Predicting Sufficient Annotation Strength for Interactive Foreground Segmentation

THE UNIVERSITY OF TEXAS AT AUSTIN

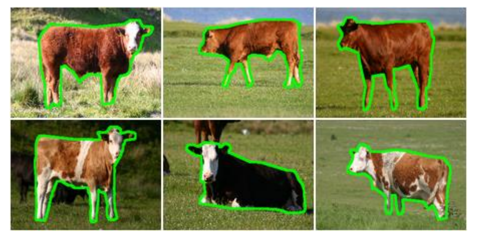
Suyog Jain and Kristen Grauman

University of Texas at Austin

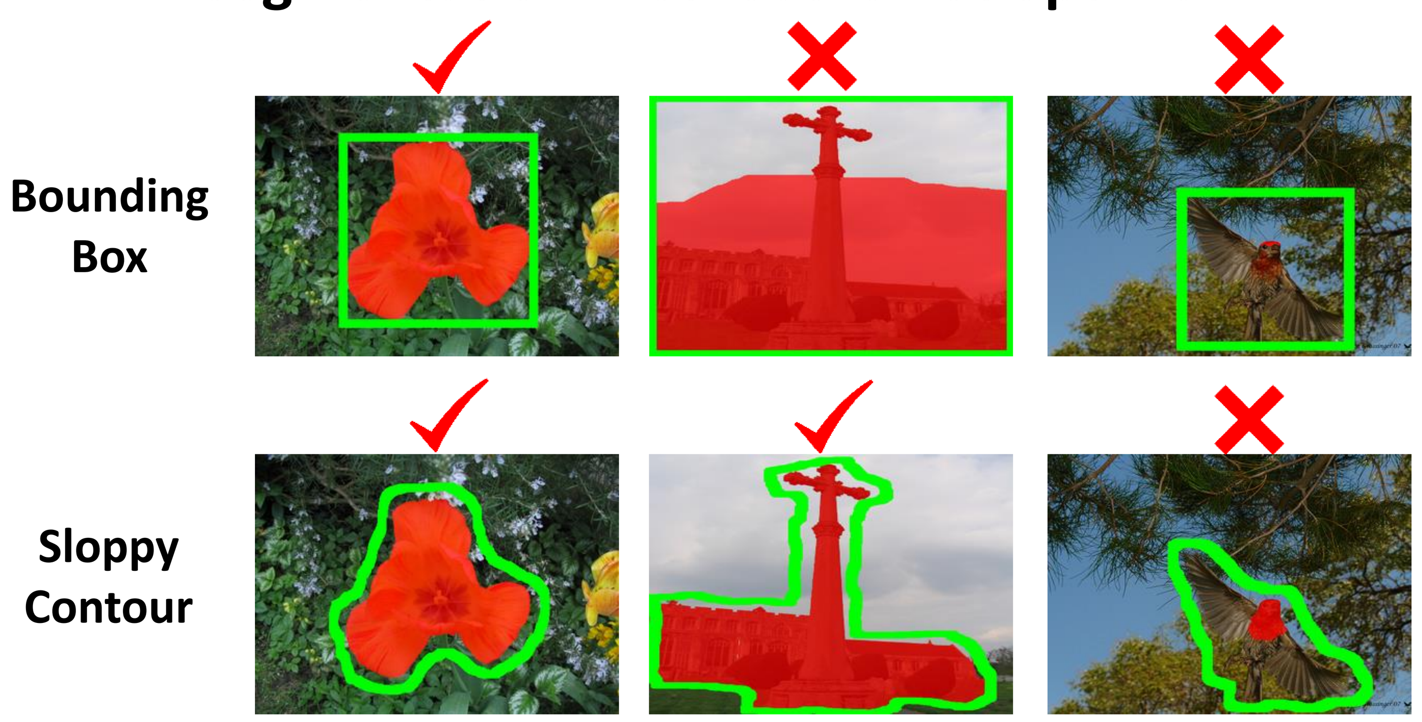
Interactive Image Segmentation

Human provides high level guidance to the segmentation algorithm.

Mobile Search **Data Collection** **Graphics**

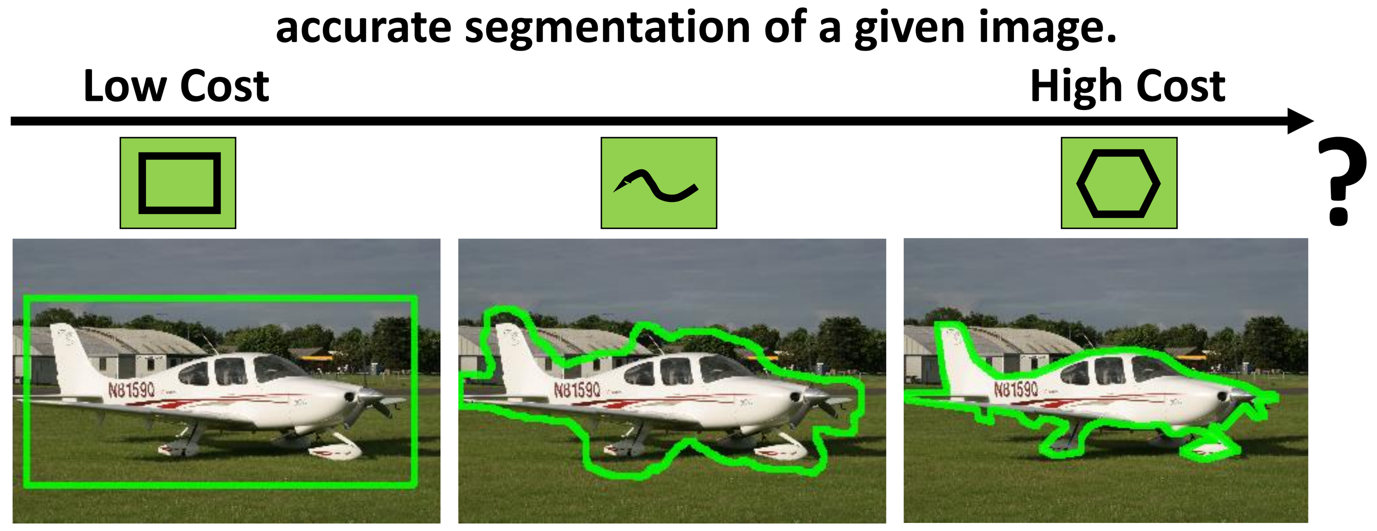


Problem: Fixing the input modality for interactive segmentation methods is not optimal



Our Goal

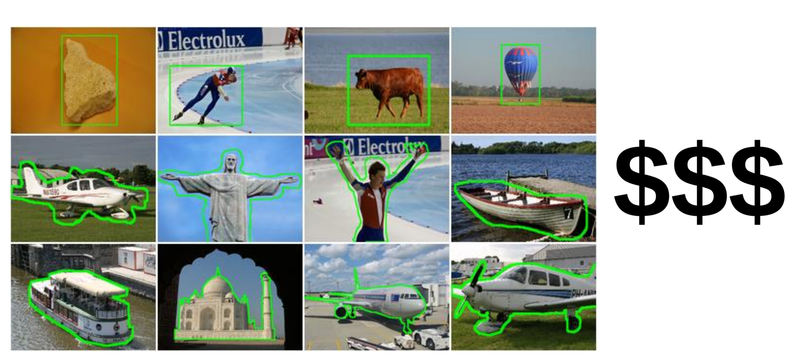
Predict the annotation modality that is sufficiently strong for accurate segmentation of a given image.



Bounding Box (7 sec) **Sloppy Contour (20 sec)** **Tight Polygon (54 sec)**

Applications:

Quick selection for a single image



Group selection with fixed budget

Segmentation model (Markov Random Field)

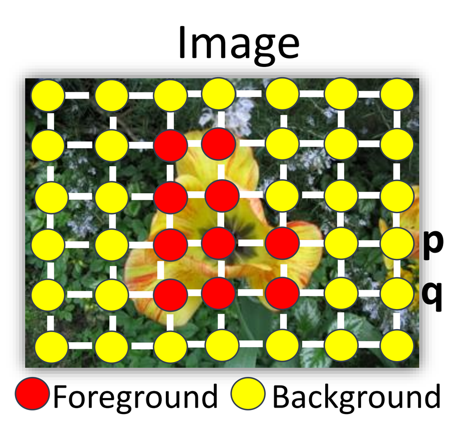
$$E(L) = \sum_{p} A_p(y_p) + \sum_{p,q \in \mathcal{N}} S_{p,q}(y_p, y_q)$$

(Data term) (Smoothness term)

$y_p \in \{1, 0\}$ is the label of pixel p

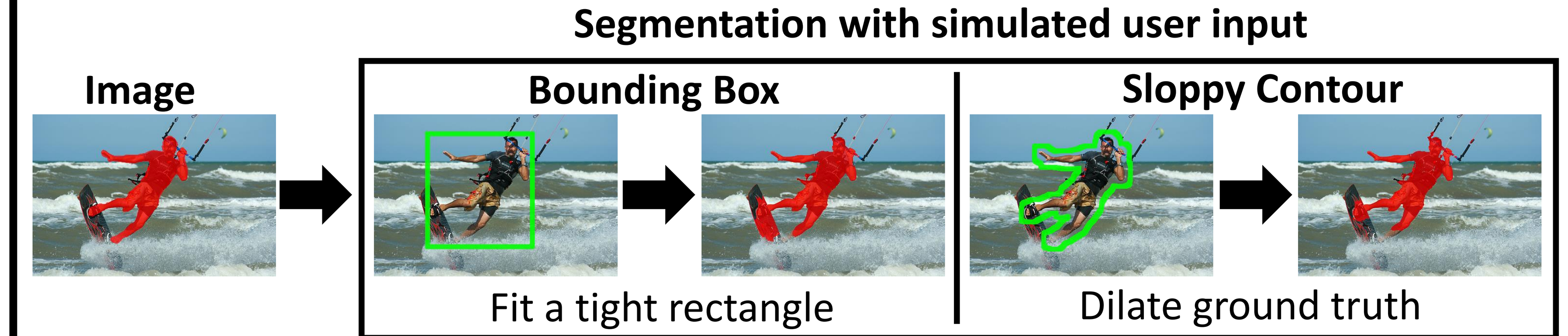
L : Labeling over entire image

[Boykov 2001, Rother 2004]

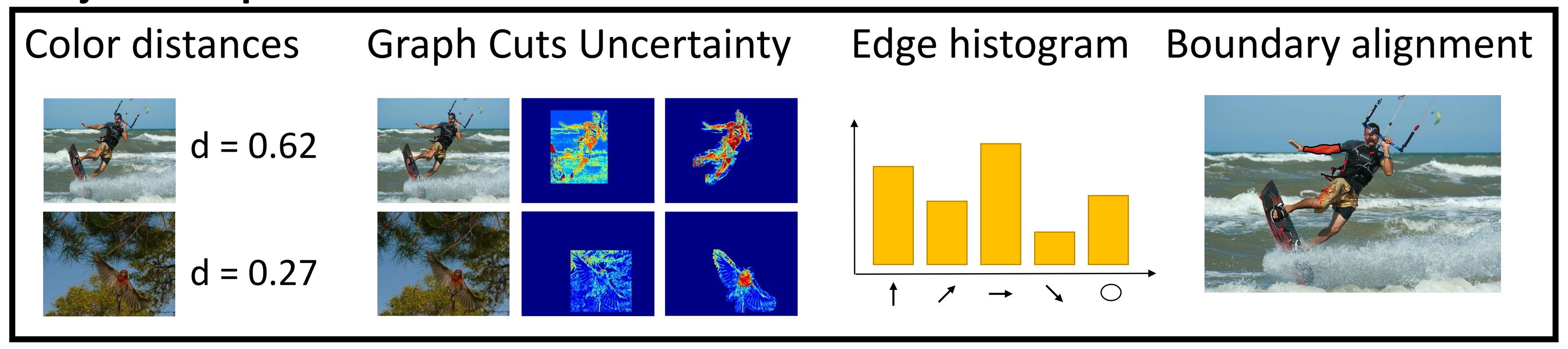


Learning to predict segmentation difficulty per modality

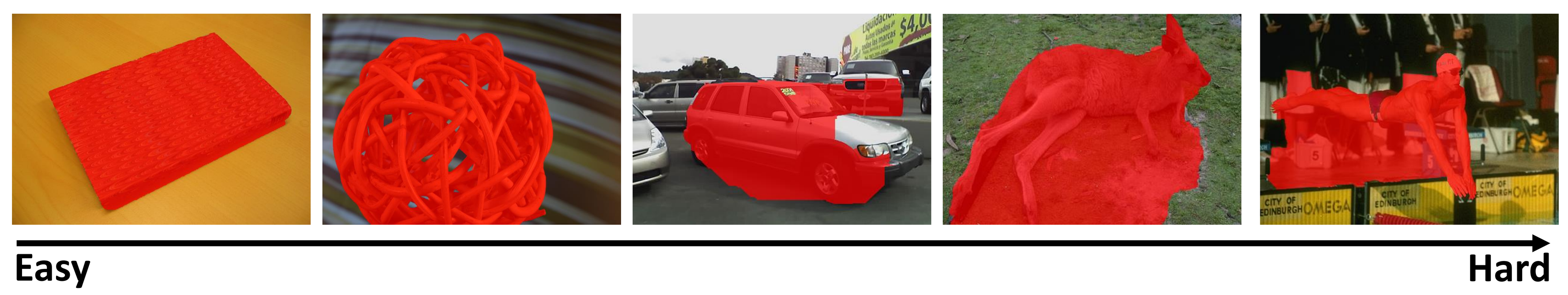
Training: Given a set of images with the foreground masks, we simulate the user input



Object independent features



Use the overlap score between the resulting segmentation and ground truth to mark an image as "easy" or "hard" and train a linear SVM classifier (for each modality).

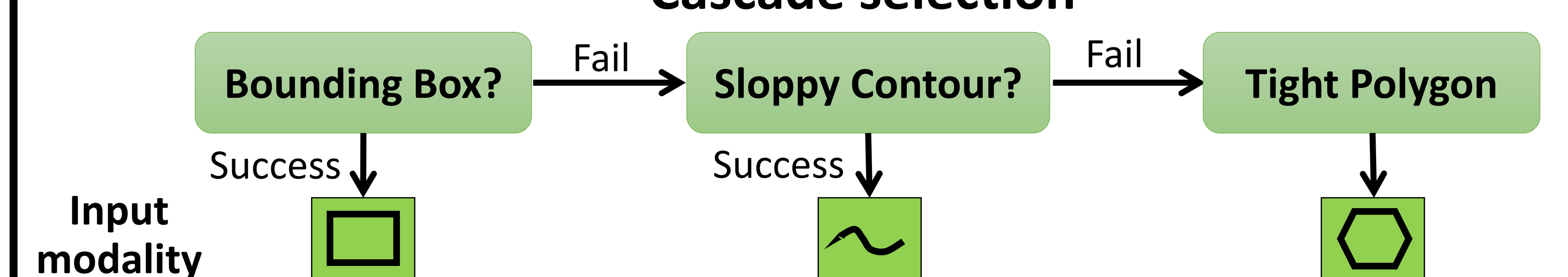


Testing: Use saliency detector to get a coarse estimate of foreground at test time.

[Liu et al. 2009]

Compute the proposed features and use trained classifiers to predict difficulty.

Cascade selection



Annotation choices with budget constraints

Goal: Given a batch of "n" images with a fixed time budget "B", we find the optimal annotation tool for each image.

Objective:
$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \sum_{k=1}^{n} p_k^b x_k^b + p_k^c x_k^c + p_k^p x_k^p,$$

$$\text{s.t.} \quad \mathbf{c}^T \mathbf{x} \leq B,$$ (Selection should not exceed budget)

Constraints
$$x_k^b + x_k^c + x_k^p = 1, \ \forall k = 1, \ldots, n,$$ (Uniqueness Constraint)

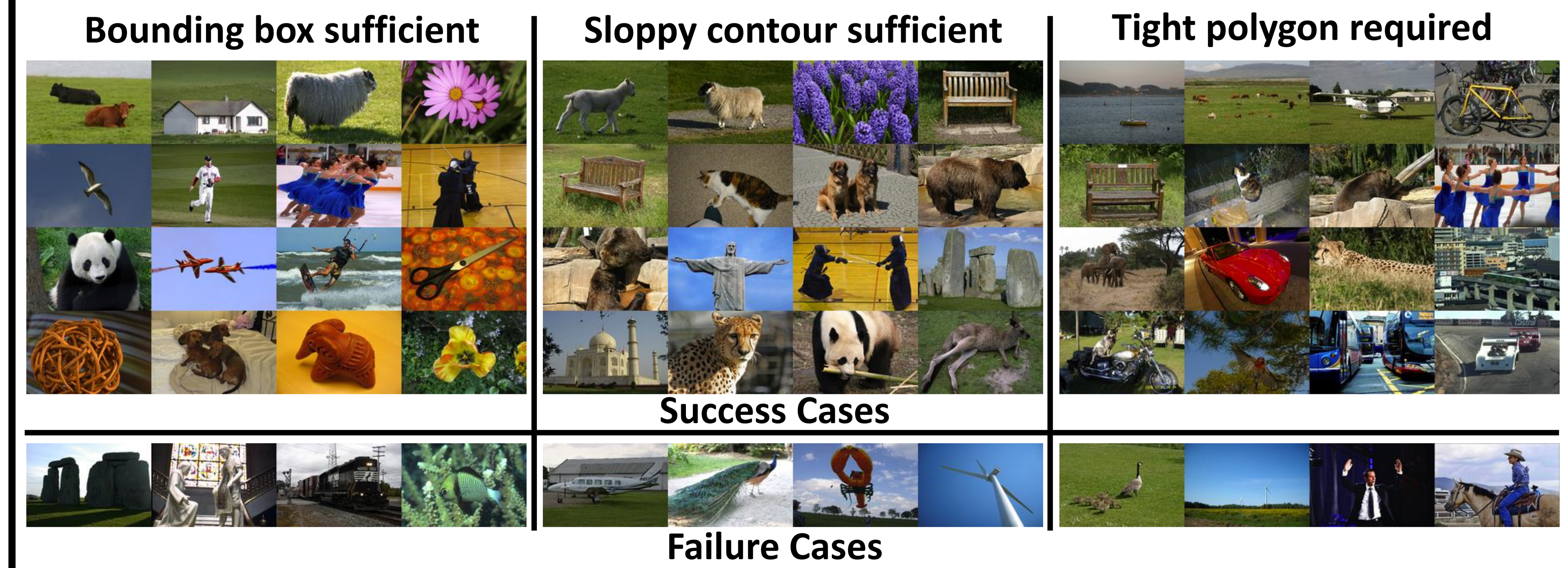
$$x_k^b, x_k^c, x_k^p \in \{0, 1\}, \ \forall k = 1, \ldots, n.$$

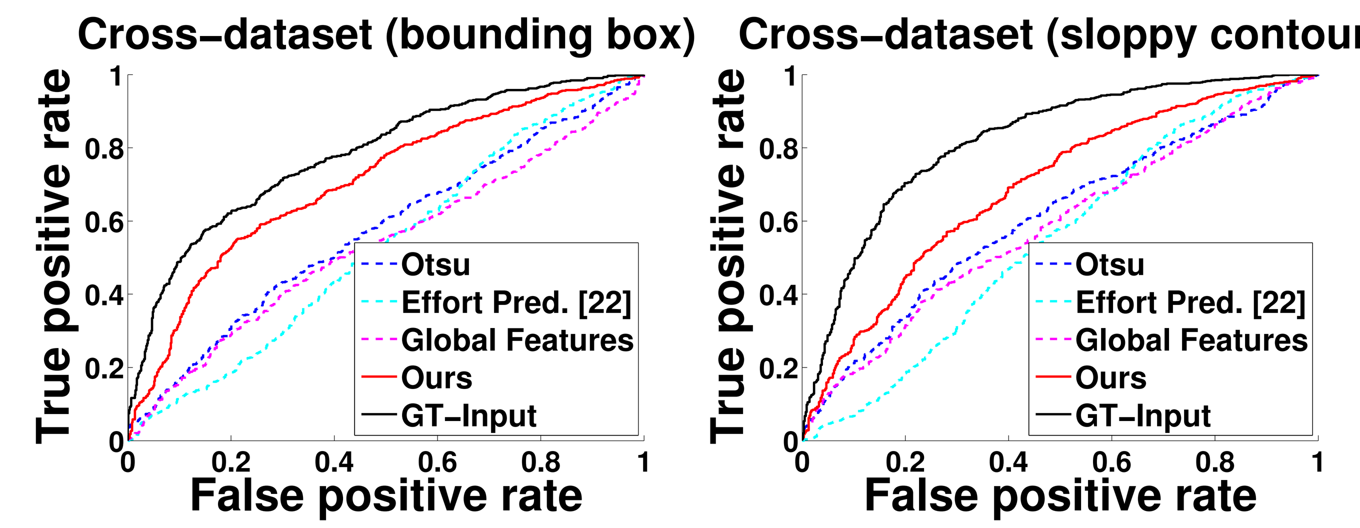
where, $\mathbf{x}$: Modality indicator vector $\mathbf{p}$: Success probability $\mathbf{c}$: Cost vector

Efficiently solved using branch and bound method.

Results

Predicting segmentation difficulty per modality





Datasets:

- MSRC (591 images)
- iCoseg (643 images)
- IIS (151 images)

Leave one dataset out cross-validation.

Our method learns generic cues to predict difficulty, not some dataset specific properties.

Baselines:

- Otsu adaptive thresholding
- SVM with global image features
- Effort Prediction (Vijayanarasimhan et al. 2009)
- Our method with Ground Truth input (upper bound)

Cascade selection – application to object recognition

Task: Given a set of images with a common object, train a classifier to separate object vs. non object regions.

How to get data labeled?

All tight: Ask the human annotator to provide pixel level masks **(status quo)**.

Ours: Use our cascade selection method to decide the best annotation for each image.

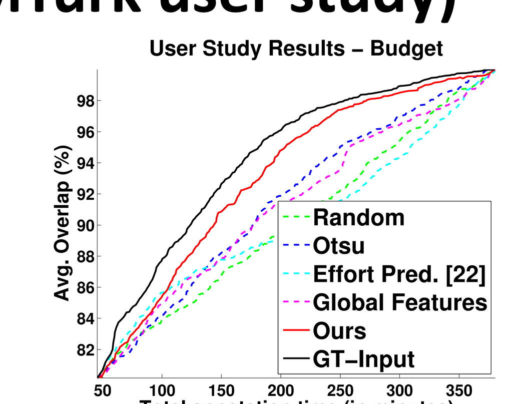
Object	Overlap Score (%)		Time Saved
	All tight	Ours	
Flower	65.09	65.6	21.2 min (73%)
Car	60.34	60.29	3.9 min (15%)
Cow	72.9	66.53	9.2 min (68%)
Cat	51.79	46.56	13.7 min (23%)
Boat	51.08	50.77	1.4 min (10%)
Sheep	75.9	75.59	17.2 min (64%)

Our method leads to substantial savings in annotation effort with minimal loss in accuracy

Annotation choices with budget constraints (MTurk user study)

- 101 MTurkers (5 per image).
- Use the median time for each image for experiments.
- Budget ranges from "all bounding boxes" to "all tight polygons"

For the same amount of annotation time, our method leads to much higher average overlap scores.



Conclusion

- A method to predict the kind of human annotation required to segment a given image.
- User study shows that explicit reasoning about segmentation difficulty is useful.