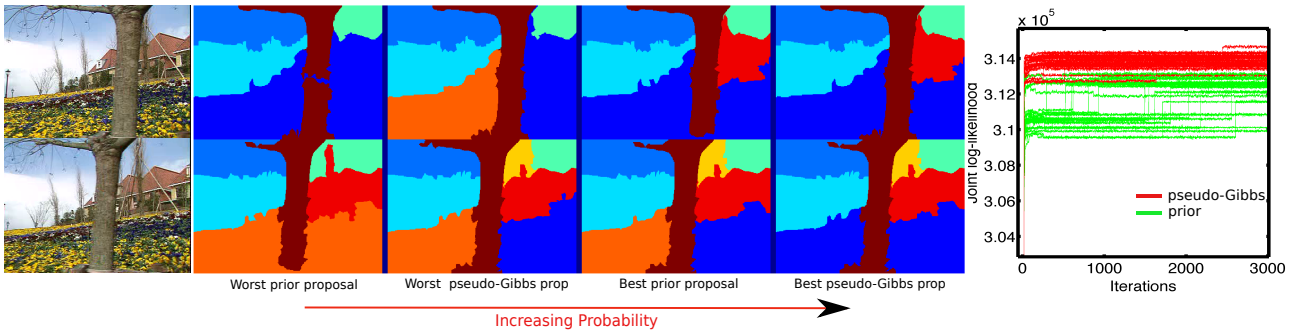

Figure 4: Data link proposal comparisons. *Left:* Two frames from the "garden" sequence, and partitions corresponding to the best and worst MAP samples using prior or pseudo-Gibbs proposals. *Right:* Joint log-likelihood trace plots for 25 trials of each proposal.

method proposed by Chang et al. (2013).[1] Each super-pixel is described by L_2 unit-normalized 120-bin HSV color and 128-bin local texton histograms. Unit normalization projects raw histograms to the surface of a hyper-sphere, where we use *von-Mises Fisher* (vMF) distributions (Mardia & Jupp, 2009) shared across all clusters of a global component. In preliminary experiments, we found that the vMF produced more accurate segmentations than multinomial models of raw histograms; similar L_2 normalizations are useful for image retrieval (Arandjelović & Zisserman, 2012). We also extracted optical flow using the "Classic+NL" algorithm (Sun et al., 2010), and associated a two-dimensional flow vector to each super-pixel, the median flow of its constituent pixels.

The color, texture, and flow features for super-pixel i in video frame g are denoted by $x_{gi} = \{x^c_{gi}, x^t_{gi}, x^f_{gi}\}$, where

$$x^c_{gi} \sim \text{vMF}(\mu^c_{z_{gi}}, \kappa^c), \mu^c_{z_{gi}} \sim \text{vMF}(\mu^c_0, \kappa^c_0), \qquad (12)$$

where κ^c, μ^c_0, and κ^c_0 are hyper-parameters controlling the concentration of color features around the direction $\mu^c_{z_{gi}}$, the mean color direction μ^c_0, and the concentration of $\mu^c_{z_{gi}}$ around μ^c_0. Texture features are generated similarly. Flow features are modeled via Gaussian distributions with conjugate, normal-inverse-Wishart priors:

$$x^f_{gi} \sim \mathcal{N}(\mu^{fg}_{z_{gi}}, \Sigma^{fg}_{z_{gi}}),\ \Sigma^{fg}_{z_{gi}} \sim \mathcal{IW}(n_0, S_0),$$
$$\mu^{fg}_{z_{gi}} \mid \Sigma^{fg}_{z_{gi}} \sim \mathcal{N}(\mu_0, \tau_0 \Sigma^{fg}_{z_{gi}}). \qquad (13)$$

Requiring all clusters in a global component, which may span several video frames, to share a single flow model is too restrictive. Instead we model the flow for each frame independently, requiring only that clusters in frame g assigned to the same component share a common flow model. Our model requires motion of a component to be locally (within a frame) coherent, but allows for large deviations between frames.[2] This assumption more closely reflects the motion statistics of objects in real videos.

Prior We used data affinities that encourage spatial neighbors not separated by strong intervening contours to connect to one another. We computed them by independently running the Pb edge detector (Martin et al., 2004) on each video frame and computing $A_{ij} = (1 - b_{ij})^3 \times \mathbf{1}[i, j]$ for each superpixel pair. Here, $0 \leq b_{ij} \leq 1$ is the maximum edge response along a straight line segment connecting the centers of superpixels i, j, and $\mathbf{1}[i, j]$ takes a value of 1 if i and j are spatial neighbors, and 0 otherwise.

Flow-based affinities, as in the earlier toy example, were used to specify the cluster affinity functions. All $\alpha_{1:G}$ and α_0 were set to 10^{-8}. The naive-hddCRP used identical data affinities and hyper-parameters, but used sequential distances between clusters (see Sec. 2.2). The hCRP used sequential affinities to govern both the data and cluster links. For a CRP, the expected number of clusters given N data points is roughly αlog(N). We set $\alpha_{1:G}$ such that the expected number of clusters in a video frame matches the number of observed ground truth clusters, and $\alpha_0 = 1$.

Data link proposals We compare the two data link proposals on 10 frames from the classic "garden" sequence. For each proposal, we ran 3000 iterations of 25 MCMC chains. The results, including MAP samples from the highest and lowest probability chains and log-likelihood trace plots, are summarized in Figure 4. The visualized MAP partitions demonstrate that all chains eventually reach reasonable configurations, but segmentations nevertheless improve qualitatively with increasing model likelihood. This suggests a correspondence between the biases captured by the hddCRP and the statistics of video partitions.

We find that pseudo-Gibbs proposals reach higher probability states more rapidly than prior proposals, and have much lower sensitivity to initialization. Overall, 24 of the 25 pseudo-Gibbs chains reach states that are more probable than the best prior proposal trial. Subsequent experiments thus focus solely on the superior pseudo-Gibbs proposal.

Empirical evaluation We compare our performance against a popular non-probabilistic *hierarchical graph-based video segmentation* (HGVS) algorithm (Grundmann et al., 2010), against the naive-hddCRP variant that was recently used for video co-segmentation (Chiu & Fritz,

[1] Chang et al. (2013) also estimate temporal correspondences between superpixels, but we do not utilize this information.

[2] See the supplement for specific hyper-parameter settings.