Int J Comput Vis
DOI 10.1007/s11263-007-0069-5

Describing Visual Scenes Using Transformed Objects and Parts

Erik B. Sudderth · Antonio Torralba · William T. Freeman · Alan S. Willsky

Received: 20 September 2005 / Accepted: 29 May 2007

Abstract We develop hierarchical, probabilistic models for objects, the parts composing them, and the visual scenes surrounding them. Our approach couples topic models originally developed for text analysis with spatial transformations, and thus consistently accounts for geometric constraints. By building integrated scene models, we may discover contextual relationships, and better exploit partially labeled training images. We first consider images of isolated objects, and show that sharing parts among object categories improves detection accuracy when learning from few examples. Turning to multiple object scenes, we propose nonparametric models which use Dirichlet processes to automatically learn the *number* of parts underlying each object category, and objects composing each scene. The resulting transformed Dirichlet process (TDP) leads to Monte Carlo algorithms which simultaneously segment and recognize objects in street and office scenes.

E.B. Sudderth (✉)
Computer Science Division, University of California, Berkeley, USA
e-mail: sudderth@eecs.berkeley.edu

A. Torralba · W.T. Freeman · A.S. Willsky
Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

A. Torralba
e-mail: torralba@csail.mit.edu

W.T. Freeman
e-mail: billf@mit.edu

A.S. Willsky
e-mail: willsky@mit.edu

Keywords Object recognition · Dirichlet process · Hierarchical Dirichlet process · Transformation · Context · Graphical models · Scene analysis

1 Introduction

Object recognition systems use the image features composing a *visual scene* to localize and categorize objects. We argue that multi-object recognition should consider the relationships between different object categories during the training process. This approach provides several benefits. At the lowest level, significant computational savings are possible if different categories share a common set of features. More importantly, jointly trained recognition systems can use similarities between object categories to their advantage by learning features which lead to better generalization (Torralba et al. 2004; Fei-Fei et al. 2004). This transfer of knowledge is particularly important when few training examples are available, or when unsupervised discovery of new objects is desired. Furthermore, contextual knowledge can often improve performance in complex, natural scenes. At the coarsest level, the overall spatial structure, or *gist*, of an image provides priming information about likely object categories, and their most probable locations within the scene (Torralba 2003; Murphy et al. 2004). In addition, exploiting spatial relationships between objects can improve detection of less distinctive categories (Fink and Perona 2004; Tu et al. 2005; He et al. 2004; Amit and Trouvé 2007).

In this paper, we develop a family of hierarchical generative models for objects, the parts composing them, and the scenes surrounding them. We focus on the so-called *basic level* recognition of visually identifiable categories, rather than the differentiation of object instances (Liter and

Bülthoff 1998). Our models share information between object categories in three distinct ways. First, parts define distributions over a common low-level feature vocabularly, leading to computational savings when analyzing new images. In addition, and more unusually, objects are defined using a common set of parts. This structure leads to the discovery of parts with interesting semantic interpretations, and can improve performance when few training examples are available. Finally, object appearance information is shared between the many scenes in which that object is found.

This generative approach is motivated by the pragmatic need for learning algorithms which require little manual supervision and labeling. While discriminative models often produce accurate classifiers, they typically require very large training sets even for relatively simple categories (Viola and Jones 2004; LeCun et al. 2004). In contrast, generative approaches can discover large, visually salient categories (such as foliage and buildings Sivic et al. 2005) without supervision. Partial segmentations can then be used to learn semantically interesting categories (such as cars and pedestrians) which are less visually distinctive, or present in fewer training images. Moreover, by employing a single hierarchy describing multiple objects or scenes, the learning process automatically shares information between categories.

Our hierarchical models are adapted from *topic models* originally used to analyze text documents (Blei et al. 2003; Teh et al. 2006). These models make the so-called *bag of words* assumption, in which raw documents are converted to word counts, and sentence structure is ignored. While it is possible to develop corresponding *bag of features* models for images (Sivic et al. 2005; Fei-Fei and Perona 2005; Barnard et al. 2003; Csurka et al. 2004), which model the appearance of detected interest points and ignore their location, we show that doing so neglects valuable information, and reduces recognition performance. To consistently account for spatial structure, we augment these hierarchies with *transformation* (Miller et al. 2000; Jojic and Frey 2001; Frey and Jojic 2003; Simard et al. 1998) variables describing the locations of objects in each image. Through these transformations, we learn parts which describe features relative to a "canonical" coordinate frame, without requiring the alignment of training or test images.

The principal challenge in developing hierarchical models for scenes is specifying tractable, scalable methods for handling uncertainty in the number of objects. This issue is entirely ignored by most existing models, which are either tested on cropped images of single objects (Weber et al. 2000; Fei-Fei et al. 2004; Borenstein and Ullman 2002), or use heuristics to combine the outputs of local "sliding window" classifiers (Viola and Jones 2004; Torralba et al. 2004; Ullman et al. 2002). *Grammars*, and related rule-based systems, provide one flexible family of hierarchical representations (Tenenbaum and Barrow 1977; Bienenstock et al. 1997). For example, several different models impose distributions on hierarchical tree-structured segmentations of the pixels composing simple scenes (Adams and Williams 2003; Storkey and Williams 2003; Siskind et al. 2004; Hinton et al. 2000; Jin and Geman 2006). In addition, an *image parsing* (Tu et al. 2005) framework has been proposed which explains an image using a set of regions generated by generic or object-specific processes. While this model allows uncertainty in the number of regions, and hence objects, its high-dimensional state space requires discriminatively trained, bottom-up proposal distributions. The BLOG language (Milch et al. 2005) provides a promising framework for *representing* unknown objects, but does not address the computational and statistical challenges which arise when *learning* scene models from training data.

We propose a different, data-driven framework for handling uncertainty in the number of object instances, based on Dirichlet processes (DPs) (Jordan 2005; Pitman 2002; Sudderth 2006). In nonparametric Bayesian statistics, DPs are used to learn mixture models whose number of components is automatically inferred from data (Escobar and West 1995; Neal 2000). A *hierarchical Dirichlet process* (HDP) (Teh et al. 2006) describes several related datasets by reusing mixture components in different proportions. We extend the HDP framework by allowing the global, shared mixture components to undergo a random set of transformations. The resulting *transformed Dirichlet process* (TDP) produces models which automatically learn the number of parts underlying each object category, and objects composing each scene.

The following section begins by reviewing prior work on feature-based image representations, and existing bag of features image models. We then develop hierarchical models which share parts among related object categories, automatically infer the number of depicted object instances, and exploit contextual relationships when parsing multiple object scenes. We evaluate these models by learning shared representations for sixteen object categories (Sect. 5), and detecting multiple objects in street and office scenes (Sect. 9).

2 Generative Models for Image Features

In this paper, we employ sparse image representations derived from local interest operators. This approach reduces dimensionality and dependencies among features, and simplifies object appearance models by focusing on the most salient, repeatable image structures. While the features we employ are known to perform well in geometric correspondence tasks (Mikolajczyk and Schmid 2005), we emphasize that our object and scene models could be easily adapted to alternative families of local descriptors.

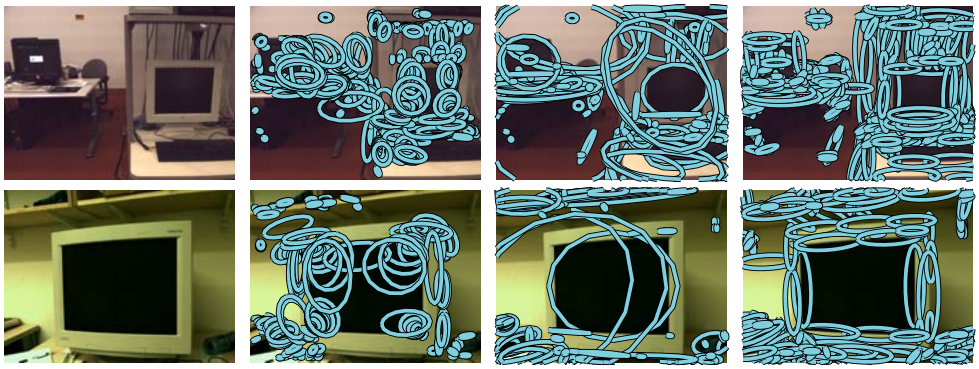

Fig. 1 Three types of interest operators applied to two office scenes: Harris-affine corners (*left*), maximally stable extremal regions (*center*), and linked sequences of Canny edges (*right*)

2.1 Feature Extraction

In each grayscale training or test image, we begin by detecting a set of elliptical interest regions (see Fig. 1). We consider three complementary criteria for region extraction. *Harris-affine* invariant regions (Mikolajczyk and Schmid 2004) detect corner-like image structure by finding pixels with significant second derivatives. The Laplacian of Gaussian operator (Lowe 2004) then provides a characteristic scale for each corner. Alternatively, *maximally stable extremal regions* (MSER) (Matas et al. 2002) are derived by analyzing the stability of a watershed segmentation algorithm. As illustrated in Fig. 1, this approach favors large, homogeneous image regions.[1] For object recognition tasks, edge-based features are also highly informative (Belongie et al. 2002). To exploit this, we find candidate edges via a Canny detector (Canny 1986), and link them into segments broken at points of high curvature (Kovesi 2005). These lines then form the major axes of elliptical interest regions, whose minor axes are taken to be 10% of that length.

Given the density at which interest regions are detected, these features provide a multiscale over-segmentation of the image. Note that low-level interest operators are inherently noisy: even state-of-the-art detectors sometimes miss salient regions, and select features which do not align with real 3D scene structure (see Fig. 1 for examples). We handle this issue by extracting large feature sets, so that many regions are likely to be salient. It is then important to design recognition algorithms which exploit this redundancy, rather than relying on a small set of key features.

2.2 Feature Description

Following several recent approaches to recognition (Sivic et al. 2005; Fei-Fei and Perona 2005; Csurka et al. 2004), we use SIFT descriptors (Lowe 2004) to describe the appearance of interest regions. SIFT descriptors are derived from windowed histograms of gradient magnitudes at varying locations and orientations, normalized to correct for contrast and saturation effects. This approach provides some invariance to lighting and pose changes, and was more effective than raw pixel patches (Ullman et al. 2002) in our experiments.

To simplify learning algorithms, we convert each raw, 128-dimensional SIFT descriptor to a vector quantized discrete value (Sivic et al. 2005; Fei-Fei and Perona 2005). For each training database, we use K-means clustering to identify a finite dictionary of W appearance patterns, where each of the three feature types is mapped to a disjoint set of *visual words*. We set the total dictionary size via cross-validation; typically, $W \approx 1\,000$ seems appropriate for categorization tasks. In some experiments, we improve discriminative power by dividing the affinely adapted regions according to their shape. Edges are separated by orientation (horizontal versus vertical), while Harris-affine and MSER regions are divided into three groups (roughly circular, versus horizontally or vertically elongated). An expanded dictionary then jointly encodes the appearance and coarse shape of each feature.

Using this visual dictionary, the i^{th} interest region in image j is described by its detected image position v_{ji}, and the discrete appearance word w_{ji} with minimal Euclidean distance (Lowe 2004). Let $\mathbf{w}_j$ and $\mathbf{v}_j$ denote the appearance and two-dimensional position, respectively, of the N_j features in image j. Figure 2 illustrates some of the visual words extracted from a database of office scenes.

2.3 Visual Recognition with Bags of Features

In many domains, there are several *groups* of data which are thought to be produced by related generative processes. For example, the words composing a text corpus are typically separated into documents which discuss partially overlapping topics (Blei et al. 2003; Griffiths and Steyvers 2004; Teh et al. 2006). Alternatively, image databases like MIT's LabelMe depict visual scenes which compose many different object categories (Russell et al. 2005). While it is simplest to analyze each group independently, doing so often

[1] Software for the detection of Harris-affine and MSER features, and computation of SIFT descriptors (Lowe 2004), was provided by the Oxford University Visual Geometry Group: http://www.robots.ox.ac.uk/~vgg/research/affine/.

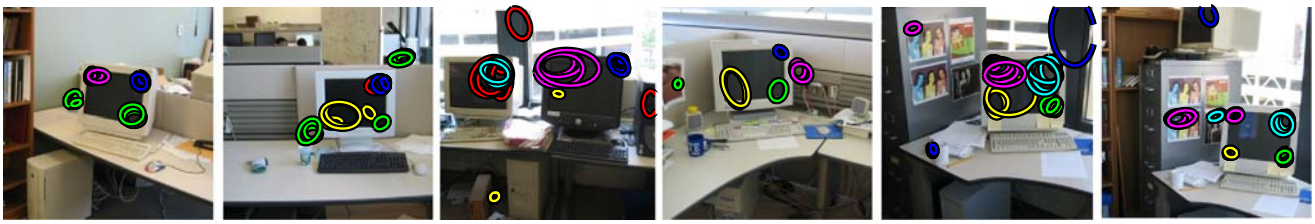

Fig. 2 A subset of the affine covariant features (*ellipses*) detected in images of office scenes. In five different colors, we show the features corresponding to the five discrete vocabulary words which most frequently align with computer screens in the training images

neglects critical information. By *sharing* random parameters among groups, hierarchical Bayesian models (Gelman et al. 2004) provide an elegant mechanism for transferring information between related documents, objects, or scenes.

Latent Dirichlet allocation (LDA) (Blei et al. 2003) provides one framework for learning mixture models which describe several related sets of observations. Given J groups of data, let $\mathbf{x}_j = (x_{j1}, \ldots, x_{jN_j})$ denote the N_j data points in group j, and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$. LDA assumes that the data within each group are *exchangeable*,[2] and independently sampled from one of K latent clusters with parameters $\{\theta_k\}_{k=1}^K$. Letting $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$ denote the mixture weights for the j^{th} group, we have

$$p(x_{ji}|\boldsymbol{\pi}_j, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_{jk} f(x_{ji}|\theta_k)$$
$$i = 1, \ldots, N_j. \tag{1}$$

Here, $f(x|\theta)$ is family of probability densities, with corresponding distributions $F(\theta)$ parameterized by θ. We later use multinomial $F(\theta)$ to model visual words, and Gaussian $F(\theta)$ to generate feature locations. LDA's use of shared mixture parameters transfers information among groups, while distinct mixture weights capture the unique features of individual groups. As discussed in Appendix 1, we improve the robustness of learning algorithms by placing *conjugate priors* (Gelman et al. 2004; Sudderth 2006) on the cluster parameters $\theta_k \sim H(\lambda)$. Mixture weights are sampled from a Dirichlet prior $\boldsymbol{\pi}_j \sim \mathrm{Dir}(\alpha)$, with hyperparameters α either tuned by cross-validation (Griffiths and Steyvers 2004) or learned from training data (Blei et al. 2003).

LDA has been used to analyze text corpora by associating groups with documents and data x_{ji} with words. The exchangeability assumption ignores sentence structure, treating each document as a "bag of words". This approximation leads to tractable algorithms which learn *topics* (clusters) from unlabeled document collections (Blei et al. 2003; Griffiths and Steyvers 2004). Using image features like those in Sect. 2, topic models have also been adapted to discover objects in simple scenes (Sivic et al. 2005) or web search results (Fergus et al. 2005), categorize natural scenes (Fei-Fei and Perona 2005; Bosch et al. 2006), and parse presegmented captioned images (Barnard et al. 2003). However, following an initial stage of low-level feature detection or segmentation, these approaches ignore spatial information, discarding positions $\mathbf{v}_j$ and treating the image as an unstructured bag of features $\mathbf{w}_j$. This paper instead develops richer hierarchical models which consistently incorporate spatial relationships.

[2]Exchangeable datasets have no intrinsic order, so that every permutation has equal joint probability (Gelman et al. 2004; Sudderth 2006).

2.4 Overview of Proposed Hierarchical Models

In the remainder of this paper, we introduce a family of hierarchical models for visual scenes and object categories. We begin by considering images depicting single objects, and develop models which share *parts* among related categories. Using spatial transformations, we then develop models which decompose scenes via a set of part-based representations of object appearance.

Fixed-Order Object Model In Sect. 3, we describe multiple object categories using a fixed number of shared parts. Results in Sect. 5 show that sharing improves detection performance when few training images are available.

Nonparametric Object Model In Sect. 4, we adapt the hierarchical Dirichlet process (Teh et al. 2006) to learn the *number* of shared parts underlying a set of object categories. The resulting nonparametric model learns representations whose complexity grows as more training images are observed.

Fixed-Order Scene Model In Sect. 6, we learn contextual relationships among a fixed number of objects, which in turn share parts as in Sect. 3. Results in Sect. 9 show that contextual cues improve detection performance for scenes with predictable, global spatial structure.

Nonparametric Scene Model In Sect. 7, we develop a transformed Dirichlet process (TDP), and use it to learn scene models which allow uncertainty in the number of visual object categories, and object instances depicted in each image. Section 8 then integrates the part-based object representations of Sect. 4 with the TDP, and thus more accurately segments novel scenes (see Sect. 9).

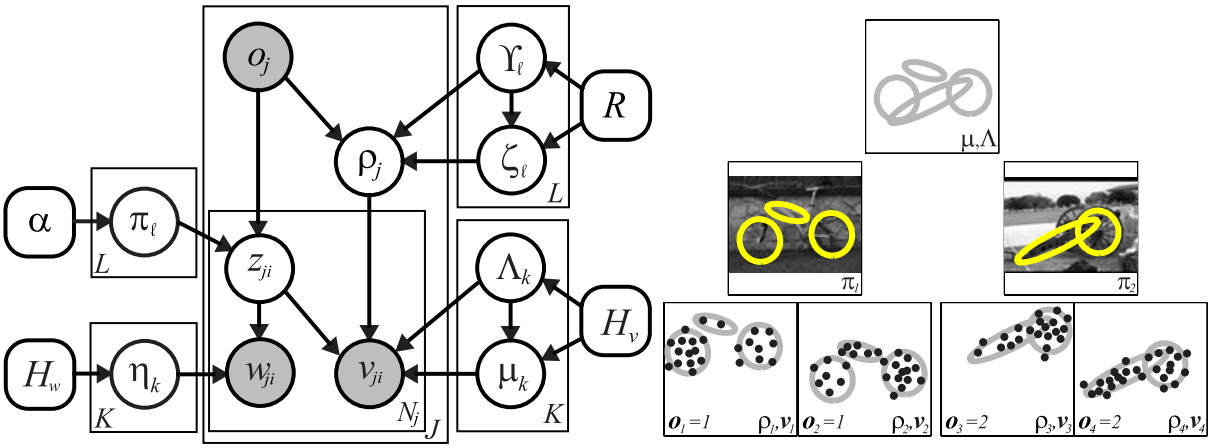

Fig. 3 A parametric, fixed-order model which describes the visual appearance of L object categories via a common set of K shared parts. The j^{th} image depicts an instance of object category o_j, whose position is determined by the reference transformation ρ_j. The appearance w_{ji} and position v_{ji}, relative to ρ_j, of visual features are determined by assignments $z_{ji} \sim \boldsymbol{\pi}_{o_j}$ to latent parts. The cartoon example illustrates how a wheel part might be shared among two categories, *bicycle* and *cannon*. We show feature positions (but not appearance) for two hypothetical samples from each category

3 Learning Parts Shared by Multiple Objects

Figure 3 illustrates a directed graphical model which extends LDA (Blei et al. 2003; Rosen-Zvi et al. 2004) to learn shared, part-based representations for multiple object categories. Nodes of this graph represent random variables or distributions, where shaded nodes are observed during training, and rounded boxes are fixed hyperparameters. Edges encode the conditional densities underlying the generative process (Jordan 2004; Sudderth 2006). To develop this model, we first introduce a flexible family of spatial transformations.

3.1 Capturing Spatial Structure with Transformations

Figure 4 illustrates the challenges in developing visual scene models incorporating feature positions. Due to variability in three-dimensional object location and pose, the absolute position at which features are observed may provide little information about their corresponding category. Recall that LDA models different groups of data by reusing *identical* cluster parameters θ_k in varying proportions. Applied directly to features incorporating both position and appearance, such topic models would need a separate global cluster for every possible location of each object category. Clearly, this approach does not sensibly describe the spatial structure underlying real scenes, and would not adequately generalize to images captured in new environments.

A more effective model of visual scenes would allow the same global cluster to describe objects at many different locations. To accomplish this, we augment topic models with *transformation* variables, thereby shifting global clusters from a "canonical" coordinate frame to the object positions underlying a particular image. Let $\tau(\theta; \rho)$ denote a family of transformations of the parameter vector θ, indexed by $\rho \in \wp$. For computational reasons, we assume that parameter transformations are invertible, and have a complementary *data transformation* $\tilde{\tau}(v; \rho)$ defined so that

$$f(v|\tau(\theta; \rho)) = \frac{1}{Z(\rho)} f(\tilde{\tau}(v; \rho)|\theta). \tag{2}$$

The normalization constant $Z(\rho)$, which is determined by the transformation's Jacobian, is assumed independent of the underlying parameters θ. Using (2), model transformations $\tau(\theta; \rho)$ are equivalently expressed by a change $\tilde{\tau}(v; \rho)$ of the observations' coordinate system. In later sections, we use transformations to translate Gaussian distributions $\mathcal{N}(\mu, \Lambda)$, in which case

$$\tau(\mu, \Lambda; \rho) = (\mu + \rho, \Lambda), \qquad \tilde{\tau}(v; \rho) = v - \rho. \tag{3}$$

Our learning algorithms use this relationship to efficiently combine information from images depicting scale-normalized objects at varying locations. For more complex datasets, we could instead employ a family of invertible affine transformations (see Sect. 5.2.2 of Sudderth 2006).

Transformations have been previously used to learn mixture models which decompose video sequences into a fixed number of layers (Frey and Jojic 2003; Jojic and Frey 2001). In contrast, the hierarchical models developed in this paper allow transformed mixture components to be shared among different object and scene categories. Nonparametric density estimates of transformations (Miller et al. 2000; Miller and Chefd'hotel 2003), and tangent approximations to transformation manifolds (Simard et al. 1998), have also been used to construct improved template-based recognition systems from small datasets. By embedding transformations in a nonparametric hierarchical model, we parse more

Fig. 4 Scale-normalized images used to evaluate two-dimensional models for visual scenes, available from the MIT LabelMe database (Russell et al. 2005). *Top*: Five of 613 images from a partially labeled dataset of street scenes, and segmented regions corresponding to cars (*red*), buildings (*magenta*), roads (*blue*), and trees (*green*). *Bottom*: Six of 315 images from a fully labeled dataset of office scenes, and segmented regions corresponding to computer screens (*red*), keyboards (*green*), and mice (*blue*)

complex visual scenes in which the number of objects is uncertain.

3.2 Fixed-Order Models for Isolated Objects

We begin by developing a parametric, hierarchical model for images dominated by a single object (Sudderth et al. 2005). The representation of objects as a collection of spatially constrained parts has a long history in vision (Fischler and Elschlager 1973). In the directed graphical model of Fig. 3, parts are formalized as groups of features that are spatially clustered, and have predictable appearances. Each of the L object categories is in turn characterized by a probability distribution $\boldsymbol{\pi}_\ell$ over a common set of K shared parts. For this *fixed-order* object appearance model, K is set to some known, constant value.

Given an image j of object category o_j containing N_j features $(\mathbf{w}_j, \mathbf{v}_j)$, we model feature positions relative to an image-specific *reference transformation*, or coordinate frame, ρ_j. For datasets in which objects are roughly scale-normalized and centered, unimodal Gaussian distributions $\rho_j \sim \mathcal{N}(\zeta_{o_j}, \Upsilon_{o_j})$ provide reasonable transformation priors. To capture the internal structure of objects, we define K distinct parts which generate features with different typical appearance w_{ji} and position v_{ji}, relative to ρ_j. The particular parts $\mathbf{z}_j = (z_{j1}, \ldots, z_{jN_j})$ associated with each feature are independently sampled from a category-specific multinomial distribution, so that $z_{ji} \sim \boldsymbol{\pi}_{o_j}$.

When learning object models from training data, we assign Dirichlet priors $\boldsymbol{\pi}_\ell \sim \mathrm{Dir}(\alpha)$ to the part association probabilities. Each part is then defined by a multinomial distribution η_k on the discrete set of W appearance descriptors, and a Gaussian distribution $\mathcal{N}(\mu_k, \Lambda_k)$ on the relative displacements of features from the object's transformed pose:

$$w_{ji} \sim \eta_{z_{ji}}, \qquad v_{ji} \sim \mathcal{N}(\tau(\mu_{z_{ji}}, \Lambda_{z_{ji}}; \rho_j)). \tag{4}$$

For datasets which have been normalized to account for orientation and scale variations, transformations are defined to shift the part's mean as in (3). In principle, however, the model could be easily generalized to capture more complex object pose variations.

Marginalizing the unobserved assignments z_{ji} of features to parts, we find that the graph of Fig. 3 defines object appearance via a finite mixture model:

$$p(w_{ji}, v_{ji}|\rho_j, o_j = \ell) = \sum_{k=1}^{K} \pi_{\ell k} \eta_k(w_{ji}) \mathcal{N}(v_{ji}; \tau(\mu_k, \Lambda_k; \rho_j)). \quad (5)$$

Parts are thus latent variables which capture dependencies in feature location and appearance, while reference transformations allow a common set of parts to model unaligned images. Removing these transformations, we recover a variant of the *author-topic model* (Rosen-Zvi et al. 2004), where objects correspond to authors, features to words, and parts to the latent topics underlying a given text corpus. The LDA model (Blei et al. 2003) is in turn a special case in which each document (image) has its own topic distribution, and authors (objects) are not explicitly modeled.

The fixed-order model of Fig. 3 shares information in two distinct ways: parts combine the same features in different spatial configurations, and objects reuse the same parts in different proportions. To learn the parameters defining these parts, we employ a Gibbs sampling algorithm (Griffiths and Steyvers 2004; Rosen-Zvi et al. 2004), which Sect. 6.2 develops in the context of a related model for multiple object scenes. This Monte Carlo method may either give each object category its own parts, or "borrow" parts from other objects, depending on the structure of the given training images.

3.3 Related Part-Based Object Appearance Models

In independent work paralleling the original development of our fixed-order object appearance model (Sudderth et al. 2005), two other papers have used finite mixture models to generate image features (Fergus et al. 2005; Loeff et al. 2006). However, these approaches model each category independently, rather than sharing parts among them. In addition, they use discrete representations of transformations and feature locations. This choice makes it difficult to learn typical transformations, a key component of the contextual scene models developed in Sect. 6. More recently, Williams and Allan have pointed out connections between so-called *generative templates of features* (Williams and Allan 2006), like the model of Fig. 3, and probabilistic voting methods such as the implicit shape model (Leibe et al. 2004).

Applied to a single object category, our approach is also related to constellation models (Fischler and Elschlager 1973; Weber et al. 2000), and in particular Bayesian training methods which share hyperparameters among categories (Fei-Fei et al. 2004). However, constellation models assume each part generates at most one feature, creating a combinatorial data association problem for which greedy approximations are needed (Helmer and Lowe 2004). In contrast, our model associates parts with expected *proportions* of the observed features. This allows several different features to provide evidence for a given part, and seems better matched to the dense, overlapping feature sets described in Sect. 2.1. Furthermore, by not placing hard constraints on the number of features assigned to each part, we develop simple learning algorithms which scale linearly, rather than exponentially, with the number of parts.

4 Sharing Parts using Nonparametric Hierarchical Models

When modeling complex datasets, it can be hard to determine an appropriate number of clusters for parametric models like LDA. As this choice significantly affects performance (Blei et al. 2003; Teh et al. 2006; Griffiths and Steyvers 2004; Fei-Fei and Perona 2005), it is interesting to explore nonparametric alternatives. In Bayesian statistics, Dirichlet processes (DPs) avoid model selection by defining priors on *infinite* models. Learning algorithms then produce robust predictions by averaging over model substructures whose complexity is justified by observed data. The following sections briefly review properties of DPs, and then adapt the hierarchical DP (Teh et al. 2006) to learn nonparametric, shared representations of multiple object categories. For more detailed introductions to Dirichlet processes and classical references, see (Pitman 2002; Jordan 2005; Teh et al. 2006; Sudderth 2006).

4.1 Dirichlet Process Mixtures

Let H be a measure on some parameter space Θ, like the conjugate priors of Appendix 1. A Dirichlet process (DP), denoted by $\mathrm{DP}(\gamma, H)$, is then a distribution over measures on Θ, where the scalar *concentration parameter* γ controls the similarity of samples $G \sim \mathrm{DP}(\gamma, H)$ to the base measure H. Analogously to Gaussian processes, DPs may be characterized by the distribution they induce on finite, measurable partitions $(T_1, \ldots, T_\ell)$ of Θ. In particular, for any such partition, the random vector $(G(T_1), \ldots, G(T_\ell))$ has a finite-dimensional Dirichlet distribution:

$$(G(T_1), \ldots, G(T_\ell)) \sim \mathrm{Dir}(\gamma H(T_1), \ldots, \gamma H(T_\ell)). \quad (6)$$

Samples from DPs are discrete with probability one, a property highlighted by the following *stick-breaking construction* (Pitman 2002; Ishwaran and James 2001):

$$G(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k), \qquad \beta'_k \sim \mathrm{Beta}(1, \gamma), \qquad \beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell). \quad (7)$$

Each parameter $\theta_k \sim H$ is independently sampled from the base measure, while the weights $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ use

beta random variables to partition a unit-length "stick" of probability mass. Following standard terminology (Teh et al. 2006; Pitman 2002), let $\beta \sim \text{GEM}(\gamma)$ denote a sample from this stick-breaking process. As γ becomes large, $\mathbb{E}[\beta'_k] = 1/(1+\gamma)$ approaches zero, and G approaches H by uniformly distributing probability mass among a densely sampled set of discrete parameters $\{\theta_k\}_{k=1}^{\infty}$.

DPs are commonly used as prior distributions for mixture models with an unknown, and potentially infinite, number of components (Escobar and West 1995; Neal 2000). Given $G \sim \text{DP}(\gamma, H)$, each observation x_i is generated by first choosing a parameter $\bar{\theta}_i \sim G$, and then sampling $x_i \sim F(\bar{\theta}_i)$. Note that we use θ_k to denote the *unique* parameters associated with distinct mixture components, and $\bar{\theta}_i$ to denote a *copy* of one such parameter associated with a particular observation x_i. For moderate concentrations γ, all but a random, finite subset of the mixture weights $\boldsymbol{\beta}$ are nearly zero, and data points cluster as in finite mixture models. In fact, mild conditions guarantee that DP mixtures provide consistent parameter estimates for finite mixture models of arbitrary order (Ishwaran and Zarepour 2002).

To develop computational methods, we let $z_i \sim \boldsymbol{\beta}$ indicate the unique component of $G(\theta)$ associated with observation $x_i \sim F(\theta_{z_i})$. Marginalizing G, these assignments $\mathbf{z}$ demonstrate an important clustering behavior (Pitman 2002). Letting N_k denote the number of observations already assigned to θ_k,

$$p(z_i | z_1, \ldots, z_{i-1}, \gamma) = \frac{1}{\gamma + i - 1}\left[\sum_k N_k \delta(z_i, k) + \gamma \delta(z_i, \bar{k})\right]. \tag{8}$$

Here, $\bar{k}$ indicates a previously unused mixture component (*a priori*, all clusters are equivalent). This process is sometimes described by analogy to a Chinese restaurant in which the (infinite collection of) tables correspond to the mixture components θ_k, and customers to observations x_i (Teh et al. 2006; Pitman 2002). Customers are social, tending to sit at tables with many other customers (observations), and each table shares a single dish (parameter). This clustering bias leads to Monte Carlo methods (Escobar and West 1995; Neal 2000) which infer the number of mixture components underlying a set of observations.

4.2 Modeling Objects with Hierarchical Dirichlet Processes

Standard Dirichlet process mixtures model observations via a single, infinite set of clusters. The *hierarchical Dirichlet process* (HDP) (Teh et al. 2006) instead shares infinite mixtures among several groups of data, thus providing a nonparametric generalization of LDA. In this section, we augment the HDP with image-specific spatial transformations, and thereby model unaligned sets of image features.

As discussed in Appendix 1, let H_w denote a Dirichlet prior on feature appearance distributions, H_v a normal-inverse-Wishart prior on feature position distributions, and $H_w \times H_v$ the corresponding product measure. To construct an HDP, a global probability measure $G_0 \sim \text{DP}(\gamma, H_w \times H_v)$ is first used to define an infinite set of shared parts:

$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k), \qquad \boldsymbol{\beta} \sim \text{GEM}(\gamma), \qquad (\eta_k, \mu_k, \Lambda_k) = \theta_k \sim H_w \times H_v. \tag{9}$$

For each object category $\ell = 1, \ldots, L$, an object-specific reweighting of these parts $G_\ell \sim \text{DP}(\alpha, G_0)$ is independently sampled from a DP with discrete base measure G_0, so that

$$G_\ell(\theta) = \sum_{t=1}^{\infty} \tilde{\pi}_{\ell t} \delta(\theta, \tilde{\theta}_{\ell t}), \qquad \tilde{\boldsymbol{\pi}}_\ell \sim \text{GEM}(\alpha), \qquad \tilde{\theta}_{\ell t} \sim G_0, \quad t = 1, 2, \ldots. \tag{10}$$

Each *local* part t (see (10)) has parameters $\tilde{\theta}_{\ell t}$ copied from some *global* part $\theta_{k_{\ell t}}$, indicated by $k_{\ell t} \sim \boldsymbol{\beta}$. Aggregating the probabilities associated with these copies, we can also directly express each object's appearance via the distinct, global parts:

$$G_\ell(\theta) = \sum_{k=1}^{\infty} \pi_{\ell k} \delta(\theta, \theta_k), \quad \pi_{\ell k} = \sum_{t | k_{\ell t} = k} \tilde{\pi}_{\ell t}. \tag{11}$$

Using (6), it can be shown that $\boldsymbol{\pi}_\ell \sim \text{DP}(\alpha, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_\ell$ are interpreted as measures on the positive integers (Teh et al. 2006). Thus, $\boldsymbol{\beta}$ determines the average importance of each global part ($\mathbb{E}[\pi_{\ell k}] = \beta_k$), while α controls the degree to which parts are reused across object categories.

Consider the generative process shown in Fig. 5 for an image j depicting object category o_j. As in the fixed-order model of Sect. 3.2, each image has a reference transformation ρ_j sampled from a Gaussian with normal-inverse-Wishart prior $(\zeta_\ell, \Upsilon_\ell) \sim R$. Each feature (w_{ji}, v_{ji}) is generated by choosing a part $z_{ji} \sim \boldsymbol{\pi}_{o_j}$, and then sampling from that part's appearance and transformed position distributions, as in (4). Marginalizing these unobserved assignments of features to parts, object appearance is defined by an *infinite* mixture model:

$$p(w_{ji}, v_{ji} | \rho_j, o_j = \ell) = \sum_{k=1}^{\infty} \pi_{\ell k} \eta_k(w_{ji}) \mathcal{N}(v_{ji}; \tau(\mu_k, \Lambda_k; \rho_j)). \tag{12}$$

This approach generalizes the parametric, fixed-order object model of Fig. 3 by defining an infinite set of potential global

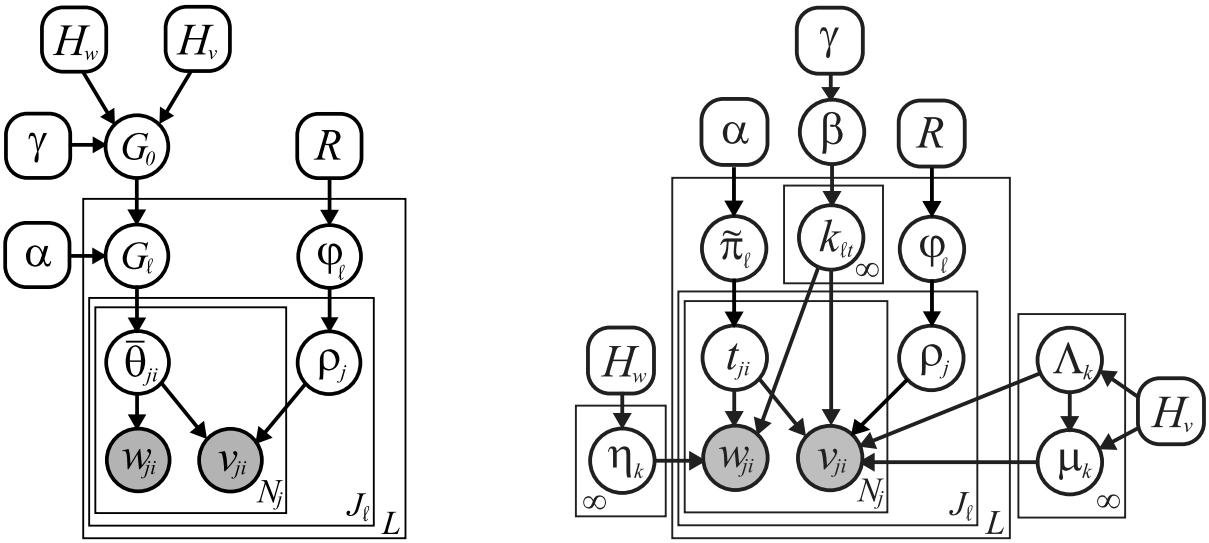

Fig. 5 Nonparametric, hierarchical DP model for the visual appearance of L object categories. The generative process is as in Fig. 3, except there are infinitely many potential parts. *Left*: Each of the J_ℓ images of object ℓ has a reference transformation $\rho_j \sim \mathcal{N}(\zeta_\ell, \Upsilon_\ell)$, where $\varphi_\ell = (\zeta_\ell, \Upsilon_\ell)$. $G_0 \sim \mathrm{DP}(\gamma, H_w \times H_v)$ then defines an infinite set of global parts, and objects reuse those parts via the reweighted distribution $G_\ell \sim \mathrm{DP}(\alpha, G_0)$. $\bar{\theta}_{ji} \sim G_\ell$ are then the part parameters used to generate feature (w_{ji}, v_{ji}). *Right*: Equivalent, Chinese restaurant franchise representation of the HDP. The explicit assignment variables $k_{\ell t}$, t_{ji} are used in Gibbs sampling algorithms (see Sect. 4.3)

parts, and using the Dirichlet process' stick-breaking prior to automatically choose an appropriate model order. It also extends the original HDP (Teh et al. 2006) by associating a different reference transformation with each training image.

The HDP follows an extension of the DP analogy known as the *Chinese restaurant franchise* (Teh et al. 2006). In this interpretation, each object or group defines a separate restaurant in which customers (observed features) (w_{ji}, v_{ji}) sit at tables (clusters or parts) t_{ji}. Each table shares a single dish (parameter) $\tilde{\theta}_{\ell t}$, which is ordered from a menu G_0 shared among restaurants (objects). Let $\mathbf{k}_\ell = \{k_{\ell t}\}$ denote the global parts assigned to all tables (local parts) of category ℓ. We may then integrate over G_0 and G_ℓ, as in (8), to find the conditional distributions of these assignment variables:

$$p(t_{ji}|t_{j1}, \ldots, t_{ji-1}, \alpha) \propto \sum_t N_{jt}\delta(t_{ji}, t) + \alpha\delta(t_{ji}, \bar{t}), \quad (13)$$

$$p(k_{\ell t}|\mathbf{k}_1, \ldots, \mathbf{k}_{\ell-1}, k_{\ell 1}, \ldots, k_{\ell t-1}, \gamma) \propto \sum_k M_k\delta(k_{\ell t}, k) + \gamma\delta(k_{\ell t}, \bar{k}). \quad (14)$$

Here, M_k is the number of tables previously assigned to θ_k, and N_{jt} the number of customers already seated at the t^{th} table in group j. As before, customers prefer tables t at which many customers are already seated (see (13)), but sometimes choose a new table $\bar{t}$. Each new table is assigned a dish $k_{\ell\bar{t}}$ according to (14). Popular dishes are more likely to be ordered, but a new dish $\theta_{\bar{k}} \sim H$ may also be selected. In this way, object categories sometimes reuse parts from other objects, but may also create a new part capturing distinctive appearance features.

4.3 Gibbs Sampling for Hierarchical Dirichlet Processes

To develop a learning algorithm for our HDP object appearance model, we consider the Chinese restaurant franchise representation, and generalize a previously proposed HDP Gibbs sampler (Teh et al. 2006) to also resample reference transformations. As illustrated in Fig. 5, the Chinese restaurant franchise involves two sets of assignment variables. Object categories ℓ have infinitely many local parts (tables) t, which are assigned to global parts $k_{\ell t}$. Each observed feature, or customer, (w_{ji}, v_{ji}) is then assigned to some table t_{ji}. By sampling these variables, we dynamically construct part-based feature groupings, and share parts among object categories.

The proposed Gibbs sampler has three sets of state variables: assignments $\mathbf{t}$ of features to tables, assignments $\mathbf{k}$ of tables to global parts, and reference transformations $\boldsymbol{\rho}$ for each training image. In the first sampling stage, summarized in Algorithm 1, we consider each training image j in turn and resample its transformation ρ_j and feature assignments $\mathbf{t}_j$. The second stage, Algorithm 2, then examines each object category ℓ, and samples assignments $\mathbf{k}_\ell$ of local to global parts. At all times, the sampler maintains dynamic lists of those tables to which at least one feature is assigned, and the global parts associated with these tables. These lists grow when new tables or parts are randomly chosen, and

Given a previous reference transformation ${\rho_j}^{(t-1)}$, table assignments ${\mathbf{t}_j}^{(t-1)}$ for the N_j features in an image depicting object category $o_j = \ell$, and global part assignments ${\mathbf{k}_\ell}^{(t-1)}$ for that object's T_ℓ tables:

1. Set $\mathbf{t}_j = {\mathbf{t}_j}^{(t-1)}$, $\mathbf{k}_\ell = {\mathbf{k}_\ell}^{(t-1)}$, and sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N_j\}$. For each $i \in \{\tau(1), \ldots, \tau(N_j)\}$, sequentially resample feature assignment t_{ji} as follows:
   (a) Decrement $N_{\ell t_{ji}}$, and remove (w_{ji}, v_{ji}) from the cached statistics for its current part $k = k_{\ell t_{ji}}$:

$$C_{kw} \leftarrow C_{kw} - 1, \quad w = w_{ji},$$
$$(\hat{\mu}_k, \hat{\Lambda}_k) \leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \ominus (v_{ji} - {\rho_j}^{(t-1)})$$

   (b) For each of the K instantiated global parts, determine the predictive likelihood

$$f_k(w_{ji} = w, v_{ji}) = \left(\frac{C_{kw} + \lambda/W}{\sum_{w'} C_{kw'} + \lambda} \right) \cdot \mathcal{N}(v_{ji} - {\rho_j}^{(t-1)}; \hat{\mu}_k, \hat{\Lambda}_k).$$

   Also determine the likelihood $f_{\bar{k}}(w_{ji}, v_{ji})$ of a potential new part $\bar{k}$.
   (c) Sample a new table assignment t_{ji} from the following $(T_\ell + 1)$-dim. multinomial distribution:

$$t_{ji} \sim \sum_{t=1}^{T_\ell} N_{\ell t} f_{k_{\ell t}}(w_{ji}, v_{ji}) \delta(t_{ji}, t) + \frac{\alpha}{\gamma + \sum_k M_k} \left[\sum_{k=1}^{K} M_k f_k(w_{ji}, v_{ji}) + \gamma f_{\bar{k}}(w_{ji}, v_{ji}) \right] \delta(t_{ji}, \bar{t}).$$

   (d) If $t_{ji} = \bar{t}$, create a new table, increment T_ℓ, and sample

$$k_{\ell \bar{t}} \sim \sum_{k=1}^{K} M_k f_k(w_{ji}, v_{ji}) \delta(k_{\ell \bar{t}}, k) + \gamma f_{\bar{k}}(w_{ji}, v_{ji}) \delta(k_{\ell \bar{t}}, \bar{k}).$$

   If $k_{\ell \bar{t}} = \bar{k}$, create a new global part and increment K.
   (e) Increment $N_{\ell t_{ji}}$, and add (w_{ji}, v_{ji}) to the cached statistics for its new part $k = k_{\ell t_{ji}}$:

$$C_{kw} \leftarrow C_{kw} + 1, \quad w = w_{ji},$$
$$(\hat{\mu}_k, \hat{\Lambda}_k) \leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \oplus (v_{ji} - {\rho_j}^{(t-1)}).$$

2. Fix ${\mathbf{t}_j}^{(t)} = \mathbf{t}_j$, ${\mathbf{k}_\ell}^{(t)} = \mathbf{k}_\ell$. If any tables are empty ($N_{\ell t} = 0$), remove them and decrement T_ℓ.
3. Sample a new reference transformation ${\rho_j}^{(t)}$ as follows:
   (a) Remove ${\rho_j}^{(t-1)}$ from cached transformation statistics for object ℓ:

$$(\hat{\zeta}_\ell, \hat{\Upsilon}_\ell) \leftarrow (\hat{\zeta}_\ell, \hat{\Upsilon}_\ell) \ominus {\rho_j}^{(t-1)}.$$

   (b) Sample ${\rho_j}^{(t)} \sim \mathcal{N}(\chi_j, \Xi_j)$, a posterior distribution determined via (45) from the prior $\mathcal{N}(\rho_j; \hat{\zeta}_\ell, \hat{\Upsilon}_\ell)$, cached part statistics $\{\hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^K$, and feature positions $\mathbf{v}_j$.
   (c) Add ${\rho_j}^{(t)}$ to cached transformation statistics for object ℓ:

$$(\hat{\zeta}_\ell, \hat{\Upsilon}_\ell) \leftarrow (\hat{\zeta}_\ell, \hat{\Upsilon}_\ell) \oplus {\rho_j}^{(t)}.$$

4. For each $i \in \{1, \ldots, N_j\}$, update cached statistics for global part $k = k_{\ell t_{ji}}$ as follows:

$$(\hat{\mu}_k, \hat{\Lambda}_k) \leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \ominus (v_{ji} - {\rho_j}^{(t-1)}),$$
$$(\hat{\mu}_k, \hat{\Lambda}_k) \leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \oplus (v_{ji} - {\rho_j}^{(t)}).$$

Algorithm 1 First stage of the Rao–Blackwellized Gibbs sampler for the HDP object appearance model of Fig. 5. We illustrate the sequential resampling of all assignments $\mathbf{t}_j$ of features to tables (category-specific copies of global parts) in the j^{th} training image, as well as that image's coordinate frame ρ_j. For efficiency, we cache and recursively update statistics $\{\hat{\zeta}_\ell, \hat{\Upsilon}_\ell\}_{\ell=1}^L$ of each object's reference transformations, counts $N_{\ell t}$ of the features assigned to each table, and appearance and position statistics $\{C_{kw}, \hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^K$ for the instantiated global parts. The $\oplus$ and $\ominus$ operators update cached mean and covariance statistics as features are added or removed from parts (see Sect. 12.1). The final step ensures consistency of these statistics following reference transformation updates

Given the previous global part assignments $\mathbf{k}_\ell{}^{(t-1)}$ for the T_ℓ instantiated tables of object category ℓ, and fixed feature assignments $\mathbf{t}_j$ and reference transformations ρ_j for all images of that object:

1. Set $\mathbf{k}_\ell = \mathbf{k}_\ell{}^{(t-1)}$, and sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, T_\ell\}$. For each $t \in \{\tau(1), \ldots, \tau(T_\ell)\}$, sequentially resample global part assignment $k_{\ell t}$ as follows:
   (a) Decrement $M_{k_{\ell t}}$, and remove all features at table t from the cached statistics for part $k = k_{\ell t}$:

$$\begin{aligned} C_{kw} &\leftarrow C_{kw} - 1 && \text{for each } w \in \mathbf{w}_t \triangleq \{w_{ji} | t_{ji} = t\}, \\ (\hat{\mu}_k, \hat{\Lambda}_k) &\leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \ominus (v - \rho_j) && \text{for each } v \in \mathbf{v}_t \triangleq \{v_{ji} \mid t_{ji} = t\}. \end{aligned}$$

   (b) For each of the K instantiated global parts, determine the predictive likelihood

$$f_k(\mathbf{w}_t, \mathbf{v}_t) = p(\mathbf{w}_t | \{w_{ji} | k_{\ell t_{ji}} = k, t_{ji} \neq t\}, H_w) \cdot p(\mathbf{v}_t | \{v_{ji} | k_{\ell t_{ji}} = k, t_{ji} \neq t\}, H_v).$$

   Also determine the likelihood $f_{\bar{k}}(\mathbf{w}_t, \mathbf{v}_t)$ of a potential new part $\bar{k}$.
   (c) Sample a new part assignment $k_{\ell t}$ from the following $(K+1)$-dim. multinomial distribution:

$$k_{\ell t} \sim \sum_{k=1}^{K} M_k f_k(\mathbf{w}_t, \mathbf{v}_t) \delta(k_{\ell t}, k) + \gamma f_{\bar{k}}(\mathbf{w}_t, \mathbf{v}_t) \delta(k_{\ell t}, \bar{k}).$$

   If $k_{\ell t} = \bar{k}$, create a new global part and increment K.
   (d) Increment $M_{k_{\ell t}}$, and add all features at table t to the cached statistics for its new part $k = k_{\ell t}$:

$$\begin{aligned} C_{kw} &\leftarrow C_{kw} + 1 && \text{for each } w \in \mathbf{w}_t, \\ (\hat{\mu}_k, \hat{\Lambda}_k) &\leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \oplus (v - \rho_j) && \text{for each } v \in \mathbf{v}_t. \end{aligned}$$

2. Fix $\mathbf{k}_\ell{}^{(t)} = \mathbf{k}_\ell$. If any global parts are unused ($M_k = 0$), remove them and decrement K.
3. Given gamma priors, resample concentration parameters γ and α using auxiliary variables (Escobar and West 1995; Teh et al. 2006).

Algorithm 2 Second stage of the Rao–Blackwellized Gibbs sampler for the HDP object appearance model of Fig. 5. We illustrate the sequential resampling of all assignments $\mathbf{k}_\ell$ of tables (category-specific parts) to global parts for the ℓ^{th} object category, as well as the HDP concentration parameters. For efficiency, we cache and recursively update appearance and position statistics $\{C_{kw}, \hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^K$ for the instantiated global parts, and counts M_k of the number of tables assigned to each part. The $\oplus$ and $\ominus$ operators update cached mean and covariance statistics as features are reassigned (see Sect. 12.1)

shrink when a previously occupied table or part no longer has assigned features. Given K instantiated global parts, the expected time to resample N features is $\mathcal{O}(NK)$.

We provide high-level derivations for the sampling updates underlying Algorithms 1 and 2 in Sect. 12.1. Note that our sampler *analytically* marginalizes (rather than samples) the weights $\boldsymbol{\beta}, \tilde{\boldsymbol{\pi}}_\ell$ assigned to global and local parts, as well as the parameters θ_k defining each part's feature distribution. Such *Rao–Blackwellization* is guaranteed to reduce the variance of Monte Carlo estimators (Sudderth 2006; Casella and Robert 1996).

5 Sixteen Object Categories

To explore the benefits of sharing parts among objects, we consider a collection of 16 categories with noticeable visual similarities. Figure 6 shows images from each category, which fall into three groups: seven animal faces, five animal profiles, and four wheeled vehicles. While training images are labeled with their category, we do *not* explicitly modify our part-based models to reflect these coarser groupings. As recognition systems scale to applications involving hundreds of objects, the inter-category similarities exhibited by this dataset will become increasingly common.

5.1 Visualization of Shared Parts

Given 30 training images from each of the 16 categories, we first extracted Harris-affine (Mikolajczyk and Schmid 2004) and MSER (Matas et al. 2002) interest regions as in Sect. 2.1, and mapped SIFT descriptors (Lowe 2004) to one of $W = 600$ visual words as in Sect. 2.2. We then used the Gibbs sampler of Algorithms 1 and 2 to fit an HDP object appearance model. Because our 16-category

Figure 6 Example images from a dataset containing 16 object categories (*columns*), available from the MIT LabelMe database (Russell et al. 2005). These categories combines images collected via web searches with the Caltech 101 (Fei-Fei et al. 2004) and Weizmann Institute (Ullman et al. 2002; Borenstein and Ullman 2002) datasets. Including a complementary background category, there are a total of 1,885 images, with at least 50 images per category

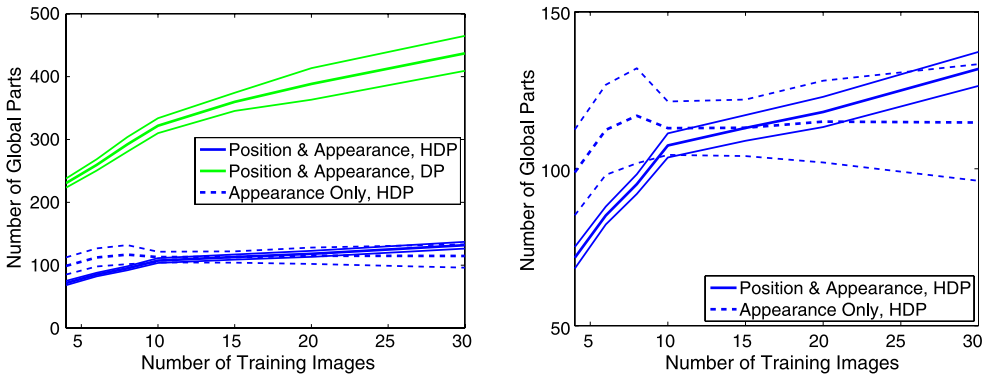

Figure 7 Mean (*thick lines*) and variance (*thin lines*) of the number of global parts created by the HDP Gibbs sampler (Sect. 4.3), given training sets of varying size. *Left*: Number of global parts used by HDP object models (*blue*), and the total number of parts instantiated by sixteen independent DP object models (*green*). *Right*: Expanded view of the parts instantiated by the HDP object models

dataset contains approximately aligned images, the reference transformation updates of Algorithm 1, steps 3–4 were not needed. Later sections explore transformations in the context of more complex scene models.

For our Matlab implementation, each sampling iteration requires roughly 0.1 seconds per training image on a 3.0 GHz Intel Xeon processor. Empirically, the learning procedure is fairly robust to hyperparameters; we chose H_υ to provide a weak ($\upsilon = 6$ degrees of freedom) bias towards moderate covariances, and $H_w = \mathrm{Dir}(W/10)$ to favor sparse appearance distributions. Concentration parameters were assigned weakly informative priors $\gamma \sim \mathrm{Gamma}(5, 0.1)$, $\alpha \sim \mathrm{Gamma}(0.1, 0.1)$, allowing data-driven estimation of appropriate numbers of global and local parts.

We ran the Gibbs sampler for 1000 iterations, and used the final assignments $(\mathbf{t}, \mathbf{k})$ to estimate the feature appearance and position distributions for each part. After an initial burn-in phase, there were typically between 120 and 140 global parts associated with at least one observation (see Fig. 7). Figure 8 visualizes the feature distributions defining seven of the more significant parts. A few seem specialized to distinctive features of individual categories, such as the spots appearing on the leopard's forehead. Many other parts are shared among several categories, modeling common aspects such as ears, mouths, and wheels. We also show one of several parts which model background clutter around image boundaries, and are widely shared among categories.

To further investigate these shared parts, we used the symmetrized KL divergence, as in (Rosen-Zvi et al. 2004), to compute a distance between all pairs of object-specific part distributions:

$$D(\boldsymbol{\pi}_\ell, \boldsymbol{\pi}_m) = \sum_{k=1}^{K} \pi_{\ell k} \log \frac{\pi_{\ell k}}{\pi_{mk}} + \pi_{mk} \log \frac{\pi_{mk}}{\pi_{\ell k}}. \tag{15}$$

In evaluating equation (15), we only use parts associated with at least one feature. Figure 9 shows the two-dimensional embedding of these distances produced by metric multidimensional scaling (MDS), as well as a dendrogram constructed via greedy, agglomerative clustering (Shepard 1980). Interestingly, there is significant sharing of parts within each of the three coarse-level groups (animal faces, animal profiles, vehicles) underlying this dataset. In addition, the similarities among the three categories of cat faces, and among those animals with elongated faces, are reflected in the shared parts.

5.2 Detection and Recognition Performance

To evaluate our HDP object appearance model, we consider two experiments. The *detection* task uses 100 images of natural scenes to train a DP background appearance model. We then use likelihoods computed as in Sect. 12.1 to classify test images as object or background. Alternatively, in the *recognition* task test images are classified

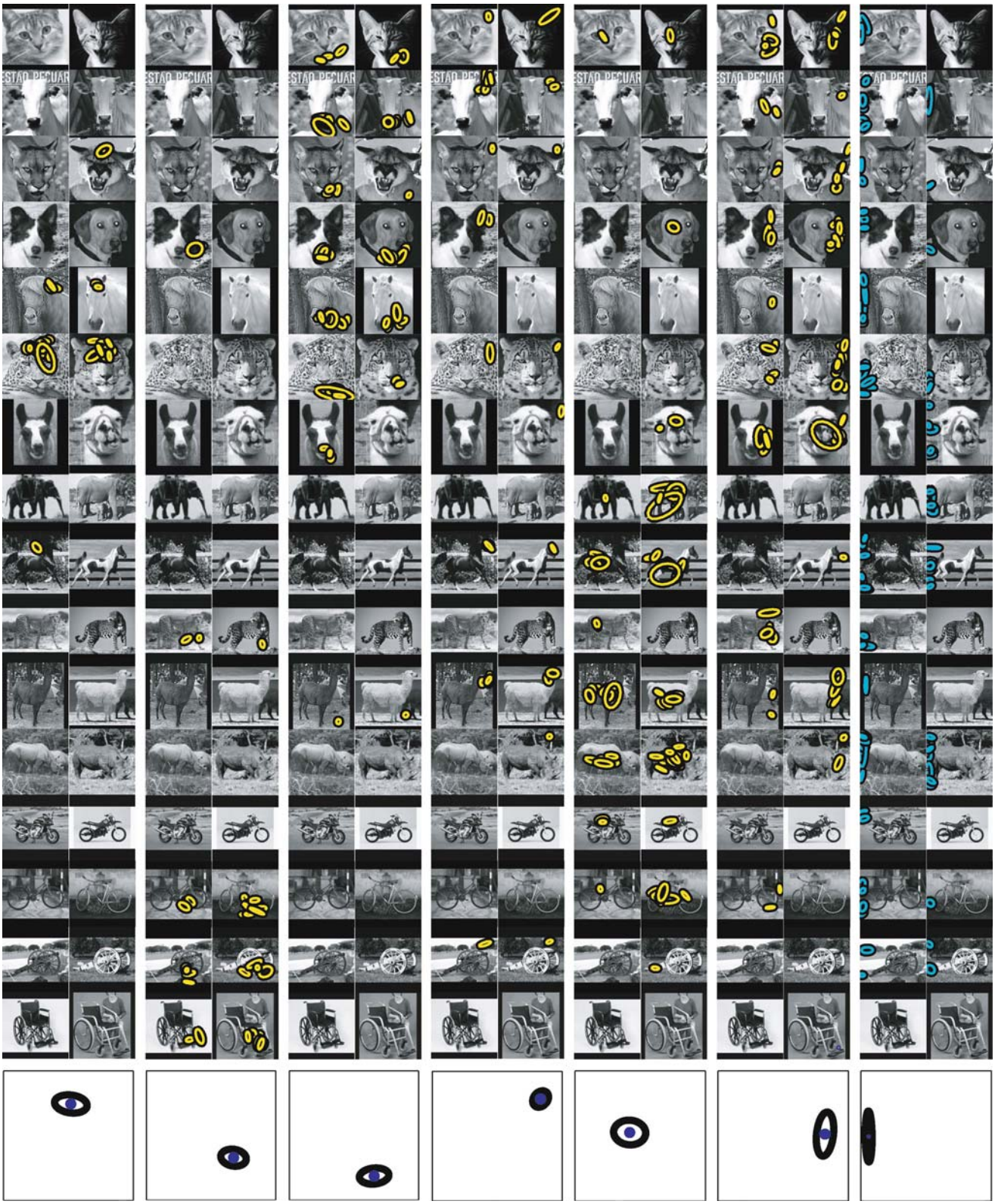

Figure 8 Seven of the 135 shared parts (*columns*) learned by an HDP model for 16 object categories (*rows*). Using two images from each category, we display those features with the highest posterior probability of being generated by each part. For comparison, we show six of the parts which are specialized to the fewest object categories (*left, yellow*), as well as one of several widely shared parts (*right, cyan*) which seem to model texture and background clutter. The *bottom row* plots the Gaussian position densities corresponding to each part. Interestingly, several parts have rough semantic interpretations, and are shared within the coarse-level object groupings underlying this dataset

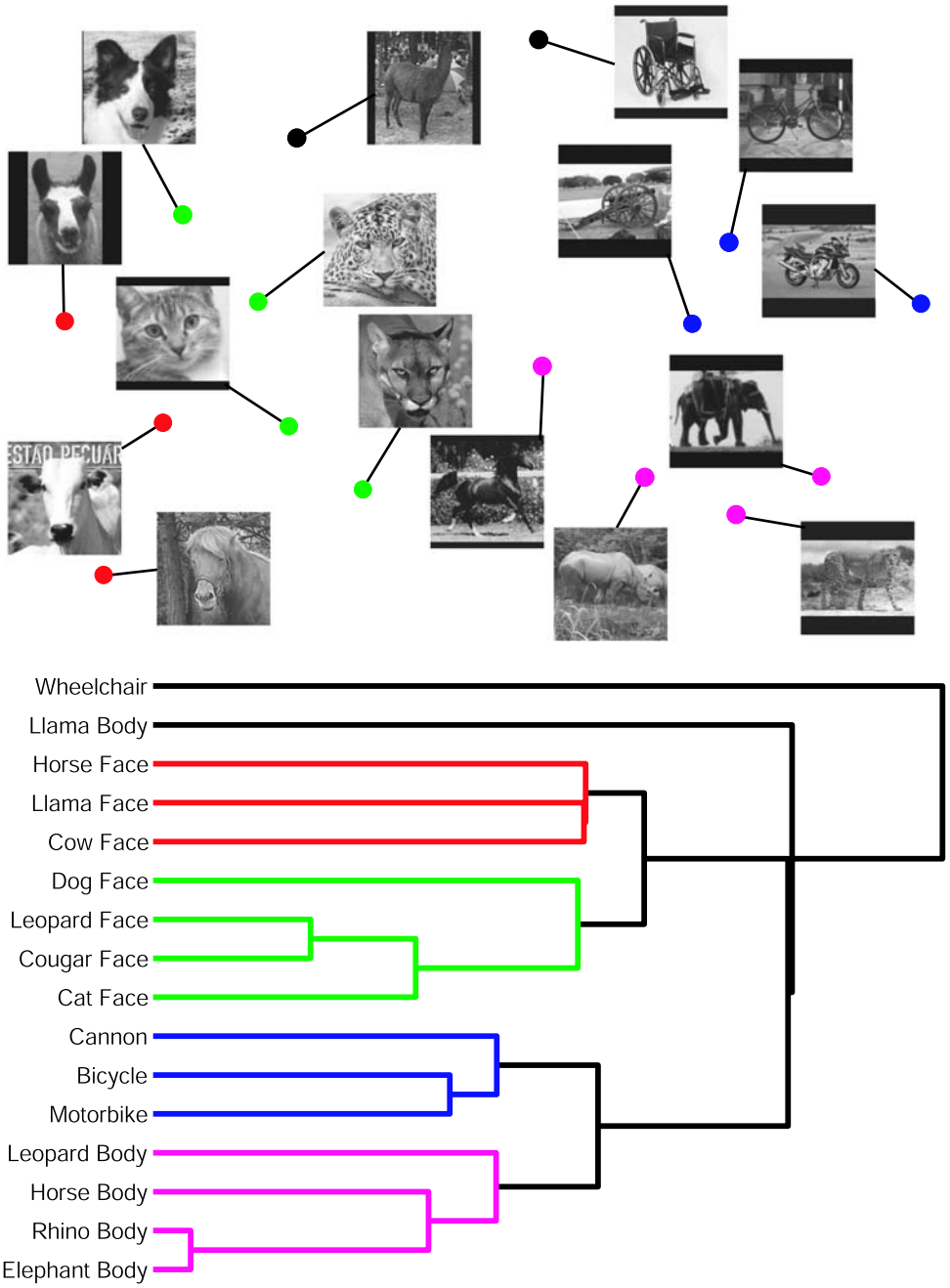

Figure 9 Two visualizations of learned part distributions $\boldsymbol{\pi}_\ell$ for the HDP object appearance model depicted in Fig. 8. *Top*: Two-dimensional embedding computed by metric MDS, in which coordinates for each object category are chosen to approximate pairwise KL distances as in (15). Animal faces are clustered *on the left*, vehicles in the *upper right*, and animal profiles in the *lower right*. *Bottom*: Dendrogram illustrating a greedy, hierarchical clustering, where branch lengths are proportional to inter-category distances. The four most significant clusters, which very intuitively align with semantic relationships among these categories, are highlighted in color

as either their true category, or one of the 15 other categories. For both tasks, we compare a *shared* model of all objects to a set of 16 *unshared*, independent DP models trained on individual categories. We also examine simplified models which ignore the spatial location of features, as in earlier bag of features approaches (Sivic et al. 2005; Csurka et al. 2004). We evaluate performance via the area under receiver operating characteristic (ROC) curves, and use nonparametric rank-sum tests (DeLong et al. 1988) to determine whether competing models differ with at least 95% confidence.

In Fig. 7, we illustrate the number of global parts instantiated by the HDP Gibbs sampler. The appearance-only HDP model learns a consistent number of parts given between 10 and 30 training images, while the HDP model of feature positions uses additional parts as more images are observed. Such data-driven growth in model complexity underlies many desirable properties of Dirichlet processes (Sudderth 2006; Jordan 2005; Ishwaran and Zarepour 2002). We also show the considerably larger number of total parts (roughly 25 per category) employed by the independent DP models of feature positions. Because we use multinomial appearance distributions, estimation of the number of parts for the

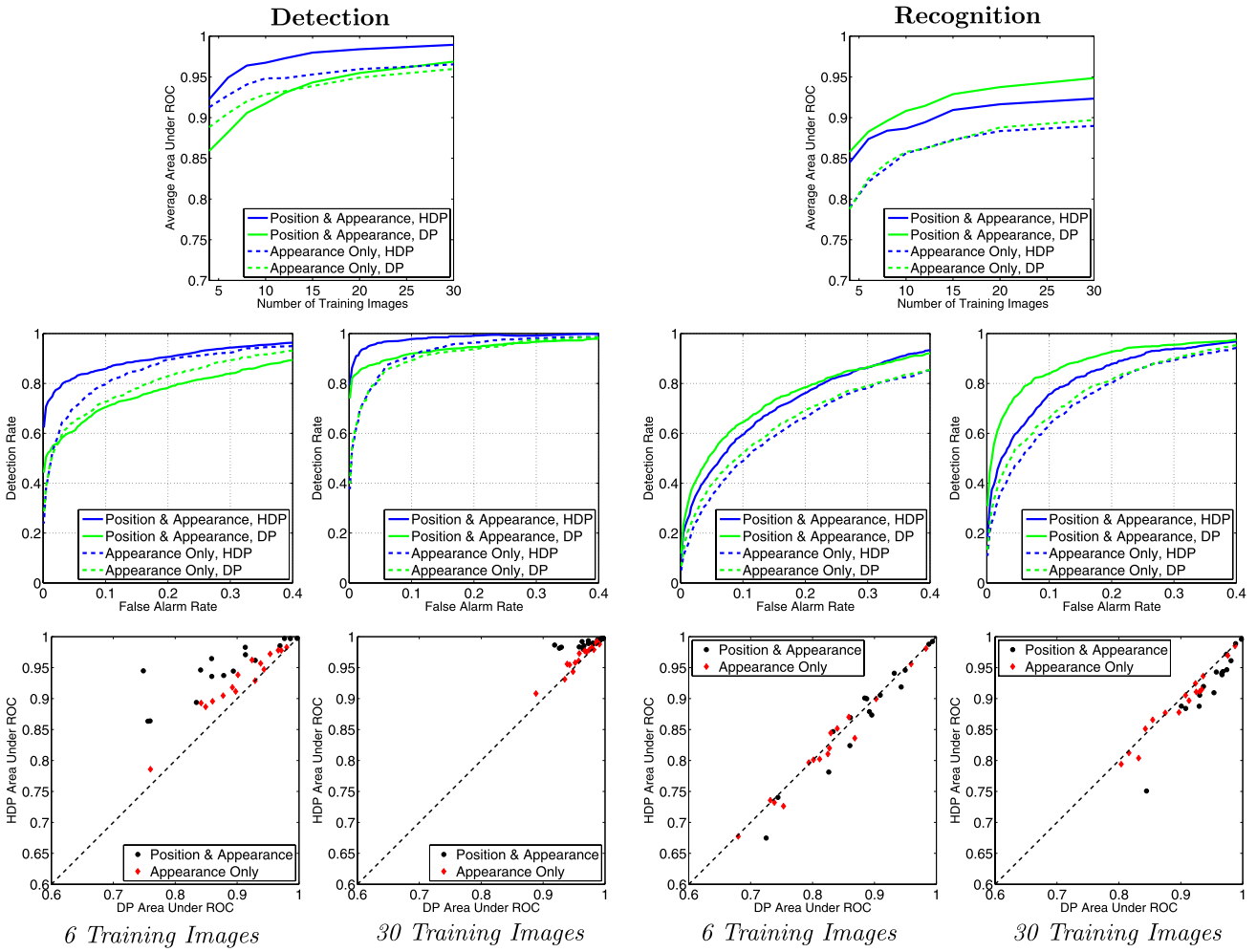

Figure 10 Performance of Dirichlet process object appearance models for the detection (*left*) and recognition (*right*) tasks. *Top*: Area under average ROC curves for different numbers of training images per category. *Middle*: Average of ROC curves across all categories (6 versus 30 training images). *Bottom*: Scatter plot of areas under ROC curves for the shared and unshared models of individual categories (6 versus 30 training images)

DP appearance-only model is ill-posed, and very sensitive to H_w; we thus exclude this model from Fig. 7.

Figure 10 shows detection and recognition performance given between 4 and 30 training images per category. Likelihoods are estimated from 40 samples extracted across 1000 iterations. Given 6 training images, shared parts significantly improve position-based detection performance for all categories (see scatter plots). Even with 30 training images, sharing still provides significant benefits for 9 categories (for the other seven, both models are extremely accurate). For the bag of features model, the benefits of sharing are less dramatic, but still statistically significant in many cases. Finally, note that with fewer than 15 training images, the unshared position-based model overfits, performing significantly worse than comparable appearance-only models for most categories. In contrast, sharing spatial parts provides superior performance for all training set sizes.

For the recognition task, shared and unshared appearance-only models perform similarly. However, with larger training sets the HDP model of feature positions is less effective for most categories than unshared, independent DP models. Confusion matrices (not shown) confirm that this small performance degradation is due to errors involving pairs of object categories with similar part distributions (see Fig. 9). Note, however, that the unshared models use many more parts (see Fig. 7), and hence require additional computation. For all categories exhibiting significant differences, we find that models incorporating feature positions have significantly higher recognition accuracy.

5.3 Comparison to Fixed-Order Object Appearance Models

We now compare the HDP object model to the parametric, fixed-order model of Sect. 3.2. Images illustrating the parts

learned by the fixed-order model, which we exclude here due to space constraints, are available in Sect. 5.4 of (Sudderth 2006). Qualitatively, the fixed-order parts are similar to the HDP parts depicted in Fig. 8, except that there is more sharing among dissimilar object categories. This in turn leads to more overlap among part distributions, and inferred object relationships which are semantically less sensible than those found with the HDP (visualized in Fig. 9).

Previous results have shown that LDA can be sensitive to the chosen number of topics (Blei et al. 2003; Teh et al. 2006; Griffiths and Steyvers 2004; Fei-Fei and Perona 2005). To further explore this issue, we examined fixed-order object appearance models with between two and thirty parts per category (32–480 shared parts versus 16 unshared 2–30 part models). For each model order, we ran a collapsed Gibbs sampler (see Sect. 12.2) for 200 iterations, and categorized test images via probabilities based on six posterior samples. We first considered part association probabilities $\boldsymbol{\pi}_\ell$ learned using a symmetric Dirichlet prior:

$$(\pi_{\ell 1}, \ldots, \pi_{\ell K}) \sim \mathrm{Dir}(\bar{\alpha}, \ldots, \bar{\alpha}) = \mathrm{Dir}(\bar{\alpha} K). \tag{16}$$

Our experiments set $\bar{\alpha} = 5$, inducing a small bias towards distributions which assign some weight to each of the K parts. Figure 11 shows the average detection and recognition performance, as measured by the area under the ROC curve, for varying model orders. Even with 15 training images of each category, shared models with more than 4–6 parts per category (64–96 total parts) overfit and exhibit reduced accuracy. Similar issues arise when learning finite mixture models, where priors as in (16) may produce inconsistent parameter estimates if K is not selected with care (Ishwaran and Zarepour 2002).

In some applications of the LDA model, the number of topics K is determined via cross-validation (Blei et al. 2003; Griffiths and Steyvers 2004; Fei-Fei and Perona 2005). This approach is also possible with the fixed-order object appearance model, but in practice requires extensive computational effort. Alternatively, model complexity can be regulated by the following modified part association prior:

$$(\pi_{\ell 1}, \ldots, \pi_{\ell K}) \sim \mathrm{Dir}\left(\frac{\alpha_0}{K}, \ldots, \frac{\alpha_0}{K}\right) = \mathrm{Dir}(\alpha_0). \tag{17}$$

For a fixed precision α_0, this prior becomes biased towards sparse part distributions $\boldsymbol{\pi}_\ell$ as K grows large (Sudderth 2006). Figure 11 illustrates its behavior for $\alpha_0 = 10$. In contrast with the earlier overfitting, (17) produces stable recognition results across a wider range of model orders K.

As $K \to \infty$, predictions based on Dirichlet priors scaled as in (17) approach a corresponding Dirichlet process (Teh et al. 2006; Ishwaran and Zarepour 2002). However, if we apply this limit directly to the model of Fig. 3, objects asymptotically associate features with *disjoint* sets of parts, and the benefits of sharing are lost. We see the beginnings of this trend in Fig. 11, which shows a slow decline in detection

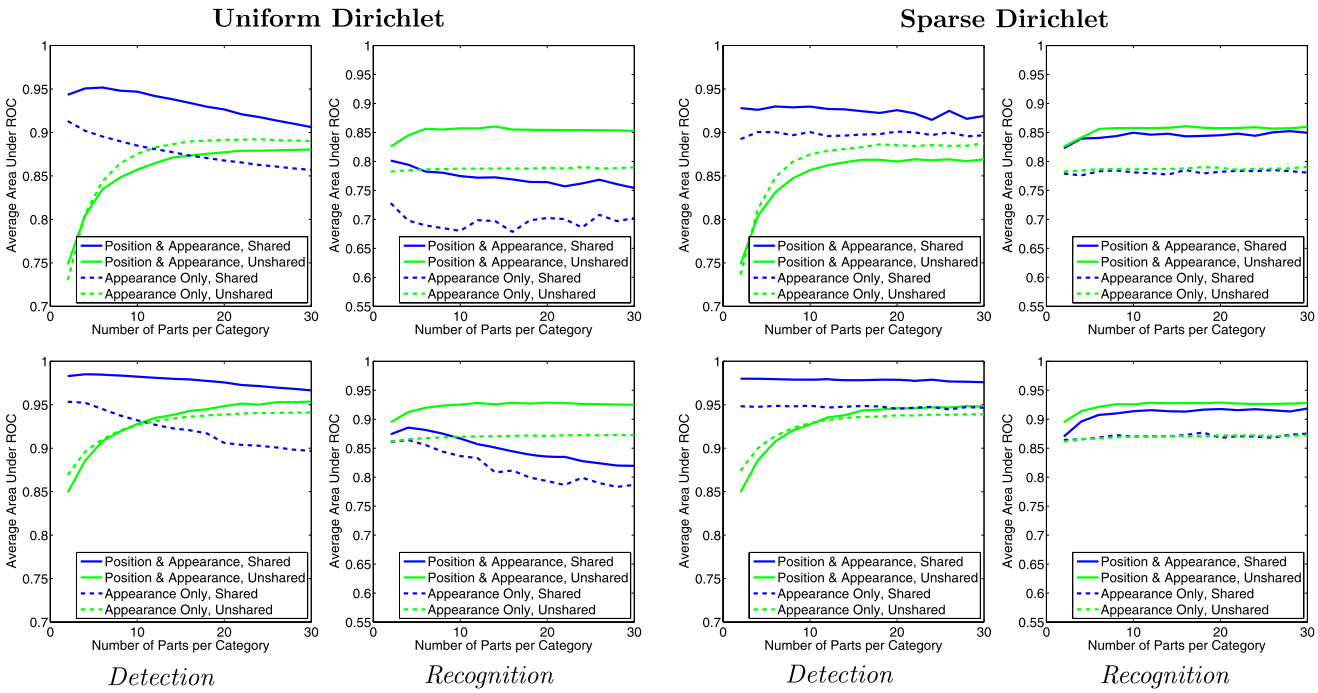

Figure 11 Performance of fixed-order object appearance models with varying numbers of parts K. Part association priors are either biased towards uniform distributions $\boldsymbol{\pi}_\ell \sim \mathrm{Dir}(\bar{\alpha} K)$ (*left block*, as in (16)), or sparse distributions $\boldsymbol{\pi}_\ell \sim \mathrm{Dir}(\alpha_0)$ (*right block*, as in (17)). We compare detection and recognition performance given 4 (*top row*) or 15 (*bottom row*) training images per category

performance as K increases. The HDP elegantly resolves this problem via the discrete global measure G_0, which explicitly couples the parts in different categories. Comparing Figs. 10 and 11, the HDP's detection and recognition performance is comparable to the *best* fixed-order model. Via a nonparametric viewpoint, however, the HDP leads to efficient learning methods which avoid model selection.

6 Contextual Models for Fixed Sets of Objects

The preceding results demonstrate the potential benefits of transferring information among object categories when learning from few examples. However, because the HDP model of Fig. 5 describes each image via a single reference transformation, it is limited to scenes which depict a single, dominant foreground object. In the following sections, we address this issue via a series of increasingly sophisticated models for *visual scenes* containing multiple objects.

6.1 Fixed-Order Models for Multiple Object Scenes

We begin by generalizing the fixed-order object appearance model of Sect. 3.2 to describe multiple object scenes (Sudderth et al. 2005). Retaining its parametric form, we assume that the scene s_j depicted in image j contains a fixed, *known* set of object categories. For example, a simple office scene might contain one computer screen, one keyboard, and one mouse. Later sections consider more flexible scene models, in which the number of object instances is also uncertain.

As summarized in Fig. 12, the scene transformation ρ_j provides a reference frame for each of L objects. For simplicity, we focus on scale-normalized datasets, so that ρ_j is a $2L$-dimensional vector specifying each object's image coordinates. Scene categories then have different Gaussian transformation distributions $\rho_j \sim \mathcal{N}(\zeta_{s_j}, \Upsilon_{s_j})$, with normal-inverse-Wishart priors $(\zeta_s, \Upsilon_s) \sim R$. Because these Gaussians have full, $2L$-dimensional covariance matrices, we learn contextual, scene-specific correlations in the locations at which objects are observed.

Visual scenes are also associated with discrete distributions $\boldsymbol{\beta}_s$ specifying the proportion of observed features generated by each object. Features are generated by sampling an object category $o_{ji} \sim \boldsymbol{\beta}_{s_j}$, and then a corresponding part $z_{ji} \sim \boldsymbol{\pi}_{o_{ji}}$. Conditioned on these assignments, the discrete appearance w_{ji} of each feature is independently sampled as in Sect. 3.2. Feature position v_{ji} is determined by shifting parts relative to the chosen object's reference transformation:

$$\begin{aligned} w_{ji} &\sim \eta_{z_{ji}}, \\ v_{ji} &\sim \mathcal{N}(\mu_{z_{ji}} + \rho_{j\ell}, \Lambda_{z_{ji}}), \qquad o_{ji} = \ell. \end{aligned} \tag{18}$$

Here, $\rho_{j\ell}$ is the subvector of ρ_j corresponding to the reference transformation for object ℓ. Marginalizing unobserved assignments z_{ji} of features to parts, we find that each object's appearance is defined by a different finite mixture

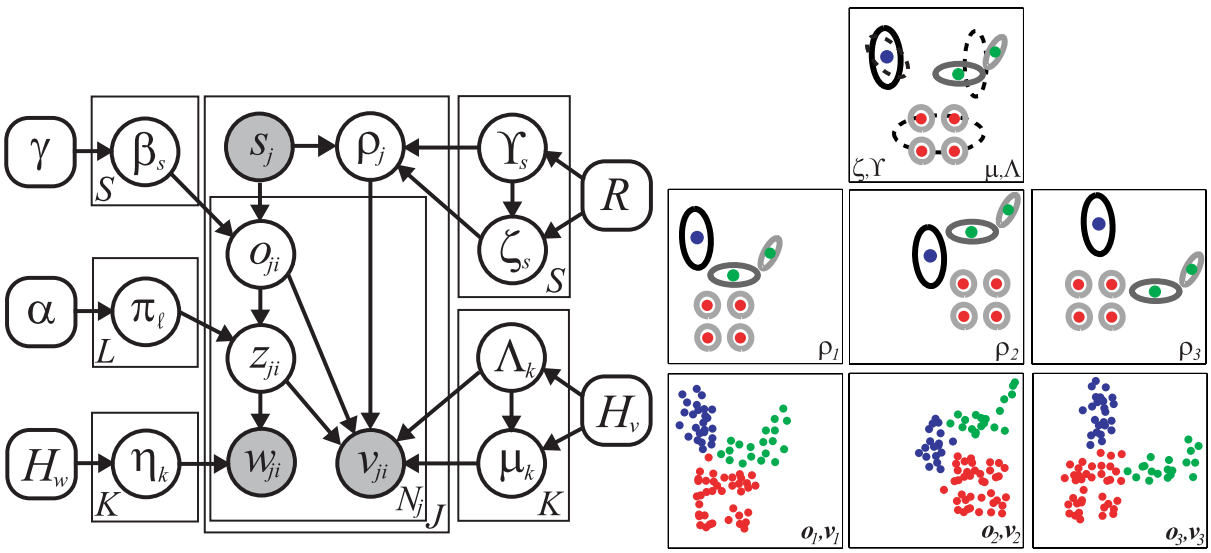

Figure 12 A parametric model for visual scenes containing fixed sets of objects. The j^{th} image depicts visual scene s_j, which combines L object categories at locations determined by the vector ρ_j of reference transformations. Each object category is in turn defined by a distribution $\boldsymbol{\pi}_\ell$ over a common set of K shared parts. The appearance w_{ji} and position v_{ji} of visual features, relative to the position of associated object o_{ji}, are then determined by assignments $z_{ji} \sim \boldsymbol{\pi}_{o_{ji}}$ to latent parts. The cartoon example defines $L = 3$ color-coded object categories, which employ one (*blue*), two (*green*), and four (*red*) of the shared Gaussian parts, respectively. Dashed ellipses indicate marginal transformation priors for each object, but the model also captures higher-order correlations in their relative spatial positions

model:

$$p(w_{ji}, v_{ji}|\rho_j, o_{ji} = \ell) = \sum_{k=1}^{K} \pi_{\ell k} \eta_k(w_{ji}) \mathcal{N}(v_{ji}; \mu_k + \rho_{j\ell}, \Lambda_k). \quad (19)$$

For scenes containing a single object, this model is equivalent to the fixed-order model of Sect. 3.2. More generally, however, (19) faithfully describes images containing several objects, which differ in their observed locations and underlying part-based decompositions. The graph of Fig. 12 generalizes the author-topic model (Rosen-Zvi et al. 2004) by incorporating reference transformations, and by not constraining objects (authors) to generate equal proportions of image features (words).

6.2 Gibbs Sampling for Fixed-Order Visual Scenes

Learning and inference in the scene-object-part hierarchy of Fig. 12 is possible via Monte Carlo methods similar to those developed for the HDP in Sect. 4.3. As summarized in Algorithm 3, our Gibbs sampler alternatively samples assignments (o_{ji}, z_{ji}) of features to objects and parts, and corresponding reference transformations ρ_j. This method, whose derivation is discussed in Sect. 12.2, generalizes a Gibbs sampler developed for the author-topic model (Rosen-Zvi et al. 2004). We have found sampling reference transformations to be faster than our earlier use of incremental EM updates (Sudderth et al. 2005; Sudderth 2006).

Given a training image containing N features, a Gibbs sampling update of every object and part assignment requires $\mathcal{O}(NLK)$ operations. Importantly, our use of Gaussian transformation distributions also allows us to *jointly* resample the positions of L objects in $\mathcal{O}(L^3)$ operations. We evaluate the performance of this contextual scene model in Sect. 9.1.

7 Transformed Dirichlet Processes

To model scenes containing an uncertain number of object *instances*, we again employ Dirichlet processes. Section 4 adapted the HDP to allow uncertainty in the number of parts underlying a set of object categories. We now develop a *transformed Dirichlet process* (TDP) which generalizes the HDP by applying a random *set* of transformations to each global cluster (Sudderth et al. 2006b). Section 8 then uses the TDP to develop robust nonparametric models for structured multiple object scenes.

7.1 Sharing Transformations via Stick-Breaking Processes

To simplify our presentation of the TDP, we revisit the hierarchical clustering framework underlying the HDP (Teh et al. 2006). Let $\theta \in \Theta$ parameterize a cluster or topic distribution $F(\theta)$, and H be a prior measure on Θ. To more flexibly share these clusters among related groups, we consider a family of parameter transformations $\tau(\theta; \rho)$, indexed by $\rho \in \wp$ as in Sect. 3.1. The TDP then employs *distributions over transformations* $\rho \sim Q(\varphi)$, with densities $q(\rho|\varphi)$ indexed by $\varphi \in \Phi$. For example, if ρ is a vector defining a translation as in (3), φ could parameterize a zero-mean Gaussian family $\mathcal{N}(\rho; 0, \varphi)$. Finally, let R denote a prior measure (for example, an inverse-Wishart distribution) on Φ.

We begin by extending the Dirichlet process' stick-breaking construction, as in (9), to define a global measure relating cluster parameters θ to transformations ρ:

$$G_0(\theta, \rho) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(\theta, \theta_\ell) q(\rho|\varphi_\ell), \quad \boldsymbol{\beta} \sim \text{GEM}(\gamma), \quad \theta_\ell \sim H, \quad \varphi_\ell \sim R. \quad (20)$$

Note that each global cluster θ_ℓ has a different, continuous transformation distribution $Q(\varphi_\ell)$. As in the HDP, we then independently draw $G_j \sim \text{DP}(\alpha, G_0)$ for each of J groups of data. Because samples from DPs are discrete with probability one, the joint measure for group j equals

$$G_j(\theta, \rho) = \sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta(\theta, \tilde{\theta}_{jt}) \delta(\rho, \rho_{jt}), \quad \tilde{\boldsymbol{\pi}}_j \sim \text{GEM}(\alpha), \quad (\tilde{\theta}_{jt}, \rho_{jt}) \sim G_0. \quad (21)$$

Each *local* cluster in group j has parameters $\tilde{\theta}_{jt}$, and corresponding transformation ρ_{jt}, derived from some global cluster. Anticipating our later identification of global clusters with object categories, we let $o_{jt} \sim \boldsymbol{\beta}$ indicate this correspondence, so that $\tilde{\theta}_{jt} = \theta_{o_{jt}}$. As summarized in Fig. 13, each observation v_{ji} is independently sampled from the *transformed* parameters of some local cluster:

$$(\bar{\theta}_{ji}, \bar{\rho}_{ji}) \sim G_j, \quad v_{ji} \sim F(\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji})). \quad (22)$$

As with standard mixtures, (22) can be equivalently expressed via a discrete variable $t_{ji} \sim \tilde{\boldsymbol{\pi}}_j$ indicating the transformed cluster associated with observation $v_{ji} \sim F(\tau(\tilde{\theta}_{jt_{ji}}; \rho_{jt_{ji}}))$. Figure 13 also shows an alternative graphical representation of the TDP, based on these explicit assignments of observations to local clusters, and local clusters to transformations of global clusters.

As discussed in Sect. 4.2, the HDP models groups by reusing an *identical* set of global clusters in different proportions. In contrast, the TDP modifies the shared, global

Given a previous reference transformation ${\rho_j}^{(t-1)}$, and object and part assignments $({\mathbf{o}_j}^{(t-1)}, {\mathbf{z}_j}^{(t-1)})$ for the N_j features in an image depicting scene $s_j = s$:

1. Set $(\mathbf{o}_j, \mathbf{z}_j) = ({\mathbf{o}_j}^{(t-1)}, {\mathbf{z}_j}^{(t-1)})$, and sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N_j\}$. For $i \in \{\tau(1), \ldots, \tau(N_j)\}$, sequentially resample feature assignments (o_{ji}, z_{ji}) as follows:
   (a) Remove feature (w_{ji}, v_{ji}) from the cached statistics for its current part and object:

$$\begin{aligned} M_{s\ell} &\leftarrow M_{s\ell} - 1, & \ell &= o_{ji}, \\ N_{\ell k} &\leftarrow N_{\ell k} - 1, & k &= z_{ji}, \\ C_{kw} &\leftarrow C_{kw} - 1, & w &= w_{ji}, \\ (\hat{\mu}_k, \hat{\Lambda}_k) &\leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \ominus (v_{ji} - {\rho_{j\ell}}^{(t-1)}). \end{aligned}$$

   (b) For each of the $L \cdot K$ pairs of objects and parts, determine the predictive likelihood

$$f_{\ell k}(w_{ji} = w, v_{ji}) = \left(\frac{C_{kw} + \lambda / W}{\sum_{w'} C_{kw'} + \lambda} \right) \cdot \mathcal{N}(v_{ji} - {\rho_{j\ell}}^{(t-1)}; \hat{\mu}_k, \hat{\Lambda}_k).$$

   (c) Sample new object and part assignments from the following $L \cdot K$-dim. multinomial distribution:

$$(o_{ji}, z_{ji}) \sim \sum_{\ell=1}^{L} \sum_{k=1}^{K} (M_{s\ell} + \gamma / L) \left(\frac{N_{\ell k} + \alpha / K}{\sum_{k'} N_{\ell k'} + \alpha} \right) f_{\ell k}(w_{ji}, v_{ji}) \delta(o_{ji}, \ell) \delta(z_{ji}, k).$$

   (d) Add feature (w_{ji}, v_{ji}) to the cached statistics for its new object and part:

$$\begin{aligned} M_{s\ell} &\leftarrow M_{s\ell} + 1, & \ell &= o_{ji}, \\ N_{\ell k} &\leftarrow N_{\ell k} + 1, & k &= z_{ji}, \\ C_{kw} &\leftarrow C_{kw} + 1, & w &= w_{ji}, \\ (\hat{\mu}_k, \hat{\Lambda}_k) &\leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \oplus (v_{ji} - {\rho_{j\ell}}^{(t-1)}). \end{aligned}$$

2. Fix $({\mathbf{o}_j}^{(t)}, {\mathbf{z}_j}^{(t)}) = (\mathbf{o}_j, \mathbf{z}_j)$, and sample a new reference transformation ${\rho_j}^{(t)}$ as follows:
   (a) Remove ${\rho_j}^{(t-1)}$ from cached transformation statistics for scene s:

$$(\hat{\zeta}_s, \hat{\Upsilon}_s) \leftarrow (\hat{\zeta}_s, \hat{\Upsilon}_s) \ominus {\rho_j}^{(t-1)}.$$

   (b) Sample ${\rho_j}^{(t)} \sim \mathcal{N}(\chi_j, \Xi_j)$, a posterior distribution determined via (52) from the prior $\mathcal{N}(\rho_j; \hat{\zeta}_s, \hat{\Upsilon}_s)$, cached part statistics $\{\hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^K$, and feature positions $\mathbf{v}_j$.
   (c) Add ${\rho_j}^{(t)}$ to cached transformation statistics for scene s:

$$(\hat{\zeta}_s, \hat{\Upsilon}_s) \leftarrow (\hat{\zeta}_s, \hat{\Upsilon}_s) \oplus {\rho_j}^{(t)}.$$

3. For each $i \in \{1, \ldots, N_j\}$, update cached statistics for part $k = z_{ji}$ as follows:

$$\begin{aligned} (\hat{\mu}_k, \hat{\Lambda}_k) &\leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \ominus (v_{ji} - {\rho_{j\ell}}^{(t-1)}), \\ (\hat{\mu}_k, \hat{\Lambda}_k) &\leftarrow (\hat{\mu}_k, \hat{\Lambda}_k) \oplus (v_{ji} - {\rho_{j\ell}}^{(t)}), \end{aligned} \qquad \ell = o_{ji}.$$

Algorithm 3 Rao–Blackwellized Gibbs sampler for the fixed-order visual scene model of Fig. 12. We illustrate the sequential resampling of all object and part assignments $(\mathbf{o}_j, \mathbf{z}_j)$ in the j^{th} training image, as well as that image's coordinate frame ρ_j. A full iteration of the Gibbs sampler applies these updates to all images in random order. For efficiency, we cache and recursively update statistics $\{\hat{\zeta}_s, \hat{\Upsilon}_s\}_{s=1}^S$ of each scene's reference transformations, counts $M_{s\ell}$, $N_{\ell k}$ of the features assigned to each object and part, and statistics $\{C_{kw}, \hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^K$ of those features' appearance and position. The $\oplus$ and $\ominus$ operators update cached mean and covariance statistics as features are added or removed from parts (see Sect. 12.1)

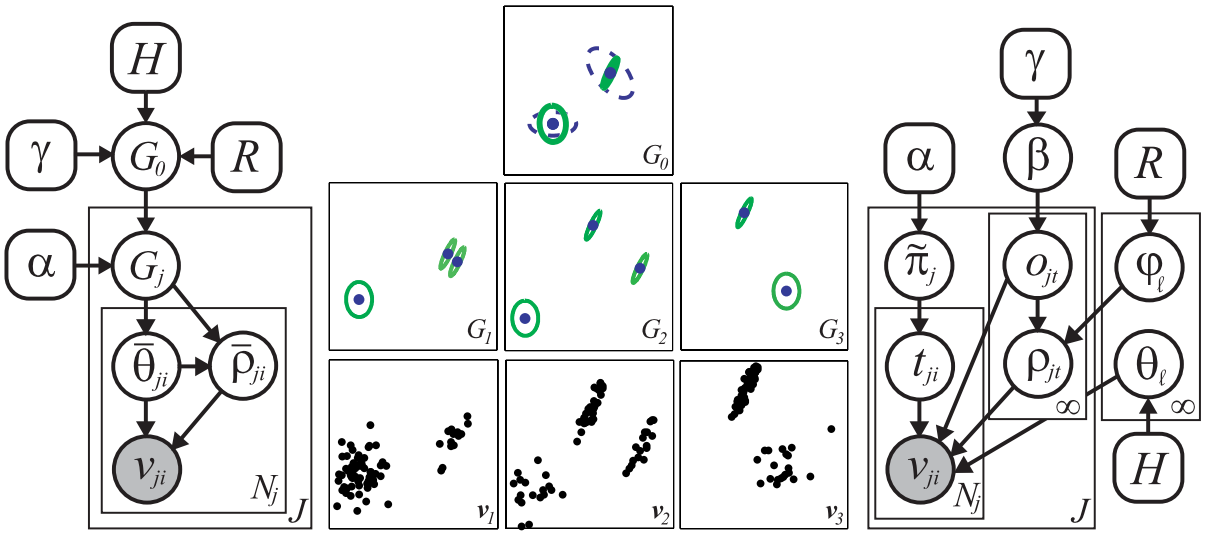

Figure 13 Directed graphical representations of a transformed Dirichlet process (TDP) mixture model. *Left*: Each group is assigned an infinite discrete distribution $G_j \sim \mathrm{DP}(\alpha, G_0)$, which is sampled from a global distribution $G_0(\theta, \rho)$ over transformations ρ of cluster parameters θ. Observations v_{ji} are then sampled from *transformed* parameters $\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji})$. *Center*: Illustration using 2D spatial data. G_0 is composed of 2D Gaussian distributions (green covariance ellipses), and corresponding Gaussian priors (blue dashed ellipses) on translations. The observations $\mathbf{v}_j$ in each of three groups are generated by transformed Gaussian mixtures G_j. *Right*: Chinese restaurant franchise representation of the TDP. Each group j has infinitely many local clusters (tables) t, which are associated with a transformation $\rho_{jt} \sim Q(\varphi_{o_{jt}})$ of some global cluster (dish) $o_{jt} \sim \boldsymbol{\beta}$. Observations (customers) v_{ji} are assigned to a table $t_{ji} \sim \tilde{\boldsymbol{\pi}}_j$, and share that table's transformed (seasoned) global cluster $\tau(\theta_{z_{ji}}; \rho_{jt_{ji}})$, where $z_{ji} = o_{jt_{ji}}$

clusters via a set of group-specific stochastic transformations. As we later demonstrate, this allows us to model richer datasets in which only a subset of the global clusters' properties are naturally shared.

7.2 Gibbs Sampling for Transformed Dirichlet Processes

To develop computational methods for learning transformed Dirichlet processes, we generalize the HDP's Chinese restaurant franchise representation (Teh et al. 2006). As in Sect. 4.2, customers (observations) v_{ji} sit at tables t_{ji} according to the clustering bias of (13), and new tables choose dishes via their popularity across the franchise (see (14)). As shown in Fig. 13, however, the dish (parameter) $\theta_{o_{jt}}$ at table t is now seasoned (transformed) according to $\rho_{jt} \sim Q(\varphi_{o_{jt}})$. Each time a dish is ordered, the recipe is seasoned differently, and each dish θ_ℓ has different typical seasonings $Q(\varphi_\ell)$.

While the HDP Gibbs sampler of Sect. 4.3 associated a single reference transformation with each image, the TDP instead describes groups via a *set* of randomly transformed clusters. We thus employ three sets of state variables: assignments $\mathbf{t}$ of observations to tables (transformed clusters), assignments $\mathbf{o}$ of tables to global clusters, and the transformations $\boldsymbol{\rho}$ associated with each occupied table. As summarized in Algorithm 4, the cluster weights $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\pi}}_j$ are then analytically marginalized.

In the TDP, each global cluster ℓ combines transformations with different likelihood parameters θ_ℓ. Thus, to adequately explain the same data with a different cluster o_{jt}, a complementary change of ρ_{jt} is typically required. For this reason, Algorithm 4 achieves *much* more rapid convergence via a blocked Gibbs sampler which simultaneously updates (o_{jt}, ρ_{jt}). See Sect. 12.3 for discussion of the Gaussian integrals which make this tractable. Finally, note that the TDP's concentration parameters have intuitive interpretations: γ controls the expected number of global clusters, while α determines the average number of transformed clusters in each group. As in the HDP sampler, Algorithm 4 uses auxiliary variable methods (Escobar and West 1995; Teh et al. 2006) to learn these statistics from training data.

7.3 A Toy World: Bars and Blobs

To provide intuition for the TDP, we consider a toy world in which "images" depict a collection of two-dimensional points. As illustrated in Fig. 14, the training images we consider typically depict one or more diagonally oriented "bars" in the upper right, and round "blobs" in the lower left. As in more realistic datasets, the exact locations of these "objects" vary from image to image. We compare models learned by the TDP Gibbs sampler of Algorithm 4 and a corresponding HDP sampler. Both models use Gaussian clusters $\theta_\ell = (\mu_\ell, \Lambda_\ell)$ with vague normal-inverse-Wishart priors H. For the TDP, transformations ρ then define translations of

Given previous table assignments $\mathbf{t}_j{}^{(t-1)}$ for the N_j observations in group j, and transformations $\boldsymbol{\rho}_j{}^{(t-1)}$ and global cluster assignments $\mathbf{o}_j{}^{(t-1)}$ for that group's T_j tables:

1. Set $\mathbf{t}_j = \mathbf{t}_j{}^{(t-1)}$, $\mathbf{o}_j = \mathbf{o}_j{}^{(t-1)}$, $\boldsymbol{\rho}_j = \boldsymbol{\rho}_j{}^{(t-1)}$, and sample a random permutation $\tau(\cdot)$ of $\{1, \ldots, N_j\}$. For each $i \in \{\tau(1), \ldots, \tau(N_j)\}$, sequentially resample data assignment t_{ji} as follows:
   (a) Decrement $N_{jt_{ji}}$, and remove v_{ji} from the cached statistics for its current cluster $\ell = o_{jt_{ji}}$:

   $$(\hat{\mu}_\ell, \hat{\Lambda}_\ell) \leftarrow (\hat{\mu}_\ell, \hat{\Lambda}_\ell) \ominus (v_{ji} - \rho_{jt_{ji}}).$$

   (b) For each of the T_j instantiated tables, determine the predictive likelihood

   $$f_t(v_{ji}) = \mathcal{N}(v_{ji} - \rho_{jt}; \hat{\mu}_\ell, \hat{\Lambda}_\ell), \quad \ell = o_{jt}.$$

   (c) For each of the L instantiated global clusters, determine the marginal likelihood

   $$g_\ell(v_{ji}) = \mathcal{N}(v_{ji}; \hat{\mu}_\ell + \hat{\zeta}_\ell, \hat{\Lambda}_\ell + \hat{\Upsilon}_\ell).$$

   Also determine the marginal likelihood $g_{\bar{\ell}}(v_{ji})$ of a potential new global cluster $\bar{\ell}$.
   a) Sample a new table assignment t_{ji} from the following $(T_j + 1)$-dim. multinomial distribution:

   $$t_{ji} \sim \sum_{t=1}^{T_j} N_{jt} f_t(v_{ji}) \delta(t_{ji}, t) + \frac{\alpha}{\gamma + \sum_\ell M_\ell} \left[\sum_{\ell=1}^{L} M_\ell g_\ell(v_{ji}) + \gamma g_{\bar{\ell}}(v_{ji}) \right] \delta(t_{ji}, \bar{t}).$$

   (e) If $t_{ji} = \bar{t}$, create a new table, increment T_j, and sample

   $$o_{j\bar{t}} \sim \sum_{\ell=1}^{L} M_\ell g_\ell(v_{ji}) \delta(o_{j\bar{t}}, \ell) + \gamma g_{\bar{\ell}}(v_{ji}) \delta(o_{j\bar{t}}, \bar{\ell}).$$

   If $o_{j\bar{t}} = \bar{\ell}$, create a new global cluster and increment L.
   (f) If $t_{ji} = \bar{t}$, also sample $\rho_{j\bar{t}} \sim \mathcal{N}(\chi_{j\bar{t}}, \Xi_{j\bar{t}})$, a posterior distribution determined via (57) from the prior $\mathcal{N}(\rho_{j\bar{t}}; \hat{\zeta}_\ell, \hat{\Upsilon}_\ell)$ and likelihood $\mathcal{N}(v_{ji}; \hat{\mu}_\ell + \rho_{j\bar{t}}, \Lambda_\ell)$, where $\ell = o_{j\bar{t}}$.
   (g) Increment $N_{jt_{ji}}$, and add v_{ji} to the cached statistics for its new cluster $\ell = o_{jt_{ji}}$:

   $$(\hat{\mu}_\ell, \hat{\Lambda}_\ell) \leftarrow (\hat{\mu}_\ell, \hat{\Lambda}_\ell) \oplus (v_{ji} - \rho_{jt_{ji}}).$$

2. Fix $\mathbf{t}_j{}^{(t)} = \mathbf{t}_j$. If any tables are empty ($N_{jt} = 0$), remove them and decrement T_j.
3. Sample a permutation $\tau(\cdot)$ of $\{1, \ldots, T_j\}$. For each $t \in \{\tau(1), \ldots, \tau(T_j)\}$, jointly resample (o_{jt}, ρ_{jt}):
   (a) Decrement $M_{o_{jt}}$, and remove all data at table t from the cached statistics for cluster $\ell = o_{jt}$:

   $$(\hat{\mu}_\ell, \hat{\Lambda}_\ell) \leftarrow (\hat{\mu}_\ell, \hat{\Lambda}_\ell) \ominus (v - \rho_{jt}) \quad \text{for each } v \in \mathbf{v}_t \triangleq \{v_{ji} | t_{ji} = t\}.$$

   a) For each of the L instantiated global clusters and a potential new cluster $\bar{\ell}$, determine the marginal likelihood $g_\ell(\mathbf{v}_t)$ via the Gaussian computations of (58).
   (c) Sample a new cluster assignment o_{jt} from the following $(L + 1)$-dim. multinomial distribution:

   $$o_{jt} \sim \sum_{\ell=1}^{L} M_\ell g_\ell(\mathbf{v}_t) \delta(o_{jt}, \ell) + \gamma g_{\bar{\ell}}(\mathbf{v}_t) \delta(o_{jt}, \bar{\ell}).$$

   If $o_{jt} = \bar{\ell}$, create a new global cluster and increment L.

Algorithm 4 Gibbs sampler for the TDP mixture model of Fig. 13. For efficiency, we cache and recursively update statistics $\{\hat{\mu}_\ell, \hat{\Lambda}_\ell, \hat{\zeta}_\ell, \hat{\Upsilon}_\ell\}_{\ell=1}^{L}$ of each global cluster's associated data and reference transformations, and counts of the number of tables M_ℓ assigned to each cluster, and observations N_{jt} to each table. The $\oplus$ and $\ominus$ operators update cached mean and covariance statistics (see Sect. 12.1)

(d) Sample a new transformation $\rho_{jt} \sim \mathcal{N}(\chi_{jt}, \Xi_{jt})$, a posterior distribution determined via (57) from the prior $\mathcal{N}(\rho_{jt}; \hat{\zeta}_\ell, \hat{\Upsilon}_\ell)$ and likelihood $\mathcal{N}(v; \hat{\mu}_\ell + \rho_{jt}, \hat{\Lambda}_\ell)$, where $\ell = o_{j\bar{t}}$ and $v \in \mathbf{v}_t$.
(e) Increment $M_{o_{jt}}$, and add all data at table t to the cached statistics for cluster $\ell = o_{jt}$:

$$(\hat{\mu}_\ell, \hat{\Lambda}_\ell) \leftarrow (\hat{\mu}_\ell, \hat{\Lambda}_\ell) \oplus (v - \rho_{jt}) \quad \text{for each } v \in \mathbf{v}_t.$$

4. Fix $\mathbf{o}_j{}^{(t)} = \mathbf{o}_j$, $\boldsymbol{\rho}_j{}^{(t)} = \boldsymbol{\rho}_j$. If any global clusters are unused ($M_\ell = 0$), remove them and decrement L.
5. Given gamma priors, resample concentration parameters γ and α using auxiliary variables (Escobar and West 1995; Teh et al. 2006).

Algorithm 4 (*continued*)

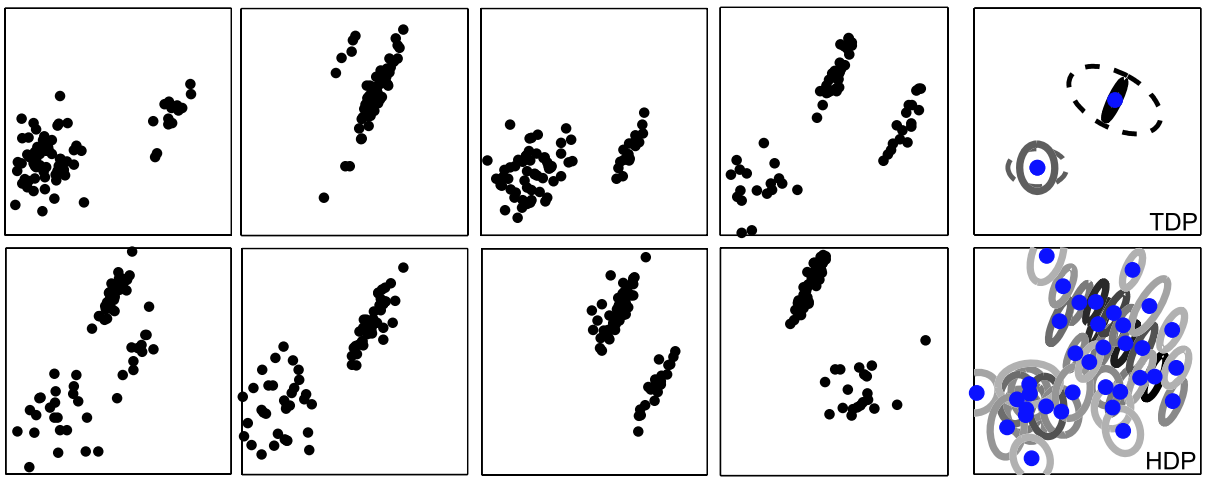

Figure 14 Learning HDP and TDP models from toy 2D spatial data. *Left*: Eight of fifty training "images" containing diagonally oriented bars and round blobs. *Upper right*: Global distribution $G_0(\theta, \rho)$ over Gaussian clusters (*solid*) and translations (*dashed*) learned by the TDP Gibbs sampler. *Lower right*: Global distribution $G_0(\theta)$ over the much larger number of Gaussian clusters (intensity proportional to probability β_ℓ) learned by the HDP Gibbs sampler

global cluster means, as in Sect. 3.1, and R is taken to be an inverse-Wishart prior on zero-mean Gaussians. For both models, we run the Gibbs sampler for 100 iterations, and resample concentration parameters at each iteration.

As shown in Fig. 14, the TDP sampler learns a global distribution $G_0(\theta, \rho)$ which parsimoniously describes these images via translations of two bar and blob-shaped global clusters. In contrast, because the HDP models absolute feature positions, it defines a large set of global clusters which discretize the range of observed object positions. Because a smaller number of features are used to estimate the shape of each cluster, they less closely approximate the true shapes of bars and blobs. More importantly, the HDP model cannot predict the appearance of these objects in new image positions. We thus see that the TDP's use of transformations is needed to adequately transfer information among different object instances, and generalize to novel spatial scenes.

7.4 Characterizing Transformed Distributions

Recall that the global measure G_0 underlying the TDP (see (20)) defines a discrete distribution over cluster parameters θ_ℓ. In contrast, the distributions $Q(\varphi_\ell)$ associated with transformations of these clusters are continuous. Each group j will thus create many copies $\tilde{\theta}_{jt}$ of global cluster θ_ℓ, but associate each with a *different* transformation ρ_{jt}. Aggregating the probabilities assigned to these copies, we can directly express G_j in terms of the distinct global cluster parameters:

$$G_j(\theta, \rho) = \sum_{\ell=1}^{\infty} \pi_{j\ell} \delta(\theta, \theta_\ell) \left[\sum_{s=1}^{\infty} \omega_{j\ell s} \delta(\rho, \check{\rho}_{j\ell s}) \right], \qquad \pi_{j\ell} = \sum_{t | o_{jt} = \ell} \tilde{\pi}_{jt}. \tag{23}$$

In this expression, we have grouped the infinite set of transformations which group j associates with each global cluster ℓ:

$$\{\check{\rho}_{j\ell s} | s = 1, 2, \ldots\} = \{\rho_{jt} | o_{jt} = \ell\}. \tag{24}$$

The weights $\boldsymbol{\omega}_{j\ell} = (\omega_{j\ell 1}, \omega_{j\ell 2}, \ldots)$ then equal the proportion of the total cluster probability $\pi_{j\ell}$ contributed by each transformed cluster $\tilde{\pi}_{jt}$ satisfying $o_{jt} = \ell$. The following proposition provides a direct probabilistic characterization of the transformed measures arising in the TDP.

Proposition *Let $G_0(\theta, \rho)$ be a global measure as in (20), and $G_j(\theta, \rho) \sim \mathrm{DP}(\alpha, G_0(\theta, \rho))$ be expressed as in (23). The marginal distributions of G_j with respect to parameters and transformations then also follow Dirichlet processes*:

$$G_j(\theta) \sim \mathrm{DP}(\alpha, G_0(\theta)), \quad G_0(\theta) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(\theta, \theta_\ell), \tag{25}$$

$$G_j(\rho) \sim \mathrm{DP}(\alpha, G_0(\rho)), \quad G_0(\rho) = \sum_{\ell=1}^{\infty} \beta_\ell Q(\varphi_\ell). \tag{26}$$

Alternatively, given any discrete parameter θ_ℓ from the global measure, we have

$$G_j(\rho | \theta = \theta_\ell) \sim \mathrm{DP}(\alpha\beta_\ell, Q(\varphi_\ell)). \tag{27}$$

The weights assigned to transformations of θ_ℓ thus follow a stick-breaking process $\boldsymbol{\omega}_{j\ell} \sim \mathrm{GEM}(\alpha\beta_\ell)$.

Proof See Sect. 6.2.2 of the doctoral thesis (Sudderth 2006). □

Examining (25), we see that the TDP induces discrete marginal distributions on parameters exactly like those arising in the HDP (Teh et al. 2006). The HDP can thus be seen as a limiting case of the TDP in which transformations are insignificant or degenerate.

As the concentration parameter α becomes a large, a Dirichlet process $\mathrm{DP}(\alpha, H)$ approaches the base measure H by distributing small weights among a large number of discrete samples (see Sect. 4.1). The result in (27) thus shows that parameters θ_ℓ with small weight β_ℓ will also have greater variability in their transformation distributions, because (on average) they are allocated fewer samples. Intuitively, the concentration parameters $\{\alpha\beta_\ell\}_{\ell=1}^{\infty}$ associated with transformations of all global clusters sum to α, the overall concentration of G_j about G_0.

7.5 Dependent Dirichlet Processes

The HDP is a special case of a very general *dependent Dirichlet process* (DDP) (MacEachern 1999) framework for introducing dependency among multiple DPs. DDPs have been previously used to model spatial data, by using a single "global" stick-breaking process to mix an infinite set of Gaussian processes (Gelfand et al. 2005) or linear (ANOVA) models (De Iorio et al. 2004). However, applied to the spatial data considered in this paper, these approaches would learn feature models which depend on absolute image coordinates. As discussed in Sect. 3.1, such approaches are poorly matched to the structure of visual scenes.

Viewing cluster parameters and transformations as one augmented parameter vector, TDPs are also a special case of the DDP framework. However, this perspective obscures the interplay between the discrete and continuous portions of the TDP base measure, and the manner in which transformations modify parameters to achieve a very rich class of dependencies.

8 Modeling Scenes with Unknown Numbers of Objects

The transformed Dirichlet process developed in Sect. 7 defines global clusters via a parametric, exponential family $F(\theta)$. As suggested by the toy example of Fig. 14, this approach could be directly used to construct simple, weakly structured models of object geometry (Sudderth et al. 2006b). However, realistic objects have complex internal structure, and significant local appearance variations. We thus extend the basic TDP of Fig. 13 to learn richer, part-based models for object categories.

8.1 Transformed DP Models for Objects and Parts

As in the single-object HDP of Sect. 4.2, each part $\theta_{\ell k} = (\eta_{\ell k}, \mu_{\ell k}, \Lambda_{\ell k})$ of object category ℓ has a Gaussian position distribution $\mathcal{N}(\mu_{\ell k}, \Lambda_{\ell k})$, and a multinomial appearance distribution $\eta_{\ell k}$. Letting $H = H_w \times H_v$ denote a prior measure on part parameters, $F_\ell \sim \mathrm{DP}(\kappa, H)$ is then an infinite discrete measure representing the potentially infinite set of parts underlying the ℓ^{th} visual category:

$$\begin{aligned} F_\ell(\theta) &= \sum_{k=1}^{\infty} \varepsilon_{\ell k} \delta(\theta, \theta_{\ell k}), \\ \boldsymbol{\varepsilon}_\ell &\sim \mathrm{GEM}(\kappa), \qquad (\eta_{\ell k}, \mu_{\ell k}, \Lambda_{\ell k}) = \theta_{\ell k} \sim H. \end{aligned} \tag{28}$$

The Gaussian parameters $(\mu_{\ell k}, \Lambda_{\ell k})$ associated with each part model feature positions in an object-centered coordinate frame. In the visual scenes considered by Sect. 9, we expect there to be little direct overlap in the appearance of different categories. For simplicity, (28) thus describes categories using independent parts, rather than hierarchically sharing parts as in Sect. 4.2.

The TDP model of Sect. 7.1 employed a global measure G_0 modeling transformations ρ of an infinite set of cluster

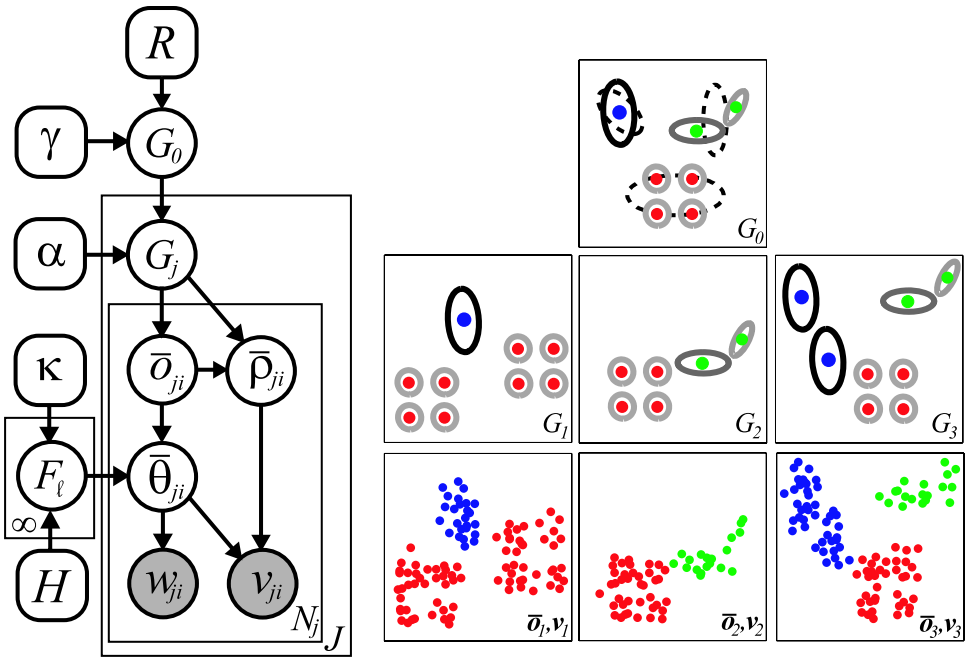

Figure 15 TDP model for 2D visual scenes (*left*), and cartoon illustration of the generative process (*right*). Global mixture G_0 describes the expected frequency and image position of visual categories, whose internal structure is represented by part-based appearance models $\{F_\ell\}_{\ell=1}^{\infty}$. Each image distribution G_j instantiates a randomly chosen set of objects at transformed locations ρ. Image features with appearance w_{ji} and position v_{ji} are then sampled from transformed parameters $\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji})$ corresponding to different parts of object $\bar{o}_{ji}$. The cartoon example defines three color-coded object categories, which are composed of one (*blue*), two (*green*), and four (*red*) Gaussian parts, respectively. *Dashed ellipses* indicate transformation priors for each category

parameters. Generalizing this construction, we allow infinitely many potential visual categories o, and characterize transformations of these part-based models as follows:

$$G_0(o,\rho) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(o,\ell) q(\rho|\varphi_\ell), \qquad \boldsymbol{\beta} \sim \mathrm{GEM}(\gamma), \qquad \varphi_\ell \sim R. \tag{29}$$

In this distribution, the random variable o indicates the part-based model, as in (28), corresponding to some category. The appearance of the j^{th} image is then determined by a set of randomly transformed objects $G_j \sim \mathrm{DP}(\alpha, G_0)$, so that

$$G_j(o,\rho) = \sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta(o, o_{jt}) \delta(\rho, \rho_{jt}), \qquad \tilde{\boldsymbol{\pi}}_j \sim \mathrm{GEM}(\alpha), \qquad (o_{jt}, \rho_{jt}) \sim G_0. \tag{30}$$

In this expression, t indexes the set of object *instances* in image j, which are associated with visual categories o_{jt}. Each of the N_j features in image j is independently sampled from some object instance $t_{ji} \sim \tilde{\boldsymbol{\pi}}_j$. This can be equivalently expressed as $(\bar{o}_{ji}, \bar{\rho}_{ji}) \sim G_j$, where $\bar{o}_{ji}$ is the global category corresponding to an object instance situated at transformed location $\bar{\rho}_{ji}$. Finally, parameters corresponding to one of this object's parts generate the observed feature:

$$(\bar{\eta}_{ji}, \bar{\mu}_{ji}, \bar{\Lambda}_{ji}) = \bar{\theta}_{ji} \sim F_{\bar{o}_{ji}}, \qquad w_{ji} \sim \bar{\eta}_{ji}, \qquad v_{ji} \sim \mathcal{N}(\bar{\mu}_{ji} + \bar{\rho}_{ji}, \bar{\Lambda}_{ji}). \tag{31}$$

In later sections, we let $k_{ji} \sim \boldsymbol{\epsilon}_{\bar{o}_{ji}}$ indicate the part underlying the i^{th} feature. Focusing on scale-normalized datasets, we again associate transformations with image-based translations.

The hierarchical, TDP scene model of Fig. 15 employs three different stick-breaking processes, allowing uncertainty in the number of visual categories (GEM(γ)), parts composing each category (GEM(κ)), and object instances depicted in each image (GEM(α)). It thus generalizes the parametric model of Fig. 12, which assumed fixed, known sets of parts and objects. As $\kappa \to 0$, each category uses a single part, and we recover a variant of the simpler TDP model of Sect. 7.1. Interestingly, if $\alpha \to 0$ and transformations are neglected, we recover a single-object model related to the recently (and independently) developed *nested Dirichlet process* (Rodriguez et al. 2006).

8.2 Gibbs Sampling for TDP Models of Visual Scenes

To learn the parameters of the visual scene model depicted in Fig. 15, we generalize the TDP Gibbs sampler of Algorithm 4. We maintain a dynamic list of the instantiated object instances t in each image j, representing each instance by a transformation ρ_{jt} of global visual category o_{jt}. Each feature (w_{ji}, v_{ji}) is then assigned to a part k_{ji} of some instance t_{ji}. Via blocked Gibbs resampling of these four sets of variables $(\mathbf{o}, \boldsymbol{\rho}, \mathbf{t}, \mathbf{k})$, we then simultaneously segment and recognize objects.

Section 12.4 describes the form of this sampler in more detail. In the first stage, we fix the object instances $(\mathbf{o}_j, \boldsymbol{\rho}_j)$

in each image and jointly resample the part and instance (k_{ji}, t_{ji}) assigned to each feature. The resulting updates combine aspects of our earlier TDP (Algorithm 4, steps 1–2) and fixed-order scene (Algorithm 3, step 1) Gibbs samplers. In the second stage, we fix assignments $\mathbf{t}_j$ of features to object instances, effectively segmenting images into independent objects. We may then jointly resample the location ρ_{jt}, visual category o_{jt}, and part assignments $\{k_{ji} | t_{ji} = t\}$ associated with each table by adapting the single-object HDP sampler of Algorithms 1–2. Note that this second stage approximates the infinite set of potential parts for category ℓ (see (28)) by the K_ℓ parts to which at least one feature is currently assigned. This can be seen as a dynamic version of the stick-breaking truncations underlying certain other DP sampling algorithms (Ishwaran and James 2001; Rodriguez et al. 2006).

9 Street and Office Scenes

To evaluate our hierarchical models for multiple object scenes, we use the two datasets depicted in Fig. 4. The first set contains 613 street scenes depicting four "objects": buildings, cars (side views), roads, and trees. To align with the assumptions underlying our 2D scene models, images were normalized so that cars appear at comparable scales. As shown in Fig. 4, some of these street scenes have labels for all four categories, while others are only partially segmented. The second dataset includes 315 pictures of office scenes containing four objects: computer screens (frontal views), keyboards, mice, and background clutter. In this case, images were normalized so that computer screens appeared at comparable scales, and all object instances were labeled.

For both datasets, we represent training and test images by the three types of interest regions described in Sect. 2.1. We estimated a separate appearance dictionary for each dataset, which after expansion to encode region shape (see Sect. 2.2) contained $W = 1600$ visual words.

9.1 Fixed-Order Scene Models

We begin by examining the fixed-order visual scene model of Fig. 12, and learn parameters via the Gibbs sampler of Algorithm 3. For training, we used 400 street scenes and 250 office scenes; the remaining images then provide a segmented test set. To estimate model parameters, we first ran the Gibbs sampler for 500 iterations using only the training images. We incorporate manual segmentations by fixing the object category assignments o_{ji} of labeled features. For unlabeled features, object assignments are left unconstrained, and sampled as in Algorithm 3. Each scene model used thirty shared parts, and Dirichlet precision parameters set as $\gamma = 4$, $\alpha = 15$ via cross-validation. The position prior H_v weakly favored parts covering 10% of the image range, while the appearance prior $\mathrm{Dir}(W/10)$ was biased towards sparse distributions.

9.1.1 *Visualization of Learned Parts*

Figure 16 illustrates learned, part-based models for street and office scenes. Although objects share a common set of parts within each scene model, we can approximately count the number of parts used by each object by thresholding the posterior part distributions $\boldsymbol{\pi}_\ell$. For street scenes, cars are allocated roughly four parts, while buildings and roads use large numbers of parts to uniformly tile regions corresponding to their typical size. Several parts are shared

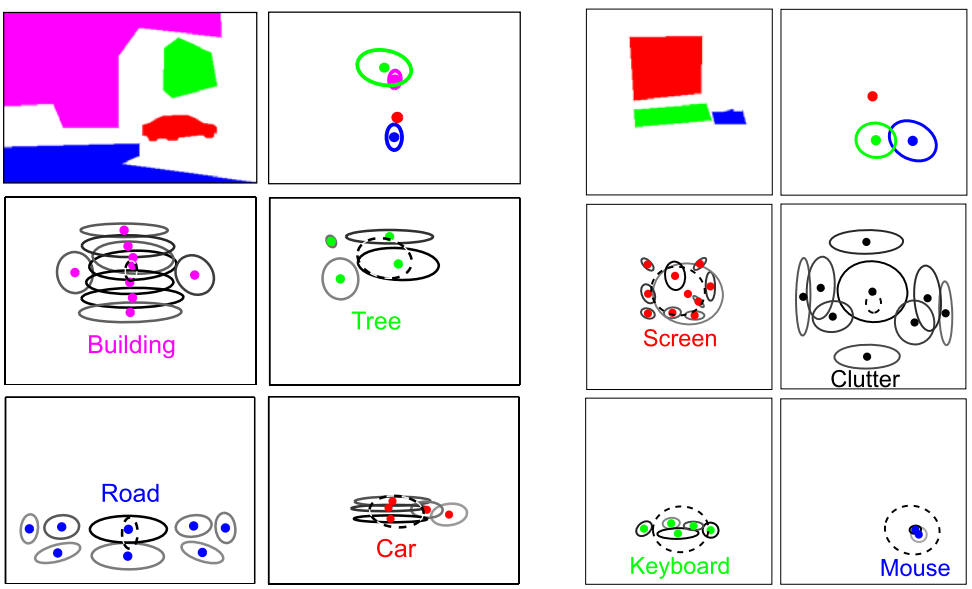

Figure 16 Learned contextual, fixed-order models for street scenes (*left*) and office scenes (*right*), each containing four objects. *Top*: Gaussian distributions over the positions of other objects given the location of the car (*left*) or computer screen (*right*). *Bottom*: Parts (*solid*) generating at least 5% of each category's features, with intensity proportional to probability. Parts are translated by that object's mean position, while the *dashed ellipses* indicate each object's marginal transformation covariance

between the tree and building categories, presumably due to the many training images in which buildings are only partially occluded by foliage. The office scene model describes computer screens with ten parts, which primarily align with edge and corner features. Due to their smaller size, keyboards are described by five parts, and mice by two. The background clutter category then uses several parts, which move little from scene to scene, to distribute features across the image. Most parts are unshared, although the screen and keyboard categories reuse a few parts to describe edge-like features.

Figure 16 also illustrates the contextual relationships learned by both scene models. Intuitively, street scenes have a vertically layered structure, while in office scenes the keyboard is typically located beneath the monitor, and the mouse to the keyboard's right.

9.1.2 Segmentation of Novel Visual Scenes

To analyze test images, we fix the part and object assignments corresponding to the final Gibbs sampling iteration on the training set. To avoid local optima, we then run the test image Gibbs sampler for 20 iterations from each of ten different random initializations. Given reference transformations sampled in this fashion, we use (50) to estimate the posterior probability that test features were generated by each candidate object category. Averaging these probabilities provides a confidence-weighted segmentation, which we illustrate by fading uncertain features to gray.

Figure 17 shows segmentations for several typical test street scenes, and transformed parts from the highest likelihood sampling iteration. Segmentations of building and road features are typically very accurate, as the contextual model learns the vertical layering inherent in street scenes.

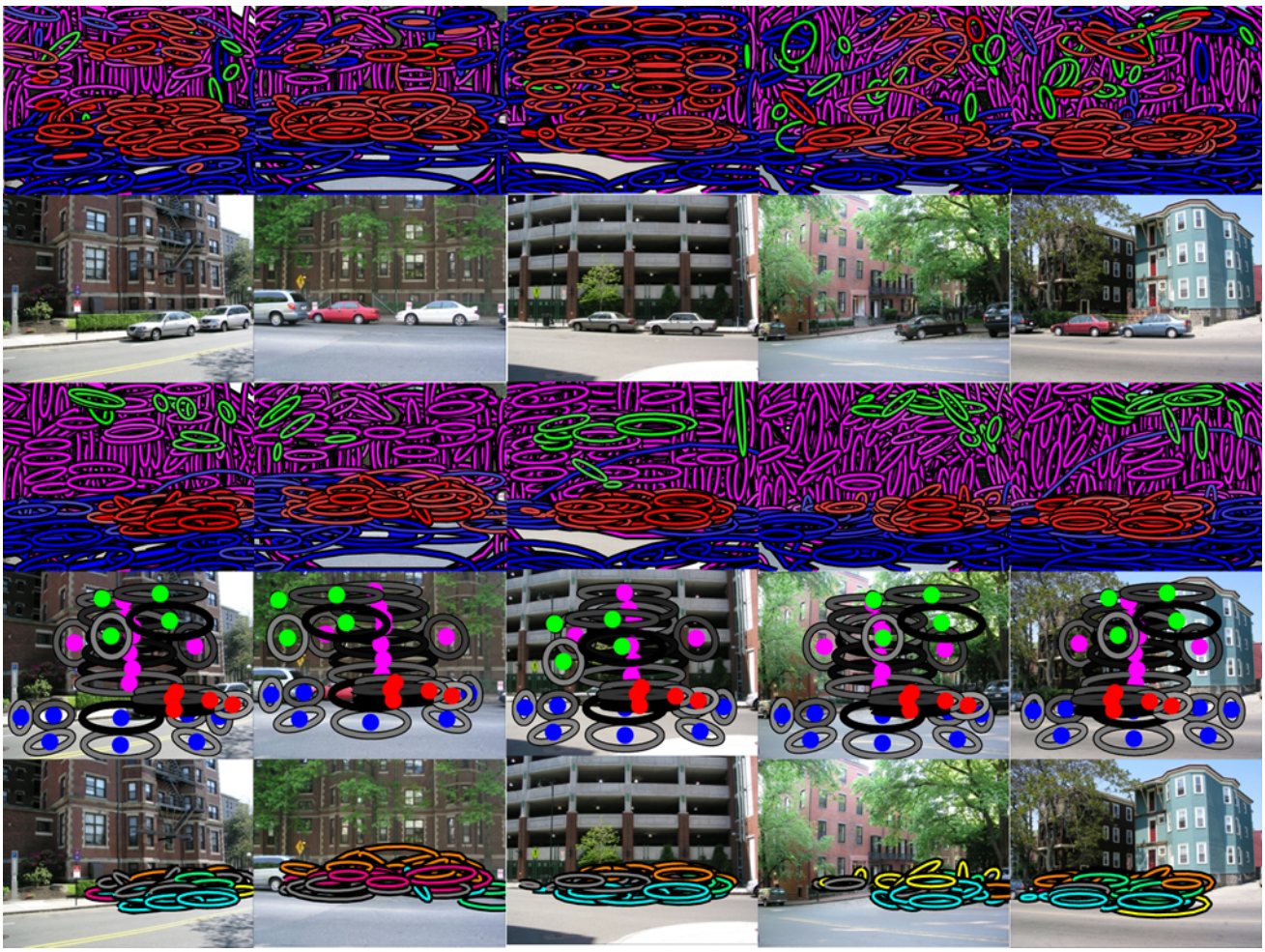

Figure 17 Feature segmentations produced by a contextual, fixed-order model of street scenes containing cars (*red*), buildings (*magenta*), roads (*blue*), and trees (*green*). For five test images (*second row*), we compare segmentations which assign features to the most probable object category for the contextual model (*third row*) and a baseline bag of features model (*first row*). We also show model parts translated according to each image's reference transformation (*fourth row*), and color-coded assignments of features to the different parts associated with cars (*fifth row*)

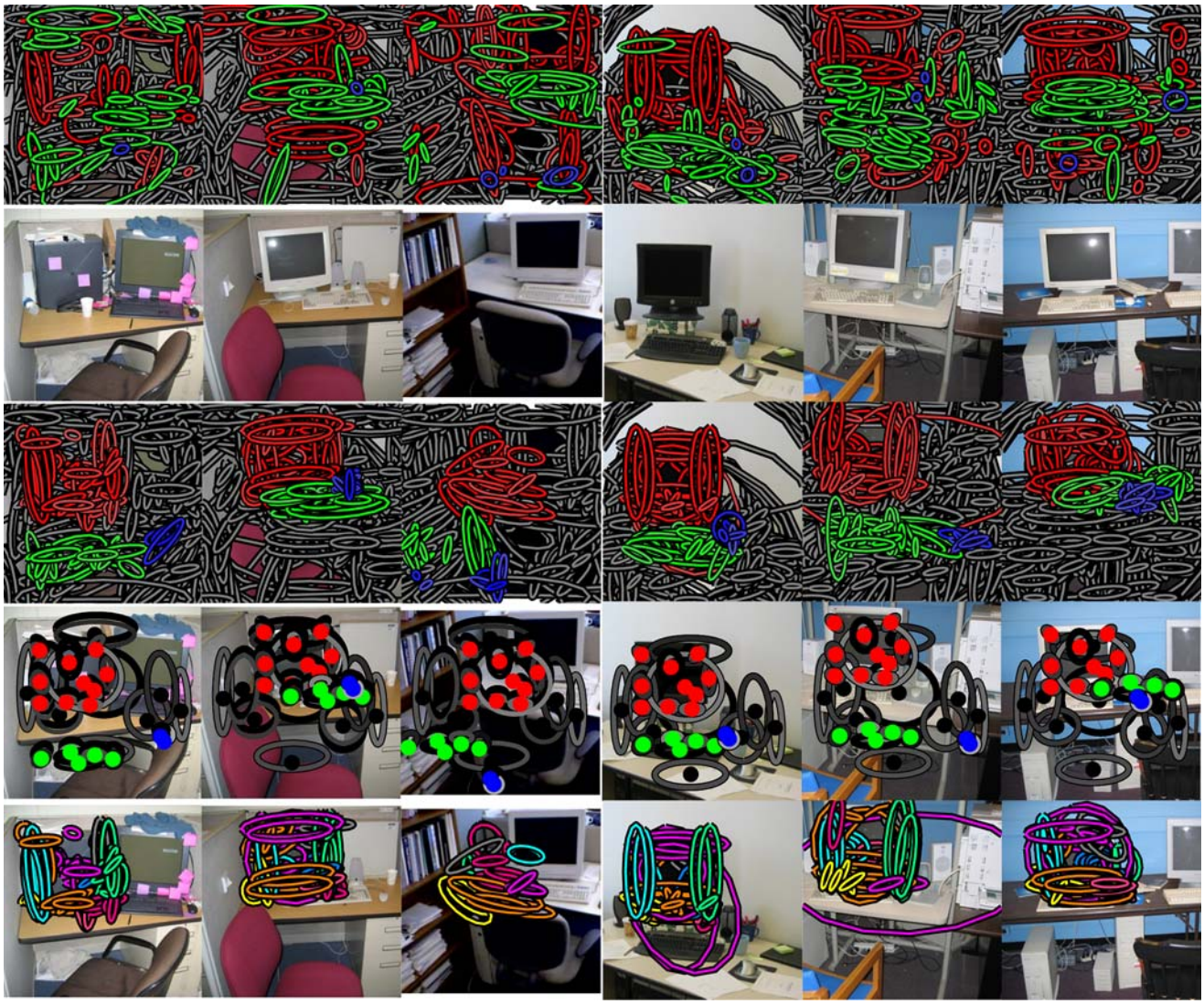

Figure 18 Feature segmentations produced by a contextual, fixed-order model of office scenes containing computer screens (*red*), keyboards (*green*), mice (*blue*), and background clutter (*gray*). For six test images (*second row*), we compare segmentations which assign features to the most probable object category for the contextual model (*third row*) and a baseline bag of features model (*first row*). We also show model parts translated according to each image's reference transformation (*fourth row*), and color-coded assignments of features to the different parts associated with computer screens (*fifth row*)

Note that a number of test images violate our parametric model's assumption that scenes depict a single instance of each object. To partially correct for this, the model learns horizontally elongated car parts which extend beyond an average car. Although this better segments adjacent cars, nearby background clutter is often mislabeled. In images containing widely separated cars, one car is usually missed entirely. The assumption that every image contains one tree is also problematic, since some features are typically classified as foliage even when no trees are present.

Figure 18 shows similar segmentation results for office scenes. Because most test images do indeed contain a single computer screen, the model's use of a fixed-order transformation causes fewer errors for office scenes. Contextual information is especially important for detecting computer mice (see Fig. 18). Very few features are detected in the region corresponding to the mouse, and they are often not distinctive. However, as screens can be reliably located, this provides a strong constraint on the expected location of the mouse. In fact, for test images in which no mouse is present the system often hallucinates one in other appropriately positioned clutter.

For comparison, Figs. 17 and 18 also show segmentation results for a bag of features model (Sivic et al. 2005), derived from the full contextual model of Fig. 12 by ignoring feature positions, and thus reference transformations. As confirmed by the ROC curves of Fig. 19, the appearance-only model is

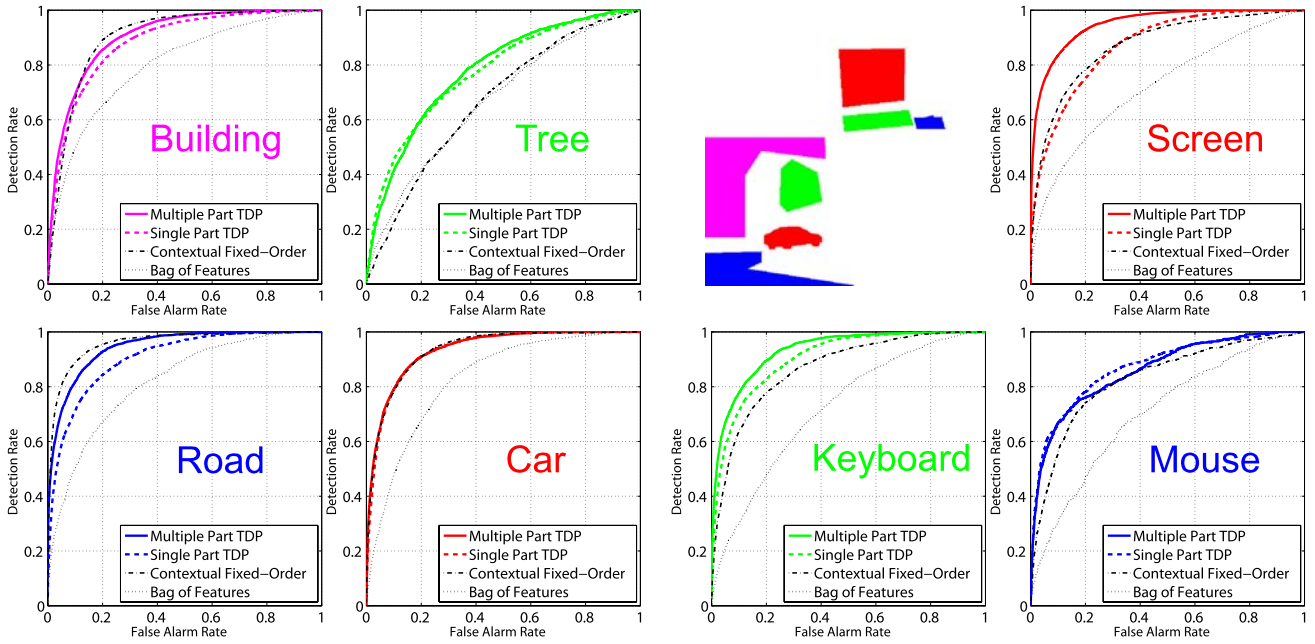

Figure 19 ROC curves summarizing segmentation performance for the features composing street scenes (*left*) and office scenes (*right*). We compare the full TDP scene model of Fig. 15 (*solid*, *colored*) to a simplified, single-part TDP model (*dashed*, *colored*), a fixed-order contextual scene model (*dash-dotted*, *black*) as in Fig. 12, and a baseline bag of features model (*dotted*, *black*)

significantly less accurate for all categories except trees. For street scenes, the full, position-based model recognizes car features reasonably well despite employing a single reference position, and roads are very accurately segmented. For office scenes, it exploits contextual relationships to detect mice and keyboards with accuracy comparable to the more distinctive computer screens. These improvements highlight the importance of spatial structure in visual scene understanding.

9.2 Transformed Dirichlet Process Scene Models

We now examine our TDP scene models via the training and test images used to evaluate the fixed-order model. To estimate model parameters, we first ran the Gibbs sampler of Sect. 8.2 for 500 training iterations using only those features with manually specified object category labels. For street scenes, we then ran another 100 Gibbs sampling iterations using all features. Empirically, this sequential training converges faster because it initializes visual categories with cleanly segmented objects. For each dataset, we compare the full TDP scene model of Fig. 15 to a simplified model which constrains each category to a single part (Sudderth et al. 2006b). This single-part TDP is similar to the model in Fig. 13, except that visual categories also have multinomial appearance distributions.

During training, we distinguish the manually labeled *object categories* from the *visual categories* composing the TDP's global distribution G_0. We restrict the Gibbs sampler from assigning different objects to the same visual category, but multiple visual categories may be used to describe different forms of a particular object. When learning TDP scene models, we also distinguish *rigid objects* (e.g., computer screens, keyboards, mice, and cars) from *textural objects* such as buildings, roads, trees, and office clutter. For rigid objects, we restrict all features composing each labeled training instance to be associated with the *same* transformed global cluster. This constraint, which is enforced by fixing the table assignments t_{ji} for features of rigid objects, ensures that the TDP learns descriptions of complete objects rather than object pieces. For textural categories, we allow the sampler to partition labeled training regions into transformed object instances, and thus automatically discover smaller regions with consistent, predictable structure.

One of the strengths of the TDP is that the learning process is reasonably insensitive to the particular values of the hyperparameters. The prior distribution H characterizing object parts was set as in Sect. 9.1, while the inverse-Wishart transformation prior R weakly favored zero-mean Gaussians covering the full image range. The concentration parameters defining the numbers of visual categories $\gamma \sim \text{Gamma}(1.0, 0.1)$ and parts per category $\kappa \sim \text{Gamma}(1.0, 0.1)$ were then assigned vague gamma priors, and resampled during the learning process. To encourage the learning of larger global clusters for textural categories, the concentration parameter controlling the num-

ber of object instances was more tightly constrained as $\alpha \sim \text{Gamma}(1.0, 1.0)$.

9.2.1 Visualization of Learned Parts

Figure 20 illustrates the global, visual categories that were learned from the dataset of street scenes. The single-part TDP uses compact global categories, and many transformed object instances, to more uniformly spread features across the image. Buildings, roads, and trees are each split into several visual categories, which describe different characteristic structural features. The full TDP scene model creates a more detailed, 9-part car appearance model. It also learns extended, multiple-part models of the large building and road regions which appear in many training images. The full part-based model thus captures some of the coarse-scale structure of street scenes, while the simpler single-part TDP is limited to modeling local feature dependencies.

As shown in Fig. 21, the single-part TDP model of office scenes is qualitatively similar to the street scene model: images are described by large numbers of compact transformed clusters. The multiple-part TDP, however, reveals interesting differences in the global structure of these scene categories. Due to their internal regularities, computer screens and keyboards are each described by detailed visual categories with many parts. To model background clutter, the TDP learns several small clusters of parts which uniformly distribute features within image regions. Because the TDP currently lacks an explicit occlusion model, it also defines a frame-like visual category which captures the background features often found at image boundaries.

9.2.2 Segmentation of Novel Visual Scenes

To analyze test images, we fix the part and object assignments from the final training Gibbs sampling iteration, and then run the test image Gibbs sampler for 50 iterations from each of ten initializations. Given the transformed object instances created at each test iteration, we use (59) to estimate the posterior probability that test features were generated by each category, and average the probabilities from different samples to produce segmentations.

Figure 22 illustrates feature segmentations for several typical test street scenes, and transformed object instances corresponding to one iteration of the Gibbs sampler. In contrast with the fixed-order model of Sect. 6, TDPs allow each object category to occur at multiple locations within a single image. This allows the TDP to correctly find multiple cars in several scenes where the fixed-order model only detects a single car. Conversely, because the TDP does not model object relationships, it sometimes incorrectly detects cars in textured regions of buildings. The fixed-order model's contextual Gaussian prior suppresses these false alarms by forcing cars to lie beneath buildings.

We show similar segmentation results for office scenes in Fig. 23. Computer screens are typically reliably detected, particularly by the multiple-part TDP model. Perhaps surprisingly, mice are also detected with reasonable accuracy, although there are more false alarms than with the contextual model. In addition to accurately segmenting screen features, the part-based TDP model correctly associates a single transformed object cluster with most screen instances. In contrast, the weaker appearance model of the single-part TDP causes it to create several transformed clusters for many computer screens, and thereby incorrectly label adjacent background features.

As confirmed by the ROC curves of Fig. 19, both TDP models improve significantly on the bag of features model. For large, rigid objects like computer screens and keyboards, including parts further increases recognition performance. The two TDP models perform similarly when segmenting cars, perhaps due to their lower typical resolution. However, the street scene interpretations illustrated in Fig. 22 show that the part-based TDP does a better job of *counting* the true number of car instances depicted in each image. While including parts leads to more intuitive global models of textural categories, for these simple datasets it does not improve segmentation accuracy.

Comparing the TDP's performance to the fixed-order scene model (see Fig. 19), we find that their complementary strengths are useful in different situations. For example, the fixed-order model's very strong spatial prior leads to improved building and road detection, but worse performance for the less structured features composing trees. The TDP more cleanly segments individual cars from the background, but also makes additional false alarms in contextually implausible regions of buildings; the overall performance of the two models is comparable. Mouse detection performance is also similar, because the rigid contextual prior cannot find mice which are not to the right of a computer screen. For computer screens, however, the TDP's allowance for multiple instances, and creation of additional parts to form a stronger appearance model, leads to significant performance improvements. Finally, we emphasize that the TDP also estimates the *number* of objects composing each scene, a task which is beyond the scope of the fixed-order model.

10 Discussion

The hierarchical models developed in this paper are designed to capture the complex structure of multiple object scenes. We provide a framework for integrating spatial relationships with "bag of features" models, and show that this leads to significant gains in recognition performance. In addition, by coupling transformations with nonparametric

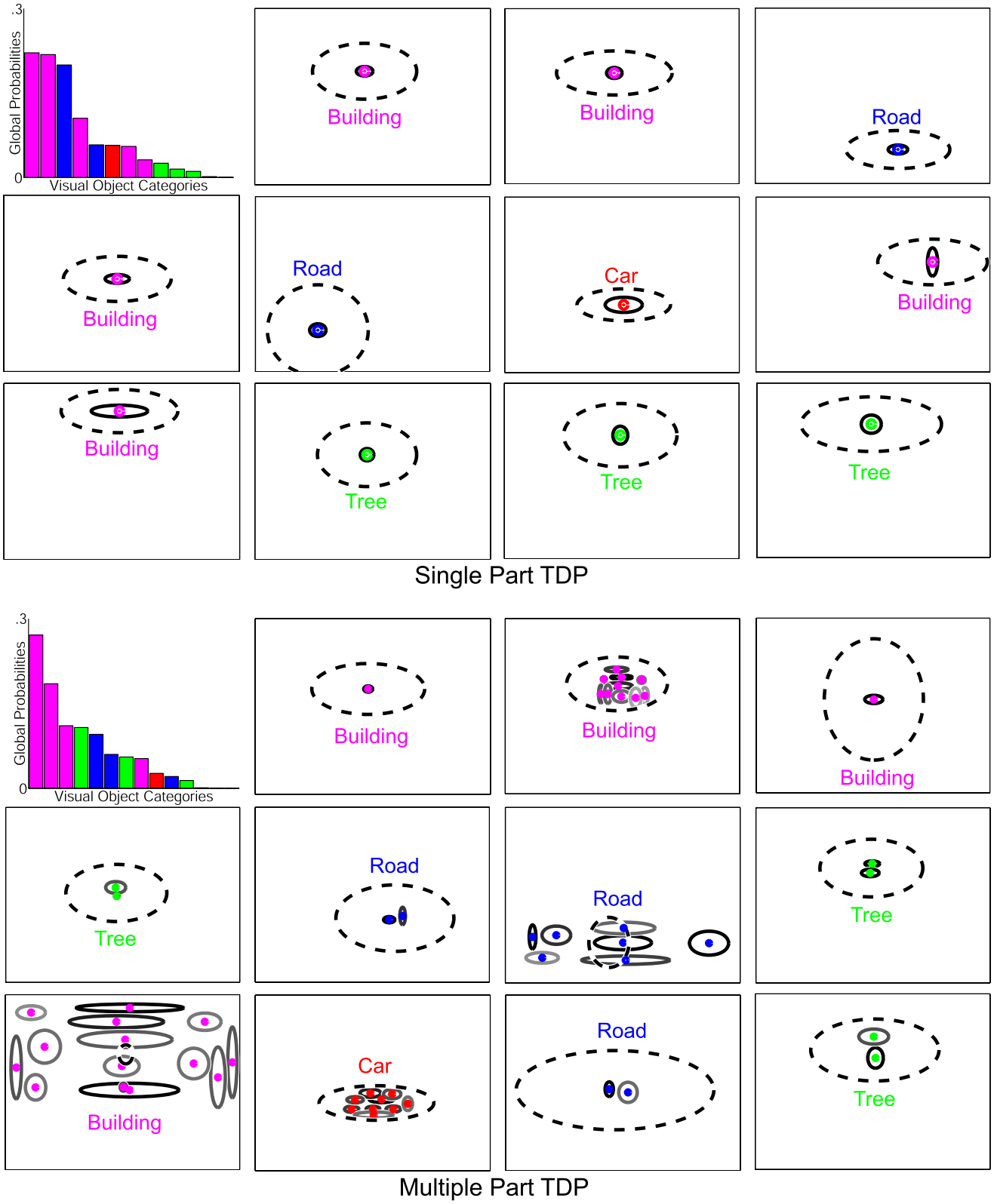

Figure 20 Learned TDP models for street scenes containing cars (*red*), buildings (*magenta*), roads (*blue*), and trees (*green*). *Top*: Simplified, single-part TDP in which the shape of each visual category is described by a single Gaussian (*solid ellipses*). We show the 11 most common visual categories at their mean positions, and also plot their transformation covariances (*dashed ellipses*). *Bottom*: Multiple-part TDP in which the number of parts (*solid ellipses*, intensity proportional to probability) underlying each category is learned automatically. We again show the 11 most probable categories, and their Gaussian transformation distributions (*dashed ellipses*)

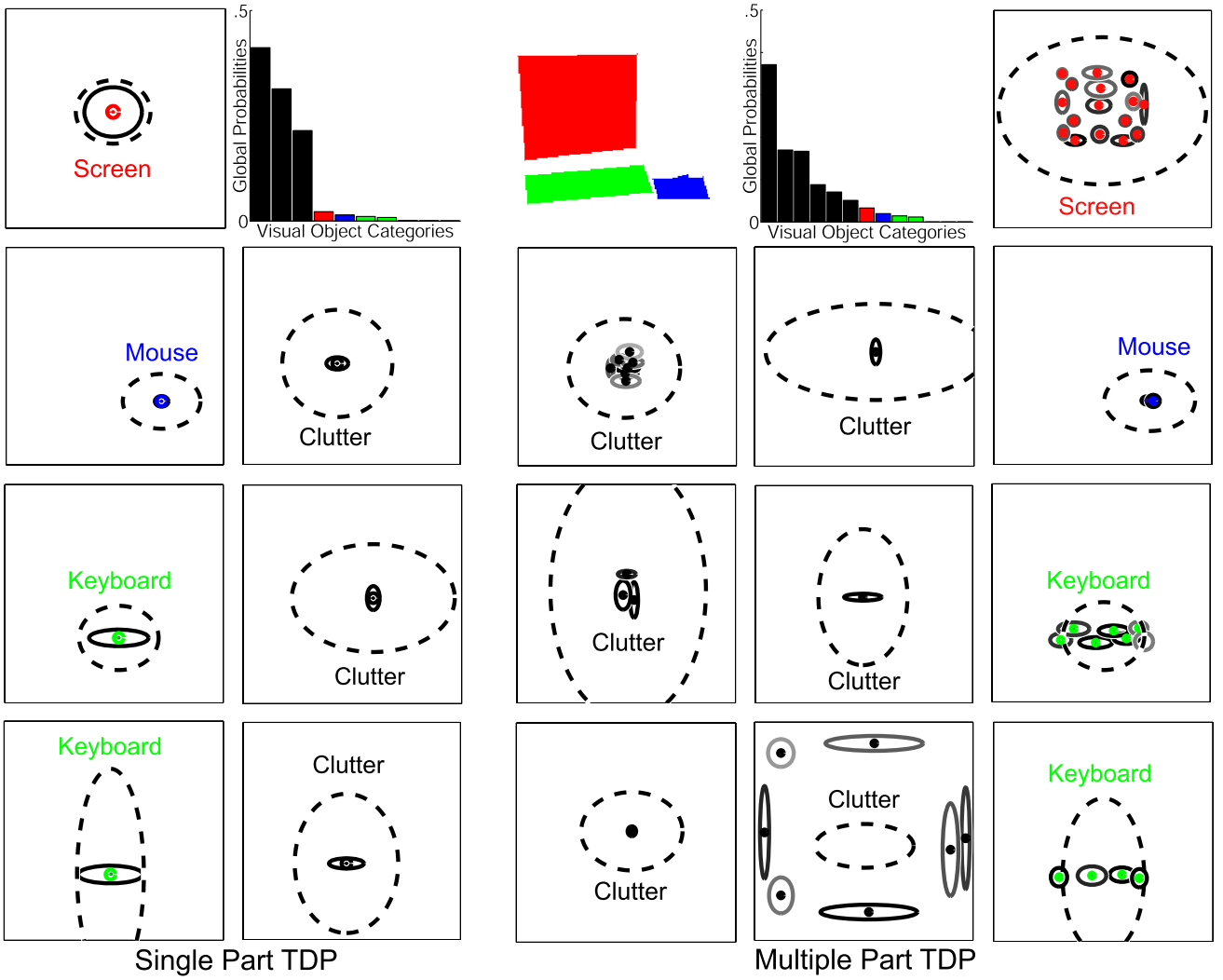

Figure 21 Learned TDP models for office scenes containing computer screens (*red*), keyboards (*green*), mice (*blue*), and background clutter (*black*). *Left*: Simplified, single-part TDP in which the shape of each visual category is described by a single Gaussian (*solid ellipses*). We show the 7 most common visual categories at their mean positions, and also plot their transformation covariances (*dashed ellipses*). *Right*: Multiple-part TDP in which the number of parts (*solid ellipses*, intensity proportional to probability) underlying each category is learned automatically. We show the 10 most probable categories, and their Gaussian transformation distributions (*dashed ellipses*)

prior distributions, the transformed Dirichlet process (TDP) allows us to reason consistently about the number of objects depicted in a given image. By addressing these issues in a generative framework, we retain an easily extendable, modular structure, and exploit partially labeled datasets. Furthermore, our nonparametric approach leads to expressive part-based models whose complexity grows as more images are observed.

Interestingly, the pair of scene models analyzed by this paper have complementary strengths. The fixed-order model learns contextual relationships among object categories and uses parts to describe objects' internal structure, but assumes that the number of parts and objects is known. In contrast, the TDP models unknown numbers of visual categories, object instances, and parts, but ignores contextual relationships. Our experimental results suggest that a model which balances the TDP's flexibility with additional global structure would prove even more effective.

More generally, the TDP framework can accommodate far richer classes of transformations. Natural candidates include spatial rotation and scaling, and also appearance transformations, which could be used to account for lighting or texture variations. In recent work building on this paper, we developed a variant of the TDP which infers three-dimensional scene structure from the predictable geometry of known objects (Sudderth et al. 2006a). Nonparametric methods may also play an important role in the design of

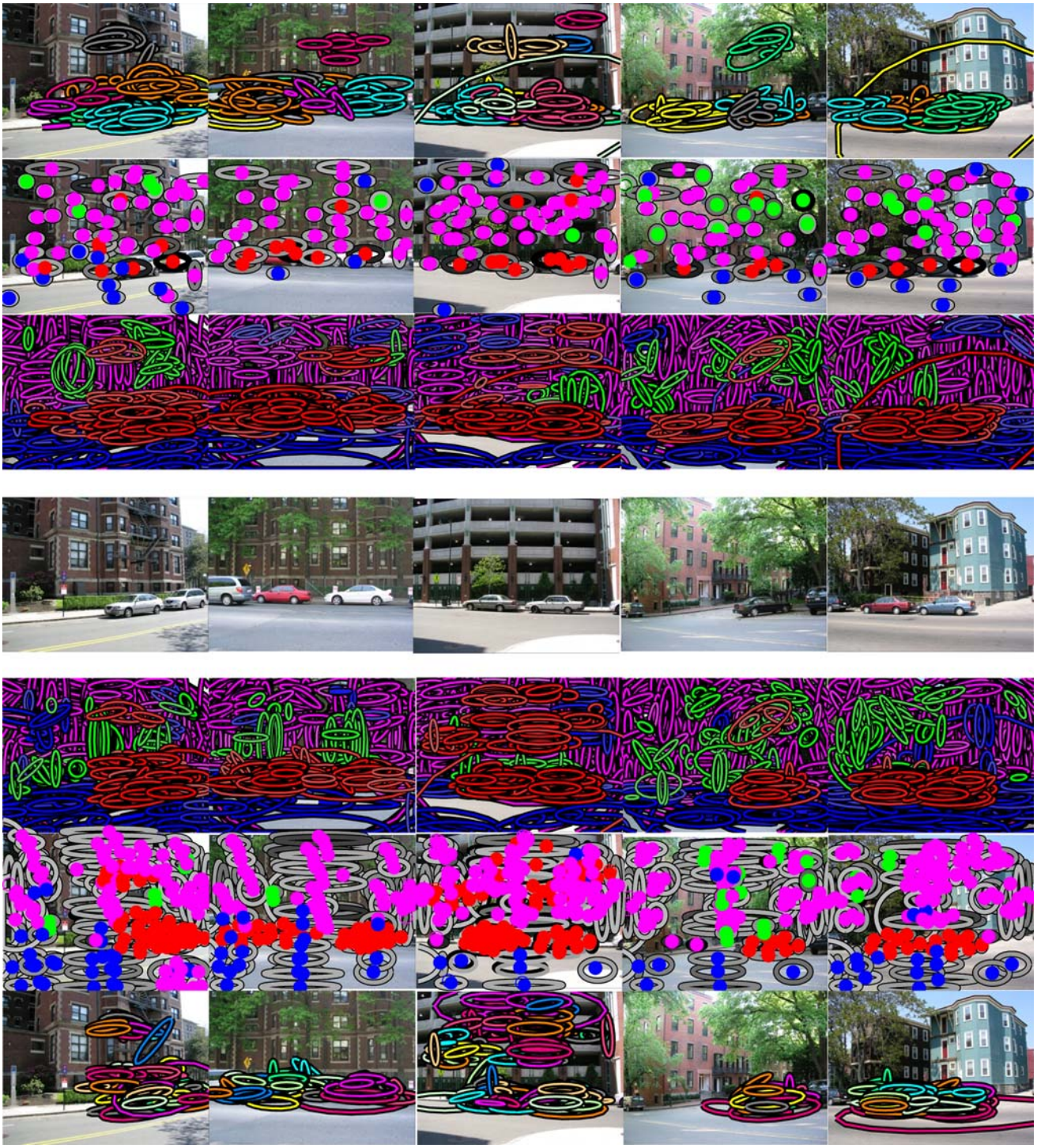

Figure 22 Feature segmentations produced by TDP models of street scenes containing cars (*red*), buildings (*magenta*), roads (*blue*), and trees (*green*). We compare a simplified TDP model which describes object shape via a single Gaussian cluster (*top rows*) to the full, multiple-part TDP model (*bottom rows*) of Fig. 15. *Row 4*: Five test images. *Rows 3 & 5*: Segmentations for each model, in which features are assigned to the object category with the highest posterior probability. *Rows 2 & 6*: Parts corresponding to the objects instantiated at a single Gibbs sampling iteration. *Rows 1 & 7*: Color-coded assignments of features to different parts and instances of the car category

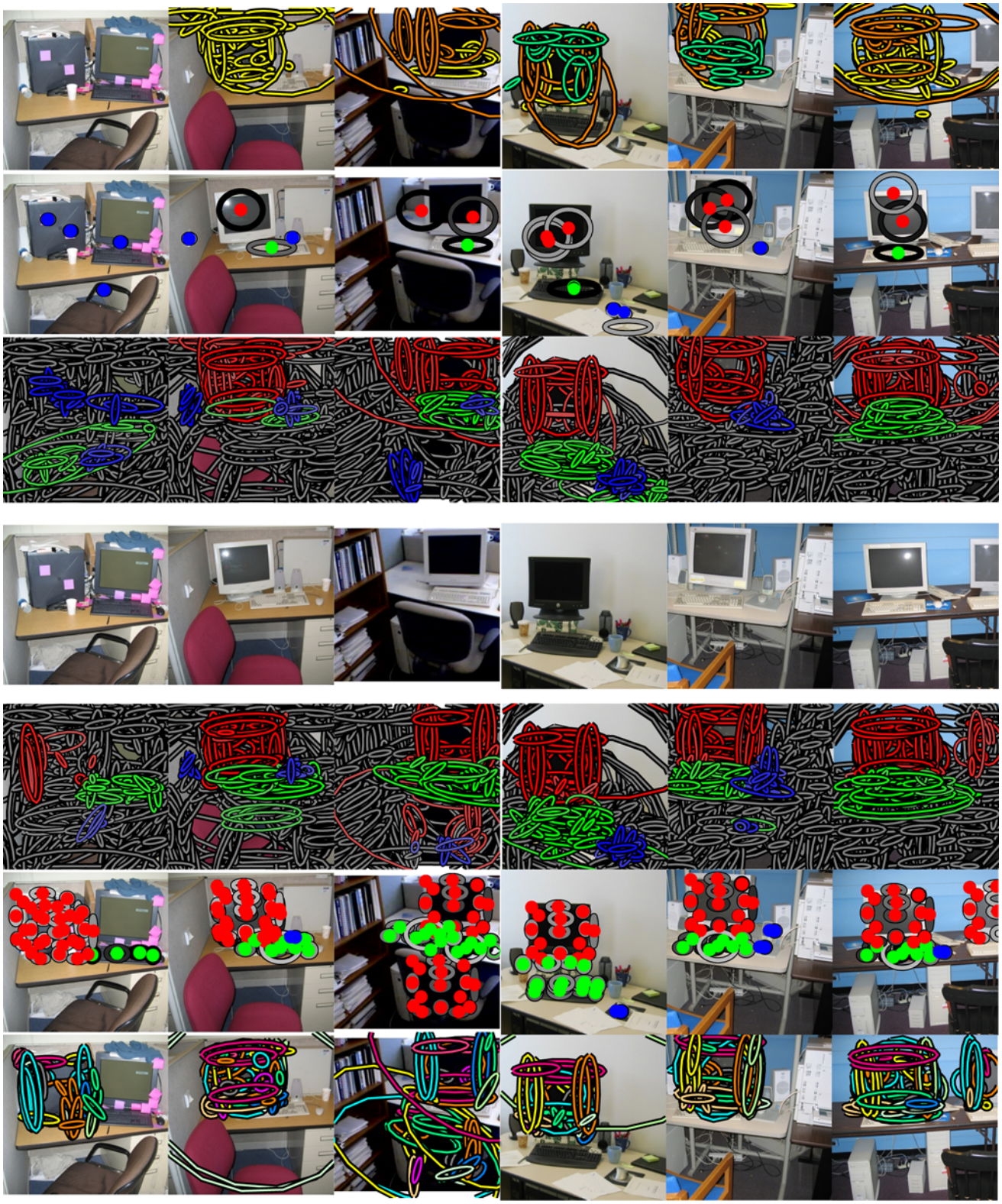

Figure 23 Feature segmentations produced by TDP models of office scenes containing computer screens (*red*), keyboards (*green*), mice (*blue*), and background clutter (*gray*). We compare a simplified TDP model which describes object shape via a single Gaussian cluster (*top rows*) to the full, multiple-part TDP model (*bottom rows*) of Fig. 15. *Row 4*: Six test images. *Rows 3 & 5*: Segmentations for each model, in which features are assigned to the object category with the highest posterior probability. *Rows 2 & 6*: Parts corresponding to the objects instantiated at a single Gibbs sampling iteration (background clutter not shown). *Rows 1 & 7*: Color-coded assignments of features to different parts and instances of the screen category

models which share more expressive, multi-layer structures among object categories.

Acknowledgements The authors thank Josh Tenenbaum, Daniel Huttenlocher, and the anonymous reviewers for helpful suggestions. This research supported in part by MURIs funded through AFOSR Grant FA9550-06-1-0324 and ARO Grant W911NF-06-1-0076.

Appendix 1 Learning with Conjugate Priors

Let $f(x|\theta)$ denote a family of probability densities, parameterized by θ, and $h(\theta|\lambda)$ a corresponding prior for the generative process. This prior is itself a member of some family with *hyperparameters* λ. Such priors are *conjugate* to $f(x|\theta)$ if, for any N independent observations $\{x_i\}_{i=1}^N$ and hyperparameters λ, the posterior distribution remains in the same family:

$$p(\theta|x_1,\ldots,x_N,\lambda) \propto h(\theta|\lambda)\prod_{i=1}^{N} f(x_i|\theta) \propto h(\theta|\bar{\lambda}). \tag{32}$$

The posterior distribution is then compactly described by an updated set of hyperparameters $\bar{\lambda}$. Conjugate priors exist for any regular *exponential family* $f(x|\theta)$ of probability distributions (Gelman et al. 2004; Sudderth 2006), and lead to efficient learning algorithms based on *sufficient statistics* of observed data.

11.1 Dirichlet Analysis of Multinomial Observations

Let x be a discrete random variable taking one of K categorical values, and $\pi_k \triangleq \Pr[x=k]$. A set of N independent samples $\{x_i\}_{i=1}^N$ then follow the *multinomial* distribution:

$$\begin{aligned} p(x_1,\ldots,x_N|\pi_1,\ldots,\pi_K) &= \frac{N!}{\prod_k C_k!}\prod_{k=1}^{K}{\pi_k}^{C_k}, \\ C_k &\triangleq \sum_{i=1}^{N}\delta(x_i,k). \end{aligned} \tag{33}$$

Counts C_k of the frequency of each category provide sufficient statistics for maximum likelihood (ML) parameter estimates $\hat{\pi}_k = C_k/N$. However, such unregularized estimates may be inaccurate unless $N \gg K$. The *Dirichlet* distribution (Gelman et al. 2004) is the multinomial's conjugate prior:

$$\mathrm{Dir}(\pi;\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}\prod_{k=1}^{K}{\pi_k}^{\alpha_k-1}, \quad \alpha_k > 0. \tag{34}$$

The Dirichlet's mean is $\mathbb{E}_\alpha[\pi_k] = \alpha_k/\alpha_0$, where $\alpha_0 \triangleq \sum_k \alpha_k$. Its variance is inversely proportional to this *precision* parameter α_0. We sometimes use $\mathrm{Dir}(\alpha_0)$ to denote a Dirichlet prior with symmetric parameters $\alpha_k = \alpha_0/K$. When $K=2$, the Dirichlet is equivalent to the *beta* distribution (Gelman et al. 2004).

Given N observations from a multinomial distribution with Dirichlet prior $\pi \sim \mathrm{Dir}(\alpha)$, the parameters' posterior distribution is $\mathrm{Dir}(\alpha_1+C_1,\ldots,\alpha_K+C_K)$, where C_k are counts as in (33). In the Monte Carlo algorithms developed in this paper, the *predictive likelihood* of a new observation $\bar{x} \sim f(x|\pi)$ is used to reassign visual features to objects or parts:

$$\begin{aligned} p(\bar{x}=k|x_1,\ldots,x_N,\alpha) &= \int_\Pi f(\bar{x}|\pi)\,p(\pi|x_1,\ldots,x_N,\alpha)\,d\pi \\ &= \frac{C_k+\alpha_k}{N+\alpha_0}. \end{aligned} \tag{35}$$

This prediction smooths the raw frequencies underlying the ML estimate by the *pseudo-counts* contributed by the Dirichlet prior. More generally, the predictive likelihood of multiple categorical observations can be expressed as a ratio of gamma functions (Griffiths and Steyvers 2004; Gelman et al. 2004).

11.2 Normal-Inverse-Wishart Analysis of Gaussian Observations

Consider a continuous-valued random variable x taking values in $\mathbb{R}^d$. A *Gaussian* or *normal* distribution (Gelman et al. 2004) with mean μ and positive definite covariance matrix Λ equals

$$\begin{aligned} &\mathcal{N}(x;\mu,\Lambda) \\ &\quad = \frac{1}{(2\pi)^{d/2}|\Lambda|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Lambda^{-1}(x-\mu)\right\}. \end{aligned} \tag{36}$$

The sums of observations and their outer products, or equivalently the sample mean and covariance, provide sufficient statistics of Gaussian data. The conjugate prior for the covariance of a zero-mean Gaussian is the *inverse-Wishart* $\mathcal{W}(\nu,\Delta)$, a multivariate extension of the scaled inverse-χ^2 density (Gelman et al. 2004). Its strength is determined by the degrees of freedom $\nu > d$, interpreted as the size of a pseudo-dataset with covariance Δ. If a Gaussian's mean μ is also uncertain, we take $\Lambda \sim \mathcal{W}(\nu,\Delta)$ and $\mu \sim \mathcal{N}(\vartheta, \Lambda/\kappa)$. Here, ϑ is the expected mean, for which we have κ pseudo-observations on the scale of observations $x \sim \mathcal{N}(\mu,\Lambda)$. The resulting *normal-inverse-Wishart* prior (Gelman et al. 2004) equals

$$\begin{aligned} &\mathcal{NW}(\mu,\Lambda;\kappa,\vartheta,\nu,\Delta) \\ &\quad \propto |\Lambda|^{-(\frac{\nu+d}{2}+1)} \\ &\qquad \times \exp\left\{-\frac{1}{2}\mathrm{tr}(\nu\Delta\Lambda^{-1}) - \frac{\kappa}{2}(\mu-\vartheta)^T\Lambda^{-1}(\mu-\vartheta)\right\}. \end{aligned} \tag{37}$$

Note that the mean and variance parameters (μ, Λ) are dependent, so that means which differ significantly from ϑ typically have larger associated variance (Gelman et al. 2004; Sudderth 2006).

Given N observations $x_i \sim \mathcal{N}(\mu, \Lambda)$ from a Gaussian with prior $(\mu, \Lambda) \sim \mathcal{NW}(\kappa, \vartheta, \nu, \Delta)$, the posterior distribution is also normal-inverse-Wishart, with updated hyperparameters $(\bar{\kappa}, \bar{\vartheta}, \bar{\nu}, \bar{\Delta})$:

$$\bar{\kappa}\bar{\vartheta} = \kappa\vartheta + \sum_{i=1}^{N} x_i, \quad \bar{\kappa} = \kappa + N, \tag{38}$$

$$\bar{\nu}\bar{\Delta} = \nu\Delta + \sum_{i=1}^{N} x_i x_i^T + \kappa\vartheta\vartheta^T - \bar{\kappa}\bar{\vartheta}\bar{\vartheta}^T, \quad \bar{\nu} = \nu + N. \tag{39}$$

To robustly determine these posterior parameters, we cache the observations' sum (see (38)), and the Cholesky decomposition of the sum of observation outer products (see (39)). Marginalizing over posterior uncertainty in the true Gaussian parameters, the predictive likelihood of a new observation $\bar{x} \sim \mathcal{N}(\mu, \Lambda)$ is multivariate Student-t with $(\bar{\nu} - d + 1)$ degrees of freedom (Gelman et al. 2004). Assuming $\bar{\nu} \gg d$, this density is well approximated by a moment-matched Gaussian (Sudderth 2006):

$$p(\bar{x}|x_1, \ldots, x_N, \kappa, \vartheta, \nu, \Delta) \approx \mathcal{N}\left(\bar{x}; \bar{\vartheta}, \frac{(\bar{\kappa}+1)\bar{\nu}}{\bar{\kappa}(\bar{\nu}-d-1)}\bar{\Delta}\right). \tag{40}$$

The predictive likelihood thus depends on *regularized* estimates of the sample mean and covariance.

Appendix 2 Posterior Inference via Gibbs Sampling

This appendix provides partial derivations for the Gibbs samplers used in earlier sections of this paper. Our algorithms combine and generalize previous Monte Carlo methods for Gaussian hierarchical models (Gelman et al. 2004), variants of LDA (Griffiths and Steyvers 2004; Rosen-Zvi et al. 2004), DP mixtures (Escobar and West 1995; Neal 2000), and the HDP (Teh et al. 2006).

12.1 Hierarchical Dirichlet Process Object Appearance Model

We first examine the HDP object appearance model of Sect. 4.2, and use the HDP's Chinese restaurant franchise representation (Teh et al. 2006) to derive Algorithms 1–2. To avoid cumbersome notation, let $z_{ji} = k_{o_j t_{ji}}$ denote the global part associated with feature (w_{ji}, v_{ji}). Note that z_{ji} is uniquely determined by that feature's table assignment $t_{ji} = t$, and the corresponding table's part assignment $k_{\ell t}$.

Table Assignment Resampling Consider the table assignment t_{ji} for feature (w_{ji}, v_{ji}), given all other variables. Letting $\mathbf{t}_{\setminus ji}$ denote all table assignments excluding t_{ji}, Fig. 5 implies that

$$\begin{aligned} &p(t_{ji}|\mathbf{t}_{\setminus ji}, \mathbf{k}, \mathbf{w}, \mathbf{v}, \mathbf{o}, \boldsymbol{\rho}) \\ &\quad \propto p(t_{ji}|\mathbf{t}_{\setminus ji}, o_j)\, p(w_{ji}|\mathbf{t}, \mathbf{k}, \mathbf{w}_{\setminus ji}) \\ &\qquad \times p(v_{ji}|\mathbf{t}, \mathbf{k}, \mathbf{v}_{\setminus ji}, \boldsymbol{\rho}). \end{aligned} \tag{41}$$

Because samples from the Dirichlet process are exchangeable (Pitman 2002), we evaluate the first term by thinking of t_{ji} as the *last* in a sequence of N_j observations, so that it follows the Chinese restaurant franchise predictive rule of (13). The second and third terms of (41) are the predictive likelihood of the i^{th} feature's appearance w_{ji} and position v_{ji}. For existing tables t, the appearance likelihood is determined via counts C_{kw} of the number of times each visual word w is currently assigned to global part $k = k_{\ell t}$ (see Sect. 11.1). The position likelihood instead depends on statistics of the *relative* displacements of image features from the current reference transformations:

$$\begin{aligned} &p(v_{ji}|z_{ji} = k, \mathbf{t}_{\setminus ji}, \mathbf{k}, \mathbf{v}_{\setminus ji}, \boldsymbol{\rho}) \\ &\quad = \iint H_v(\mu_k, \Lambda_k) \\ &\qquad \times \prod_{j'i'|z_{j'i'}=k} \mathcal{N}(v_{j'i'}; \tau(\mu_k, \Lambda_k; \rho_{j'}))\, d\mu_k\, d\Lambda_k \\ &\quad \propto \iint H_v(\mu_k, \Lambda_k) \\ &\qquad \times \prod_{j'i'|z_{j'i'}=k} \mathcal{N}(\tilde{\tau}(v_{j'i'}; \rho_{j'}); \mu_k, \Lambda_k)\, d\mu_k\, d\Lambda_k. \end{aligned} \tag{42}$$

Here, the data transformation of (2) allows us to describe all observations of part k in a common coordinate frame. Because H_v is normal-inverse-Wishart, the predictive likelihood of (42) is multivariate Student-t. We approximate this via a Gaussian $\mathcal{N}(v_{ji}; \hat{\mu}_k, \hat{\Lambda}_k)$, with parameters determined via regularized moment-matching of *transformed* observations $\tilde{\tau}(v_{ji}; \rho_j)$ as in (40). For compactness, we define $(\hat{\mu}_k, \hat{\Lambda}_k) \oplus v_{ji}$ to be an operator which updates a normal-inverse-Wishart posterior based on a new feature v_{ji} (see (38, 39)). Similarly, $(\hat{\mu}_k, \hat{\Lambda}_k) \ominus v_{ji}$ removes v_{ji} from the posterior statistics of part k. Algorithm 1 uses these operators to recursively update likelihood statistics as table assignments and transformations change.

When computing the likelihood of new tables, Algorithm 1 marginalizes over potential global part assignments (Teh et al. 2006). If a new table is instantiated $(t_{ji} = \bar{t})$, we also choose a corresponding global part $k_{\ell\bar{t}}$. Exchangeability again implies that this assignment is biased by the number of other tables M_k assigned to each global part, as in the Chinese restaurant franchise of (14).

Reference Transformation Resampling Fixing all assignments $(\mathbf{t}, \mathbf{k})$, each feature is associated with a unique global part. While marginalization of part parameters (μ_k, Λ_k) improves efficiency when resampling feature assignments, it complicates transformation resampling. We thus employ an *auxiliary variable* method (Neal 2000), and sample a single position parameter from the posterior of each part associated with at least one observation:

$$(\hat{\mu}_k, \hat{\Lambda}_k) \sim p(\mu_k, \Lambda_k | \{(v_{ji} - \rho_j) | z_{ji} = k\}), \quad k = 1, \ldots, K. \tag{43}$$

Sampling from these normal-inverse-Wishart distributions (see Sect. 11.2) is straightforward (Gelman et al. 2004). To determine the current transformation prior for object category $o_j = \ell$, we similarly sample $(\hat{\zeta}_\ell, \hat{\Upsilon}_\ell)$ given fixed transformations $\{\rho_{j'} | o_{j'} = \ell\}$ for all other images of object ℓ.

Given these auxiliary part parameters, and assuming transformations are chosen to translate image features as in (3), the posterior distribution for transformation ρ_j factors as follows:

$$p(\rho_j | o_j = \ell, \mathbf{t}, \mathbf{k}, \mathbf{v}, \{\hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^K, \hat{\zeta}_\ell, \hat{\Upsilon}_\ell) \propto \mathcal{N}(\rho_j; \hat{\zeta}_\ell, \hat{\Upsilon}_\ell) \prod_{k=1}^{K} \prod_{i | z_{ji} = k} \mathcal{N}(v_{ji} - \rho_j; \hat{\mu}_k, \hat{\Lambda}_k). \tag{44}$$

Reference transformations for other images induce a Gaussian prior on ρ_j, while feature assignments in image j effectively provide Gaussian observations. The posterior transformation distribution is thus also Gaussian, with mean χ_j and covariance Ξ_j expressed in information form (Gelman et al. 2004):

$$\begin{aligned} \Xi_j^{-1} &= \hat{\Upsilon}_\ell^{-1} + \sum_{k=1}^{K} \sum_{i | z_{ji} = k} \hat{\Lambda}_k^{-1}, \\ \Xi_j^{-1} \chi_j &= \hat{\Upsilon}_\ell^{-1} \hat{\zeta}_\ell + \sum_{k=1}^{K} \sum_{i | z_{ji} = k} \hat{\Lambda}_k^{-1} (v_{ji} - \hat{\mu}_k). \end{aligned} \tag{45}$$

Note that Ξ_j^{-1} adds one multiple of $\hat{\Lambda}_k^{-1}$ for each feature assigned to part k. After resampling $\rho_j \sim \mathcal{N}(\chi_j, \Xi_j)$, the auxiliary part and transformation parameters are discarded. Because our datasets have many training images, these auxiliary variables are well approximated by modes of their corresponding normal-inverse-Wishart posteriors. For simplicity, Algorithm 1 thus directly uses the Gaussian parameters implied by cached statistics when resampling transformations.

Global Part Assignment Resampling We now consider the assignments $k_{\ell t}$ of tables to global parts, given fixed associations $\mathbf{t}$ between features and tables. Although each category ℓ has infinitely many tables, we only explicitly sample assignments for the T_ℓ tables occupied by at least one feature ($N_{\ell t} > 0$). Because $k_{\ell t}$ determines the part for all features assigned to table t, its posterior distribution depends on their joint likelihood (Teh et al. 2006). Let $\mathbf{w}_{\ell t} = \{w_{ji} | t_{ji} = t, o_j = \ell\}$ denote the appearance features for table t, and $\mathbf{w}_{\backslash \ell t}$ all other features. Defining $\mathbf{v}_{\ell t}$ and $\mathbf{v}_{\backslash \ell t}$ similarly, we have

$$p(k_{\ell t} | \mathbf{k}_{\backslash \ell t}, \mathbf{t}, \mathbf{w}, \mathbf{v}, \boldsymbol{\rho}) \propto p(k_{\ell t} | \mathbf{k}_{\backslash \ell t}) p(\mathbf{w}_{\ell t} | \mathbf{t}, \mathbf{k}, \mathbf{w}_{\backslash \ell t}) p(\mathbf{v}_{\ell t} | \mathbf{t}, \mathbf{k}, \mathbf{v}_{\backslash \ell t}, \boldsymbol{\rho}). \tag{46}$$

Via exchangeability, the first term follows from the Chinese restaurant franchise of (14). The joint likelihood of $\mathbf{w}_{\ell t}$ is determined by those features assigned to the same part:

$$p(\mathbf{w}_{\ell t} | k_{\ell t} = k, \mathbf{t}, \mathbf{k}_{\backslash \ell t}, \mathbf{w}_{\backslash \ell t}) \propto \int p(\eta_k | \{w_{j'i'} | z_{j'i'} = k, t_{j'i'} \neq t\}) \times \prod_{j,i | t_{ji} = t} p(w_{ji} | \eta_k) d\eta_k. \tag{47}$$

As discussed in Sect. 11.1, this likelihood has a closed form for conjugate Dirichlet priors. The likelihood of $\mathbf{v}_{\ell t}$ has a similar form, except that part statistics are determined by transformed feature positions as in (42). Evaluating these likelihoods for each of the K currently instantiated parts, as well as a potential new global part $\bar{k}$, we may then resample $k_{\ell t}$ as summarized in Algorithm 2.

Concentration Parameter Resampling The preceding sampling equations assumed fixed values for the concentration parameters γ and α defining the HDP's stick-breaking priors (see (9, 10)). In practice, these parameters noticeably impact the number of global and local parts learned by the Gibbs sampler. As with standard Dirichlet process mixtures (Escobar and West 1995), it is thus preferable to choose weakly informative gamma priors for these concentration parameters. Auxiliary variable methods may then be used to resample α and γ following each Gibbs iteration (Teh et al. 2006).

Likelihoods for Object Detection and Recognition To use our HDP object model for recognition tasks, we compute the likelihood that a test image j is generated by each candidate object category o_j. Because images are independently sampled from a common parameter set, we have

$$p(\mathbf{w}_j, \mathbf{v}_j | o_j, \mathcal{J}) = \int p(\mathbf{w}_j, \mathbf{v}_j | o_j, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\varphi}) p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \mathcal{J}) d\boldsymbol{\pi} d\boldsymbol{\theta} d\boldsymbol{\varphi}.$$

In this expression, $\mathcal{J}$ denotes the set of training images, $\boldsymbol{\theta} = \{\eta_k, \mu_k, \Lambda_k\}_{k=1}^{\infty}$ the part position and appearance parameters, and $\boldsymbol{\varphi} = \{\zeta_\ell, \Upsilon_\ell\}_{\ell=1}^{L}$ the reference transformation parameters. The Gibbs samplers of Algorithms 1 and 2 provide samples $(\mathbf{t}^{(a)}, \mathbf{k}^{(a)}, \boldsymbol{\rho}^{(a)})$ approximately distributed according to $p(\mathbf{t}, \mathbf{k}, \boldsymbol{\rho}|\mathcal{J})$. Given A such samples, we approximate the test image likelihood as

$$p(\mathbf{w}_j, \mathbf{v}_j|o_j, \mathcal{J}) \approx \frac{1}{A}\sum_{a=1}^{A} p(\mathbf{w}_j, \mathbf{v}_j|o_j, \boldsymbol{\pi}^{(a)}, \boldsymbol{\theta}^{(a)}, \boldsymbol{\varphi}^{(a)}). \tag{48}$$

Here, $(\boldsymbol{\pi}^{(a)}, \boldsymbol{\theta}^{(a)}, \boldsymbol{\varphi}^{(a)})$ denote parameters sampled from the posterior distributions induced by $(\mathbf{t}^{(a)}, \mathbf{k}^{(a)}, \boldsymbol{\rho}^{(a)})$, which have simple forms (Sudderth 2006; Teh et al. 2006; Ishwaran and James 2001).

In practice, we approximate the infinite stick-breaking process of (7) by only sampling parameters for the $K^{(a)}$ global parts to which $(\mathbf{t}^{(a)}, \mathbf{k}^{(a)})$ assigns at least one feature. Ignoring reference transformations, image features are then conditionally independent:

$$p(\mathbf{w}_j, \mathbf{v}_j|o_j = \ell, \boldsymbol{\pi}^{(a)}, \boldsymbol{\theta}^{(a)}) = \prod_{i=1}^{N_j}\sum_{k=1}^{K^{(a)}} \hat{\pi}_{\ell k}\hat{\eta}_k(w_{ji})\mathcal{N}(v_{ji}; \hat{\mu}_k, \hat{\Lambda}_k). \tag{49}$$

Here, $\boldsymbol{\theta}^{(a)} = \{\hat{\eta}_k, \hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^{K^{(a)}}$, and $\hat{\pi}_{\ell k}$ indicates the total weight assigned to global part k by the tables of object ℓ, as in (11). This expression calculates the likelihood of N_j features in $\mathcal{O}(N_j K^{(a)})$ operations. To account for reference transformations, we run the Gibbs sampler of Algorithm 1 on the test image, and then average the feature likelihoods implied by sampled transformations.

12.2 Fixed-Order Models for Objects and Scenes

In this section, we extend methods developed for the author-topic model (Rosen-Zvi et al. 2004) to derive a Gibbs sampler for the fixed-order visual scene model of Sect. 6.1. A special case of this sampler, as summarized in Algorithm 3, is also used for learning in the fixed-order, single object model of Sect. 3.2.

To improve convergence, we develop a blocked Gibbs sampler which jointly resamples the object o_{ji} and part z_{ji} associated with each feature. Fixing transformations ρ_j, Fig. 12 implies that

$$p(o_{ji}, z_{ji}|\mathbf{o}_{\setminus ji}, \mathbf{z}_{\setminus ji}, \mathbf{w}, \mathbf{v}, \mathbf{s}, \boldsymbol{\rho}) \propto p(o_{ji}|\mathbf{o}_{\setminus ji}, s_j)p(z_{ji}|\mathbf{z}_{\setminus ji}, o_{ji})p(w_{ji}|\mathbf{z}, \mathbf{w}_{\setminus ji}) \times p(v_{ji}|\mathbf{z}, \mathbf{v}_{\setminus ji}, \mathbf{o}, \boldsymbol{\rho}). \tag{50}$$

Because $\boldsymbol{\beta}_s \sim \mathrm{Dir}(\gamma)$ is assigned a Dirichlet prior, (35) shows that the first term depends on the number $M_{s\ell}$ of features that $\mathbf{o}_{\setminus ji}$ assigns to object ℓ in images of scene s. Similarly, because $\boldsymbol{\pi}_\ell \sim \mathrm{Dir}(\alpha)$, the second term depends on the number $N_{\ell k}$ of features simultaneously assigned to object ℓ and part k. Finally, the appearance and position likelihoods are identical to those in the HDP object model (see Sect. 12.1), except that each object ℓ has its own reference location $\rho_{j\ell}$. Note that features associated with different objects contribute to a common set of K shared parts.

We resample reference transformations ρ_j via an extension of the auxiliary variable method of Sect. 12.1. Given sampled parameters for parts $\{\hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^{K}$ and the $2L$-dim. reference prior distribution $(\hat{\zeta}_s, \hat{\Upsilon}_s)$, the posterior distribution of ρ_j factors as follows:

$$p(\rho_j|s_j = s, \mathbf{o}, \mathbf{z}, \mathbf{v}, \{\hat{\mu}_k, \hat{\Lambda}_k\}_{k=1}^{K}, \hat{\zeta}_s, \hat{\Upsilon}_s) \propto \mathcal{N}(\rho_j; \hat{\zeta}_s, \hat{\Upsilon}_s)\prod_{k=1}^{K}\prod_{i|z_{ji}=k}\mathcal{N}(v_{ji} - \rho_{jo_{ji}}; \hat{\mu}_k, \hat{\Lambda}_k). \tag{51}$$

Each feature v_{ji} provides a Gaussian observation of the *subvector* of ρ_j corresponding to its assigned object o_{ji}. Transformations thus have a Gaussian posterior, with mean χ_j and covariance Ξ_j:

$$\begin{aligned}
\Xi_j{}^{-1} &= \hat{\Upsilon}_{s_j}^{-1} + \mathrm{blkdiag}\left\{\sum_{k=1}^{K}\sum_{\substack{i|z_{ji}=k\\ o_{ji}=1}}\hat{\Lambda}_k^{-1}, \ldots, \sum_{k=1}^{K}\sum_{\substack{i|z_{ji}=k\\ o_{ji}=L}}\hat{\Lambda}_k^{-1}\right\},\\
\Xi_j{}^{-1}\chi_j &= \hat{\Upsilon}_{s_j}^{-1}\hat{\zeta}_{s_j} + \left[\sum_{k=1}^{K}\sum_{\substack{i|z_{ji}=k\\ o_{ji}=1}}\hat{\Lambda}_k^{-1}(v_{ji} - \hat{\mu}_k), \ldots, \sum_{k=1}^{K}\sum_{\substack{i|z_{ji}=k\\ o_{ji}=L}}\hat{\Lambda}_k^{-1}(v_{ji} - \hat{\mu}_k)\right]^T.
\end{aligned} \tag{52}$$

By caching statistics of features, we may then sample a new reference transformation $\rho_j \sim \mathcal{N}(\chi_j, \Xi_j)$ in $\mathcal{O}(L^3)$ operations. As in Sect. 12.1, Algorithm 3 approximates the auxiliary variables underlying this update by modes of the Gaussian parameters' normal-inverse-Wishart posteriors.

12.3 Transformed Dirichlet Process Mixtures

We now generalize the HDP Gibbs sampler of Sect. 12.1 to learn parameters for the TDP mixture model of Sect. 7.1. As summarized in Algorithm 4, we first fix assignments o_{jt} of tables to global clusters, and corresponding transforma-

tions ρ_{jt}. From the graphical TDP representation of Fig. 13, we have

$$p(t_{ji}|\mathbf{t}_{\setminus ji}, \mathbf{o}, \mathbf{v}, \boldsymbol{\rho}) \propto p(t_{ji}|\mathbf{t}_{\setminus ji}) p(v_{ji}|\mathbf{t}, \mathbf{o}, \mathbf{v}_{\setminus ji}, \boldsymbol{\rho}). \tag{53}$$

As in the HDP Gibbs sampler of Sect. 12.1, the Chinese restaurant process (see (13)) expresses the first term via the number N_{jt} of other observations currently assigned to each table. For existing tables, the likelihood term may then be evaluated by using the data transformation $\tilde{\tau}(v_{ji}; \rho_{jt_{ji}})$ to describe observations in a common coordinate frame. This approach is analogous to (42), except that the TDP indexes reference transformations by tables t rather than images j.

For new tables $\bar{t}$, we improve sampling efficiency by integrating over potential assignments $o_{j\bar{t}}$ to global clusters (Teh et al. 2006). As in Sects. 12.1 and 12.2, we assume fixed transformation parameters $\hat{\varphi}_\ell$ and observation parameters $\hat{\theta}_\ell$ for each of the L instantiated global clusters; these may either be sampled as auxiliary variables (Neal 2000), or approximated by corresponding posterior modes. Marginalizing over transformations $\rho_{j\bar{t}}$, the overall likelihood of a new table equals

$$\begin{aligned} &p(v_{ji}|t_{ji} = \bar{t}, \mathbf{o}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}}) \\ &\quad \propto \sum_\ell p(o_{j\bar{t}} = \ell|\mathbf{o}) \int_{\wp} f(\tilde{\tau}(v_{ji}; \rho)|\hat{\theta}_\ell) q(\rho|\hat{\varphi}_\ell) d\rho. \end{aligned} \tag{54}$$

The prior probability of each global cluster follows from the Chinese restaurant franchise prediction rule (see (14)). The integral of (54) is tractable when $\hat{\theta}_\ell = (\hat{\mu}_\ell, \hat{\Lambda}_\ell)$ parameterizes a Gaussian distribution, and $\hat{\varphi}_\ell = (\hat{\zeta}_\ell, \hat{\Upsilon}_\ell)$ a Gaussian prior on translations (see Sect. 3.1). We then have

$$\begin{aligned} &\int_{\wp} \mathcal{N}(v_{ji} - \rho; \hat{\mu}_\ell, \hat{\Lambda}_\ell) \mathcal{N}(\rho; \hat{\zeta}_\ell, \hat{\Upsilon}_\ell) d\rho \\ &\quad = \mathcal{N}(v_{ji}; \hat{\mu}_\ell + \hat{\zeta}_\ell, \hat{\Lambda}_\ell + \hat{\Upsilon}_\ell). \end{aligned} \tag{55}$$

For more complex transformations, numerical or Monte Carlo approximations may be needed.

A related approach is used to jointly resample assignments o_{jt} of tables to global clusters, and corresponding transformations ρ_{jt}, given fixed associations $\mathbf{t}$ between observations and tables:

$$\begin{aligned} p(o_{jt} = \ell, \rho_{jt}|\mathbf{o}_{\setminus jt}, \boldsymbol{\rho}_{\setminus jt}, \mathbf{t}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}}) &\propto p(o_{jt} = \ell|\mathbf{o}_{\setminus jt}) q(\rho_{jt}|\hat{\varphi}_\ell) \\ &\quad \times \prod_{i|t_{ji}=t} f(\tilde{\tau}(v_{ji}; \rho_{jt})|\hat{\theta}_\ell). \end{aligned} \tag{56}$$

Suppose again that $\hat{\theta}_\ell = (\hat{\mu}_\ell, \hat{\Lambda}_\ell)$, $\hat{\varphi}_\ell = (\hat{\zeta}_\ell, \hat{\Upsilon}_\ell)$ parameterize Gaussian distributions. Conditioning on this table's assignment to some global cluster $o_{jt} = \ell$, the posterior distribution of the transformation ρ_{jt} is Gaussian as in Sect. 12.1, with mean χ_{jt} and covariance Ξ_{jt} equaling

$$\begin{aligned} {\Xi_{jt}}^{-1} &= \hat{\Upsilon}_\ell^{-1} + \sum_{i|t_{ji}=t} \hat{\Lambda}_\ell^{-1}, \\ {\Xi_{jt}}^{-1} \chi_{jt} &= \hat{\Upsilon}_\ell^{-1} \hat{\zeta}_\ell + \sum_{i|t_{ji}=t} \hat{\Lambda}_\ell^{-1} (v_{ji} - \hat{\mu}_\ell). \end{aligned} \tag{57}$$

Using standard manipulations of Gaussian random variables, we may then marginalize ρ_{jt} to determine the overall likelihood that $o_{jt} = \ell$:

$$\begin{aligned} &p(o_{jt} = \ell|\boldsymbol{\rho}_{\setminus jt}, \mathbf{t}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}}) \\ &\quad \propto \left(\frac{|\Xi_{jt}|}{|\hat{\Lambda}_\ell|^{N_{jt}} |\hat{\Upsilon}_\ell|} \right)^{1/2} \\ &\qquad \times \exp\left\{ -\frac{1}{2} \sum_{i|t_{ji}=t} (v_{ji} - \hat{\mu}_\ell)^T \hat{\Lambda}_\ell^{-1} (v_{ji} - \hat{\mu}_\ell) \right. \\ &\qquad \left. - \frac{1}{2} \hat{\zeta}_\ell^T \hat{\Upsilon}_\ell^{-1} \hat{\zeta}_\ell + \frac{1}{2} {\chi_{jt}}^T {\Xi_{jt}}^{-1} \chi_{jt} \right\}. \end{aligned} \tag{58}$$

Note that we evaluate this expression with a *different* (χ_{jt}, Ξ_{jt}), computed as in (57), for each candidate global cluster ℓ. Step 3 of Algorithm 4 first uses this marginalized likelihood to choose o_{jt}, and then samples a corresponding transformation ρ_{jt} from the Gaussian of (57).

12.4 Transformed DP Models for Objects and Scenes

This section generalizes the TDP Gibbs sampler of Sect. 12.3 to learn parameters for the full TDP scene model of Sect. 8.1. Because visual categories are defined by different sets of parts, blocked resampling of instance and part assignments (t_{ji}, k_{ji}) is necessary. Figure 15 implies that

$$\begin{aligned} &p(t_{ji}, k_{ji}|\mathbf{t}_{\setminus ji}, \mathbf{k}_{\setminus ji}, \mathbf{w}, \mathbf{v}, \mathbf{o}, \boldsymbol{\rho}) \\ &\quad \propto p(t_{ji}|\mathbf{t}_{\setminus ji}) p(k_{ji}|\mathbf{k}_{\setminus ji}, \mathbf{t}, \mathbf{o}) p(w_{ji}|\mathbf{t}, \mathbf{k}, \mathbf{o}, \mathbf{w}_{\setminus ji}) \\ &\qquad \times p(v_{ji}|\mathbf{t}, \mathbf{k}, \mathbf{o}, \mathbf{v}_{\setminus ji}, \boldsymbol{\rho}). \end{aligned} \tag{59}$$

The first term encourages assignments to object instances t associated with many other features N_{jt}, as in (13). Similarly, the second term is derived from the stick-breaking prior $\boldsymbol{\varepsilon}_\ell \sim \text{GEM}(\kappa)$ on the probabilities associated with each visual category's parts:

$$\begin{aligned} &p(k_{ji}|t_{ji} = t, o_{jt} = \ell, \mathbf{k}_{\setminus ji}, \mathbf{t}_{\setminus ji}, \mathbf{o}_{\setminus jt}) \\ &\quad \propto \sum_{k=1}^{K_\ell} B_{\ell k} \delta(k_{ji}, k) + \kappa \delta(k_{ji}, \bar{k}). \end{aligned} \tag{60}$$

Here, $B_{\ell k}$ denotes the number of *other* features currently assigned to each of the K_ℓ instantiated parts of object ℓ, and $\bar{k}$ a potential new part. The appearance likelihood is as in

Sect. 12.1, except that we maintain counts $C_{\ell k w}$ of the number of times appearance descriptor w is assigned to each instantiated category ℓ and part k. Finally, our position likelihood computation extends the scheme of Algorithm 4 to cache statistics $(\hat{\mu}_{\ell k}, \hat{\Lambda}_{\ell k})$ of the transformed features for each category and part. To sample from (59), we evaluate these likelihoods for every existing part, and a potential new part, of each object instance. We also determine the likelihood of creating a new object instance by marginalizing potential category assignments and transformations as in (54), (55).

The second phase of our Gibbs sampler fixes object assignments $\mathbf{t}$, and considers potential reinterpretations of each instance t using a new global object category o_{jt}. Because parts and transformations are defined differently for each category, blocked resampling of $(o_{jt}, \rho_{jt}, \{k_{ji} | t_{ji} = t\})$ is necessary. As in Sect. 12.3, we resample transformations by instantiating auxiliary parameters for parts $(\hat{\eta}_{\ell k}, \hat{\mu}_{\ell k}, \hat{\Lambda}_{\ell k})$ and category-specific transformation priors $(\hat{\zeta}_{\ell}, \hat{\Upsilon}_{\ell})$. Suppose first that $o_{jt} = \ell$ is fixed. Due to the exponentially large number of joint assignments of this instance's features to parts, the marginal distribution of ρ_{jt} is intractable. However, given ρ_{jt}, part assignments k_{ji} have conditionally independent posteriors as in (41). Alternatively, given fixed part assignments for all features, ρ_{jt} follows the Gaussian posterior of (44), which arose in the single-object HDP sampler. Intuitively, fixing $\mathbf{t}$ effectively segments the scene's features into independent objects.

For each candidate visual category o_{jt}, we first perform a small number of auxiliary Gibbs iterations which alternatively sample part assignments $\{k_{ji} | t_{ji} = t\}$ and the transformation ρ_{jt}. Fixing the final ρ_{jt}, part assignments are then marginalized to compute the likelihood of o_{jt}. Typically, the posterior distribution of ρ_{jt} is tightly concentrated given fixed $\mathbf{t}$, and 3–5 auxiliary iterations provide an accurate approximation. Combining this likelihood with the global DP clustering bias of (14), we resample o_{jt}, and then conditionally choose $(\rho_{jt}, \{k_{ji} | t_{ji} = t\})$.

References

Adams, N. J., & Williams, C. K. I. (2003). Dynamic trees for image modelling. *Image and Vision Computing*, *21*, 865–877.

Amit, Y., & Trouvé, A. (2007). Generative models for labeling multi-object configurations in images. In J. Ponce, et al. (Ed.), *Toward category-level object recognition*. Berlin: Springer.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 509–522.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, MDL priors, and object recognition. In *Neural information processing systems 9* (pp. 838–844). Cambridge: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borenstein, E., & Ullman, S. (2002). Class-specific, top-down segmentation. In *European conference on computer vision* (Vol. 2, pp. 109–122).

Bosch, A., Zisserman, A., & Muñoz, X. (2006). Scene classification via pLSA. In *European conference on computer vision* (pp. 517–530).

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(6), 679–698.

Casella, G., & Robert, C. P. (1996). Rao–Blackwellisation of sampling schemes. *Biometrika*, *83*(1), 81–94.

Csurka, G., et al. (2004). Visual categorization with bags of keypoints. In *ECCV workshop on statistical learning in computer vision*.

De Iorio, M., Müller, P., Rosner, G. L., & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, *99*(465), 205–215.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 524–531).

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPR workshop on generative model based vision*.

Fergus, R., Fei-Fei, L., Perona, P., & Zisserman, A. (2005). Learning object categories from Google's image search. In *International conference on computer vision* (Vol. 2, pp. 1816–1823).

Fink, M., & Perona, P. (2004). Mutual boosting for contextual inference. In *Neural information processing systems 16*. Cambridge: MIT Press.

Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, *22*(1), 67–92.

Frey, B. J., & Jojic, N. (2003). Transformation-invariant clustering using the EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(1), 1–17.

Gelfand, A. E., Kottas, A., & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, *100*(471), 1021–1035.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. London: Chapman & Hall.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228–5235.

He, X., Zemel, R. S., & Carreira-Perpiñán, M. A. (2004). Multiscale conditional random fields for image labeling. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 695–702).

Helmer, S., & Lowe, D. G. (2004). Object class recognition with many local features. In *CVPR workshop on generative model based vision*.

Hinton, G. E., Ghahramani, Z., & Teh, Y. W. (2000). Learning to parse images. In *Neural information processing systems 12* (pp. 463–469). Cambridge: MIT Press.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Ishwaran, H., & Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, *12*, 941–963.

Jin, Y., & Geman, S. (2006). Context and hierarchy in a probabilistic image model. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 2145–2152).

Jojic, N., & Frey, B. J. (2001). Learning flexible sprites in video layers. In *IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 199–206).

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*(1), 140–155.

Jordan, M. I. (2005). Dirichlet processes, Chinese restaurant processes and all that. Tutorial at *Neural Information Processing Systems*.

Kovesi, P. (2005). MATLAB and Octave functions for computer vision and image processing. Available from http://www.csse.uwa.edu.au/~pk/research/matlabfns/.

LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 97–104).

Leibe, B., Leonardis, A., & Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *ECCV workshop on statistical learning in computer vision*.

Liter, J. C., & Bülthoff, H. H. (1998). An introduction to object recognition. *Zeitschrift für Naturforschung*, *53c*, 610–621.

Loeff, N., Arora, H., Sorokin, A., & Forsyth, D. (2006). Efficient unsupervised learning for localization and detection in object categories. In *Neural information processing systems 18* (pp. 811–818). Cambridge: MIT Press.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings section on Bayesian statistical science* (pp. 50–55). Alexandria: American Statistical Association.

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *British machine vision conference* (pp. 384–393).

Mikolajczyk, K., & Schmid, C. (2004). Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *International joint conference on artificial intelligence 19* (pp. 1352–1359)

Miller, E. G., & Chefd'hotel, C. (2003). Practical nonparametric density estimation on a transformation group for vision. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 114–121).

Miller, E. G., Matsakis, N. E., & Viola, P. A. (2000). Learning from one example through shared densities on transforms. In *IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 464–471).

Murphy, K., Torralba, A., & Freeman, W. T. (2004). Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Neural information processing systems 16*. Cambridge: MIT Press.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Pitman, J. (2002). *Combinatorial stochastic processes*. Technical Report 621, U.C. Berkeley Department of Statistics, August 2002.

Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2006). *The nested Dirichlet process*. Working Paper 2006-19, Duke Institute of Statistics and Decision Sciences.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Uncertainty in artificial intelligence 20* (pp. 487–494). Corvallis: AUAI Press.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2005). *LabelMe: A database and web-based tool for image annotation*. Technical Report 2005-025, MIT AI Lab.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–398.

Simard, P. Y., LeCun, Y. A., Denker, J. S., & Victorri, B. (1998). Transformation invariance in pattern recognition: Tangent distance and tangent propagation. In B. O. Genevieve & K. R. Müller (Eds.), *Neural networks: tricks of the trade* (pp. 239–274). Berlin: Springer.

Siskind, J. M., Sherman, J., Pollak, I., Harper, M. P., & Bouman, C. A. (2004, submitted). Spatial random tree grammars for modeling hierarchal structure in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., & Freeman, W. T. (2005). Discovering objects and their location in images. In *International conference on computer vision* (Vol. 1, pp. 370–377).

Storkey, A. J., & Williams, C. K. I. (2003). Image modeling with position-encoding dynamic trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(7), 859–871.

Sudderth, E. B. (2006). *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2005). Learning hierarchical models of scenes, objects, and parts. In *International conference on computer vision* (Vol. 2, pp. 1331–1338).

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2006a). Depth from familiar objects: A hierarchical model for 3D scenes. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 2410–2417).

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2006b). Describing visual scenes using transformed Dirichlet processes. In *Neural information processing systems 18* (pp. 1297–1304). Cambridge: MIT Press.

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

Tenenbaum, J. M., & Barrow, H. G. (1977). Experiments in interpretation-guided segmentation. *Artificial Intelligence*, *8*, 241–274.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, *53*(2), 169–191.

Torralba, A., Murphy, K. P., & Freeman, W. T. (2004). Sharing features: Efficient boosting procedures for multiclass object detection. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 762–769).

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S. C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, *63*(2), 113–140.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–687.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. In *European conference on computer vision* (pp. 18–32).

Williams, C. K. I., & Allan, M. (2006). *On a connection between object localization with a generative template of features and pose-space prediction methods*. Informatics Research Report 719, University of Edinburgh.