Active and Interactive Image and Video Segmentation

Kristen Grauman

University of Texas at Austin

Work with Suyog Jain and Danna Gurari

THE UNIVERSITY OF TEXAS AT AUSTIN

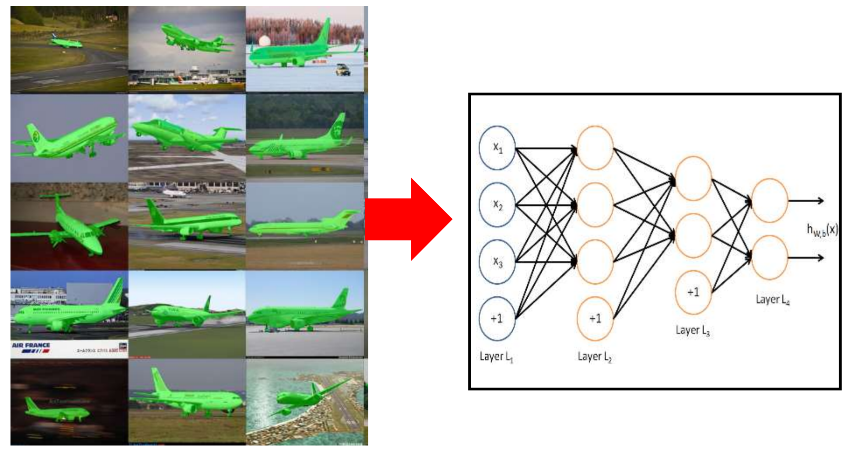
Foreground Object Segmentation

Task: Generate pixel level masks for the foreground objects in an image or video

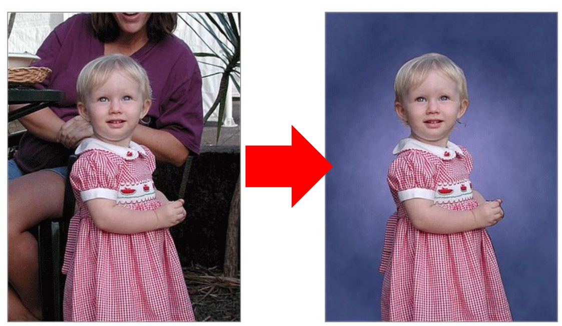
Kristen Grauman, UT Austin

Why Foreground Object Segmentation?

Visual search



Recognition



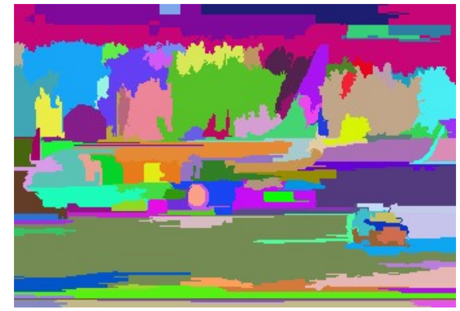
Graphics

Spectrum of automatic segmentation methods

Unsupervised methods

[Shi & Malik 2000, Felzenszwalb 2004, Martin 2004, Wang 2005, Arbeláez 2011, …]

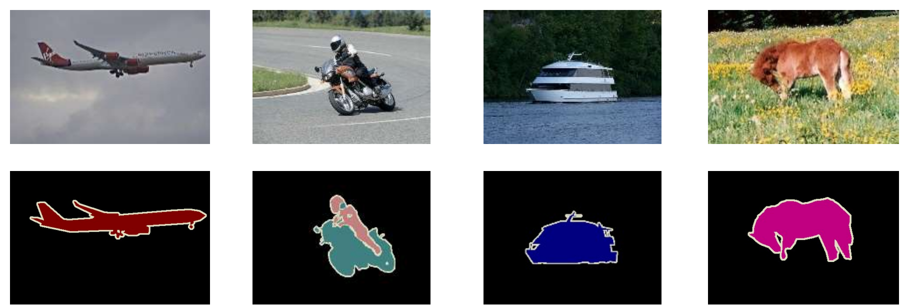
Image



Bottom-up Segmentation

Fully supervised methods

[Borenstein 2002, Kumar 2005, Shotton 2006, Pantofaru 2008, Ladicky 2009, Fulkerson 2009, …]

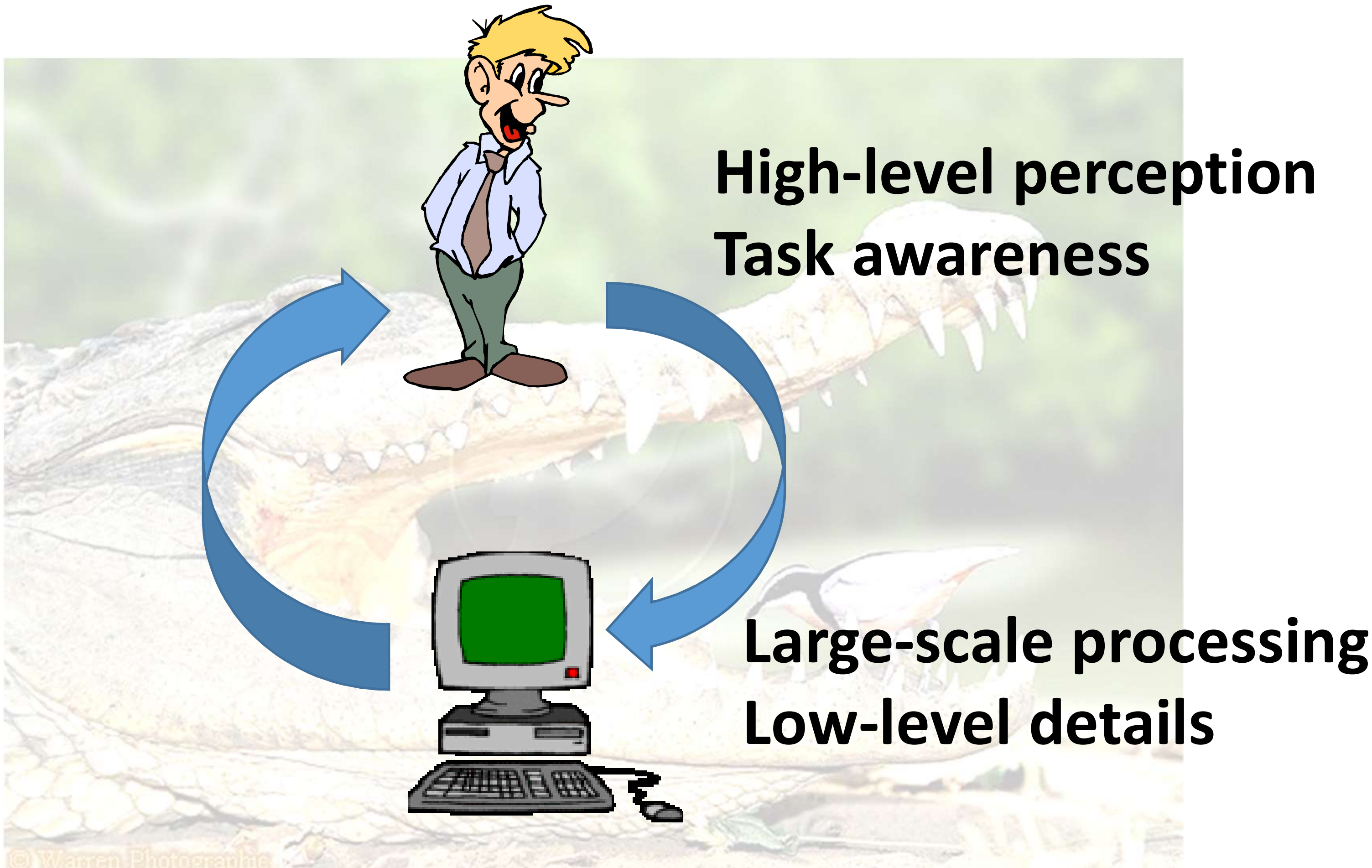


Airplane | Motorbike & Person | Ship | Cow

Symbiosis in Segmentation

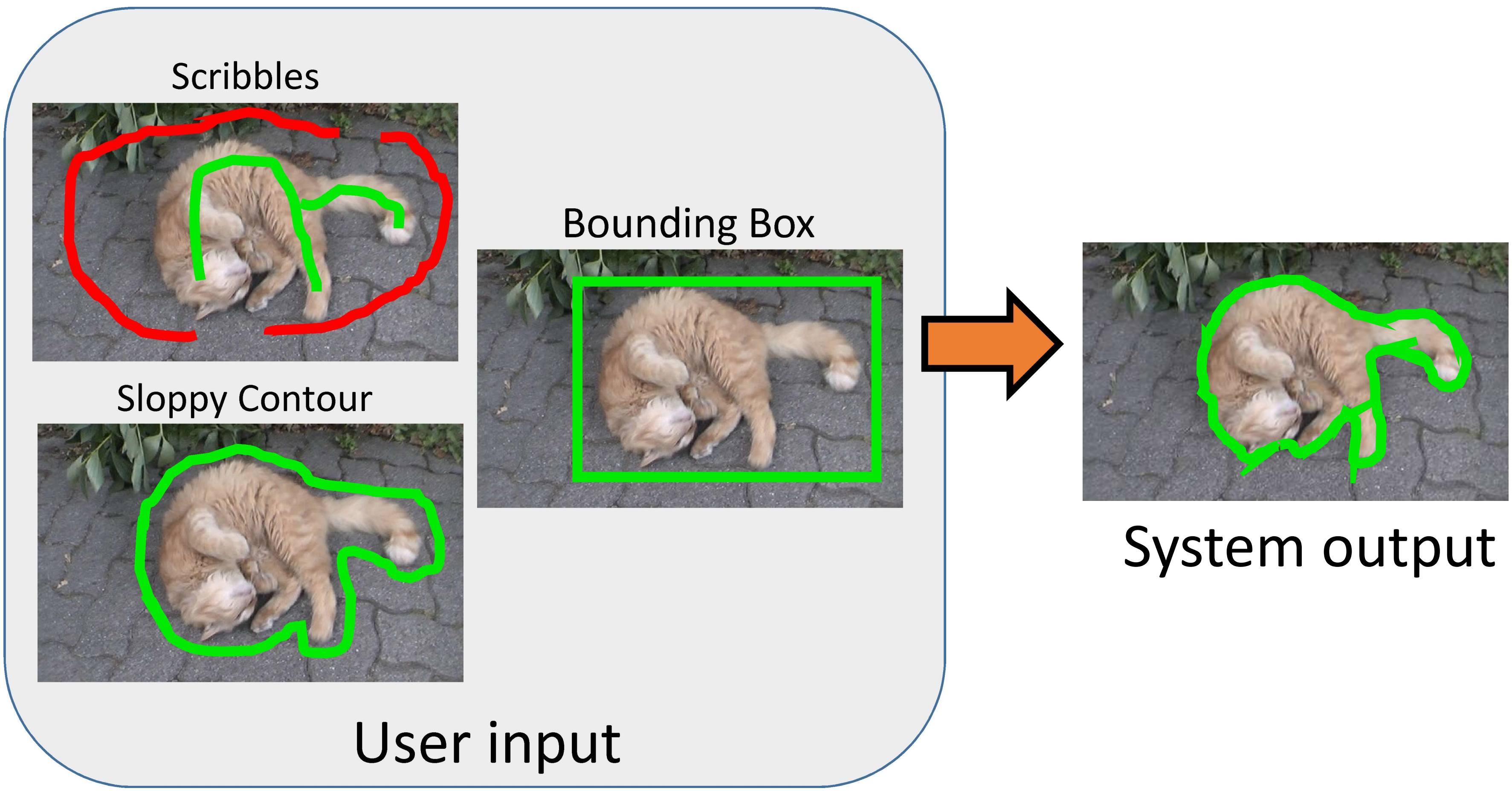
Kristen Grauman, UT Austin

Symbiosis in Segmentation



Interactive Segmentation

Main idea in existing methods: Use "light" annotations to infer more precise boundaries



[Boykov 2001, Zabih 2001, Rother 2004, Kohli 2008,…]

Kristen Grauman, UT Austin

Limiting assumptions in existing work

- One-size-fits-all annotation modalities
- Human always knows best
- Constant human in the loop to monitor video segmentation

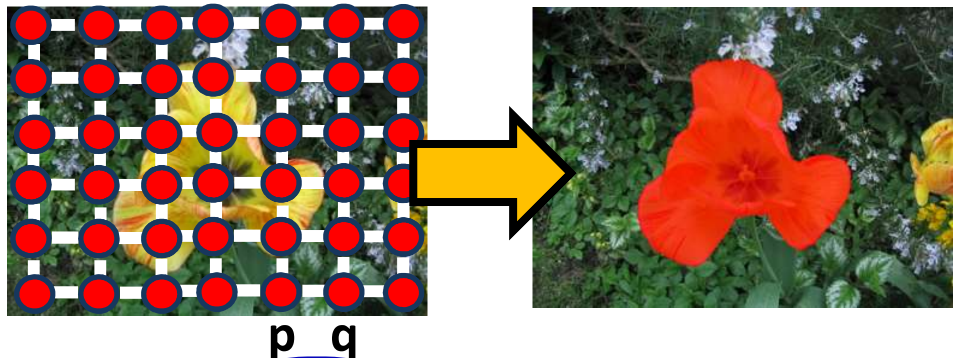
Our goal

Active and interactive segmentation methods to predict exactly where and how human intervention is needed

This talk:

1. Given an image, what strength of annotation is needed?
2. Given a collection of images, which ones need human input?
3. Given a video, how to propagate minimal human input?

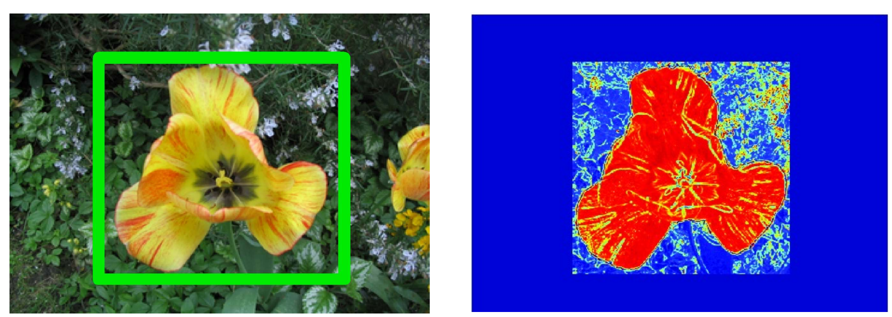
Background: a typical MRF segmentation model



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

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

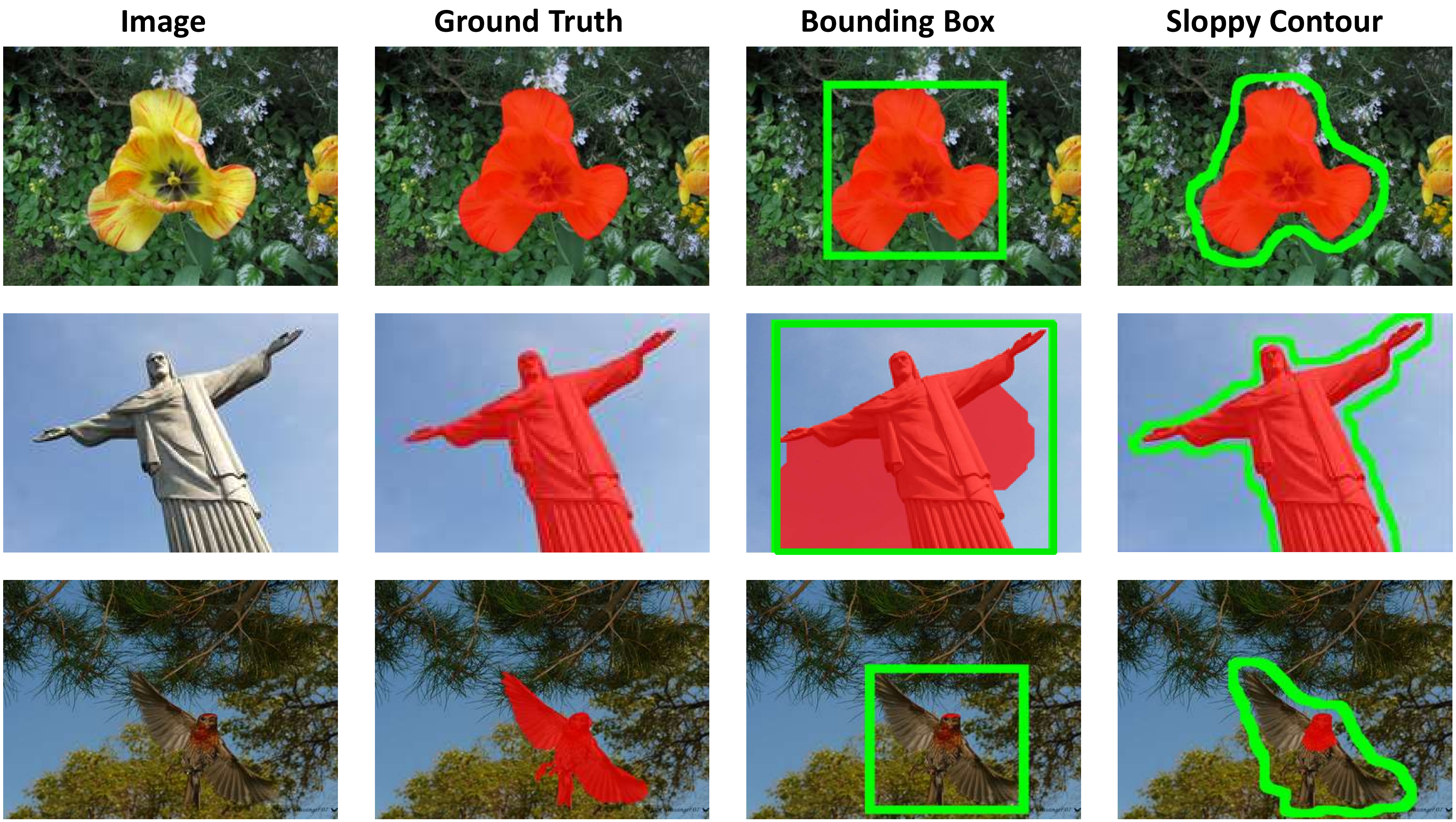
User input leads to foreground likelihoods



[Boykov 2001, Zabih 2001, Gulshan 2010, Kohli 2008, …]

Kristen Grauman, UT Austin

Problem



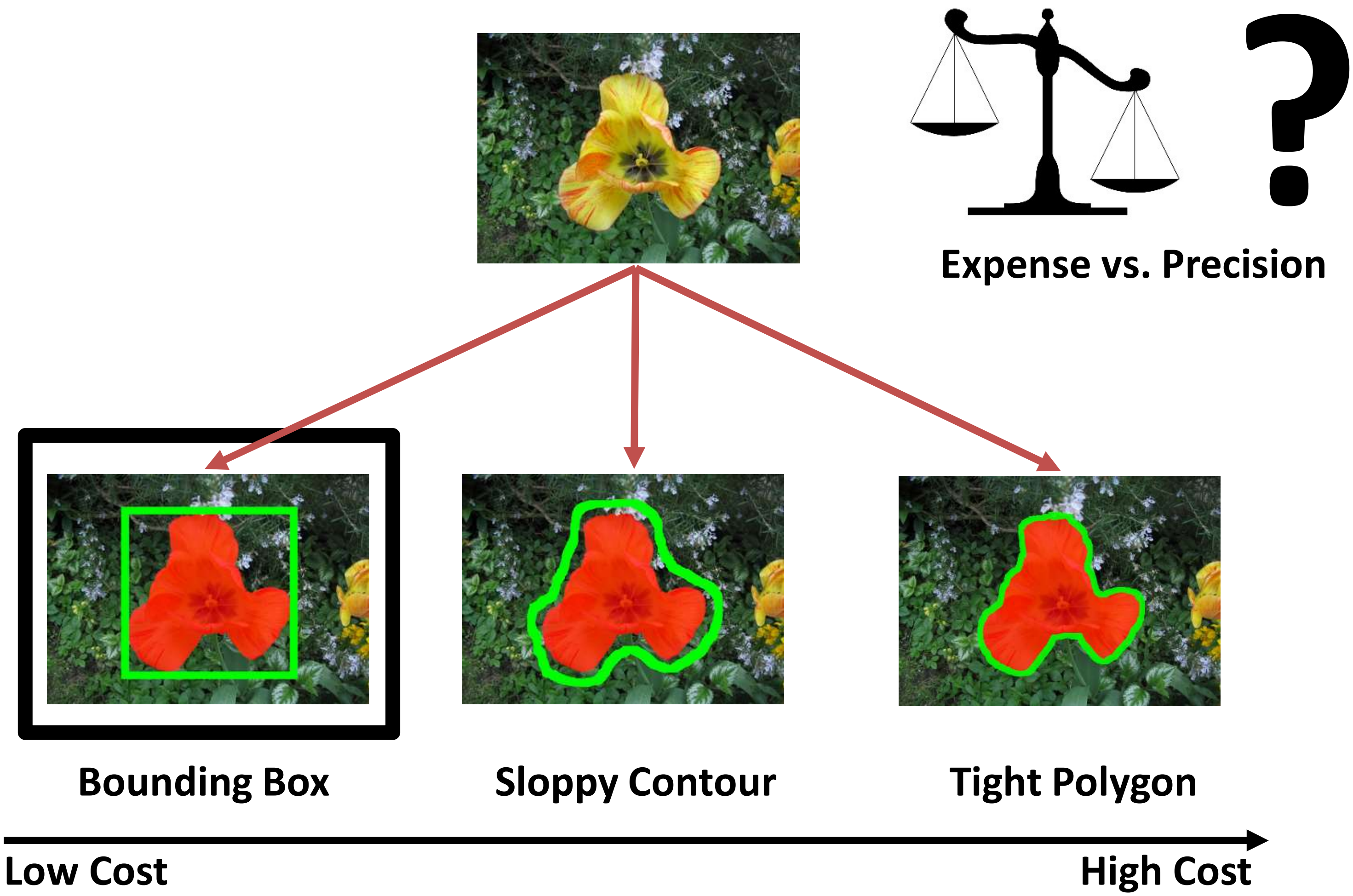
Fixing the input modality leads to a suboptimal trade-off between human and machine effort!

Kristen Grauman, UT Austin


I MUST BE WILLING TO TAKE ON EXTRA RESPONSIBILITIES?
FALSE. THAT'S NOT IN THE JOB DESCRIPTION.
quickmeme.com

Our Idea

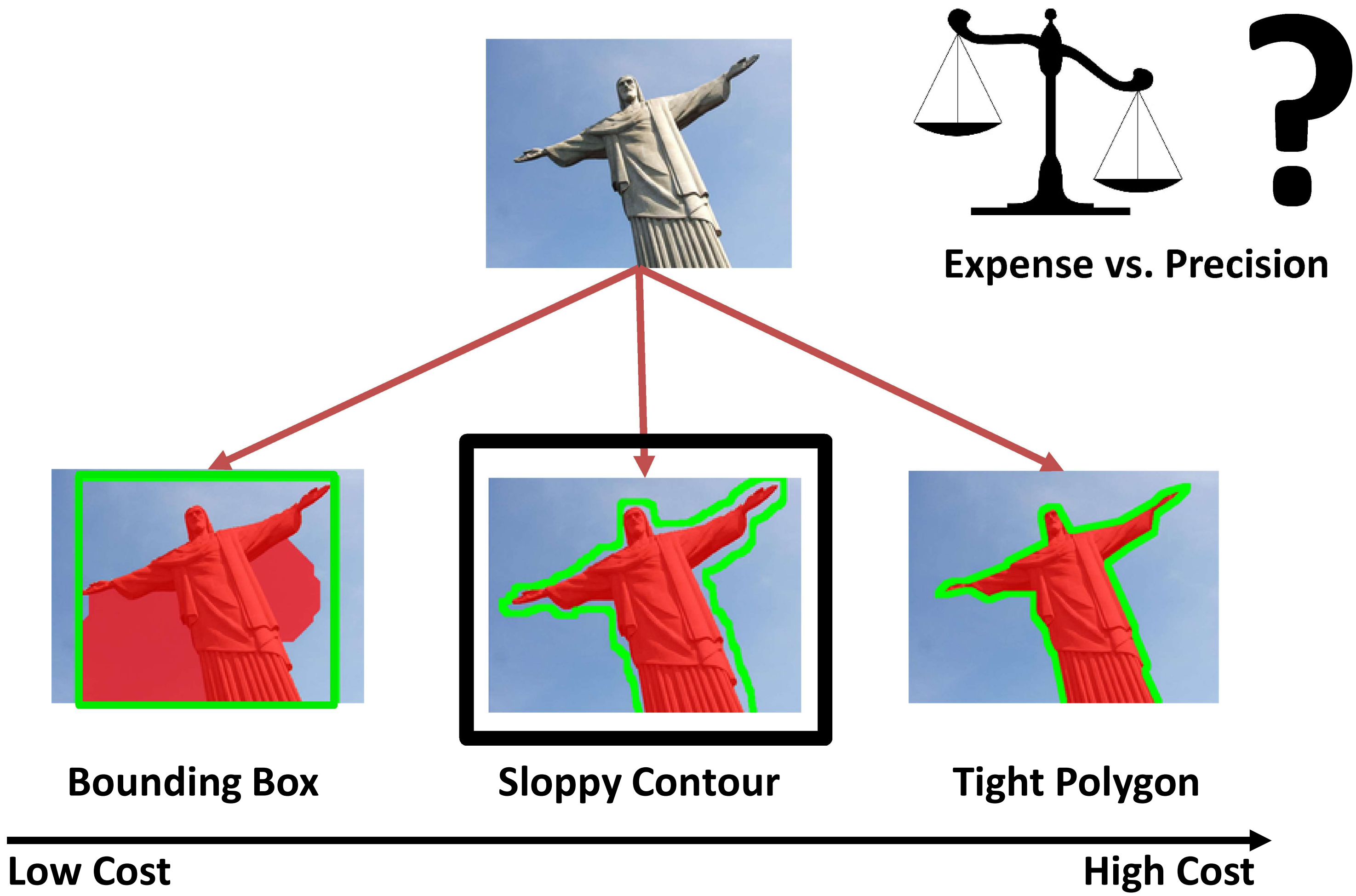
Predict the annotation modality that is sufficiently strong for accurate segmentation



Jain & Grauman, ICCV 2013

Our Idea

Predict the annotation modality that is sufficiently strong for accurate segmentation



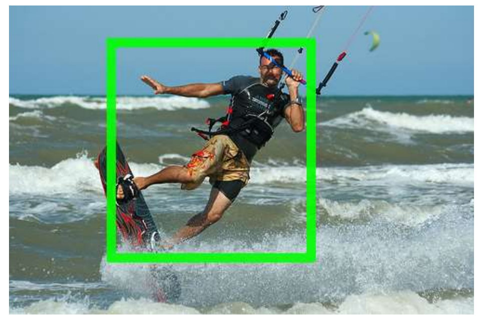
Jain & Grauman, ICCV 2013

Training Phase

- Given ground truth foreground, simulate the user input.

Ground Truth

Bounding Box



Fit a tight rectangle

Sloppy Contour

Dilate the ground truth mask

Graph Cuts Segmentation

Compute object independent features which capture the foreground separability.

Jain & Grauman, ICCV 2013

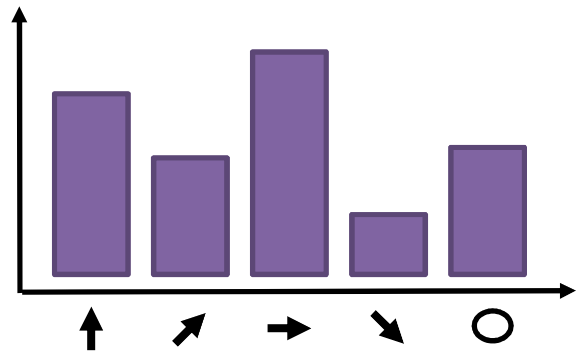
Training Phase:
Learn Image Cues Indicative of Difficulty

- **Color Separability**

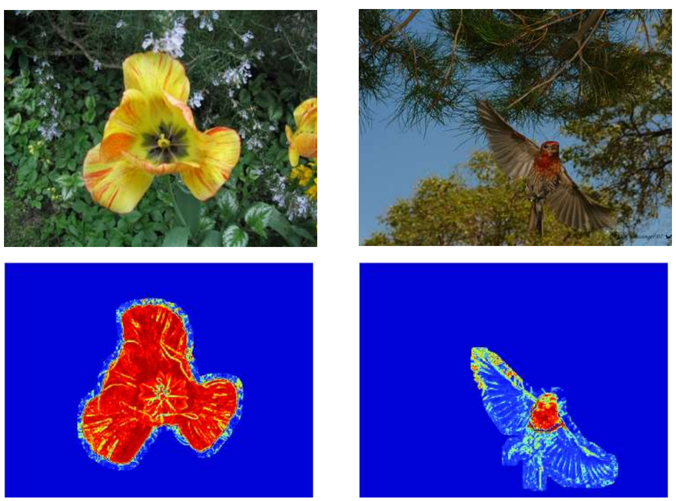
d = 0.6269

d = 0.2764

- **Edge Complexity**



- **Label Uncertainty**



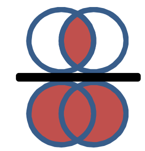
- **Boundary alignment**

Kristen Grauman, UT Austin

Jain & Grauman, ICCV 2013

Training Phase:
Learn Image Cues Indicative of Difficulty

- Train model to predict difficulty for each input modality.

- Easiness = segmentation overlap score $\left(\frac{Pred. \cap GT}{Pred. \cup GT}\right)$ 

E.g. for bounding box:

Easy → **Hard**

Jain & Grauman, ICCV 2013

Testing Phase: Will a Given Modality Succeed?

- Given novel image, salient object detector (Liu et al. 2009) to roughly localize probably foreground

Predict whether each modality would succeed:

1. Compute bounding boxes/sloppy contours from mask
2. Apply graph cut segmentation.
3. Extract features and predict the difficulty.

Jain & Grauman, ICCV 2013

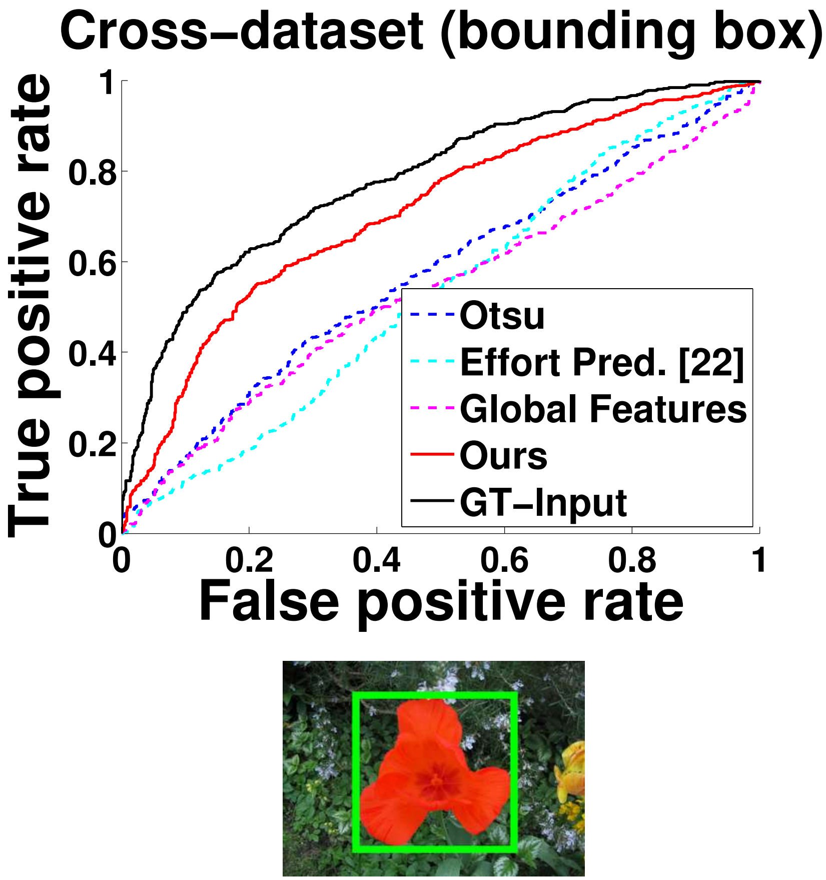
Datasets

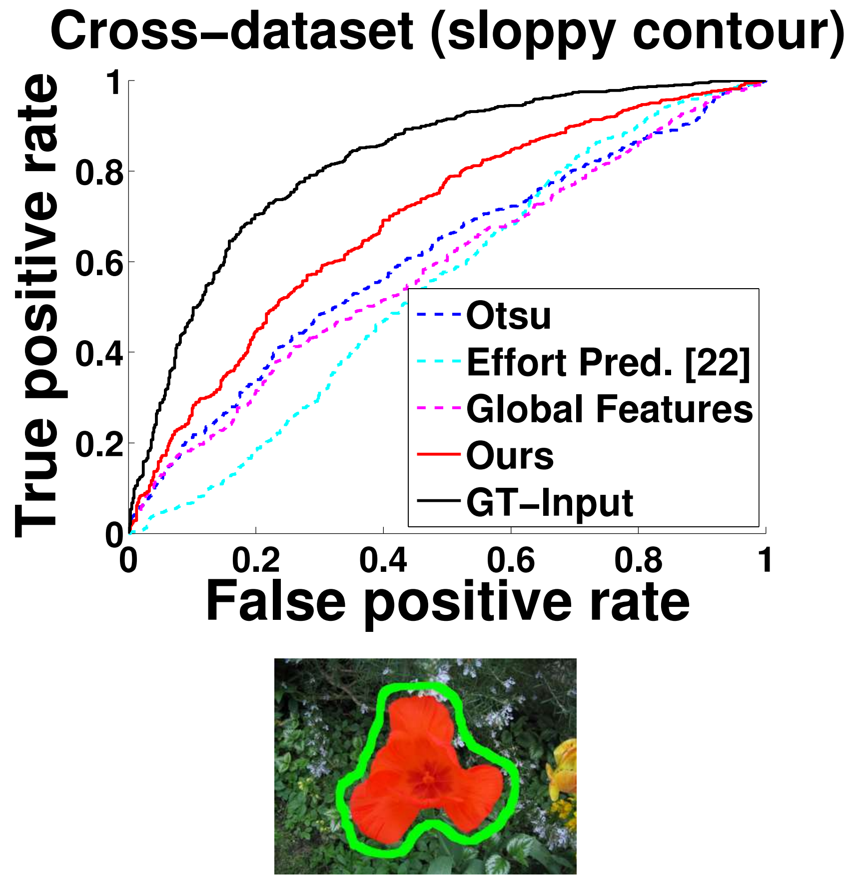
- **MSRC (591 images, 20 classes)**

- **CMU – Cornell iCoseg (643 images, 38 groups)**

- **Interactive Image Segmentation (151 unrelated images)**

How well can we detect difficult images?





Our method learns generic cues, not dataset-specific features.

Jain & Grauman, ICCV 2013

Qualitative Results – Success Cases

Bounding Box sufficient

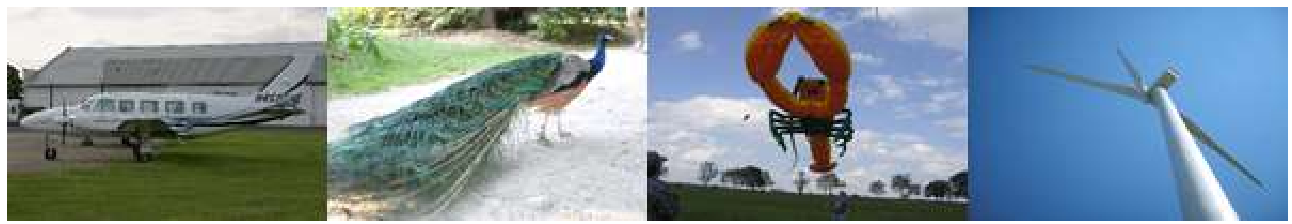
Sloppy contour sufficient

Tight Polygon required

Qualitative Results – Failure Cases

Bounding Box sufficient

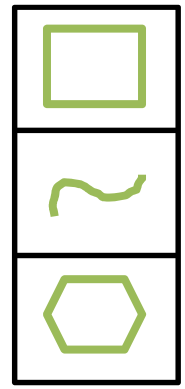
Sloppy contour sufficient



Tight Polygon required

Using difficulty predictions to intelligently gather annotations

1 **Cascaded Selection**



Bounding box

Sloppy Contour

Tight Polygon

Using difficulty predictions to intelligently gather annotations

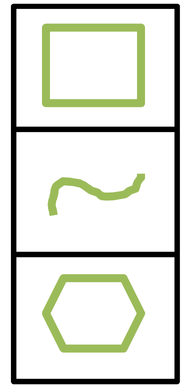
How accurate is the resulting recognition system?

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%)

For almost no loss in accuracy, our method leads to substantial savings in annotation effort.

Using difficulty predictions to intelligently gather annotations

1 Cascaded Selection



Bounding box

Sloppy Contour

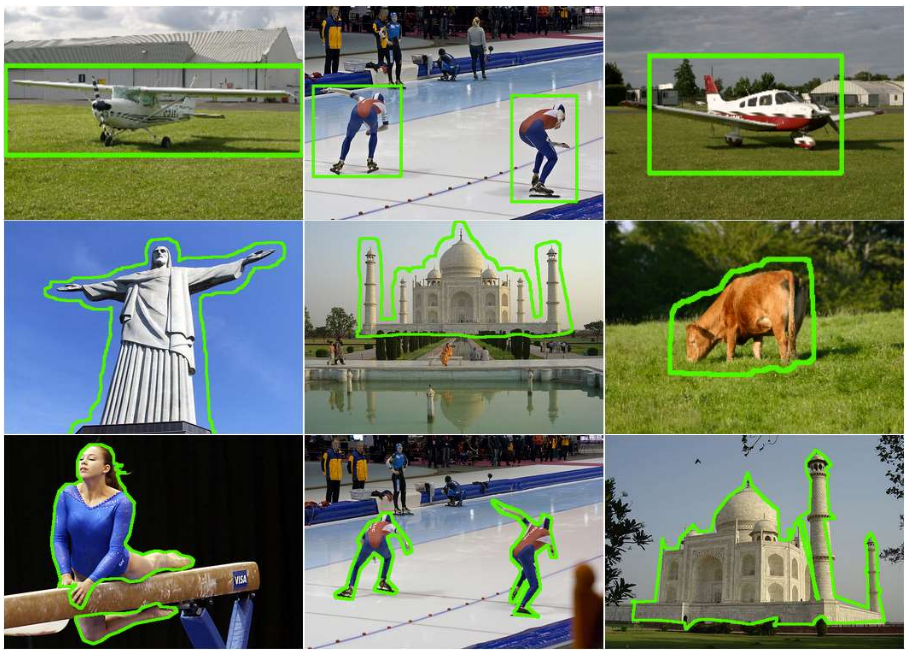
Tight Polygon

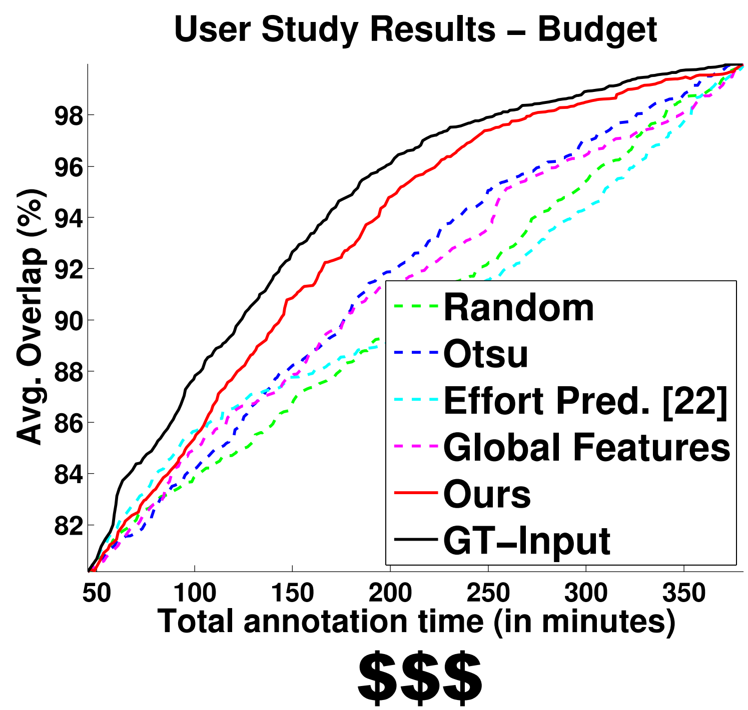
2 Budgeted Selection

$$$

Using difficulty predictions to intelligently gather annotations

Given a cost **budget**, can we maximize the accuracy crowd will achieve in collaboration with algorithm?

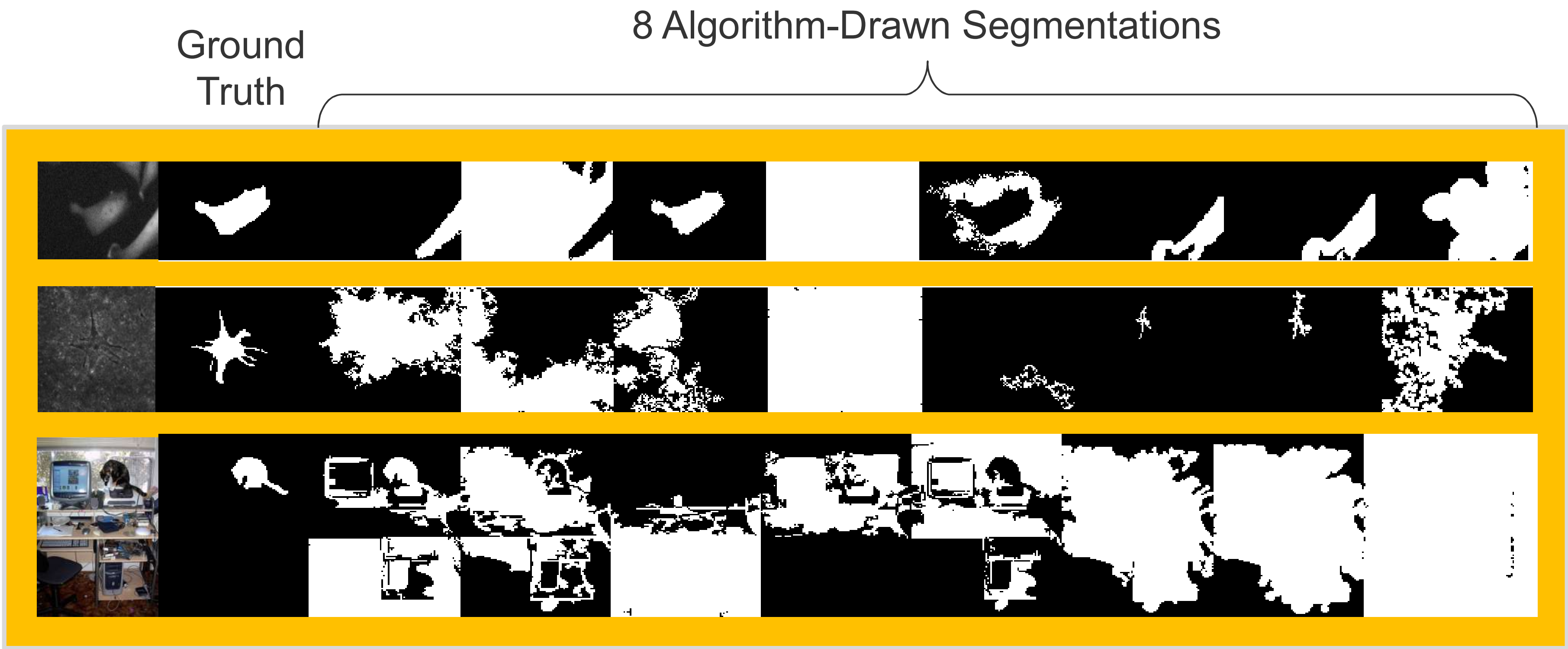




101 Turkers contribute annotations

Kristen Grauman, UT Austin

Learning the failure behavior per segmentation algorithm



Pinpoint which method new image should go to…or when to “pull the plug” and go to human annotator.

Gurari et al. CVPR 2016

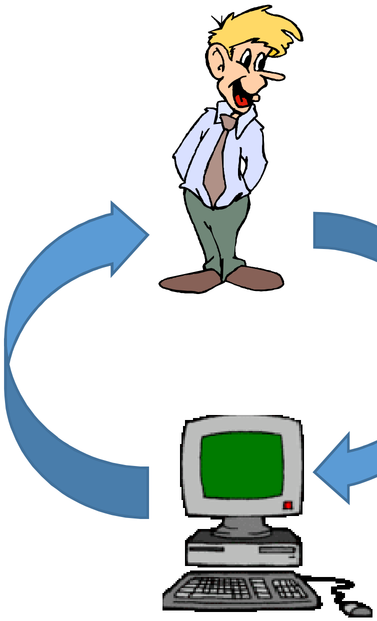
Our goal

Active and interactive segmentation methods to predict exactly where and how human intervention is needed

This talk:

1. Given an image, what strength of annotation is needed?
2. Given a collection of images, which ones need human input?
3. Given a video, how to propagate minimal human input?

Symbiosis in Segmentation

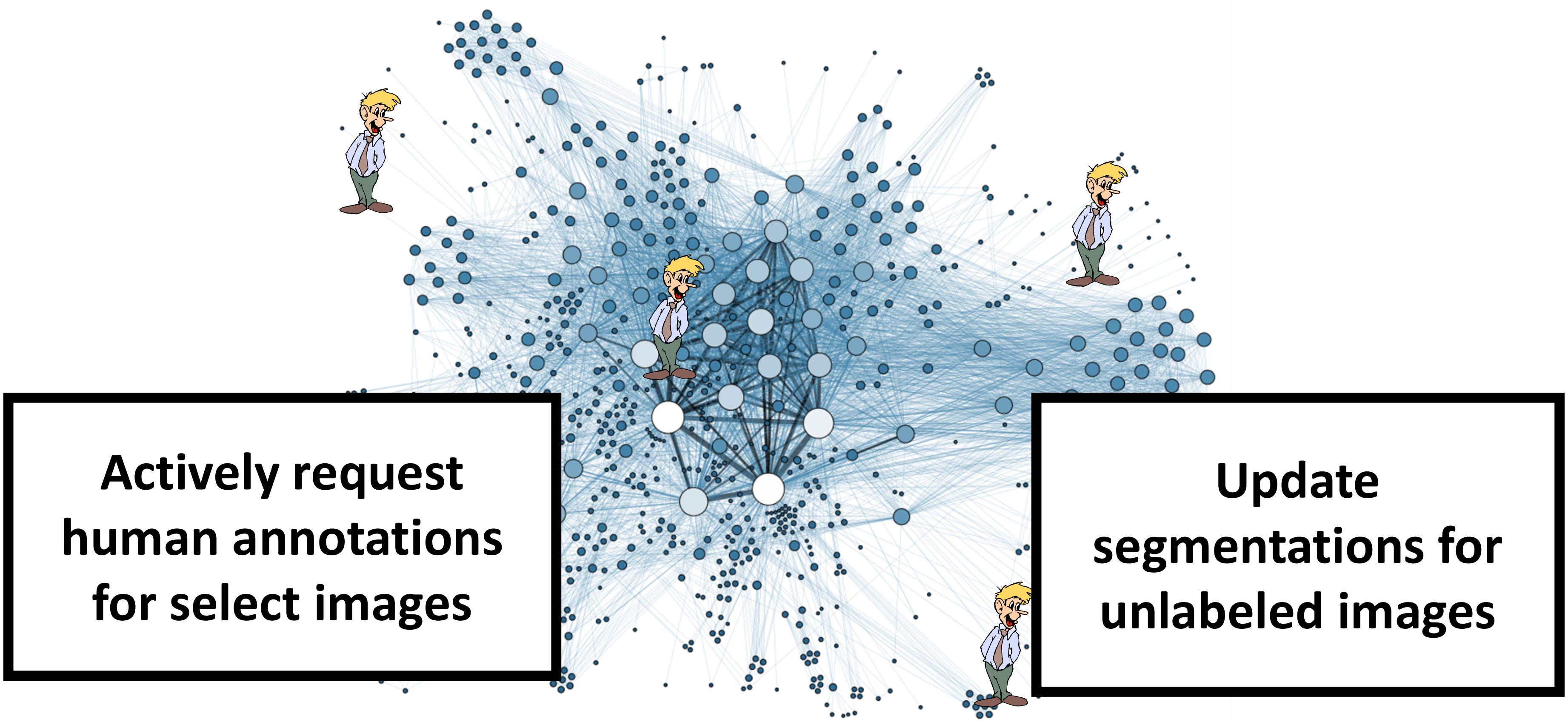


Traditional approach: Propagate human input within the image.

[Rother et al. 2004, Boykov & Jolly 2001, Mortensen & Barrett 1995, Tang et al. 2013 …]

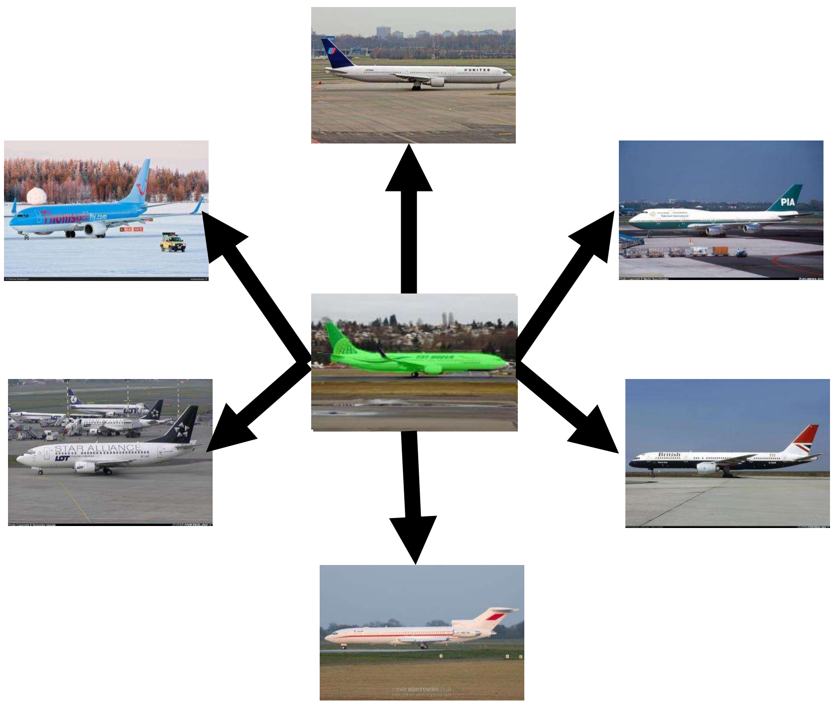
Our goal: Active propagation

How to propagate human input segmentations across multiple images/frames?

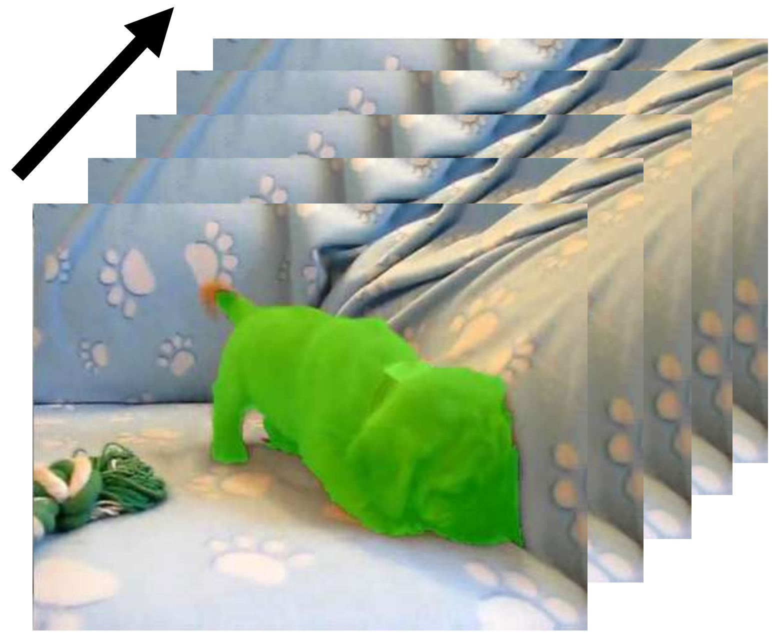


Key question 1: How to propagate?

Given some subset of labeled data, how to propagate to unlabeled data

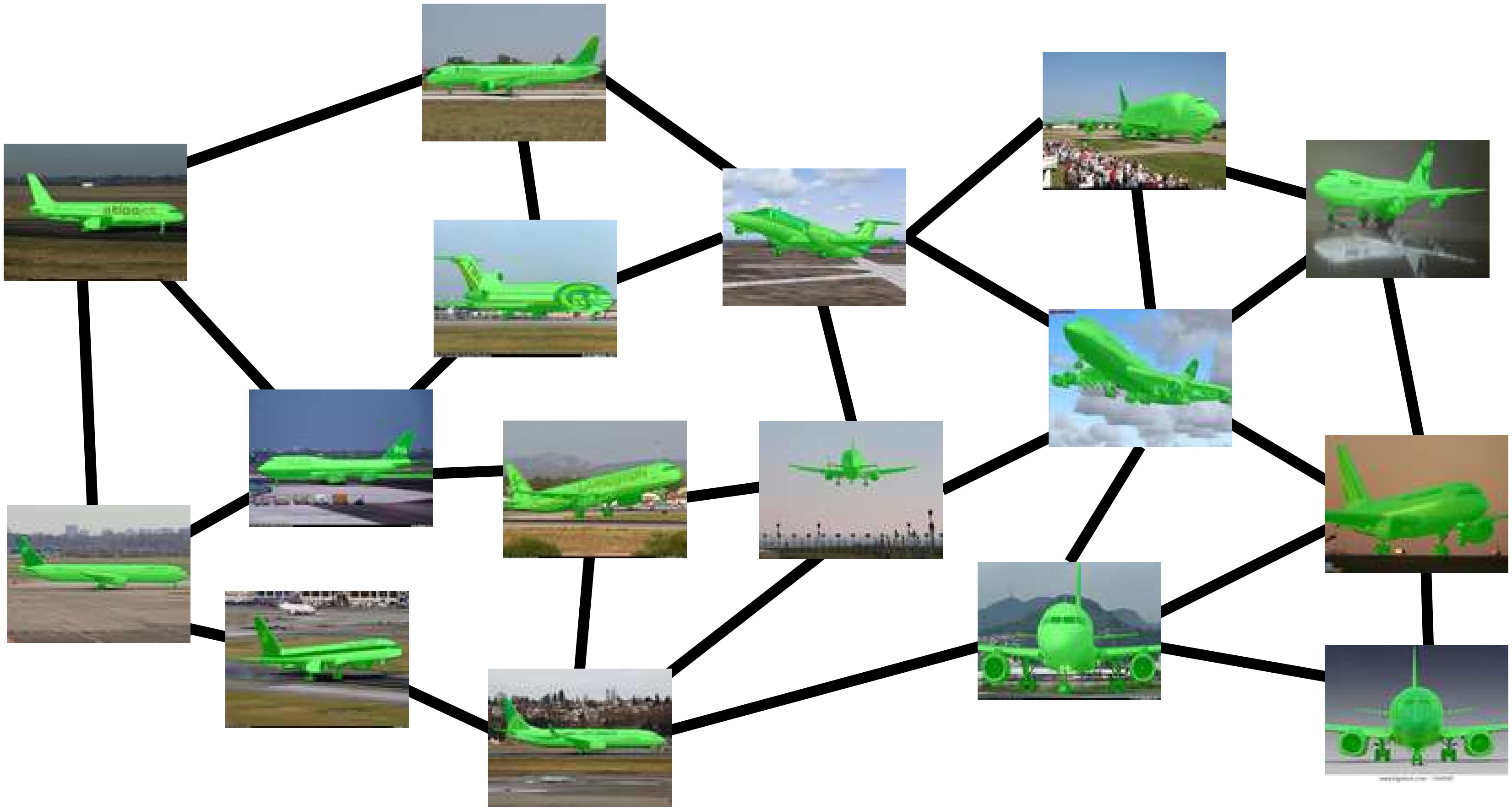


Images



Video

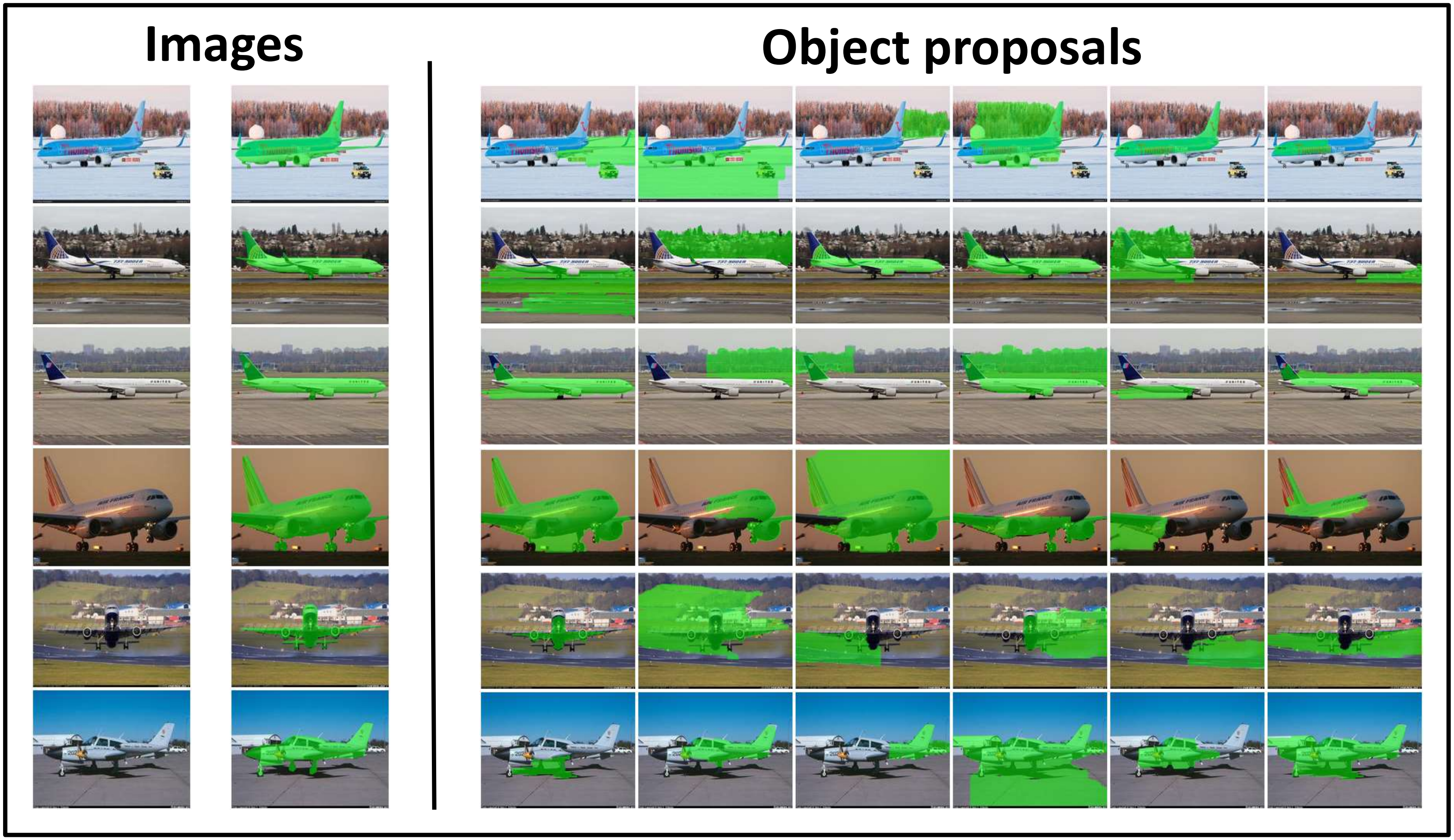
Weakly Supervised Scenario



Exploit repeated patterns by jointly segmenting out the foreground object

Approach – Segmentation Propagation

Generate bottom-up object proposals for each image

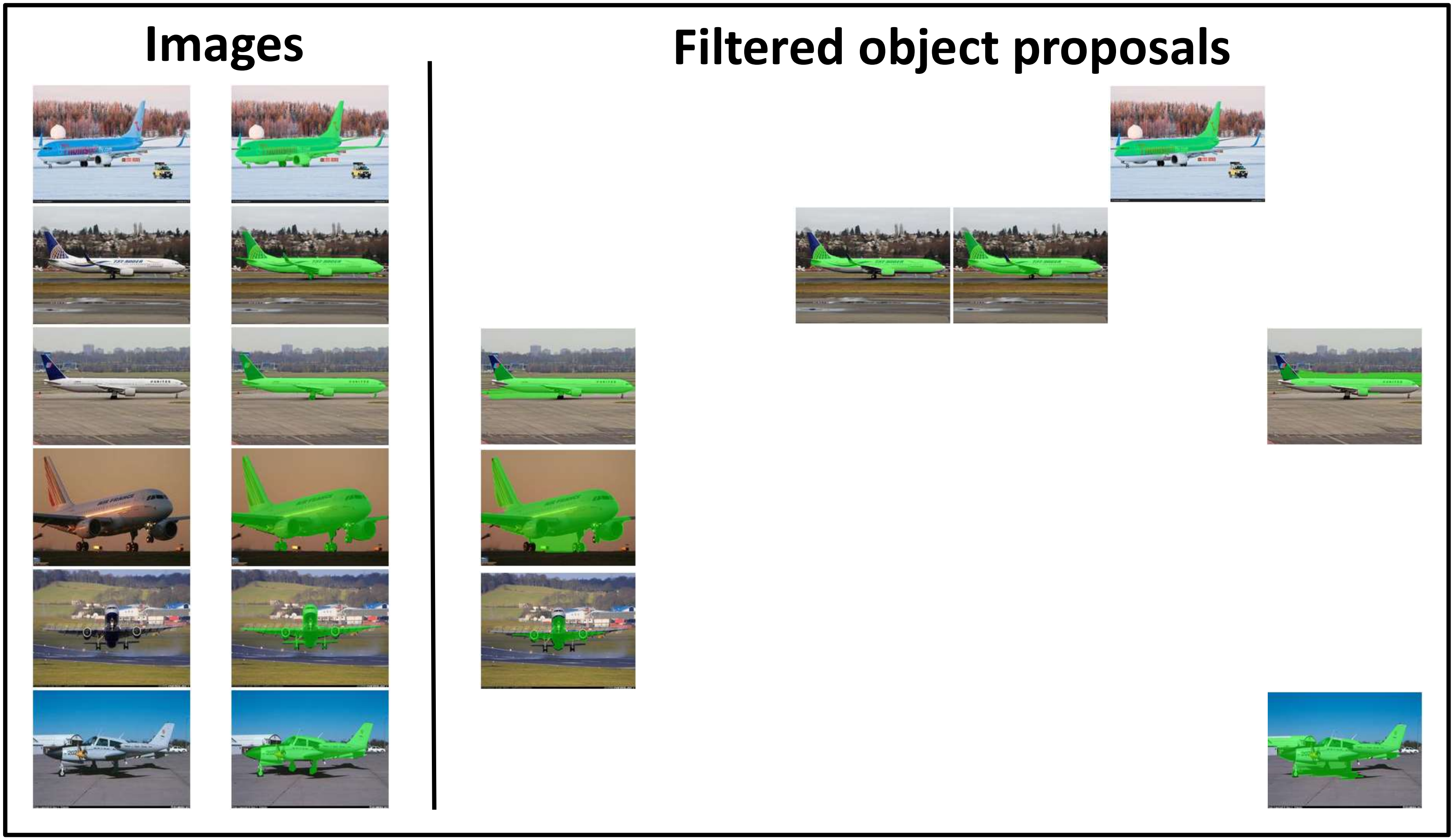


[Carreira 2012, Arbelaez 2014]

[Jain & Grauman, CVPR 2016]

Approach – Segmentation Propagation

Goal: Select "good" proposals in each image



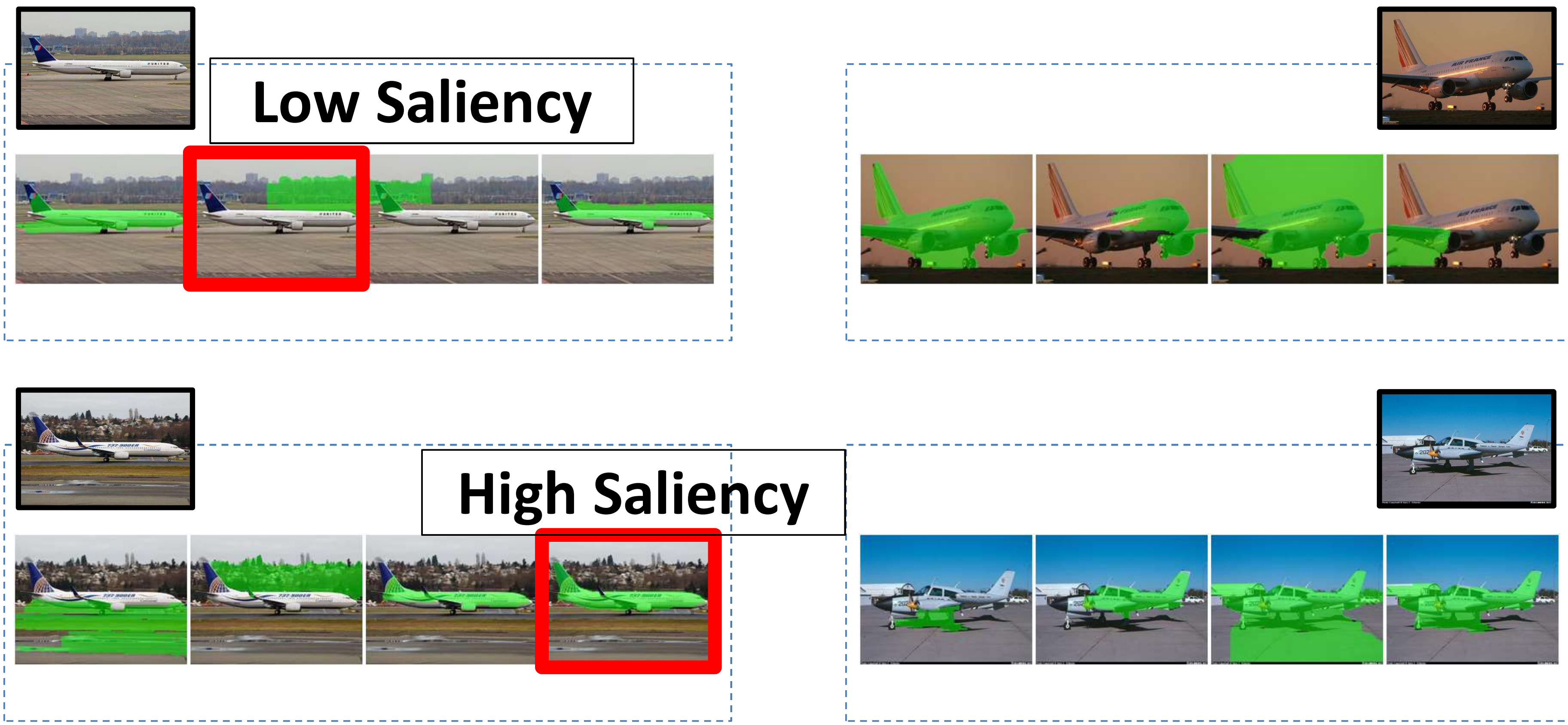
[Carreira 2012, Arbelaez 2014]

[Jain & Grauman, CVPR 2016]

Approach – MRF Joint Segmentation

Define a joint segmentation graph over region proposals

Unary Term: Average region saliency

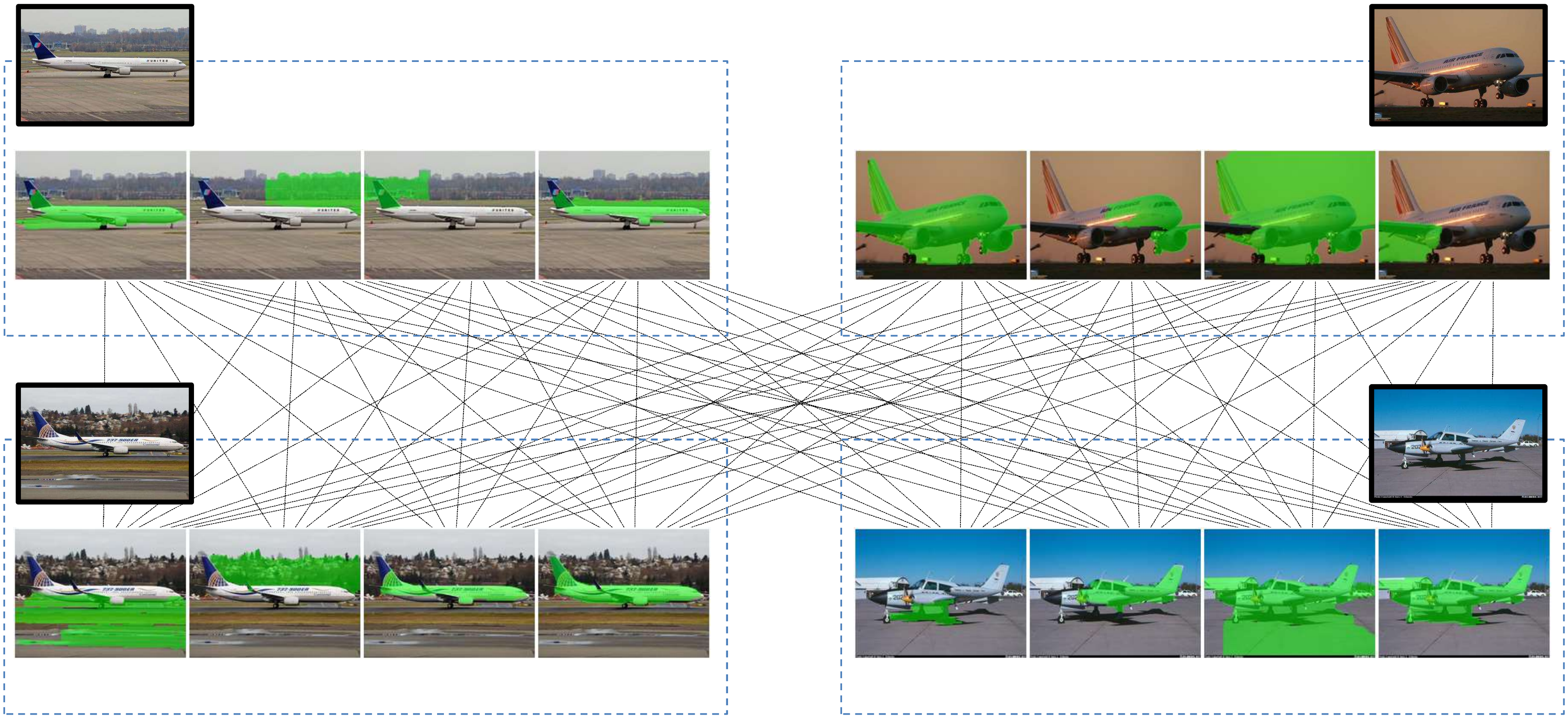


$$E(\mathcal{Y}) = \boxed{\sum_{R_{ij}} -\log \Phi(Y_{ij})} + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})$$

[Jain & Grauman, CVPR 2016]

Approach – MRF Joint Segmentation

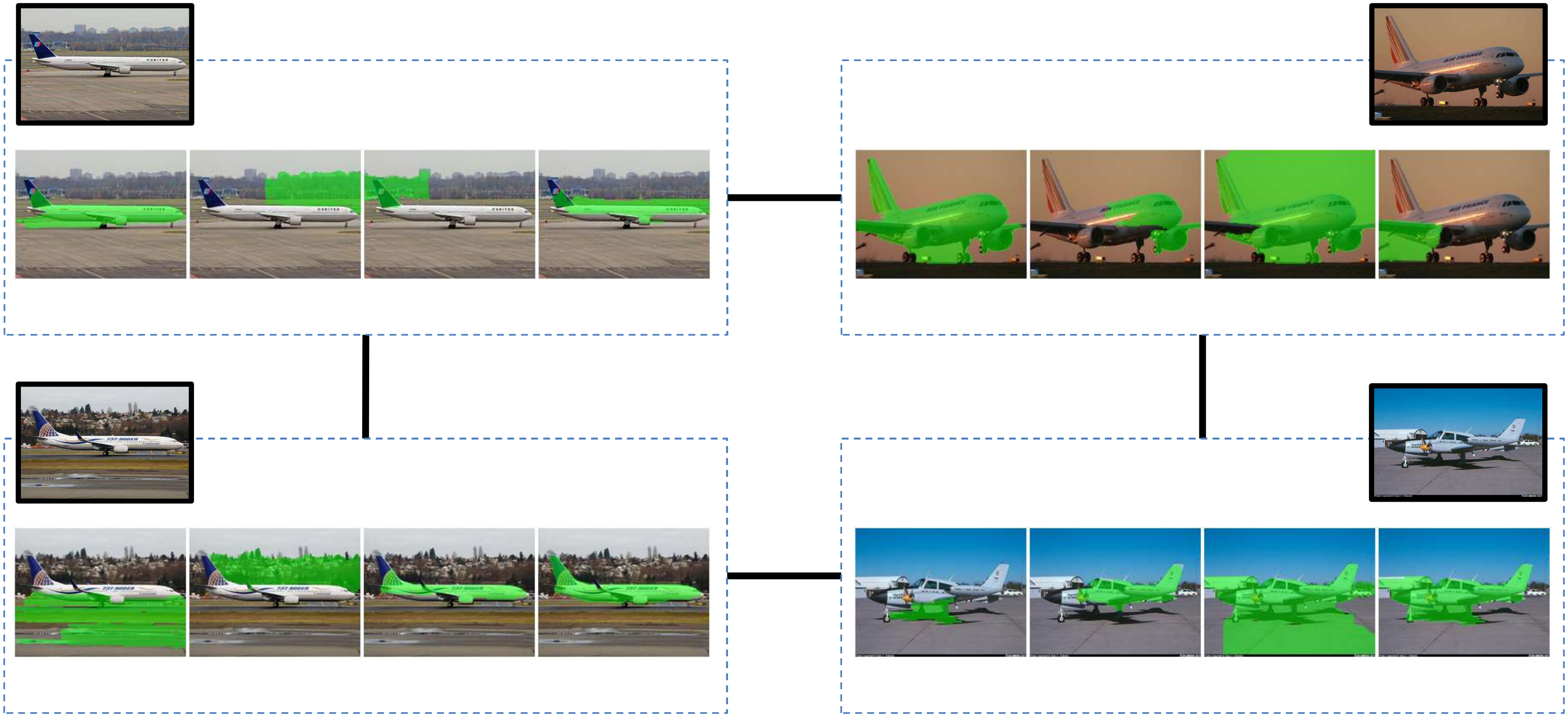
Pairwise connections between all region proposals



$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \boxed{\sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})}$$

Approach – MRF Joint Segmentation

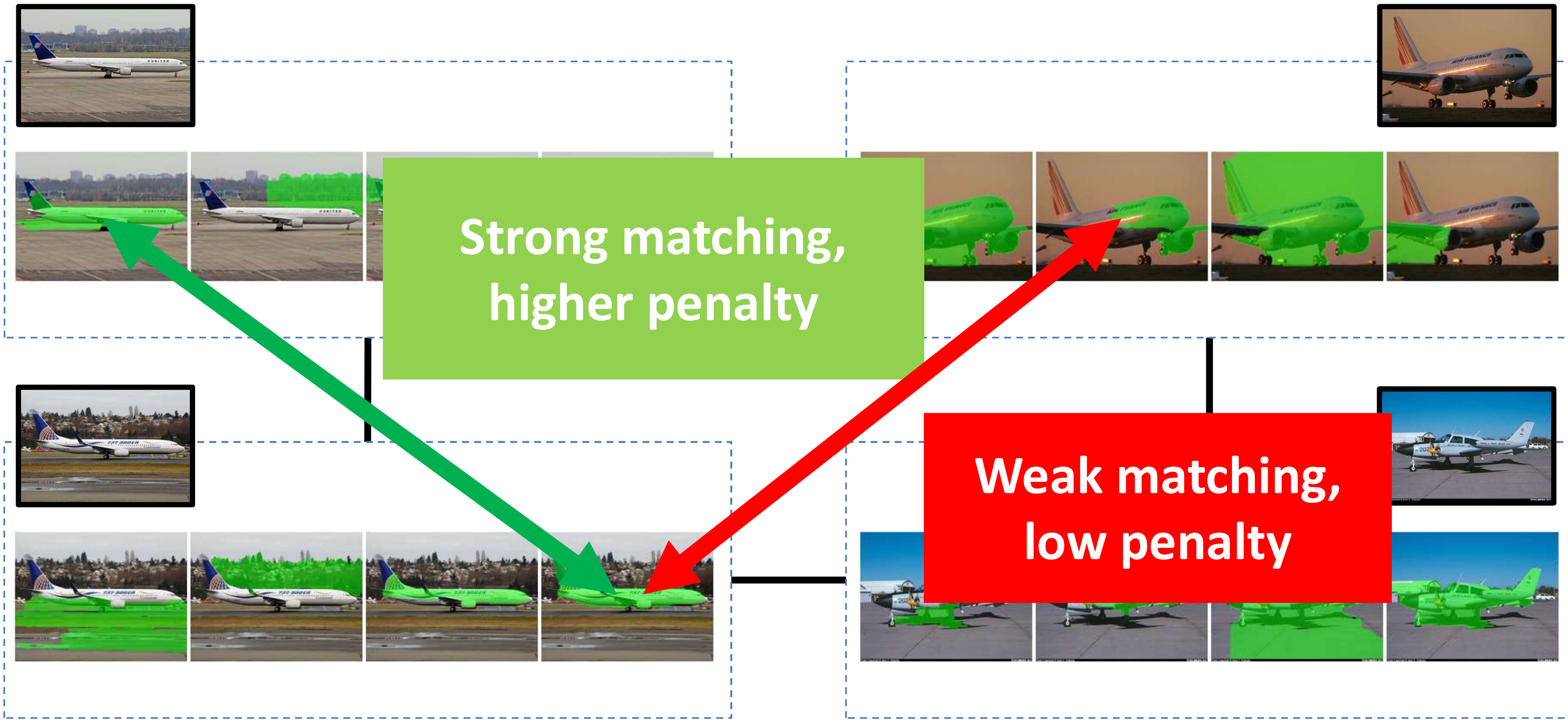
Pairwise connections between all region proposals



$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \boxed{\sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})}$$

Approach – MRF Joint Segmentation

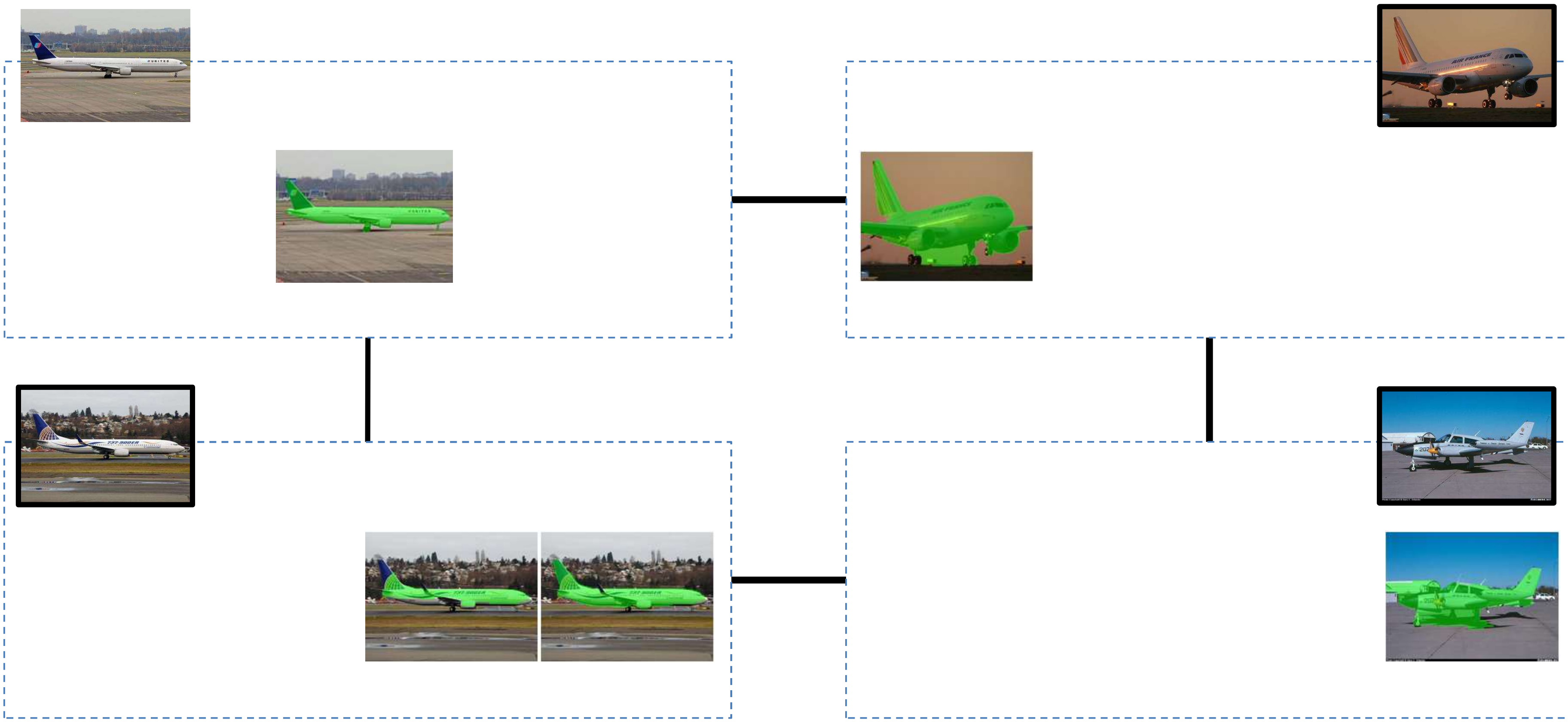
Pairwise connections between all region proposals



$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \boxed{\sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})}$$

Approach – MRF Joint Segmentation

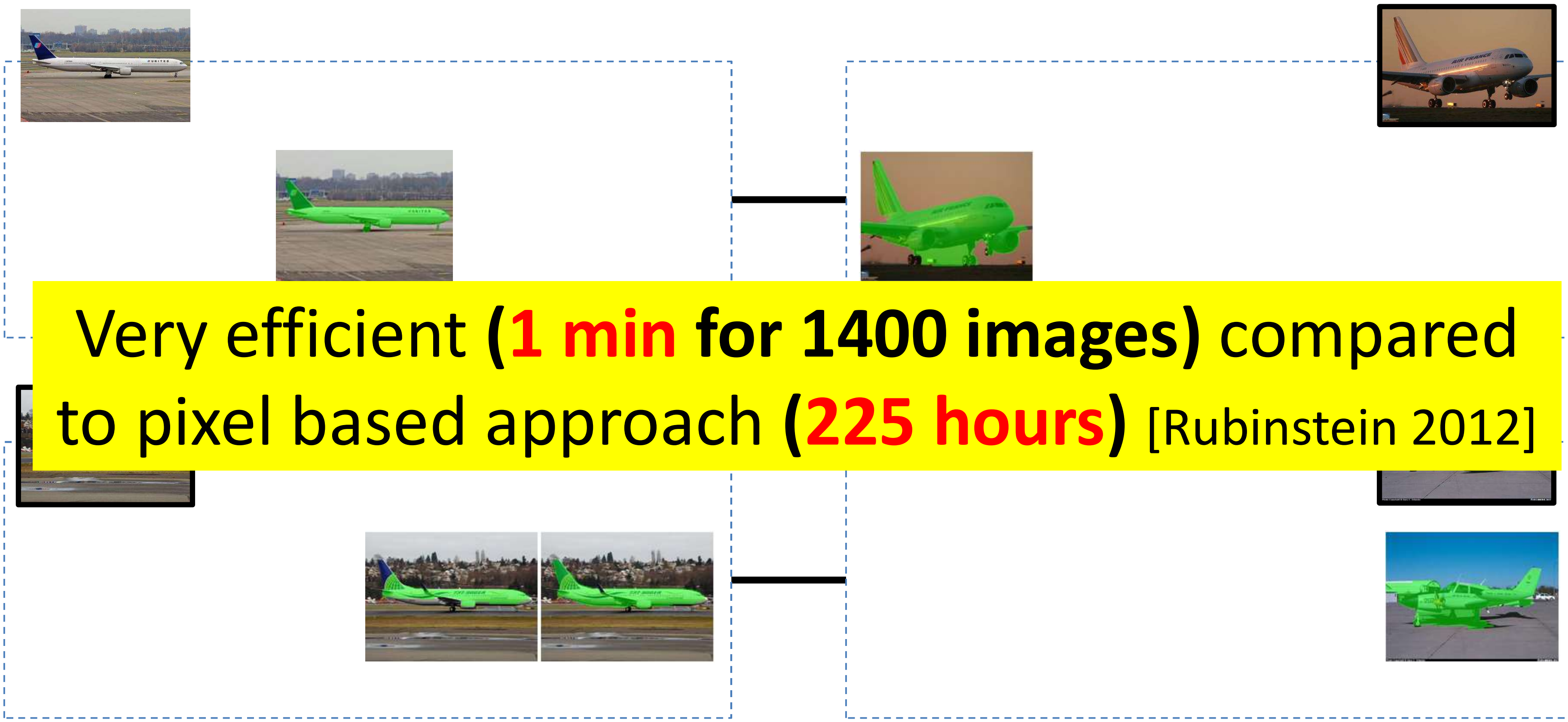
Energy minimization using Graph-cuts



$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})$$

Approach – MRF Joint Segmentation

Energy minimization using Graph-cuts

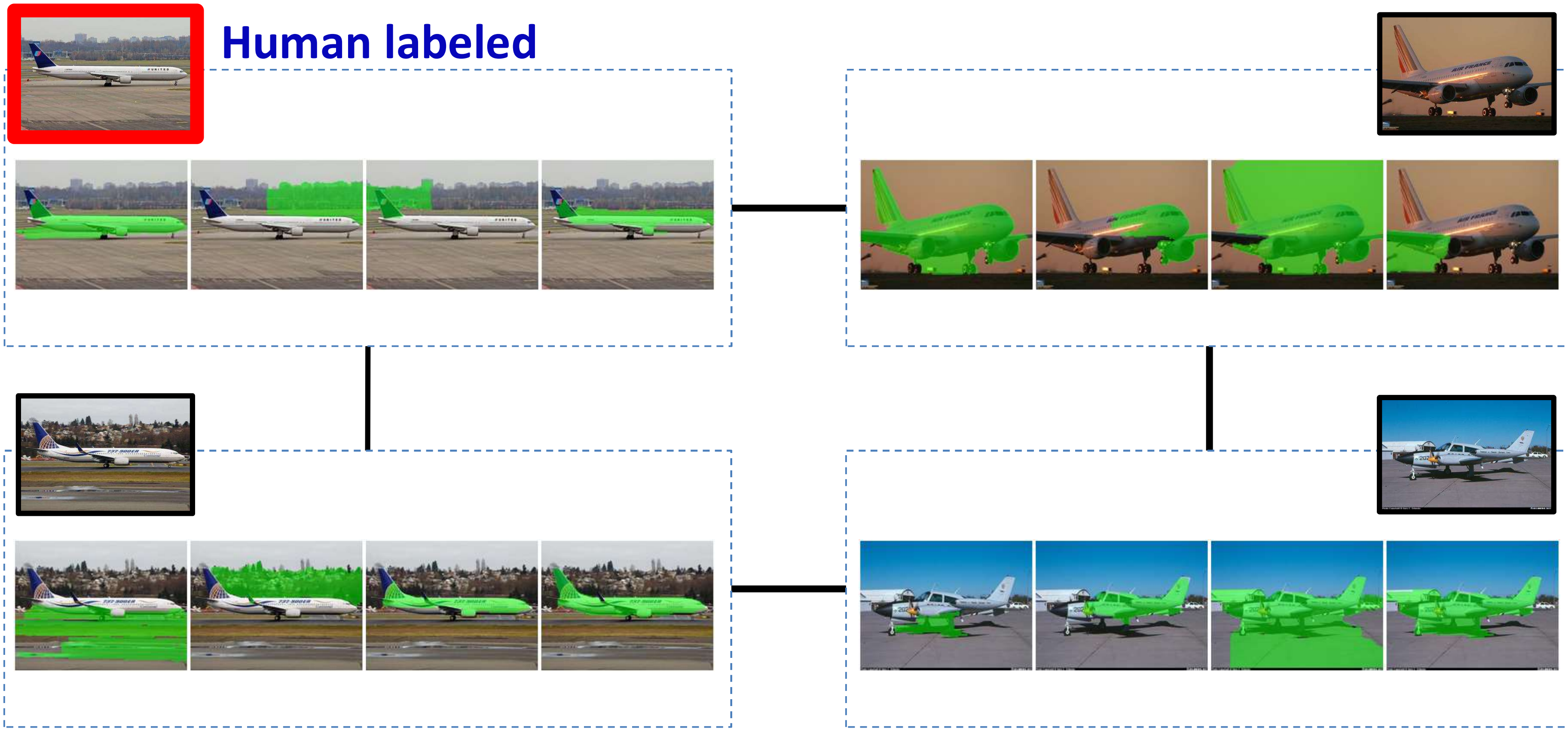


Very efficient (**1 min for 1400 images**) compared to pixel based approach (**225 hours**) [Rubinstein 2012]

[Jain & Grauman, CVPR 2016]

Approach – MRF Joint Segmentation

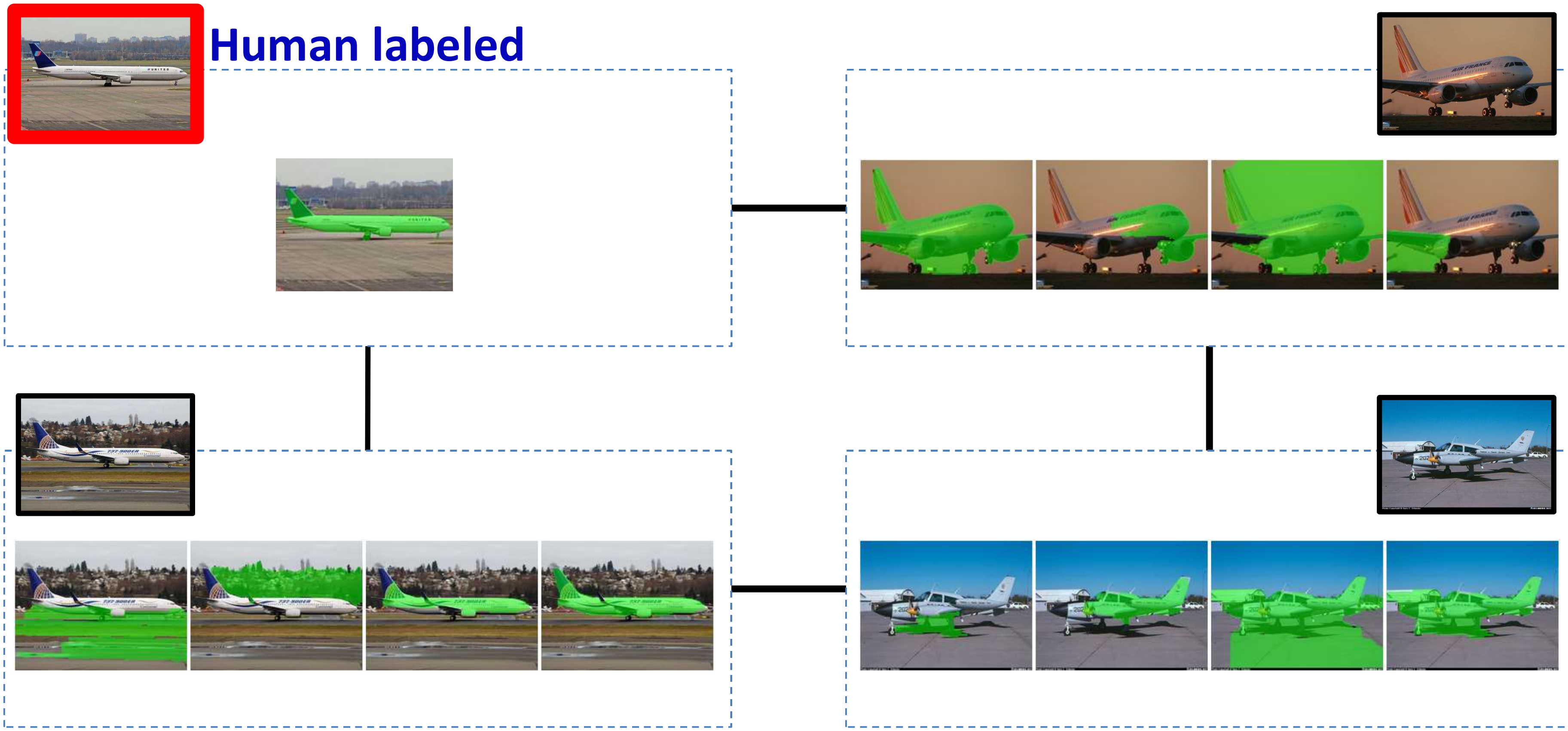
Actively choose an image to be labeled by humans



$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \boxed{\sum_{R_{ij},R'_{ij}\in\mathcal{E}} \Psi(Y_{ij}, Y'_{ij})}$$

Approach – MRF Joint Segmentation

Inject human-labeled regions in the joint graph



$$E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \boxed{\sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})}$$

Weakly Supervised Segmentation

ImageNet dataset (~1M images, 3624 classes) **[Deng 2009]**

Methods	ImageNet dataset		
	Top obj. box [64]	Tang et al. [64]	Ours
BBox-CorLoc	37.42	53.20	**57.64**

We correctly localize 41,715 more images than next best approach.

Weakly Supervised Segmentation

MIT Object Discovery Dataset [Rubinstein 2012]



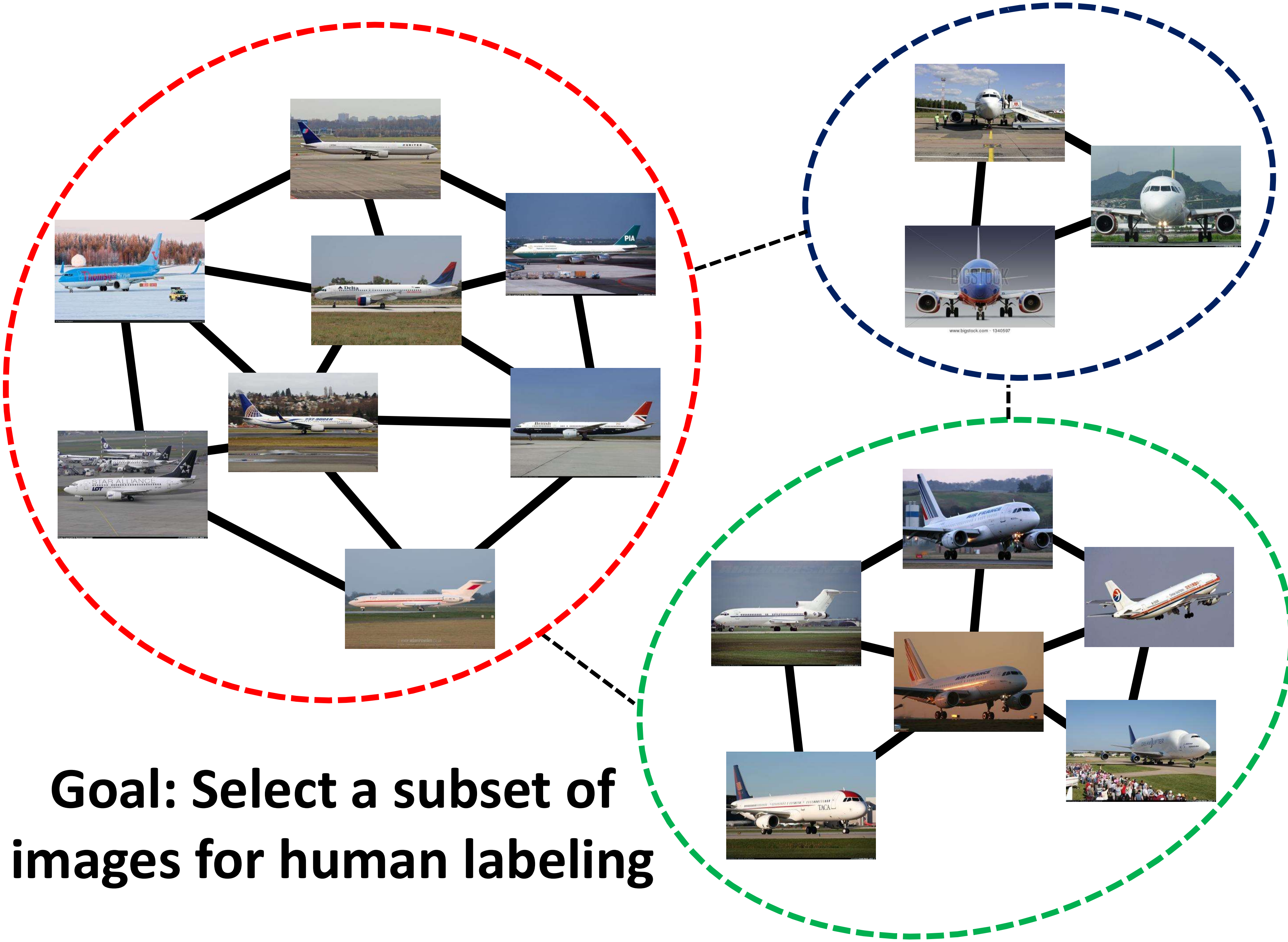
Methods	MIT dataset (subset)			MIT dataset (full)		
	Airplane	Car	Horse	Airplane	Car	Horse
# Images	82	89	93	470	1208	810
Joulin et al. [34]	15.36	37.15	30.16	n/a	n/a	n/a
Joulin et al. [35]	11.72	35.15	29.53	n/a	n/a	n/a
Kim et al. [37]	7.9	0.04	6.43	n/a	n/a	n/a
Rubinstein et al. [59]	55.81	64.42	51.65	55.62	63.35	53.88
Chen et al. [16]	54.62	**69.2**	44.46	60.87	62.74	**60.23**
Ours	**58.65**	66.47	**53.57**	**62.27**	**65.3**	55.41

Consistently good performance that boosts state of the art in most cases

Key question 2: Which to annotate?

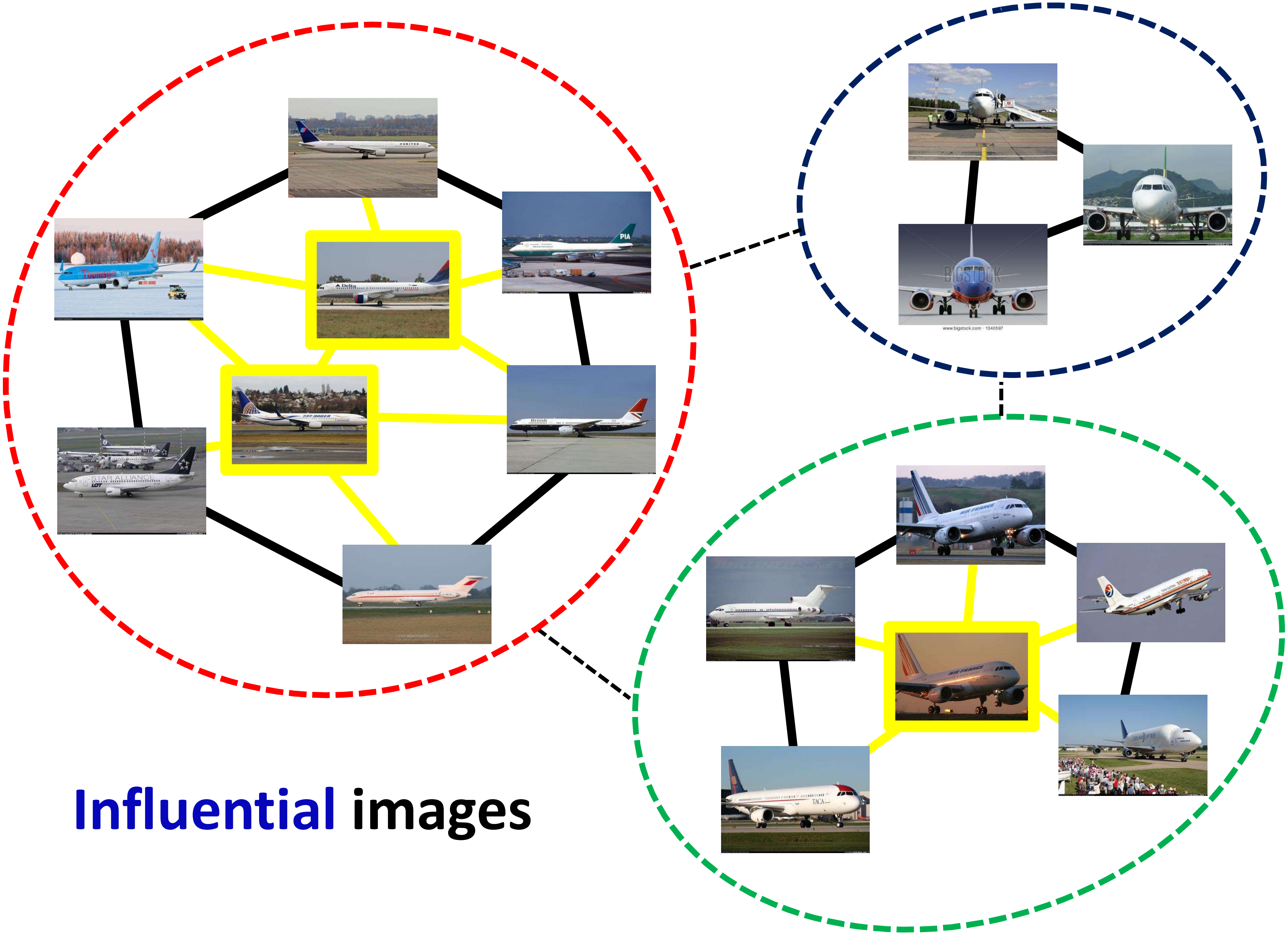
Given an annotation budget, which ones ought to be labeled by human annotators?

Active Selection



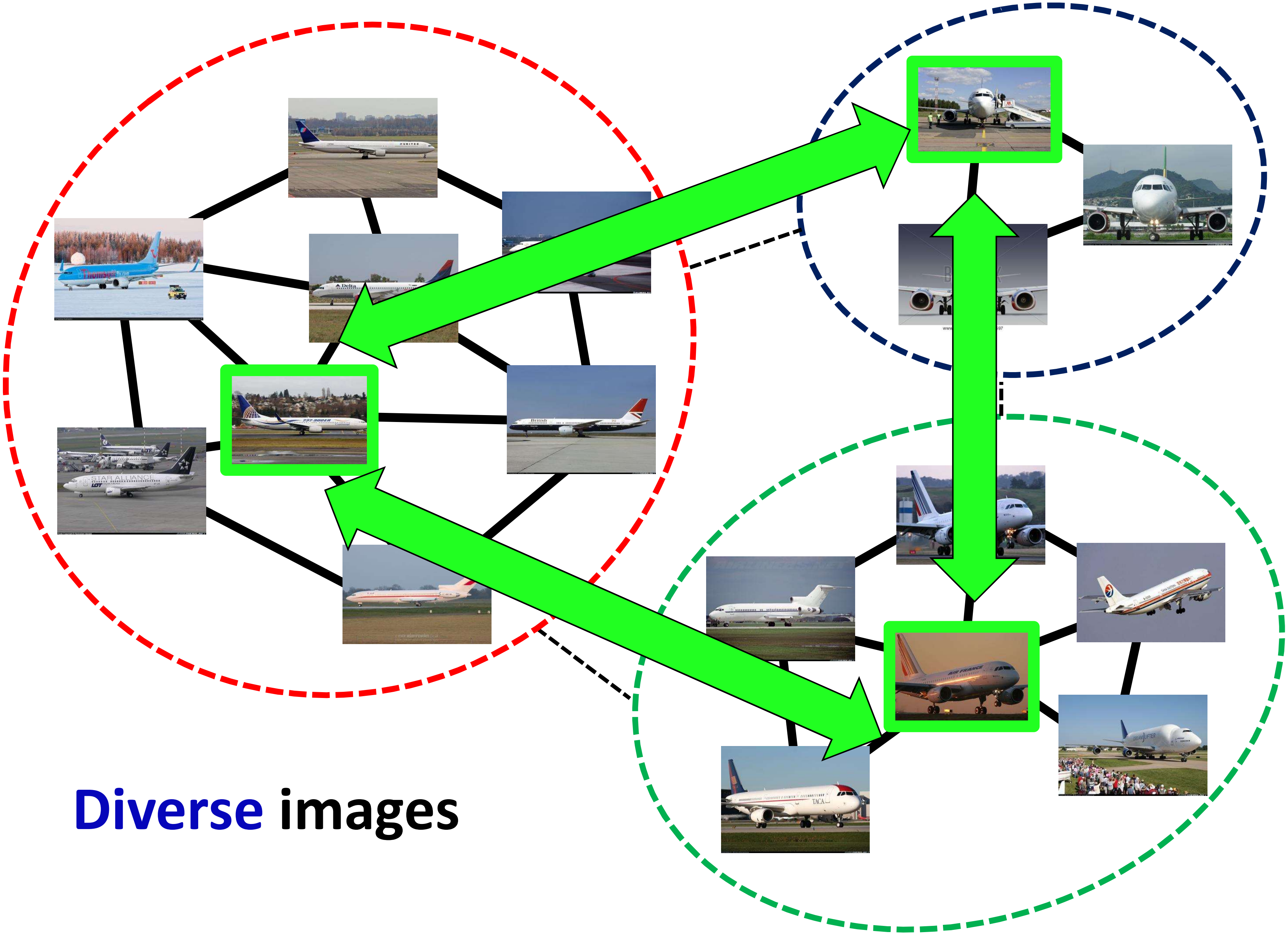
Goal: Select a subset of images for human labeling

Active Selection

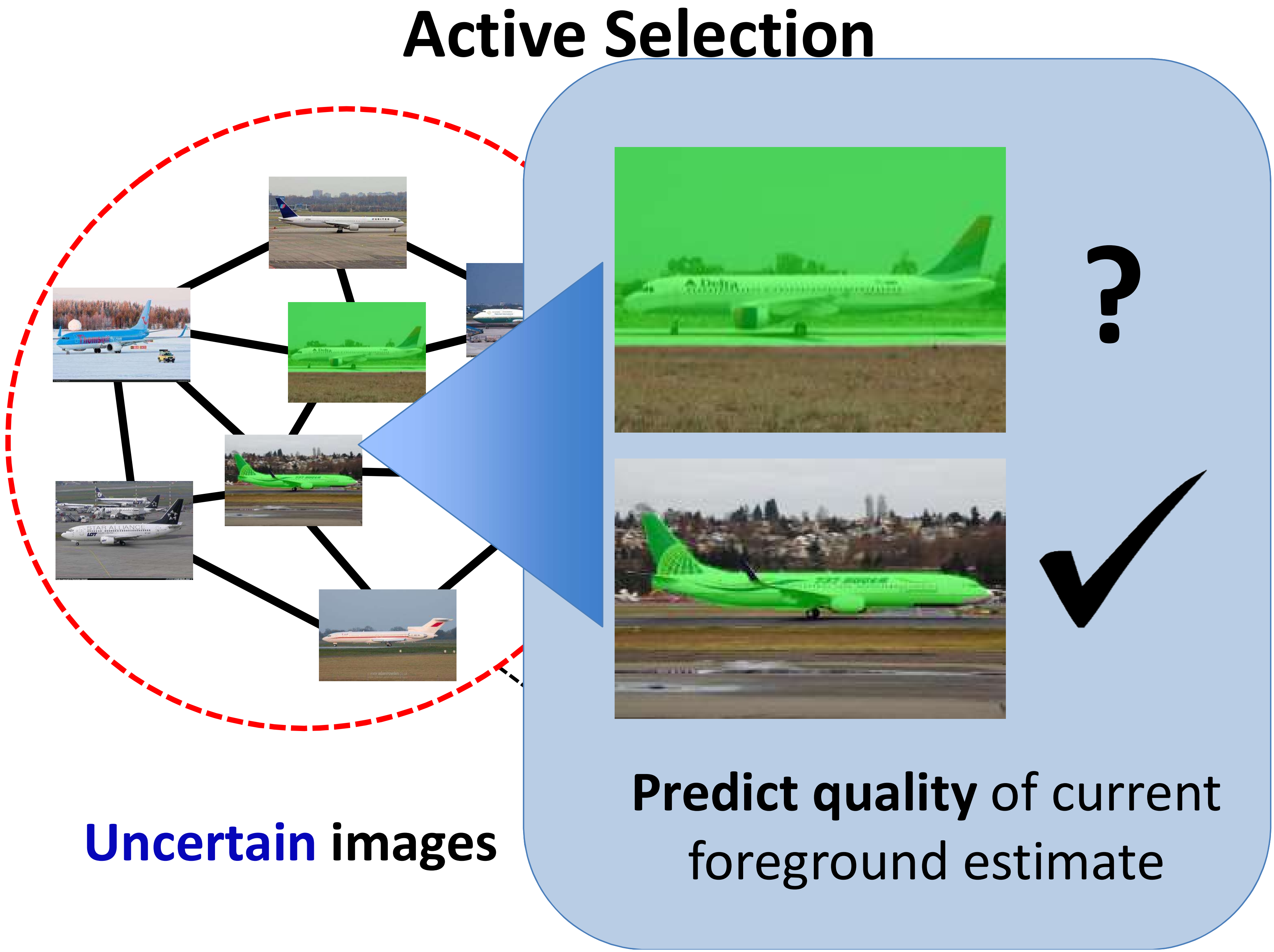


Influential images

Active Selection

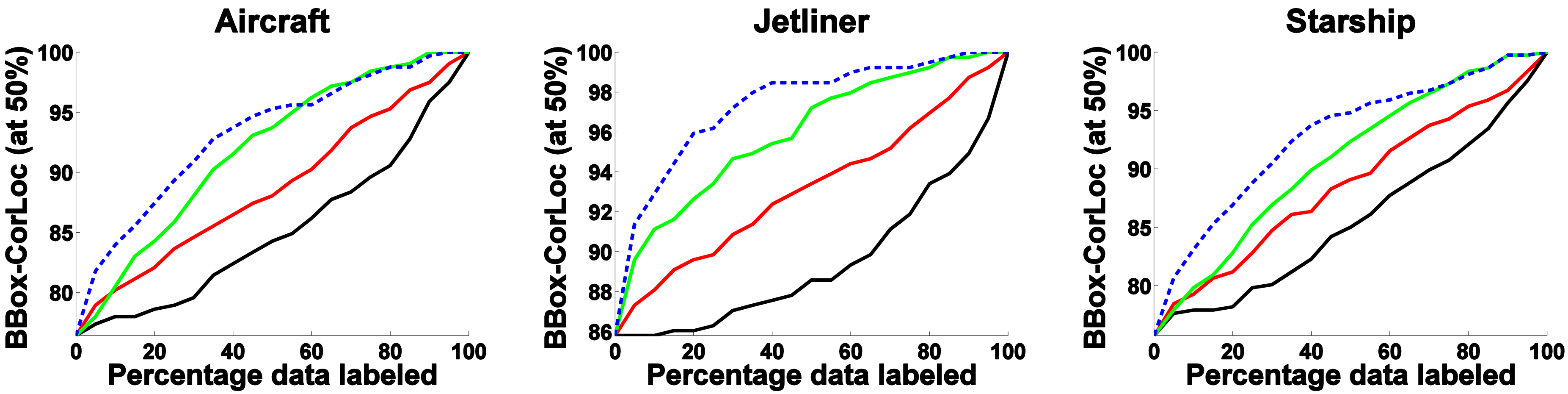


Diverse images

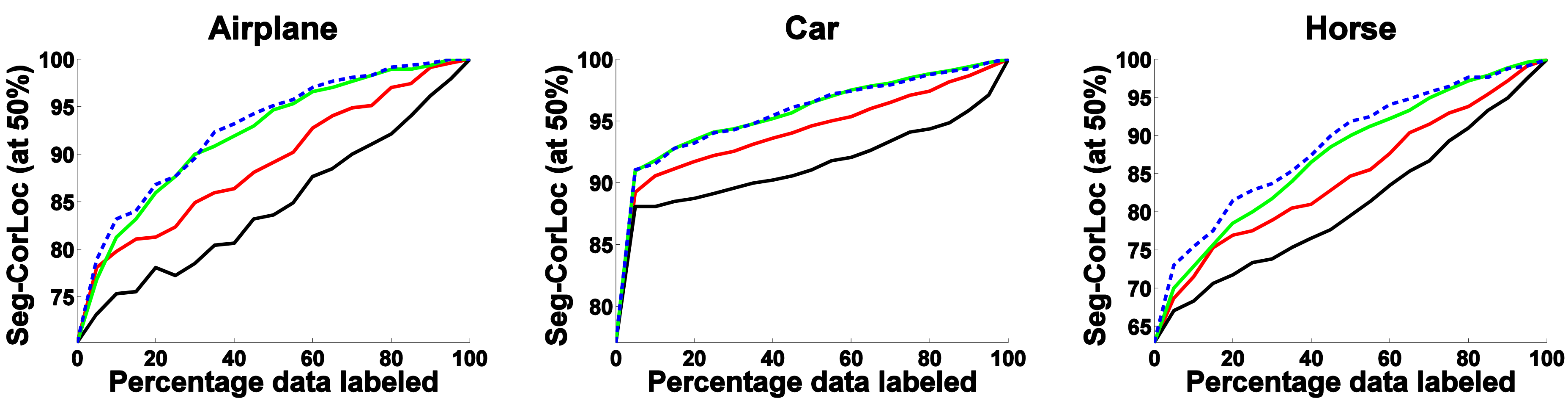

Active Selection
Uncertain images
Predict quality of current foreground estimate

Active Segmentation Propagation

ImageNet Dataset



MIT Object Discovery Dataset



Random — PageRank [Rubinstein 2012] — Ours without uncertainty — Ours

Our goal

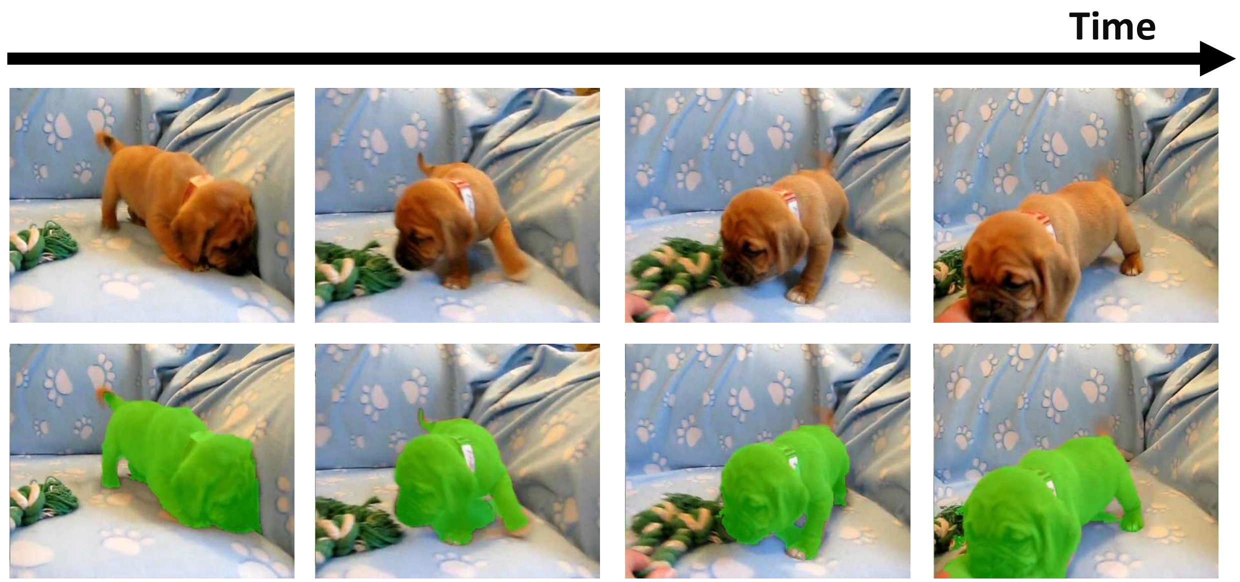
Active and interactive segmentation methods to predict exactly where and how human intervention is needed

This talk:

1. Given an image, what strength of annotation is needed?
2. Given a collection of images, which ones need human input?
3. Given a video, how to propagate minimal human input?

Propagation in Video: Problem

Existing methods [Tsai 2010, Fathi 2011, Vijayanarasmihan 2012] **can only enforce local consistency** in space and time (using pairwise connections).



Labeled frame **Automatic propagation of object labels**

Robust foreground propagation requires **capturing long range dependencies** as object evolves in shape over time.

Propagation in video

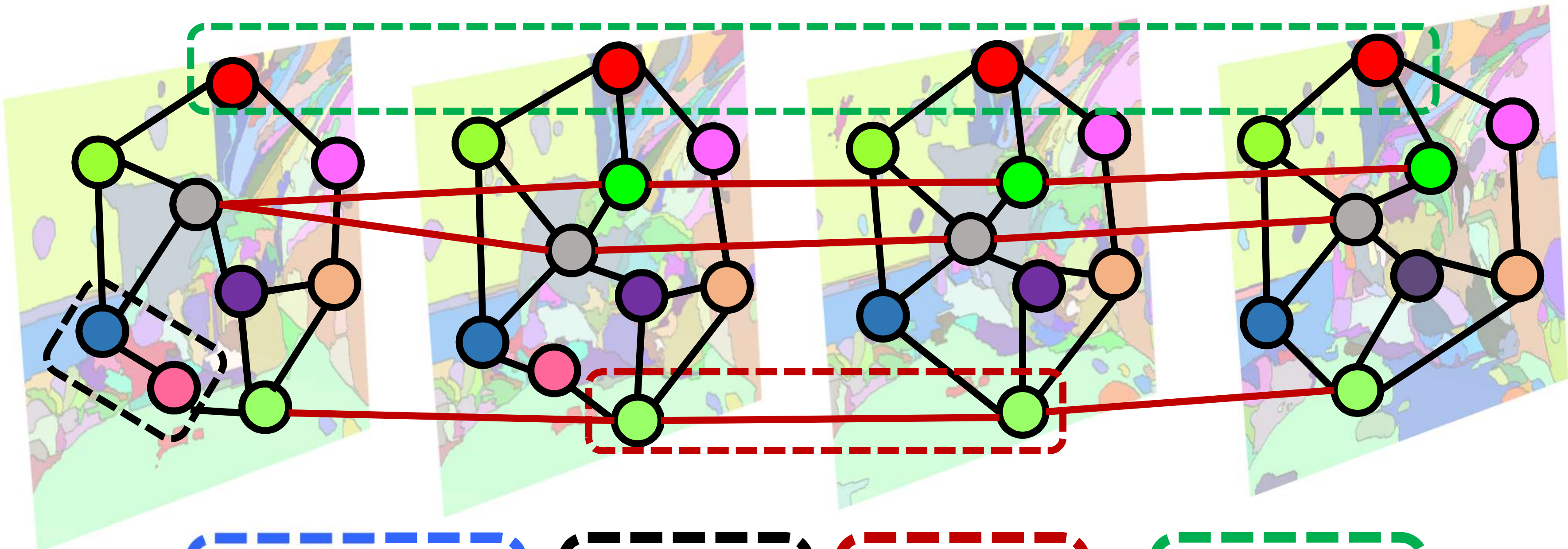
Supervoxels: bottom-up space-time regions

[Grundmann 2010, Xu 2012]

Our idea: Higher order potentials over supervoxels to enforce long term temporal consistency

S. Jain, K. Grauman, Supervoxel-Consistent Foreground Propagation in Video, ECCV 2014

Propagation in video



Appearance & Spatial Prior

Spatial Cliques

Temporal Cliques

Supervoxel Cliques

$$E(\mathcal{Y}) = \underbrace{\sum_{(t,i)\in\mathcal{X}} \Phi_t^i(y_t^i)}_{\textit{Unary potential}} + \underbrace{\sum_{\substack{[(t,i),(t',j)]\in\mathcal{E}\\ t'\in\{t,t+1\}}} \Phi_{t,t'}^{i,j}\left(y_t^i, y_{t'}^j\right)}_{\textit{Pairwise potential}} + \underbrace{\sum_{v\in\mathcal{S}} \Phi_v\left(y_v\right)}_{\textit{Higher order potential}}$$

Assign soft preferences for label consistency within supervoxels

Robust P^n model [Kohli 2008]

Results

Video **Supervoxels**

PF-MRF
[Vijayanarasimhan 2012]

Ours
[Jain & Grauman, ECCV 2014]

Click Carving for video segmentation

- Interactively segment the frame to be propagated: boundary clicks fetch relevant object proposals

[Jain & Grauman, HCOMP 2016]

Click Carving for video segmentation

- Results achieved with average of 2 user clicks

[Jain & Grauman, HCOMP 2016]

Summary

Active human-machine collaboration for foreground object segmentation in images and video

- Active selection of sufficiently strong annotation modality to initialize interactive image segmentation
- Active segmentation propagation for large weakly supervised image collections
- Click carving and high order supervoxel potentials for segmentation propagation in video

References

- **Click Carving: Segmenting Objects in Video with Point Clicks**. S. D. Jain and K. Grauman. In Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP), Austin, TX, October 2016.
- **Active Image Segmentation Propagation**. S. Jain and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 2016.
- **Pull the Plug? Predicting If Computers or Humans Should Segment Images**. D. Gurari, S. Jain, M. Betke, and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 2016.
- **Supervoxel-Consistent Foreground Propagation in Video**. S. Jain and K. Grauman. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, Sept 2014.
- **Predicting Sufficient Annotation Strength for Interactive Foreground Segmentation**. S. Jain and K. Grauman. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, December 2013.