ShapeOdds: Variational Bayesian Learning of Generative Shape Models

Shireen Elhabian and Ross Whitaker
Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA
{shireen,whitaker}@sci.utah.edu

Abstract

Shape models provide a compact parameterization of a class of shapes, and have been shown to be important to a variety of vision problems, including object detection, tracking, and image segmentation. Learning generative shape models from grid-structured representations, aka silhouettes, is usually hindered by (1) data likelihoods with intractable marginals and posteriors, (2) high-dimensional shape spaces with limited training samples (and the associated risk of overfitting), and (3) estimation of hyperparameters relating to model complexity that often entails computationally expensive grid searches. In this paper, we propose a Bayesian treatment that relies on direct probabilistic formulation for learning generative shape models in the silhouettes space. We propose a variational approach for learning a latent variable model in which we make use of, and extend, recent works on variational bounds of logistic-Gaussian integrals to circumvent intractable marginals and posteriors. Spatial coherency and sparsity priors are also incorporated to lend stability to the optimization problem by regularizing the solution space while avoiding overfitting in this high-dimensional, low-sample-size scenario. We deploy a type-II maximum likelihood estimate of the model hyperparameters to avoid grid searches. We demonstrate that the proposed model generates realistic samples, generalizes to unseen examples, and is able to handle missing regions and/or background clutter, while comparing favorably with recent, neural-network-based approaches.

1. Introduction

Shape modeling deals with learning statistical properties of a shape population. This is typically accomplished by estimating a probability distribution from a set of *i.i.d.* training samples drawn from the true, unknown distribution, treating individual data points as samples in a high-dimensional *shape space*. Shape models are an enabling technology for a variety of vision and imaging applications, such as feature localization [1–3], object recognition [4, 5], pose estimation [6, 7], object detection [8–10], image segmentation [11–15], tracking [16–18], object reconstruction [19, 20], animation [21–23], and shape synthesis [24]. Image segmentation, for instance, often benefits from incorporating expectations of particular classes of objects (*e.g.,* birds, animals, faces), in the form of *shape priors* to guide/constrain the segmentation process [25].

This paper addresses the problem of learning generative shape models from *grid-structured* representations in which data points in the shape space are represented as *binary* functions defined over a discrete image domain, *i.e.,* silhouettes. There is a rich history of work on learning shape statistics from silhouettes in which the main distinction is capturing *local* (*i.e.,* low-level) versus *global* (*i.e.,* high-level) correlations. *Local* structure interactions between pixels typically capture generic properties, *e.g.,* smoothness and continuity (often via Markov random fields – MRFs, *e.g.,* [26–29]). Here we focus on *global* models designed to capture complex high-level shape structure (*e.g.,* facial parts, horse legs, vehicle wheels), which may also be complemented by low-level spatial priors. Learning global shape models in the silhouettes space is challenging because the binary variables entail *non-Gaussian data likelihoods*, which often lead to intractable marginals and posteriors. *High-dimensional shape space* with limited training samples further increases the tendency to overfit. Additionally, the *hyperparameters* associated with model complexity often result in computationally expensive discrete searches.

Here we rely on a *generative* model, where a silhouette is a realization of a spatially coherent field of Bernoulli random variables – characterized by a *parameter map* – defined on the image domain. This probabilistic representation of shape yields globally optimal solutions for certain problems, *e.g.,* segmentation and tracking, due to the convexity of the parameter maps space [30]. Learning a probability distribution over the silhouettes space amounts to estimating the parameter map of a silhouette. Because the shape space is a *unit hypercube*, such a learning task does not benefit from a *vector space* structure. *LogOdds*, as an alternative to probabilities, places parameter maps in a vector space where addition and scalar multiplications have probabilistic interpretations [31]. Consequently, most existing approaches have resorted to modeling shape variability *indirectly* on a space of some predefined implicit function, including signed distance maps (SDMs) and Gaussian

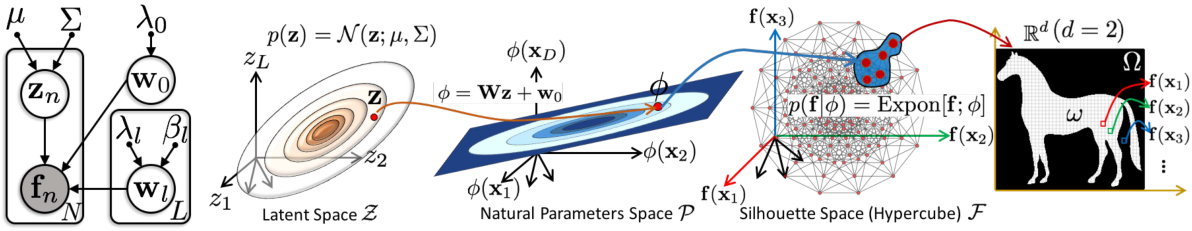

Figure 1. ShapeOdds: shape-generating process

smoothed silhouettes [31], whose zero level set reflects the shape's boundary. However, such representations typically do not have a statistical foundation, and therefore do not benefit from optimal estimation strategies.

Dimensionality reduction techniques (*e.g.,* linear [30–34] and nonlinear [35–39]) are often applied to those intermediate representations to *parameterize* the underlying shape variation. *Nonparameteric* approaches (*e.g.,* [40–45]), on the other hand, avoid making assumptions about the form of the underlying density function by explicitly relying on the available training samples, hence promoting a local influence of individual samples being encoded in the estimator kernel width, but creating challenges for the robustness and generalization of the fitted nonparametric model with inherent tradeoff between the estimate bias and variance [46, 47]. Nonetheless, these modeling approaches fail to define a proper generative model, which is advantageous in handling unbiased noise (*e.g.,* missing regions and/or background clutter) [48–50]. They also do not readily lead to hierarchical/layered architectures (*e.g.,* deep learning), which promise to capture different levels of representation abstraction [51, 52]. Moreover, these approaches rely on maximum-likelihood estimation of the principal subspace and thus ignore uncertainties associated with the estimated low-dimensional representation [50, 51, 53]. Welling *et al.* [54] have argued that non-Bayesian approaches "optimize a questionable objective", and are prone to overfitting in high-dimensional spaces with a small number of training samples—a very common circumstance in training from silhouettes. Another adverse consequence of non-Bayesian point estimates is the sensitivity to regularization parameters, requiring careful discrete searches [55, 56].

Recently, *stochastic neural nets*, in particular restricted Boltzmann machines (RBMs) [57, 58] and their deep/layered architectures [59–62], have offered more flexible *undirected* models for binary inputs without relying on any intermediate implicit representation. Efficient maximum-likelihood learning and inference algorithms are available via the omission of lateral connections in the same layer [52, 63]. Motivated by the pragmatic development of tractable algorithms [64], these models mostly rely on a generic type of deep network structure that does not inject any domain knowledge of the modeling problem at hand; *i.e.,* there is no attempt to model a particular generative process. Consequently, an exponential number of hidden units and a large amount of training data are typically required to approximate an arbitrary binary distribution [65]. The *shape Boltzmann machine* (ShapeBM) [60] was recently proposed to alleviate the need for large training data by heuristically partitioning the shape space using axis-aligned overlapping boxes combined with a weight-sharing scheme. Nonetheless, ShapeBM inherits the lack of a specific generating process. Further, the unsupervised, data-driven, learning scheme of such networks typically comes with high demands on practitioner expertise and computational costs to determine the ideal network architecture and associated hyperparameters for a particular data set.

In this paper, we propose a method to learn the underlying variability of a silhouettes population that is derived from a generative model, thus defining a propability density function *directly* over the silhouettes space while not relying on heuristics for the computation of parameter maps. We present a *Bayesian* treatment of a *latent variable model* – *ShapeOdds* – for generative shape modeling in which an *observed* high-dimensional shape space is assumed to be generated from an underlying *latent* low-dimensional process. We extend recent works in the machine learning literature on variational bounds of logistic-Gaussian integrals designed to circumvent the intractable marginal likelihood and latent posterior leading to *deterministic* learning. The proposed variational formulation further reduces the sensitivity to hyperparameters by modeling posterior uncertainties [66]. This extension to computer vision applications makes use of general-purpose priors [52]—in particular, *spatial coherency* and *sparsity*—to untangle the underlying factors of shape variation revealed by the data. ShapeOdds is further equipped with a data-driven hyperprior that automatically estimates model hyperparameters – with closed-form re-estimation expressions – without the need for discrete searches and cross validation. In contrast to RBMs and their deep variants, ShapeOdds benefits from the *explaining-away* property of *directed* probabilistic models, yielding parsimonious posteriors in which latent variables compete and collaborate to explain the observed shape instance [52]. Such a property could be achieved with undirected models in the presence of lateral connections in the observed and hidden layers [52] at the expense of not benefiting from efficient *sampling-based* training algorithms (*e.g.,* [67–70]) that are associated with RBMs and increase the parameter space dimensionality, thereby exac-

erbating the risk of overfitting. Experiments demonstrate that ShapeOdds is able to generate realistic samples, generalize to unseen data, and handle unbiased noise. ShapeOdds paves the way to a rich class of shape models with which deep architectures of latent models can be introduced to capture more complex shape distributions.

2. Latent Gaussian Model for Silhouettes

Consider a raster defined over a spatial domain $\Omega \subset \mathbb{R}^d$ (here $d = 2$) containing D pixels. The foreground object $\omega \subset \Omega$ is represented by a silhouette $\mathbf{f} \in \{0,1\}^D$, where $\mathbf{f}(\mathbf{x}) = 1$, iff $\mathbf{x} \in \omega \ \ \forall\, \mathbf{x} \in \Omega$. In a *generative* sense, $\mathbf{f}$ is a realization of a *spatially correlated* field of D *Bernoulli* random variables defined on Ω with a pixelwise *parameter* $\mathbf{q}(\mathbf{x}) \in [0,1]$ where $\mathbf{q}(\mathbf{x}) = p(\mathbf{x} \in \omega)$. Spatial regularity on the silhouettes, typically modeled as MRFs, help describe local correlations between nearby pixels. The Bernoulli likelihood has an equivalent form in terms of the *exponential family distributions* that is parameterized by a field of real values $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}$, known as *natural parameters*, where $\boldsymbol{\phi}(\mathbf{x}) = \text{logit}[\mathbf{q}(\mathbf{x})]$ with $\mathbf{q}(\mathbf{x})$ being the first moment of this form and hence denoted a *expectation* parameters. The merit in considering such an equivalence is casting any parameter estimation problem as an *unconstrained* optimization in the natural parameters space. Hence, the *generative model* of a silhouette includes a pixelwise Bernoulii likelihood and the MRF spatial prior.

$$p(\mathbf{f}|\boldsymbol{\phi}) = \left\{ \prod_{\mathbf{x}\in\Omega} p(\mathbf{f}(\mathbf{x})|\boldsymbol{\phi}(\mathbf{x})) \right\} \times \frac{1}{Z} \exp\left(-\frac{1}{T} U(\mathbf{f}) \right) \quad (1)$$

$$\text{with} \quad p(\mathbf{f}(\mathbf{x})|\boldsymbol{\phi}(\mathbf{x})) = \exp\left[\mathbf{f}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}) - \text{llp}[\boldsymbol{\phi}(\mathbf{x})]\right] \quad (2)$$

where $\text{llp}[\phi] = \log\left(1 + e^{\phi}\right)$ is the *logistic-log-partition* function, $U(\mathbf{f})$ are clique potentials that favor spatially coherent silhouettes, Z is a Gibbs distribution normalization constant, and T is its temperature [71].

Consider an unknown *shape distribution* $p(\mathbf{f})$ in the silhouettes space $\mathcal{F}$, of which we have only an ensemble of silhouettes $\mathrm{F} = \{\mathbf{f}_n\}_{n=1}^N \subset \mathcal{F}$. In latent variable formalism, this distribution is governed by a low-dimensional *shape-generating process* of L independent latent variables $\mathbf{z} \in \mathbb{R}^L$ where $L \ll D$. Here we consider a class of latent Gaussian models (LGMs) to capture correlations between observed pixels through Gaussian latent variables. In particular, a point $\mathbf{z}$ in the *latent space* $\mathcal{Z}$ is *generated* according to a Gaussian prior disitribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \Sigma)$, where $\mu \in \mathbb{R}^L$ and $\Sigma \in \mathbb{R}^{L\times L}$, which is mapped onto the natural parameters space $\mathcal{P}$ by a smooth mapping $h : \mathcal{Z} \to \mathcal{P}$. The logit function further maps $\mathcal{P}$ to the expectation parameters space $\mathcal{Q}$. A natural parameters map $\boldsymbol{\phi} \in \mathbb{R}^D$ is assumed to be confined to a linear subspace in $\mathcal{P}$ parameterized by a *factor loading* matrix $\mathbf{W} \in \mathbb{R}^{D\times L}$ and an *offset* vector $\mathbf{w}_0 \in \mathbb{R}^D$ where $\boldsymbol{\phi} = h(\mathbf{z}) = \mathbf{W}\mathbf{z} + \mathbf{w}_0$. A $\boldsymbol{\phi}-$map thus induces a distribution $p(\mathbf{f}|\boldsymbol{\phi})$ of silhouettes in $\mathcal{F}$.

The corresponding natural parameters of F are $\Phi = \{\boldsymbol{\phi}_n\}_{n=1}^N \subset \mathcal{P}$. Although they lie in a linear subspace in $\mathcal{P}$, typically they correspond to a nonlinear manifold in $\mathcal{F}$. In a high-dimensional setting, suboptimal local maxima of the log-likelihood will result in a mapping h that induces a badly twisted manifold in $\mathcal{F}$, giving rise to a multimodal posterior distribution in $\mathcal{Z}$. Penalizing highly twisted mappings is usually achieved through *regularization* [72], which from a Bayesian viewpoint requires introducing a smoothness prior on the mapping parameters $\mathbf{W}$ and $\mathbf{w}_0$ controlled by *hyperparameters*. This prior serves a *computational* purpose by lending stability to the learning process through regularizing the solution space and a *statistical* purpose, which is to avoid overfitting in this high-dimensional low-sample-size scenario [72]. We introduce a Gaussian MRF (GMRF) prior over individual loading/offset vectors $\{\mathbf{w}_l\}_{l=0}^L$ with $\mathbf{w}_l : \Omega \to \mathbb{R}^D$ where the prior on the mapping h can be factored out as $p(\mathbf{W}, \mathbf{w}_0) = \prod_{l=0}^L p(\mathbf{w}_l|\lambda_l)$. The smoothness prior over a vector $\mathbf{w}_l$ can be written as a Gibbs distribution, $p(\mathbf{w}_l|\lambda_l) \propto \exp\{-\lambda_l E(\mathbf{w}_l)\}$ where $\lambda_l > 0$ is a hyperparameter that controls the generalizability aspect of the resultant mapping. Gibbs energy $E(\mathbf{w}_l)$, hence, is chosen to favor smooth vectors by penalizing abrupt edges. We use Laplacian-square energy, *i.e.,* $E(\mathbf{w}_l) = \|\Delta\mathbf{w}_l\|_2^2$ to quantify the edges within $\mathbf{w}_l$.

The intrinsic dimensionality of the silhouettes manifold is determined by the choice of the latent dimensionality L. Nonetheless, an exhaustive grid search over this choice can become computationally intractable, especially when extending the proposed shape model to mixtures or even deep architectures of latent models. The probabilistic formulation of LGMs allows this discrete model selection to be handled within the Bayesian paradigm [73]. We make use of the sparsity-inducing automatic relevance determination (ARD) prior to further regularize the solution space via a parameterized data-driven prior distribution that effectively prunes away irrelevant factors of variations, as data reveals [74]. We introduce an ARD prior on the loading vectors $\{\mathbf{w}_l\}_{l=1}^L$ with L set to the maximum allowed dimensionality, *i.e.,* $L = N - 1$ and $N \ll D$. ARD is a zero-mean isotropic Gaussian prior parameterized by $\beta_l \in \mathbb{R}_{>0}$ such that $p(\mathbf{w}_l|\beta_l) = \mathcal{N}(\mathbf{w}_l; \mathbf{0}_D, \beta_l^{-1}\mathbf{I}_D)$ where $\mathbf{0}_D$ and $\mathbf{I}_D$ are the zero vector and identity matrix in $\mathbb{R}^D$, respectively. During the learning process, $\beta_l \to \infty$ for irrelevant factors to remove the unnecessary complexity of the resulting model [75]. These sparsity and smoothness priors impose a special structure on the loading matrix $\mathbf{W}$ that enables model identifiability [76], an inherent property of LGMs. These priors do not necessarily ensure unique model parameters, but they encourage interpretable solutions [77].

ShapeOdds thus refers to the shape-generating process with model parameters $\Theta = \{\mu, \Sigma, \mathbf{W}, \mathbf{w}_0\}$ and priors hyperparameters $\Psi = \{\boldsymbol{\lambda}, \boldsymbol{\beta}\}$ where $\boldsymbol{\lambda} \in \mathbb{R}_{>0}^{L+1}$ and $\boldsymbol{\beta} \in \mathbb{R}_{>0}^{L}$. ShapeOdds defines a data-driven mapping of silhouettes to

the $\mathcal{P}-$space with a vector space structure. The underlying generative process can be defined as follows:

$$\mathbf{z}_n \sim \mathcal{N}(\mu, \Sigma), \quad \mathbf{f}_n|\mathbf{z}_n, \Theta \sim \text{Expon}[\boldsymbol{\phi}_n]\,\text{Mrf}[\nu] \tag{3}$$

$$\text{Expon}[\boldsymbol{\phi}_n] \doteq \prod\nolimits_{\mathbf{x}\in\Omega} \text{Expon}[\boldsymbol{\phi}_n(\mathbf{x})] \tag{4}$$

$$\boldsymbol{\phi}_n = \mathbf{W}\mathbf{z}_n + \mathbf{w}_0, \quad \mathbf{w}_0|\lambda_0 \sim \text{GMrf}[\lambda_0] \tag{5}$$

$$\mathbf{w}_l|\lambda_l, \beta_l \sim \text{GMrf}[\lambda_l]\,\text{Ard}[\beta_l] \tag{6}$$

with $\text{GMrf}[\lambda] \doteq \mathcal{N}(\mathbf{0}_D, \lambda^{-1}\mathbf{S})$ where $\mathbf{S}$ is the structure matrix containing the stencil of the negative bi-Laplacian operator; the first variation of the Laplacian-square energy $E(\mathbf{w}_l)$. The MRF prior in (3), whose hyperparameter $\nu > 0$ is related to the temperature in (1), reflects the spatial regularity of the given silhouettes. Note that the choice of ν does not affect the model learning process. Eq (4) is due to the axiom of local/conditional independence [78, 79], *i.e.,* the observed variables are conditionally independent given the latent variables, where $\text{Expon}[\boldsymbol{\phi}(\mathbf{x})]$ is given by (2). The graphical model of ShapeOdds is given in Figure 1.

3. Variational Learning of ShapeOdds

The potential gain of ShapeOdds is twofold: (1) autodetection of the latent dimensionality to avoid grid searches and (2) autoregularization of the solution space to promote model generalizability. Nonetheless, this treatment comes with an extra degree of intractability; the Gaussian prior is not conjugate to the Bernoulli likelihood, resulting in an intractable logistic-Gaussian integral. A maximum likelihood point estimate for the posterior ignores associated uncertainties, resulting in overfitting even with careful regularization [54]. Instead, we propose a *variational* approximation to the marginal likelihood to derive a tractable and deterministic expectation-maximization (EM) algorithm for model learning while preserving posterior uncertainty.

The *marginal likelihood* in $\mathcal{F}-$space can be obtained by *integrating* the joint density $p(\mathbf{f}, \mathbf{z})$ in the product space $\mathcal{F} \times \mathcal{Z}$ over the latent space $\mathcal{Z}$. To obtain a tractable integral, we restrict the posterior distribution $p(\mathbf{z}_n|\mathbf{f}_n, \Theta)$ to a tractable family. Let $q(\mathbf{z}_n|\boldsymbol{\gamma}_n) = \mathcal{N}(\mathbf{z}_n|\mathbf{m}_n, \mathbf{V}_n)$, where $\boldsymbol{\gamma}_n = \{\mathbf{m}_n, \mathbf{V}_n\}$, be a Gaussian approximate to the posterior with mean $\mathbf{m}_n \in \mathbb{R}^L$ and covariance $\mathbf{V}_n \in \mathbb{R}^{L\times L}$. The *lower bound* to the log-marginal likelihood can be obtained by dividing and multiplying by the posterior approximate and then applying the Jensen inequality.

$$\mathcal{L}(\Theta) \geq \underline{\mathcal{L}}^J(\Theta, \boldsymbol{\gamma}) = \sum\nolimits_{n=1}^{N} \mathbb{E}_{q(\mathbf{z}_n|\boldsymbol{\gamma}_n)}\left[\log \frac{p(\mathbf{z}_n|\Theta)}{q(\mathbf{z}_n|\boldsymbol{\gamma}_n)}\right] + \mathbb{E}_{q(\mathbf{z}_n|\boldsymbol{\gamma}_n)}\left[\log p(\mathbf{f}_n|\mathbf{z}_n, \Theta)\right] \tag{7}$$

The *first expectation* term in (7) is the negative Kullback-Leibler (KL) divergence that pushes the variational posterior to the Gaussian prior. Its closed form is given by

$$-\,\text{KL}_n[q\|p] = {}^1\!/_2\left\{\log|\mathbf{V}_n\Sigma^{-1}| - \text{Tr}[\mathbf{V}_n\Sigma^{-1}] - (\mathbf{m}_n - \mu)^T\Sigma^{-1}(\mathbf{m}_n - \mu) + L\right\} \tag{8}$$

Using the mapping $h(\mathbf{z})$ and the conditional independence in (4), the Gaussian approximate posterior $q(\mathbf{z}_n|\boldsymbol{\gamma}_n)$ in $\mathcal{Z}$ induces a per-pixel Gaussian posterior $q(\boldsymbol{\phi}_n(\mathbf{x})|\widetilde{\boldsymbol{\gamma}}_n^{\mathbf{x}})$ in $\mathcal{P}$ with $\widetilde{\boldsymbol{\gamma}}_n^{\mathbf{x}} = \{\widetilde{\mathbf{m}}_n^{\mathbf{x}}, \widetilde{\mathbf{V}}_n^{\mathbf{x}}\}$ where $\mathbf{w}_0^{\mathbf{x}} \in \mathbb{R}$ and $\mathbf{W}^{\mathbf{x}} \in \mathbb{R}^L$.

$$\widetilde{\mathbf{m}}_n^{\mathbf{x}} = \mathbf{W}^{\mathbf{x}}\mathbf{m}_n + \mathbf{w}_0^{\mathbf{x}}, \quad \widetilde{\mathbf{V}}_n^{\mathbf{x}} = \mathbf{W}^{\mathbf{x}}\mathbf{V}_n(\mathbf{W}^{\mathbf{x}})^T \tag{9}$$

Note that the spatial coherency is still promoted through the GMRF prior on the offset and loading vectors. Using the exponential form of the Bernoulli likelihood in (2), the *second expectation* term in (7) can be expressed in $\mathcal{P}$ as

$$\begin{aligned}&\sum\nolimits_{\mathbf{x}\in\Omega} \mathbb{E}_{q(\boldsymbol{\phi}_n(\mathbf{x})|\widetilde{\boldsymbol{\gamma}}_n^{\mathbf{x}})}\left[\log p(\mathbf{f}_n(\mathbf{x})|\boldsymbol{\phi}_n(\mathbf{x}))\right] \\ &= \sum\nolimits_{\mathbf{x}\in\Omega}\left\{\mathbf{f}_n(\mathbf{x})\widetilde{\mathbf{m}}_n^{\mathbf{x}} - \mathbb{E}_{q(\boldsymbol{\phi}_n(\mathbf{x})|\widetilde{\boldsymbol{\gamma}}_n^{\mathbf{x}})}\left[\text{llp}[\boldsymbol{\phi}_n(\mathbf{x})]\right]\right\} \quad (10)\\ &\geq \sum\nolimits_{\mathbf{x}\in\Omega} \underline{\mathcal{B}}_n(\mathbf{x}) := \left\{\mathbf{f}_n(\mathbf{x})\widetilde{\mathbf{m}}_n^{\mathbf{x}} - \overline{\mathcal{B}}(\widetilde{\boldsymbol{\gamma}}_n^{\mathbf{x}}, \boldsymbol{\alpha}_n^{\mathbf{x}})\right\} \quad (11)\end{aligned}$$

Eq (10) is intractable due to the llp function and can be lower-bounded in $\mathcal{P}-$space by defining an upper bound $\overline{\mathcal{B}}$ for the expectation of the llp function with *local*, *i.e.,* per-pixel, variational parameters $\boldsymbol{\alpha}_n^{\mathbf{x}}$. The new bound reads as

$$\underline{\mathcal{L}}(\Theta, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \sum\nolimits_{n=1}^{N} \underbrace{\left\{-\,\text{KL}_n[q\|p] + \sum\nolimits_{\mathbf{x}\in\Omega} \underline{\mathcal{B}}_n(\mathbf{x})\right\}}_{\underline{\mathcal{L}}_n(\Theta, \boldsymbol{\gamma}_n, \boldsymbol{\alpha}_n)} \tag{12}$$

To avoid the recomputation of per-pixel/per-sample $\boldsymbol{\alpha}_n^{\mathbf{x}}$, we use a *fixed* piecewise quadratic upper bound for the llp function recently proposed in [80] as a proven tight bound compared to other quadratic bounds, *e.g.,* [81, 82], where $\boldsymbol{\alpha}_n^{\mathbf{x}} = \boldsymbol{\alpha}\ \forall\ n, \mathbf{x}$. Consider a quadratic bound with $R-$intervals defined by $R+1$ control points $\tau_0 = -\infty < \tau_1 < \ldots < \tau_R = +\infty$, whose parameters $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_r\}_{r=1}^R$ and $\boldsymbol{\alpha}_r = [a_r, b_r, c_r]$ are estimated via a minimax optimization to ensure a tight bound [80] (here we use $R = 20$, where the error was shown to approach zero [80]). The upper bound $\overline{\mathcal{B}}$ can thus be expressed in terms of truncated Gaussian moments, due to the approximate Gaussian posterior, whose closed-form expressions along with their gradients are available [80].

$$\overline{\mathcal{B}}(\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\alpha}) = \sum_{r=1}^{R} \int_{\tau_{r-1}}^{\tau_r} \mathcal{N}(\phi; \widetilde{\mathbf{m}}, \widetilde{\mathbf{V}})\left[a_r\phi^2 + b_r\phi + c_r\right] d\phi \tag{13}$$

We propose a variational EM algorithm that optimizes the rigorous lower bound defined in (12) using the fixed upper bound in (13). The E-step in (15) optimizes the variational posterior means and covariances at an iteration i given the current guess of model parameters $\Theta^{(i-1)}$. The M-step chooses the next guess of $\Theta^{(i)}$ to maximize the *regularized* variational bound in (19). Iterating between these two steps involves concave optimizations due to the concavity of the lower bound in (12) [80, 83] and the semi-positive definiteness of the bi-Laplacian operator, for which we can use gradient-based optimization (see Algorithm 1 for gradient expressions). The maximum-a-posteriori (MAP) objective of the offset and loading vectors, after removing constant terms, can be written as

$$\begin{aligned}\mathcal{E}(\mathbf{W}, \mathbf{w}_0|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \Psi) &= -\underline{\mathcal{L}}(\Theta, \boldsymbol{\gamma}, \boldsymbol{\alpha}) + ({}^{\lambda_0}\!/_2)\,\mathbf{w}_0^T\mathbf{S}\mathbf{w}_0 \\ &+ \sum\nolimits_{l=1}^{L}\left\{({}^{\lambda_l}\!/_2)\,\mathbf{w}_l^T\mathbf{S}\mathbf{w}_l + ({}^{\beta_l}\!/_2)\,\mathbf{w}_l^T\mathbf{w}_l\right\}\end{aligned} \tag{14}$$

The first variation of (14) w.r.t. the offset and loading vectors reads as in (23) and (24): the vectors $\mathbf{g}_n^{\mathbf{m}}$ and $\mathbf{G}_n^{\mathbf{V}}$ are bound gradients in (18), $\odot$ refers to a Hadamard product, $\mathbf{m}_{n,l}$ is the $l-$th entry of $\mathbf{m}_n$, and $\mathbf{V}_{n,l}$ is the $l-$th column of $\mathbf{V}_n$. To enable large time steps Δt while maintaining stable updates, we use a *semi-implicit scheme* with finite-forward time marching to define iterative updates for $\mathbf{w}_l$'s in (26), where spatial convolution $\otimes$ can be efficiently performed as multiplication in the Fourier domain.

Hyperparameters: To complete our Bayesian treatment, we formulate an *evidence approximation*, aka type-II maximum likelihood, in which we marginalize over Ψ. We consider a Jeffery prior to construct a hyper-hyperparameters-free noninformative hyperprior on the hyperparameters, leading to analytic integrals. The noninformative hyperpriors on λ_l and β_l are $p(\lambda_l) \propto 1/\lambda_l$ and $p(\beta_l) \propto 1/\beta_l$, respectively. The marginalization over the hyperparameters involves λ_l-integrals and β_l-integrals, each with an analytic form $\frac{\Gamma(D/2)|\mathbf{S}|^{1/2}}{\pi^{D/2}(\mathbf{w}_l^T\mathbf{S}\mathbf{w}_l)^{D/2}}$ and $\frac{\Gamma(D/2)|\mathbf{S}|^{1/2}}{\pi^{D/2}(\mathbf{w}_l^T\mathbf{w}_l)^{D/2}}$, respectively, using a $\Gamma-$function integral form. For a given Ψ, the gradient of $\log p(\Theta|\Psi)$ w.r.t. $\mathbf{w}_l$ should coincide with that of the marginal $p(\Theta)$ [84]. Hence, the effective values of the hyperparameters in (30) re-estimate Ψ after each pair of E- and M-steps to compute the new Ψ given the current guess of Θ. This re-estimation mechanism can be viewed as an iterative regularization similar to [85] in which an appropriate sequence of regularizers is used to facilitate the convergence of the M-step in high-dimensional scenarios.

Robust inference: Inference contexts, *e.g.,* segmentation and tracking, involve querying the learned shape model to infer the $\phi-$map (and the corresponding $\mathbf{q}-$map) of the closest silhouette to a given corrupted one. Missing foreground regions and/or background clutter are rendered as a lack of compliance with the learned model, *i.e., outliers*, introducing an erroneous maximum likelihood estimate of the approximate posterior due to assigning higher weights to outlying pixels during the inference process. To increase the *robustness* of the estimated posterior, we deploy a functional, aka $\rho-$functions in the robust statistics field [86], of the marginal bound that is more forgiving of outlying pixels. We formulate the inference problem in an optimization context without relying on explicitly detecting the outliers' spatial support to be discarded from the inference process. The lower bound in (15) can be rewritten as a pixelwise bound as follows:

$$\underline{\mathcal{L}}(\boldsymbol{\gamma};\Theta,\boldsymbol{\alpha}) = \sum\nolimits_{\mathbf{x}\in\Omega} -\mathcal{E}_{\mathbf{x}}(\boldsymbol{\gamma}) \doteq -\,\mathrm{KL}_{\mathbf{x}}[q||p] + \underline{\mathcal{B}}_n(\mathbf{x}) \quad (36)$$

where the negative KL term in (8) can be defined in $\mathcal{P}-$space as in (37) with $\widetilde{\mu}^{\mathbf{x}} = \mathbf{w}_0^{\mathbf{x}}$ and $\widetilde{\sigma}^{\mathbf{x}} = \mathbf{W}^{\mathbf{x}}(\mathbf{W}^{\mathbf{x}})^T$.

$$-\,\mathrm{KL}_{\mathbf{x}}[q||p] = \frac{1}{2}\left\{\log\left|\frac{\widetilde{\mathbf{V}}^{\mathbf{x}}}{\widetilde{\sigma}^{\mathbf{x}}}\right| - \frac{\widetilde{\mathbf{V}}^{\mathbf{x}}}{\widetilde{\sigma}^{\mathbf{x}}} - \frac{(\widetilde{\mathbf{m}}^{\mathbf{x}} - \widetilde{\mu}^{\mathbf{x}})^2}{\widetilde{\sigma}^{\mathbf{x}}} + 1\right\} \quad (37)$$

To assign less weight to pixels poorly supported by Θ,

Algorithm 1 Variational EM for learning ShapeOdds

E-Step:

$$\gamma_n^{(i)} = \mathrm{argmax}_{\boldsymbol{\gamma}_n}\, \underline{\mathcal{L}}_n(\Theta^{(i-1)}, \boldsymbol{\gamma}_n, \boldsymbol{\alpha}) \quad \forall\, n \in 1, ..., N \quad (15)$$

$$\frac{\partial \underline{\mathcal{L}}_n}{\partial \mathbf{m}_n} = -\Sigma^{-1}(\mathbf{m}_n - \mu) + \sum\nolimits_{\mathbf{x}\in\Omega} g_n^{\mathbf{m}}(\mathbf{x})(\mathbf{W}^{\mathbf{x}})^T \quad (16)$$

$$\frac{\partial \underline{\mathcal{L}}_n}{\partial \mathbf{V}_n} = \frac{1}{2}\left[\mathbf{V}_n^{-1} - \Sigma^{-1}\right] + \sum\nolimits_{\mathbf{x}\in\Omega} G_n^{\mathbf{V}}(\mathbf{x})(\mathbf{W}^{\mathbf{x}})^T\mathbf{W}^{\mathbf{x}} \quad (17)$$

$$\text{where} \quad g_n^{\mathbf{m}}(\mathbf{x}) = \frac{\partial \underline{\mathcal{B}}_n(\mathbf{x})}{\partial \widetilde{\mathbf{m}}_n^{\mathbf{x}}}, \quad G_n^{\mathbf{V}}(\mathbf{x}) = \frac{\partial \underline{\mathcal{B}}_n(\mathbf{x})}{\partial \widetilde{\mathbf{V}}_n^{\mathbf{x}}} \quad (18)$$

M-Step:

$$\Theta^{(i)} = \arg\max_{\Theta}\left\{\sum\nolimits_{n=1}^{N} \underline{\mathcal{L}}_n(\Theta, \boldsymbol{\gamma}_n^{(i)}, \boldsymbol{\alpha})\right\} + \log p(\Theta|\Psi) \quad (19)$$

$$\text{where} \quad p(\Theta|\Psi) = p(\mathbf{w}_0|\lambda_0)\prod\nolimits_{l=1}^{L} p(\mathbf{w}_l|\lambda_l)p(\mathbf{w}_l|\beta_l) \quad (20)$$

$$\mu = \frac{1}{N}\sum\nolimits_{n=1}^{N} \mathbf{m}_n \quad (21)$$

$$\Sigma = \frac{1}{N}\sum\nolimits_{n=1}^{N}\left\{\mathbf{V}_n + (\mathbf{m}_n - \mu)(\mathbf{m}_n - \mu)^T\right\} \quad (22)$$

$$\frac{d\mathcal{E}}{d\mathbf{w}_0} = \frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_0} + \lambda_0\mathbf{S}\mathbf{w}_0, \quad \frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_0} = -\sum\nolimits_{n=1}^{N} \mathbf{g}_n^{\mathbf{m}} \quad (23)$$

$$\frac{d\mathcal{E}}{d\mathbf{w}_l} = -\left\{\frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_l} + \lambda_l\mathbf{S}\mathbf{w}_l + \beta_l\mathbf{w}_l\right\} \quad (24)$$

$$\frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_l} = \sum\nolimits_{n=1}^{N} \mathbf{g}_n^m\mathbf{m}_{n,l} + 2\mathbf{G}_n^{\mathbf{V}} \odot \mathbf{W}\mathbf{V}_{n,l} \quad (25)$$

$$\mathbf{w}_l^{(t)} = \left\{\frac{1}{1 + \Delta t\lambda_l\mathbf{S}}\right\} \otimes \left\{\mathbf{w}_l^{(t-1)} + \Delta t\left[\frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_l} - \delta(l > 0)\beta_l\mathbf{w}_l^{(t-1)}\right]\right\}, \; \forall\, l = \{0, 1, ..., L\} \quad (26)$$

H-Step:

$$p(\Theta) = \prod\nolimits_{l=0}^{L}\left\{\int_0^{\infty} p(\mathbf{w}_l|\lambda_l)p(\lambda_l)d\lambda_l\right\} \times \prod\nolimits_{l=1}^{L}\left\{\int_0^{\infty} p(\mathbf{w}_l|\beta_l)p(\beta_l)d\beta_l\right\} \quad (27)$$

$$p(\mathbf{w}_l|\lambda_l) = \frac{\lambda_l^{D/2}|\mathbf{S}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{\lambda_l}{2}\mathbf{w}_l^T\mathbf{S}\mathbf{w}_l\right\} \quad (28)$$

$$p(\mathbf{w}_l|\beta_l) = \frac{\beta_l^{D/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{\beta_l}{2}\mathbf{w}_l^T\mathbf{w}_l\right\} \quad (29)$$

$$\lambda_l = \frac{D}{\mathbf{w}_l^T\mathbf{S}\mathbf{w}_l}, \quad \beta_l = \frac{D}{\mathbf{w}_l^T\mathbf{w}_l} \quad \text{where} \quad \mathbf{w}_l^T\mathbf{S}\mathbf{w}_l = ||\Delta\mathbf{w}_l||^2 \quad (30)$$

Robust E-Step:

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}=\{\mathbf{m},\mathbf{V}\}} \mathcal{E}_R(\boldsymbol{\gamma};\kappa) := \sum\nolimits_{\mathbf{x}\in\Omega} \rho\left(\mathcal{E}_{\mathbf{x}}(\boldsymbol{\gamma});\kappa\right) \quad (31)$$

$$\frac{\partial \mathcal{E}_R}{\partial \mathbf{m}} = \sum\nolimits_{\mathbf{x}\in\Omega} \psi\left(\mathcal{E}_{\mathbf{x}}(\boldsymbol{\gamma});\kappa\right)\frac{\partial \mathcal{E}_{\mathbf{x}}}{\partial \mathbf{m}} \quad (32)$$

$$\frac{\partial \mathcal{E}_R}{\partial \mathbf{V}} = \sum\nolimits_{\mathbf{x}\in\Omega} \psi\left(\mathcal{E}_{\mathbf{x}}(\boldsymbol{\gamma});\kappa\right)\frac{\partial \mathcal{E}_{\mathbf{x}}}{\partial \mathbf{V}} \quad (33)$$

$$\frac{\partial \mathcal{E}_{\mathbf{x}}}{\partial \mathbf{m}} = \left\{\frac{\widetilde{\mathbf{m}}^{\mathbf{x}} - \widetilde{\mu}^{\mathbf{x}}}{\widetilde{\sigma}^{\mathbf{x}}} - \mathbf{g}^m(\mathbf{x})\right\}(\mathbf{W}^{\mathbf{x}})^T \quad (34)$$

$$\frac{\partial \mathcal{E}_{\mathbf{x}}}{\partial \mathbf{V}} = \left\{-\frac{1}{2}\left[\frac{1}{\widetilde{\mathbf{V}}^{\mathbf{x}}} - \frac{1}{\widetilde{\sigma}^{\mathbf{x}}}\right] - \mathbf{G}^V(\mathbf{x})\right\}(\mathbf{W}^{\mathbf{x}})^T\mathbf{W}^{\mathbf{x}} \quad (35)$$

we use the robust $\rho-$function of Bianco and Yohai [87] (modified in [88] to ensure boundedness) tailored for logistic functions. The *influence* function reads as $\psi(k;\kappa) = \rho'(k;\kappa) = e^{-\sqrt{\kappa}}\delta(k \le \kappa) + e^{-\sqrt{k}}\delta(k > \kappa)$, where the tuning parameter $\kappa > 0$ provides a compromise between robustness and efficiency. Higher κ values yield a more efficient estimate by considering all pixels as inliers, yet less robust. The robust inference formulation can thus be written as in (31) with gradient expressions in (32) and (33). The *negative* of the marginal log-likelihood bound in (15) is convex w.r.t. the posterior variational parameters

$\boldsymbol{\gamma}$. Nonetheless, its robust formulation in (31) is no longer convex. Fortunately, the scale parameter κ allows for *continuation* methods to be used to find a globally optimal solution for the nonconvex $\mathcal{E}_R$. A global solution can be achieved by constructing successive convex approximations of $\mathcal{E}_R$ that can be readily minimized using gradient-based methods, *e.g.,* LBFGS. To construct such a sequence, we use a variant of the graduated nonconvexity algorithm [89]. The minimization can thus begin with $\kappa^{(0)} = \max \mathcal{E}_x(\boldsymbol{\gamma})$, chosen so there are no outliers, *i.e.,* $\rho''(k;\kappa) > 0 \;\; \forall \; k$. Outliers can then be gradually introduced by lowering the value of κ and repeating the minimization, starting from the solution of the previous approximation.

4. Experiments

We assessed the performance of ShapeOdds, as compared to *baseline* models, w.r.t. generating valid shapes (*realism*), modeling unseen shapes (*generalization*), and recovering valid shapes from corrupted ones (*robustness*).

Datasets: We considered two datasets that represent different challenging aspects in shape modeling. (1) The *Weizmann horse* dataset [90] contains 328 silhouettes of horses facing to the left with significant pose variation, cropped and normalized to 32×32 pixels as in [60]. The challenge of this dataset is the limited number of training samples as compared to the underlying shape variability, revealed by the different positions of heads, tails, and legs. (2) The *Caltech-101 motorbike* dataset [91] contains 798 silhouettes, cropped and normalized to 64×64 pixels. We use this dataset to manifest learning ShapeOdds in high-dimensional silhouette space with limited training samples.

Baseline models: For comparison, we considered the state-of-the-art ShapeBM [60], which learns shape models directly in the silhouettes space without relying on any intermediate representation. We used the same hyperparameters settings in [60] for the two datasets. Using the implementation provided by Stavros Tsogkas *et al*. [92], we trained ShapeBM with the overlap of four pixels using pre-training and 3000 epochs. Similar to [60], we used $2,000$ and 100 hidden units for the first and second layers, respectively, for the horse dataset. For the motorbike dataset, we used $1,200$ and 50 units. We further considered current practices that use intermediate embeddings such as signed distance maps (SDMs) and Gaussian smoothed silhouettes (GAUSS). Since any monotonic transformation of SDMs can be considered as a valid LogOdds representation [31], SDMs-based representation has a scalar free parameter that controls the smoothness of the resulting natural parameters map. Further, GAUSS-based representation has its kernel width as a free parameter. As such, for a fair comparison, the multiplicative factor for SDMs and the width of the Gaussian kernel were optimized using cross-validation over the training data. We considered learning shape models using PCA in the LogOdds space, similar to [31], and in the expectation parameters space, similar to [30]. For nonparametric models, we considered the kernel density estimate (KDE) using SDMs as in [42] where we fixed the kernel width to be the mean squared nearest-neighbor distance.

Realism: We sampled a set of horses and motorbikes from the learned models, see Figure 2, where we visualized the corresponding $\mathbf{q}-$maps. ShapeOdds can be sampled using its directed model where we start with sampling the Gaussian prior $p(\mathbf{z})$ and then mapping a latent point $\mathbf{z}$ to the $\mathcal{P}-$space using the mapping $h(\mathbf{z})$. ShapeBM was sampled using extended Gibbs sampling similar to [60]. Samples from PCA-based models were generated by sampling the within-subspace Gaussian distribution whose covariance structure is defined by the eigenvalues of the estimated principal subspace. One can note the poor samples generated from shape models learned in the *expectation* parameters space, *i.e.,* PCA-Prob-SDMs and PCA-Prob-GAUSS. The smoothness and ghosting artifacts that are evident in the generated samples are due to the *iterative projection* scheme [30] required to project a given shape onto the expectation parameters space, which amounts to clamping all values to the $[0, 1]-$interval. Learning shape models in the LogOdds space does not suffer from such artifacts. However, models fail to learn enough shape variability, leading to samples with similar shape that do not preserve shape class features such as horse legs. ShapeBM can generate sharply defined samples with different horse poses and motorbike shapes. Nonetheless, thin shape features, *e.g.,* horse legs, are not well defined, especially with highly variable poses. On the other hand, ShapeOdds can generate crisp $\mathbf{q}-$maps with significant shape variability while preserving shape details such as horse legs and motorbike handle bars.

Generalization: We considered a variant of the *generalization* measure in [93] to assess whether a learned model can represent *unseen* shape instances and quantify the ability of the learned density function to spread out between and around the training shapes. Rather than using sample reconstruction error as in [93], we used *cross-entropy* as in [94] to measure how likely an unseen sample $\mathbf{u}$ follows an Expon distribution with a parameter map $\mathbf{q}$ reconstructed from a shape model $\Theta^{(N)}$, trained over $N-$samples. The generalization measure reads as $G(\mathbf{u};\mathbf{q}) \doteq -\frac{1}{D}\left\{\mathbf{u}^T \log[\mathbf{q}] + (\mathbf{1}_D - \mathbf{u})^T \log[1-\mathbf{q}]\right\}$. Figure 3 reports the generalization statistics for both datasets as a function of the training sample size where training subsets of $N = \{15\%, 35\%, 55\%, 75\%\}$ were randomly drawn 10 times. Since KDE-SDMs does not attempt to learn the silhouettes distribution, one can observe its poor generalization, which reveals its tendency to overfit with sparse training samples in high-dimensional space. SDMs- and GAUSS-based models make use of more training samples for better generalization. Nonetheless, poorer performance indicates that they lead to suboptimal gener-

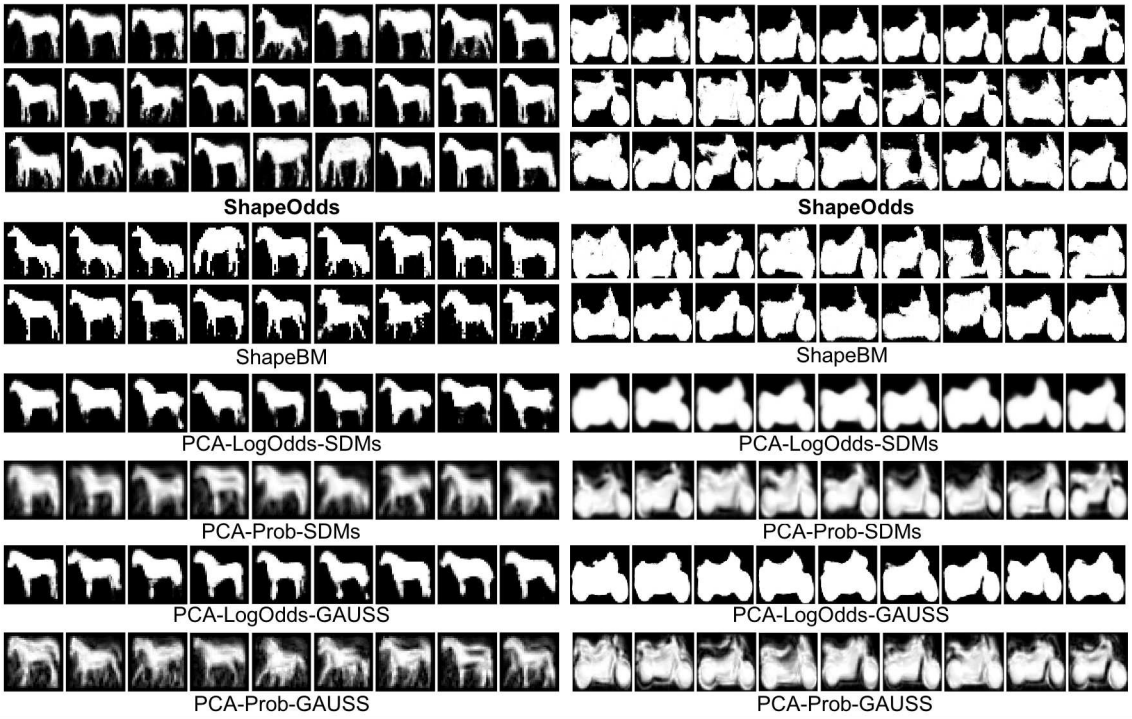

Figure 2. **Realism** – sampled shapes from (left) horses and (right) motorbikes datasets using ShapeOdds and other baseline models.

ative models that do not generalize well on unseen data. In particular, signed distance to the shape's boundary is a geometric representation that does not correlate well with the underlying generative process. Further, blurring silhouettes lose the ability to model the distribution of the given population due to the blind smoothing along shape boundaries irrespective of the underlying shape variability. The effect of such smoothing is more evident when learning the shape model in the LogOdds space, especially with small N, indicating that blurring silhouettes is not a statistically principled approach to embed silhouettes in the $\mathcal{P}-$space. ShapeOdds compares favorably against all baseline models and shows better generalization performance even with small training sizes compared to the underlying variability. However, ShapeBM shows slightly better generalization on the horse dataset with $N = 49$ samples. The main reason is that ShapeBM advocates an axis-aligned shape space partitioning with a weight-sharing scheme to balance the number of parameters to estimate and the generality of the model. Extending ShapeOdds to mixtures would achieve such a balance in a *data-driven* manner and a statistically principled approach by parameterizing each mixture component by its dominant subspace. Figure 4 demonstrates that ShapeOdds generalizes to unseen examples in nontrivial ways. One can observe the crisp $\mathbf{q}-$maps that are recovered from ShapeOdds compared to other baseline models.

Robustness: We further assessed ShapeOdds capability to recover valid shapes from corrupted ones. Here we consider *unbiased* noise where there is no prior assumption about the

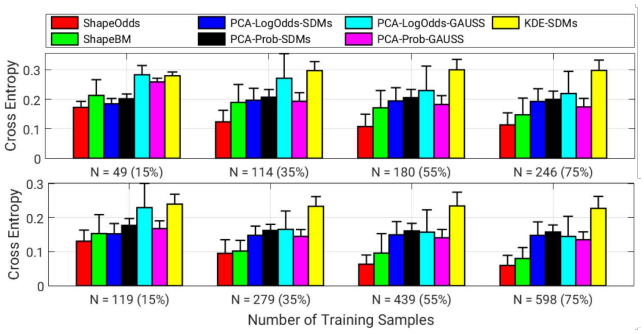

Figure 3. Mean and std of the generalization measure on the horse (top) and motorbike (bottom) datasets. Lower is better.

corrupted region, *i.e.,* foreground and/or background. An *unseen* silhouette is assumed to be corrupted by another Bernoulli random field with parameter map $\mathbf{q}_c$ and with contamination rate $\epsilon \in [0, 1]$. We generated correlated noise masks using $\mathbf{q}_c$'s simulated by convolving random noise with a Gaussian kernel of standard deviation $\sigma_c = 2.0$ and mapping the resulting field to $[0, 1]-$interval. Missing foreground regions and/or background clutter were determined by thresholding the simulated $\mathbf{q}_c$ using the threshold that results in a contamination rate of $\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Figure 5 shows the cross-entropy of the recovered $\mathbf{q}-$maps for corrupted horse silhouettes as a function of ϵ. Figure 6 demonstrates sample inference results of corrupted horses for different corruption scenarios. Motorbike dataset showed a similar performance but was omitted due to space limitations. Note that the best performance for ShapeBM is handling missing foreground regions, *e.g.,* occlusion. How-

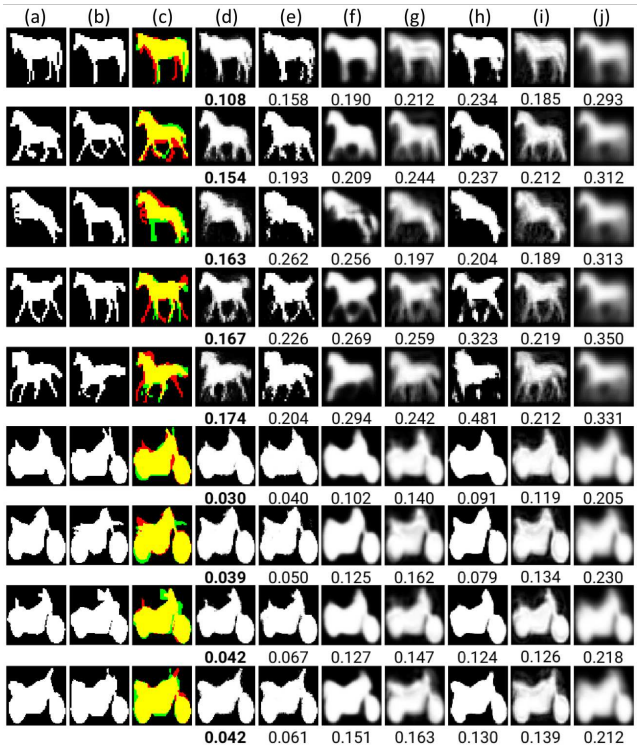

Figure 4. **Generalization**: (a) unseen silhouette, (b) closest silhouette in the training dataset, (c) overlay of (a) and (b) (red pixels are present only in the unseen sample, green pixels are present only in the training sample, and yellow pixels are present in both), reconstructed $\mathbf{q}-$maps from (d) ShapeOdds, (e) ShapeBM, (f) PCA-LogOdds-SDMs, (g) PCA-Prob-SDMs, (h) PCA-LogOdds-GAUSS, (i) PCA-Prob-GAUSS, and (j) KDE-SDMs. Generarlization measure $G(\mathbf{u};\mathbf{q})$ (lower is better) is reported where bold indicates best generalization.

ever, qualitative and quantitative results are indicative of its poor performance in cases of background clutter and general unbiased corruption. KDE-SDMs constructs a nonparametric estimate of the $\phi-$maps based on the similarity of the given corrupted silhouette to each training sample. However, it tends to recover similar, over-smoothed, $\mathbf{q}-$maps for all corrupted silhouettes. This is a typical mode of failure for this model and appears to be an inability to find, through optimization, a good set of weights on training samples to, in turn, recover a good parameter map. ShapeOdds shows some success in handling low noise levels, but it fails to properly recover valid $\mathbf{q}-$maps for highly contaminated silhouettes. The proposed robust inference, on the other hand, maintains good performance even with high levels of foreground and/or background corruption.

5. Conclusion and Future Work

We presented a probabilistic generative shape model – ShapeOdds – that can capture variability patterns directly in the silhouettes space. Our formulation offers a tractable and deterministic EM-like model learning that avoids overfitting in high-dimensional shape spaces with closed-form re-estimation formulas for the hyperparameters. Experiments have shown that ShapeOdds can generate realistic-looking shapes, generalize to unseen samples in nontrivial ways, and recover shape instances from corrupted ones. In the future, we plan to pursue several extensions to ShapeOdds, including data-driven soft partitioning of the shape space by learning mixtures of ShapeOdds, transformation-invariant model learning to relax the assumption of aligned training shapes, learning with outlying shape instances by using heavy-tailed latent priors, joint modeling of shape and appearance, and deep latent models to allow scaling to higher resolution silhouettes while avoiding overfitting.

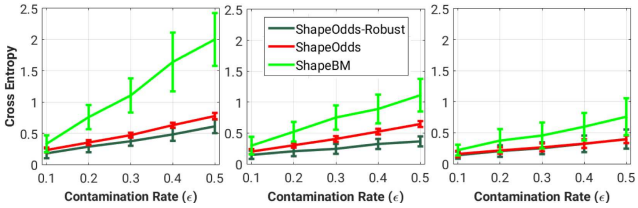

Figure 5. **Robustness** – horse: (left) missing foreground regions and background clutter, (middle) background clutter only, and (right) missing foreground regions only. Lower is better.

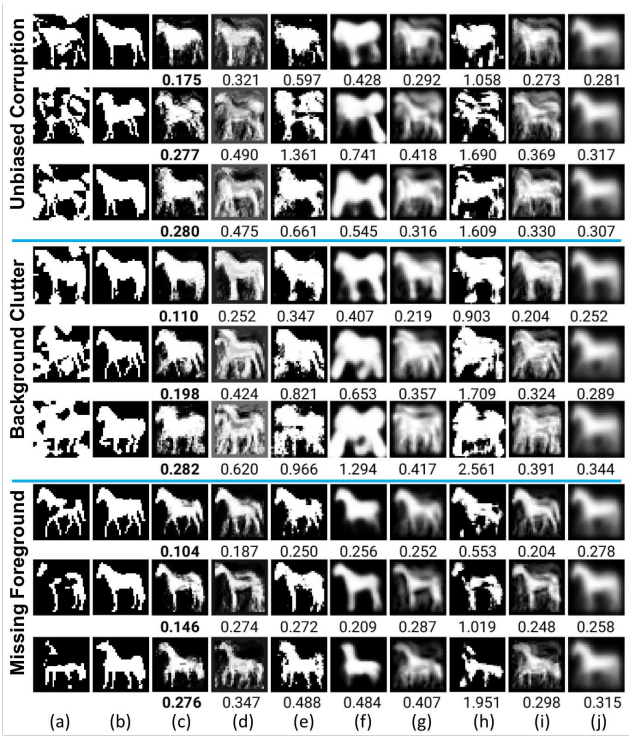

Figure 6. **Robustness**: (a) corrupted silhouette, (b) groundtruth silhouette, $\mathbf{q}-$maps recovered from (c) ShapeOdds-Robust, (d) ShapeOdds, (e) ShapeBM, (f) PCA-LogOdds-SDMs, (g) PCA-Prob-SDMs, (h) PCA-LogOdds-GAUSS, (i) PCA-Prob-GAUSS, and (j) KDE-SDMs. Cross entropy (lower is better) is reported where bold indicates best performance.

References

[1] S. Milborrow and F. Nicolls, "Locating facial features with an extended active model," in *European conference on computer vision*, pp. 504–513, Springer, 2008.

[2] Y. Li, L. Gu, and T. Kanade, "A robust shape model for multi-view car alignment," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2466–2473, IEEE, 2009.

[3] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *European Conference on Computer Vision*, pp. 413–426, Springer, 2008.

[4] N. H. Trinh and B. B. Kimia, "Skeleton search: Category-specific object recognition and segmentation using a skeletal shape model," *International Journal of Computer Vision*, vol. 94, no. 2, pp. 215–240, 2011.

[5] A. Toshev, A. Makadia, and K. Daniilidis, "Shape-based object recognition in videos using 3d synthetic object models," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 288–295, IEEE, 2009.

[6] O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black, "Contour people: A parameterized model of 2d articulated human shape," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 639–646, IEEE, 2010.

[7] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1385–1392, IEEE, 2011.

[8] V. Ferrari, F. Jurie, and C. Schmid, "From images to shape models for object detection," *International journal of computer vision*, vol. 87, no. 3, pp. 284–303, 2010.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[10] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3d geometric model," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1688–1695, IEEE, 2010.

[11] B. Alexe, T. Deselaers, and V. Ferrari, "Classcut for unsupervised class segmentation," in *European conference on computer vision*, pp. 380–393, Springer, 2010.

[12] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, "A bayesian model of shape and appearance for subcortical brain segmentation," *Neuroimage*, vol. 56, no. 3, pp. 907–922, 2011.

[13] D. Grosgeorge, C. Petitjean, J.-N. Dacher, and S. Ruan, "Graph cut segmentation with a statistical shape model in cardiac mri," *Computer Vision and Image Understanding*, vol. 117, no. 9, pp. 1027–1035, 2013.

[14] N. Vu and B. Manjunath, "Shape prior segmentation of multiple objects with graph cuts," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[15] M. P. Kumar, P. H. Torr, and A. Zisserman, "Objcut: Efficient segmentation using top-down and bottom-up cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 530–545, 2010.

[16] M. Fussenegger, P. Roth, H. Bischof, R. Deriche, and A. Pinz, "A level set framework using a new incremental, robust active shape model for object segmentation and tracking," *Image and Vision Computing*, vol. 27, no. 8, pp. 1157–1168, 2009.

[17] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2610–2617, IEEE, 2012.

[18] D. Cremers, "Dynamical statistical shape priors for level set-based tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1262–1273, 2006.

[19] M. Paladini, A. Bartoli, and L. Agapito, "Sequential non-rigid structure-from-motion with the 3d-implicit low-rank shape model," in *European Conference on Computer Vision*, pp. 15–28, Springer, 2010.

[20] I. Kemelmacher-Shlizerman and R. Basri, "3d face reconstruction from a single image using a single reference face shape," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 394–405, 2011.

[21] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," in *Computer Graphics Forum*, vol. 28, pp. 337–346, Wiley Online Library, 2009.

[22] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 41, 2013.

[23] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM Transactions on Graphics (TOG)*, vol. 24, pp. 408–416, ACM, 2005.

[24] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun, "A probabilistic model for component-based shape synthesis," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, p. 55, 2012.

[25] M. Rousson and N. Paragios, "Prior knowledge, level set representations & visual grouping," *International Journal of Computer Vision*, vol. 76, no. 3, pp. 231–243, 2008.

[26] M. Szummer, P. Kohli, and D. Hoiem, "Learning crfs using graph cuts," in *Computer Vision–ECCV 2008*, pp. 582–595, Springer, 2008.

[27] P. Kohli, P. H. Torr, *et al.*, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[28] S. Nowozin and C. H. Lampert, "Global connectivity potentials for random field models," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 818–825, IEEE, 2009.

[29] C. Rother, P. Kohli, W. Feng, and J. Jia, "Minimizing sparse higher order energy functions of discrete variables," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1382–1389, IEEE, 2009.

[30] D. Cremers, F. R. Schmidt, and F. Barthel, "Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–6, IEEE, 2008.

[31] K. M. Pohl, J. Fisher, S. Bouix, M. Shenton, R. W. McCarley, W. E. L. Grimson, R. Kikinis, and W. M. Wells, "Using the logarithm of odds to define a vector space on probabilistic atlases," *Medical Image Analysis*, vol. 11, no. 5, pp. 465–477, 2007.

[32] M. E. Leventon, W. E. L. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *CVPR*, vol. 1, pp. 316–323, IEEE, 2000.

[33] A. Tsai, A. Yezzi Jr, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 137–154, 2003.

[34] M. Rousson, N. Paragios, and R. Deriche, "Implicit active shape models for 3d segmentation in mr imaging," in *MICCAI*, pp. 209–216, Springer, 2004.

[35] D. Cremers, T. Kohlberger, and C. Schnörr, "Nonlinear shape statistics in mumfordshah based segmentation," *Computer VisionECCV 2002*, pp. 516–518, 2002.

[36] D. Cremers, T. Kohlberger, and C. Schnörr, "Shape statistics in kernel space for variational image segmentation," *Pattern Recognition*, vol. 36, no. 9, pp. 1929–1943, 2003.

[37] S. Dambreville, Y. Rathi, and A. Tannen, "Shape-based approach to robust image segmentation using kernel pca," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 977–984, IEEE, 2006.

[38] S. Dambreville, Y. Rathi, and A. Tannenbaum, "A framework for image segmentation using shape models and kernel space shape priors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 8, pp. 1385–1399, 2008.

[39] F. Lecumberry, Á. Pardo, and G. Sapiro, "Simultaneous object classification and segmentation with high-order multiple shape models," *IEEE Transactions on Image Processing*, vol. 19, no. 3, pp. 625–635, 2010.

[40] M. Rousson and D. Cremers, "Efficient kernel density estimation of shape and intensity priors for level set segmentation," in *MICCAI*, pp. 757–764, Springer, 2005.

[41] D. Cremers and M. Rousson, "Efficient kernel density estimation of shape and intensity priors for level set segmentation," in *Deformable Models*, pp. 447–460, Springer, 2007.

[42] D. Cremers, S. J. Osher, and S. Soatto, "Kernel density estimation and intrinsic alignment for shape priors in level set segmentation," *IJCV*, vol. 69, no. 3, pp. 335–351, 2006.

[43] J. Kim, M. Çetin, and A. S. Willsky, "Nonparametric shape priors for active contour-based image segmentation," *Signal Processing*, vol. 87, no. 12, pp. 3021–3044, 2007.

[44] A. Wimmer, G. Soza, and J. Hornegger, "A generic probabilistic active shape model for organ segmentation," in *MICCAI*, pp. 26–33, Springer, 2009.

[45] D. M. Gavrila, "A bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1408–1421, 2007.

[46] J. H. Friedman, W. Stuetzle, and A. Schroeder, "Projection pursuit density estimation," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 599–608, 1984.

[47] S. Dasgupta, "Learning mixtures of gaussians," in *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 634–644, IEEE, 1999.

[48] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[49] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[50] A. I. Schein, L. K. Saul, and L. H. Ungar, "A generalized linear model for principal component analysis of binary data.," in *AISTATS*, vol. 3, p. 10, 2003.

[51] C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 281–293, 1998.

[52] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[53] H. Kiers, "Multivariate analysis, part 2: Classification, covariance structure, and repeated measurements, by wj krzanowski and fhc marriott," *Journal of Classification*, vol. 15, no. 2, pp. 294–297, 1998.

[54] M. Welling, C. Chemudugunta, and N. Sutter, "Deterministic latent variable models and their pitfalls," *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 196–207, 2008.

[55] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *NIPS*, vol. 20, pp. 1–8, 2011.

[56] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proceedings of the 25th international conference on Machine learning*, pp. 880–887, ACM, 2008.

[57] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting crfs with boltzmann machine shape priors for image labeling," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2019–2026, IEEE, 2013.

[58] Y. Li, D. Tarlow, and R. Zemel, "Exploring compositional high order pattern potentials for structured output learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 49–56, IEEE, 2013.

[59] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *International Conference on Artificial Intelligence and Statistics*, pp. 448–455, 2009.

[60] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn, "The shape boltzmann machine: a strong model of object shape," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 155–176, 2014.

[61] S. Eslami and C. Williams, "A generative model for parts-based object segmentation," in *Advances in Neural Information Processing Systems*, pp. 100–107, 2012.

[62] F. Chen, H. Yu, R. Hu, and X. Zeng, "Deep learning shape priors for object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1870–1877, 2013.

[63] J. Yang, S. Safar, and M.-H. Yang, "Max-margin boltzmann machines for object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 320–327, 2014.

[64] Y. Gal and Z. Ghahramani, "On modern deep learning and variational inference," in *Advances in Approximate Bayesian Inference workshop, NIPS*, 2015.

[65] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," in *Advances in Neural Information Processing Systems*, pp. 912–919, 1992.

[66] M. E. Khan, G. Bouchard, K. P. Murphy, and B. M. Marlin, "Variational bounds for mixed-data factor analysis," in *Advances in Neural Information Processing Systems*, pp. 1108–1116, 2010.

[67] M. A. Carreira-Perpinan and G. Hinton, "On contrastive divergence learning.," in *AISTATS*, vol. 10, pp. 33–40, Citeseer, 2005.

[68] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[69] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071, ACM, 2008.

[70] T. Tieleman and G. Hinton, "Using fast weights to improve persistent contrastive divergence," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1033–1040, ACM, 2009.

[71] S. Z. Li, *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[72] M. J. Wainwright, "Structured regularizers for high-dimensional problems: Statistical and computational issues," *Annual Review of Statistics and Its Application*, vol. 1, pp. 233–253, 2014.

[73] C. M. Bishop, "Bayesian pca," *Advances in neural information processing systems*, pp. 382–388, 1999.

[74] D. P. Wipf and S. S. Nagarajan, "A new view of automatic relevance determination," in *Advances in neural information processing systems*, pp. 1625–1632, 2008.

[75] R. M. Neal, "Bayesian learning for neural networks," 1996.

[76] C. Archambeau and F. R. Bach, "Sparse probabilistic projections," in *Advances in neural information processing systems*, pp. 73–80, 2009.

[77] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[78] D. J. Bartholomew, "The foundations of factor analysis," *Biometrika*, vol. 71, no. 2, pp. 221–232, 1984.

[79] B. Everett, *An introduction to latent variable models*. Springer Science & Business Media, 2013.

[80] B. M. Marlin, M. E. Khan, and K. P. Murphy, "Piecewise bounds for estimating bernoulli-logistic latent gaussian models.," in *ICML*, pp. 633–640, 2011.

[81] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 197–200, 1992.

[82] T. Jaakkola and M. I. Jordan, "A variational approach to bayesian logistic regression models and their extensions," in *Sixth International Workshop on Artificial Intelligence and Statistics*, vol. 82, 1997.

[83] E. Khan, S. Mohamed, and K. P. Murphy, "Fast bayesian inference for non-conjugate gaussian process regression," in *Advances in Neural Information Processing Systems*, pp. 3140–3148, 2012.

[84] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[85] X. Yi and C. Caramanis, "Regularized em algorithms: A unified framework and statistical guarantees," in *Advances in Neural Information Processing Systems*, pp. 1567–1575, 2015.

[86] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, vol. 114. John Wiley & Sons, 2011.

[87] A. M. Bianco and V. J. Yohai, "Robust estimation in the logistic regression model," in *Robust statistics, data analysis, and computer intensive methods*, pp. 17–34, Springer, 1996.

[88] C. Croux and G. Haesbroeck, "Implementing the bianco and yohai estimator for logistic regression," *Computational statistics & data analysis*, vol. 44, no. 1, pp. 273–295, 2003.

[89] 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, no. 1, pp. 75–104, 1996.

[90] E. Borenstein and S. Ullman, "Combined top-down/bottom-up segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2109–2125, 2008.

[91] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[92] S. Tsogkas, I. Kokkinos, G. Papandreou, and A. Vedaldi, "Semantic part segmentation with deep learning," *arXiv preprint arXiv:1505.02438*, 2015.

[93] R. Davies, C. Taylor, *et al.*, *Statistical models of shape: Optimisation and evaluation*. Springer Science & Business Media, 2008.

[94] Z. Zivkovic and J. Verbeek, "Transformation invariant component analysis for binary images," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 254–259, IEEE, 2006.