 0, s_{z_{jc}}\sigma_j^2) \tag{3.8}$$

and sampling this discrete distribution is straightforward.

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. 3.1), a more rigorous procedure for monitoring convergence is still desirable. We use the popular approach by Gelman and Rubin [104], which relies on running several Markov chains in parallel and initializing them at different over-dispersed starting points. By computing the within-sequence variance σ_w^2 and the between-sequence variance σ_b^2 of a scalar estimand (here, the model energy), one can monitor convergence by estimating the potential scale reduction

$$\hat{R} = \sqrt{\frac{(n-1)\sigma_w^2 + \sigma_b^2}{n\sigma_w^2}}, \tag{3.9}$$

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 [104] for details.

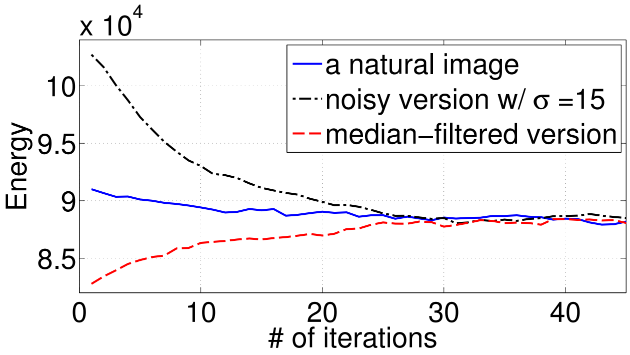


Figure 3.1: Fast mixing of the Gibbs sampler when sampling an image of 100×100 pixels from the learned pairwise MRF prior. Three chains are initialized with dispersed starting points (an image and processed versions). Approximate convergence is reached after 28 iterations ($\hat{R} < 1.1$).

3.2 Model evaluation via generative properties

Here we revive and extend the methodology of Zhu and Mumford [6], 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-independent way of evaluating MRFs.

3.2.1 Evaluation based on marginal statistics

According to [8], the filter-based MRFs (cliques corresponding to the filters) discussed in this thesis take the form of a maximum entropy distribution, so the statistics of the in-build features (filter responses) should be preserved. Note that [6] only analyzed the derivative statistics and did not perform quantitative measurements. We here propose to analyze generative models regarding these properties and to quantitatively measure the Kullback-Leibler (KL) divergences between the discretized marginal distributions of images with respect to filters $P_{\hat{\mathbf{x}}}$ and those of model samples $P_{\hat{\mathbf{s}}}$:

$$D_{\text{KL}} = \sum_{k \in \mathcal{K}} P_{\hat{\mathbf{s}}}(k) \cdot \ln \frac{P_{\hat{\mathbf{s}}}(k)}{P_{\hat{\mathbf{x}}}(k)}, \tag{3.10}$$

where $\mathcal{K}$ denotes the probability space. The lower the KL-divergence values, the better the model captures the statistical properties of images.

Minimizing the KL-divergence between the model and the data distribution is equivalent to maximizing the likelihood of the model to the data. While the most popular learning objective for image priors is maximum likelihood (ML), we have also developed a direct and quantitative measurement for evaluating the effectiveness of learning.

3.2.2 Conditional sampling

In order to avoid extreme values at the less constrained boundary pixels [35] during model analysis, we rely on conditional sampling. In particular, we sample the pixels $\mathbf{x}_A \in \mathbb{R}^{N_A}$ given fixed pixels $\mathbf{x}_B \in \mathbb{R}^{N_B}$, $N_A + N_B = N$, and scales $\mathbf{z}$ according to the conditional Gaussian distribution

$$p(\mathbf{x}_A | \mathbf{x}_B, \mathbf{z}; \boldsymbol{\Omega}), \tag{3.11}$$

where A and B denote the index sets of the respective pixels. Without loss of generality, we assume that

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix}, \quad \boldsymbol{\Sigma} = \left(\mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^{\mathrm{T}}\right)^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^{\mathrm{T}} & \mathbf{B} \end{bmatrix}^{-1}, \tag{3.12}$$

where the square sub-matrices $\mathbf{A}$ and $\mathbf{B}$ have the dimensions $N_A \times N_A$ and $N_B \times N_B$, respectively. The conditional distribution of interest can now be derived as

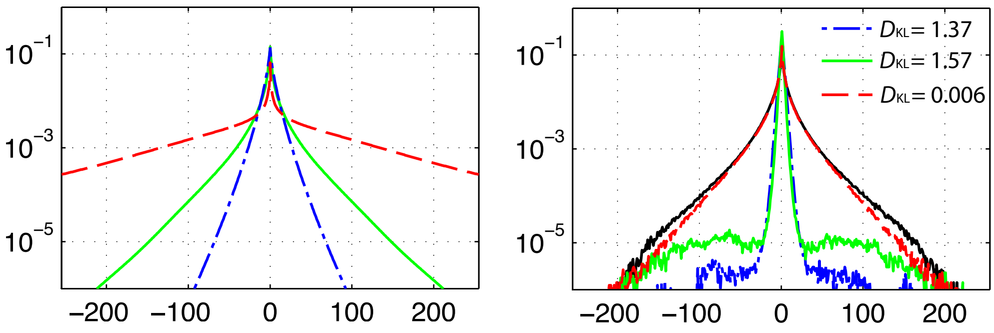
$$\begin{aligned} p(\mathbf{x}_A | \mathbf{x}_B, \mathbf{z}; \boldsymbol{\Omega}) &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^{\mathrm{T}} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix} \right) \\ &\propto \exp\left(-\frac{1}{2} \left(\mathbf{x}_A + \mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B\right)^{\mathrm{T}} \mathbf{A} \left(\mathbf{x}_A + \mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B\right) \right) \\ &\propto \mathcal{N}\left(\mathbf{x}_A; -\mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B, \mathbf{A}^{-1}\right). \end{aligned} \tag{3.13}$$

The matrix $\mathbf{A}$ is given by the appropriate sub-matrices of $\mathbf{W}$ and $\boldsymbol{\Lambda}$, and allows for using the same efficient sampling scheme. The mean $\boldsymbol{\mu} = -\mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B$ of the Gaussian can also be computed by solving a least-squares problem $\mathbf{A}\boldsymbol{\mu} = -\mathbf{C}\mathbf{x}_B$. Sampling the conditional distribution of scales $p(\mathbf{z}|\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\Omega}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\Omega})$ remains as before.

3.2.3 Properties of common 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 (3.2) to the target potential by minimizing their KL-divergence through gradient-based nonlinear optimization of the weights ω_{jk}. We achieve very good fits through a wide range of different potential shapes ($D_{\mathrm{KL}} < 0.0005$).

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 [105], 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 pixels. To avoid boundary artifacts during sampling, we follow [35] and condition on fixed image boundaries from a separate set of 3000



(a) potentials

(b) derivative marginals

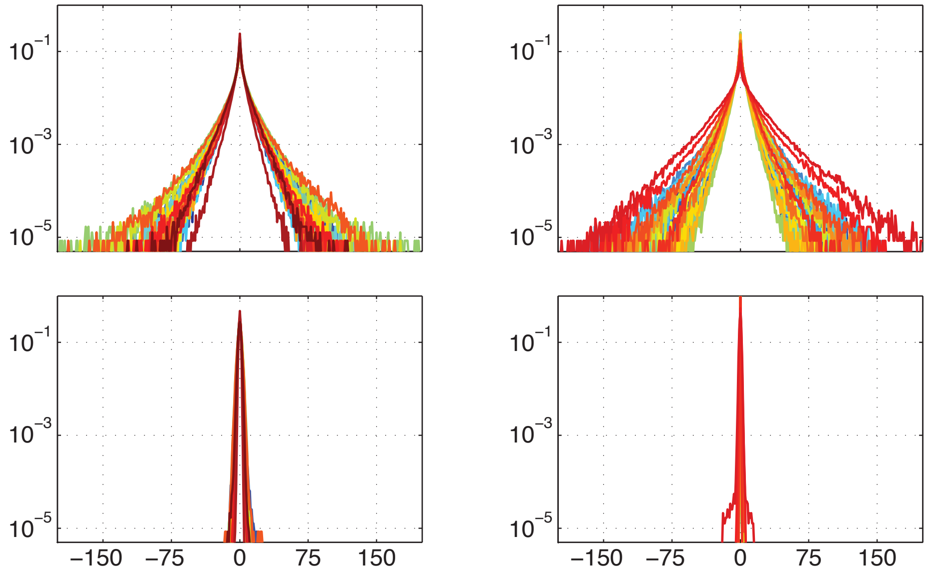
Figure 3.2: Pairwise MRFs. (a) Potentials: generalized Laplacian [5] (blue, dash-dotted), fit of the marginals [31] (green, solid), and proposed flexible potential (red, dashed). (b) Derivative histograms of samples from corresponding MRFs, and statistics of natural images (black, solid). Typical models [31, 5] lead to an incorrect representation of image statistics.

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. 3.2), which motivates the use of heavy-tailed potentials with shapes similar to the empirical derivative statistics [46, 29, 5]. It has also become common to fit potential functions directly to the derivative histogram [31, 39]. 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 [29, 5] (here, $\beta = 0.5, \gamma = 0.7$). We also consider GSM potentials that have been directly fitted to the empirical marginals, similar to [31, 39]. As Fig. 3.2 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



(a) FoE of Roth and Black [7], avg. $D_{\mathrm{KL}} = 2.19$

(b) FoE of Weiss and Freeman [39], avg. $D_{\mathrm{KL}} = 5.26$

Figure 3.3: Filter statistics of natural images (top) and filter marginals of FoE models (bottom). (Filters are normalized for ease of display.)

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*. [6]), the shape of the potential function is the only cause of the problem.

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 [6] 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 [7, 102, 39] differ by their parametric expert functions and learned filters, and have shown to be among the best-performing image priors. We analyze their generative properties by inspecting 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 [7] and the GSM-based FoE model of [39]. The study of natural images has found that even arbitrary zero-mean filters have heavy-tailed statistics [67], which also holds for the learned filters (see Fig. 3.3, top). We confirm and extend the findings of Levi [42], that the original FoE model [7] does not capture the filter statistics (Fig. 3.3, 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 [39] 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 [6] that constrain the statistics of the bank of learned filters, this means either the parametric form of the experts or the learning procedure is insufficient.

3.2.4 Visual inspection of model samples

Although visual inspection is not very suitable for evaluating the "quality" of model samples, it can provide some intuition about which image structures a model captures. Fig. 3.4 shows three subsequent samples (after reaching the equilibrium distribution) from the common MRFs discussed above in Sec. 3.2.3. It can been seen that samples from common pairwise models appear too "grainy", while those from previous FoE models are too smooth and without discontinuities.

3.3 Learning comprehensive MRFs

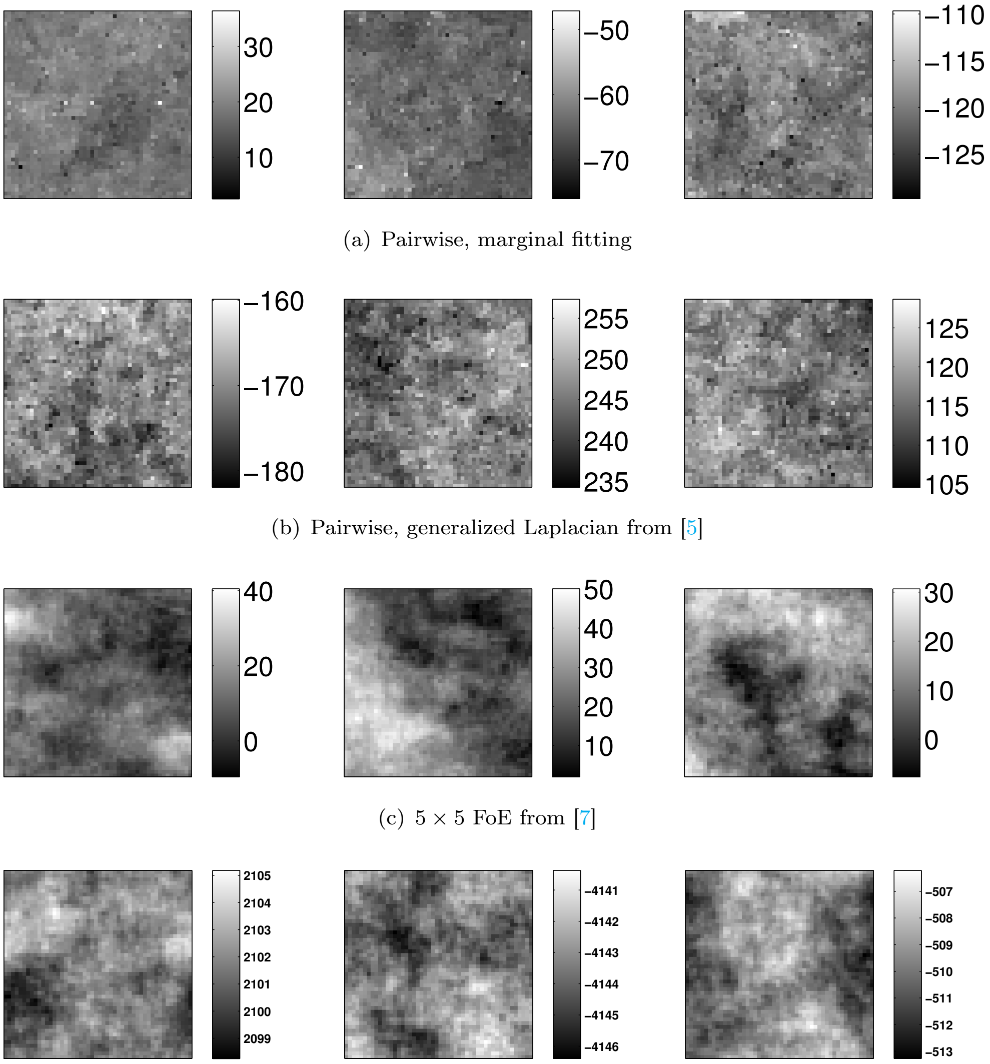
We have shown that popular pairwise and high-order MRFs are quite poor generative models and can not even capture the statistics of the inbuilt features. This appears in contradiction to the maximum entropy interpretation of filter-based MRFs [6], which means that the potential functions of these models are chosen or learned inappropriately. As the Gaussian scale mixtures (GSMs) are flexible to represent a wide variety of heavy-tailed potentials, in this section we investigate the model learning procedures.

3.3.1 Standard learning procedure

Learning the model parameters $\boldsymbol{\Omega}$ from data involves estimating the weights ω_{jk} of the GSM, and in case of Fields of Experts also the filters $\mathbf{f}_j$. The classical learning objective for training models of natural images is maximum likelihood (ML). A gradient ascent on the log-likelihood for a parameter Ω_i leads to the update

$$\Omega_i^{(t+1)} = \Omega_i^{(t)} + \eta \left[\left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_p - \left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{\mathbf{X}^0} \right], \tag{3.14}$$

where E is the unnormalized Gibbs energy according to $p(\mathbf{x};\boldsymbol{\Omega}) = e^{-E(\mathbf{x};\boldsymbol{\Omega})}/Z(\boldsymbol{\Omega})$, η is the learning rate, $\langle\cdot\rangle_{\mathbf{X}^0}$ denotes the average over the training data $\mathbf{X}^0$, and $\langle\cdot\rangle_p$ denotes the expectation value w.r.t. the model distribution $p(\mathbf{x};\boldsymbol{\Omega}^{(t)})$.



(a) Pairwise, marginal fitting

(b) Pairwise, generalized Laplacian from [5]

(c) 5 × 5 FoE from [7]

(d) 15 × 15 FoE from [39] (no pixels removed due to circular boundary handling)

Figure 3.4: Three subsequent samples (left to right) from various MRF models after reaching the equilibrium distribution. The boundary pixels are removed for better visualization.

One conceptual advantage is that (3.14) minimizes the Kullback-Leibler (KL) divergence between the model and the data distribution and, in principle, makes the model statistics as close as possible to those of natural images, which means a maximum entropy model. Various difficulties, however, arise in practice. First, there is no closed form expression for the model expectation, and exact computation is intractable. Approximate inference, e.g., using sampling, must thus be used. Markov chain Monte Carlo (MCMC) approximations are historically most common (e.g., [6]), but very inefficient. Consequently, ML estimation itself is frequently approximated by contrastive divergence (CD) [33], which avoids costly equilibrium samples: Samplers are initialized with the training data $\mathbf{X}^0$ and only run for n (usually a small number) MCMC iterations to yield the sample set $\mathbf{X}^n$. Then $\langle \partial E / \partial \Omega_i \rangle_{\mathbf{X}^n}$ is used to replace $\langle \partial E / \partial \Omega_i \rangle_p$ in (3.14). A second challenge is the speed of mixing, which is usually addressed with efficient sampling methods. For a detailed review of the standard maximum likelihood learning procedure, readers are refereed to Sec. 2.2.3.

3.3.2 Best practices for learning

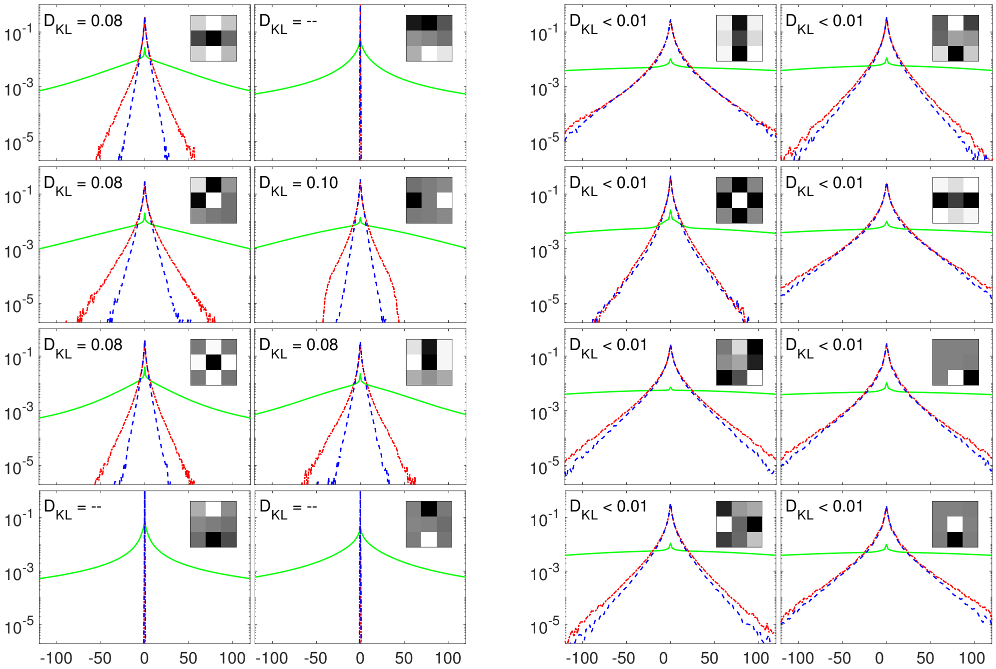
While the standard learning procedure with the efficient auxiliary-variable Gibbs sampling (see Sec. 3.1.2) allows to learn simple pairwise MRFs that capture derivative statistics correctly (*cf*. Fig. 3.2), the high-order case (e.g., FoEs) is more problematic (*cf*. Fig. 3.5(a)): Marginal statistics of model samples were found not as heavy-tailed as those of natural images [50]. Moreover, the models are limited to moderate clique sizes and a comparatively small number of filters.

We here show that such a standard learning procedure is insufficient to learn accurate high-order MRF models of natural images, and propose best practices for an improved learning scheme.

PCD instead of CD

Although contrastive divergence (CD) is a reasonably good and formally justified approximation of maximum likelihood [33], it may still incur a training bias. While using n-step CD (with large n) may reduce the bias, learning becomes much less efficient. We instead use *persistent contrastive divergence* (PCD) [106], in which the samplers are not reinitialized each time the parameters are updated. Contrastingly, the samples from the previous iteration are retained and used for initializing the next iteration. Combined with a small learning rate, the samplers are held close to the stationary distribution:

$$\left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{\mathbf{X}^{\text{PCD}}} \approx \left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{\mathbf{X}^{\infty}} \approx \left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_p . \tag{3.15}$$



(a) 3×3 FoE trained with standard learning procedure.

(b) 3×3 FoE trained with improved learning procedure.

Figure 3.5: Filter-based MRFs and image statistics: Image filter (top right) with corresponding learned potential function (solid, green), marginal histograms of natural images (dash-dotted, red), and model marginals obtained by sampling (dashed, blue). The proposed learning scheme leads to a better match to natural image statistics (top left, marginal KL-divergence).

Thus each parameter update closely approximates a true ML update step as in (3.14). Even with a small learning rate, PCD has an efficiency comparable to that of 1-step CD, but substantially reduces bias as the experiments below show. Note that while PCD has been used to train restricted Boltzmann machines [106] and filter-based MRFs with Student-t potentials [107], this is the first time that PCD has been investigated in conjunction with more flexible GSM potentials.

Replacing CD with PCD not only reduces training bias, but more importantly improves the models' properties significantly. To demonstrate this, we use a 1-step CD trained 3×3 FoE [50] as a basis and retrain the potentials with PCD, while keeping the filters fixed. The resulting marginal statistics of the inbuilt model features match those of natural images well; all marginal KL-divergences are below 0.01. Fig. 3.6 shows in detail the parts where PCD affects the potential shape the most and most improves the resulting model marginal. Another notable benefit of using PCD is that it enables the following improved boundary handling scheme.

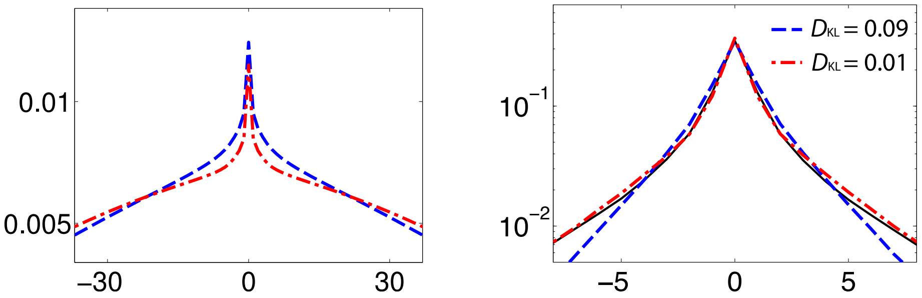


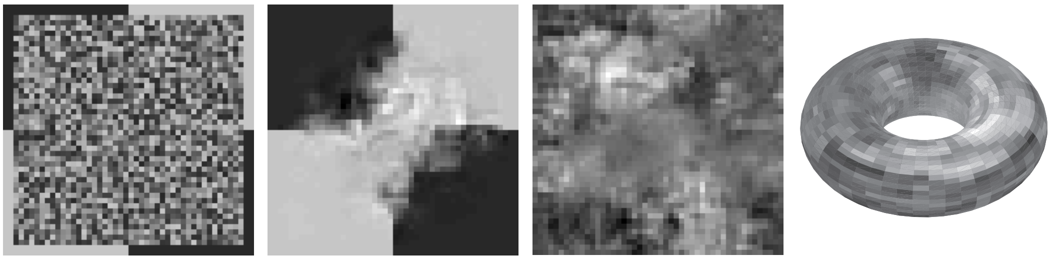
Figure 3.6: Difference between (*left*) potentials trained with 1-step CD (dashed, blue) and PCD (dash-dotted, red), as well as (*right*) resulting model marginals (magnified for display). The marginal KL-divergence is given w.r.t. natural images (solid, black).

Boundary handling

Boundary pixels are a common source of problems in MRFs, since they are overlapped by fewer cliques, making them less constrained than those in the image interior. When sampling the model, boundary pixels of the samples tend to take extreme values, which affects both learning and analysis of the model through sampling. Norouzi *et al.* [35] proposed to use conditional sampling, i.e. keeping a small number of pixels around the boundary fixed and conditionally sampling the interior. The drawback of this scheme is that the boundary pixels will significantly diffuse into the interior during sampling, which can be seen from the example in Fig. 3.7(a,b). To reduce bias in learning and evaluation of the model, a thick boundary from the samples thus has to be discarded. The disadvantage is that this lowers the accuracy and the efficiency of learning.

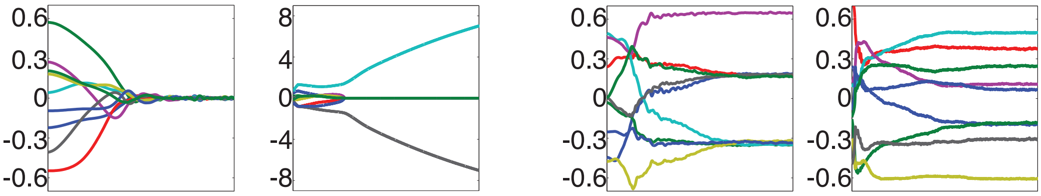
To address this, we propose toroidal sampling, in which the cliques are extended to also cover the boundary pixels in a wrap-around fashion. The toroidal topology used during sampling is shown in Fig. 3.7(d). The obvious benefit of using this topology is the absence of any boundary pixels; all pixels are overlapped by the same number of cliques, and there are as many cliques as pixels. Since all pixels are constrained equally, boundary artifacts are avoided, and bias from the boundaries during learning is avoided.

Fig. 3.7(c) shows how toroidal sampling is less affected by its initialization and can quickly explore a large space. The generated samples will in turn make learning more accurate, while not requiring boundary pixels to be discarded. This increases the learning efficiency, because fewer parallel samplers suffice to estimate the likelihood gradient accurately. It is important to note that while PCD allows using toroidal sampling, the more common CD does not. This is because CD repeatedly initializes the samplers from the training images, which usually do not satisfy periodic boundary conditions.



(a) (b) (c) (d)

Figure 3.7: Effect of boundary handling on samples: (a) Initialization of the sampler; (b) typical sample generated by conditional sampling (note how the boundaries affect the interior of the sample); (c, d) typical sample and its topology generated by the proposed toroidal sampling.



(a) Without normalization (b) With normalization

Figure 3.8: Filter coefficients may decay or disperse during learning (a). This problem can be solved with filter normalization (b).

Filter normalization

Some researchers (e.g. [40, 2]) have suggested to impose constraints on the norms of filters, because filters may otherwise become "inactive", i.e. decay to zero during training. As zero filters and the corresponding potentials do not contribute to the model at all, this is an issue, especially when a large number of filters are trained [40, 2]. But even with fewer filters as used here (the flexibility of the GSM potentials imposes limits on the attainable number of filters), we observe that filter coefficients may decay or disperse during learning (Fig. 3.8). To address this, we normalize the coefficients of each filter to unit ℓ^2 norm after each parameter update. This incurs no loss of generality due to the redundancy between the GSM scales and the filter coefficient norm: GSM potentials with an infinitely large range of scales can in principle adapt to filters with arbitrary coefficient norm. The necessarily limited range of GSM scales in practice, however, does not allow to properly model the potentials if the filters take extreme values. Moreover, removing the parameter redundancy increases robustness, and in turn enables learning more filters. Fig. 3.5(b) shows an example in which all 8 filters are "active" and contribute to the model. Unlike previous work [40, 2], combining filter normalization with more flexible potentials enables learning different, heavy-tailed potentials. This notably improves the ability to capture the marginal statistics of the inbuilt features.

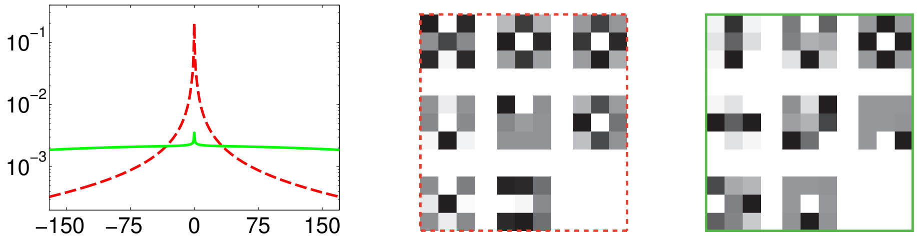


Figure 3.9: Typical uniform initialization of GSM weights (dashed, red) leads to filters with fewer patterns (*middle*). Broad (Dirac δ-like) initialization (solid, green) leads to more diverse filters (*right*).

Initialization of parameters

Initialization is crucial due to the non-convexity of the data likelihood. Specifically, we found that the initialization of the potential shape (GSM weights) can significantly affect learning, including the filters. A uniform initialization of the GSM weights (Fig. 3.9, red curve) is problematic. This overly constrains the pixel values and makes model samples spatially flat. The filter responses on the samples thus fall into a much smaller range than those on training images. The learning algorithm aims to reduce this difference by changing the filters toward patterns that reduce the filter-response range on natural images. The effect is that filters, particularly the Laplacian (Fig. 3.9, middle), are redundantly learned.

We alleviate this by initializing the potentials such that the pixels are initially less constrained than they should be. To that end, we initialize the potentials with a broad Dirac δ-like shape (Fig. 3.9, green curve). We find that this improves the robustness of learning and enables training a more diverse set of filters that captures different kinds of spatial structures (Fig. 3.9, right). Our findings indicate that the filters, on the other hand, are best initialized randomly. Initializing them to interpretable filters, such as Gabor filters, is counterproductive as these are usually not optimal for FoEs, and cause training to get stuck in poor local optima.

3.3.3 Evaluation of learned MRFs

Based on 1000 randomly cropped 48 × 48 image patches from the training images of [105], we trained (1) a pairwise MRF with fixed horizontal and vertical derivative filters and a single GSM potential, and (2) high-order FoEs with different size of cliques and different number of GSM experts including filters. We use a fixed base variance and a wide range of 15 scales $s = \exp(0, \pm1, \pm2, \pm3, \pm4, \pm5, \pm7, \pm9)$ to support a broad range of shapes. The remaining details of the learning procedure are similar to [7]: We use a fixed learning rate, exponential smoothing of the gradient, zero-mean filters, and stochastic gradient descent with mini-batches of 100 images. Fig. 3.2 shows the learned pairwise MRF and Fig. 3.5(b) shows a leaned 3 × 3 FoE

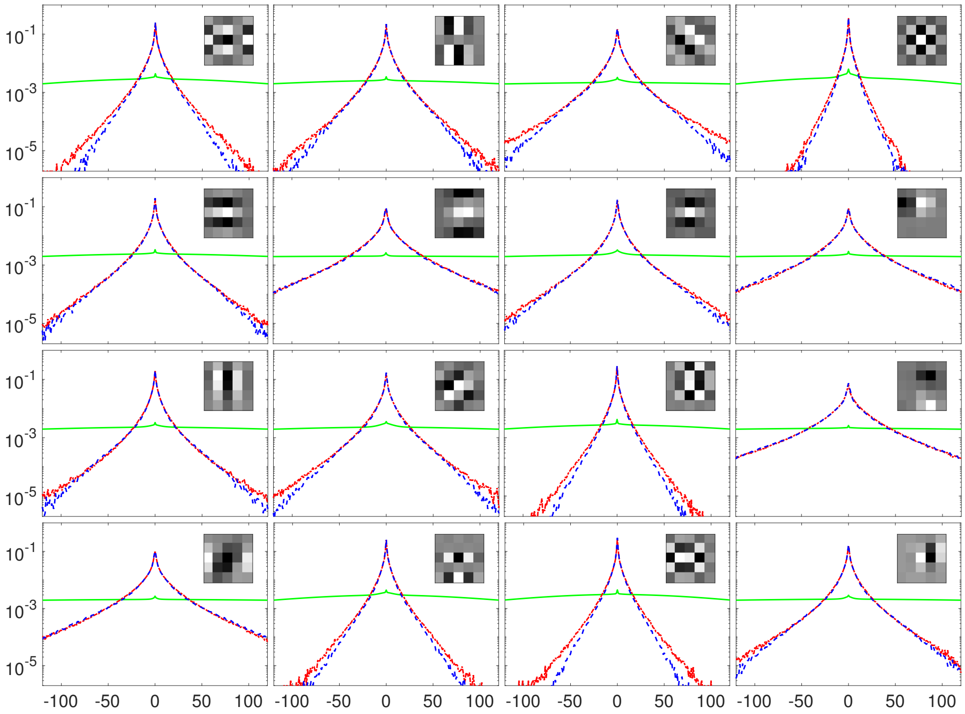


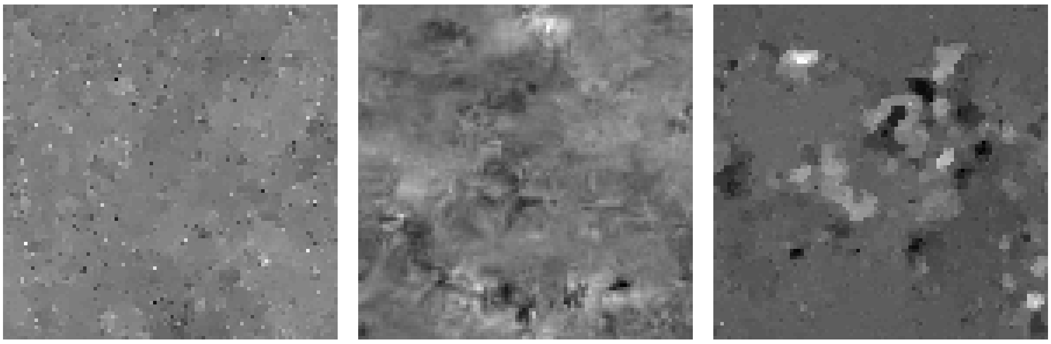
Figure 3.10: Learned 5 × 5 FoE with 16 experts and filters. Image filter (top right) with corresponding learned potential function (solid, green), marginal histograms of natural images (dash-dotted, red), and model marginals obtained by sampling (dashed, blue). All marginal KL-divergence values are smaller than 0.006. Note the more structured appearance of filters compared to that of the learned FoE filters in [7] (*cf*. Fig. 2.3).

with 8 experts and filters. To showcase the benefits of the improved learning scheme, a 5 × 5 model with 16 experts and filters is also learned (see Fig. 3.10). We use the same strategy in Sec. 3.2 based on model samples to analyze our learned models.

Model features

Fig. 3.2(a) shows the learned pairwise potential, which is *significantly heavier-tailed* than the marginal derivative statistics and looks similar to a Student-t distribution [46]. In contrast to the popular pairwise MRFs from above, it well captures the marginal derivative statistics of natural images (Fig. 3.2(b), marginal $D_{\text{KL}} = 0.006$). Because of the maximum entropy interpretation of MRFs, this potential shape is *optimal* for generative pairwise MRF image models.

In case of the FoEs, we find fully "active" filters and *very broad experts with a small, narrow peak* (*cf*. Fig. 3.5(b)). Their almost δ-like shape is in contrast to the experts used before [7, 39] (*cf*. Fig. 2.3(b)). Fig. 3.5(b) shows that these learned experts lead to a good match between the filter statistics of natural images and the filter marginals of the learned model. This is also true for our 5×5 models (Fig. 3.10).



(a) pairwise MRF (b) 5×5 FoE (c) MGSM-FRAME

Figure 3.11: Samples from our learned models. (a) The learned pairwise model yields locally uniform samples with occasional discontinuities that appear spatially isolated ("speckles"). (b) The learned high-order FoE model leads to smoothly varying samples with spatially correlated discontinuities. (c) Our reimplemented FRAME model [6] with expert functions represented by mixtures of GSMs (MGSM) yields samples with large uniform areas as well as edge-like discontinuities.

The resulting priors are thus truly *maximum entropy* models. We can conclude that the Student-t experts of [7] were not heavy-tailed enough, and that fitting experts to marginal statistics [39] is not appropriate. Instead, GSM expert functions prove to be sufficiently flexible for achieving good generative properties.

Other important statistics

To fully comprehend the modeling power of MRF priors, it is instructive to go beyond the model's features. Since natural images exhibit heavy-tailed statistics even for the marginals of *random linear filters* [67], we evaluate our models in this regard with random filters of 4 different sizes (8 of each size). Fig. 3.12 (top) shows the average responses to these random filters for natural images, as well as the learned models. Moreover, natural images have been found to exhibit *scale invariant derivative statistics* [108, 109]. Hence, we check the marginal statistics of derivatives at 3 image scales (powers of 2), which are shown in Fig. 3.12 (middle). Natural images have also been found to have characteristic *conditional distributions* of two image features, with a particular bow-tie shape [32]. Fig. 3.12 (bottom) shows the conditional histograms of neighboring image derivatives.

From Figs. 3.12(b), we can see that the learned pairwise MRF only captures the statistics of small random 3×3 filters and derivatives at the finest scale well. The model marginals of larger random filters and coarser-scale derivatives, however, tend toward Gaussians. The conditional distribution of neighboring derivatives also deviates much from that of natural images. On the other hand, the learned high-order FoEs (Figs. 3.12(c) and (d)), well capture the characteristics of natural images across a much wider range of random filter sizes and derivative scales, as well as the conditional statistics, which clearly demonstrates the improved modeling power. This also becomes apparent by visually comparing samples from these models (see Fig. 3.11). Moreover, we can observe that larger and more filters can further improve the model's statistics.

To the best of our knowledge, this is the first time that such close matches between model and natural image statistics have been reported for MRF image priors. Importantly, this also demonstrates that FoEs are indeed capable of capturing a large number of key statistical properties of natural images.

3.4 Learning MRFs with multi-modal potentials

In the previous sections, we have discussed MRFs with unimodal potenitals, such as the Student-t, the generalized Laplacian or the more flexible GSMs. The intuition is that important image features usually show heavy-tailed distributions. However, there are MRFs that require multimodal potentials. One example are the seminal FRAME models of Zhu *et al.* [6, 8], which use a discrete non-parametric representation for the potential functions that allows to represent multimodal functions. Heess *et al.* [9] proposed bi-modal potentials for modeling textures in a Fields-of-Experts framework. Unlike unimodal potentials, which proved insufficient for modeling specific kinds of textures, their bimodal potentials enabled the model to learn the characteristic properties of textures.

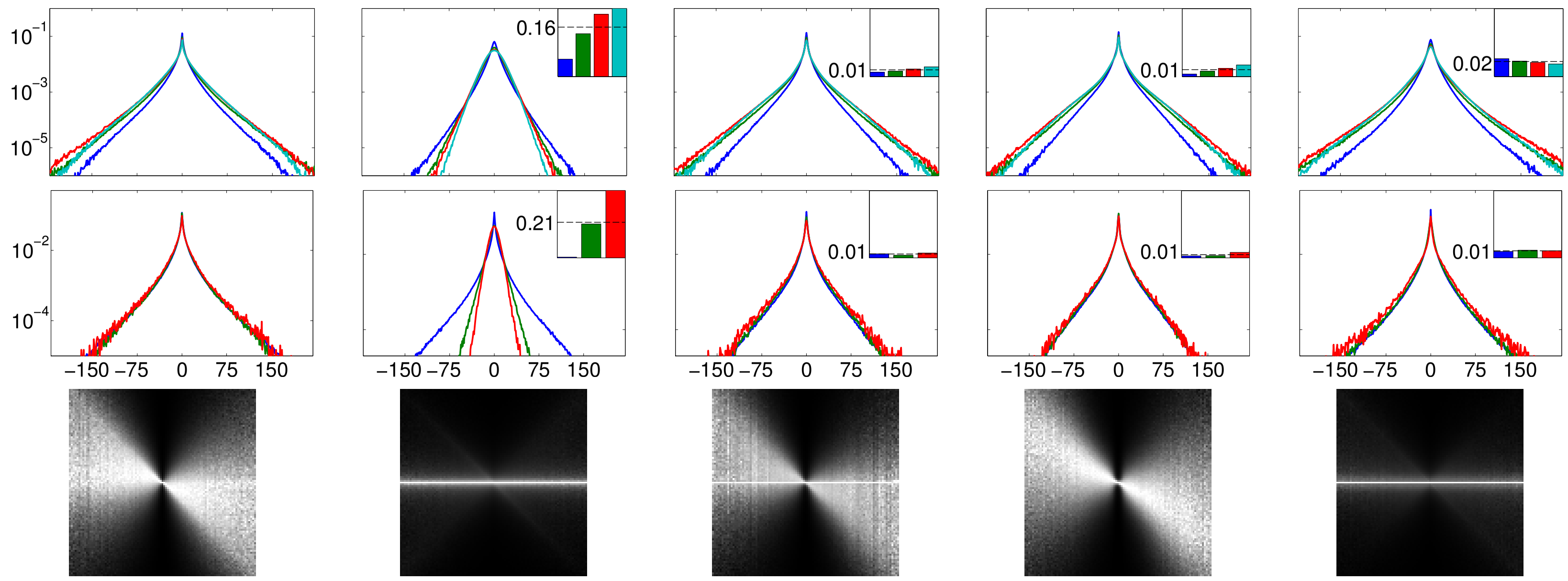
3.4.1 Mixtures of GSMs for modeling potentials

In Sec. 3.1.1 we have chosen flexible Gaussian scale mixtures (GSM) as the potentials in the MRFs. Formally, GSMs are specified as an infinite mixture of Gaussians (MoG) with shared mean μ (typically zero), but different variance. In most applications, finite variants are used, e.g.,

$$\phi_{\text{GSM}}(x) = \sum_{k=1}^{K} \alpha_k \cdot \mathcal{N}(x; \mu, s_k \cdot \sigma^2), \tag{3.16}$$

where $\alpha_k \geq 0, \sum_k \alpha_k = 1$ are the mixture weights of the scales s_k. Fig. 3.13 (left) illustrates a heavy-tailed distribution modeled as a GSM and its Gaussian components.

Here, we propose mixtures of GSMs (MGSMs) as an alternative, more flexible



(a) Natural images (b) Pairwise MRF (c) 3×3 FoE (8 filters) (d) 5×5 FoE (16 filters) (e) MGSM-FRAME

Figure 3.12: Random filter, multiscale derivative and conditional statistics. *(top)* Average marginal histograms of 8 random filters (0-mean, unit norm) of various sizes: 3×3 (blue), 5×5 (green), 7×7 (red), 9×9 (cyan). *(middle)* Derivative statistics at three spatial scales: 0 (blue), 1 (green), 2 (red); scales are powers of 2 with 0 being the original scale. (top right – KL-divergence) *(bottom)* Conditional histograms of neighboring derivatives. Brightness corresponds to probability.

representation. MGSMs are composed of different GSMs with differing means:

$$\phi_{\mathrm{MGSM}}(x) = \sum_{l=1}^{L} \sum_{k=1}^{K} \alpha_{lk} \cdot \mathcal{N}(x; \mu_l, s_{lk} \cdot \sigma^2). \tag{3.17}$$

The benefit of this formulation is that it allows representing multimodal distributions that are highly kurtotic with tight peaks (see Fig. 3.13 (right)).

To simplify our following treatment, we write the MGSM potential as a general Gaussian mixture:

$$\phi(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; \boldsymbol{\omega}_j) = \sum_k \omega_{jk} \cdot \mathcal{N}(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; \mu_{jk}, \sigma_{jk}^2), \tag{3.18}$$

where ω_{jk} are the normalized weights of the Gaussian component with mean μ_{jk} and variance σ_{jk}^2. It is important to note that this does not imply that we treat the potential as a general Gaussian mixture during learning, as this would lead to difficulties with learning and inference as described below.

For sampling the MGSM-based model we still rely on the Gibbs sampler described in Sec. 3.1.2. The conditional distribution $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\Omega})$ is slightly different:

$$p(z_{jc}|\mathbf{x}; \boldsymbol{\Omega}) \propto \omega_{jz_{jc}} \cdot \mathcal{N}\big(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; \mu_{jz_{jc}}, \sigma_{jz_{jc}}^2\big). \tag{3.19}$$

However, the conditional distribution $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\Omega})$ is more complex:

$$\begin{aligned}
p(\mathbf{x}|\mathbf{z}; \boldsymbol{\Omega}) &\propto \prod_c \prod_j \mathcal{N}\big(\mathbf{f}_j^{\mathrm{T}} \mathbf{x}_{(c)}; \mu_{jz_{jc}}, \sigma_{jz_{jc}}^2\big) \\
&\propto \exp\left(-\frac{1}{2} \sum_c \sum_j \frac{(\mathbf{w}_{jc}^{\mathrm{T}} \mathbf{x} - \mu_{jz_{jc}})^2}{\sigma_{jz_{jc}}^2} \right) \\
&\propto \exp\left(-\frac{1}{2} \mathbf{x}^{\mathrm{T}} \left(\sum_c \sum_j \frac{\mathbf{w}_{jc} \mathbf{w}_{jc}^{\mathrm{T}}}{\sigma_{jz_{jc}}^2} \right) \mathbf{x} + \mathbf{x}^{\mathrm{T}} \left(\sum_c \sum_j \frac{\mu_{jz_{jc}}}{\sigma_{jz_{jc}}^2} \mathbf{w}_{jc} \right) \right) \\
&\propto \exp\left(-\frac{1}{2} \mathbf{x}^{\mathrm{T}} \underbrace{(\mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^{\mathrm{T}})}_{\boldsymbol{\Sigma}^{-1}} \mathbf{x} + \mathbf{x}^{\mathrm{T}} (\mathbf{W} \boldsymbol{\Lambda} \tilde{\boldsymbol{\mu}}) \right) \\
&\propto \mathcal{N}\big(\mathbf{x}; \underbrace{\boldsymbol{\Sigma} \mathbf{W} \boldsymbol{\Lambda} \tilde{\boldsymbol{\mu}}}_{\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}}, \boldsymbol{\Sigma}\big),
\end{aligned} \tag{3.20}$$

where $\mathbf{W}$ and $\boldsymbol{\Lambda}$ take the same forms as in (3.6), and $\tilde{\boldsymbol{\mu}} = [\cdots, \mu_{jz_{jc}}, \cdots]^{\mathrm{T}}$ is composed of the means of GSM components. Note that the conditional mean $\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}$ is in general different from zero. Sampling (3.20) still remains efficient by solving a least-squares problem.

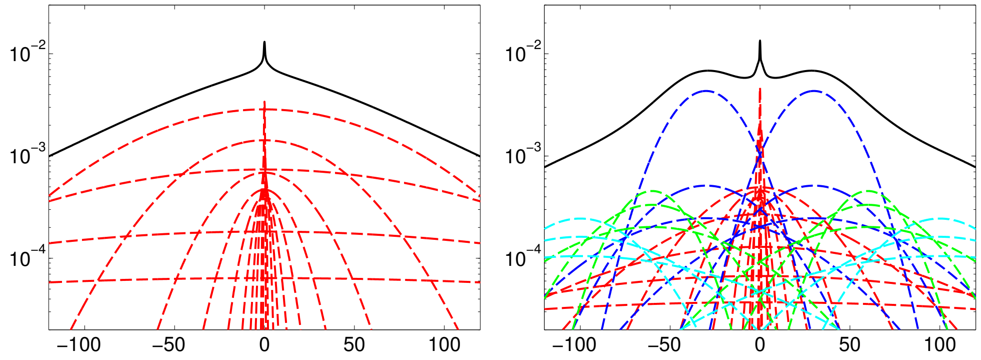


Figure 3.13: Potential functions (solid lines) represented by *(left)* unimodal Gaussian scale mixture (GSM) and *(right)* multimodal mixture of GSMs (MGSM). The dashed lines show the Gaussian components. Colors indicate GSMs with different means.

MGSMs *vs.* general Gaussian mixtures

Even though the model as well as the sampling procedure in principle allow using potentials modeled as arbitrary mixtures of Gaussians (*cf.* [42]), we found this not to work well in practice. In a general MoG, not only the weights, but also the component means and variances have to be learned, which leads to many local optima in the learning objective. In fact, even re-fitting a trained MGSM model with a general MoG using expectation maximization fails: The crucial peak at zero is only poorly captured, which leads to models with incorrect statistical properties. The MGSM model proposed here avoids these difficulties and facilitates learning broad and at the same time tightly peaked potentials.

3.4.2 Revisiting FRAME

The FRAME model for natural images [6] is a combination of pairwise MRFs at four image scales, thus a high-order MRF. Under this model the energy of a natural image is defined as

$$E(\mathbf{x}) = \sum_{s=0}^{3} \sum_{c \in \mathcal{C}_s} \left(\phi_{xs}(\mathbf{f}_x^{\mathrm{T}} \mathbf{x}_{(c)}^{[s]}) + \phi_{ys}(\mathbf{f}_y^{\mathrm{T}} \mathbf{x}_{(c)}^{[s]}) \right), \tag{3.21}$$

where s are four scales of an image pyramid, $\mathcal{C}_s$ defines the cliques at each scale, $\mathbf{f}_x$ and $\mathbf{f}_y$ are first-order x- and y-derivative filters with corresponding potentials ϕ_{xs} and ϕ_{ys}, and $\mathbf{x}_{(c)}^{[s]}$ denotes the pixels of clique c at image scale s.

Even though conceptually clear and simple, the model was hindered significantly by its implementation and the computational capabilities of computers at the time: Discrete, non-parametric representations with a small number of bins were used to model the potentials, images were reduced to 32 gray levels, and a single-site Gibbs sampler with discrete-valued pixels was used for learning[1]. As a result, this family of models did not find widespread use despite of being very flexible in principle. Finally, inference was performed in a PDE framework, which required to approximate the learned potentials and only led to modest levels of image restoration performance.

Another hinderance in adoption, as, e.g., noted by Weiss and Freeman [39], is the unintuitive shape of the learned potentials. While the potentials for the derivative features at the finest scale are heavy-tailed, those for the coarser scales were found to hand an "inverted" shape. This seems to contradict essentially all other MRF image models in the literature. Nonetheless, when evaluating the model by sampling from it, Zhu and Mumford [6] found the model to reproduce the scale invariant derivative statistics of natural images.

MGSM-FRAME

The MGSMs described above can represent multimodal potentials. In contrast to the non-parametric representation as used in the original FRAME models [6, 8], MGSMs are continuous-valued, thus do not require any gray value discretization, and moreover admit faster mixing auxiliary variable Gibbs samplers. With the flexible MGSM potentials, it is thus possible to analyze the FRAME model with modern inference and learning techniques and to reassess its performance. The FRAME model can be "translated" into an MRF in our framework:

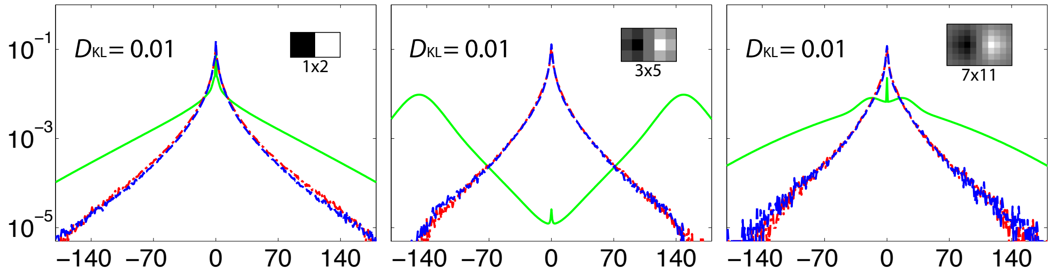
$$p_{\mathrm{FRAME}}(\mathbf{x};\mathbf{\Omega}) \propto \prod_{s=0}^{S} \prod_{c \in \mathcal{C}_s} \phi(\mathbf{f}_{xs}^{\mathrm{T}} \mathbf{x}_{(c)}; \omega_s) \cdot \phi(\mathbf{f}_{ys}^{\mathrm{T}} \mathbf{x}_{(c)}; \omega_s). \tag{3.22}$$

Dealing with S different image scales is done by omitting the downsampling step for the coarser scales, and instead increasing the stride between neighboring cliques $\mathcal{C}_s$ (2^s pixels at scale s). The filters $\mathbf{f}_{xs}$ and $\mathbf{f}_{ys}$ for $s > 0$ are given by convolutions of horizontal and vertical first derivative filters and the binomial kernels stemmed from $[1/4, 1/2, 1/4]^{\mathrm{T}}$ that are used for generating the image pyramid.

As larger filters will make the precision matrix in (3.20) denser and thus slow down the sampling, for simplicity and efficiency, we only learn the MGSM-FRAME model with 3 scales and use the same potentials for horizontal and vertical filters at each scale. We use 3 GSM components – one with fixed zero mean and two with non-zero means to be learned – to enable learning tri-modal potentials.

Generative properties. Fig. 3.14 shows the potentials and the marginal statistics of the learned model. The in-build feature, i.e., multi-scale derivative statistics being well captured suggests that the learning objective is reached. While the random filter

[1] In other regards, FRAME was well ahead of its time. E.g., it used persistent contrastive divergence 10 years before it was rediscovered [106].



(a) original scale (b) scale 1 (c) scale 2

Figure 3.14: Learned potentials (solid, green) of MGSM-FRAME, model marginals (dashed, blue) and marginals of natural images (dash-dotted, red). Top right: corresponding filters (only horizontal ones shown). Learned potentials at scale 1 have an "inverted" shape, which was also observed in [6].

statistics are much better than those of the pairwise MRF, the conditional statistics are still relatively poor (see Fig. 3.12(e)). Important image structures are still not well captured by multi-scale derivative filters. A sample from the learned MGSM-FRAME model is shown in Fig. 3.11(c). As can be seen, the sample contains large uniform areas as well as simple edge-like discontinuities.

About the "inverted" potentials. Although the potentials are represented differently in our MGSM-FRAME model, we can confirm the finding of "inverted" potentials in the original FRAME for natural images [6]. Comparing Fig. 3.14 with Fig. 3.12 (b, middle), it is not hard to understand the unintuitive shape of the learned potentials. To make the derivative statistics at coarser scales of a pairwise MRF right, applying inverted-shape derivative regularizers at these scales can encourage the Gausian-like statistics to be more heavy-tailed, which results in the same structure as the FRAME. Also note that central peaks of the potentials at coarser scales are important for right model statistics. While the peaks are also observed in the original FRAME, it is ignored by the fitting functions in the applications.

3.4.3 Discussion

The 5×5 filters of the FoEs that we have learned in Sec. 3.3 can only cover image structures with moderate sizes. Pushing the FoEs even further to learn larger filters is difficult due to the less-sparse linear equation systems in sampling. The FRAME model, taking another route, provides a solution to regularize large image structures. Since the mixture of GSMs provide a more flexible representation of potential functions, and the FRAME model can be regarded as a hierarchical extension of common pairwise MRF, is it worthwhile to learn a hierarchical FoE in a similar way? For the case of pairwise MRFs, as the model statistics at coarser scales are unsatisfactory, applying potentials on these scales will be beneficial. The FoEs, however, have already depicted good multi-scale statistics, which suggests that further applying coarser-scale potentials will not contribute no matter how the shapes of such potential functions are. For this reason, under the filter-based MRF model structure and learning procedures discussed in this thesis, a hierarchical extension of FoEs does not help; at least no appropriate coarser-scale filters can be learned. Therefore, further gains of natural image priors are likely challenging and may require new model design.

3.5 Modeling visual textures with MRFs

When only considering specific types of images or scenes, somewhat different challenges arise, since the specific structure of the data needs to be taken into account. For example, visual textures, even though playing a large part in the composition of natural images, cannot be modeled well by directly applying the methods for building generic image priors. The major reason is that generic image priors mainly consider the smoothness and continuity of the image, while texture models have to capture the specific textural structures.

To this end, the seminal FRAME texture model [8] uses Markov random fields (MRFs) with non-parametric, multi-modal potentials to allow for spatial structure generation. Heess *et al.* [9] suggested an MRF with parametric bi-modal potentials, which can be learned alongside the filters (features). Another class of probabilistic texture models extends restricted Boltzmann machines (RBMs) [110] toward capturing the spatial structure of textures, e.g., work by Kivinen *et al.* [36] and Luo *et al.* [37]. Common to these MRF- or RBM-based texture models is that they can be interpreted as a conditional Gaussian random field whose parameters are controlled by discrete latent variables. Moreover, all of them simultaneously perform regularization and generate textural structures through modeling the conditional covariance and mean, respectively. Due to their complex "mean+covariance" construction, these models are not easy to train in practice. Some compromises toward stabilization of training, e.g., tiled-convolutional weight-sharing [36, 37], can be detrimental to the quality of the generated textures. Moreover, the relative importance of the mean vs. the covariance component of these models is unclear in light of modeling textures.

3.5.1 Mean units

Mean units in MRFs

Eq. (3.1) is a typical formulation of a MRF prior for a generic image $\mathbf{x}$ through modeling the response to some linear filters with potential functions. As the filter responses usually have heavy-tailed empirical distributions around zero, the potential functions $\phi(\cdot;\boldsymbol{\omega}_j)$ are also chosen to be heavy-tailed (e.g., Student-t). Many heavy-tailed potentials can be formulated as Gaussian scale mixtures (GSMs) [32]. For better understanding and more efficient inference, such GSM potentials allow augmenting the prior with discrete-valued hidden variables $\mathbf{z} = (z_{jc})_{j,c}$, one for each filter $\mathbf{f}_j$ and clique c, which represent the index of the Gaussian mixture component modeling the filter response [50]. It holds that $p_{\text{MRF}}(\mathbf{x};\boldsymbol{\Omega}) \propto \sum_{\mathbf{z}} p_{\text{MRF}}(\mathbf{x},\mathbf{z};\boldsymbol{\Omega})$. Given the hidden variables, the conditional distribution for the image is a zero-mean Gaussian

$$p_{\text{MRF}}(\mathbf{x}|\mathbf{z};\boldsymbol{\Omega}) \propto \mathcal{N}\big(\mathbf{x};\mathbf{0},\boldsymbol{\Sigma}(\mathbf{z},\boldsymbol{\Omega})\big). \tag{3.23}$$

As changing the hidden units only changes the conditional covariance, such basic image priors focus on modeling the covariance structure of the image, which is intuitive as they are primarily aimed at regularization.

Heess *et al.* [9] showed that such generic MRF priors for natural images are not suitable for textures, and propose to extend them using bi-modal potential functions. Multi-modal potentials can also be modeled with Gaussian mixtures; however, the components may no longer all have zero means. Given the hidden units, the conditional distribution of such an MRF texture model

$$p_{\text{MRF}}(\mathbf{x}|\mathbf{z};\boldsymbol{\Omega}) \propto \mathcal{N}\big(\mathbf{x};\boldsymbol{\mu}(\mathbf{z},\boldsymbol{\Omega}),\boldsymbol{\Sigma}(\mathbf{z},\boldsymbol{\Omega})\big) \tag{3.24}$$

shows that bi-modal potentials capture not only the covariance structure, but also the local mean intensities. The seminal FRAME texture model [8] with its non-parametric potentials is also consistent with this observation. Comparing (3.24) with (3.23) suggests that modeling the conditional mean is a particular trait of texture models. The intuitive explanation is that the model does not "just" perform regularization, but instead generates textural structure.

Note that in these models the filters are applied in a convolutional manner across the entire image. Since filters can be understood as weights connecting visible and hidden units, this is called convolutional weight sharing (*cf.* Fig. 3.15). Importantly, this keeps the number of parameters manageable, even on images of an arbitrary size, and also gives rise to the model's spatial invariance.

Nonetheless, learning such a "mean+covariance" model is difficult in practice, since the hidden units affect both conditional mean and covariance in complex ways. Since the filters need to be sufficiently large to generate coherent structures, the resulting covariance matrix will furthermore be quite dense, making both learning and inference rather inefficient. Moreover, the learned texture filters from [9] lack clear structure (see Fig. 3.16), making them difficult to interpret.

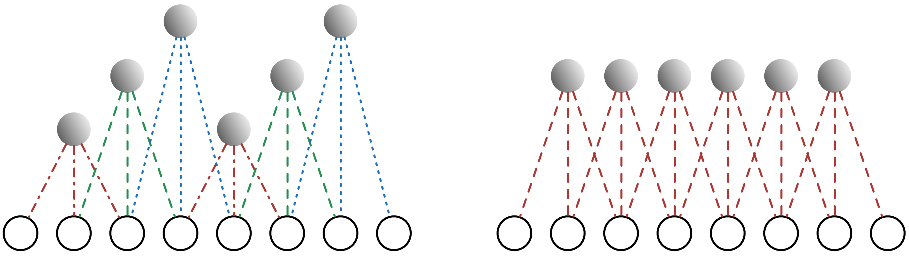


Figure 3.15: Tiled-convolutional (left) and full-convolutional (right) weight sharing. Lines converging to the hidden units (shaded) are the filters; they share their parameters when indicated by the same color or line type.

Mean units in RBMs

Models derived from restricted Boltzmann machines (RBMs) take a different route. A Gaussian RBM [110] models an image by defining an energy function over visible units $\mathbf{x}$ (here, the image pixels) and binary hidden units $\mathbf{h}$. The random variables have a Boltzmann distribution $p_{\text{RBM}}(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\{-E_{\text{RBM}}(\mathbf{x}, \mathbf{h})\}$, where Z is the partition function and E denotes energy. Gaussian RBMs have the property that the conditional distribution of the visible units given the hidden ones is a Gaussian

$$p_{\text{RBM}}(\mathbf{x}|\mathbf{h}; \boldsymbol{\Omega}) \propto \mathcal{N}\big(\mathbf{x}; \boldsymbol{\mu}(\mathbf{h}, \boldsymbol{\Omega}), \boldsymbol{\Sigma}\big), \tag{3.25}$$

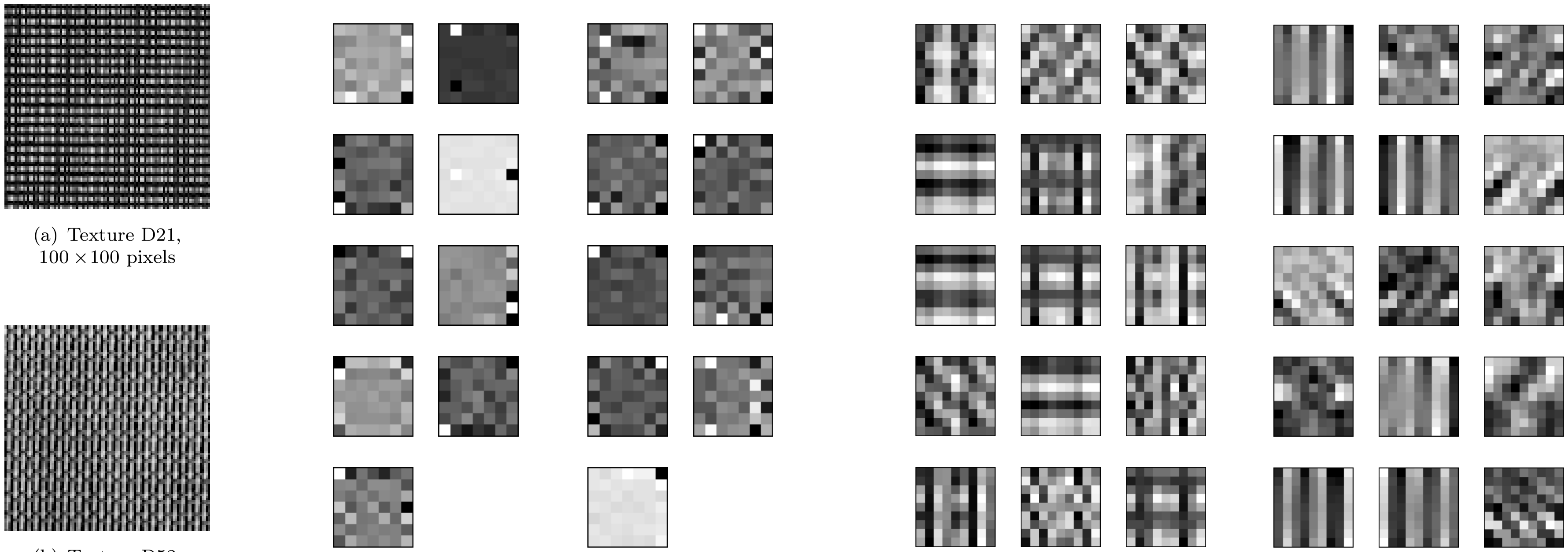
in which only the conditional mean depends on the hidden variables.

More recent variants of Boltzmann machines for texture modeling [36, 37] not only model the conditional mean, but also the covariance akin to (3.24). [36] experimentally compared three models: Gaussian RBMs, Products of Student-t (PoT) [2], and their combination, which corresponds to modeling conditional mean, covariance, and mean+covariance, respectively. The results revealed the importance of the conditional mean for texture synthesis and inpainting.

Note that both [36, 37] adopt tiled-convolutional weight sharing [107] (see Fig. 3.15). The apparent reason is that the states of hidden units are less correlated, thus making training of the models easier. Unfortunately, tiled-convolutional models involve many parameters, since several sets of features (filters) must be learned. For example, [36, 37] learn and use more than 300 features for every texture, which are moreover not spatially invariant. Consequently, tiled-convolutional weight sharing requires copious training data, which for textures is often not available.

3.5.2 Learning convolutional Gaussian RBMs for textures

Mean units appear to be an important component of many texture models. As these models also include covariance units and/or complex weight sharing, it is not



(a) Texture D21, 100 × 100 pixels

(b) Texture D53, 100 × 100 pixels

(c) D21 BiFoE filters

(d) D53 BiFoE filters

(e) D21 our cGRBM filters

(f) D53 our cGRBM filters

Figure 3.16: Brodatz textures [52] and learned filters using BiFoE [9] and our cGRBM.

clear how important the mean units are. Below, we investigate this and explore the capability of "mean-only" Gaussian RBMs for textures.

Convolutional Gaussian RBM

A spatially invariant model is obtained through applying the Gaussian RBM to all overlapping cliques of a large texture image in a convolutional manner. The energy function of the convolutional Gaussian RBM (cGRBM) is then written as

$$E_{\text{cGRBM}}(\mathbf{x},\mathbf{h}) = \frac{1}{2\gamma}\mathbf{x}^{\mathrm{T}}\mathbf{x} - \sum_{c}\sum_{j} h_{jc}\big(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_c + b_j\big), \tag{3.26}$$

where we add a weight γ to the quadratic term. Here, $\mathbf{w}_j$ determine the interaction between pairs of visible units $\mathbf{x}_c$ and hidden units h_{jc}. Thus, $\mathbf{w}_j$ are the features or filters, b_j are the biases, c and j are indices for all overlapping image cliques and filters, respectively. The conditional distribution of $\mathbf{x}$ given $\mathbf{h}$ is a Gaussian

$$p_{\text{cGRBM}}(\mathbf{x}|\mathbf{h}) \propto \mathcal{N}\Big(\mathbf{x}; \gamma\sum_{c}\sum_{j} h_{jc}\mathbf{w}_{jc}, \gamma\mathbf{I}\Big), \tag{3.27}$$

where the vector $\mathbf{w}_{jc}$ is defined as $\mathbf{w}_{jc}^{\mathrm{T}}\mathbf{x} = \mathbf{w}_j^{\mathrm{T}}\mathbf{x}_c$. The conditional distribution of $\mathbf{h}$ given $\mathbf{x}$ is a simple logistic sigmoid function

$$p_{\text{cGRBM}}(h_{jc}|\mathbf{x}) \propto \text{logsig}(\mathbf{w}_{jc}^{\mathrm{T}}\mathbf{x} + b_j). \tag{3.28}$$

Probability density of an image under cGRBM

Based on the energy function of the cGRBM in (3.26), we have the joint distribution of the visible units $\mathbf{x}$ (image) and the hidden units $\mathbf{h}$

$$p_{\text{cGRBM}}(\mathbf{x},\mathbf{h}) \propto \mathcal{N}(\mathbf{x};\mathbf{0},\gamma\mathbf{I})\cdot\prod_{j}\prod_{c}\exp\big\{h_{jc}(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_c+b_j)\big\}. \tag{3.29}$$

The probability density of $\mathbf{x}$ can be obtained by marginalizing out the hidden units

$$\begin{aligned} p_{\text{cGRBM}}(\mathbf{x}) &\propto \mathcal{N}(\mathbf{x};\mathbf{0},\gamma\mathbf{I})\cdot\prod_{j}\prod_{c}\Big(1+\exp\{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_c+b_j\}\Big) \\ &\propto \mathcal{N}(\mathbf{x};\mathbf{0},\gamma\mathbf{I})\cdot\prod_{j}\prod_{c}\varphi\big(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_c;b_j\big), \end{aligned} \tag{3.30}$$

where φ are filter specific potential functions. As can be seen, the cGRBM is also a Markov random field (MRF) model.

Texture data

For a fair comparison with other models [9, 36, 37], we follow their use of the Brodatz texture images [52] for training and testing our models. The images are rescaled to either 480×480 or 320×320 pixels, while preserving the major texture features, and then are normalized to zero mean and unit standard deviation. We also divide each image into a top half for training and a bottom half for testing.

Learning

As the partition function of the model is intractable, we perform approximate maximum likelihood (ML) learning based on persistent contrastive divergence (PCD) [106]. Model samples are obtained using efficient block Gibbs sampling, which alternately samples the visible units $\mathbf{x}$ or the hidden units $\mathbf{h}$ given the other. Through marginalizing the hidden units we obtain the free energy

$$E(\mathbf{x}) = \frac{1}{2\gamma}\mathbf{x}^{\mathrm{T}}\mathbf{x} - \sum_j \sum_c \log\left(1 + \exp\{\mathbf{w}_{jc}^{\mathrm{T}}\mathbf{x} + b_j\}\right). \tag{3.31}$$

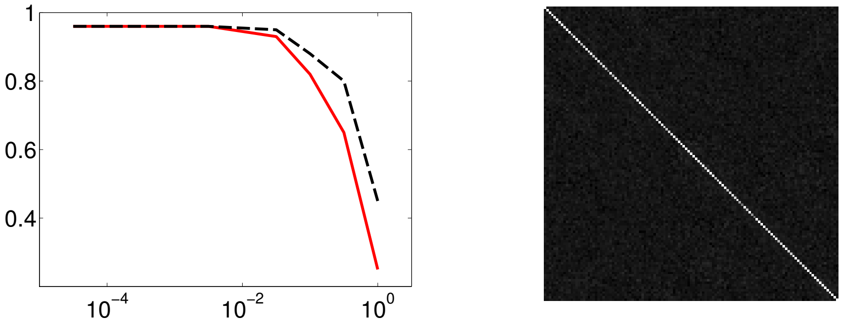
The parameters (i.e., the filters) are updated using gradient ascent

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} + \eta\left[\left\langle\frac{\partial E(\mathbf{x})}{\partial \mathbf{w}_j}\right\rangle_{\mathbf{X}^{\mathrm{PCD}}} - \left\langle\frac{\partial E(\mathbf{x})}{\partial \mathbf{w}_j}\right\rangle_{\mathbf{X}^0}\right], \tag{3.32}$$

where η is the learning rate, $\langle\cdot\rangle$ denotes the average over the training data $\mathbf{X}^0$ or the samples $\mathbf{X}^{\mathrm{PCD}}$.

The standard learning procedure [110], however, does not ensure that a good convolutional RBM texture model is learned in practice. For example, even the simple mean-only RBM baseline of [36] stabilizes learning using tiled-convolutional weight sharing and a slowly mixing Hybrid Monto Carlo (HMC) sampler. Consequently, care must be taken to be able to train a cGRBM for textures.

Choice of parameter γ. The typical best practice in Gaussian RBMs is to set the weight γ to 1 when the training data is normalized [110]. But we find that $\gamma = 1$ is far from the optimum and its value can greatly affect the generative properties of the trained texture models. Fig. 3.17(a) shows how γ changes the texture similarity score (TSS) [9] (see Sec. 3.5.3 for details) of model samples and synthesized textures. Actually, since a texture sample drawn with the Gibbs sampler is a sum of the conditional mean and *i.i.d.* Gaussian noise, $\gamma = 1$ will lead to the textural structures being dominated by noise. But even if we synthesize textures by taking the conditional mean of the final sampling iteration, as we do here, we see that γ can greatly affect the quality of the texture and that the previous best practice of $\gamma = 1$ does not work



(a)

(b)

Figure 3.17: *(a)* Texture similarity scores (TSS) of synthesized textures (black, dashed) and model samples (red, solid) *vs*. the choice of γ. *(b)* Covariance matrix of 1000 samples of $\mathbf{h}$ for $\gamma = 0.03$. Results are based on D21.

well. This may be the reason why other texture models considered more complex pixel covariances and/or rely on less well-mixing samplers for stabilizing training.

Although Fig. 3.17(a) suggests that smaller values of γ should be preferred, an overly small γ will lead to a small covariance for the Gaussian in (3.24) and consequently to slow mixing of the Gibbs sampler. We find $\gamma = 0.03$ to be a good trade-off. To illustrate this, Fig. 3.17(b) shows the covariance matrix computed from 1000 consecutive samples of $\mathbf{h}$ corresponding to one feature. As it is close to a diagonal matrix, the variables in $\mathbf{h}$ are approximately independent, thus the sampler mixes well. We use $\gamma = 0.03$ for all our experiments.

Other best practices. To obtain structured filters and – in our experience – also better texture models, we have to impose some constraints on the filters. As usual, the filters are initialized with random values, but their coefficient norms are ensured to be small initially. During training they are moreover constrained to have zero mean and limited to not increase above an empirical threshold of $\frac{0.05}{\gamma}$. Otherwise, the filters will often get stuck in poor local optima without any clear structure. Since the biases do not change significantly during learning, we fix them to $b_j = -\frac{1}{3}\|\mathbf{w}_j\|$, similar to [35]. The bias depends on the current norm of filter coefficients to keep a reasonable portion of the hidden units being "on" (*cf*. (3.28)).

Also note that the typical whitening of the training data cannot be applied for textures, even if it is common for natural image priors. Since whitening will remove the major structural pattern of a single texture, it is in our experience difficult for the RBM to represent the remaining spatial patterns.

The learned models. We trained cGRBM models for several Brodatz textures, each of which is trained based on 40 patches of size 76×76, randomly cropped from the corresponding preprocessed training image. As all the training images are rescaled,

Table 3.1: Means and standard deviations of TSS of the synthesized textures.

Model	D6	D21	D53	D77
BiFoE [9]	0.757 ± 0.059	0.871 ± 0.032	0.827 ± 0.087	0.646 ± 0.022
Tm [36]	0.930 ± 0.021	0.890 ± 0.079	0.849 ± 0.061	0.866 ± 0.008
TmPoT [36]	0.933 ± 0.036	0.896 ± 0.070	0.853 ± 0.056	0.870 ± 0.008
TssRBM [37]	0.937 ± 0.047	0.948 ± 0.025	0.941 ± 0.022	0.841 ± 0.012
DBN [37]	0.952 ± 0.016	0.947 ± 0.032	0.950 ± 0.026	0.864 ± 0.160
cGRBM (ours)	$\mathbf{0.963} \pm 0.005$	$\mathbf{0.961} \pm 0.008$	$\mathbf{0.965} \pm 0.004$	$\mathbf{0.875} \pm 0.013$

we simply fix the filter size to 9×9 for all models. The models for textures D6, D21 and D53 consist of 15 learned filters, while the model for D77 has 20 filters due to its slightly more complex pattern. Examples of the learned filters are shown in Fig. 3.16. We can observe clearly apparent structure, e.g., unlike [9].

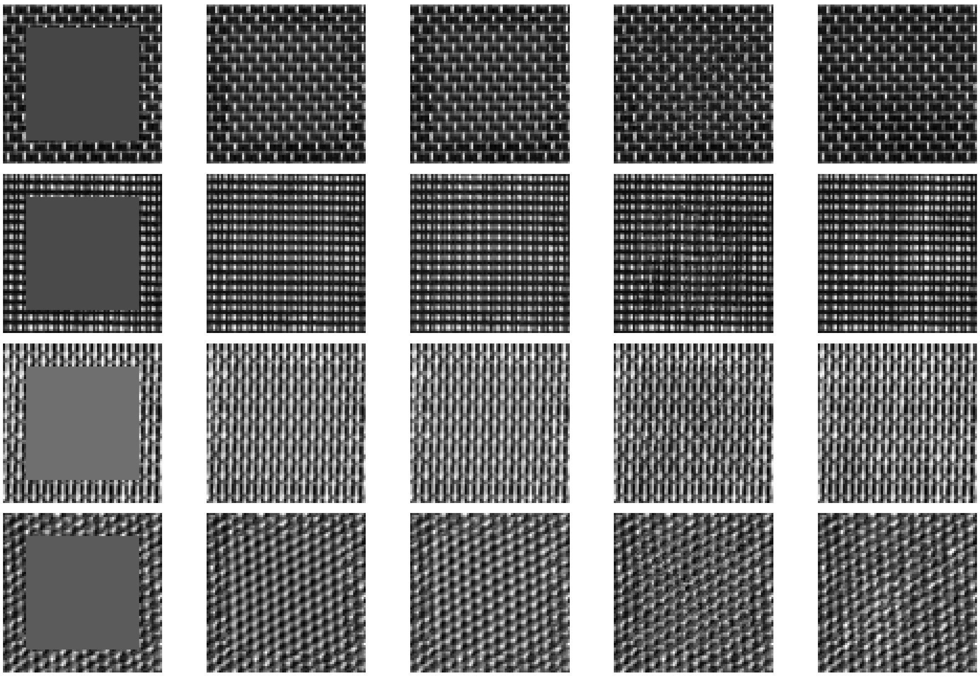
3.5.3 Generative properties

To evaluate the generative performance of the learned cGRBM texture models, we quantitatively compute texture similarity scores (TSS) [9] between the synthesized textures and real texture patches, which is defined based on the maximum normalized cross correlation

$$\mathrm{TSS}(\mathbf{s}, \mathbf{x}) = \max_i \frac{\mathbf{s}^{\mathrm{T}} \mathbf{x}_{(i)}}{\|\mathbf{s}\| \cdot \|\mathbf{x}_{(i)}\|}, \tag{3.33}$$

where $\mathbf{s}$ is the synthesized texture and $\mathbf{x}_{(i)}$ denotes the patch of the same size as $\mathbf{s}$ within the texture image $\mathbf{x}$, at location i.

We collect 100 samples of size 76×76 for each model (each texture) using Gibbs sampling. Since in Gibbs sampling the texture samples are obtained by summing the final conditional mean and *i.i.d.* Gaussian noise, we use the conditional means from the last sampling step as the synthesized texture. For computing the TSS, only the center 19×19 pixels are considered (the same size as in [9, 36, 37]). Tab. 3.1 shows the means and standard deviations of TSS. Thanks to the full-convolutional weight sharing scheme, our simple cGRBM models only require 15 (or 20 for D77) features (filters) to exceed the generative properties of much more complex Boltzmann machine models [36, 37] with many more (> 300) features. Note the considerable performance difference between our learned cGRBMs and the Gaussian RBM baseline "Tm" of [36], which is based on a standard learning procedure and tiled-convolutional weight sharing. Our cGRBMs even outperform the deep belief networks (DBNs) of [37]. Meanwhile, the 9×9 filter size of our models is also smaller than that of [36, 37] (11×11). The BiFoE models [9] only use 9 filters of size 7×7, but the paper argues that more and larger filters do not lead to a large difference in model quality, but greatly reduce the efficiency of inference.



(a) Inpainting frames (b) Our cGRBM (c) Our deterministic (d) Efros & Leung [113] (e) Ground truth

Figure 3.18: Examples of inpainting results. From top to bottom: D6, D21, D53, D77.

3.5.4 Texture synthesis

Hao *et al.* [111] modified high-order Gaussian gated Boltzmann machines for texture modeling, and also directly model the dependencies between two visible units. They learned 1000 features with convolutional weight sharing and achieved good texture classification performance. The performance in texture generation are less satisfactory, however, with block artifacts appearing in texture inpainting results. Efros et al. [112, 113] proposed non-parametric methods for texture synthesis. In [113] the synthesized image is grown by one pixel at a time according to the best matches between its neighbors and patches from the texture. [112] instead stitch together small patches directly.

Texture inpainting

Following previous work [36, 37], we take 76 × 76 patches from the testing texture images and create a 54 × 54 square hole in the middle of each patch by setting the intensity values to zero. The task is to generate texture in the square hole that is consistent with the given boundary.

Inpainting is done through sampling conditioned on the given boundaries (*cf.*, Sec. 3.2.2). This procedure is quite efficient when using a block Gibbs sampler. For

Table 3.2: Means and standard deviations of MSSIM scores of the inpainted textures.

Model	D6	D21	D53	D77
Tm [36]	0.858 ± 0.016	0.866 ± 0.019	0.849 ± 0.023	0.764 ± 0.027
TmPoT [36]	0.863 ± 0.018	0.874 ± 0.012	0.860 ± 0.023	0.767 ± 0.032
TssRBM [37]	0.888 ± 0.023	0.912 ± 0.014	0.916 ± 0.024	0.763 ± 0.031
DBN [37]	0.889 ± 0.025	0.906 ± 0.016	0.924 ± 0.029	0.774 ± 0.023
cGRBM (ours)	$\mathbf{0.909} \pm 0.017$	$\mathbf{0.928} \pm 0.012$	$\mathbf{0.933} \pm 0.010$	$\mathbf{0.783} \pm 0.027$
Efros & Leung [113]	0.827 ± 0.028	0.801 ± 0.029	0.863 ± 0.018	0.632 ± 0.041
Deterministic (ours)	0.899 ± 0.019	0.918 ± 0.014	0.926 ± 0.016	0.775 ± 0.034

Algorithm 1 Deterministic Texture Inpainting

```
Require: Image $\mathbf{x}$ to be inpainted
  repeat
    $h_{jc} \leftarrow H(\mathbf{w}_{jc}^{\mathrm{T}}\mathbf{x} + b_j)$        ▷ where $H$ is a unit step function
    $\mathbf{x} \leftarrow \gamma \sum_{j,c} h_{jc}\mathbf{w}_{jc}$
  until no change of $\mathbf{x}$ (or $\mathbf{h}$)
  return $\mathbf{x}$
```

each texture, we use 20 different inpainting frames and perform inpainting 5 times with different initializations, leading to 100 inpainting results. The quality of the inpainted texture is measured by the mean structural similarity (MSSIM) score [114] between the inpainted region and the ground truth. Fig. 3.18 shows examples of inpainting results and Tab. 3.2 gives a quantitative comparison with other models[2], which we outperform considerably despite a simpler model architecture.

Deterministic texture inpainting method

From Sec. 3.5.2 we know that the value for γ in our cGRBM model is small. Looking at (3.27), this means that the sample will not deviate significantly from the conditional mean. Moreover, the norms of filter coefficients must be large to balance the small γ, which implies that most values of $\mathbf{w}_{jc}^{\mathrm{T}}\mathbf{x} + b_j$ will fall outside of the smooth transition area of the logistic sigmoid in (3.28). This suggests that, in applications, it may be possible to use deterministic functions to replace sampling the two conditionals. In particular, we apply a unit step function on $\mathbf{w}_{jc}^{\mathrm{T}}\mathbf{x} + b_j$, then use the obtained binary auxiliary variable to modulate the filters to reconstruct the image, and repeat the procedures until convergence (Algorithm 1). Note that this is equivalent to a block coordinate descent on the model energy in (3.26). Since this scheme only works well if some reference pixels are given, such as in texture inpainting, we use it in this context. While slightly worse than sampling the cGRBM, the performance of the deterministic approach is still better than other models. In our inpainting experiment, our deterministic method only needed $30 \sim 50$ iterations to reach convergence, while sampling the cGRBM usually required about 100 iterations. It is

[2] Our implementation of Efros & Leung [113] uses a window size of 15×15.

moreover quite efficient, because the computation in each iteration is very simple. By contrast, nonparametric methods (e.g., [113]) are often not as efficient due to the necessary matching step.

3.5.5 Discussion

Compared to more complex (and deep) texture models with latent variables controlling the conditional covariance [36, 37], the Gaussian RBMs trained in a convolutional fashion not only yield higher quality of the synthetic textures, but also show advantages such as computational efficiency, spatially invariant and clearly structured learned filters. Note that there is still some redundancy in the leaned few filters (see Fig. 3.16 (e, f)). An open question is to automatically determine the optimal number of filters based on the complexity of the respective texture. Moreover, although a mean-only model can capture repetitive textures well, covariances might have to be considered for general textures. Training such two kinds of filters simultaneously is difficult in practice.

3.6 Summary

Based on a flexible representation of MRF potentials using Gaussian scale mixtures (GSM), we developed an efficient auxiliary-variable Gibbs sampler (Sec. 3.1). In Sec. 3.2, we proposed to evaluate probabilistic MRF priors in an application-independent manner by checking their generative properties using model samples. We found that the popular pairwise and high-order MRF priors capture image statistics quite crudely and exhibit poor generative properties. We further developed new learning strategies for training expressive high-order MRFs in Sec. 3.3. The learned MRFs well capture the statistics of the inbuilt features, thus being truly maximum entropy models. More importantly, the high-order FoE also show a number of other key statistics of natural images, which outlines the capabilities of MRFs. By extending the representation of potentials to be multimodal (Sec. 3.4), we were able to revisit the seminal FRAME model and gained more insights of MRF variants. In Sec. 3.5 we analyzed and compared the “mean” and “covariance” units in MRF models. We trained compact “mean”-only fully-convolutional RBMs that can model visual repetitive textures even better than more complex and deep “mean+covariance” models.

Chapter 4

Image analysis using learned MRF priors

MRF image models have found widespread use across low-level and high-level vision as reviewed in Chapters 1 and 2. In this chapter, we introduce new approaches with the learned comprehensive Fields of Experts (FoE) priors to solve three different real-world image analysis problems. First, we propose a new sampling-based method for natural image denoising, which moves away from computing the prevalent maximum a-posteriori (MAP) solution, but uses the Bayesian minimum mean squared error (MMSE) estimate (Sec. 4.1). This method was published in [50, 51]. Second, for microscopy image deconvolution, we further extend the sampling-based inference method to deal with more complex likelihoods. This allows super-resolution, deconvolution and denoising to be performed jointly (Sec. 4.2). This work was published in [58]. Third, we present a global optical flow-based method for non-rigid registration of cell nuclei in live cell time-lapse microscopy. Rather than using the MRFs as priors, we employ FoEs and convolutional RBMs to train image filters for extracting important image features. Registration in learned feature spaces yields higher accuracies (Sec. 4.3). This work was published in [59, 60]. In addition, elasticity properties are integrated into the MRF prior to achieve more robust registration performance, which was published in [61, 62].

4.1 Natural image denoising

Digital images always contain some random variation of brightness or color information, which is called image noise. It generally comes directly from the image sensor and circuitry of the imaging device. While different types of images have different noise sources and characteristics, here we consider typical Gaussian noise. The process of an image $\mathbf{x}$ being degraded by Gaussian noise is commonly formulated as

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{4.1}$$

where $\mathbf{n}$ is i. i. d. Gaussian noise with zero mean and variance σ^2, and $\mathbf{y}$ is the resulting noisy image. The goal of denoising is to recover the hidden noise-free image from the observed noisy image.

MRF-based denoising models are one major type of image denoising approaches. In Chapters 1 and 2, we have reviewed different MRFs for denoising (e.g., discriminative MRFs [102, 14, 11], generative MRFs [39, 7, 3]). Most MRF denoising methods compute the maximum a-posteriori (MAP) estimation, which is relatively efficient. However, theoretical and empirical results [115, 56] have pointed to deficiencies of MAP estimation (e.g., ad-hoc modifications of the generative model setting [7], inherent bias toward δ-like marginals [56]). The Bayesian minimum mean squared error (MMSE) estimate, which is equivalent to the expectation of the posterior, is also used in image denoising applications. Since approximating the MMSE estimate by sampling is computationally very expensive (e.g., perfect sampling in [47]), variational Bayesian methods, in particular, mean field approximation, have been used in MRF denoising methods [48, 49] with slightly decreased accuracy.

Below, we also mention some related image denoising methods that are not based on MRFs. BLS-GSM proposed by Portilla *et al.* [32] builds a GSM-based model of wavelet coefficients and computes the MMSE estimate from the posterior. The non-local means method [116] exploits the inherent self-similarities of natural images. A noise-free pixel is restored by a weighted average of non-local pixels based on contextual similarities. The result can be regarded as an approximation to the MMSE estimate. BM3D [55] is a popular image denoising approach based on block-matching of image fragments followed by 3D filtering. Non-local sparse coding (NLSC) [54] combines the non-local means approach and sparse coding for denoising. For a detailed review of previous work on image denoising, we refer to a recent survey [117].

In this section, we model the denoising problem in a Bayesian framework with the learned high-order MRFs for natural images from Sec. 3.3 as the prior in a purely generative setting. The auxiliary-variable Gibbs sampler introduced in Sec. 3.1.2 is extended to efficiently and accurately infer the posterior mean or the MMSE estimate. Experiments show that our approach outperforms previous MRF-based models [39, 7, 102] and can compete with popular denoising methods such as NLSC [54] and BM3D [55]. Moreover, for the proposed denoising method, the MMSE estimate not only substantially outperforms MAP, but also avoids several of its problems. In contrast to [56], our approach does not require a modification of the MRF framework. We demonstrate that a rigorous probabilistic interpretation and good generative properties of MRFs can go hand-in-hand with very good denoising performance.

The work was published in [50, 51].

4.1.1 Sampling-based denoising approach

We formulate the denoising problem as a posterior distribution in a Bayesian framework

$$p(\mathbf{x}|\mathbf{y};\boldsymbol{\Omega}) \propto p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x};\boldsymbol{\Omega}), \tag{4.2}$$

where $\mathbf{y}$ is the observed noisy image, $\mathbf{x}$ is the hidden noise-free image to be restored, and $\boldsymbol{\Omega}$ denotes model parameters. Due to the Gaussian noise assumption in (4.1), the application-specific likelihood $p(\mathbf{y}|\mathbf{x})$ is a Gaussian

$$p(\mathbf{y}|\mathbf{x}) \propto \exp\Big(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|^2\Big), \tag{4.3}$$

where $\|\cdot\|$ denotes the ℓ_2 norm. For the prior model $p(\mathbf{x};\boldsymbol{\Omega})$, we use the learned generative MRFs for natural images described in Sec. 3.3. Note that (4.2) is a purely generative setting. Though discriminative approaches that directly model the posterior (e.g., [102, 11]) are popular, and the learning and inference can be very efficient, they are limited to specific applications and lack the statistical interpretability. By contrast, our learned application-independent priors $p(\mathbf{x};\boldsymbol{\Omega})$ are versatile. While a large part of the denoising literature assumes Gaussian noise, our learned MRF priors and the denoising approach can be straightforwardly adapted to other types of noise.

MMSE estimation

We propose to restore the noise-free image $\mathbf{x}$ through computing the Bayesian minimum mean squared error (MMSE) estimate of the posterior

$$\hat{\mathbf{x}} = \arg\min_{\tilde{\mathbf{x}}} \int ||\tilde{\mathbf{x}}-\mathbf{x}||^2 p(\mathbf{x}|\mathbf{y};\boldsymbol{\Omega})\, \mathrm{d}\mathbf{x}, \tag{4.4}$$

which is equal to the mean of the posterior distribution and generally differs from the maximum in the case of non-Gaussian posteriors as used here due to the non-Gaussian prior $p(\mathbf{x};\boldsymbol{\Omega})$. Contrary to the maximum a-posteriori (MAP) estimate, 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*. [115]).

We instead approximate the MMSE estimate, or the posterior mean, by averaging the samples from the posterior distribution. To make our sampling-based method practical, we extend the auxiliary-variable Gibbs sampler introduced in Sec. 3.1.2 to efficiently sample the posterior in (4.2). The scales of the Gaussian scale mixtures

(GSMs) used for representing the potential functions of the MRF prior $p(\mathbf{x};\mathbf{\Omega})$ are retained as hidden variable $\mathbf{z}$. To draw samples from the posterior, we alternate between sampling the hidden scales according to $p(\mathbf{z}|\mathbf{x},\mathbf{y};\mathbf{\Omega}) = p(\mathbf{z}|\mathbf{x};\mathbf{\Omega})$ (remaining as before in (3.8)) and sampling the image according to

$$\begin{aligned} p(\mathbf{x}|\mathbf{y},\mathbf{z};\mathbf{\Omega}) &\propto p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x}|\mathbf{z};\mathbf{\Omega}) \\ &\propto \exp\Big(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|^2\Big) \cdot \exp\Big(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x}\Big) \\ &\propto \exp\Big(\frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{y} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\Big(\frac{1}{\sigma^2}\mathbf{I} + \mathbf{\Sigma}^{-1}\Big)\mathbf{x}\Big) \\ &\propto \mathcal{N}\Big(\mathbf{x}; \tfrac{1}{\sigma^2}\widetilde{\mathbf{\Sigma}}\mathbf{y}, \widetilde{\mathbf{\Sigma}}\Big), \end{aligned} \tag{4.5}$$

where $\widetilde{\mathbf{\Sigma}} = \big(\frac{1}{\sigma^2}\mathbf{I} + \mathbf{\Sigma}^{-1}\big)^{-1}$ with $\mathbf{\Sigma}$ defined as in (3.6). The conditional distribution in (4.5) is also a Gaussian and thus the sampling is very efficient. Note that the conditional mean $\frac{1}{\sigma^2}\widetilde{\mathbf{\Sigma}}\mathbf{y}$ in (4.5) corresponds to the case of MRF priors with potential functions represented by Gaussian scale mixtures (GSM, *cf*. Eq. (3.2)). For the case of potentials represented by mixtures of GSMs (MGSMs, *cf*. Sec. 3.4.1) or general mixtures of Gaussians (MoGs), the conditional mean can be derived analogously. In practice, to make inference more efficient, we approximate the MMSE estimate using the Rao-Blackwellized estimator [118], which averages the conditional expectations from $p(\mathbf{x}|\mathbf{y},\mathbf{z};\mathbf{\Omega})$ (e.g., the mean of the Gaussian $\frac{1}{\sigma^2}\widetilde{\mathbf{\Sigma}}\mathbf{y}$ in (4.5)) during Gibbs sampling.

MAP estimation

To highlight the benefits of using the MMSE estimate, we also compute the maximum a-posteriori (MAP) estimate which is more common in the literature

$$\hat{\mathbf{x}} = \arg\max_{\tilde{\mathbf{x}}} p(\mathbf{x}|\mathbf{y};\mathbf{\Omega}). \tag{4.6}$$

However, maximizing the posterior probability is not easy in our case of using the learned priors. To do this, a gradient ascent on the logarithm of the posterior $p(\mathbf{x}|\mathbf{y};\mathbf{\Omega})$ leads to the update (*cf*. (2.16))

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \tau\Big[\sum_j \mathbf{f}_j * \boldsymbol{\varphi}'(\tilde{\mathbf{f}}_j * \mathbf{x}^{(t)}; \boldsymbol{\omega}_j) + \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{x}^{(t)})\Big] \tag{4.7}$$

where $\tilde{\mathbf{f}}_j$ is the flipped version of filter $\mathbf{f}_j$, "$*$" denotes convolution, $\boldsymbol{\varphi}'(\cdot)$ is the vector of derivatives of all log-experts $\log\phi(\cdot_i)$, and τ is the stepsize.

In our experiments, instead of directly using this gradient ascent procedure, we

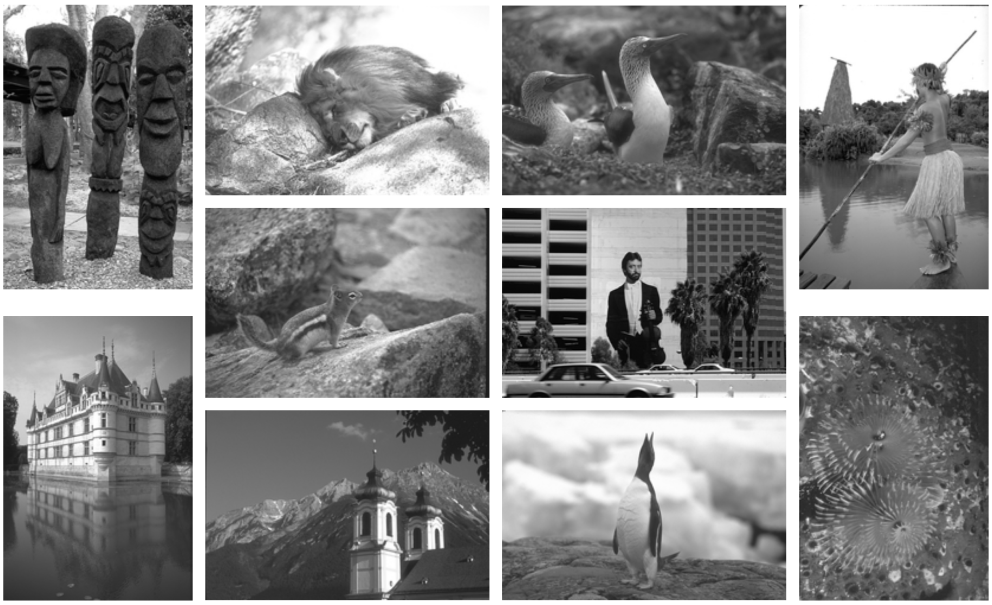


Figure 4.1: The test set of 10 images [46].

employ a more efficient conjugate gradients (CG) method implemented by Rasmussen [119]. Moreover, we maximize the posterior $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x})^{\lambda}$ where λ is an optional regularization weight, which has frequently been employed to obtain good application performance (e.g., [7] and common variational methods).

4.1.2 Experiments

We test the denoising performance of our learned MRF priors (see Sec. 3.3) on two different test sets from the Berkeley segmentation dataset [105]: A set of 10 images (160×240 pixels) used in [46] (see Fig. 4.1), and a set of 68 images (320×480 pixels) used in [120, 7, 102].

To evaluate the quality of restored images, we use two measurements: The peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) index [114]. The PSNR is a widely used evaluation criterion. It is defined via the mean squared (pixel-wise) image error $\bar{e^2}$

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{I_{\max}^2}{\bar{e^2}}\right), \tag{4.8}$$

where $I_{\max}$ is the maximum intensity value of an image. For an 8-bit image, $I_{\max} = 255$. According to the definition, a higher PSNR value corresponds to a better quality of the image. As the PSNR can not fully reflect the perceptual quality of an image to human eyes, we also employ the SSIM, which provides a perceptually more plausible measure. SSIM values range between 0 and 1. A higher value means better image quality.

Table 4.1: Denoising results for 10 test images [46].

(a) Average PSNR in dB

Model	MAP		MAP with λ		MMSE	
	σ=10	σ=20	σ=10	σ=20	σ=10	σ=20
pairwise (marginal fit [31])	28.35	23.96	30.98	26.92	29.70	24.72
pairwise (generalized Lapl. [5])	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 (Student-t [46])	28.19	24.04	31.22	26.70	29.38	24.68
pairwise (**ours**)	**30.27**	**26.48**	30.41	26.55	**32.09**	**28.32**
5×5 FoE from [7]	27.92	23.81	**32.63**	**28.92**	29.38	24.95
15×15 FoE from [39]	22.51	20.45	32.27	28.47	23.22	21.47
3×3 FoE (**ours**)	**30.49**	**25.83**	31.82	27.39	**32.94**	**29.04**

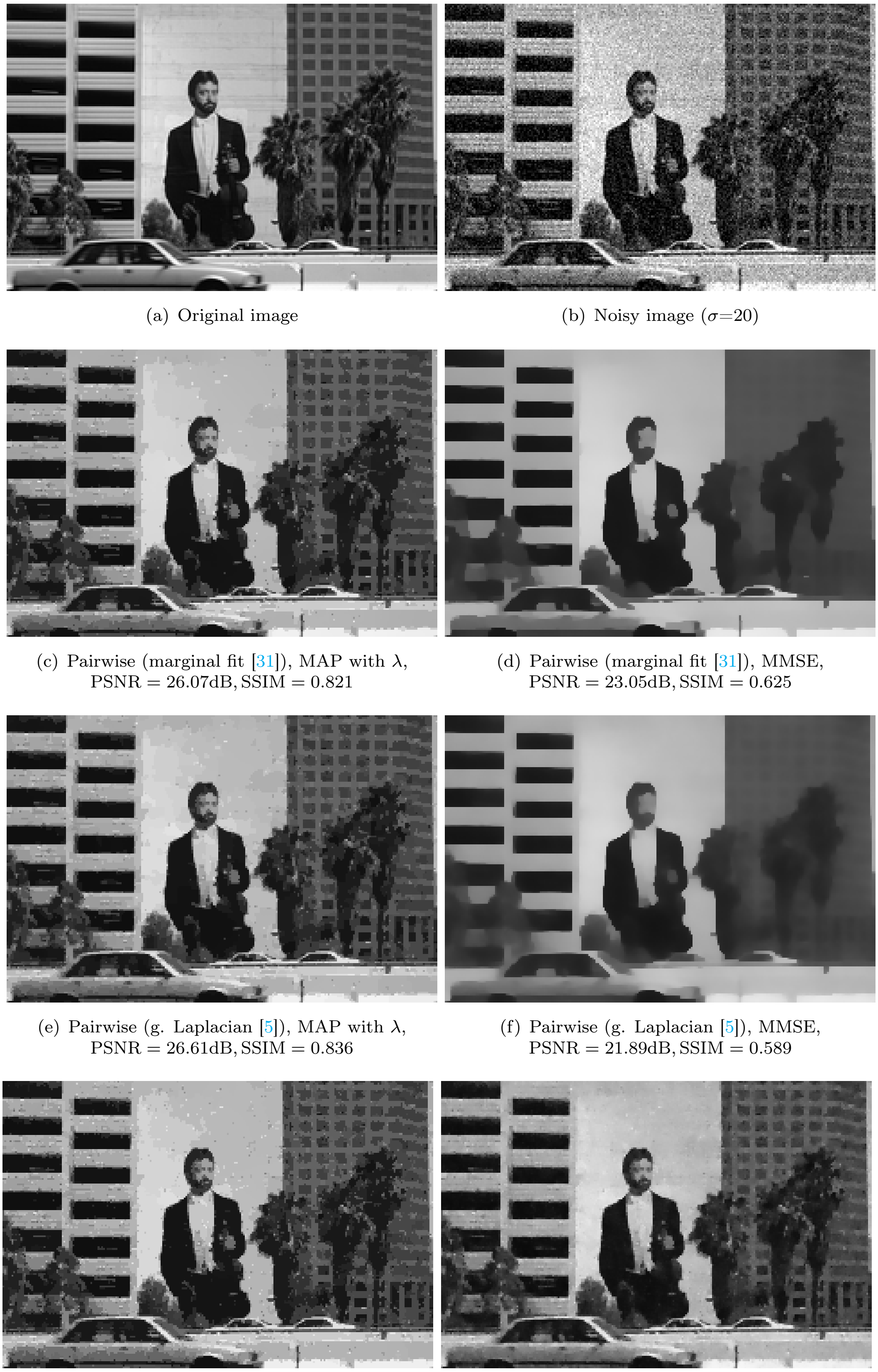
(b) Average SSIM [114]

Model	MAP		MAP with λ		MMSE	
	σ=10	σ=20	σ=10	σ=20	σ=10	σ=20
pairwise (marginal fit [31])	0.787	0.599	0.873	0.748	0.833	0.631
pairwise (generalized Lapl. [5])	0.745	0.559	0.886	0.777	0.804	0.609
pairwise (Laplacian)	0.832	0.624	**0.897**	**0.801**	0.869	0.703
pairwise (Student-t [46])	0.784	0.602	0.887	0.746	0.825	0.631
pairwise (**ours**)	**0.855**	**0.720**	0.854	0.725	**0.904**	**0.808**
5×5 FoE from [7]	0.753	0.595	**0.913**	**0.833**	0.826	0.657
15×15 FoE from [39]	0.515	0.445	0.903	0.820	0.564	0.489
3×3 FoE (**ours**)	**0.846**	**0.677**	0.899	0.781	**0.924**	**0.842**

Results on the test set of 10 images

We compare our learned models against pairwise MRFs with four potential functions (standard and generalized Laplacian [5], Student-t [46], and a marginal fit as in [31]) as well as two FoE models [7, 39]. As can be seen in Tab. 4.1, when using MAP estimation without the regularization weight, our learned models perform rather poorly despite good generative properties. With an optional regularization weight λ (optimized on the test set), improved denoising results can be achieved, but are still worse than those of previous models. Moreover, such a regularization weight deteriorates the generative properties (at least for our models).

To approximate the MMSE estimate, we run four 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 (see Sec. 3.1.2). Fig. 4.3 shows that the Gibbs sampler mixes rapidly when sampling the posterior. After discarding the "burn-in" samples from the first few sampling iterations, we average all conditional means of (4.5) from subsequent



(a) Original image

(b) Noisy image (σ=20)

(c) Pairwise (marginal fit [31]), MAP with λ, $\mathrm{PSNR} = 26.07\mathrm{dB}, \mathrm{SSIM} = 0.821$

(d) Pairwise (marginal fit [31]), MMSE, $\mathrm{PSNR} = 23.05\mathrm{dB}, \mathrm{SSIM} = 0.625$

(e) Pairwise (g. Laplacian [5]), MAP with λ, $\mathrm{PSNR} = 26.61\mathrm{dB}, \mathrm{SSIM} = 0.836$

(f) Pairwise (g. Laplacian [5]), MMSE, $\mathrm{PSNR} = 21.89\mathrm{dB}, \mathrm{SSIM} = 0.589$

(g) Pairwise (ours), MAP with λ, $\mathrm{PSNR} = 25.79\mathrm{dB}, \mathrm{SSIM} = 0.806$

(h) Pairwise (ours), MMSE, $\mathrm{PSNR} = 27.06\mathrm{dB}, \mathrm{SSIM} = 0.851$

Figure 4.2: Denoising examples of different pairwise MRFs based on MAP (with λ) and MMSE.

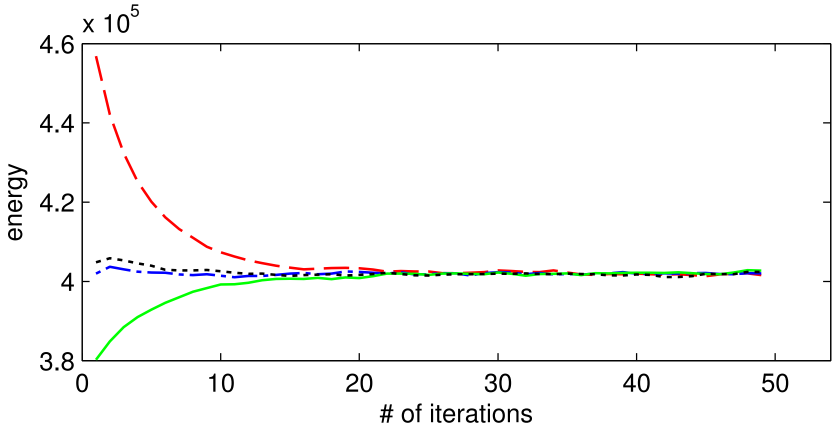


Figure 4.3: Fast mixing of the Gibbs sampler when sampling an image of 160 × 240 pixels from the denoising posterior (a Gaussian likelihood $\sigma = 20$ with the learned pairwise MRF prior). Four chains with over-dispersed initializations (red, dashed: noisy image; blue, dash-dotted: Gauss filtered version; green, solid: median filtered version; black, dotted: Wiener filtered version). Approximate convergence is reached after 24 iterations (the potential scale reduction $\hat{R} < 1.1$, *cf*. Sec. 3.1.2).

iterations until the average images from the four samplers are sufficiently close to one another. Finally, the restored image is obtained by averaging these four average images.

The results of MMSE estimation are compared with those of MAP estimation in Tab. 4.1. We can see that the MMSE outperforms MAP without a regularization weight. When applied to good generative models such as our learned ones, the MMSE is even superior to MAP with an optimal regularization weight. Note that MMSE estimation operates in a purely generative setting with no regularization weight required. Fig. 4.2 shows denoising examples of different pairwise MRFs based on MAP (with optimal regularization weight λ) and MMSE.

Denoising results on the test set of 68 images

More extensive experiments on 68 test images [7, 102] with additive Gaussian noise confirm our findings. Table 4.2 shows that, using MMSE-based denoising, even our learned pairwise MRF outperforms the FoE of [7] using MAP, despite their much larger 5×5 cliques and noise-adaptive regularization weight. Our learned FoE models improve the denoising performance consistently with larger and more filters, which coincides with the improvements of the model statistics (*cf*. Fig. 3.12). With MMSE denoising using posterior sampling, our 5 × 5 FoE even outperforms the results of [102] by 0.4dB. This is remarkable since this discriminative approach explicitly maximizes the denoising performance of MAP estimates, and furthermore uses more experts. In consequence, MMSE estimation enables application-independent generative MRFs to

Table 4.2: Denoising results for 68 test images [7, 102] ($\sigma = 25$).

Model	Inference	avg. PSNR	avg. SSIM
5×5 FoE from [7] (24 filters)	MAP w/λ	27.44	0.746
5×5 FoE from [102] (24 filters)	MAP	27.86	0.776
Pairwise MRF (ours)	MMSE	27.54	0.758
MGSM-FRAME (ours)	MMSE	27.80	0.779
3×3 FoE (ours, 8 filters)	MMSE	28.19	0.795
5×5 FoE (ours, 8 filters)	MMSE	28.22	0.797
5×5 FoE (ours, 16 filters)	MMSE	**28.26**	**0.799**
Non-local means [116]	(MMSE)	27.50	0.734
BLS-GSM [32]	MMSE	28.02	0.789
NLSC [54]	–	28.28	**0.799**
BM3D [55]	–	**28.35**	0.797

be competitive with MAP-based denoising-specific discriminative MRFs.

We also compared our method to two MMSE-related methods that are not based on MRFs: non-local means [116] (with tuned parameters), and the wavelet-based BLS-GSM [32]; we clearly outperform them despite not being limited to denoising. More importantly, our approach is competitive with popular BM3D (particularly in SSIM [114]) as well as NLSC [54]. As far as we are aware, this is the first time such competitive denoising performance has been achieved with any generative, global model of natural images.

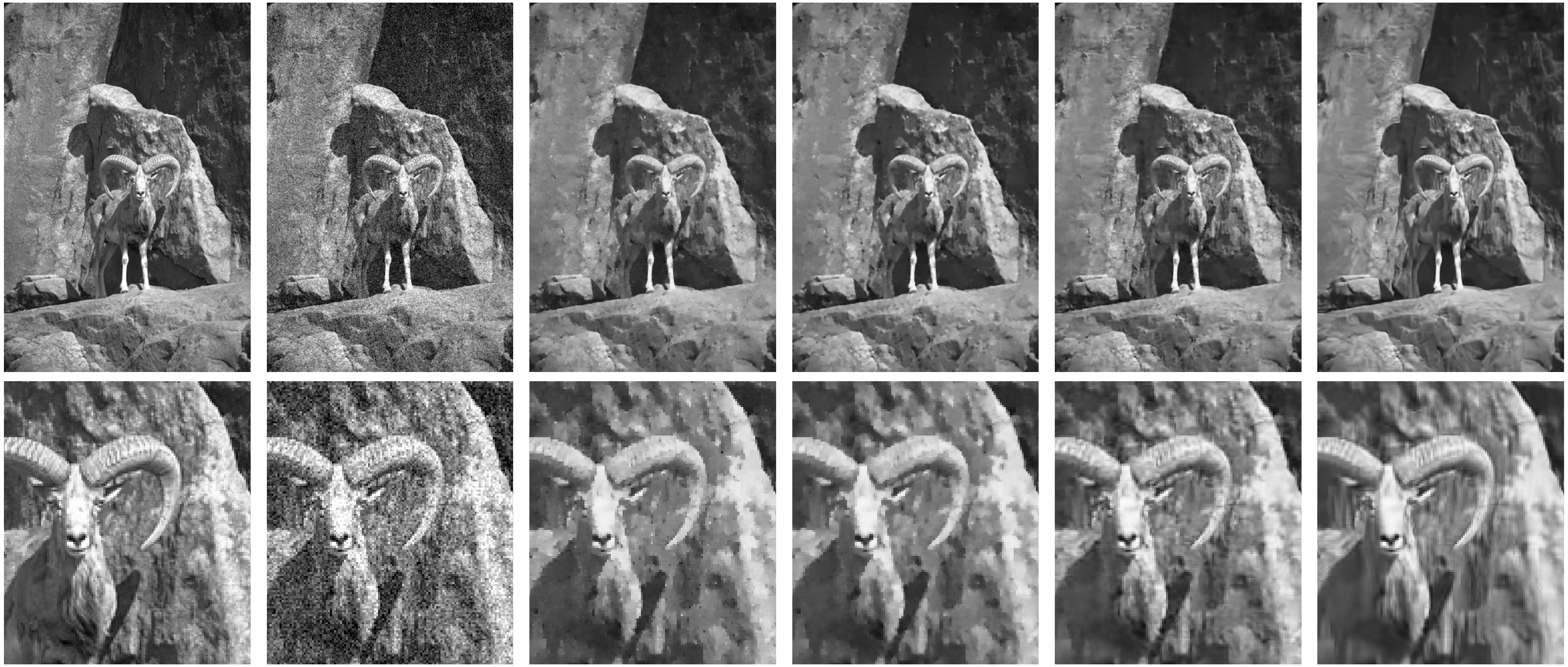
In addition, we evaluate our reimplemented FRAME model [6] using MGSM-based potential functions (see Sec. 3.4 for details). Note that the MGSM-FRAME model with MMSE estimation not only achieves much better denoising results than the original GRADE procedure in [6], but also performs competitively at a level that is not much below the popular methods despite of its age. The MGSM-FRAME is inferior to our learned high-order FoE models, which may be attributed to fewer potential functions it uses, or to the fact that its filters are not learned.

Figs. 4.4, 4.5 and 4.6 show denoising results for three of the 68 images, for which the average performance is reported in Tab. 4.2. Note that in contrast to the tested previous approaches, combining our learned models with MMSE leads to good performance on relatively smooth as well as on strongly textured images.

4.1.3 Discussion: MMSE *vs.* MAP

Problems of MAP

Our experiments have shown that, despite very good generative properties, our learned models perform rather poorly when using MAP estimation no matter whether a regularization weight is used or not. Nikolova [115] showed this to be an intrinsic



(a) Original

(b) Noisy image (σ=25)

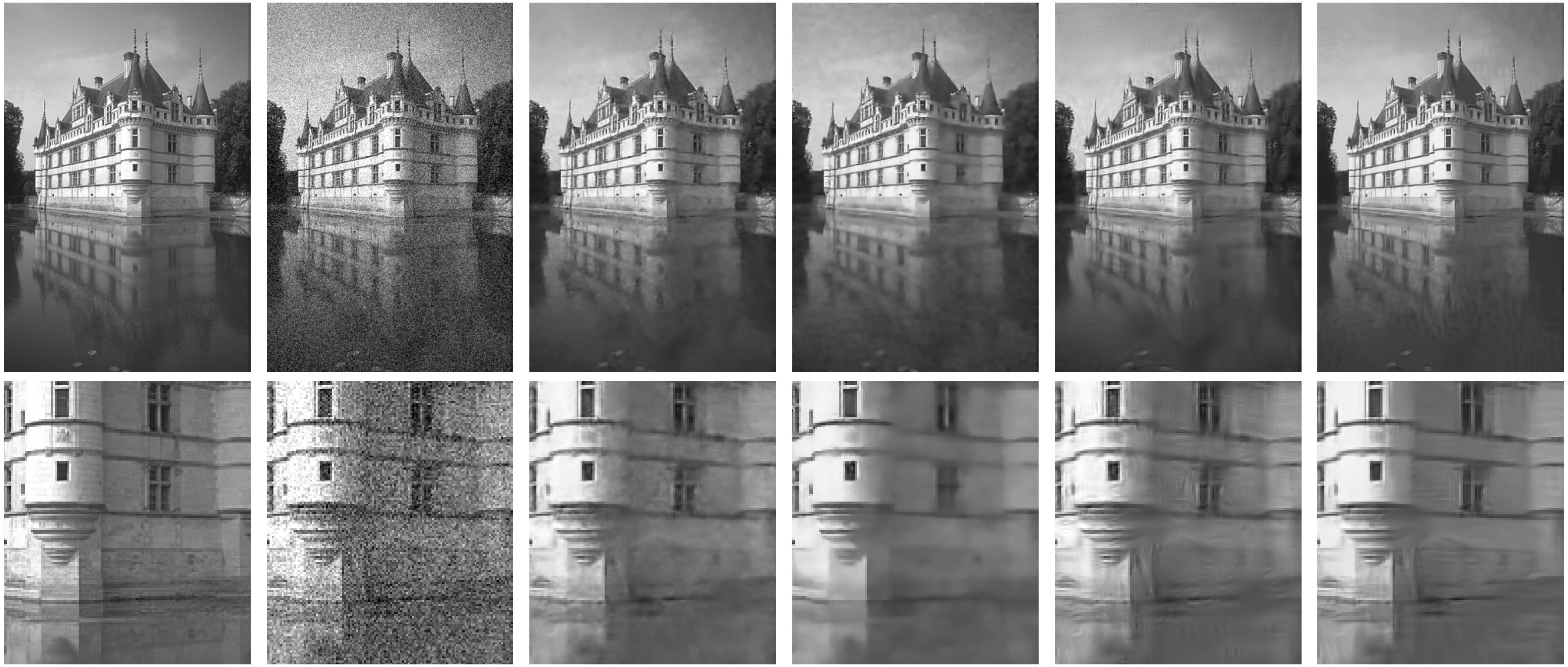
(c) Our pairwise MRF, PSNR = 26.12, SSIM = 0.685

(d) Our MGSM-FRAME, PSNR = 26.24, SSIM = 0.689

(e) Our 5×5 FoE, PSNR = 26.46, SSIM = 0.700

(f) BM3D [55], PSNR = 26.23, SSIM = 0.674

Figure 4.4: Image denoising example. *(top)* Full image; *(bottom)* detail.



(a) Original

(b) Noisy image (σ=25)

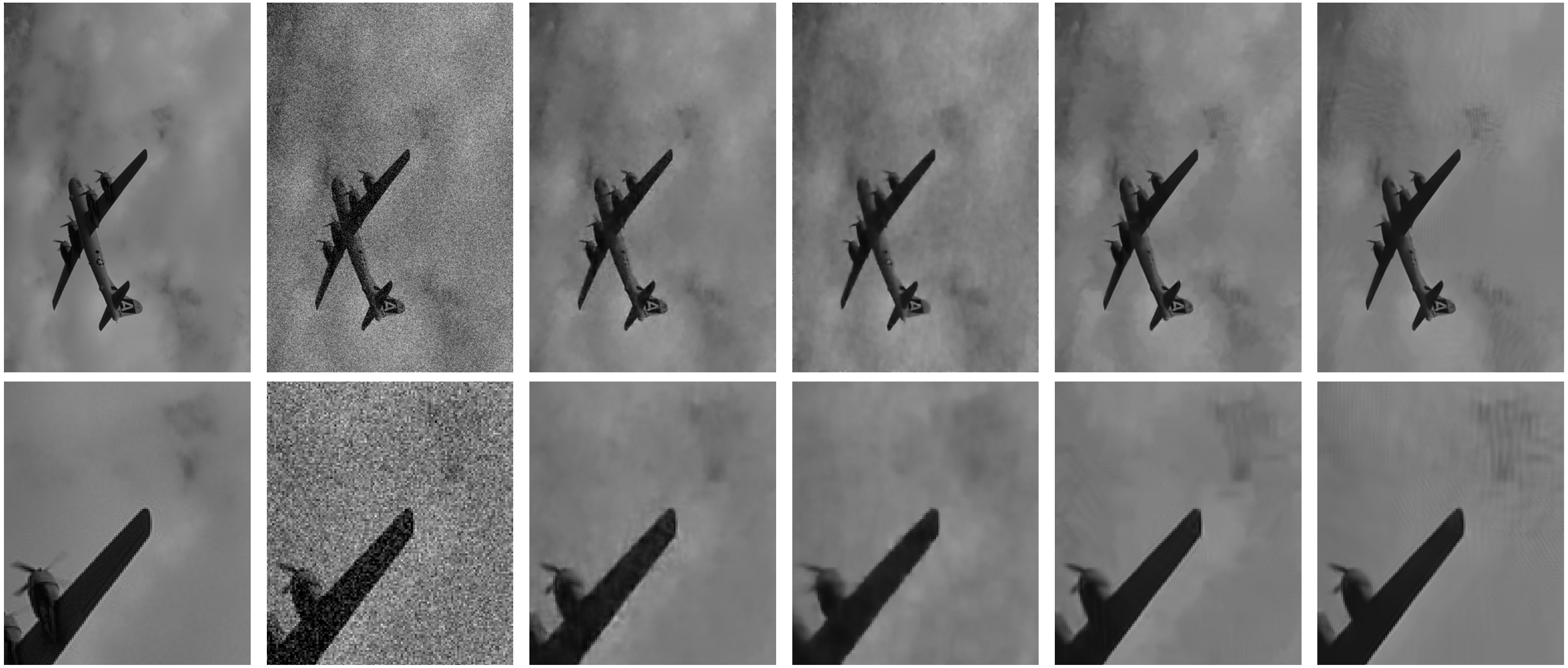
(c) Our 5×5 FoE,
PSNR = 29.23,
SSIM = 0.841

(d) 5×5 FoE from [7],
PSNR = 28.52,
SSIM = 0.816

(e) BLS-GSM [32],
PSNR = 28.99,
SSIM = 0.830

(f) NLSC [54],
PSNR = 29.30,
SSIM = 0.844

Figure 4.5: Image denoising example. *(top)* Full image; *(bottom)* detail.



(a) Original

(b) Noisy image (σ=25)

(c) Our 5×5 FoE, PSNR = 36.08, SSIM = 0.948

(d) 5×5 FoE from [7], PSNR = 35.00, SSIM = 0.938

(e) BM3D [55], PSNR = 36.65, SSIM = 0.951

(f) NLSC [54], PSNR = 35.08, SSIM = 0.942

Figure 4.6: Image denoising example. *(top)* Full image; *(bottom)* detail.

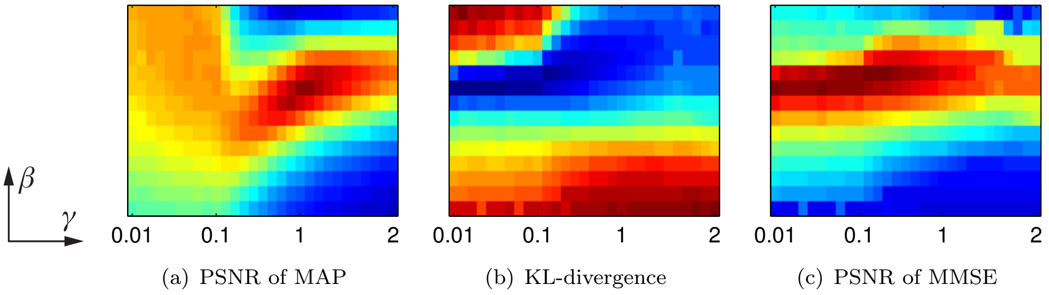
problem of MAP estimation. To better understand this, we analyze the denoising performance of pairwise MRFs with a wide range of potentials from the family of generalized Laplacians $\phi(y; \beta, \gamma) = \exp\bigl(-\beta|y^2 + \epsilon|^{\frac{\gamma}{2}}\bigr)$, where β controls the width of the 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. 4.7 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 [115], 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.

More benefits from MMSE estimation

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 (see [56] and Fig. 4.8). Woodford *et al.* [56] developed models that explicitly enforce certain statistical properties of the MAP estimate, but needed to abolish the well-understood MRF framework and had to rely on a rather complex inference procedure. From a practical point of view, Fig. 4.8 shows that we do not need to replace MRFs, but that replacing MAP with MMSE estimation is already sufficient to achieve the desired statistics of the output image and circumventing this long-standing problem. Second, Fig. 4.7 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.

Efficiency analysis of sampling-based MMSE approximation

Fig. 4.9 shows the evolution of the PSNR over the number of samples for one example. Computing the MMSE using sampling is practical, despite our simple implementation, and much superior to using only a single sample [42]. Running multiple samplers results in less correlation between model samples, thus improving the model performance (faster convergence of the denoised image), even when using sequential computing (Fig. 4.9(a)). Since communication between the samplers is not needed during the Gibbs sampling, parallel computing is easy to implement for accelerating the computation. Fig. 4.9(b) shows that, in the case of parallel computing (one sampler per computing core), faster convergence will also be achieved.



(a) PSNR of MAP

(b) KL-divergence

(c) PSNR of MMSE

Figure 4.7: Correlation between generative properties and denoising performance: Pairwise MRF with generalized Laplacian potential and different parameters. Test image is "Lena" of 128×128 pixels, $\sigma = 20$. *(a)* MAP denoising with conjugate gradient (red: high PSNR); $\text{PSNR}_{\text{max}} = 28.07\text{dB}$. *(b)* KL-divergence between derivative statistics of images and model marginals (blue: low D_{KL}). *(c)* MMSE denoising with sampling; $\text{PSNR}_{\text{max}} = 28.26\text{dB}$. While the correlation between D_{KL} and PSNR of MAP is low (normalized cross-correlation NCC $= -0.43$), the D_{KL} and PSNR of MMSE are highly correlated (NCC $= -0.84$).

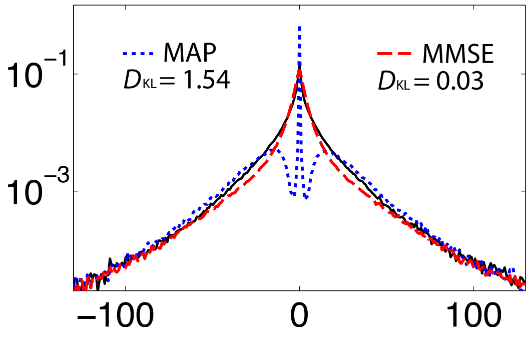
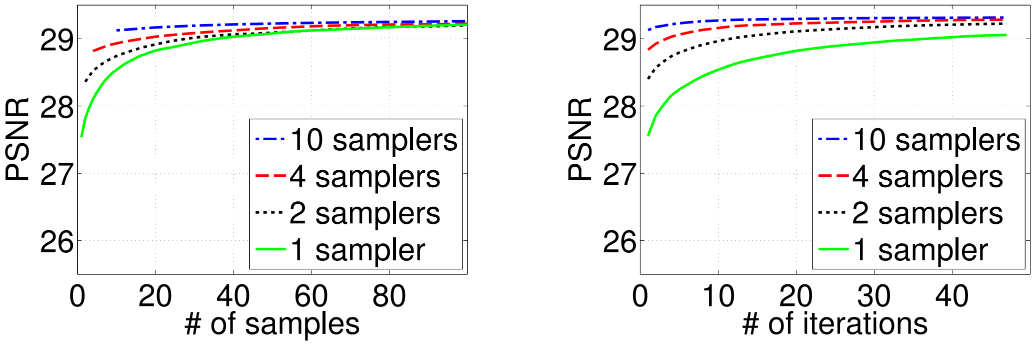


Figure 4.8: Derivative statistics of 10 denoised test images and of corresponding noise-free originals (black, solid), σ=10, 20.



(a) Sequential computing

(b) Parallel computing

Figure 4.9: Efficiency of sampling-based MMSE denoising with different number of samplers. Learned pairwise MRF, $\sigma = 20$.

4.2 Microscopy image deconvolution

4.2.1 Introduction and related work

Images acquired by photon-counting devices in applications such as biology, medicine, and astronomy are usually degraded by noise with mixed Poisson-Gaussian (PG) statistics and by blurring characterized by the system point spread function. In previous work, besides denoising methods dedicated to PG noise (e.g., [121, 122, 123]), methods for joint deconvolution and PG denoising (e.g., [124, 125, 126, 127, 128, 49]) have been proposed. Recently, Gazagnes *et al.* [129] proposed a joint super-resolution and denoising method for single-molecule localization microscopy (SMLM). However, there Poisson noise is treated as additive Gaussian noise (by using a quadratic data fidelity term).

Most of the methods mentioned above are formulated within a maximum a posteriori (MAP) framework [123, 124, 125, 126, 127, 129]. To achieve best performance, a regularization parameter which defines the trade-off between the data fidelity term and the regularization term must be tuned manually. In contrast to MAP, [121] and [128] avoid such problem by approximating the minimum mean square error (MMSE) estimate. More recently, Marnissi *et al.* [49] proposed a Bayesian deconvolution method that computes the MMSE estimate and the regularization parameter through variational Bayesian approximation. In general, as has been shown in Sec. 4.1, the usage of an application-dependent regularization parameter suggests that common regularizers (e.g., total variation) employed in these MAP-based methods as well as in [49] are insufficient as a generative image prior.

A Bayesian MMSE estimate, which is computed based on the whole model distribution, exploits more information and is statistically more sound than an MAP estimate. However, computing the MMSE estimate is generally difficult for non-convex problems. In methods for denoising and deblurring for additive Gaussian noise [50, 130], a trained non-convex probabilistic generative image prior is used and the MMSE estimate is approximated by averaging model samples based on an efficient block Gibbs sampler. As the MMSE estimation operates in a purely generative setting, a regularization parameter is not needed. However, this efficient Gibbs sampler cannot handle the case of PG noise due to the more complex statistics.

In this section, we propose a Bayesian deconvolution and denoising model for mixed Poisson-Gaussian noise and determine the MMSE solution. We explore the statistics of PG noise and find that its distribution can be well approximated using mixtures of Gaussians (MoGs). A high-order MRF image prior from Sec. 3.3 is employed and a regularization parameter is not needed. The MoG-based likelihood (data fidelity) and the Gaussian scale mixture (GSM)-based prior allow to augment the model using hidden variables and define an efficient block Gibbs sampler. The degraded images are restored using the MMSE estimate that is approximated by averaging multiple samples from our probabilistic model. We also integrate super-resolution into our model framework and simultaneously achieve super-resolution, deconvolution, and denoising. We have applied our approach to different types of synthetic data and performed a comparison with previous methods. Experiments demonstrate that our joint method is superior to a sequential scheme, and that the deconvolution performance can compete with the popular methods. Our method is also applied to real microscopy images of telomeres acquired via stimulated emission depletion (STED) nanoscopy. The work was published in [58].

4.2.2 Joint super-resolution, deconvolution and denoising

Model formulation

We formulate image restoration from a single image as a posterior distribution in a Bayesian framework

$$p(\mathbf{x}|\mathbf{y};\mathbf{\Theta},\mathbf{\Omega}) \propto p(\mathbf{y}|\mathbf{x};\mathbf{\Theta}) \cdot p(\mathbf{x};\mathbf{\Omega}), \tag{4.9}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the given degraded image (noisy, blurry, and low-resolution), $\mathbf{x} \in \mathbb{R}^{r^2 N}$ is the image to be restored (with a scaling factor r in case of super-resolution), N is the number of pixels in the image, and $\mathbf{\Theta}$ and $\mathbf{\Omega}$ are the likelihood and prior parameters, respectively.

Likelihood model. For the case of Poisson-Gaussian noise, the intensity of each pixel y_i in the observed image $\mathbf{y}$ can be modeled as

$$y_i = \alpha t_i + n_i, \tag{4.10}$$

where t_i follows a Poisson distribution

$$t_i \sim \text{Poi}\left(\tfrac{1}{\alpha}(\mathbf{S}(\mathbf{k} * \mathbf{x}))_i\right) = \text{Poi}\left((\tfrac{1}{\alpha}\mathbf{H}\mathbf{x})_i\right) := \text{Poi}(\tilde{x}_i), \tag{4.11}$$

and $\mathbf{H}$ is the matrix representing the convolution of $\mathbf{x}$ with some point spread function $\mathbf{k}$ and down-sampling by $\mathbf{S} \in \mathbb{R}^{N \times r^2 N}$, α is a gain factor that controls the Poisson noise strength, and n_i is additive Gaussian noise.

$$n_i \sim \mathcal{N}(\mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2). \tag{4.12}$$

We further assume that pixels in $\mathbf{y}$ are conditionally independent. The likelihood model can then be written as

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x};\boldsymbol{\Theta}) &= \prod_i p(y_i|\tilde{x}_i) \\
&= \prod_i \left(\sum_{t_i=0}^{+\infty} p(y_i|t_i) \cdot p(t_i|\tilde{x}_i) \right) \\
&= \prod_i \left(\sum_{t_i=0}^{+\infty} \frac{e^{-\frac{(y_i-\alpha t_i-\mu_\mathrm{n})^2}{2\sigma_\mathrm{n}^2}}}{\sqrt{2\pi}\sigma_\mathrm{n}} \frac{\tilde{x}_i^{t_i} \cdot e^{-\tilde{x}_i}}{t_i!} \right).
\end{aligned}
\tag{4.13}
$$

Prior model. We use the high-order MRF image prior from Sec. 3.1 in which the potential functions are represented by Gaussian scale mixtures (GSM)

$$
\begin{aligned}
p(\mathbf{x};\boldsymbol{\Omega}) &\propto \prod\nolimits_i \prod\nolimits_j \phi\big((\mathbf{f}_j*\mathbf{x})_i;\boldsymbol{\omega}_j\big) \\
&\propto \prod\nolimits_i \prod\nolimits_j \Big(\sum\nolimits_k \omega_{jk} \cdot \mathcal{N}\big((\mathbf{f}_j*\mathbf{x})_i; 0, s_k \cdot \sigma_j^2\big)\Big),
\end{aligned}
\tag{4.14}
$$

where $\mathbf{f}_j$ are linear filters and "$*$" denotes convolution. In the GSM-based potential functions $\phi\big(\cdot;\boldsymbol{\omega}_j\big)$, the parameters $\boldsymbol{\omega}_j$ include the mixture weights $\omega_{jk}>0, \sum_k \omega_{jk}=1$ of the zero-mean Gaussian components with scales s_k and base variance σ_j^2. The model parameters $\boldsymbol{\Omega}=\{\mathbf{f}_j, \omega_{jk}\}$ can be learned unsupervisedly from training images (see Sec. 3.3 for details on the learning procedure).

Sampling-based inference

In Sec. 4.1.3 we have shown that, when applying our learned MRF priors to a Bayesian posterior model, the MMSE estimate is superior to the MAP estimate with additional good properties (e.g., desired image statistics, no regularization parameter required). As there is no closed form expression for the MMSE estimate of our model, we develop a sampling-based approximation method.

Given $\mathbf{y}$, the likelihood $p(y_i|\tilde{x}_i)$ is a function of $\tilde{x}_i$, which has an asymmetric bell shape (Fig. 4.10, the black solid curve). Despite the complex mathematical expression, the model likelihood in (4.13) can be well approximated by fitting these functions using general mixtures of Gaussians (MoGs):

$$
p\big(y_i|\tilde{x}_i\big) \propto p\big(\tilde{x}_i; y_i\big) = \sum\nolimits_{l=1}^{M} \pi_{y_i l} \cdot \mathcal{N}\big(\tilde{x}_i; \tilde{\mu}_{y_i l}, \tilde{\sigma}_{y_i l}^2\big),
\tag{4.15}
$$

where $\pi_{y_i l}>0$, $\tilde{\mu}_{\cdot l}$ and $\tilde{\sigma}_{\cdot l}^2$ are the means and variances of the Gaussian components, respectively. Since in practice the value of y_i is usually within a finite set (or can be discretized for real values), the Gaussian mixture parameters can be obtained in advance before inference. Note that the MoGs in (4.15) are different to the GSMs

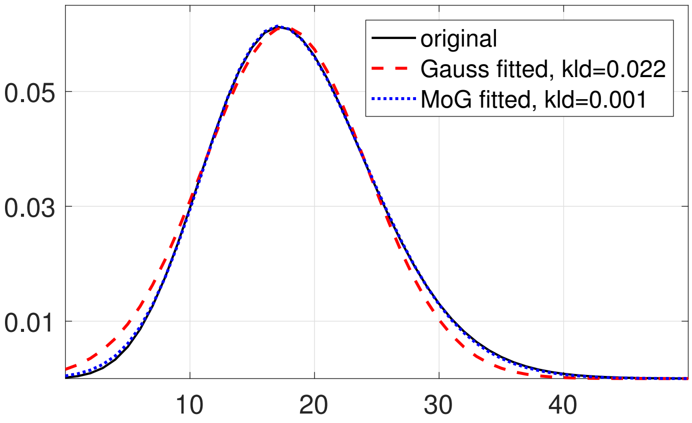


Figure 4.10: Likelihood values $p(y_i|\tilde{x}_i)$ of (4.15) w.r.t. $\tilde{x}_i$ (black, solid) for $y_i = 20$ with $\alpha = 2$, $\mu_n = -15$, and $\sigma_n = 10$. Model fitting using a mixture of Gaussians with three components (blue, dotted) is more accurate than using a Gaussian (red, dashed). "kld" means KL-divergence (a smaller value corresponds to a better fit).

in (4.14); for the latter, the means of the Gaussian components are set to zero.

The scales of the GSMs in (4.14) are retained as discrete-valued hidden random variable $z_{ji} \in \{1, ..., K\}$ such that $p(\mathbf{x}; \mathbf{\Omega}) = \sum_{\mathbf{z}} p(\mathbf{z}, \mathbf{x}; \mathbf{\Omega})$, allowing to define a rapidly mixing auxiliary-variable Gibbs sampler (*cf*. Sec. 3.1.2). Following the strategy, we further augment the mixture model in (4.15) using the discrete-valued hidden variable $h_i \in \{1, ..., M\}$, which represents the index of the Gaussian mixture component, such that $p(\mathbf{x}; \mathbf{y}) = \sum_{\mathbf{h}} p(\mathbf{h}, \mathbf{x}; \mathbf{y})$. The following conditional distributions can be derived

$$p(h_i|\mathbf{x}, y_i) \propto \pi_{y_i h_i} \cdot \mathcal{N}\big((\tfrac{1}{\alpha}\mathbf{H}\mathbf{x})_i; \tilde{\mu}_{y_i h_i}, \tilde{\sigma}^2_{y_i h_i}\big), \tag{4.16}$$

$$p(z_{ji}|\mathbf{x}) \propto \omega_{j z_{ji}} \cdot \mathcal{N}\big((\mathbf{f}_j * \mathbf{x})_i; 0, s_{z_{ji}} \sigma_j^2\big), \tag{4.17}$$

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}, \mathbf{h}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x}, \mathbf{h}) \cdot p(\mathbf{x}|\mathbf{z}) \\ &\propto \prod\nolimits_i \mathcal{N}\big((\tfrac{1}{\alpha}\mathbf{H}\mathbf{x})_i; \mu_{y_i h_i}, \sigma^2_{y_i h_i}\big) \cdot \prod\nolimits_i \prod\nolimits_j \mathcal{N}\big(\mathbf{f}_{ji}^{\mathrm{T}}\mathbf{x}; 0, s_{z_{ji}} \sigma_j^2\big) \\ &\propto \mathcal{N}\Big(\tfrac{1}{\alpha}\mathbf{H}\mathbf{x}; \mathbf{b}, \mathbf{Q}^{-1}\Big) \cdot \mathcal{N}\Big(\mathbf{x}; \mathbf{0}, \mathbf{P}^{-1}\Big) \\ &\propto \mathcal{N}\Big(\mathbf{x}; \tfrac{1}{\alpha}\mathbf{\Sigma}\mathbf{H}^{\mathrm{T}}\mathbf{Q}\mathbf{b}, \mathbf{\Sigma}\Big), \end{aligned} \tag{4.18}$$

where $\mathbf{f}_{ji}$ is defined as $\mathbf{f}_{ji}^{\mathrm{T}}\mathbf{x} = (\mathbf{f}_j * \mathbf{x})_i$, $\mathbf{b} = [..., \tilde{\mu}_{y_i h_i}, ...]^{\mathrm{T}}$, $\mathbf{Q} = \mathrm{diag}\{1/\tilde{\sigma}^2_{y_i h_i}\}$, $\mathbf{P} = \sum_i \sum_j \frac{1}{s_{z_{ji}} \sigma_j^2} \mathbf{f}_{ji} \mathbf{f}_{ji}^{\mathrm{T}}$ and $\mathbf{\Sigma} = \big(\frac{1}{\alpha^2}\mathbf{H}^{\mathrm{T}}\mathbf{Q}\mathbf{H} + \mathbf{P}\big)^{-1}$.

We can then define a rapidly mixing block Gibbs sampler that alternatively samples $\mathbf{h}, \mathbf{z}$, and $\mathbf{x}$ according to (4.16), (4.17) and (4.18), respectively. This is very efficient, as (4.16) and (4.17)) are discrete distributions conditionally independent to each other, and (4.18)) is a multivariate Gaussian that can be sampled through solving a linear system of equations. The MMSE estimate $\hat{\mathbf{x}}$ can be approximated by

Algorithm 2 Sampling-based MMSE estimate approximation

```
initialization of $\mathbf{x}^{(0)} \in \mathbb{R}^{r^2N}$;
for $t = 1, 2, \ldots$ do
    sample $\mathbf{h}^{(t)} \sim p(\mathbf{h}|\mathbf{x}^{(t-1)}, \mathbf{y})$ according to (4.16);
    sample $\mathbf{z}^{(t)} \sim p(\mathbf{z}|\mathbf{x}^{(t-1)})$ according to (4.17);
    update $\mathbf{b}^{(t)}, \mathbf{Q}^{(t)}$ and $\mathbf{\Sigma}^{(t)}$ in (4.18);
    sample $\mathbf{x}^{(t)} \sim p(\mathbf{x}|\mathbf{z}^{(t)}, \mathbf{h}^{(t)}, \mathbf{y})$ according to (4.18);
    if $t > B$ then                                ▷ $B$: no. of "burn-in" iterations
        compute $\hat{\mathbf{x}}^{(t)} = \frac{1}{t-B} \sum_{i=B+1}^{t} \frac{1}{\alpha} \mathbf{\Sigma}^{(i)} \mathbf{H}^{\mathrm{T}} \mathbf{Q}^{(i)} \mathbf{b}^{(i)}$;
        if $\|\hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}^{(t-1)}\| \leqslant \epsilon$ then
            return $\hat{\mathbf{x}}^{(t)}$;
```

averaging multiple samples of $\mathbf{x}$. To achieve faster convergence, instead we average the conditional expectations from (4.18) during Gibbs sampling, which is the Rao-Blackwellized estimator [118]. Algorithm 2 provides the pseudo-code of the sampling-based MMSE estimate approximation. Note that it is straightforward to parallelize sample drawing by running multiple samplers, which can further greatly improve the efficiency of inference.

4.2.3 Experiments

We applied our method to both natural images and fluorescence microscopy images. For the prior model, we use the 3×3 MRF with 8 filters learned from natural images in Sec. 3.3. We also studied retraining the parameters using fluorescence microscopy images. However, experimental results show that this does not improve the result. For the likelihood model, unless otherwise mentioned, we use mixtures of Gaussians (MoGs) with three components. Image noise parameters α, $\mu_{\mathbf{n}}$, $\sigma_{\mathbf{n}}$ and $\mathbf{H}$ are assumed to be given (for synthetic data) or are estimated from the images. No additional parameters need to be tuned. For a performance comparison, we choose popular denoising and deconvolution methods with authors' code available.

Denoising. In this case, $\mathbf{H}$ is set to an identity matrix. As can be seen from Tab. 4.3, using MoGs (3 components) for likelihood approximation yields a gain of the peak signal-to-noise ratio (PSNR) up to 0.45dB compared to a single Gaussian, which suggests that the accuracy of likelihood modeling is crucial for the performance. Our method outperforms the popular Poisson-Gaussian denoising method PURE-LET [121] and generates less artifacts in the restored images, especially when the noise strength is large. Fig. 4.11 shows example images of the denoising results.

Deconvolution. Here $\mathbf{H}$ is a convolution matrix generated from a blur kernel. Tab. 4.4 and Fig. 4.12 show that our method outperforms GILAM [131], which is a

Table 4.3: Denoising results (PSNR in dB) for test images with different noise parameters.

Poiss. param. α	1		2		5	
Gauss. param. σ_{n}	10	20	10	20	10	20
Method	*Cameraman* 256 × 256					
Noisy input	24.79	21.24	22.89	20.32	19.79	18.39
PURE-LET [121]	31.29	28.92	30.14	28.39	28.48	27.35
Ours ($M = 1$)	31.63	29.22	30.38	28.69	28.48	27.58
Ours ($M = 3$)	**31.66**	**29.26**	**30.41**	**28.74**	**28.64**	**27.68**
Method	*Crym* 256 × 256					
Noisy input	25.74	21.59	24.22	20.97	21.53	19.56
PURE-LET [121]	29.77	27.31	28.84	26.95	27.23	26.09
Ours ($M = 1$)	29.83	27.36	28.84	26.97	27.09	26.05
Ours ($M = 3$)	**29.87**	**27.39**	**28.90**	**27.01**	**27.27**	**26.14**
Method	*Moon* 512 × 512					
Noisy input	26.13	21.99	24.51	21.31	21.69	19.76
PURE-LET [121]	29.67	26.85	28.77	26.56	27.34	25.92
Ours ($M = 1$)	29.76	27.32	28.87	27.12	27.00	26.18
Ours ($M = 3$)	**29.85**	**27.40**	**28.98**	**27.26**	**27.36**	**26.45**
Method	*Fluocells* 512 × 512					
Noisy input	26.92	22.44	25.68	21.98	23.19	20.79
PURE-LET [121]	33.92	30.50	33.25	30.27	31.97	29.91
Ours ($M = 1$)	34.09	31.26	33.45	31.35	31.67	30.83
Ours ($M = 3$)	**34.18**	**31.36**	**33.56**	**31.44**	**32.12**	**31.00**

deconvolution method for Poissonian images. GILAM fails when the Gaussian noise part of the PG noise gets stronger. Compared to the method in [125] (using their settings), our method is slightly better (less outlier pixels in the restored image, see Fig. 4.13).

Super-resolution. When **H** is a matrix for performing Gaussian blurring and downsampling, our method can perform super-resolution and denoising simultaneously. Fig. 4.14 shows a comparison of our joint scheme and a sequential scheme of denoising (our method with **H** as identity matrix) followed by bicubic interpolation. It can be seen that our joint scheme yields sharper image edges.

Real microscopy images

In practical applications, when the noise parameters are not available, they can be directly estimated from the original images using different methods (e.g., [132, 133]). We performed joint super-resolution (factor of two), deconvolution, and denoising on real microscopy images of telomeres acquired with an easySTED system [57]. Examples are shown in Fig. 4.15. Noise parameters $\alpha \approx 3.2$ and $\sigma_{\mathrm{n}} \approx 4.9$ were estimated

Table 4.4: Deconvolution results (PSNR in dB) for Gaussian blur with variance 3 and different noise parameters.

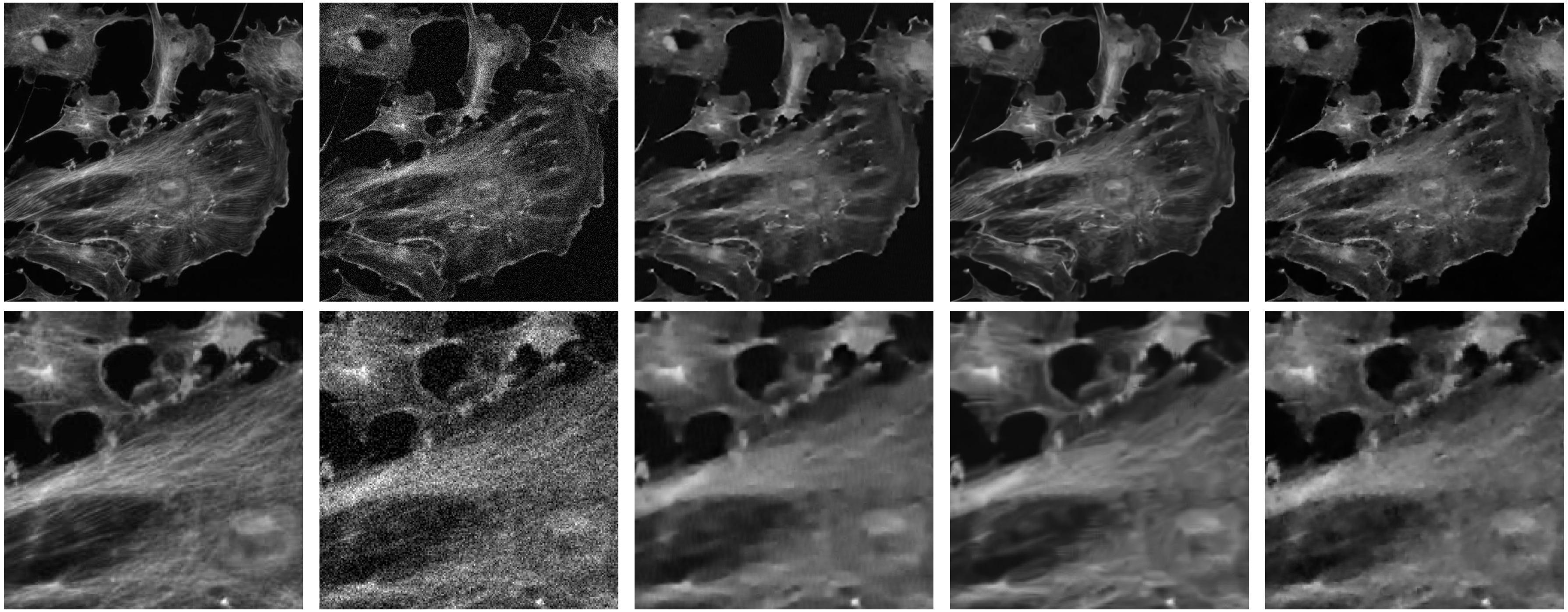
Pois. param. α	1		5	
Gaus. param. σ_{n}	0	10	0	10
Method	*Cameraman* 256×256			
Input	22.34	21.36	18.79	18.36
GILAM [131]	25.50	23.00	23.69	23.50
Our method	**25.70**	**25.08**	**24.35**	**24.12**
Method	*Crym* 256×256			
Input	23.86	22.49	20.66	19.98
GILAM [131]	25.87	22.38	23.74	23.51
Our method	**26.17**	**25.50**	**24.73**	**24.49**
Method	*Fluocells* 512×512			
Input	28.92	25.84	23.87	22.72
GILAM [131]	31.20	22.93	29.13	27.97
Our method	**32.84**	**31.55**	**30.51**	**30.40**

using [133]. It can be seen that the noise as well as the blur have been significantly reduced. The resolution of the resulting images after applying our joint method is a factor of two higher compared to the original images. We employed these images for segmentation and shape analysis of telomeres using the model fitting approach in [134]. The model in this approach is based on a Fourier representation, and by fitting the model to an image the Fourier coefficients are determined. We applied the model fitting approach both to the original images and the high resolution images. It turned out that for the high resolution images from our joint method, the Fourier coefficients reflecting the shape were determined more robustly and accurately compared to using the original images. This is important for distinguishing different experimental conditions.

Discussion

In the denoising experiment, we have seen that the accuracy of likelihood modeling is crucial for the model performance. Thus, the likelihood distribution must be accurately estimated in applications. For a likelihood that is defined (e.g., derived from some regularization of the data term), the optimal parameters can be obtained by training from the data.

Although we have developed an efficient Gibbs sampler and run multiple samplers in parallel for an acceleration, our method is still slower than the methods used in our comparison, which is a common issue of sampling-based methods. To further reduce the computation time, some variational Bayesian method, e.g., mean field approximation, can be employed for the inference in our method, with a slight loss



(a) Original

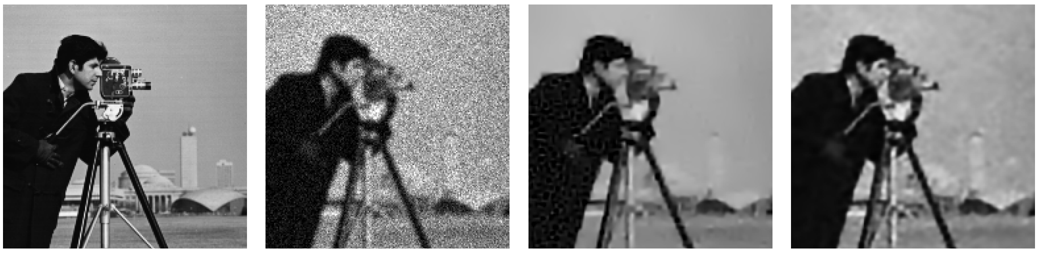
(b) With Poisson-Gaussian noise, PSNR = 20.79dB

(c) PURE-LET [121], PSNR = 29.91dB

(d) GAT-BM3D [122], PSNR = 30.67dB

(e) Our method, PSNR = 31.00dB

Figure 4.11: Restoration of *Fluocells* degraded by Poisson-Gaussian noise ($\alpha = 5, \mu_{\mathtt{n}} = 0, \sigma_{\mathtt{n}} = 20$). *(top)* Full image; *(bottom)* detail.



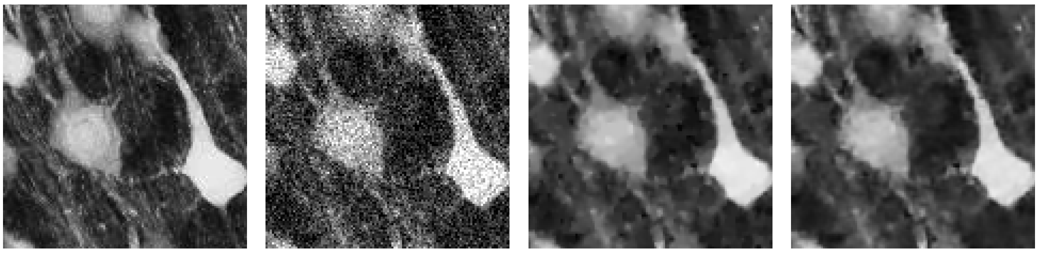
(a) Original

(b) Blurry and noisy, PSNR = 18.36dB

(c) GILAM [131], PSNR = 23.50dB

(d) Our method, PSNR = 24.12dB

Figure 4.12: Deconvolution of *Cameraman* (cropped) degraded by Gaussian blur ($\sigma^2 = 3$) and Poisson-Gaussian noise ($\alpha = 5, \mu_{\mathtt{n}} = 0, \sigma_{\mathtt{n}} = 10$).



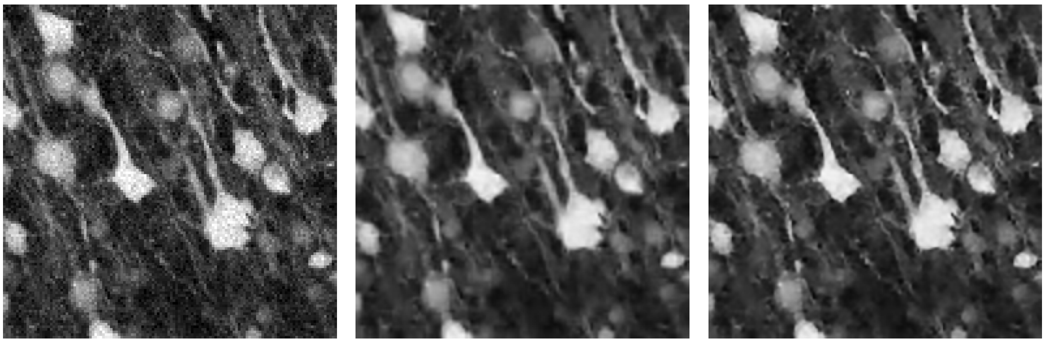
(a) Original

(b) Blurry and noisy, PSNR = 18.16dB

(c) Method in [125], PSNR = 24.59dB

(d) Our method, PSNR = 24.67dB

Figure 4.13: Deconvolution of *Crym* (cropped) degraded by Gaussian blur (σ^2=0.25) and Poisson-Gaussian noise (α=4.25, $\mu_{\mathtt{n}}$=0, $\sigma_{\mathtt{n}}$=25.5).



(a) Downsampled with noise, 128×128 pixels

(b) Restored, 256×256 pixels, PSNR = 24.35dB

(c) Restored, 256×256 pixels, PSNR = 24.63dB

Figure 4.14: Super-resolution (2x) and denoising of *Crym* using our method. Poisson-Gaussian noise ($\alpha = 1, \mu_{\mathtt{n}} = 0, \sigma_{\mathtt{n}} = 10$) is applied to (a) the downsampled image. (b) Denoising followed by bicubic interpolation; (c) Joint super-resolution and denoising. PSNR is computed based on the original image.

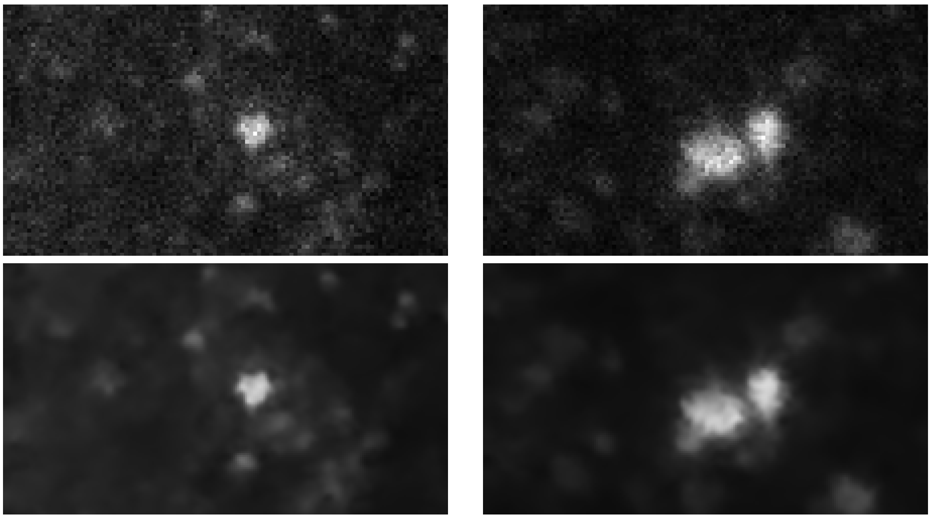


Figure 4.15: Super-resolution (2x), deconvolution, and denoising of STED microscopy images using our method. *(Top)* original images; *(Bottom)* restored images. Noise parameters are estimated using [133]: $\alpha \approx 3.2$ and $\sigma_{\mathrm{n}} \approx 4.9$.

of the accuracy. Examples of mean field for high-order MRFs can be found in [48].

In our experiments, the model with a MRF prior trained on natural images performed better than that trained on fluorescence microscopy images. Due to the significantly higher noise in microscopy images compared to natural images, a prior model might not be well trained. On the other hand, a reason is probably that microscopy images have similar local image structures as natural images, or natural images exhibit scale invariant derivative statistics [108], which also represent microscopy images.

4.3 Non-rigid registration of live cell nuclei

4.3.1 Introduction and related work

Time-lapse live cell microscopy enables studying the dynamics of subcellular structures to understand cellular processes. As the observed motion of intranuclear particles (proteins) is superimposed on the motion of cell nuclei, to analyze the relative motion patterns of particles, the particle motion must be decoupled from the deformation and movement of cell nuclei. This can be achieved by registration of cell nuclei in microscopy images at different time points. Usually the intranuclear particles and cell nuclei are acquired in two different channels (see Fig. 4.16). The deformation fields are first estimated according to the cell nucleus channel, and then applied to the particle channel to accomplish the decoupling of the motion.

Early methods for registration of cell or cell nuclei images mostly rely on rigid

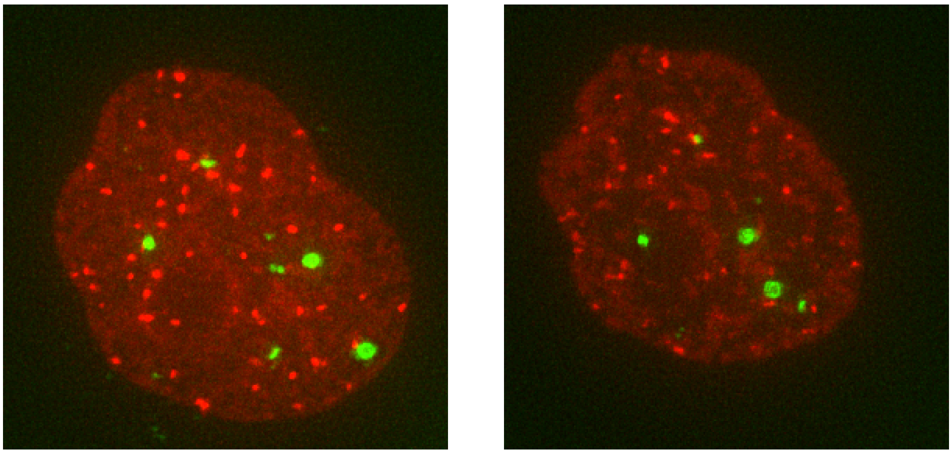


Figure 4.16: Two images at different time points from a time-lapse multi-channel live cell image sequence: (*left*) $t = 0$; (*right*) $t = 50$. The red channel shows the cell nucleus and the green channel shows the intranuclear particles.

and affine transformations (e.g., [135, 136, 137, 138, 139, 140]) and work well in applications with relatively simple conditions. As the patterns of cell deformations are usually complex and cannot be well described by linear transformations, non-rigid registration approaches have been proposed. Mattes *et al.* [141] described a method based on landmarks and thin-plate splines. Yang *et al.* [142] used segmented images and a variant of the demons algorithm with a multiresolution scheme to register static and live cell nuclei images. De Vylder *et al.* [143] proposed a registration method based on the deformation of the nuclear contour. More recently, Sorokin *et al.* [144] used contour matching and a physical prior to model the deformation based on elasticity theory. However, the approaches in [142, 143, 144] do not directly exploit the image intensity. Also, since the deformation fields are determined using interpolation, complex local deformations in the inner part of a cell nucleus are difficult to capture.

Another type of non-rigid cell image registration methods directly exploits the image intensity and employs *local* optical flow models to compute the deformation. These methods are based on the observation that the deformation between two consecutive image frames changes gradually over time and shows temporal coherence. Kim *et al.* [145] extended the Lucas-Kanade optical flow model [146] to estimate dense deformation fields between consecutive frames and then register each frame to the reference frame using an incremental scheme. This method was extended by Tektonidis *et al.* [147], where a multi-frame approach was proposed to improve the robustness to image noise. Tektonidis and Rohr [148] introduced a diffeomorphic multi-frame approach, which guarantees invertibility and continuity of the computed transformations. The major advantage of optical flow-based registration methods is that they can cope with general motion which is not necessarily constrained by a physical prior. All above mentioned approaches for non-rigid registration of cell nuclei using optical flow models [145, 147, 148] are *local*, since the deformation vectors are computed based on local image patches. Besides low efficiency, using local patches implies that only some basic regularization on the result is included.

In contrast, *global* optical flow models explicitly integrate the optical flow constraint (e.g., brightness constancy) and the regularization of optical flow vectors into one single objective function. The optical flow vectors for all pixels, i.e., a dense optical flow field, are computed simultaneously and efficiently based on the whole images. Interpolation as often used in local models is not required. By appropriately choosing the type of regularization, global models can generally improve the performance compared to local models. For a survey on optical flow models, readers are referred to [149]. However, popular global optical flow models have been originally proposed for video images of natural scenes (e.g., street traffic). For cell microscopy images, the image structures are generally more blurry (e.g., out-of-focus local regions, scattering of fluorescence) and the image noise is more significant. Another characteristic is that occlusions (due to 3D-2D projection) generally do not occur in contrast to video images of natural scenes. Due to different properties, direct application of these global optical flow models for estimating the deformation fields in fluorescence microscopy images generally does not yield optimal results.

In addition, previous optical flow-based registration methods for cell microscopy images [140, 145, 147, 148] assume brightness constancy. This means that the image intensities of moving structures do not change over time, which is a relatively strong assumption and often violated in real data. In the field of optical flow estimation for images of natural scenes, extensions of brightness constancy to gradient constancy and high-order feature constancy were proposed to improve, for example, the robustness to varying illumination (e.g., [150]). While it is often hard to handcraft the most suitable features, they could be determined by some feature learning method. In [150], estimation of optical flow for images of natural scenes relied on discriminatively learned features through maximizing the conditional likelihood of the data fidelity term. However, for cell microscopy image data, due to the lack of ground truth of deformations, an unsupervised feature learning method should be employed. Note that, since the feature-constancy data term in an optical flow model implies (conditional) independence among features, an ideal feature leaning model should be "coupled" with the registration model (data term), i.e., their hidden assumptions should be consistent. Unfortunately, common unsupervised feature learning methods (e.g., PCA, autoencoder, sparse coding [85]) and patch-based image models with features learned simultaneously (e.g., product of Student-t model [2]) do not fulfill this requirement.

For registration of temporal medical images (e.g., CT, MR and US images of organs), group-wise approaches have been proposed (e.g., [151, 152, 153, 154, 155]), which simultaneously exploit all frames of an image sequence. Typically periodic movement is considered (e.g., lung, heart) and often parametric registration approaches are employed (e.g., using B-splines). For registration of live cell microscopy image sequences, it was shown that optical flow-based methods with an incremental scheme [147, 148] perform better than a group-wise method [151]. A main reason is that an incremental scheme better copes with the heterogeneous changes of the intensity structure over time, since the deformations are computed by concatenating deformation fields estimated from a limited number of consecutive frames.

In this section, we introduce *global* optical flow models with appropriate regularization of the deformation fields for non-rigid registration of cell nuclei in live-cell fluorescence microscopy images using an incremental scheme. We compare different regularizers and show that a convex quadratic function is more suitable than non-convex functions widely used for optical flow estimation in video images of natural scenes. To increase the robustness to noise, we propose an adaptive weighting scheme derived from the mixed Poisson-Gaussian statistics of typical noise in microscopy images. The deformation fields can be efficiently computed by solving linear systems of equations.

Moreover, we extend the global models by high-order filter banks obtained through learning the Field of Experts (FoE) and the convolutional Gaussian RBM priors (see Sec. 3.1 and Sec. 3.5). We show that these two models are consistent with the hidden assumption in the feature-constancy data term of the registration model. An advantage is that these two models only require moderate amount of training data due to the convolutional model structure. To cope with the problem that learned coefficients in the filters might shift and distort image correspondences in the feature space (note that this is not an issue in image deconvolution and denoising applications), we augment the learning procedure (Sec. 3.3 and Sec. 3.5) by an additional constraint. In addition, our adaptive weighting scheme based on brightness constancy can also be extended to the global model using high-order features to increase the robustness to noise. To deal with outliers and complex noise not following Poission-Gaussian statistics, we also employ a combined local-global (CLG) scheme [66].

To evaluate the performance of our proposed global methods, we use multiple data sets including both 2D and 3D real live cell microscopy image sequences as well as synthetic image data. Experimental results demonstrate that our global method significantly improves the registration accuracy compared to existing state-of-the-art non-rigid registration methods such as [148] and [144]. It also turns out that our proposed method based on global optical flow models is much more efficient than existing methods based on local models. Also, we present results for 3D live-cell microscopy images and for synthetic image data. The work was published in [59, 60].

4.3.2 Global optical flow-based method

Our non-rigid registration method for live cell microscopy images is based on a global optical flow model for estimating the deformations between consecutive image frames and uses an incremental scheme to register a temporal image sequence.

Incremental temporal registration

Given temporal cell microscopy image sequences, the goal is to determine the dense deformation field $\mathbf{w}^{0,t}$ to register the frame $\mathbf{x}^t$ at time point t to the reference frame $\mathbf{x}^0$ at the first time point. However, directly estimating all $\mathbf{w}^{0,t}, t=1,...,T$, is not a good choice, since the deformations over a long time interval are generally too complex. Instead, we estimate all deformation fields $\mathbf{w}^{t-1,t}$ between pairs of consecutive image frames and then concatenate them incrementally to obtain

$$\mathbf{w}^{0,t} = \mathbf{w}^{t-1,t} \circ \mathbf{w}^{0,t-1} = \mathbf{w}^{t-1,t} + \tilde{\mathbf{w}}^{0,t-1}, \tag{4.19}$$

where $\tilde{\mathbf{w}}^{0,t-1}$ is the result of using $\mathbf{w}^{t-1,t}$ to warp $\mathbf{w}^{0,t-1}$ using bilinear interpolation, and $\mathbf{w}^{0,0} \triangleq \mathbf{0}$. Below, we describe methods based on a global optical flow model for computing a deformation field $\mathbf{w}^{t-1,t}$ from two consecutive frames. The superscript of $\mathbf{w}$ will be dropped for easier readability.

Global optical flow-based models

The deformation field $\mathbf{w}$ between two consecutive images $\mathbf{x}^0 \in \mathbb{R}^N$ and $\mathbf{x}^1 \in \mathbb{R}^N$ (N is the number of pixels in an image frame) is modeled via the posterior density in a Bayesian framework

$$p(\mathbf{w}|\mathbf{x}^0, \mathbf{x}^1) \propto p_D(\mathbf{x}^0|\mathbf{w}, \mathbf{x}^1) \cdot p_R(\mathbf{w}), \tag{4.20}$$

where the data-dependent likelihood $p_D(\mathbf{x}^0|\mathbf{w}, \mathbf{x}^1)$ models how $\mathbf{x}^0$ is generated from $\mathbf{w}$ and $\mathbf{x}^1$, and the prior $p_R(\mathbf{w})$ imposes regularization on $\mathbf{w}$. For 2D image sequence, $\mathbf{w} \in \mathbb{R}^{2N}$ consists of x- and y-components $\mathbf{u}$ and $\mathbf{v}$.

Data term. When assuming brightness constancy, the data fidelity (likelihood) term can be written as

$$p_D(\mathbf{x}^0|\mathbf{w}, \mathbf{x}^1) \propto \prod_{i=1}^{N} \exp\left\{ -\tfrac{1}{\alpha_i} \cdot \psi_D(\mathbf{x}_i^0 - \mathbf{x}_{\mathbf{w}i}^1) \right\}, \tag{4.21}$$

where α_i is the regularization weight and $\mathbf{x}_{\mathbf{w}}^1$ is the warped (deformed) image of $\mathbf{x}^1$ towards $\mathbf{x}^0$ by the deformation field $\mathbf{w}$ to be estimated. The penalty function ψ_D can take an arbitrary form. When a *convex* quadratic function $\psi_D(x) = x^2$ is used, (4.21) will be a Gaussian function, meaning that the difference $\mathbf{x}^0_i - \mathbf{x}^1_{\mathbf{w}i}$ is assumed to follow a Gaussian distribution. In popular optical flow methods for video images of natural scenes (e.g., [150]), *non-convex* functions for ψ_D are used (e.g., a Lorentzian function). The reason is that, due to, for example, occlusion in video images of natural scenes, the brightness constancy assumption is violated, leading to heavy-tailed histograms with a sharp peak of the intensity differences. In contrast, for cell microscopy images, there exist less violations of the brightness constancy assumption (e.g., occlusions due to 3D-2D projection as in video images generally do not occur), but significant noise. Note that for images with strong noise (e.g., microscopy images), the histogram of intensity differences has a more rounded peak [150]. Therefore, the histogram of intensity differences for cell microscopy images is expected to have a shape that is largely similar to a Gaussian function. Thus, a convex quadratic function for ψ_D should be well suited, and we use it in our method. In addition, a quadratic data term eases the inference procedure.

Regularization term. For the regularization term (prior), we use a Markov random field (MRF) formulation

$$p_R(\mathbf{w}) \propto \prod_{r=1}^{2} \prod_{i=1}^{N} \exp\{-\psi_R((\mathbf{d}_r * \mathbf{u})_i) - \psi_R((\mathbf{d}_r * \mathbf{v})_i)\}, \tag{4.22}$$

where $\mathbf{d}_r$ are first-order derivative filters in x- and y-directions. The choice of the regularization function ψ_R in (4.22) should depend on the statistical properties of $\mathbf{w}$. Unfortunately, for cell microscopy images, due to the lack of the ground truth, the statistical properties of $\mathbf{w}$ are not known. Tseng *et al.* [156] found that the cell nucleus is stiffer and more elastic than the cytoplasm. We accordingly assume that the deformation vectors in the nuclei change gradually in space and that strong discontinuities (e.g., edges) in $\mathbf{w}$ are unlikely, which means that the filter responses will hardly have large values and their histograms will not show a heavy-tailed shape. Thus, it is expected that a convex quadratic regularization $\psi_R(x) = x^2$, which has no edge-preserving property, is more appropriate. As a comparison, we also consider two robust regularizers widely used in popular optical flow models for video images of natural scenes: the Charbonnier function $\psi_R(x) = \sqrt{x^2 + \epsilon^2}$ [66], which is a differentiable variant of the $L1$ norm, and the non-convex Lorentzian function $\psi_R(x) = \log(1 + \frac{x^2}{2\sigma^2})$ [157]. The latter arises in the context of a Student t-distribution for one degree of freedom, and is also a common potential function or regularizer in image modeling and analysis (e.g., [30, 7]).

Inference. The deformation field $\mathbf{w}$ is computed by minimizing the model energy $E(\mathbf{w}) = -\log p(\mathbf{w}|\mathbf{x}^0, \mathbf{x}^1)$, which is equivalent to the maximum a-posteriori (MAP) estimate. For small deformations, the data term can be linearized using a first order Taylor expansion. When both ψ_D and ψ_R are quadratic functions and a single regularization weight $\alpha_i = \alpha_0$ is applied, the model energy is

$$E(\mathbf{w}) = \left\|\mathbf{\Gamma u} + \mathbf{\Lambda v} + \mathbf{x}^1 - \mathbf{x}^0\right\|^2 + \alpha_0 \cdot \sum_r \sum_i \left((\mathbf{d}_{ir}^{\mathrm{T}}\mathbf{u})^2 + (\mathbf{d}_{ir}^{\mathrm{T}}\mathbf{v})^2\right) + c, \tag{4.23}$$

where $\mathbf{\Gamma} = \mathrm{diag}(\mathbf{x}_x^1)$ and $\mathbf{\Lambda} = \mathrm{diag}(\mathbf{x}_y^1)$ are diagonal matrices containing the partial image derivatives $\mathbf{x}_x^1$ and $\mathbf{x}_y^1$ in x- and y-directions, $\mathbf{d}_{ir}$ are vectors defined as $\mathbf{d}_{ir}^{\mathrm{T}}\mathbf{u} = (\mathbf{d}_r * \mathbf{u})_i$, and c is a constant. To minimize (4.23), we need to solve the linear system of equations

$$\left(\begin{bmatrix} \mathbf{\Gamma\Gamma} & \mathbf{\Gamma\Lambda} \\ \mathbf{\Gamma\Lambda} & \mathbf{\Lambda\Lambda} \end{bmatrix} + \alpha_0 \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}\right) \cdot \mathbf{w} = \begin{bmatrix} \mathbf{\Gamma} \\ \mathbf{\Lambda} \end{bmatrix} \cdot (\mathbf{x}^0 - \mathbf{x}^1), \tag{4.24}$$

where

$$\mathbf{D} = \sum_r \sum_i \mathbf{d}_{ir}\mathbf{d}_{ir}^{\mathrm{T}}.$$

As the coefficient matrix on the left side (comprising $\mathbf{\Gamma}, \mathbf{\Lambda}$, and $\mathbf{D}$) is sparse and positive definite, (4.24) can be efficiently solved using Cholesky factorization. For the Lorentzian and Charbonnier functions, we follow [158] and use a graduated non-convexity (GNC) scheme.

Handling noise and outliers

To cope with significant noise and outliers in cell microscopy images, we use an adaptive weighting scheme and a combined local-global scheme as well as a combination of these two schemes.

Adaptive weighting (AW) scheme. The noise in cell microscopy images largely follows mixed Poisson-Gaussian statistics. An observed image can be written as $\mathbf{x}_i = g_0\mathbf{t}_i + \mathbf{n}_i$, where g_0 is the gain of the overall electronic system, $\mathbf{t}_i \in \mathbb{Z}_0^+$ is the number of collected photo-electrons at pixel i which is a random variable following a Poisson distribution $\mathbf{t}_i \sim \mathrm{Poi}(\boldsymbol{\lambda}_i)$, $\boldsymbol{\lambda}_i \in \mathbb{R}^+$, and $\mathbf{n}_i \in \mathbb{R}$ is Gaussian noise $\mathbf{n}_i \sim \mathcal{N}(\mu_n, \sigma_n^2)$ (e.g., [159]). The corresponding noise-free image is

$$\hat{\mathbf{x}}_i = g_0\boldsymbol{\lambda}_i + \mu_n. \tag{4.25}$$

Assuming brightness constancy for the (hidden) noise-free images $\hat{\mathbf{x}}_i^0 = \hat{\mathbf{x}}_{\mathbf{w}i}^1$, we have $\boldsymbol{\lambda}_i^0 = \boldsymbol{\lambda}_{\mathbf{w}i}^1$, thus $\mathbf{t}_{\mathbf{w}i}^1 \sim \mathrm{Poi}(\boldsymbol{\lambda}_{\mathbf{w}i}^1) = \mathrm{Poi}(\boldsymbol{\lambda}_i^0)$ following the same Poisson distribution as $\mathbf{t}_i^0 \sim \mathrm{Poi}(\boldsymbol{\lambda}_i^0)$, and therefore the distribution of their difference can be approximated by a Gaussian [160]

$$(\mathbf{t}_i^0 - \mathbf{t}_{\mathbf{w}i}^1) \sim \mathcal{N}(0, 2\boldsymbol{\lambda}_i^0). \tag{4.26}$$

With $(\mathbf{n}_i^{\mathrm{o}} - \mathbf{n}_{\mathbf{w}i}^{\mathrm{1}}) \sim \mathcal{N}(0, 2\sigma_n^2)$, the intensity difference of the observed images in the *data term* (4.21)

$$\mathbf{x}_i^{\mathrm{o}} - \mathbf{x}_{\mathbf{w}i}^{\mathrm{1}} = g_0(\mathbf{t}_i^{\mathrm{o}} - \mathbf{t}_{\mathbf{w}i}^{\mathrm{1}}) + (\mathbf{n}_i^{\mathrm{o}} - \mathbf{n}_{\mathbf{w}i}^{\mathrm{1}}) \tag{4.27}$$

has a Gaussian distribution with varying variance

$$\begin{aligned}(\mathbf{x}_i^{\mathrm{o}} - \mathbf{x}_{\mathbf{w}i}^{\mathrm{1}}) &\sim \mathcal{N}\big(0, 2g_0^2\boldsymbol{\lambda}_i^{\mathrm{o}} + 2\sigma_n^2\big) \\ &\propto \exp\Big\{-\frac{1}{4(g_0^2\boldsymbol{\lambda}_i^{\mathrm{o}}+\sigma_n^2)}\cdot(\mathbf{x}_i^{\mathrm{o}} - \mathbf{x}_{\mathbf{w}i}^{\mathrm{1}})^2\Big\}.\end{aligned} \tag{4.28}$$

Comparing (4.28) and (4.21) suggests using a quadratic penalty in the data term as well as a varying regularization weight at each pixel location

$$\alpha_i = \alpha_0(g_0^2\boldsymbol{\lambda}_i^{\mathrm{o}} + \sigma_n^2). \tag{4.29}$$

In practice, however, estimating the parameters in (4.29) is difficult and an approximation is needed.

We now consider the *regularization term.* In our fluorescence microscopy images, cell nuclei have more or less homogeneous intensity and include bright spot-like structures. The spots generally preserve their sizes and do not expand or shrink. Thus, it is reasonable to assume that the deformation vectors of the pixels inside one spot are very similar. Such deformation continuity can be achieved by using a higher weight for the regularization term to more strongly penalize the variation of the deformation. Since pixels inside the spots have high intensities, and higher weights are needed at these locations, we suggest using weights that are proportional to the image intensities:

$$\alpha_i = \alpha_0\hat{\mathbf{x}}_i^{\mathrm{o}} = \alpha_0(g_0\boldsymbol{\lambda}_i^{\mathrm{o}} + \mu_n^{\mathrm{o}}). \tag{4.30}$$

Thus, the weight α_i should be determined so that it is consistent with the weighting schemes for the data term in (4.29) and the regularization term in (4.30). We notice that the relations in (4.29) and (4.30) have the same structure: In both cases the weights are linear functions of $\boldsymbol{\lambda}_i$ and have the same sign of derivatives (simultaneously increasing or decreasing). Therefore, we propose determining the weights using the relation in (4.30), where the noise-free image is roughly approximated by applying Gaussian smoothing on the original image $\mathbf{x}^{\mathrm{o}}$

$$\alpha_i = \alpha_0(\hat{\mathbf{x}}_i^{\mathrm{o}} + b) \approx \alpha_0\big((\mathbf{g}_\sigma * \mathbf{x}^{\mathrm{o}})_i + b\big), \tag{4.31}$$

where b is a positive constant and $\mathbf{g}_\sigma$ is a Gaussian filter with kernel width σ.

For the proposed adaptive weighting scheme, the model energy is still convex.

The deformation field is computed by solving

$$\left(\begin{bmatrix}\boldsymbol{\Gamma\Gamma} & \boldsymbol{\Gamma\Lambda}\\ \boldsymbol{\Gamma\Lambda} & \boldsymbol{\Lambda\Lambda}\end{bmatrix} + \begin{bmatrix}\mathbf{D}_\alpha & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_\alpha\end{bmatrix}\right)\cdot\mathbf{w} = \begin{bmatrix}\boldsymbol{\Gamma}\\ \boldsymbol{\Lambda}\end{bmatrix}\cdot(\mathbf{x}^0 - \mathbf{x}^1), \tag{4.32}$$

where $\mathbf{D}_\alpha = \sum_{i,r}\alpha_i\mathbf{d}_{ir}\mathbf{d}_{ir}^{\mathrm{T}}$.

Combined local-global (CLG) scheme. The statistics of noise and outliers in cell microscopy images are sometimes more complex and cannot be accurately modeled using mixed Poisson-Gaussian as in the case of the adaptive weighting scheme. An alternative to improve the robustness to noise and outliers is the combined local-global (CLG) method [66] which exploits local image information within a global model.

The deformation field $\mathbf{w}$ is then determined by solving

$$\begin{aligned}&\left(\begin{bmatrix}\mathrm{diag}\{\mathbf{g}_\rho*(\mathbf{x}_x^1\odot\mathbf{x}_x^1)\} & \mathrm{diag}\{\mathbf{g}_\rho*(\mathbf{x}_x^1\odot\mathbf{x}_y^1)\}\\ \mathrm{diag}\{\mathbf{g}_\rho*(\mathbf{x}_x^1\odot\mathbf{x}_y^1)\} & \mathrm{diag}\{\mathbf{g}_\rho*(\mathbf{x}_y^1\odot\mathbf{x}_y^1)\}\end{bmatrix} + \alpha_0\begin{bmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{bmatrix}\right)\mathbf{w}\\ &= \begin{bmatrix}\mathrm{vec}\left\{\mathbf{g}_\rho*\left(\mathbf{x}_x^1\odot(\mathbf{x}^0-\mathbf{x}^1)\right)\right\}\\ \mathrm{vec}\left\{\mathbf{g}_\rho*\left(\mathbf{x}_y^1\odot(\mathbf{x}^0-\mathbf{x}^1)\right)\right\}\end{bmatrix},\end{aligned} \tag{4.33}$$

where "$\odot$" denotes element-wise product and $\mathbf{g}_\rho$ is a Gaussian kernel with standard deviation ρ. Note that, as the sparsity of the coefficient matrix in (4.33) is the same as in (4.24), the solution can still be computed efficiently.

Combined AW+CLG scheme. The AW and CLG schemes described above can also be combined. To this end we extended the linear system of equations of the CLG scheme in (4.33) by using the weight matrix $\mathbf{D}_\alpha$ from the AW scheme in (4.32) instead of the matrix $\mathbf{D}$.

Extension to 3D image data

For temporal 3D image sequences (volume images over time), the deformation field $\mathbf{w}\in\mathbb{R}^{3N}$ includes an additional z-component $\mathbf{h}\in\mathbb{R}^N$. While the formulation of the data term (4.21) is the same as in the 2D case, the regularization term needs to be extended to regularize all components of $\mathbf{w}$

$$p_R(\mathbf{w}) \propto \prod_{r=1}^{3}\prod_{i=1}^{N}\exp\left\{-\psi_R\big((\mathbf{d}_r*\mathbf{u})_i\big) - \psi_R\big((\mathbf{d}_r*\mathbf{v})_i\big) - \psi_R\big((\mathbf{d}_r*\mathbf{h})_i\big)\right\}, \tag{4.34}$$

where $\mathbf{d}_3$ is a first-order derivative filter in z-direction. The linear system of equations for computing the 3D deformation field $\mathbf{w}$ for a global model (*cf.* (4.23)) is

$$\left(\begin{bmatrix} \mathbf{\Gamma\Gamma} & \mathbf{\Gamma\Lambda} & \mathbf{\Gamma\Delta} \\ \mathbf{\Gamma\Lambda} & \mathbf{\Lambda\Lambda} & \mathbf{\Lambda\Delta} \\ \mathbf{\Gamma\Delta} & \mathbf{\Lambda\Delta} & \mathbf{\Delta\Delta} \end{bmatrix} + \alpha_0 \begin{bmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D} \end{bmatrix}\right) \mathbf{w} = \begin{bmatrix} \mathbf{\Gamma} \\ \mathbf{\Lambda} \\ \mathbf{\Delta} \end{bmatrix} (\mathbf{x}^0 - \mathbf{x}^1), \tag{4.35}$$

where $\mathbf{\Delta} = \text{diag}(\mathbf{x}_z^1)$ with $\mathbf{x}_z$ denoting the partial derivatives in z-direction. $\mathbf{\Gamma}$, $\mathbf{\Lambda}$, and $\mathbf{D}$ are defined analogously as in the 2D case described above. Note that, as the dimension of the coefficient matrix in (4.35) is much higher than that for the 2D model, using Cholesky factorization to compute the exact solution is not efficient. Thus, for the 3D model we instead use the preconditioned conjugate gradient (PCG) method [161] to obtain an approximate solution. To handle noise and outliers, the adaptive weighting or the CLG scheme can be straightforwardly applied.

4.3.3 Registration in the learned feature space

For estimating the optical flow or deformation field between two images $\mathbf{x}^0$ and $\mathbf{x}^1$, often brightness constancy is assumed and the brightness residual after image warping $\boldsymbol{\delta}_i = \mathbf{x}_i^0 - \mathbf{x}_{\mathbf{w}i}^1$ is penalized in the data energy term $E_D(\mathbf{w}) = \sum_i \xi(\boldsymbol{\delta}_i)$, where ξ is some penalty function. To improve the performance under varying illumination, gradient constancy and high-order feature constancy were proposed for images of natural scenes (e.g., [150]). Suppose $\mathbf{f}_j$ are filters for extracting image derivatives or high-order features, the residual is then $\boldsymbol{\delta}_{ji} = (\mathbf{f}_j * \mathbf{x}^0)_i - (\mathbf{f}_j * \mathbf{x}^1)_{\mathbf{w}i}$. When several features are jointly considered, all energy terms of the features are summed up yielding the data term $E_D(\mathbf{w}) = \sum_j \sum_i \xi(\boldsymbol{\delta}_{ji})$, whose corresponding probabilistic formulation (the likelihood) is

$$\begin{aligned} p_D(\mathbf{w}) &\propto \exp\{-E_D(\mathbf{w})\} \\ &\propto \prod_{j=1}^{J} \prod_{i=1}^{N} \exp\{-\xi(\boldsymbol{\delta}_{ji})\} \\ &\propto \prod_{j=1}^{J} \prod_{i=1}^{N} p\big((\mathbf{f}_j * \mathbf{x}^0)_i \big| \mathbf{w}, (\mathbf{f}_j * \mathbf{x}^1)\big). \end{aligned} \tag{4.36}$$

This factorization of the likelihood into a product of conditional distributions implies that the filter responses $(\mathbf{f}_j * \mathbf{x}^0)_i$ are assumed to be (conditionally) independent. However, in practice the requirement of independence cannot be fulfilled, as there are more filter responses than pixels. Note that orthogonality of filters does not lead to independence of filter responses (e.g., image derivatives in x- and y-directions are generally correlated).

Learning features using generative image models

According to (4.36), high-order features with least dependence are best. In addition, the features should be specific for the considered images. To determine filters that capture the most important and least dependent features, we train filter-based generative image models implying the independence assumption. We consider two generative image models: Fields of Experts (FoE) and convolutional Gaussian restricted Boltzmann machine (cGRBM). These two kinds of models not only have been widely used in computer vision, but also have the advantage that they only require moderate amount of training data due to the convolutional model structure.

Fields of Experts. The FoE [50, 51] (see also Sec. 2.2.2) is a filter-based, high-order Markov random field (MRF) image prior, where both potential functions and filters are trained unsupervisedly . The probability density of an image $\mathbf{x}$ is defined as

$$p_{\text{FoE}}(\mathbf{x}) \propto \prod_{j=1}^{J} \prod_{i=1}^{N} \phi\big((\mathbf{f}_j * \mathbf{x})_i; \boldsymbol{\omega}_j\big), \tag{4.37}$$

where $\mathbf{f}_j$ are linear filters, and $\phi(\cdot; \boldsymbol{\omega}_j)$ are filter-specific potential functions (or experts) with parameters $\boldsymbol{\omega}_j$.

Comparing (4.37) with (4.36), it can be seen that the FoE prior is a generalization of the data term (4.36). Proper training procedures will not only yield filters that capture specific image features of the training data, but also enforce the filter responses to be as independent as possible, making the learned filters well suitable for the model in (4.36).

Convolutional Gaussian RBM. The cGRBM (see also Sec. 2.3.1) is a variant of a convolutional RBM [35, 162], for which a convolutional weight sharing scheme (as for convolutional neural networks) is used for a Gaussian RBM [110]. The shared weights can be regarded as filters. In a cGRBM, an image is modeled by defining an energy function of real-valued visible units $\mathbf{x}$ (here, the image pixels) and binary hidden units $\mathbf{h}$:

$$E_{\text{cGRBM}}(\mathbf{x}, \mathbf{h}) = \frac{1}{2} \sum_{i=1}^{N} \mathbf{x}_i^2 - \sum_{j=1}^{J} \sum_{i=1}^{N} h_{ji}\big((\mathbf{f}_j * \mathbf{x})_i + b_j\big), \tag{4.38}$$

where $\mathbf{f}_j$ are the weights or filters, and b_j denote the biases.

The probability density of $\mathbf{x}$ can be obtained by marginalizing out the hidden units

$$
\begin{aligned}
p_{\text{cGRBM}}(\mathbf{x}) &\propto \mathcal{N}(\mathbf{x};\mathbf{0},\mathbf{I})\cdot\prod_{j=1}^{J}\prod_{i=1}^{N}\left(1+\exp\{(\mathbf{f}_j*\mathbf{x})_i+b_j\}\right) \\
&\propto \mathcal{N}(\mathbf{x};\mathbf{0},\mathbf{I})\cdot\prod_{j=1}^{J}\prod_{i=1}^{N}\varphi\big((\mathbf{f}_j*\mathbf{x})_i;b_j\big),
\end{aligned}
\tag{4.39}
$$

where φ are filter specific potential functions. From (4.39), it can be seen that the cGRBM is also consistent with the independence property of model in (4.36).

Learning. Although we are only interested in the coefficients of the filters $\mathbf{f}_j$, all model parameters (including $\boldsymbol{\omega}_j$ for the FoE and b_j for the cGRBM) must be learned simultaneously. As the partition functions of both models are intractable, learning is non-trivial. In our approach we use sampling-based approximate maximum likelihood learning and follow the best practices suggested in Sec. 3.3 and Sec. 3.5. We noticed that these learning procedures do not constrain the learned coefficients from shifting within the filters, which is not an issue in other applications (e.g., image restoration [163, 90, 89]). However, shifting of the filter coefficients generally leads to shifted extracted image features (see Fig. 4.17), which may distort image correspondences. To address this, we include an additional constraint in the learning procedure. In our approach, after each iteration of the filter update, the filter coefficients are shifted so that their gravity center is always the center of the filter, and thus the extracted features are not shifted.

Non-rigid registration using learned features

Based on the learned filters $\mathbf{f}_j$ using FoE or cGRBM, we define the data term assuming feature constancy

$$
p_D(\mathbf{x}^0|\mathbf{w},\mathbf{x}^1)\propto\prod_{j=1}^{J}\prod_{i=1}^{N}\exp\Big\{-\frac{1}{\alpha_{ji}}\big((\tilde{\mathbf{x}}_j^0)_i-(\tilde{\mathbf{x}}_{j\mathbf{w}}^1)_i\big)^2\Big\},
\tag{4.40}
$$

where $\tilde{\mathbf{x}}_j \triangleq \mathbf{f}_j*\mathbf{x}$ and $\tilde{\mathbf{x}}_{j\mathbf{w}}$ denotes the warped (high-order feature) image by $\mathbf{w}$. We use a quadratic penalty function according to the analysis in Sec. 4.3.2 above. If we use a single regularization weight $\alpha_{ji}=\alpha_0$ in (4.40), $\mathbf{w}$ can be computed by solving the linear system of equations

$$
\left(\alpha_0\begin{bmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{bmatrix}+\sum_{j=1}^{J}\begin{bmatrix}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j & \boldsymbol{\Gamma}_j\boldsymbol{\Lambda}_j\\ \boldsymbol{\Gamma}_j\boldsymbol{\Lambda}_j & \boldsymbol{\Lambda}_j\boldsymbol{\Lambda}_j\end{bmatrix}\right)\cdot\mathbf{w}=\sum_{j=1}^{J}\begin{bmatrix}\boldsymbol{\Gamma}_j\\ \boldsymbol{\Lambda}_j\end{bmatrix}\cdot\left(\tilde{\mathbf{x}}_j^0-\tilde{\mathbf{x}}_j^1\right),
\tag{4.41}
$$

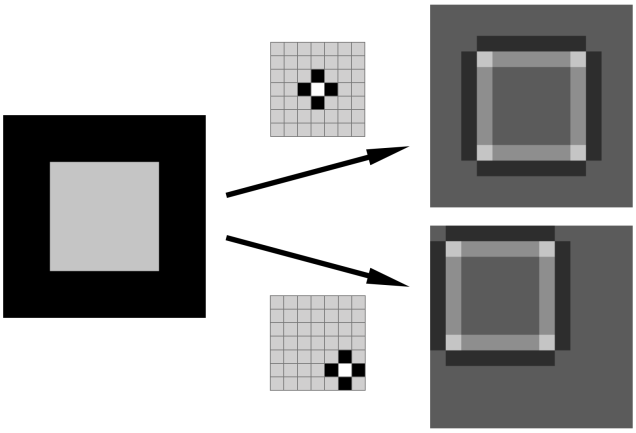


Figure 4.17: Application of two filters (middle) to an input image (left). The result (right) shows that the extracted features are shifted using the filter with shifted coefficients (middle-bottom).

where the diagonal matrices $\mathbf{\Gamma}_j = \mathrm{diag}\{\tilde{\mathbf{x}}^1_{jx}\}$ and $\mathbf{\Lambda}_j = \mathrm{diag}\{\tilde{\mathbf{x}}^1_{jy}\}$ are composed of the partial derivatives in x- and y-directions of $\tilde{\mathbf{x}}^1_j$. As the sparsity of the coefficient matrix is the same as for the global model using brightness constancy in (4.24), solving (4.41) is still very efficient.

Constancy assumptions based on high-order image features are more sensitive to noise. Although Poisson-Gaussian distributed noise is more complex in feature space (compared to brightness), due to the linearity of learned filters, the adaptive weighting scheme for high-order features can be derived analogously as for brightness. In addition, the combined local-global (CLG) scheme can also be used.

Data

We use multiple data sets of real time-lapse fluorescence microscopy images of live cells to evaluate our proposed methods (see Tab. 4.5). Data set “S” consists of four 2D image sequences (S1-S4) and data set “T” contains two 3D volume image sequences (T1 and T2). All these sequences were acquired by a widefield microscope and consist of two channels, one for nuclei of live human cells (U2OS cell line) with different chromatin stainings (H2A-mCherry, YFP-SP100) and the other for subcellular particles (CFP stained PML bodies). Strong deformations occur since the cells are going into mitosis (cell nucleus division). 9 spot-like structures from each 2D sequence and 6 from each 3D sequence in the nucleus channel were manually annotated and tracked.

The other two data sets “A” and “B” were acquired with a confocal microscope [144, 164, 165][1]. In sequences A1 and A2, the nuclei of U2OS cells stained with mCherry-BP1-2 were UV-irradiated in a stripe-like region. The nuclei undergo sig-

[1] The data and annotations are available at https://cbia.fi.muni.cz/CellRegistration

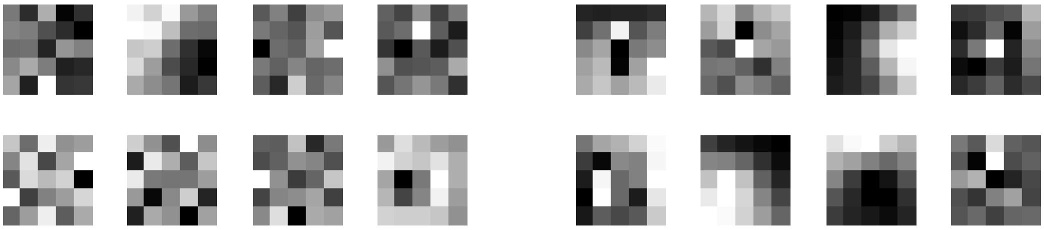
Table 4.5: Real time-lapse microscopy images for evaluation.

Sequence	Length (time points)	Dimension (pixel)	Pixel/voxel size (nm)
S1	150	512×512	216×216
S2	200	384×384	216×216
S3	100	512×512	216×216
S4	149	512×512	216×216
T1	100	512×512×15	216×216×500
T2	149	512×512×10	216×216×1500
A1	25	512×512	240.5×240.5
A2	38	512×512	240.5×240.5
B1	42	287×356	470×470
B2	30	512×512	490×490
B3	42	279×318	540×540
B4	35	322×304	540×540

nificant translation and rotation along with deformations. Key points of the UV-irradiated stripe were manually tracked by three independent annotators, considering that image noisy is rather significant and structures are blurry. Sequences B1-B4 depict HeLa cells with histone H2B tagged by GFP. In each sequence bleached regions form four intersecting line structures, which can be automatically detected (see Fig. 4.23). These line structures are stable w.r.t. the nucleus and represent the motion and deformation of the nucleus, and thus can be used for performance evaluation.

4.3.4 Experiments

We also used synthetic image data for a direct evaluation of the estimated deformation fields. We generated four 2D synthetic sequences (G1-G4). The simulated time-lapse image sequences of cell nuclei should be as realistic as possible and the used deformations should be coherent and consistent with the cell shape. To achieve this, we used deformation fields estimated by an independent registration approach (pairwise local method [148]) from real data (sequences S1-S4) as ground truth for the deformations. The synthetic image sequences were generated using the first frame of the real data and progressively deforming this frame by the ground truth deformation. The sequences comprise between 100 and 200 time points. Note that the ground truth deformation fields represent large deformations. To verify this we computed the mean and standard deviation (STD) of the magnitude of the deformation vectors between the first and the last frame for each synthetic sequence (see Tab. 4.11, second and third rows). It can be seen that the mean and STD are up to 44.3 pixels and 23.8 pixels, respectively, and the values differ largely for the different sequences, which corresponds to different levels of complexity of the deformations.


(a) FoE

(b) cGRBM

Figure 4.18: Learned 2D filters (5×5 pixels) from two generative image models.

Implementation details

For our global non-rigid registration method we employ a standard coarse-to-fine strategy for optimization to deal with large deformations. Gaussian pyramids with 5 scales and a downsampling factor of 0.5 are used. At each image scale, the deformation field is initialized using the result from the coarser scale, and then iteratively updated 5 times using the increment calculated between an image and the warped subsequent image (by the current estimated deformation field).

To train the two generative image models FoE and cGRBM and determine the features, we use 2000 image patches of size 50×50 pixels, which are randomly selected (excluding patches that do not contain any cell nuclei) from the nucleus channel in data set “S”. Since the images in data sets “A” and “B” contain more significant noise and rather blurry structures, which degrades the learning result, we exclude them from the training images. We have trained eight 2D filters of size 5×5 pixels for each model (see Fig. 4.18). It can be seen that, while the learned cGRBM filters appear better spatially structured, the FoE filters are more complex and more difficult to interpret. 3D filters were not used in our experiments due to limited data of data set “T” for learning 3D image models.

For the adaptive weighting and CLG schemes we use empirically determined fixed parameters $b = 10$ and $\sigma = 1$ in (4.31), and $\rho = 1$ in (4.33) in all experiments for both real and synthetic images.

Performance evaluation using real image data

Geometric registration error. Since ground truth of the deformation fields for the real microscopy image sequences is not available, it is not possible to directly evaluate the estimated deformation fields for registration. Evaluation of the intensity similarity between registered images is also inappropriate due to significant noise, intensity changes over time as well as appearing or vanishing structures. Instead, we evaluate the proposed methods by the geometric registration error as the Euclidean distance of an image structure in the registered image to its corresponding location

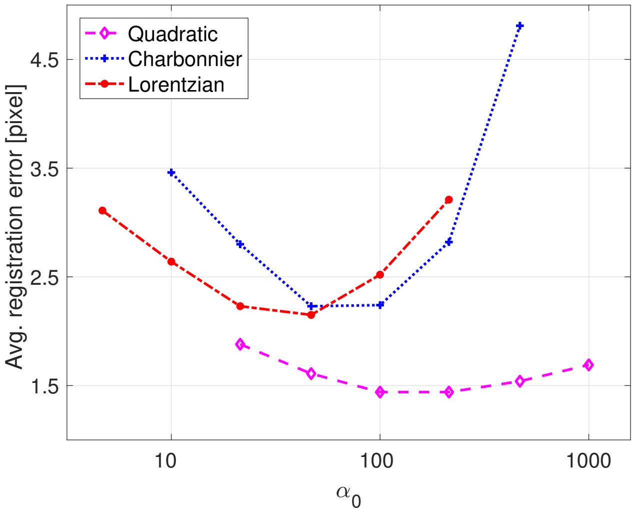


Figure 4.19: Average registration error (in pixel) as a function of the regularization weight α_0 for global models based on brightness constancy and different regularizers for data set "S" (S1-S4).

in the reference image[2]

$$e_{mt} = \|\mathbf{p}_{m0} - \mathbf{p}_{mt}(\mathbf{w}^{0,t})\|_2, \tag{4.42}$$

where $\mathbf{p}_{m0}$ denotes the coordinates of the mth image structure in the reference image and $\mathbf{p}_{mt}(\mathbf{w}^{0,t})$ represents those in the image at time point t transformed by the estimated deformation $\mathbf{w}^{0,t}$. We use manually or automatically annotated image structures in each sequence and compute the mean and standard deviation of the registration errors for these structures.

Comparison of different regularizers. We first compare basic global models (assuming brightness constancy, see Sec. 4.3.2) with three different regularizers, namely quadratic, Charbonnier (with $\epsilon = 0.01$) and Lorentzian (with $\sigma = 0.1$). As the regularization parameter α_0 is the only free model parameter (see (4.23)) and its optimal value depends on the type of deformation field, we applied the models to all sequences in data set "S" with different values of α_0. As can be seen in Fig. 4.19, the quadratic regularizer not only yields the lowest registration error, but is also more robust than the others due to its flatter error curve. This is consistent with our analysis in Sec. 4.3.2 and confirms that strong discontinuities in the deformation fields are unlikely.

Comparison of multiple model variants. We also compared the performance of multiple variants of our global method (using a quadratic regularizer) for data set "S": Models based on brightness constancy ("Br") without and with adaptive

[2] An exception is that, for line features in the sequences B1-B4, the Fréchet distance is computed as the registration error as in [144].

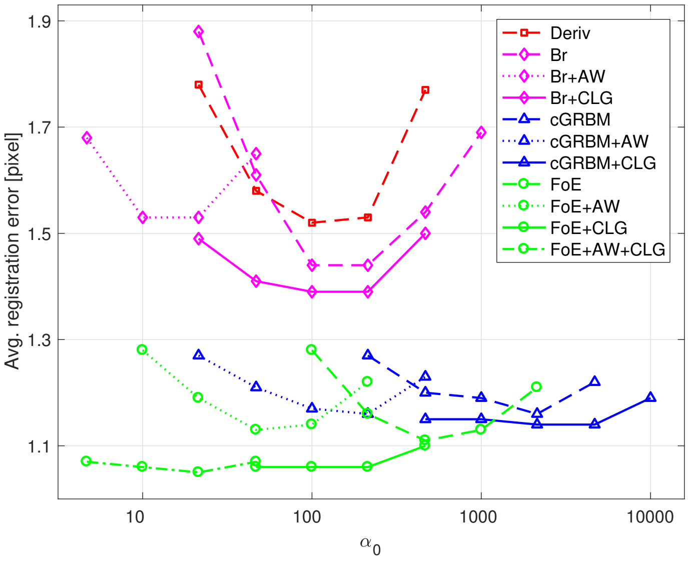
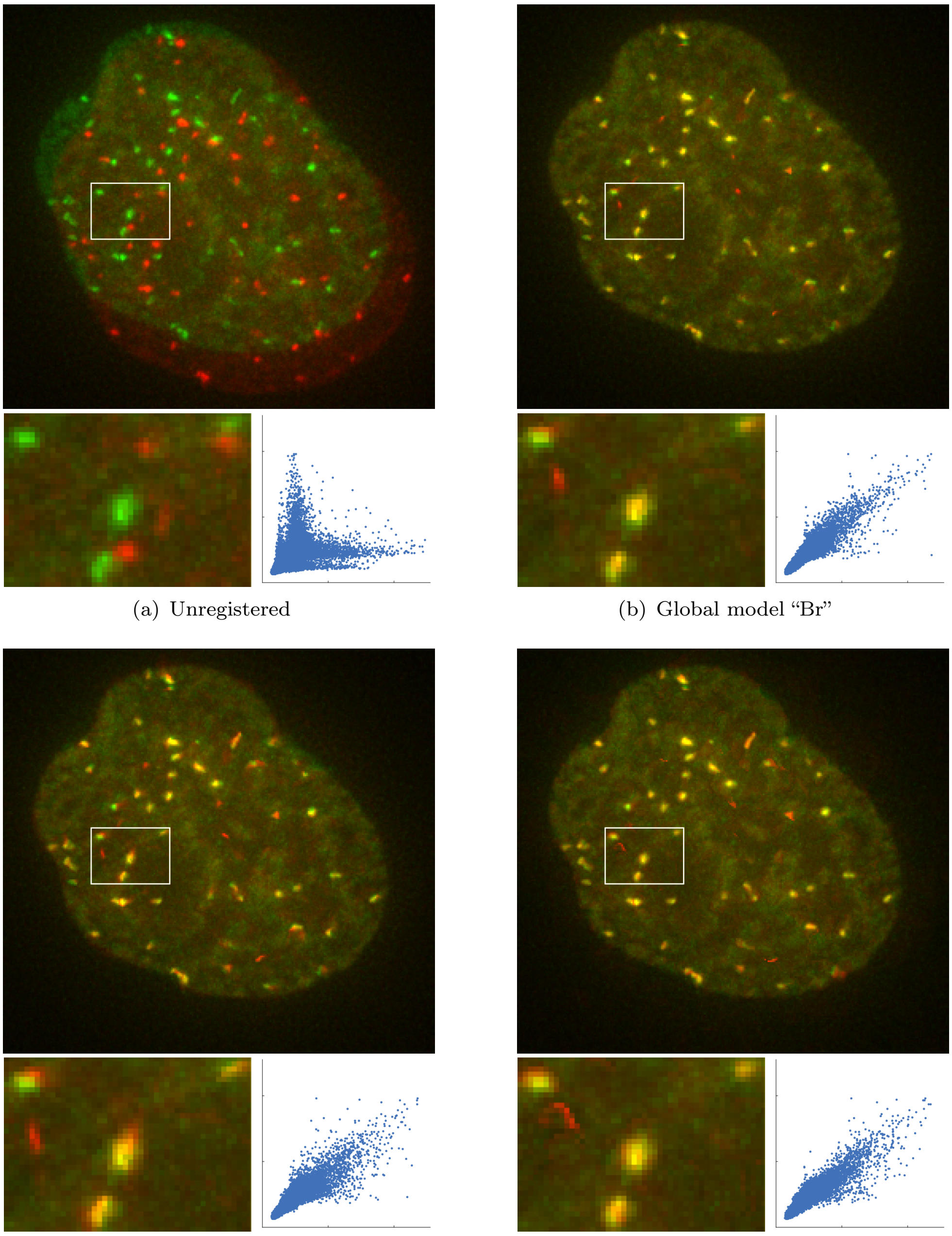


Figure 4.20: Average registration error as a function of the regularization weight α_0 for different model variants of our global method for data set "S" (S1-S4).

weighting ("+AW") or combined local-global ("+CLG"), model based on first-order derivative (gradient) constancy ("Deriv"), and models based on high-order feature constancy ("FoE" using learned FoE features, and "cGRBM" using learned cGRBM features) without and with adaptive weighting and/or CLG.

Fig. 4.20 shows the average registration errors of these models w.r.t. the regularization parameter α_0. It can be seen that the model based on gradient constancy ("Deriv") performs worse than that based on brightness constancy ("Br"), which suggests that first-order derivative features are insufficient to capture structures in noisy microscopy images. Using the learned FoE or cGRBM features significantly improves the result and also increases the robustness w.r.t. α_0, which can be attributed to the more accurate modeling of the data term using learned image features. For FoE we also tested the impact of the scheme for shifting the filter coefficients in our approach (*cf.* Sec. 4.3.3 above). Without using the scheme, we obtained for S1-S4 an average registration error of 1.13 pixels with an average STD of 0.65 pixels, while using the scheme yielded a lower average error of 1.11 pixels and a lower average STD of 0.58 pixels. For this data set, the CLG scheme further improves the performance, while the AW scheme does not yield an improvement. This is not surprising since the noise and outliers in the images have a more complex pattern, but adaptive weighting assumes Poisson-Gaussian noise. The combined FoE+AW+CLG scheme yields a slightly lower average error compared to CLG. In Tab. 4.6 the registration errors and standard deviations of different variants of our global model with optimal α_0 (w.r.t. the whole data set) for all sequences of "S" are provided. The models using learned FoE features are slightly better than (or similar to) those using cGRBM features. An example of registration results for the sequence S4 can be seen in Fig. 4.21. For



(a) Unregistered

(b) Global model “Br”

(c) Global model “Deriv”

(d) Global model “FoE+CLG”

Figure 4.21: Overlay and scatterplot of registered 20th frame (green) with 1st frame (red) of sequence S4, using different model variants of our global method.

Table 4.6: Registration errors and standard deviations (in pixel) for annotated spot-like structures in 2D image sequences S1-S4. Results for proposed global models and previous methods.

Sequence	S1	S2	S3	S4	Avg.	Avg. STD
Unregistered	6.56	13.97	26.65	27.84	18.76	9.53
Local, multi-frame weighting [147]	2.99	2.45	3.56	5.31	3.58	2.11
Local, diffeomorphic multi-frame [148]	2.10	1.38	2.19	2.93	2.15	1.64
Contour-based method [144]	-	-	6.73	-	-	-
Our global model “Br”	1.31	0.95	1.74	1.76	1.44	0.89
Our global model “Br+AW”	1.36	1.03	1.86	1.87	1.53	0.97
Our global model “Br+CLG”	1.28	0.95	1.76	1.58	1.39	0.87
Our global model “cGRBM”	1.10	**0.78**	**1.73**	1.02	1.16	0.68
Our global model “cGRBM+AW”	1.12	0.81	**1.73**	0.99	1.16	0.69
Our global model “cGRBM+CLG”	1.05	0.79	1.83	0.89	1.14	0.67
Our global model “FoE”	1.04	**0.78**	**1.73**	0.89	1.11	**0.58**
Our global model “FoE+AW”	1.06	0.80	1.79	0.87	1.13	0.59
Our global model “FoE+CLG”	0.94	**0.78**	1.78	**0.73**	1.06	0.66
Our global model “FoE+AW+CLG”	**0.92**	**0.78**	1.76	0.74	**1.05**	0.66

our global model “FoE+CLG” more yellow regions in the overlay of registered images are observable, which indicates a higher registration accuracy.

Comparison with existing registration methods. We also performed a comparison with two local methods (local multi-frame weighting [147], local diffeomorphic multi-frame [148]) and a recent contour-based method [144]. For a fair comparison between our global models and previous methods, for data sets “S” and “T”, the model parameter α_0 is optimized for the whole data set (as done for the local models in [147, 148]), and for data sets “A” and “B”, α_0 is optimized for every sequence (as for the contour-based method in [144]). Below, we present results using the optimal value of α_0 obtained by grid search.

For all sequences of data set “S”, it can be seen in Tab. 4.6 that all our global models outperform previous local methods [147, 148] and the contour-based approach [144]. The best result is obtained by the model “FoE+AW+CLG” based on learned FoE features with a combined AW and CLG scheme for handling noise and outliers. The average registration error is decreased by more than 50% compared to [148].

Tab. 4.7 shows the results for the 3D image sequences of data set “T”. As trained 3D image features are not available, only the results for our global model using brightness (“Br” and “Br+CLG”) are provided. Compared to the local models in [147] and [148], our global model achieves a significant decrease of the registration error (75% and 60%, respectively). Fig. 4.22 shows the annotated spot-like structures for sequence T1 after registration by our global model “Br+CLG” compared to the unregistered case, which illustrates the registration accuracy (much smaller variation of the spot locations).

Table 4.7: Registration errors and standard deviations (in pixel) for annotated structures in two 3D image sequences. Results for proposed global models and previous methods.

Sequence	T1	T2	Avg.	Avg. STD
Unregistered	24.94	21.24	23.09	11.69
Multi-frame weighting [147]	6.98	4.63	5.81	3.75
Diffeomorphic multi-frame [148]	4.90	2.41	3.66	2.66
Our global model “Br”	1.83	1.43	1.63	0.96
Our global model “Br+CLG”	**1.67**	**1.27**	**1.47**	**0.83**

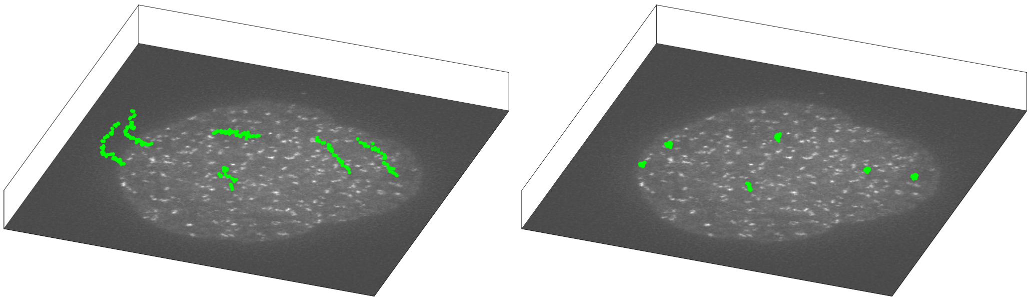


Figure 4.22: Annotated spot-like structures (green dots) in 3D sequence T1. Unregistered (left) and registered (right) by our global model “Br+CLG”.

Sequences in data set “A” are more difficult to register due to significant noise, blurry image structures as well as large intensity variation (see Fig. 4.25). The registration results are presented in Tab. 4.8. It can be seen that our global model generally yields better results than the previous contour-based approach [144]. The best result is obtained for the global model with learned FoE features, for which the average registration error is about 15% lower than that of [144]. It also turns out that the adaptive weighting scheme is superior to the CLG scheme for this data set. The reason is that adaptive weighting more appropriately models the image noise (significant noise but less outliers due to appearing or vanishing structures). Using FoE+AW+CLG improves the result for one image sequence, but the average error for both sequences is slightly worse compared to AW.

Next we consider data set “B”. For performance evaluation we follow [144] and calculate the Fréchet distance between corresponding line features, and the Euclidean distances between corresponding intersection points and end points of the line structures. The quantitative results in Tab. 4.9 show that our global model yields better results than the previous contour-based approach [144]. Using learned features for FoE or cGRBM improves the result compared to the global model with brightness, while the cGRBM features perform slightly better than FoE features. Fig. 4.23 shows an example. Using adaptive weighting or CLG the result is not further improved. This is probably due to the bleached regions which significantly alter the image structure.

Table 4.8: Registration errors (in pixel) for annotated structures in sequences of data set "A".

Sequence	A1				A2			
Annotator	H_1	H_2	H_3	Avg.	H_1	H_2	H_3	Avg.
Unregistered	30.17	30.75	31.29	30.74	49.42	50.20	49.05	49.55
Cont.-based [144]	5.71	7.18	5.78	6.22	7.95	9.10	8.78	8.61
"Br+CLG"	6.87	7.60	5.63	6.70	8.47	8.90	9.60	8.99
"Br+AW"	6.69	7.42	5.38	6.50	7.19	7.85	7.80	7.61
"cGRBM+CLG"	7.27	7.19	5.37	6.61	7.77	8.18	7.88	7.94
"cGRBM+AW"	6.23	6.86	5.14	6.08	7.26	7.74	**7.68**	7.56
"FoE+CLG"	6.01	6.53	4.69	5.74	7.47	7.67	8.65	7.93
"FoE+AW"	5.98	6.81	**3.93**	5.57	**6.45**	**6.78**	7.86	**7.03**
"FoE+AW+CLG"	**5.69**	**6.47**	4.26	**5.47**	6.71	6.95	7.87	7.18

Computation time. We have implemented our approach using Matlab without optimization and performed the experiments on a Linux workstation with an Intel Xeon E5 (2.90GHz) CPU. As the inference of our model involves solving a series of linear systems of equations with sparse coefficient matrix, computation is very efficient. Registration of sequence A2 (all 38 images) takes 1min 37sec by our global method "FoE+AW", while the contour-based approach [144] requires 3min 46sec. Registration of the first 100 images of sequence S2 takes 1min 14sec to 2min 20sec depending on the model variant of our method (see Tab. 4.10 for details). By contrast, the local method [148] needs 1h 37min to 2h 41min depending on the method variant. For the 3D data, while [148] requires more than 27h, our method only needs 8min.

Performance evaluation using synthetic data

Since for the synthetic data the ground truth deformations are known, we evaluate the proposed methods by the endpoint error (EE) defined as the Euclidean distance between an estimated deformation vector and the corresponding ground truth vector.

Tab. 4.11 shows the mean EEs (over all pixels within the region of the cell nuclei and over all time points) of the estimated deformation fields for the four synthetic image sequences (G1-G4). Our global models yield a much better result than the multi-frame local model in [148]. Note that the sequences comprise different levels of complexity of the deformations (see Tab. 4.11, second and third rows for the mean and STD of the magnitude of the deformation vectors). Compared to [148], our model using FoE features decreases the mean EE by more than 50%. As the original synthetic images were generated without adding noise, we did not apply the model variants with AW or CLG.

To evaluate the robustness of our global models w.r.t. different spatial and temporal resolutions, we downsample the synthetic sequences spatially and temporally

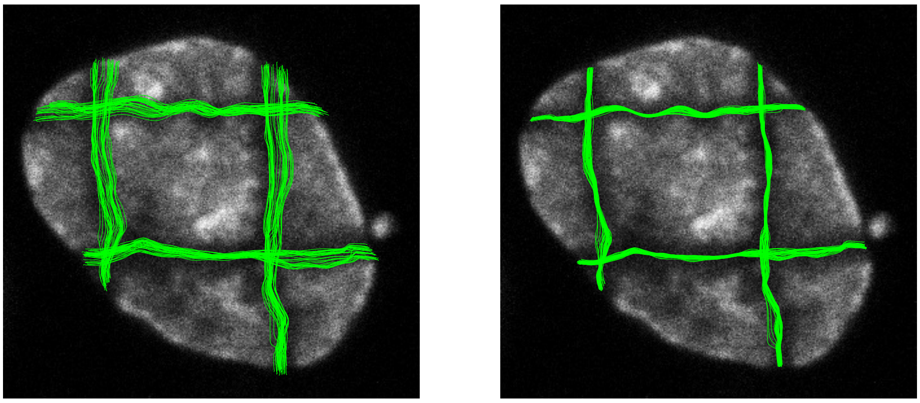


Figure 4.23: Distributions of line structures in sequence B4. Unregistered (left) and registered (right) by our global model "cGRBM".

Table 4.9: Registration errors (in pixel) for annotated structures in sequences of data set "B".

Sequence	B1	B2	B3	B4	Avg.
Line features					
Unregistered	22.91	11.54	12.43	11.92	14.70
Contour-based [144]	9.72	**7.82**	5.32	6.93	7.45
Our "Br"	10.05	8.20	**4.75**	6.46	7.36
Our "FoE"	**8.82**	7.87	5.79	6.51	7.25
Our "cGRBM"	8.97	**7.82**	4.94	**6.37**	**7.02**
Intersection points					
Unregistered	11.30	5.71	8.00	6.79	7.95
Contour-based [144]	7.92	5.22	2.76	3.04	4.73
Our "Br"	7.56	5.40	**2.34**	2.90	4.55
Our "FoE"	**6.51**	**4.80**	2.58	3.06	4.24
Our "cGRBM"	6.71	4.99	2.35	**2.85**	**4.22**
End points					
Unregistered	18.15	6.40	9.30	8.13	10.50
Contour-based [144]	3.69	2.02	2.17	2.43	2.58
Our "Br"	5.84	1.69	**1.81**	1.93	2.82
Our "FoE"	3.58	1.56	1.92	2.08	2.28
Our "cGRBM"	**3.36**	**1.42**	1.82	**1.86**	**2.12**

Table 4.10: Computation time (in seconds) of different model variants and extra computation time compared to "Br" for registering the first 100 images of sequence S2.

Br	Br+AW	Br+CLG	FoE	FoE+AW	FoE+CLG	FoE+AW+CLG
74	78 (+5%)	77 (+4%)	121 (+64%)	126 (+70%)	140 (+89%)	140 (+89%)

Table 4.11: Mean endpoint errors (in pixel) of the estimated deformations for four synthetic image sequences (without noise), and mean as well as STD of the magnitude of the ground truth deformation vectors.

Sequence	G1	G2	G3	G4	Avg.
Mean of vector magnitude	11.9	18.8	44.3	32.7	-
STD of vector magnitude	5.68	8.42	23.8	17.5	-
Multi-frame local method [148]	0.75	1.07	1.89	2.06	1.44
Our global model “Br”	0.47	0.49	0.95	1.25	0.79
Our global model “cGRBM”	0.44	0.48	**0.77**	1.20	0.72
Our global model “FoE”	**0.42**	**0.47**	0.80	**1.11**	**0.70**

Table 4.12: Mean endpoint errors (in pixel) of the estimated deformations for four downsampled synthetic sequences.

Sequence downsampling	Original	1/2 spatial	1/2 temporal
Our model “Br”	0.79	0.96	0.69
Our model “cGRBM”	0.72	0.94	0.68
Our model “FoE”	**0.70**	**0.91**	**0.65**

by a factor of 1/2. From Tab. 4.12 it can be seen that for 1/2 spatial downsampling, the mean EE is only slightly increased, while for 1/2 temporal downsampling, the mean EE is even decreased, since fewer image frames are used. In addition, the results for the different variants of the global model are consistent with those for the original synthetic data (see Tab. 4.11).

We also investigate the robustness of the proposed adaptive weighting (AW) scheme against image noise. We add mixed Poisson-Gaussian (PG) noise, which is typical for fluorescence microscopy images, to the synthetic image sequences with different noise levels according to $x' = s \cdot t + n_N$ and $t \sim \text{Poi}(x/s)$, where x and x' denote intensity values of noise-free and noisy images, s is a gain factor controlling the Poisson noise strength, and n_N is additive Gaussian noise with variance σ_N^2 (example images are shown in Fig. 4.24). The registration results are provided in Tab. 4.13. It can be seen that the proposed adaptive weighting scheme consistently improves the registration accuracy (although we used empirically determined parameter values without knowing the exact noise levels), while the CLG scheme even slightly degrades the performance for strong noise. This observation is consistent with the results for the data set “A” of real images which include significant noise. The combined AW+CLG scheme improves the result compared to CLG and AW except for a high level of mixed PG noise, where AW is best.

Table 4.13: Mean endpoint errors (in pixel) of the estimated deformations for four synthetic sequences with different levels of noise.

Noise level	Gaussian	Poisson	Mixed PG	Mixed PG
	$\sigma_N\!=\!3$	$s\!=\!0.3,\sigma_N\!=\!0$	$s\!=\!0.8,\sigma_N\!=\!3$	$s\!=\!1.2,\sigma_N\!=\!5$
"Br"	1.58	2.02	3.16	3.87
"Br+CLG"	1.50	1.97	3.18	3.95
"Br+AW"	1.43	1.82	2.98	**3.71**
"Br+AW+CLG"	**1.38**	**1.77**	**2.96**	3.80

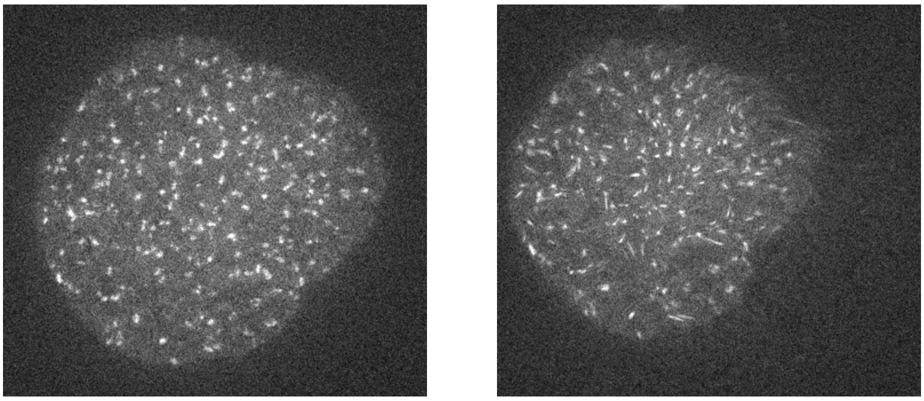


Figure 4.24: Images from the synthetic sequence G3, with mixed Poisson-Gaussian noise ($s\!=\!1.2, \sigma_N\!=\!5$) for $t\!=\!0$ (left) and $t\!=\!50$ (right).

Discussion

From experiments it turned out that our global models for non-rigid registration of cell microscopy images generally outperform existing methods. Best results are achieved by our global models using learned image features (FoE or cGRBM). In general, when the image structures are not artificially altered (e.g., by bleaching as for data set "B"), the registration error using FoE features is lower compared to cGRBM features. The learned FoE image model can better capture the statistics of training images, though FoE filters are less spatially structured and difficult to interpret (see Fig. 4.18). The adaptive weighting (AW) and CLG schemes generally further improve the registration accuracy. For images with significant noise (e.g., data set "A" and synthetic data with Poisson-Gaussian noise), AW is more suitable as the statistics of the noise is considered; for images containing many outliers (e.g., data set "S", where irregular spot-like structures may vanish or arise), CLG yields better results as neighborhood information is utilized.

In practice, it is generally difficult to decide whether AW or CLG is better suited for certain data. If the noise statistics are known (e.g., significant noise or not), one could decide based on this information. Otherwise, a small amount of annotated data could be used to decide which of the two schemes should be applied. Alternatively, the combined AW+CLG scheme can be used. We found that the combined scheme improves the result for certain image sequences, and the average error for the data sets is similar to CLG or AW, while the computation time is nearly the same as for CLG (see Tab. 4.10). An advantage of the combined scheme is that it is not necessary to decide between AW and CLG.

In our experiments we used different types of ground truth depending on the data set. For data set “S” manual annotation was used, which is relatively accurate since localizing spot-like structures and finding correspondences are relatively easy. Data set “A” contains significant noise and blurry structures, thus the annotation was performed by three independent annotators to increase the reliability. For data set “B”, intersecting line structures of bleached regions were automatically detected and used for the evaluation, since these line structures are found to be stable w.r.t. the nucleus. However, the artificial line structures due to bleaching change the properties and statistics of cell microscopy images. As the bleached regions are dark and lack spatial structures, local deformations are hardly captured within these regions. Therefore, the results for data set “B” should be considered with caution.

In our global model we have employed an MRF as the regularizer of the deformation fields, which is not only effective, but also allows arbitrary deformations. Although the regularizer can be straightforwardly extended by additional terms, one should be careful using other constraints or physical prior (e.g., elasticity), because they might be too strict in the case of vanishing or appearing structures in real data.

FoE and cGRBM used for learning features in our global model have a sound probabilistic interpretation and the learned filters can capture important features of the considered images. Their convolutional model structure is not only consistent with the (hidden) independence assumption in the registration model, but also requires moderate amount of training data, which is important for many biological problems with limited data. In our experiments, we found that the models using FoE features yield slightly better registration results than those using cGRBM features, however, training of FoE is slower compared to cGRBM.

4.3.5 Integrating elasticity into the prior

In previous work on non-rigid registration of cell nuclei, often intensity-based methods (e.g., [140, 145, 148] and our approaches described above in Sec. 4.3) were used since changes of image intensity are direct evidence for estimating motion and deformation. As it was found that cell nuclei show elasticity properties and the intranuclear region is stiffer and more elastic than the cytoplasm [156], contour-based methods that exploit elasticity properties were proposed. In [143], the deformation of the nuclear contour is determined from the image data, and deformation vectors inside the nucleus are interpolated using polyharmonic splines. Linear elasticity was explicitly modeled in [144] and combined with contour matching. Final deformation fields are interpolated using thin-plate splines or partial differential equations. However, in [143, 144] deformation estimation only depends on the nucleus contours, while image intensity information is not directly exploited. Thus, local deformations inside the cell nucleus cannot be captured.

Below, we further extend the above introduced global optical flow-based approaches for non-rigid registration of cell nuclei (Sec. 4.3) by integrating elasticity constraints. We analyze the Navier equation from linear elasticity theory and derive an elasticity prior term. Compared to previous contour-based methods for non-rigid registration that use an elasticity model [144], our method exploits intensity changes inside cell nuclei that act as body forces and provide additional information. Experiments using real cell images demonstrate increased registration accuracy of the proposed method. The work was published in [61, 62].

From elasticity theory to an elasticity prior

In linear elasticity theory, the deformation field (displacement vector field) $\mathbf{w}$ for isotropic and homogeneous material can be determined by solving the Navier equation given the body forces $\mathbf{f}$ [166]:

$$\mu\nabla^2\mathbf{w} + (\lambda+\mu)\nabla(\nabla\cdot\mathbf{w}) + \mathbf{f} = 0, \tag{4.43}$$

where μ and λ are the Lamé constants, and ∇ is the nabla operator.

To derive an elasticity prior that is consistent with (4.43), we consider an energy function of a 2D deformation field $\mathbf{w}\in\mathbb{R}^{2N}$ (N is the number of pixels in an image) and the given forces $\mathbf{f}\in\mathbb{R}^{2N}$

$$E(\mathbf{w};\mathbf{f}) = -\mathbf{f}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mu s + \frac{1}{2}(\lambda+\mu)t, \tag{4.44}$$

$$\text{with}\quad s = \|\mathbf{D}_x\mathbf{u}\|^2+\|\mathbf{D}_y\mathbf{u}\|^2+\|\mathbf{D}_x\mathbf{v}\|^2+\|\mathbf{D}_y\mathbf{v}\|^2, \tag{4.45}$$

$$t = \|\mathbf{D}_x\mathbf{u}+\mathbf{D}_y\mathbf{v}\|^2. \tag{4.46}$$

Here $\mathbf{u}\in\mathbb{R}^N$ and $\mathbf{v}\in\mathbb{R}^N$ are x- and y-components of $\mathbf{w}$, respectively, $\mathbf{D}_x$ and $\mathbf{D}_y$ are matrices for computing the first order partial derivatives of $\mathbf{u}$ and $\mathbf{v}$ in x- and y-direction, and $\|\cdot\|$ is the ℓ_2 norm. The necessary condition for the minimization of

(4.44) is

$$
\begin{aligned}
\mathbf{0} &= \nabla_{\mathbf{w}} E(\mathbf{w};\mathbf{f}) \\
&= -\mathbf{f} + \mu \begin{bmatrix} \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_x + \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_x + \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_y \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} + (\lambda + \mu) \begin{bmatrix} \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_x & \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_y \\ \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_x & \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_y \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \\
&= -\mathbf{f} - \mu \begin{bmatrix} \mathbf{u}_{xx} + \mathbf{u}_{yy} \\ \mathbf{v}_{xx} + \mathbf{v}_{yy} \end{bmatrix} - (\lambda + \mu) \begin{bmatrix} \mathbf{u}_{xx} + \mathbf{v}_{yx} \\ \mathbf{u}_{xy} + \mathbf{v}_{yy} \end{bmatrix},
\end{aligned}
\tag{4.47}
$$

where double subscripts of $\mathbf{u}$ and $\mathbf{v}$ denote second order partial derivatives. Since (4.47) is equivalent to (4.43), linear elasticity is included in the energy function in (4.44) to model the displacements given body forces. We can accordingly establish a probabilistic posterior model of displacements under linear elasticity in a Bayesian framework with a likelihood term

$$p(\mathbf{f}|\mathbf{w}) \propto \exp\{\mathbf{f}^{\mathrm{T}}\mathbf{w}\} \tag{4.48}$$

and an elasticity prior

$$p(\mathbf{w};\mu,\lambda) \propto \exp\{-\tfrac{1}{2}\mu s\} \cdot \exp\{-\tfrac{1}{2}(\lambda+\mu)t\}. \tag{4.49}$$

In (4.49), the first term is a common Gaussian Markov random field where the partial derivatives of $\mathbf{u}$ and $\mathbf{v}$ are assumed to be independent (e.g., as in [60]), while the second term is a multivariate Gaussian coupling the partial derivatives of $\mathbf{u}$ in x-direction with the partial derivatives of $\mathbf{v}$ in y-direction.

Deformation estimation using the elasticity prior

We employ the derived elasticity prior in (4.49) to extend the global model under a Bayesian framework for estimating the deformation of nuclei. Given two consecutive frames $\mathbf{x}^0$ and $\mathbf{x}^1$, the likelihood term is modeled based on image intensities under a brightness constancy assumption

$$p(\mathbf{x}^0|\mathbf{w},\mathbf{x}^1) \propto \prod_i \exp\Big\{-\frac{1}{\alpha_i}(\mathbf{x}_i^0 - \mathbf{x}_{\mathbf{w}i}^1)^2\Big\}, \tag{4.50}$$

where i is the index of image pixel locations, α_i the regularization weight, and $\mathbf{x}_{\mathbf{w}}^1$ the warped image of $\mathbf{x}^1$ towards $\mathbf{x}^0$ by the deformation field $\mathbf{w}$. To handle the problem that the relative motion of sub-cellular structures deteriorates the estimation of nucleus deformation, we detect sub-cellular structures in the image data and ignore the corresponding pixels by setting $\alpha_i = \infty$. For the other pixels, we use an adaptive scheme (as in Sec. 4.3.2) to determine the weights α_i and improve the robustness to noise and outliers.

For the regularization term, we employ the derived elasticity prior

$$p(\mathbf{w}) \propto \exp\{-\beta s\} \cdot \exp\{-\gamma t\}, \tag{4.51}$$

where $\beta = \frac{1}{2}\mu$ and $\gamma = \frac{1}{2}(\lambda + \mu)$ are model parameters, and s and t are defined in (4.45) and (4.46). Compared to the regularization term in (4.22), we have an extra term that penalizes the divergence of the deformation field.

The deformation field $\mathbf{w}$ between $\mathbf{x}^0$ and $\mathbf{x}^1$ is computed using the maximum a-posteriori (MAP) estimate of the posterior $p(\mathbf{w}|\mathbf{x}^0, \mathbf{x}^1) \propto p(\mathbf{x}^0|\mathbf{w}, \mathbf{x}^1) \cdot p(\mathbf{w})$. A first order Taylor expansion is used to linearize the likelihood term in (4.50) assuming small deformation. The MAP estimate can be obtained by solving the linear system of equations

$$\left(\begin{bmatrix} \mathbf{\Lambda\Gamma}_{\mathbf{x}^1_x} \\ \mathbf{\Lambda\Gamma}_{\mathbf{x}^1_y} \end{bmatrix} \begin{bmatrix} \mathbf{\Gamma}_{\mathbf{x}^1_x} \\ \mathbf{\Gamma}_{\mathbf{x}^1_y} \end{bmatrix}^{\mathrm{T}} + \beta \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{bmatrix} + \gamma \begin{bmatrix} \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_x & \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_y \\ \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_x & \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_y \end{bmatrix} \right) \mathbf{w} = \begin{bmatrix} \mathbf{\Lambda\Gamma}_{\mathbf{x}^1_x} \\ \mathbf{\Lambda\Gamma}_{\mathbf{x}^1_y} \end{bmatrix} (\mathbf{x}^0 - \mathbf{x}^1), \tag{4.52}$$

where $\mathbf{\Gamma}_{\mathbf{x}^1_x} = \mathrm{diag}(\mathbf{x}^1_x)$ and $\mathbf{\Gamma}_{\mathbf{x}^1_y} = \mathrm{diag}(\mathbf{x}^1_y)$ are diagonal matrices containing the first order partial derivatives of $\mathbf{x}^1$ in x- and y-direction, $\mathbf{\Lambda} = \mathrm{diag}([\cdots, {}^1\!/\!\alpha_i, \cdots])$ represents the weights α_i, and $\mathbf{F} = \mathbf{D}_x^{\mathrm{T}}\mathbf{D}_x + \mathbf{D}_y^{\mathrm{T}}\mathbf{D}_y$. Solving (4.52) is very efficient since the coefficient matrix on the left side is sparse and positive definite. To deal with large deformations, we use a standard coarse-to-fine strategy.

Extension for 3D images

We have noticed that a direct extension of the registration method with elasticity constraints for 3D images by following the strategy described in Sec. 4.3.2 yields unsatisfactory registration results under complex conditions, for example, strongly different resolution in x- and y-direction compared to the z-direction, much fewer slices compared to image height and width (resulting in difficulties of using a coarse-to-fine strategy), large deformation along the z-axis, and reduced intensity information due to excluding some image structures. To address these issues, we suggest three extensions. First, we use different weights in the regularization term of the deformation model for the partial derivatives in z-direction than in x- and y-direction according to the difference in resolution. Second, we employ a special 3D image pyramid for the coarse-to-fine strategy to handle large deformations. We use multiple scales for the x- and y-direction, but only one scale for the z-direction; and use a varying weight for each pyramid scale. Third, we perform an affine pre-registration.

Table 4.14: Registration errors (in pixel) for annotated feature points in image sequences of dataset A.

Sequence	A1				A2			
Annotator	H_1	H_2	H_3	Avg.	H_1	H_2	H_3	Avg.
Unregistered	30.17	30.75	31.29	30.74	49.42	50.20	49.05	49.55
Contour-based [144] (static)	5.96	7.20	6.01	6.39	8.23	9.45	8.97	8.89
Contour-based [144] (dynamic)	5.71	7.18	5.78	6.22	7.95	9.10	8.78	8.61
Ours (Br+AW)	6.69	7.42	5.38	6.50	7.19	7.85	7.80	7.61
Ours (FoE+AW)	5.98	6.81	3.93	5.57	6.45	6.78	7.86	7.03
Ours (Br+AW+Elasticity)	**5.36**	**5.29**	**3.75**	**4.80**	**5.63**	**6.62**	**6.67**	**6.31**

Experiments

For a quantitative evaluation of the proposed method we use two datasets of live cell microscopy image sequences: The first is dataset A (see Sec. 4.3.3 for details) and the second (dataset D) is an image sequence including 10 frames of 350×350 pixels generated by a confocal microscope with a resolution of 104nm×104nm, showing replication foci (expressed by fluorescently tagged PCNA) in nuclei of HeLa cells during S-Phase [167].

The model parameters β and γ in (4.51) were optimized by first choosing a suitable value for β, and then setting γ empirically to $5 \sim 10$ times larger than β. For dataset D, to exclude replication foci from deformation estimation, we used a Laplacian-of-Gaussian (LoG) spot detector with $\sigma = 2$ to locate foci. For all pixels with a distance of no more than 5 pixels to detected foci, α_i in (4.50) was set to ∞. For the coarse-to-fine strategy, 5 and 4 image scales were used for datasets A and D, respectively. The registration of an entire image sequence is accomplished based on an incremental scheme (see Sec. 4.3.2).

We compared the registration results of the proposed method with those of the contour-based method [144] which also includes an elasticity model (static or dynamic) and the above introduced global optical flow methods with the brightness feature (Sec. 4.3.2) and learned high-order FoE features (Sec. 4.3.3). The results are provided in Tab. 4.14. It can be seen that the registration error of the proposed method with elasticity constraints is much lower than that of the previous methods. We attribute this to the fact that the proposed method exploits intensity information inside cell nuclei to capture local deformation. Note that the proposed method with elasticity constraints uses only brightness information but outperforms the model using learned high-order FoE features (Sec. 4.3.3). This demonstrates that it is important to use an elasticity model for registration. Fig. 4.25 shows the registration result of the proposed method for sequence A2 illustrated by the distribution of feature points in all 38 frames of the sequence from one annotator. It can be seen that the variation of the feature points in the registered images is much smaller.

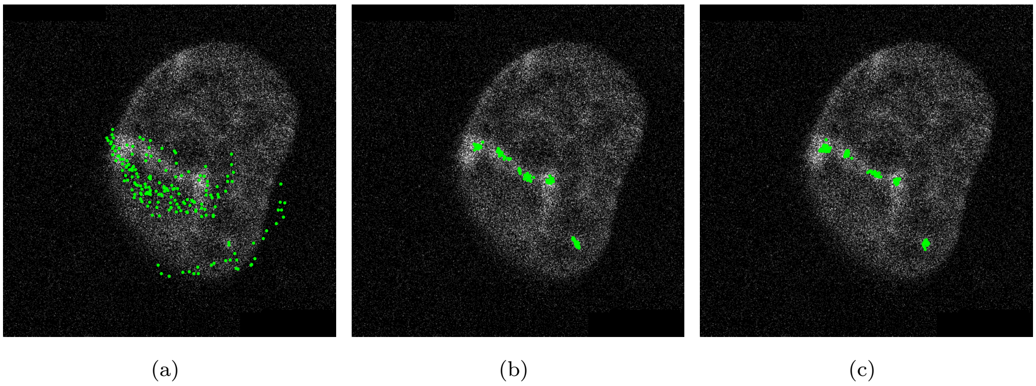


Figure 4.25: Example registration result. Distribution of feature points in all 38 frames of sequence A2 from one annotator, overlaid with the 1st frame. (a) Unregistered; (b) registered by our global model "FoE+AW"; (c) registered by our method with elasticity constraints.

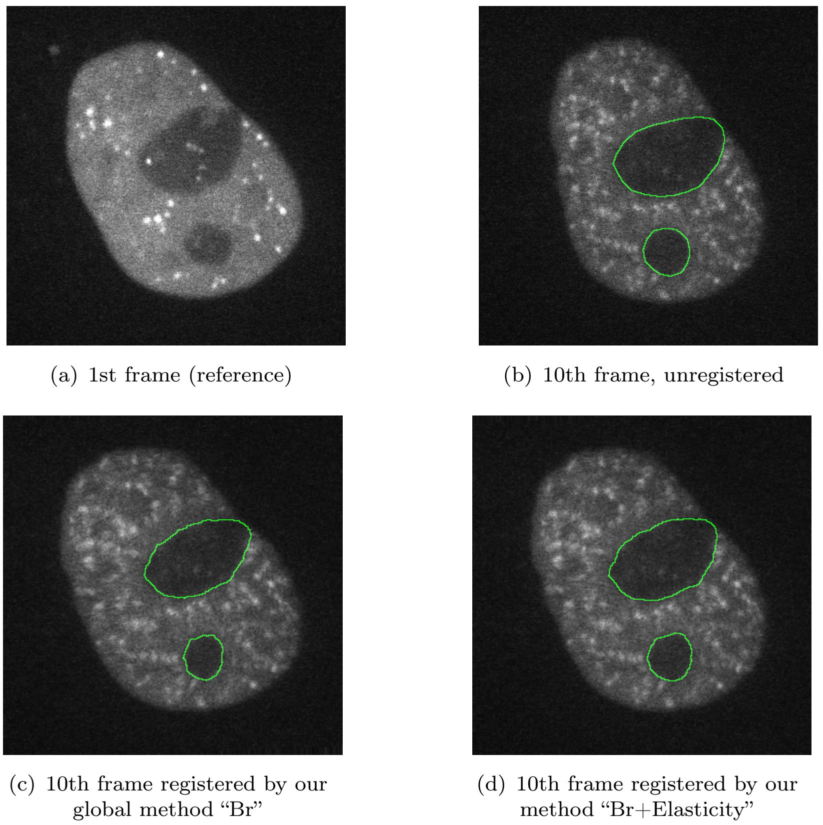
For dataset D, it is very difficult to define corresponding feature points for evaluation due to appearing and disappearing foci structures. However, there exist hole structures inside cell nuclei (nucleoli), which can be straightforwardly annotated (Fig. 4.26(b)). Exploiting the warped annotations of nucleoli in the registered images (Fig. 4.26(c,d)), the registration accuracy can be quantified. For a comparison, we computed the area of the nucleoli. The results are shown in Fig. 4.27. It can be seen that the proposed method with elasticity constraints much better preserves the size of the nucleoli in the registered images compared to our previous method using only the brightness feature.

The computation time of the proposed method with elasticity constraints is only a little bit higher than that of our base global model (e.g., 1min 32sec *vs.* 1min 20sec for registration of sequence A2 on a Linux PC with an Intel Core i9 3.60GHz CPU). The reason is that the coefficient matrix of the linear system of equations in (4.52) is slightly less sparse. As the deformation fields between all pairs of consecutive frames are estimated independently, parallelization is straightforward and can be used to further reduce the computation time.

4.4 Summary

In this chapter, we introduced new methods for three real-world image analysis problems. Each method exploits the learned high-order MRFs, either as image priors or for image feature extraction.

For natural image denoising, we proposed an efficient sampling-based method for Bayesian minimum mean squared error (MMSE) estimation (Sec. 4.1). The auxiliary-variable block Gibbs sampler was extended to allow the MMSE estimate to


(a) 1st frame (reference)

(b) 10th frame, unregistered

(c) 10th frame registered by our global method “Br”

(d) 10th frame registered by our method “Br+Elasticity”

Figure 4.26: Registration results of the 1st and 10th frame of dataset D. Green contours indicate nucleoli inside the nucleus in the unregistered and registered frames.

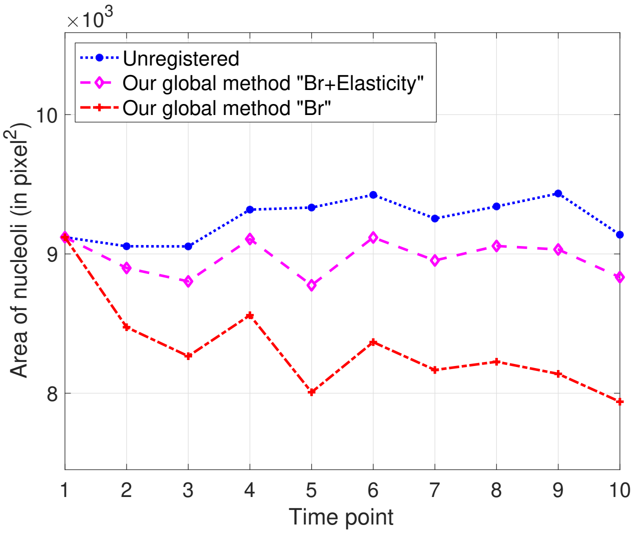


Figure 4.27: Quantified area of nucleoli in the registered images of dataset D.

be efficiently approximated by averaging the samples from the posterior distribution. We demonstrated that our approach with a purely generative setting (without any ad-hoc modifications) substantially outperforms MAP estimation and can compete with popular denoising methods.

The microscopy image deconvolution problem is a typical example that involves a complex likelihood model (Sec. 4.2). The mixed Poisson-Gaussian statistics of noise in microscopy images were approximated by mixtures of Gaussians to model the likelihood, while the learned high-order MRFs acted as the prior. To efficiently sample the posterior to approximate the MMSE estimate, the auxiliary-variable Gibbs sampler was further extended. Super-resolution, deconvolution, and denoising of images can be jointly performed in our proposed method, which can be straightforwardly adapted to other image restoration problems with arbitrary linear degradation model as well as other types of noise. We also found that, for a posterior model with the learned generative MRF as the prior, an accurate modeling of the likelihood is crucial for the model performance.

In Sec. 4.3 we introduced a global optical flow-based method to estimate the complex deformation of live cell nuclei and accomplish non-rigid registration by an incremental scheme. Based on the study of noise statistics in fluorescence microscopy images, an adaptive weighting scheme was developed to increase model robustness. We further extended the global model by exploiting high-order image features of cell nuclei beyond brightness, which were extracted using filter banks obtained by training high-order MRFs, in particular, the FoEs and the convolutional Gaussian RBMs. Using multiple data sets of real 2D and 3D live cell microscopy image sequences as well as synthetic image data, we showed that our proposed approach outperforms previous methods in terms of both registration accuracy and computational efficiency. In addition, elasticity properties were integrated into the MRF prior to achieve more robust registration performance.

Chapter 5

Summary and Outlook

5.1 Summary

Markov random fields (MRF) based on linear filter responses, also termed filter-based MRFs in this dissertation, are one of the most popular forms for modeling image priors. Note that regularization in variational methods can generally also be interpreted as applying a filter-based MRF prior. The rigorous probabilistic interpretation of MRFs has several distinct advantages. For example, the modeling choices can be made based on statistical properties of the data, and the most suitable model parameters can be determined using learning algorithms. Compared to local models (e.g., patch-based image priors), MRFs are global models of entire images (regardless of image size) and can quite easily be applied to a wide range of problems without any adaptation. Bayesian probabilistic models with MRFs allow advanced statistical inference methods to be exploited to find the solution. Meanwhile, information from different sources can be combined in a principled way and the uncertainty of the solution is maintained.

In this dissertation, we have explored how well filter-based MRFs can model images and developed new image analysis methods based on the learned high-order MRF priors.

Application-independent evaluation, effective learning, and understanding of MRF image priors

To directly and quantitatively evaluate MRF priors in an application-independent way, we developed an efficient auxiliary-variable Gibbs sampler for a general filter-based MRF with potentials represented by flexible Gaussian scale mixtures (GSM). Multiple generative properties of MRF priors can thus be analyzed using model samples. We found that popular pairwise and high-order MRFs, despite success in various applications, were not able to capture the statistics of the inbuilt features. This appeared in contradiction to the maximum entropy interpretation of filter-based MRFs, implying that the potentials of these models were not properly chosen or learned. In fact, it is the incoincidence between learning objective and model evaluation that often leads to the unawareness of failures in learning or model choices.

We used our developed auxiliary-variable Gibbs sampler to learn pairwise MRFs and high-order Fields of Experts (FoE) with flexible GSM potentials for generic image priors. A maximum likelihood learning procedure based on sampling and contrastive divergence only worked well for pairwise MRFs. We further proposed strategies to solve the learning issues for the high-order models. We found that all learned pairwise MRF and high-order FoEs capture the statistics of model features correctly, thus being real maximum entropy models. Moreover, the learned FoEs also quite accurately represent multi-scale derivative statistics, random filter statistics as well as joint feature statistics of natural images. This demonstrates that FoEs are indeed capable of capturing a large number of key statistical properties of natural images.

Despite being flexible, the GSM-based potentials are uni-modal, which prevents the MRFs from well modeling some special types of images, e.g., visual textures. We proposed using a mixture of multiple GSMs to enable multi-modal potentials, thus improving model expressiveness of MRFs. Such extension helped us to find more insights of MRFs through revisiting the seminal FRAME model. To further understand the importance of "covariance" units that impose regularization in typical MRFs for visual textures, we proposed to learn "mean"-only fully convolutional Gaussian restricted Boltzmann machines (RBM) for modeling Brodatz textures. Our learned models exhibit several favorable properties as well as structured and more interpretable features; yet they outperform more complex and deep models in texture synthesis and in-painting. The results demonstrated that the "mean" units in MRFs actually take the most important role in modeling textures.

Image analysis methods based on learned MRF priors

For natural image denoising, we defined a typical posterior model in a Bayesian framework, where the learned high-order FoEs for generic images were used as the prior in a purely generative setting. We first showed that popular MAP estimation does not yield satisfactory results. To take full advantage of generative MRFs, we then adapted the auxiliary-variable Gibbs sampler to sample the posterior distribution and infer the MMSE estimate. Experiments showed that our approach can compete with popular denoising methods. Moreover, the MMSE estimate not only substantially outperforms MAP, but also avoids several of its problems (e.g., ad-hoc modifications and inherent bias toward δ-like marginals). We also demonstrated that a rigorous probabilistic interpretation and good generative properties can go hand-in-hand with very good application performance.

Regarding microscopy image deconvolution, we first studied the statistics of mixed Poisson-Gaussian noise in microscopy images and approximated the distributions using mixtures of Gaussians (MoG). A Bayes deconvolution model composed of a learned high-order MRF prior and an MoG-based likelihood allowed a further extension of the developed efficient block Gibbs sampler. The degraded microscopy images are then restored by approximating the MMSE estimate using samples from the complex posterior. Under the framework of our model, super-resolution, deconvolution, and denoising can be jointly performed. Experiments demonstrated that the deconvolution performance of our method can compete with state-of-the-art methods. Our method was also applied to real microscopy images of telomeres acquired via stimulated emission depletion (STED) nanoscopy for distinguishing different experimental conditions.

Registration of live cell nuclei in time-lapse microscopy images is difficult not only due to complex nucleus deformation but also due to significant noise in the images. We introduced a global optical flow-based method to estimate the deformation of cell nuclei and accomplish non-rigid registration by an incremental scheme. Our experiences from leaning MRF image priors helped to rapidly determine the most suitable regularizer of the deformation fields in the model. Based on a study of noise statistics in fluorescence microscopy images, an adaptive weighting scheme was developed to increase model robustness. We further extended the global model by exploiting high-order image features beyond brightness. As the model formulations of FoEs and convolutional Gaussian RBMs are both consistent with the assumption of high-order feature constancy in the registration model, we trained filter banks with these generative MRF models for extracting high-order features of cell nuclei, thus achieving increased registration accuracy. Using multiple data sets of real 2D and 3D live cell microscopy image sequences as well as synthetic image data, we showed that our proposed approach outperforms previous methods in terms of both registration accuracy and computational efficiency. This also demonstrated that high-order MRFs are good choices for image feature learning, since good generative properties imply that important image features have been captured by the learned filters.

5.2 Outlook and future work

In this dissertation, we have seen that the high-order MRFs, in particular FoEs, not only show impressive power of modeling image statistics when properly trained, but also exhibit competitive performance in a variety of image analysis problems with a pure generative setting. However, the learned MRF priors and our approaches still have a number of limitations. Below, we briefly discuss some of them which may motivate future work in this area.

Although our learned FoEs accurately capture important statistics of natural images, the model samples look not as realistic as real images: Typical image structures, e.g., sharp edges, closed contours, and flat regions, are hardly observed (*cf*. Fig. 3.11). Pushing further to learn more and larger filters is not practical. Many filters lead to slower mixing, larger ones to less-sparse linear equation systems in sampling. Some work attributed the reason of non-realistic samples to be that the FoEs with GSM potentials do not have the "mean" units and thus can not appropriately provide some structure information. We thus tried learning FoEs with mixtures of GSMs-based multi-modal potentials to allow for mean units, but found this not very helpful. We have also exploited the idea of applying filters at multiple image scales from the FRAME model and trained FoEs with such a hierarchical structure. This model was not only difficult to train, but yielded "inactive" filters and experts, suggesting this structure can not further improve the FoE as a generic image prior. This observation was not surprising because multi-scale derivative statistics have already be well captured by GSM-FoEs. Therefore, further gains as well as realistic model samples are challenging and may require new model designs.

Another reason for the non-realistic samples might be that our FoEs are homogeneous. The same clique potential function is used throughout the entire image due to the convolutional structure, which limits the number of model parameters and makes them manageable. As different image content (e.g., sky, trees, cars) has different statistical properties, content information (e.g., from a higher-level model of scene understanding) will be useful for selecting best clique potentials to obtain inhomogeneous or content-aware MRFs. Exploiting structure tensors as a content indicator, we have trained a content-aware MRF based on the steerable random fields and the multi-modal mixture of GSM potentials as well as proposed learning strategies. While model samples contain distinct image structures, the MRF was trained in the framework of conditional random fields (CRF), thus not being a pure prior. In addition, due to smaller cliques and fewer experts, its application performance is not as good as our learned FoEs. The scene-aware MRF [13] employs an explicit scene coefficient and is not a pure prior either. To learn a pure content-aware MRF prior, additional latent variables may need to be defined, while these variables are associated with some spatially dependent prior and may be marginalized out.

One problem that we encountered in the applications with a generative setting is the modeling of the application-specific likelihood. The denoising experiment of microscopy image deconvolution application (*cf*. Fig. 4.10 and Tab. 4.3) showed that the accuracy of likelihood modeling is crucial for the model performance. In practice, it is hard to ensure high accuracy and many likelihoods have actually remained hand-tuned. Even when the likelihood is learned from training data, the choice of likelihood representation functions is still an issue, because it directly affects the modeling accuracy and the feasibility of probabilistic inference. On the other hand, the likelihood also determines which estimate should be used for inferring the solution. When the likelihood function imposes relatively strong constraints and only has one (major) modality (e.g., denoising of images with additive Gaussian noise), an MMSE estimation is generally more suitable. When the posterior distribution has two or more modalities, an MAP estimate from any modality probably makes more sense than simply averaging all modalities.

In recent years deep neural networks are getting increasingly popular. Deep probabilistic or generative models of generic images have been proposed (see Sec. 2.3.4), aiming at capturing high-level image structures and generating realistic samples for various purposes such as artwork, super-resolution, and colorization. However, these deep generative models are hard to be used like MRF priors with a generative setting in a Bayesian framework. Actually, MRF priors studied in this dissertation can be combined with deep networks, sometimes quite straightforwardly, for a number of vision tasks. For example, deep networks can be trained as efficient solvers for inference with MRFs (see Sec. 2.3.3). Combining knowledge or physics-driven approaches (e.g., the MRFs in Sec. 4.3.5) and data-driven deep learning is one of the directions for our future research.

Bibliography

[1] A. Hyvärinen, P. O. Hoyer, and J. Hurri, "Extensions of ICA as models of natural images and visual processing," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 963–974, 2003.

[2] M. Welling, G. E. Hinton, and S. Osindero, "Learning sparse topographic representations with products of Student-t distributions," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1359–1366, 2002.

[3] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 479–486, 2011.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[5] M. F. Tappen, B. C. Russell, and W. T. Freeman, "Exploiting the sparse derivative prior for super-resolution and image demosaicing," in *Proceedings of the 3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.

[6] S. C. Zhu and D. Mumford, "Prior learning and Gibbs reaction-diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1236–1250, 1997.

[7] S. Roth and M. J. Black, "Fields of experts," *International Journal of Computer Vision*, vol. 82, no. 2, pp. 205–229, 2009.

[8] S. C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, 1998.

[9] N. Heess, C. K. I. Williams, and G. E. Hinton, "Learning generative texture models with extended Fields-of-Experts," in *British Machine Vision Conference (BMVC)*, 2009.

[10] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[11] Y. Chen, R. Ranftl, and T. Pock, "Insights into analysis operator learning: From patch-based sparse models to higher order MRFs," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1060–1072, 2014.

[12] S. Roth and M. J. Black, "Steerable random fields," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 377–387, 2007.

[13] D. Gong, Y. Zhang, Q. Yan, and H. Li, "Learning scene-aware image priors with high-order Markov random fields," *Applied Informatics*, vol. 4, no. 12, 2017.

[14] J. Sun and M. F. Tappen, "Learning non-local range markov random field for image restoration," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2745–2752, 2011.

[15] Z. Wu, D. Lin, and X. Tang, "Deep Markov random field for image modeling," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 295–312, 2016.

[16] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep learning markov random field for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1814–1828, 2017.

[17] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8934–8943, 2018.

[18] Z. Ren, J. Yan, X. Yang, A. Yuille, and H. Zha, "Unsupervised learning of optical flow with patch consistency and occlusion estimation," *Pattern Recognition*, vol. 103, no. 107191, 2020.

[19] B. D. de Vos, F. F. Berendsen, M. A. Viergeve, H. Sokooti, M. Staring, and I. Isgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.

[20] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, and J. Guttag, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, pp. 1788–1800, 2019.

[21] Y. Zhu, Z. Zhou, G. Liao, and K. Yuan, "A new unsupervised learning method for 3D deformable medical image registration," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 908–912, 2021.

[22] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1712–1722, 2019.

[23] H. Chen, Y. Zhang, Y. Chen, J. Zhang, W. hua Zhang, H. Sun, Y. Lv, P. Liao, J. Zhou, and G. Wang, "LEARN: Learned experts' assessment-based reconstruction network for sparse-data CT," in *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1333–1347, 2018.

[24] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International Conference on Machine Learning (ICML)*, pp. 1747–1756, 2016.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014.

[27] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning (ICML)*, vol. 97 of *PMLR*, pp. 7354–7363, 2019.

[28] G. E. Hinton, "Products of experts," in *International Conference on Artificial Neural Networks (ICANN)*, vol. 1, pp. 1–6, 1999.

[29] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1971, 2009.

[30] Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton, "Energy-based models for sparse overcomplete representations," *Journal of Machine Learning Research*, vol. 4, pp. 1235–1260, 2003.

[31] H. Scharr, M. J. Black, and H. W. Haussecker, "Image statistics and anisotropic diffusion," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 840–847, 2003.

[32] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, 2003.

[33] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[34] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

[35] M. Norouzi, M. Ranjbar, and G. Mori, "Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2735–2742, 2009.

[36] J. J. Kivinen and C. K. I. Williams, "Multiple texture Boltzmann machines," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 638–646, 2012.

[37] H. Luo, P. L. Carrier, A. Courville, and Y. Bengio, "Texture modeling with convolutional spike-and-slab RBMs and deep extensions," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 415–423, 2013.

[38] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1222–1239, Nov. 2001.

[39] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[40] U. Köster, J. T. Lindgren, and A. Hyvärinen, "Estimating Markov random field potentials for natural images," in *Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)* (T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, eds.), vol. 5441 of *Lecture Notes in Computer Science*, pp. 515–522, Springer, 2009.

[41] S. Lyu and E. P. Simoncelli, "Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 693–706, 2009.

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

[43] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 900–907, 2003.

[44] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields," in *Proceedings of European Conference on Computer Vision (ECCV)*, vol. 2, pp. 16–29, 2006.

[45] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[46] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black, "Efficient belief propagation with learned higher-order Markov random fields," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 269–282, 2006.

[47] C. Fox and G. K. Nicholls, "Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited," in *AIP Conference Proceedings*, vol. 568, 2001.

[48] K. Schelten and S. Roth, "Mean field for continuous high-order MRFs," in *Pattern Recognition, Proceedings of DAGM-Symposium*, pp. 52–61, 2012.

[49] Y. Marnissi, Y. Zheng, E. Chouzenoux, and J.-C. Pesquet, "A variational Bayesian approach for image restoration. Application to image deblurring with Poisson-Gaussian noise," *IEEE Trans. Comput. Imaging*, vol. 3, no. 4, pp. 722–737, 2017.

[50] U. Schmidt, Q. Gao, and S. Roth, "A generative perspective on MRFs in low-level vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1751–1758, 2010.

[51] Q. Gao and S. Roth, "How well do filter-based MRFs model natural images?," in *Pattern Recognition, Proceedings of DAGM-Symposium*, vol. 7476 of *Lecture Notes in Computer Science*, pp. 62–72, Springer, 2012.

[52] "www.ux.uis.no/~tranden/brodatz.html."

[53] Q. Gao and S. Roth, "Texture synthesis: from convolutional RBMs to efficient deterministic algorithms," in *IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition (S+SSPR)*, vol. 8621, pp. 434–443, Springer, 2014.

[54] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2272–2279, 2009.

[55] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[56] O. J. Woodford, C. Rother, and V. Kolmogorov, "A global perspective on MAP inference for low-level vision," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2319–2326, 2009.

[57] F. Görlitz, P. Hoyer, H. J. Falk, L. Kastrup, J. Engelhardt, and S. W. Hell, "A STED microscope designed for routine biomedical applications," *Progress In Electromagnetics Research*, vol. 147, pp. 57–68, 2014.

[58] Q. Gao, S. Eck, J. Matthias, I. Chung, J. Engelhardt, K. Rippe, and K. Rohr, "Bayesian joint super-resolution, deconvolution, and denoising of images with Poisson-Gaussian noise," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 938–942, 2018.

[59] Q. Gao and K. Rohr, "Optical flow-based non-rigid registration of cell nuclei: Global model with adaptively weighted regularization," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 420–423, 2017.

[60] Q. Gao and K. Rohr, "A global method for non-rigid registration of cell nuclei in live cell time-lapse images," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2259–2270, 2019.

[61] Q. Gao, V. O. Chagin, M. C. Cardoso, and K. Rohr, "Non-rigid registration of live cell nuclei using global optical flow with elasticity constraints," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1457–1460, 2021.

[62] Q. Gao, V. Chagin, M. C. Cardoso, and K. Rohr, "Quantifying newly appearing replication foci in cell nuclei based on 3D non-rigid registration," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, 2022.

[63] S. Z. Li, *Markov Random Field Modeling in Image Analysis.* Springer, 2nd ed., 2001.

[64] J. Besag, "Spatial interaction and the statistical analysis of lattices," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 36, no. 2, pp. 192–236, 1974.

[65] F. R. Kschischang, B. J. Frey, and H.-A. Loelinger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, pp. 498–519, Feb. 2001.

[66] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[67] J. Huang, *Statistics of Natural Images and Models.* PhD thesis, Brown University, 2000.

[68] Z. Tu and S.-C. Zhu, "Image segmentation by data-driven Markov chain Monte Carlo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 657–673, May 2002.

[69] A. Barbu and S.-C. Zhu, "Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1239–1253, 2005.

[70] M. Á. Carreira-Perpiñán and G. E. Hinton, "On contrastive divergence learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 33–40, 2005.

[71] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 695–703, 2004.

[72] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, pp. 179–201, June 2006.

[73] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, pp. 5–43, Jan. 2003.

[74] R. Fletcher, *Practical Methods of Optimization.* John Wiley & Sons, 2nd ed., 1987.

[75] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 82–96, 2002.

[76] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision*, vol. 40, pp. 24–47, Oct. 2000.

[77] J. Sun, N.-N. Zhen, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 787–800, July 2003.

[78] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[79] W. Kim, J. Park, and K. M. Lee, "Stereo matching using population-based MCMC," *International Journal of Computer Vision*, vol. 83, pp. 195–209, June 2009.

[80] I. Sutskever and T. Tieleman, "On the convergence properties of contrastive divergence," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 789–795, 2010.

[81] H. Chen and A. Murray, "Continuous restricted Boltzmann machine with an implementable training algorithm," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 150, no. 3, pp. 153–158, 2003.

[82] U. Schmidt and S. Roth, "Learning rotation-aware features: From invariant priors to equivariant descriptors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2050–2057, 2012.

[83] I. T. Jolliffe, *Principal Component Analysis.* New York, New York: Springer, 2nd ed., 2002.

[84] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: Computation in Neural Systems*, vol. 7, pp. 333–339, May 1996.

[85] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[86] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1137–1144, 2006.

[87] I. J. Goodfellow, Q. V. Le, A. M. Saxe, H. Lee, and A. Y. Ng, "Measuring invariances in deep networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 646–654, 2009.

[88] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223, 2011.

[89] U. Schmidt, J. Jancsary, S. Nowozin, S. Roth, and C. Rother, "Cascades of regression tree fields for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 677–689, 2016.

[90] K. Schelten, S. Nowozin, J. Jancsary, C. Rother, and S. Roth, "Interleaved regression tree field cascades for blind image deconvolution," in *Proc. IEEE Wint. Conf. on Appl. of Comp. Vis.*, pp. 494–501, 2015.

[91] G. van Tulder and M. de Bruijne, "Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted Boltzmann machines," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1262–1272, 2016.

[92] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solutions of ill-posed problems in computational vision," *Journal of the American Statistical Association*, vol. 82, pp. 76–89, Mar. 1987.

[93] A. Blake and A. Zisserman, *Visual Reconstruction.* MIT Press, 1987.

[94] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, Nov. 1992.

[95] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger, “Robust anisotropic diffusion,” *IEEE Transactions on Image Processing*, vol. 7, pp. 421–432, Mar. 1998.

[96] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[97] C. Schnörr, “Unique reconstruction of piecewise-smooth images by minimizing strictly convex nonquadratic functionals,” *Journal of Mathematical Imaging and Vision*, vol. 4, pp. 189–198, May 1994.

[98] R. Szeliski, “Bayesian modeling of uncertainty in low-level vision,” *International Journal of Computer Vision*, vol. 5, pp. 271–301, Dec. 1990.

[99] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, “Quicksilver: Fast predictive image registration – A deep learning approach,” *NeuroImage*, vol. 158, pp. 378–396, 2017.

[100] H. Li and Y. Fan, “Non-rigid image registration using self-supervised fully convolutional networks without training data,” in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1075–1078, 2018.

[101] D. P. Kingma and M. Welling, “An introduction to variational autoencoders,” *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[102] K. G. G. Samuel and M. F. Tappen, “Learning optimized MAP estimates in continuously-valued MRF models,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 477–484, 2009.

[103] Y. Weiss, “Sampling from natural image priors without quantization.” Personal communication., Nov. 2004.

[104] A. Gelman and D. B. Rubin, “Inference from iterative simulation using multiple sequences,” *Statistical Science*, vol. 7, pp. 457–472, Nov. 1992.

[105] 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,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 416–423, 2001.

[106] T. Tieleman, “Training restricted Boltzmann machines using approximations to the likelihood gradient,” in *International Conference on Machine Learning (ICML)*, pp. 1064–1071, 2008.

[107] M. Ranzato, V. Mnih, and G. E. Hinton, “Generating more realistic images using gated MRF’s,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2002–2010, 2010.

[108] D. L. Ruderman, “The statistics of natural images,” *Network: Computation in Neural Systems*, vol. 5, pp. 517–548, Nov. 1994.

[109] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, “On advances in statistical modeling of natural images,” *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, Jan. 2003.

[110] G. Hinton, “A practical guide to training restricted Boltzmann machines,” Tech. Rep. UTML TR 2010–003, University of Toronto, 2010.

[111] T. Hao, T. Raiko, A. Ilin, and J. Karhunen, “Gated Boltzmann machine in texture modeling,” in *International Conference on Artificial Neural Networks (ICANN)*, vol. 7553 of *Lecture Notes in Computer Science*, pp. 124–131, Springer, 2012.

[112] A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in *ACM SIGGRAPH*, pp. 341–346, 2001.

[113] A. A. Efros and T. K. Leung, “Texture synthesis by non-parametric sampling,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1033–1038, 1999.

[114] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[115] M. Nikolova, “Model distortions in Bayesian MAP reconstruction,” *AIMS Journal on Inverse Problems and Imaging*, vol. 1, no. 2, pp. 399–422, 2007.

[116] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 60–65, 2005.

[117] B. Goyala, A. Dograb, S. Agrawala, B. Sohic, and A. Sharma, “Image denoising review: From classical to state-of-the-art approaches,” *Information Fusion*, vol. 55, pp. 220–244, 2020.

[118] G. Papandreou and A. L. Yuille, “Gaussian sampling by local perturbations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1858–1866, 2010.

[119] C. Rasmussen, *minimize.m – Conjugate Gradient Minimization.* 2006. www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize.

[120] A. Barbu, “Learning real-time MRF inference for image denoising,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1574–1581, 2009.

[121] F. Luisier, T. Blu, and M. Unser, “Image denoising in mixed Poisson-Gaussian noise,” *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 696–708, 2011.

[122] M. Mäkitalo and A. Foi, “Optimal inversion of the generalized Anscombe transformation for Poisson-Gaussian noise,” *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 91–103, 2013.

[123] S. A. Haider, A. Cameron, P. Siva, D. Lui, M. J. Shafiee, A. Boroomand, N. Haider, and A. Wong, “Fluorescence microscopy image noise reduction using a stochastically-connected random field model,” *Scientific Reports*, vol. 6, no. 20640, 2016.

[124] A. Jezierska, E. Chouzenoux, J.-C. Pesquet, and H. Talbot, “A primal-dual proximal splitting approach for restoring data corrupted with Poisson-Gaussian noise,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1085–1088, 2012.

[125] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, “A convex approach for image restoration with exact Poisson-Gaussian likelihood,” *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2662–2682, 2015.

[126] J. li, Z. Shen, R. Yin, and X. Zhang, “A reweighted L2 method for image restoration with Poisson and mixed Poisson-Gaussian noise,” *Inverse Problems and Imaging*, vol. 9, no. 3, pp. 875–894, 2015.

[127] B. Bajic, J. Lindblad, and N. Sladoje, “Blind restoration of images degraded with mixed Poisson-Gaussian noise with application in transmission electron microscopy,” in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 123–127, 2016.

[128] J. Li, F. Luisier, and T. Blu, “PURE-LET image deconvolution,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 92–105, 2018.

[129] S. Gazagnes, E. Soubies, and L. Blanc-Feraud, “High density molecule localization for super-resolution microscopy using CEL0 based sparse approximation,” in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 28–31, 2017.

[130] U. Schmidt, K. Schelten, and S. Roth, “Bayesian deblurring with integrated noise estimation,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2625–2632, 2011.

[131] D. Chen, “Regularized generalized inverse accelerating linearized alternating minimization algorithm for frame-based Poissonian image deblurring,” *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 716–739, 2014.

[132] A. Jezierska, J.-C. Pesquet, H. Talbot, and C. Chaux, “Iterative Poisson-Gaussian noise parametric estimation for blind image denoising,” in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 2819–2823, 2014.

[133] M. Mäkitalo and A. Foi, “Noise parameter mismatch in variance stabilization, with an application to Poisson-Gaussian noise estimation,” *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5348–5359, 2014.

[134] S. Eck, S. Wörz, K. Müller-Ott, M. Hahn, A. Biesdorf, G. Schotta, K. Rippe, and K. Rohr, “A spherical harmonics intensity model for 3D segmentation and 3D shape analysis of heterochromatin foci,” *Medical Image Analysis*, vol. 32, pp. 18–31, 2016.

[135] B. Rieger, C. Molenaar, R. Dirks, and L. V. Vliet, “Alignment of the cell nucleus from labeled proteins only for 4D in vivo imaging,” *Microscopy Research and Technique*, vol. 64, pp. 142–150, 2004.

[136] A. P. Goobic, J. Tang, and S. T. Acton, “Image stabilization and registration for tracking cells in the microvasculature,” *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 2, pp. 287–299, 2005.

[137] C. A. Wilson and J. A. Theriot, "A correlation-based approach to calculate rotation and translation of moving cells," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1939–1951, 2006.

[138] P. Matula, P. Matula, M. Kozubek, and V. Dvorak, "Fast point-based 3-D alignment of live cells," *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2388–2396, 2006.

[139] O. Dzyubachyk, J. Essers, W. A. van Cappellen, C. Baldeyron, A. Inagaki, W. J. Niessen, and E. Meijering, "Automated analysis of time-lapse fluorescence microscopy images: from live cell images to intracellular foci," *Bioinformatics*, vol. 26, no. 19, pp. 2424–2430, 2010.

[140] S. Ozere, P. Bouthemy, F. Spindler, P. Paul-Gilloteaux, and C. Kervrann, "Robust parametric stabilization of moving cells with intensity correction in light microscopy image sequences," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 468–471, 2013.

[141] J. Mattes, J. Nawroth, P. Boukamp, R. Eils, and K. M. Greulich-Bode, "Analyzing motion and deformation of the cell nucleus for studying co-localizations of nuclear structures," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1044–1047, 2006.

[142] S. Yang, D. Kohler, K. Teller, T. Cremer, P. L. Baccon, E. Heard, R. Eils, and K. Rohr, "Nonrigid registration of 3-D multichannel microscopy images of cell nuclei," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 493 – 499, 2008.

[143] J. De Vylder, W. H. De Vos, E. M. Manders, and W. Philips, "2D mapping of strongly deformable cell nuclei based on contour matching," *Cytometry Part A*, vol. 79A, pp. 580–588, 2011.

[144] D. V. Sorokin, I. Peterlik, M. Tektonidis, K. Rohr, and P. Matula, "Non-rigid contour-based registration of cell nuclei in 2-D live cell microscopy images using a dynamic elasticity model," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 173–184, 2018.

[145] I. Kim, Y. M. Chen, D. L. Spector, R. Eils, and K. Rohr, "Nonrigid registration of 2-D and 3-D dynamic cell nuclei images for improved classification of subcellular particle motion," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 1011–1022, 2011.

[146] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 674–679, 1981.

[147] M. Tektonidis, I. Kim, Y. M. Chen, R. Eils, D. L. Spector, and K. Rohr, "Non-rigid multi-frame registration of cell nuclei in live cell fluorescence microscopy image data," *Medical Image Analysis*, vol. 19, pp. 1–14, 2015.

[148] M. Tektonidis and K. Rohr, "Diffeomorphic multi-frame non-rigid registration of cell nuclei in 2D and 3D live cell images," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1405–1417, 2017.

[149] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: A survey," *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, 2015.

[150] D. Sun, S. Roth, J. P. Lewis, and M. J. Black, "Learning optical flow," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 83–97, 2008.

[151] C. Metz, S. Klein, M. Schaap, T. van Walsum, and W. Niessen, "Nonrigid registration of dynamic medical imaging data using nD+t B-splines and a groupwise optimization approach," *Medical Image Analysis*, vol. 15, pp. 238–249, 2011.

[152] M. Yigitsoy, C. Wachinger, and N. Navab, "Temporal groupwise registration for motion modeling," in *Proceedings of International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 648–659, 2011.

[153] S. Durrleman, X. Pennec, A. Trouvé, J. Braga, G. Gerig, and N. Ayache, "Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 22–59, 2013.

[154] W. Shi, M. Jantsch, P. Aljabar, L. Pizarro, W. Bai, H. Wang, D. O'Regan, X. Zhuang, and D. Rueckert, "Temporal sparse free-form deformations," *Medical Image Analysis*, vol. 17, pp. 779–789, 2013.

[155] M. Polfliet, S. Klein, W. Huizinga, M. M. Paulides, W. J. Niessen, and J. Vandemeulebroucke, "Intrasubject multimodal groupwise registration with the conditional template entropy," *Medical Image Analysis*, vol. 46, pp. 15–25, 2018.

[156] Y. Tseng, J. S. H. Lee, T. P. Kole, I. Jiang, and D. Wirtz, "Micro-organization and visco-elasticity of the interphase nucleus revealed by particle nanotracking," *J. Cell Sci.*, vol. 117, pp. 2159–2167, 2004.

[157] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, pp. 75–104, Jan. 1996.

[158] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2439, 2010.

[159] J. Boulanger, C. Kervrann, P. Bouthemy, P. Elbau, J.-B. Sibarita, and J. Salamero, "Patch-based nonlocal functional for denoising fluorescence microscopy image sequences," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 442–454, 2010.

[160] J. O. Irwin, "The frequency distribution of the difference between two independent variates following the same Poisson distribution," *Journal of the Royal Statistical Society*, vol. 100, no. 3, pp. 415–416, 1937.

[161] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. V. der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, 1994.

[162] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *International Conference on Machine Learning (ICML)*, pp. 609–616, 2009.

[163] H. Zhang, Y. Zhang, H. Li, and T. S. Huang., "Generative Bayesian image super resolution with natural image prior," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4054–4067, 2012.

[164] D. V. Sorokin, J. Suchánková, E. Bartova, and P. Matula, "Visualizing stable features in live cell nucleus for evaluation of the cell global motion compensation," *Folia biologica*, vol. 60, pp. 45–49, 2014.

[165] V. Foltankova, P. Matula, D. V. Sorokin, S. Kozubek, and E. Bartova, "Hybrid detectors improved time-lapse confocal microscopy of PML and 53BP1 nuclear body colocalization in DNA lesions," *Microscopy and Microanalysis*, vol. 19, no. 2, pp. 360–369, 2013.

[166] M. H. Sadd, *Elasticity: Theory, Applications, and Numerics.* Elsevier, third ed., 2014.

[167] V. O. Chagin, C. S. Casas-Delucchi, M. Reinhart, L. Schermelleh, Y. Markaki, A. Maiser, J. J. Bolius, A. Bensimon, M. Fillies, P. Domaing, Y. M. Rozanov, H. Leonhardt, and M. C. Cardoso, "4D visualization of replication foci in mammalian cells corresponding to individual replicons," *Nature Communications*, vol. 7, no. 11231, 2016.