## Appendix: Parameter Sweeps

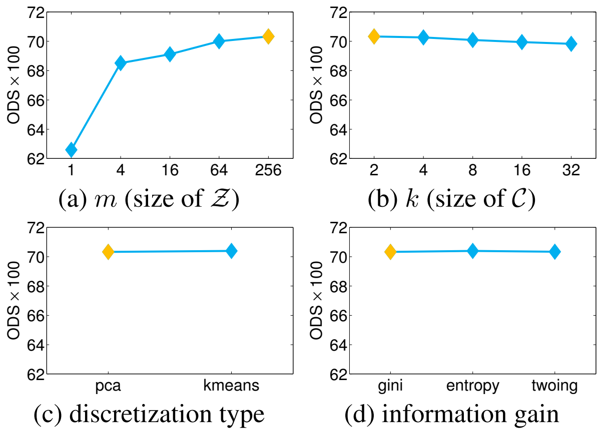


Figure 1. **Splitting Parameter Sweeps.** See text.

We set all parameters with the help of the BSDS validation set. In Figures 1-3 we explore the effect of choices of **splitting**, **model** and **feature** parameters. For each experiment we train on the 200 image BSDS training set and measure edge detection accuracy on the 100 image BSDS validation set (using the standard ODS performance metric). All results are averaged over 5 trials.

By default all parameters are set to the values described in the main text (and indicated by orange markers in the plots). Then, keeping all but one parameter fixed, we explore the effect on edge detection accuracy as a single parameter is varied. For computational reasons, however, for the validation experiments we sample fewer image patches ($10^5$ versus $10^6$) and train fewer trees (4 versus 8).

With these default setting, SE achieves an ODS of ∼.70 on the validation set. This is lower than the performance of our full model (ODS=.74) for three reasons: (1) we use fewer patches and trees, (2) the validation set is slightly more challenging than the test set, and (3) we use a faster evaluation procedure (evaluating at only 10 thresholds).

**Splitting Parameters:** In Figure 1 we explore how best to measure information gain over structured labels. Recall we utilize a two-stage approach of mapping $\mathcal{Y} \rightarrow \mathcal{Z}$ followed by $\mathcal{Z} \rightarrow \mathcal{C}$. Plots (a) and (b) demonstrate that $m = |\mathcal{Z}|$ should be large and $k = |\mathcal{C}|$ small. Results are robust to both the discretization method and the discrete measure of information gain as shown in plots (c) and (d).

**Model Parameters:** In Figure 2 we plot the influence of parameters governing the model and training data. Plots (a) and (b) show the effect of image and label patch sizes on accuracy, $32 \times 32$ image patches and $16 \times 16$ label patches perform best. Plots (c)-(e) show that increasing the number of patches, training images, and trees, respectively, leads to improved accuracy. Plot (f) shows that training each tree with a fraction of the total features has only a minor impact on accuracy (but results in proportionally lower memory usage). In (g) and (h) we see that deep trees pruned so every node has at least 8 training samples give best performance.

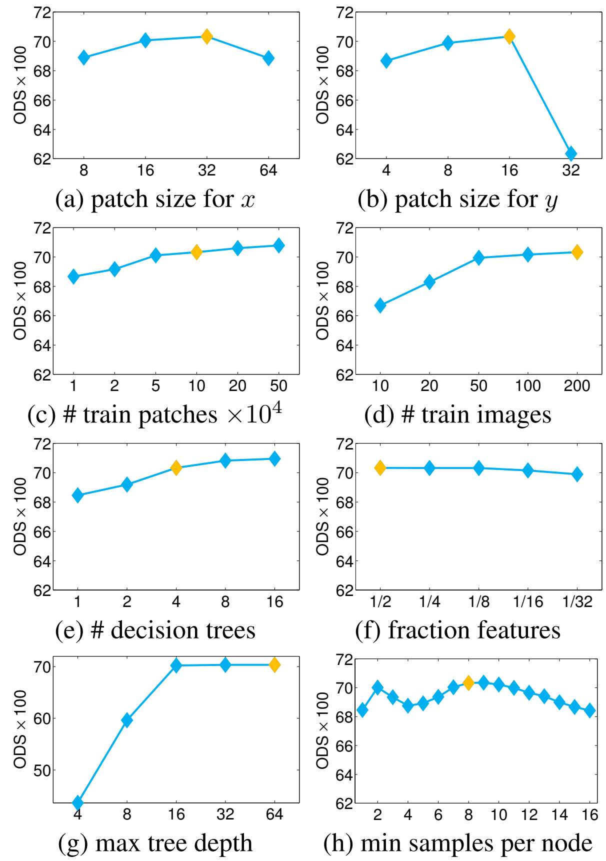


Figure 2. **Model Parameter Sweeps.** See text.

**Feature Parameters:** Figure 3 shows how varying the channel features affects accuracy. We refer readers to the main text for details, here we only note that performance is relatively insensitive to a broad range of parameter settings.

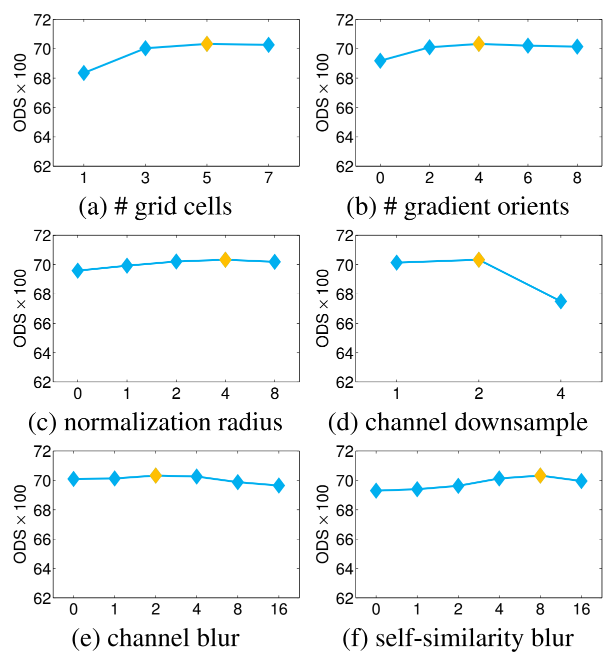


Figure 3. **Feature Parameter Sweeps.** See text.