cess induces a clustering bias, and leads to efficient Monte Carlo methods which automatically learn the number of clusters underlying a particular set of observations [2, 11].

3.2. Transformed Dirichlet Processes

As we demonstrate later, the DP mixture of eq. (8) leads to effective part–based models for the internal geometry of rigid objects. More generally, we expect multiple object scenes to share local features, but differ significantly in their global spatial structure. The *hierarchical Dirichlet process* (HDP) [16] was developed to address the related problem of partially sharing topics among text documents. Applied to spatial data, the HDP chooses a globally shared mixture $G_0 \sim \mathrm{DP}(\gamma, H)$ as in eqs. (6, 7). Each image is then assigned a mixture $G_j \sim \mathrm{DP}(\alpha, G_0)$, reusing the same Gaussian clusters θ_ℓ in different proportions:

$$G_j(\theta) = \sum_{\ell=1}^{\infty} \pi_{j\ell}\delta(\theta, \theta_\ell) \qquad \boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}) \tag{9}$$

The global mixture weights $\boldsymbol{\beta}$ determine the expected cluster proportions $\boldsymbol{\pi}_j$, while α specifies the variability from image to image. This construction assumes that images differ only in the proportion of observed features for each spatial cluster, rather than the location and shape of those clusters. Because objects are not observed in consistent locations relative to the camera, a standard HDP would thus not adequately generalize to novel visual scenes.

Motivated by these difficulties, we consider a family of *transformations* $\tau(\theta; \rho)$ of the global mixture components θ, indexed by ρ. For spatial data, we associate these transforms [6] with shifts of the mean location of global clusters:

$$\tau(\mu, \Lambda; \rho) = (\mu + \rho, \Lambda) \tag{10}$$

The *transformed Dirichlet process* (TDP) [15] generalizes the HDP to more flexibly share spatial structure among images. The TDP is derived from *distributions over transformations* $q(\rho \mid \phi)$, indexed by $\phi \in \Phi$. Let R denote a prior measure on the space of transformation distributions Φ, which we later constrain to be zero–mean Gaussians.

We begin by augmenting the Dirichlet process stick–breaking construction of eq. (7) to define a global measure describing both parameters θ and transformations ρ:

$$G_0(\theta, \rho) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(\theta, \theta_\ell) q(\rho \mid \phi_\ell) \qquad \begin{matrix} \theta_\ell \sim H \\ \phi_\ell \sim R \end{matrix} \tag{11}$$

As before, $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$. Note that each cluster θ_ℓ has a different transformation distribution $q(\rho \mid \phi_\ell)$. We then independently sample $G_j \sim \mathrm{DP}(\alpha, G_0)$ for each image. Because samples from DPs are discrete with probability one, this joint measure can be written as

$$G_j(\theta, \rho) = \sum_{\ell=1}^{\infty} \pi_{j\ell}\delta(\theta, \theta_\ell) \left[\sum_{t=1}^{\infty} \lambda_{j\ell t}\delta(\rho, \rho_{j\ell t}) \right] \tag{12}$$

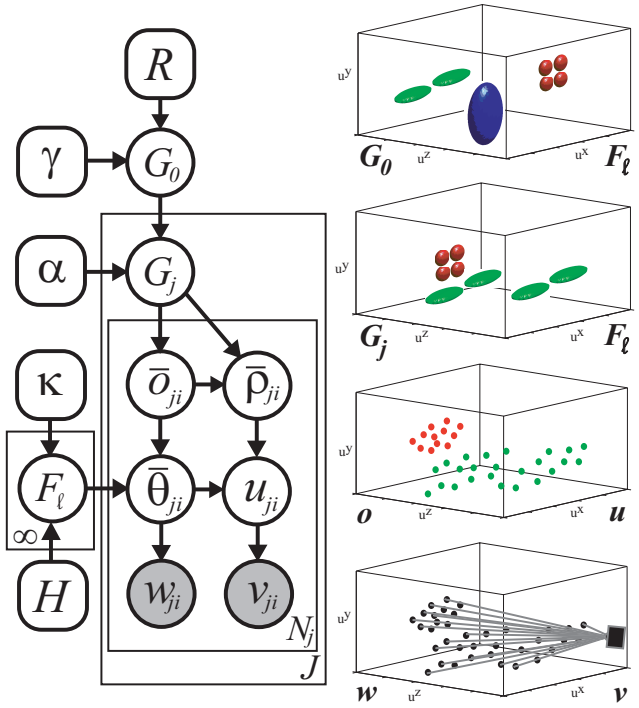

Figure 3. TDP model for 3D scenes (left), and cartoon illustration of the generative process (right). Global mixture G_0 describes the expected frequency and location of visual categories, whose internal structure is represented by part–based appearance models $\{F_\ell\}_{\ell=1}^{\infty}$. Each image mixture G_j instantiates a randomly chosen set of objects at transformed locations ρ. 3D feature positions u_{ji} are sampled from transformed parameters $\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji})$ corresponding to parts of object $\bar{o}_{ji}$. The camera observes projections v_{ji} of these features, with part–dependent appearance w_{ji}.

where $\sum_t \lambda_{j\ell t} = 1$. As $G_0(\theta, \rho)$ only has support at a discrete set of cluster parameters, $G_j(\theta, \rho)$ will associate many different transformations $\rho_{j\ell t}$ with each distinct θ_ℓ.

In the simplest case, each 3D feature in image j is now generated by sampling $(\bar{\theta}_{ji}, \bar{\rho}_{ji}) \sim G_j$, and then choosing $u_{ji} \sim \mathcal{N}(\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji}))$ from a transformed Gaussian [15]. Intuitively, global mixture components θ_ℓ define object geometry in a "canonical" coordinate frame, while the random set of transformations ρ determine the object instances within a particular scene. Critically, the TDP allows uncertainty in the *number* of objects depicted by each image. For instance, in the toy example of Fig. 3, the green object appears twice, while the blue does not appear at all.

3.3. Part–Based Object Appearance Models

Applied directly, the TDP model of eqs. (11, 12) describes the geometry of each global object cluster by a single Gaussian. This representation poorly captures the complex structure of many real objects, and does not model local variations in object appearance. In this section, we show how Dirichlet processes may also be adapted to learn richer,