Active Image Segmentation Propagation

Suyog Dutt Jain Kristen Grauman
University of Texas at Austin
suyog@cs.utexas.edu, grauman@cs.utexas.edu

Abstract

We propose a semi-automatic method to obtain foreground object masks for a large set of related images. We develop a stagewise active approach to propagation: in each stage, we actively determine the images that appear most valuable for human annotation, then revise the foreground estimates in all unlabeled images accordingly. In order to identify images that, once annotated, will propagate well to other examples, we introduce an active selection procedure that operates on the joint segmentation graph over all images. It prioritizes human intervention for those images that are uncertain and influential in the graph, while also mutually diverse. We apply our method to obtain foreground masks for over 1 million images. Our method yields state-of-the-art accuracy on the ImageNet and MIT Object Discovery datasets, and it focuses human attention more effectively than existing propagation strategies.

1. Introduction

Large-scale labeled image datasets have had a transformative impact on computer vision in recent years, most notably for image classification. However, image annotation remains a costly undertaking in terms of both time and money. In particular, gathering high quality *spatial annotations*—pixel-level foreground masks—is challenging. First of all, the physical mousing actions required are time intensive (e.g., compared to simply labeling which object is present). Furthermore, non-expert annotators exhibit inconsistencies in how precisely they mark object boundaries, which means leveraging the crowd typically requires some finessing and "re-dos".

As a result, datasets with spatial annotations lag seriously behind their category-labeled counterparts. For example, while ImageNet is comprised of an impressive 14M labeled images, there are orders of magnitude fewer spatial annotations—only 1M images (7% of the dataset) offer bounding box annotations, and only 4K images (0.03%) have foreground segmentation masks [12]. While the new Microsoft COCO dataset [27] has spatial annotations for

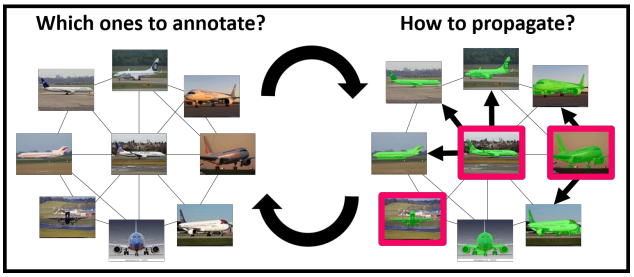

Figure 1: Our active image segmentation propagation method alternates between: (1) Actively choosing images which once annotated by humans will likely be most useful in propagating segmentations to other images and (2) Given human annotations on actively chosen images (marked in pink), propagating them (dark arrows) to generate segmentations for other unlabeled images. Best viewed in color.

2.5M object instances, these were obtained only by investing staggering amounts of time and money; at over 22 person hours per 1,000 segmentations [27], that's more than $400,000 if paying minimum wage.

The difficulty in generating foreground spatial annotations for image collections is problematic given their high potential utility. For example, they are useful to build training sets for region-based object detectors, or improve image retrieval by focusing on the region of interest. Aside from serving as training data, there are also applications that directly use a well-segmented image database, such as data-driven image synthesis or retargeting.

Aware of this need, researchers have developed an array of weakly supervised segmentation algorithms [2, 10, 44, 20, 36, 39, 35, 42, 12]. The main idea is to take a pool of images known to contain the same object category, and exploit the repeated patterns to jointly segment out the foreground per image. On the one hand, this paradigm is attractive for its low manual effort, especially since such weakly labeled images are readily available on the Web via keyword search. On the other hand, the resulting segmentations are imperfect. No matter the method, the foreground masks hit a ceiling of accuracy since the segmentation task is underconstrained even with weak supervision.

We propose an intermediate solution. Rather than rely solely on human-provided segmentations (accurate but too expensive) or automatic segmentations (inexpensive but

too inaccurate), we develop a *semi-automatic segmentation propagation* approach. The idea is to actively request human annotations for select images that, once labeled with their foreground, are most expected to help co-segment the remaining unlabeled images. The propagation engine proceeds in stages, each time (1) using the recently annotated images to revise foreground estimates in all unlabeled images, and (2) using those results to determine the next best batch of images to present to human annotators. In this way, we neither restrict ourselves to the saturation point of the fully automatic methods, nor do we get large volumes of data labeled by humans (see Figure 1).

To achieve this goal, we develop an active selection approach tailored to foreground propagation. It operates on a graph constructed over all images in the collection. Our active selection process favors choosing images that are *uncertain*—poorly explained by any images labeled so far, as well as *influential*—similar to many unlabeled images, making their foreground mask transferrable—and mutually *diverse*—so as to avoid redundant human effort. A critical part of our method design is its stagewise propagation, which permits both human-annotated *and* automatically annotated images to influence the system's view of what most needs human attention next.

Our framework differs in important ways from existing work on both active learning and segmentation propagation. Active learning methods for recognition aim to train a model that will make accurate category label predictions on unseen test images (e.g., [41, 46, 43]). In contrast, our goal is to get all available images spatially annotated by semi-automatic propagation (i.e., ours is a transductive setting). Whereas active learning iteratively refines a single classifier, our method iteratively refines the foreground estimates for a collection of images. There is very limited prior work on segmentation propagation, and existing methods are either passive [12] or only select annotations to initialize the algorithm [36]. A key insight of our technical approach is to repeatedly analyze the current segmentations (both human-*and* algorithm-provided) to actively decide on subsequent annotations.

Our main contributions are (1) a scalable approach for semi-automatic segmentation propagation in image collections, (2) a stage-wise active selection algorithm to determine the images which appear most valuable for human annotation and (3) a large-scale empirical demonstration that actively allocating human effort can lead to substantial savings in annotation costs for the segmentation problem.

We evaluate our approach on ImageNet [37] and the MIT Object Discovery [35] dataset. Applying our method to more than 1 million images, we show that intelligently focusing human effort leads to significantly better foreground extraction. As a secondary result, we show that in its fully automatic form, our model produces state-of-the-art foreground segmentation accuracy on these widely used datasets.

2. Related Work

Fully supervised and unsupervised methods Current segmentation methods can be organized according to the human supervision they assume. One extreme consists of strongly supervised semantic segmentation methods (e.g., [40, 28, 47]), which train object models from manually segmented multi-label images. Such methods demand substantial labeled data for training, which could be more efficiently acquired with the help of our approach. The other extreme consists of fully unsupervised methods that use unlabeled images to discover object categories (e.g., [38, 26]). Whereas ambiguity about the object(s) of interest poses a significant challenge for those methods, we work in the "weakly labeled" setting.

Weakly supervised foreground segmentation Our work is more related to *weakly supervised* methods, which aim to segment the foreground object(s) while exploiting the fact that all input images contain instances of the same object category [2, 10, 44, 20, 21, 36, 39, 35, 9].[1] Depending on the method, the output segmentation might be pixel-level masks [2, 44, 20, 21, 39, 35] or bounding boxes [10, 42]. Recent advances include ways to accommodate noisily labeled inputs [35, 42], multi-class data [21, 20], and object proposal regions [44, 10, 1]. While typically the entire weakly labeled set is treated as a whole, some methods aim to limit the influence of co-segmentation to closely related images [36, 35, 9, 16].

Our basic co-segmentation engine builds on this rich body of work, with refinements that (as we will see in results) improve the state-of-the-art when applied without any manual foreground labels. In particular, our idea for selecting and fusing multiple region proposals per image offers important advantages. More importantly, *active* segmentation propagation is new; the existing weakly supervised methods above use no human intervention.

Segmentation propagation Most closely related to our work are methods for *segmentation propagation*, which use labeled seeds to propagate foreground masks to other images in the weakly labeled set [36, 12]. Our method has two key novel aspects. First, we actively select which images should next receive foreground labels from human annotators. In contrast, existing methods are either opportunistic (and hence passive) about the labeled seeds, using only existing labeled data [12], or else select them in a one-shot manner without reacting to the impact of previously annotated examples [36]. Second, our stagewise procedure

[1] This class of techniques can also be described as *co-segmentation* or *joint segmentation* or *object discovery* or *co-localization* methods; in all cases, a set of related images is used to discover the common foreground.

constantly re-evaluates the impact of new labels, revising the current foreground estimates on all images. In contrast, [12] assumes that propagation will proceed best among the closest semantically related classes in an external object hierarchy (ImageNet), and [36] assumes that propagation will proceed best among each image's GIST neighbors. Empirically, our approach compares favorably to both existing propagation methods (cf. Section 4).

Interactive segmentation Another way to make image segmentation semi-automatic is to let a human guide the segmentation of an individual image. This is the concept behind the popular GrabCut [34] method: a user's bounding box coarsely localizes the foreground, and the system completes the pixel-level mask. Recent work considers how to guide a user to regions where a "scribble" would be most valuable [6], or predict which type of input (bounding box, contour) is best suited for an image [15]. Our method also aims to intelligently engage annotators, but our objective is to segment an entire batch of images based on minimal manual foreground masks. Whereas existing work [6, 15] considers uncertainty only within individual images, our method reasons about the image collection as a whole.

Active learning with images Active learning has been explored for object recognition and image classification [45, 46, 41, 11, 43, 4]. The goal is to focus human labels on those images that will most reduce the uncertainty of the classifier, such that it can generalize well to novel images. Selection strategies include reducing the classifier's expected error [46, 43, 4] or maximizing the diversity among the selected images [14, 11]. As discussed above, all such methods are closely coupled to their classifier of interest, and they aim to find good images to label by category. This is the case even for those that operate on image regions [45, 41, 43]. In contrast, our task is to select images from which *segmentation will propagate well*, and we aim to find good images to annotate with foreground masks.

3. Approach

Let $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ be a collection of weakly supervised images, all of which contain instances of the same object category. Our goal is to jointly segment these images, yielding a foreground object mask $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$ for each one.

We first describe the regions and descriptors we use to construct the image graph (Sec. 3.1). Then we define our joint segmentation procedure to simultaneously solve for all foreground masks, given foreground annotations on only a subset of the images (Sec. 3.2). Finally, we introduce our active procedure for identifying the set of images that should be annotated next (Sec. 3.3). Figure 2 visually illustrates all the steps.

3.1. Region proposals and descriptors

We define our segmentation graph over *region proposals*. Region proposals are "object-like" segments that are prioritized among all bottom-up regions as those being most likely to agree with true object boundaries [8, 5]. We assume that at least *some* of them capture the foreground object well—and possibly more than one per image. Thus, the goal of our joint segmentation procedure will be to identify the subset of region proposals that are good, and fuse them to obtain the final segmentation (see Sec. 3.2 for details). Apart from being more efficient than traditional pixel-based graphs (e.g., [35]), we show that a region-based representation lets us define strong pairwise consistency potentials based on regions matched across images.

Existing region proposal methods typically produce $\sim$ 500 regions per image, a large sample that may include redundant candidates and background objects. To refine the set of proposals, we develop the following filtering steps. First we generate the generic object proposals and compute a saliency map using [18]. Next we obtain two ranked lists of these proposals using saliency and objectness scores [8], respectively. We retain the union of the top 30% from each list. Then, we cluster the reduced set into r clusters. To capture shape and spatial alignment, respectively, we use the regions' HOG similarity and spatial overlap (IoU metric), and cluster with k-medoids. The r cluster centers (typically r=10) form the final set of proposals for each image. We found that this careful filtering was much more accurate than constraining the number of region proposals using the objectness scores directly. For example, on the MIT dataset our filtering step results in a mean average best score (MABO) of 72.2 with only 10 proposals. In contrast, simply retaining the top 10 proposals using scores from [8] results in a MABO of 64.95. The clustering step selects diverse proposals, leading to higher recall with fewer proposals.

Let $\mathcal{R} = \{R_{ij}\}$ denote the set of all region proposals in all N images, where R_{ij} denotes the j-th region for image I_i. Our joint segmentation approach, to be defined next, relies on both image- and region-level features. For each image I_i, we extract a global appearance descriptor denoted I_i^{c}. For each region R_{ij}, we extract two features: a saliency rating R_{ij}^{s}, and a region appearance descriptor R_{ij}^{c}.[2]

3.2. Semi-automatic joint foreground segmentation

We define a Markov Random Field (MRF) joint segmentation graph $\mathcal{G} = (\mathcal{R}, \mathcal{E})$ based on the filtered region proposals across all images in the collection. Each region $R_{ij} \in \mathcal{R}$ forms a node and the edges $\mathcal{E}$ connect pairs of regions. During segmentation, the edges will encourage consistent labels for similar regions, while the nodes will encourage

[2]One could choose from a variety of features; we employ off-the-shelf CNN-based descriptors and saliency metrics (see Sec. 4 for details).

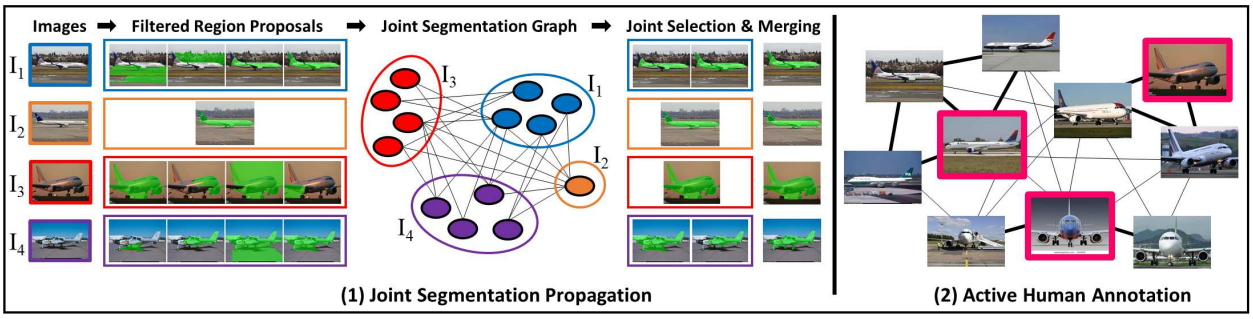

Figure 2: Approach: **(1) Joint segmentation propagation:** Given a set of images $\{I_1, I_2, I_3, I_4\}$ with I_2 already segmented by a human, the goal is to generate foreground segmentations for the remaining images. We first generate a set of filtered region proposals for each image. Next, a joint segmentation graph over these region proposals (edges = region similarity) is defined. An energy function defined over this graph is minimized to obtain a set of good proposals for each image, which are then fused to obtain the final segmentation. **(2) Active human annotation:** Our active selection method works over a joint graph defined over all images in the collection (**darker edges** = high similarity). These pairwise similarities allow us to identify influential images (most useful for others) and also help in enforcing diversity in selection (to avoid redundancy). We also account for uncertainty (not depicted here) by predicting the quality of the current segmentation. Example selections by our method are shown in pink. Best viewed in color.

foreground labels for salient regions that are consistent with well-segmented exemplars. We keep a sparse set of edges $\mathcal{E}$ by only connecting regions whose similarity exceeds a threshold τ. No edges connect regions in the same image.

Let $\mathcal{Y} = \{Y_{ij}\}$ be a set of binary region labels, where:

$$Y_{ij} = \begin{cases} 1 & \text{if proposal } R_{ij} \text{ is a good segmentation for } I_i \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Let $\mathcal{S} \subseteq \mathcal{I}$ denote the current subset of images labeled with foreground masks by human annotators. (We explain in Sec. 3.3 how the composition of this set is iteratively and actively defined.) Once an image I_s has been labeled, meaning it first appears in $\mathcal{S}$, we adjust the graph accordingly. First, we replace all nodes R_{sj} by the single mask region given by the human annotator, denoted $\bar{R}_s$, and we clamp its label $Y_s = 1$. Then, we modify the edge set $\mathcal{E}$ appropriately, such that in image s, only the mask $\bar{R}_s$ has edges to similar regions in unlabeled images.[3] These updates inject the human-labeled regions into the segmentation pipeline, allowing us to propagate the valuable information through the pairwise terms (defined below).

There are several ways to use the human-labeled masks to guide the joint segmentation. One could use them to train a foreground appearance model (e.g., as in iCoseg [6]). However, this is most effective only in the stricter co-segmentation setting where the same exact foreground object instance repeats across images. An alternative could be to directly transfer the segmentation from labeled images to unlabeled images, e.g., using dense matching [28, 47]. However, due to variations in scale and shape of foreground objects, global alignment is difficult in many cases.

Instead, our approach relies on strong matches discovered between foreground regions in human-labeled images and region proposals in unlabeled images. The intuition is that a good region proposal (i.e., one close to the actual foreground object segment) will strongly match a human-labeled ground truth region. On the contrary, a bad proposal will have weaker matches.

We define the following energy function $E(\mathcal{Y})$ for jointly segmenting the image collection $\mathcal{I}$:

$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) \; + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}). \quad (2)$$

The unary term is defined as

$$\Phi(Y_{ij}) = \begin{cases} Y_{ij} & \text{if } i \in \mathcal{S} \\ \alpha^{\text{s}} \Phi^{\text{s}}(Y_{ij}) \; + \alpha^{\text{m}} \Phi^{\text{m}}(Y_{ij}) & \text{if } i \in \mathcal{I} \backslash \mathcal{S}. \end{cases}$$

This unary prefers to label as foreground those regions that are (1) *salient* and/or (2) form a good *match* with some previously labeled foreground mask. The variables α^{s} and α^{m} weight the influence of the saliency and matching terms, respectively. The *saliency* term is defined using the saliency region feature (R^{s}_{ij}) as:

$$\Phi^{\text{s}}(Y_{ij}) = Y_{ij} R^{\text{s}}_{ij} + (1 - Y_{ij})(1 - R^{\text{s}}_{ij}), \quad (3)$$

so that we favor assigning $Y_{ij} = 1$ if R_{ij} is very salient.

The *match* component of the unary term encodes that a region proposal with a good ground truth region match is likely foreground. In particular, we identify matches for a region by considering its "local neighborhood" of images in the graph. For each unlabeled image I_i, we retrieve its p nearest neighbors from the labeled set $\mathcal{S}$ using the image-level features I^{c}_i. Denote that set $\mathcal{N}(I_i, \mathcal{S})$. Then, for each region proposal R_{ij}, we find the best matching ground truth foreground region among these p neighbors, and use this matching score in the unary term:

$$\Phi^{\text{m}}(Y_{ij}) = Y_{ij} R^{\text{m}}_{ij} + (1 - Y_{ij})(1 - R^{\text{m}}_{ij}), \text{where} \quad (4)$$

[3] For simplicity of notation, below we continue to use R_{ij} for all regions unless strictly required; it should be understood that $\forall I_i \in \mathcal{S}$ there is only one proposal, instead of r proposals.

$$R^{\mathrm{m}}_{ij} = \max_{p \in \mathcal{N}(\mathcal{I}_i, \mathcal{S})} sim(R^{\mathrm{c}}_{ij}, \bar{R}^{\mathrm{c}}_{p}), \quad (5)$$

and sim is the cosine similarity, and $\bar{R}_p$ denotes the p-th ground truth region.

The pairwise term in Eq (2) encourages similar-looking regions to take the same label:

$$\Psi(Y_{ij}, Y'_{ij}) = \delta(Y_{ij} \neq Y'_{ij})\ sim(R^{\mathrm{c}}_{ij}, R'^{\mathrm{c}}_{ij}). \quad (6)$$

This term enforces consistency in our joint selection of good region proposals, since we incur a penalty proportional to region similarity if the two regions receive different labels.

The minimum energy solution $\mathcal{Y}^* = \arg\min_{\mathcal{Y}} E(\mathcal{Y})$ yields a set of good region proposals for each image in the collection. Note that we do not constrain only one proposal to be selected per image. We purposely allow selecting *multiple* good regions per image, for two reasons. First, an image can naturally have multiple good region proposals (e.g. covering different object parts). As we will see next, our fusion step can take these multiple partial proposals to obtain a single accurate segmentation. Second, it allows us to efficiently and exactly minimize our energy function using graph-cuts [7]. We found this works much better in practice than approximate inference techniques. A complete round of propagation for $N = 1,400$ images takes just **1 minute** on a single CPU (excluding feature extraction). In contrast, the state-of-the-art propagation method of [36] would take 225 hours to propagate labels (excluding both feature extraction and SIFT-Flow).

To obtain the final segmentation mask M_i, we fuse the chosen good region proposals Y^*_i. We use the selected regions as a rough prior for the object's spatial extent, and then use that to build an image-specific foreground appearance model. Specifically, for each chosen proposal in I_i we retrieve the p nearest human-labeled masks. We transfer those masks into I_i (we use simple resizing and transfer, similar to [25, 22]), average the transferred masks of all proposals, and mean threshold the result to obtain a spatial prior. We then build a GMM over RGB color values for all pixels in the spatial prior. Finally, the combined appearance and spatial prior are used to define an image-specific MRF, which is minimized using graph cuts to obtain M_i.

In summary, our semi-supervised segmentation propagation algorithm is designed to be accurate (through careful filtering of regions and use of sparse actively chosen human annotations) and efficient (by avoiding expensive dense matching steps [35] and by using an efficient graph cuts energy minimization framework instead of costly approximate inference techniques as in [10]).

3.3. Active selection for propagation

We now describe our stagewise algorithm to actively select images for annotation. The active selection procedure takes as input the image collection $\mathcal{I}$, an annotation budget k specifying the number of images to get labeled per stage, and the number of total annotation stages T. In each stage t, we solicit annotations for the actively chosen batch $\mathcal{S}_t$, augment $\mathcal{S}$ with that newly labeled data ($\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_t$), and propagate the segmentation as described above. The output after T rounds is the resulting propagated masks $\mathcal{M}$ on all images. Note that throughout the stages, each unlabeled mask is continually refined, and its intermediate results affect subsequent stages' active selections.

Our active selection algorithm accounts for three criteria—influence, diversity, and uncertainty. The former two criteria account for relationships between images that are relevant to propagation, while the latter accounts for the inherent difficulty of individual images.

An image *influential* for propagation is similar to many other images in the collection. Intuitively, labeling such a "hub" image can directly improve the mask quality of the related images, particularly given our match-based unaries and localized image neighborhoods (Eq (5) and Eq (6)). We measure the influence of a candidate batch $\mathcal{S}_t$ as:

$$\textsc{Influence}(\mathcal{S}_t) = \frac{1}{|\mathcal{S}'_t|} \sum_{I_i \in \mathcal{S}_t} \sum_{I_j \in \mathcal{S}'_t} sim(I^{\mathrm{c}}_i, I^{\mathrm{c}}_j), \quad (7)$$

where $\mathcal{S}'_t$ denotes all unlabeled images not in the candidate batch $\mathcal{S}_t$ and sim is the cosine similarity.

A batch of images that are *diverse* ensures broad coverage over the entire collection. Selecting images which are influential but also very similar would not lead to a large information gain. Hence, we also add a penalty for selecting mutually similar images:

$$\textsc{Diversity}(\mathcal{S}_t) = -\frac{1}{|\mathcal{S}_t|} \sum_{I_i \in \mathcal{S}_t} \sum_{I_j \in \mathcal{S}_t} sim(I^{\mathrm{c}}_i, I^{\mathrm{c}}_j). \quad (8)$$

An image that is *uncertain*—inherently difficult to segment automatically—is also a good candidate for human supervision. We quantify the uncertainty of a batch as:

$$\textsc{Uncertainty}(\mathcal{S}_t) = \frac{1}{|\mathcal{S}_t|} \sum_{I_i \in \mathcal{S}_t} D(M_i), \quad (9)$$

where $D(\cdot)$ is a learned predictor of image difficulty. This prediction function is trained to infer when an image is badly segmented. Taking inspiration from prior work [33, 8, 15], we devise a set of descriptors suggestive of segmentation quality, and train a regression function using images for which we know each region's overlap with the true foreground. Given a region, the predictor returns its expected normalized overlap with the ground truth.

Specifically, we train a random forest regressor using 1,385 images from the MSRC [29], iCoseg [6], and IIS [13] datasets. The regression target is the overlap score with ground truth. To generate training samples, we sample CPMC [8] region proposals whose overlap falls in the

Algorithm 1 Active Selection Algorithm

procedure ACTIVESELECTION
  Input: $\mathcal{I}, \mathcal{I}^u = \mathcal{I}, \mathcal{I}^l = \phi$;
  Define: $\mathcal{F}(\mathcal{S}) = \text{INFLUENCE}(\mathcal{S}) + \text{DIVERSITY}(\mathcal{S}), \mathcal{S} \subseteq \mathcal{I}$;
  for each stage t = 1,2, ..., T **do**
    Candidate set: $\mathcal{I}_t^u = \phi$;
    for i = 1,2, ..., K **do**
      $s^* = \arg\max_{s \in \mathcal{I}^u \setminus \mathcal{I}_t^u} D(M_s^{t-1})$; $\mathcal{I}_t^u = \mathcal{I}_t^u \cup s^*$;
    end for
    $\mathcal{S}_t = \phi, \mathcal{S}_t^{'} = \mathcal{I}_t^u$;
    for i = 1,2, ..., k **do**
      $s^* = \arg\max_{s \in \mathcal{S}_t^{'}} \mathcal{F}(\mathcal{S}_t \cup s) - \mathcal{F}(\mathcal{S}_t)$;
      $\mathcal{S}_t = \mathcal{S}_t \cup s^*$; $\mathcal{S}_t^{'} = \mathcal{S}_t^{'} \setminus s^*$;
    end for
    $\mathcal{I}^l = \mathcal{I}^l \cup \mathcal{S}_t$; $\mathcal{I}^u = \mathcal{I}^u \setminus \mathcal{S}_t$;
  end for
end procedure

top and bottom 5% of all proposals. We use the following features as indicators of segmentation quality: (1) boundary alignment between the input region and superpixel boundaries, (2) the number of connected components, which reflects segment coherence, (3) color separability of the region from the background based on χ^2 distance on RGB histograms, and (4) region compactness, as measured by the ratios of the region's area to its tight bounding box and its convex hull. See Supp. for details.

We would like to identify the set maximizing all three criteria simultaneously. This is a combinatorial problem over all subsets $\mathcal{S}_t \subseteq \mathcal{I}$ and impractical to solve optimally. We instead employ a greedy approach to account for all factors. First, we extract the $K > k$ most uncertain unlabeled images, as judged using the predictor $D(M_i)$ applied to the current mask estimated at the end of the previous stage. From among that pool, we select the subset $\mathcal{S}_t$, accounting for both influence and diversity. Starting with an empty set, we iteratively add an image at a time until we reach the budget k. The selected image is the one giving the maximal marginal increase for $\text{INFLUENCE}(\mathcal{S}_t) + \text{DIVERSITY}(\mathcal{S}_t)$. See Algorithm 1 for complete pseudocode.

Our greedy algorithm is inspired by the maximization procedure typically used for monotone submodular functions, which offers theoretical gurantees [23]. Due to the diversity penalty, our objective is non-monotonic, hence known approximation guarantees do not apply; nonetheless, it works well in practice. It is also fast: for a pool of 1,400 unlabeled images, our active selection requires just seconds.

4. Results

Datasets: We evaluate on two datasets:

- **ImageNet:** We conduct a large-scale evaluation of our approach using ImageNet [37] (∼1M images, 3,624 classes). We follow the setup of [42], and consider all images with bounding box annotations available.[4]
- **MIT Object Discovery:** This challenging dataset consists of Airplanes, Cars and Horses [35]. Its intra-class appearance variation is much greater than that of older co-segmentation datasets (MSRC [29] or iCoseg [6]).

Baselines: Apart from an ablated version of our method (i.e., w/o uncertainty), we compare with these baselines:

- **Passive:** This is a simple passive baseline where at every stage, we randomly pick k images from the unlabeled set to be labeled by humans.
- **PageRank Selection [36]:** This is the only active propagation method in the literature, making it critical for comparison. It uses PageRank importance ranking and clustering to pick k good images at each stage.
- **Semantic Propagation [12]:** An existing propagation method that promotes propagation between semantically related classes. It seeds the propagation with labeled images from existing datasets.
- **State-of-the art weakly supervised methods:** We compare the special case of our method (only weak supervision) with several existing approaches [35, 9, 19, 20, 21, 42]. Other weakly supervised methods [30, 31, 32] for semantic segmentation consider multi-label data, and so are not directly comparable.

Evaluation metrics: We use: (1) **Jaccard Score:** Standard intersection-over-union (IoU) metric between predicted and ground truth segmentation masks (for MIT) and between bounding boxes (for ImageNet), and (2) **CorLoc Score:** Percentage of images correctly localized according to PASCAL criterion (i.e IoU > 0.5) used in [42, 10]. For MIT we use the segmentation masks (Seg-CorLoc) and for ImageNet we use bounding boxes (BBox-CorLoc) since it lacks ground truth masks.

Implementation details: We generate region proposals for MIT using CPMC [8] and for ImageNet using MCG [5] (due to efficiency). For global appearance I_i^{c}, we extract 4096-dim Convolutional Neural Network (CNN) features [24] using Caffe [17]. For saliency R_{ij}^{s}, we average the region's pixel-level saliency values from [18]. For region appearance R_{ij}^{c}, we extract a CNN feature for the region's tight bounding box. We set: $\tau = 0.7, p = 5, \alpha^{\text{s}} = \alpha^{\text{m}} = 0.5$, # rounds $T = 20, k = (\text{ \# images}/T), K = 4 * k$. All parameters were set after manual inspection of few images, then fixed for all experiments. In all experiments human annotation is simulated using ground truth data. Our run-time

[4] Since ImageNet lacks segmentation ground truth for all images, (1) we evaluate our masks against the bounding boxes, using a tight bounding box around the predicted segmentation and (2) when our method requests a human-drawn segmentation, it gets the region proposal with maximum overlap with the ground-truth bounding box.

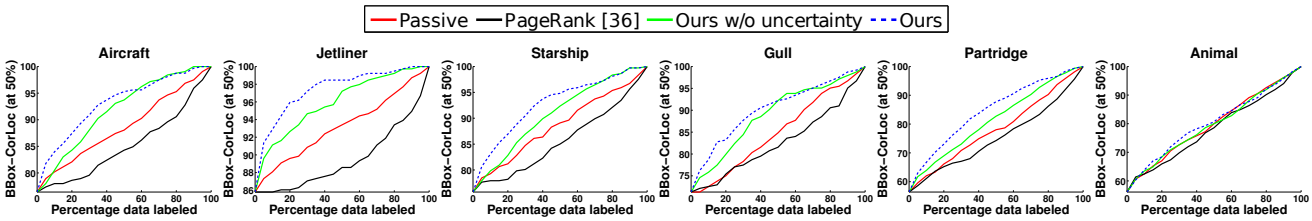

Figure 3: Active propagation for varying amounts of human annotation on a subset of the 3,624 ImageNet total synsets we tested (more in Supp.). Since only bounding box ground truth is available, we show bounding-box localization (BBox-CorLoc) accuracy (see Supp. for bounding-box Jaccard plots). Last plot (Animal) shows a failure case. Best viewed in color.

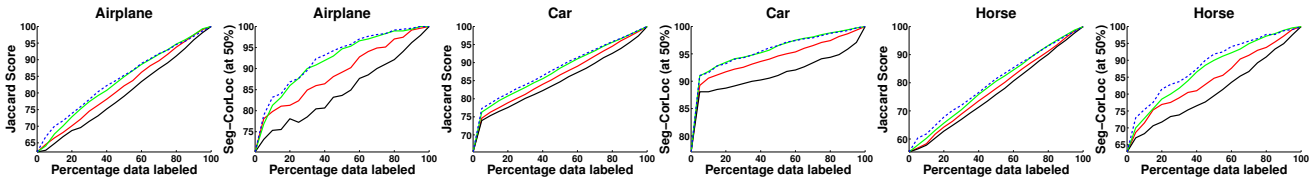

Figure 4: Active propagation results for varying amounts of human annotation for MIT Object Discovery dataset. We show both segmentation overlap (Jaccard) and segmentation localization (Seg-CorLoc) accuracy for each of the three classes. Best viewed in color.

Figure 5: Example active annotation choices for the 3 image collections (Airplane, Car, Horse) in the MIT dataset during the first stage with k = 10. The algorithm selects influential and diverse images (e.g., prototypical shapes) with some relatively difficult/unusual ones (best viewed in color).

is dominated by the cost of computing pairwise similarities between region proposals, $O((Nr)^2)$ for N images and r region proposals per image.

4.1. Active segmentation propagation

First we present results for active selection. In this setting we iteratively request annotators to provide true segmentations for a subset of images. We then use these labeled images to improve the joint segmentation of other unlabeled images in the collection.

Figure 5 shows qualitative examples of annotation choices made by our active selection algorithm. We find the impact of all the components quite visible in the choices. Several influential and diverse images which provide good coverage over the collection are chosen, along with some relatively difficult and unusual ones.

Figures 3 and 4 show the quantitative results. On the extreme left, we have the performance of the purely weakly supervised setting (no human input) and on the extreme right, annotators provide ground-truth segmentations for all images in the collection. In between we see the trade-off between actively allocating human effort versus other baselines. Since this is a transductive setting where the goal is to generate segmentations for all images, we plot average results over all the images in the collection (whether human or computer segmented). This scoring protocol has an additional advantage of averaging over the same number of images after each round of annotation, making trends on the x-axis easy to interpret.

For all metrics and datasets, the proposed approach outperforms all baselines. While all methods naturally improve with more labeled data, the slope of our improvement curve is substantially sharper using minimal human effort—sometimes dramatically so (e.g., Jetliner on ImageNet or Airplane on MIT). It is important to note that all methods are using identical CNN features and the same propagation algorithm, hence our gains exactly show the impact of making wiser annotation choices.

Surprisingly, we find that the Passive baseline outperforms the active PageRank method employed in [36]. We believe this is because PageRank emphasizes the influence property more, and, despite its clustering component, fails to select sufficiently diverse examples[5] (in [36] no comparison with a passive baseline is shown). On the other hand, our method takes into account influence, diversity, and uncertainty to choose good candidates for annotation. This leads to better annotation choices and in turn better propagation. We also see that omitting uncertainty from our approach decreases accuracy, showing the value of this segmentation-specific active selection component.

While all methods fare better on the "easier" task of localization (vs estimating pixel-perfect masks), our gains are actually substantially higher for localization (as measured by Seg-CorLoc and BBox-CorLoc). In addition, for both datasets, our gains are much higher for larger collections ($>$ 100 images). Larger collections exhibit both greater redun-

[5] Restricting our proposed method to use "influence" alone also performs worse than passive and comparable to [36].

dancy as well as several modes within the data. Our method successfully exploits these patterns while making annotation choices. For example, for MIT "Airplanes", we correctly localize 90% of the images with only 30% of the data labeled by annotators. In contrast, the Passive and active PageRank baselines require significantly more annotations (55% and 70%, respectively) to achieve the same accuracy.

Figure 3 also shows an interesting failure case for the ImageNet "Animal" class. Upon inspection, we found that it contains images from several different animal types with very little structural similarity; in this case, our active annotation method did not fare any better than the baselines.

We stress that, to our knowledge, [36] represents the only prior attempt to incorporate active selection with segmentation propagation. Before any inference, that method seeds a dense-flow graph with images chosen with a PageRank sampling. Our stage-wise method takes a very different strategy, iteratively self-inspecting its own estimates and redirecting human attention accordingly. As seen in Figure 3 & 4, our approach significantly outperforms the one-shot PageRank approach [36] in all experiments, and our propagation method is orders of magnitude faster (cf. Sec 3.2).

We also compare with the other state of the art segmentation propagation approach from Guillaumin et al. [12]. For this, we consider all images which are common between our experimental setup and that of [12]. This gives us a total of 99,020 images across 352 ImageNet classes. From the data provided by the authors, we found that ground-truth bounding boxes for 67,029 of those images were used to seed the propagation in [12]. For the same amount of labeled data our active segmentation propagation approach achieves a Jaccard score of 65% as opposed to 62.63% by [12]. More importantly, reducing the supervision budget for our method, we achieve the same accuracy as this (passive) state of the art propagation method [12] when using 26% *less* human-annotated data. This large savings in human effort shows the clear value of actively determining where human guidance is most needed.

4.2. Weakly supervised foreground segmentation

Next we test our method in a purely weakly supervised setting against several existing methods. In this special case, weak supervision (i.e., all images have an object from the same category) is the only information available. No additional human annotation is requested. This corresponds to setting $\mathcal{S} = \emptyset$, $\alpha^{\mathrm{s}} = 1$ and $\alpha^{\mathrm{m}} = 0$.

Table 1 compares our approach to several existing methods [35, 9, 19, 20, 21] on the MIT (subset from [35] and full) dataset. Our approach outperforms all existing methods in 4 out of 6 cases and has consistently good accuracy in all cases. This is really encouraging because our joint segmentation model is simpler and more efficient than existing methods (e.g [35] uses dense matching, [9] uses negative training data to train detectors). The key strengths of our propagation design lie in carefully selecting region proposals that have good coverage over the objects and are not redundant (without this we see a 8% drop in performance on average, see Supp. for details), combined with the region-based matching potentials. Jointly selecting good region proposals then helps in discovering similar pattern configurations over the entire collection. The method of [9] possibly benefits from stronger discriminative exemplar-appearance models for the Horse class in MIT (full).

Methods	MIT dataset (subset)			MIT dataset (full)		
	Airplane	Car	Horse	Airplane	Car	Horse
# Images	82	89	93	470	1208	810
Joulin et al. [19]	15.36	37.15	30.16	n/a	n/a	n/a
Joulin et al. [20]	11.72	35.15	29.53	n/a	n/a	n/a
Kim et al. [21]	7.9	0.04	6.43	n/a	n/a	n/a
Rubinstein et al. [35]	55.81	64.42	51.65	55.62	63.35	53.88
Chen et al. [9]	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

Table 1: Comparison with state-of-the-art methods on MIT dataset for weakly supervised joint foreground segmentation (Metric: Jaccard score).

ImageNet dataset	
# Classes	# Images
3,624	939,516

Methods	BBox-CorLoc
Top obj. box [3]	37.42
Tang et al. [42]	53.20
Ours	**57.64**

Table 2: Comparison with state-of-the-art methods on ImageNet for weakly supervised joint foreground segmentation (Metric: Avg. BBox-CorLoc).

Table 2 shows results on ImageNet. The "Top obj" baseline is the result of taking the top Objectness window [3], as reported in [42]. Our method outperforms the state of the art [42] by a considerable margin, which again highlights the strengths of our joint segmentation graph. With nearly 1 million images, a performance gain of 4.44% means that we correctly localize 41,715 more images than [42].

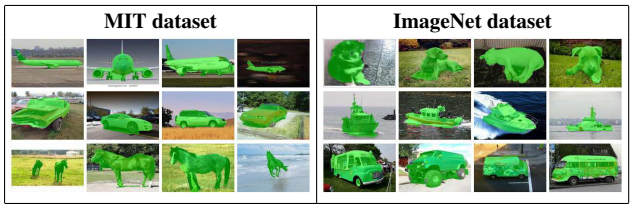

Figure 6: Qualitative results for weakly supervised joint segmentation. The segmentation result is shown with a green overlay over the image. See Supp. for more results (incl. failure cases). Best viewed in color.

Figure 6 shows qualitative results. Our method is able to segment objects well in spite of large intra-class variations. Because of the joint segmentation graph, our method can successfully segment some challenging instances where the object is not easily separable from the background but matches well with similar regions in easier images.

Conclusions We proposed a scalable approach to actively solicit foreground annotations useful to propagate segmentations in large image collections. Our results demonstrate its effectiveness: we improve the state of the art in multiple datasets for both weakly supervised segmentation and active propagation.

Acknowledgments: This research is supported in part by ONR YIP N00014-12-1-0754.

References

[1] E. Ahmed, S. Cohen, and B. Price. Semantic object selection. In *CVPR*, June 2014.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Classcut for unsupervised class segmentation. In *ECCV*, 2010.

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an Object? In *CVPR*, 2010.

[4] O. Aodha, N. Campbell, J. Kautz, and G. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, 2014.

[5] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[6] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. iCoseg: Interactive Co-segmentation with Intelligent Scribble Guidance. In *CVPR*, 2010.

[7] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(9):1124–1137, Sept. 2004.

[8] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(7):1312–1328, 2012.

[9] X. Chen, A. Shrivastava, and A. Gupta. Enriching Visual Knowledge Bases via Object Discovery and Segmentation. In *CVPR*, 2014.

[10] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100(3):275–293, September 2012.

[11] E. Elhamifar, G. Sapiro, A. Yan, and S. Sastry. A convex optimization framework for active learning. In *ICCV*, 2013.

[12] M. Guillaumin, D. Küttel, and V. Ferrari. ImageNet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014.

[13] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *CVPR*, 2010.

[14] S. Hoi, R. Jin, J. Zhu, and M. Lyu. Semi-supervised SVM Batch Mode Active Learning with Applications to Ima ge Retrieval. *ACM Transactions on Information Systems*, 1(1), 2009.

[15] S. Jain and K. Grauman. Predicting sufficent annotation strength for interactive foreground segmentation. In *ICCV*, 2013.

[16] S. Jain and K. Grauman. Which image pairs will cosegment well? predicting partners for cosegmentation. In *ACCV*, 2014.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[18] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang. Saliency detection via absorbing markov chain. In *ICCV*, 2013.

[19] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[20] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012.

[21] G. Kim, E. Xing, L. Fei Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011.

[22] J. Kim and K. Grauman. Shape sharing for object segmentation. In *ECCV*, 2012.

[23] A. Krause and D. Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2013.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*. 2012.

[25] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012.

[26] Y. J. Lee and K. Grauman. Collect-Cut: Segmentation with top-down cues discovered in multi-object images. In *CVPR*, 2010.

[27] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[28] C. Liu, J. Yuen, and A. Torralba. Sift flow: dense correspondence across different scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5), 2011.

[29] T. Malisiewicz and A. A. Efros. Spatial support for objects via multiple segmentations. In *BMVC*, 2007.

[30] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[31] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR Workshop*, 2015.

[32] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[33] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003.

[34] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[35] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[36] M. Rubinstein, C. Liu, and W. T. Freeman. Annotation propagation in large image databases via dense image correspondence. In *ECCV*, 2012.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[38] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. In *CVPR*, 2006.

[39] J. Serrat, A. Lopez, N. Paragios, and J. C. Rubio. Unsupervised co-segmentation through region matching. In *CVPR*, 2012.

[40] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation. In *ECCV*, 2006.

[41] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Cl ass Active Learning. In *CVPR*, 2010.

[42] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014.

[43] A. Vezhnevets, J. M. Buhmann, and V. Ferrari. Active learning for semantic segmentation with expected change. In *CVPR*, 2012.

[44] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011.

[45] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*, 2009.

[46] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 2010.

[47] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. In *ECCV*, 2010.