Collect-Cut: Segmentation with Top-Down Cues Discovered in Multi-Object Images

Yong Jae Lee and Kristen Grauman

University of Texas at Austin

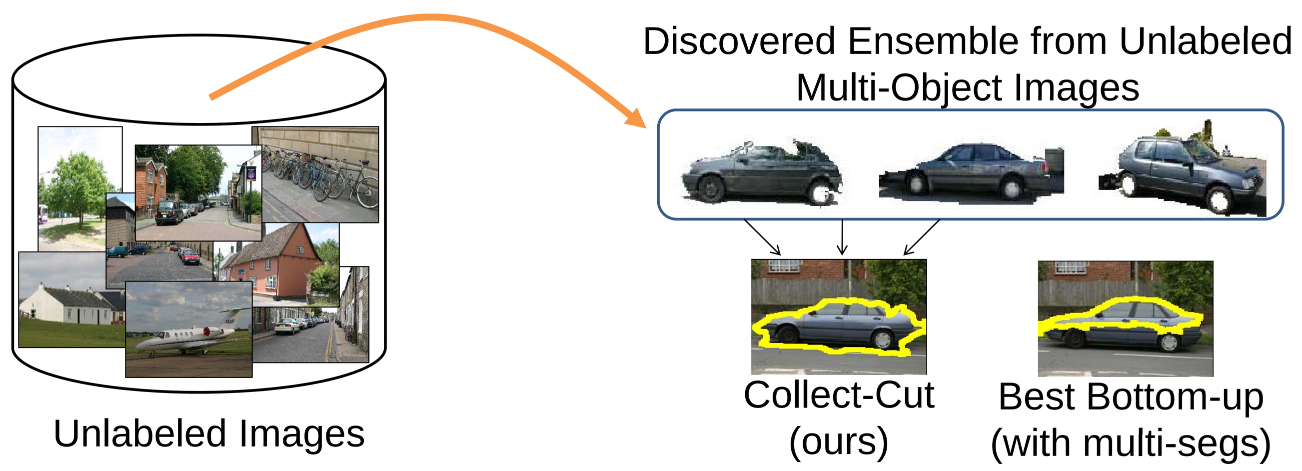
Problem

Unsupervised category discovery in multi-object images is directly influenced by segmentation quality.

Bottom-up methods cannot always produce object-like segments, even with multiple-segmentations.

A problem for discovery, since the system will never have the chance to detect recurring objects that do not have good segments.

Main idea

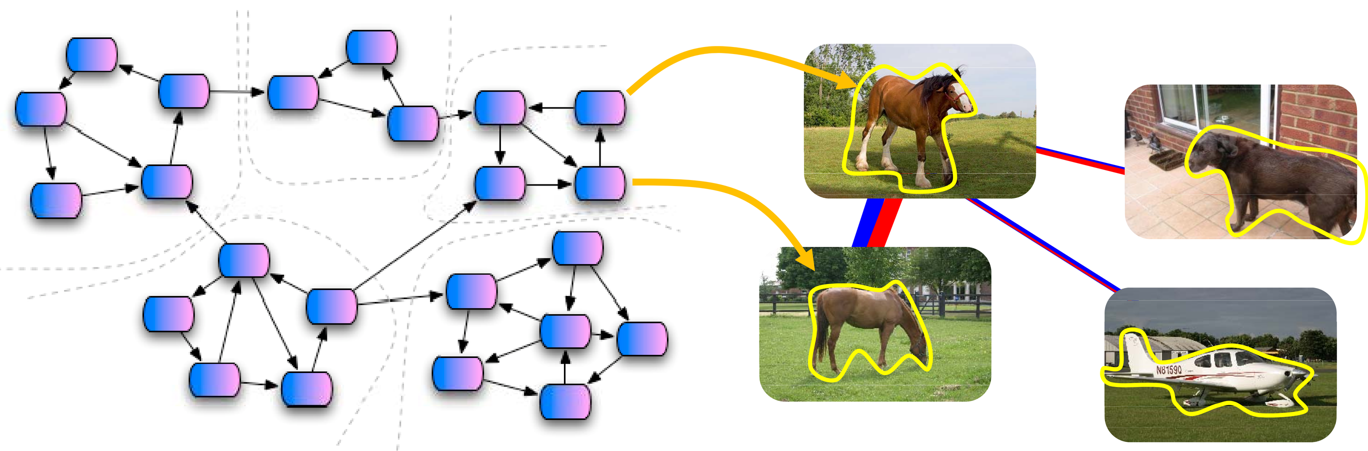


- Discover shared top-down cues from a *collection* of unlabeled multi-object images, and use them to refine both the segments and discovered objects.
- Design an energy function that can be minimized with graph cuts to revise the spatial extent of each segment.

Algorithm overview

Group regions with similar appearance and context [*Object-Graphs*, Lee & Grauman, CVPR 2010] → Form ensemble model with representative instances → Collective graph-cut segment refinement → Re-cluster with refined regions

Context-aware region grouping [Lee & Grauman, CVPR 2010]

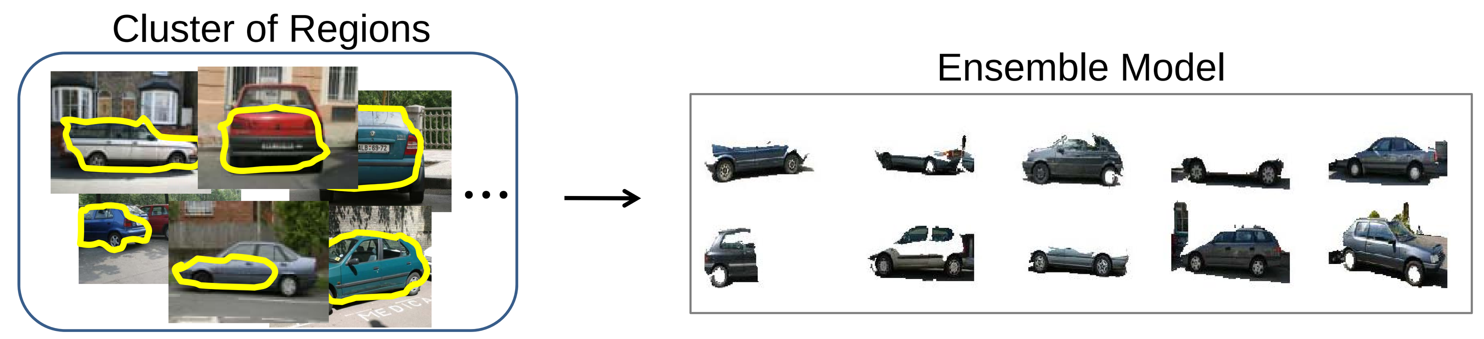


$$K(s_i, s_j) = K_{app}(s_i, s_j) + K_{obj-graph}(s_i, s_j)$$

Group regions with similar appearance *and* object-level context. Clusters will be more inclusive of intra-class appearance variations.

Assembling ensemble models

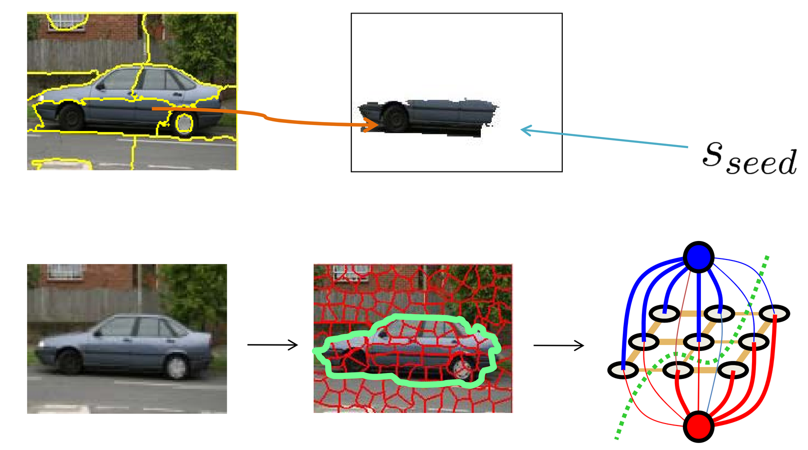
Clusters may contain partial objects, or heterogeneous instances of objects in similar contexts.



Extract *r* exemplars with highest intra-cluster affinity as top-down model of appearance.

Collective graph-cut segment refinement

Given discovered ensemble models, take each initial "seed" region and refine its segmentation.



Define a graph over image's superpixels

Energy minimization → image segmentation

***Energy function to minimize*:**

$$E(f, s_{seed}) = \sum_{i \in \mathcal{S}} D_i(f_i) + \sum_{i,j \in \mathcal{N}} V_{i,j}(f_i, f_j)$$

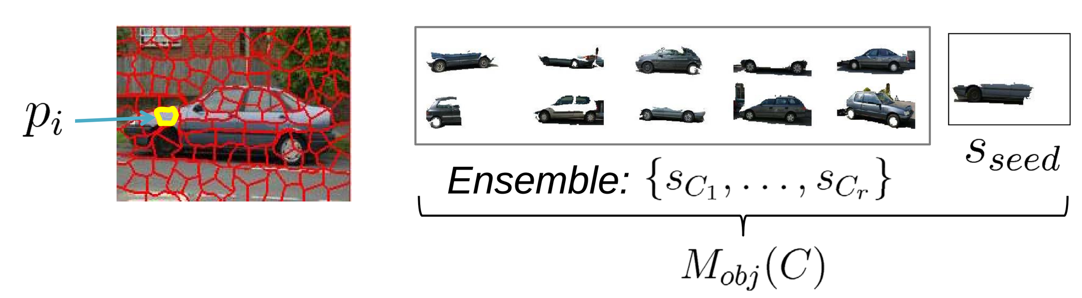
labeling of superpixels · data term · smoothness term

- **Data term:**

$$D_i(f_i) = \begin{cases} \exp\left(-d(p_i, M_{obj}(C))\right), & \text{if } f_i = 0; \\ \exp\left(-d(p_i, M_{bg}(I))\right), & \text{if } f_i = 1. \end{cases}$$

high cost when a superpixel is labeled as background but has low distance to the ensemble model

$$d(p_i, M_{obj}(C)) = \min_j \chi^2(p_i, s_{C_j}), \text{ for } s_{C_j} \in M_{obj}(C)$$



Find superpixel's best matching region, compute its distance. Exploit diversity of object parts in ensemble; each instance contributes only when needed.

high cost when a superpixel is labeled as object but has low distance to the image's background

$$d(p_i, M_{bg}(I)) = \chi^2(p_i, s_{I_{k^*}}), \text{ where } k^* = \operatorname*{argmin}_k w_k \, \chi^2(p_i, s_{I_k})$$

Modulate distances with function on entropy s.t. objects from familiar categories (i.e., with low entropy) are selected as most similar bg region.

- **Smoothness term:**

$$V_{i,j}(f_i, f_j) = |f_i - f_j| \cdot \exp(-\beta \cdot z(p_i, p_j)), \text{ where}$$

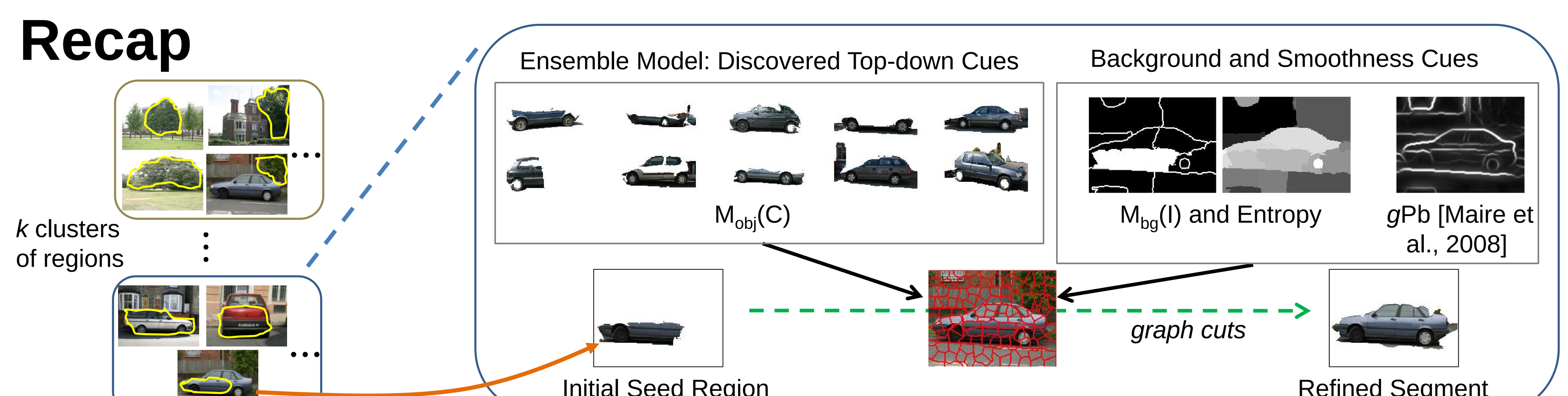
$$z(p_i, p_j) = \frac{1}{2}\left(\chi^2(p_i, p_j) + \mathrm{Pb}(p_i, p_j)\right)$$

Favor assigning same label to neighboring superpixels unless they

have very different color

have high probability of an intervening contour

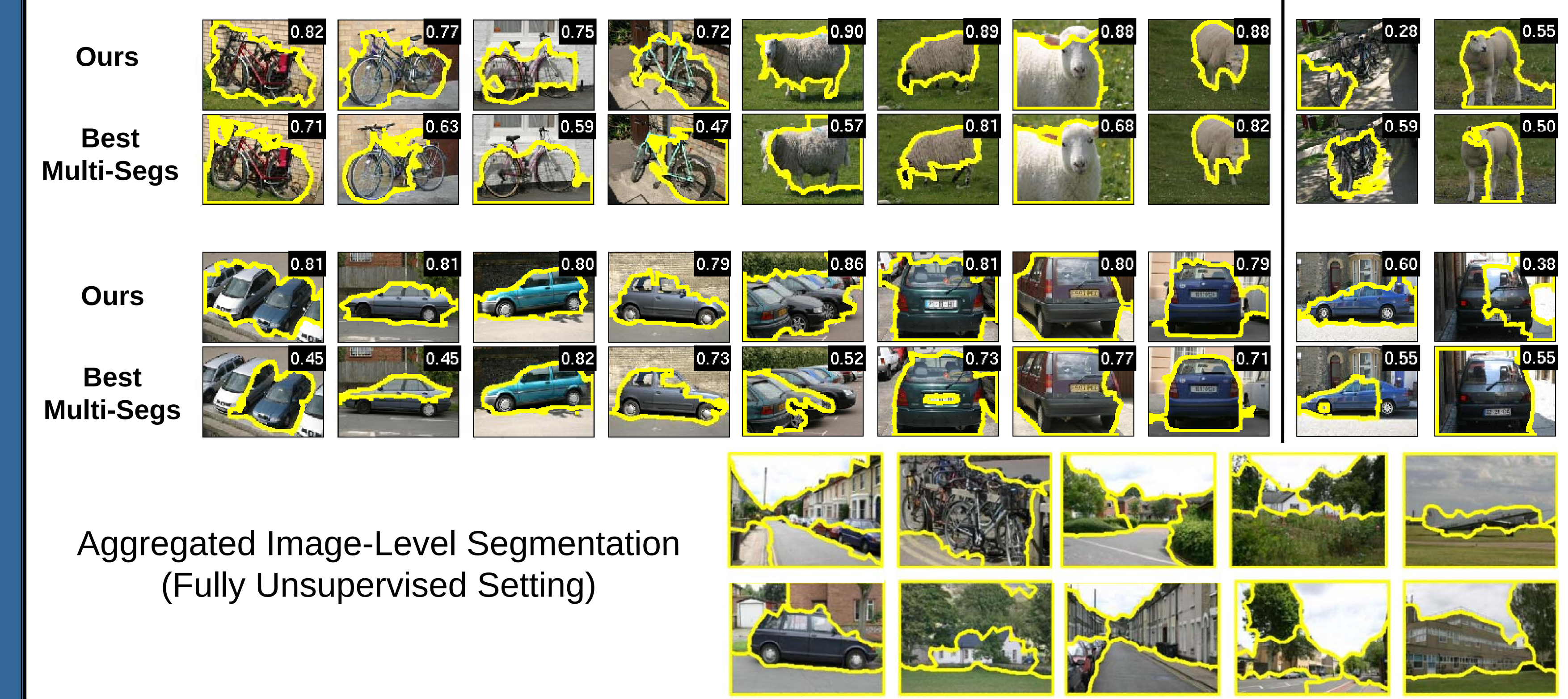
Recap



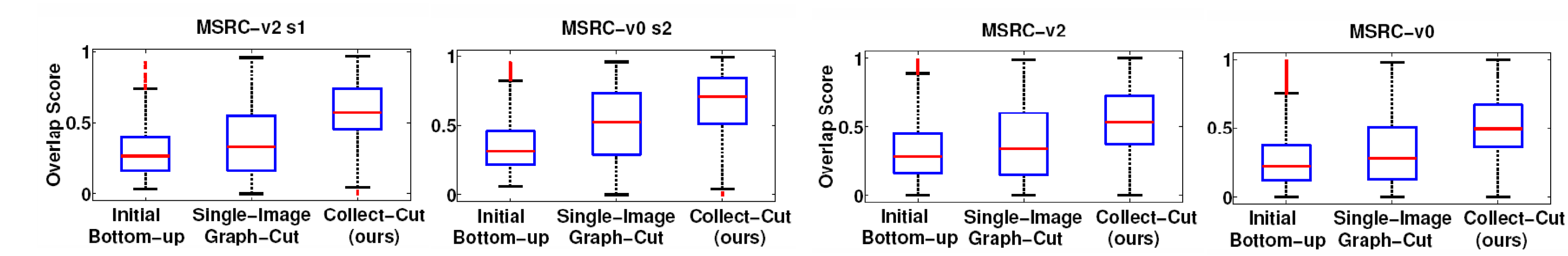
Results

Datasets: MSRC-v2 (21 classes, 591 imgs), MSRC-v0 (21 classes, 3457 imgs).

Qualitative segmentation examples



Object segmentation accuracy



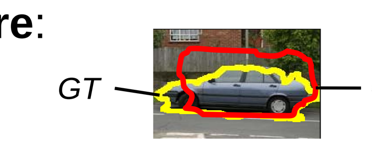
(a) Semi-supervised via familiar-object context (b) Fully Unsupervised

By leveraging the shared structure in the collection of images, our method produces significantly better segmentations than either baseline.

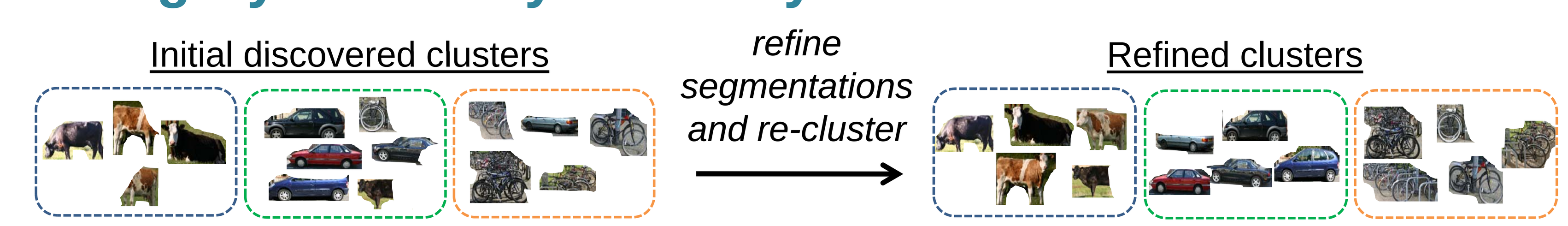
Fully Unsupervised: no previously learned category models. Use LDA [Russell et al., 2006], prefer more distant regions to be background.

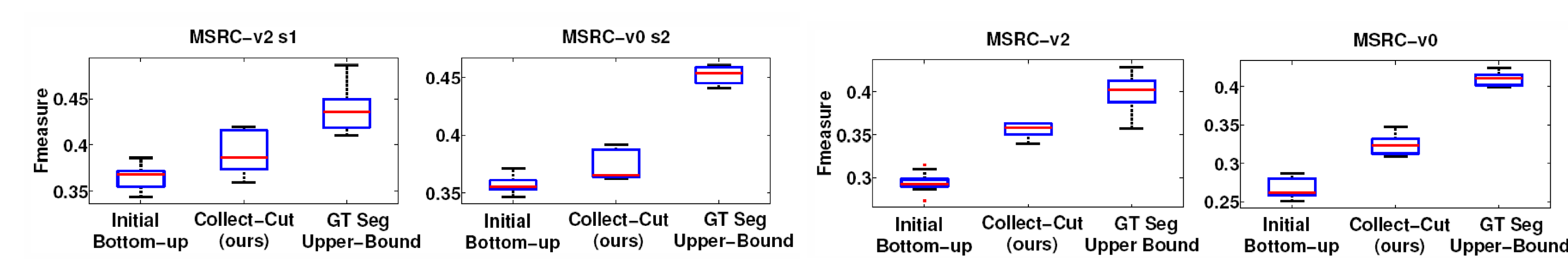
Initial Bottom-up: NCuts multiple-segs.
Single-Image Graph-Cut: object model: initial seed region, bg model: outermost regions.

Overlap Score: $OS = \frac{|GT \cap R|}{|GT \cup R|}$



Category discovery accuracy





(a) Semi-supervised via familiar-object context (b) Fully Unsupervised

Spatial extent of refined regions more closely matches true objects, allowing more complete app. features to be extracted per region, leading to better groupings.

Conclusion

- Discover shared structure in unlabeled set of images to refine the object-level segmentations and category-level groupings.
- Optionally introduce knowledge about previously learned categories.