Figure-Ground Image Segmentation helps Weakly-Supervised Learning of Objects

Katerina Fragkiadaki and Jianbo Shi

GRASP Laboratory, University of Pennsylvania
3330 Walnut St., Philadelphia, PA-19104
katef@seas.upenn.edu, jshi@cis.upenn.edu

Abstract. Given a collection of images containing a common object, we seek to learn a model for the object without the use of bounding boxes or segmentation masks. In linguistics, a single document provides no information about location of the topics it contains. On the contrary, an image has a lot to tell us about where foreground and background topics lie. Extensive literature on modelling bottom-up saliency and pop-out aims at predicting eye fixations and allocation of visual attention in a single image, prior to any recognition of content. Most salient image parts are likely to capture image foreground. We propose a novel probabilistic model, *shape and figure-ground aware model* (sFGmodel) that exploits bottom-up image saliency to compute an informative prior on segment topic assignments. Our model exploits both figure-ground organization in each image separately, as well as feature re-occurrence across the image collection. Since we use image dependent topic prior, during model learning we optimize a *conditional* likelihood of the image collection given the image bottom-up saliency information. Our discriminative framework can tolerate larger intraclass variability of objects with fewer training data. We iterate between bottom-up figure-ground image organization and model parameter learning by accumulating image statistics from the entire image collection. The model learned influences later image figure-ground labelling. We present results of our approach on diverse datasets showing great improvement over generative probabilistic models that do not exploit image saliency, indicating the suitability of our model for weakly-supervised visual organization.

1 Introduction

Given a collection of images containing a common object, we seek to learn a model for detection and segmentation of the object without additional supervision. The absence of figure-ground segmentation ahead of time makes this task challenging. However, learning of object models with minimum amount of supervision is necessary for scaling vision systems to large number of object categories.

Models for unsupervised learning rely on the figure consistency principle: *foreground features tend to re-occur and co-occur more consistently across images than background features.* This permits their separation from the background and incorporation into the model built for the common object. However, the task remains challenging mainly for the following reasons:

1. *Heavy clutter.* The more cluttered the images, the harder to dig out the common object.

2. *Persistent co-occurrence of foreground with its semantically related background.* Examples are car and road, giraffe and grass, swan and water. So, in practice, the backgrounds in the image collection are not random. Rather they are highly correlated with the common figure, making it difficult for a generative process to segment it from the background.
3. *Large intraclass variation* of many object categories due to articulation, deformation, change of view point. This violates the figure consistency principle.

We propose a novel approach that deals with the above challenges by coupling figure-ground image segmentation and learning of the common object. To our knowledge, this is the first work that exploits image saliency and figure-ground organization for weakly-supervised learning of objects.

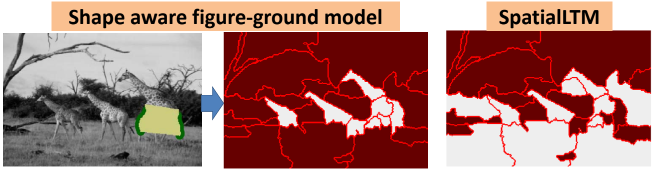

Fig. 1. The baseline model ([1]) does not discriminate between Giraffe and background due to persistent co-occurrence of Giraffe and ground in the image collection and wide variation of Giraffe shape. Wide intraclass variability is a common phenomenon in the visual domain. Our model exploits figure-ground information and effectively learns to segment the object.

We set our problem as topic discovery in the image collection: we aim at assigning image segments to visually coherent topics and learn the models for the common object (single foreground topic) and its background (possibly multiple background topics). We employ an iterative algorithm. Initially, we extract purely bottom-up figure-ground cues from each image, represented as *multiple soft figure-ground maps*. We score these maps using bottom-up image saliency. The map scores are not fixed, they change according to feature re-occurrence: figure-ground maps that propose foreground found most consistent across the image collection will iteratively get higher scores. At each iteration, we sample the highest scoring map in each image and obtain a prior on segment figure-ground labels. We perform a constrained probabilistic segment topic assignment by assigning different topics to segments that have different figure-ground labels. We accumulate image statistics and update the model parameters accordingly. Model update influences the scores of figure-ground maps and thus the figure-ground segment labels. Thus, figure-ground segmentation changes according to the model being built.

Our model has the following advantages:

- The object is naturally repulsed by its background and frequent co-occurrence of object and its semantically related background is no longer a problem. Optimizing segment topics *given* image saliency cues gives a discriminative flavor and offers robustness towards purely generative models.
- *Segment independence is not part of our assumptions*. Bottom-up saliency and figure-ground organization are operations that involve competition among segments

in each image and thus segment independence does not hold (see also fig. 2). This models the visual domain more accurately than most of the probabilistic models in previous work.
- We are not restricted to a fixed figure-ground segmentation. Our input is a set of soft figure-ground maps and is part of the learning process to choose the best one. The *loop* from feature re-occurrence back to bottom-up image saliency cues deals effectively with the presence of multiple foreground objects in each image.

The paper is organized as follows: We discuss related work in section 2. We present our model in section 3. In section 3.2 we present our representation for figure-ground image organization. Learning and inference in our model are presented in sections 4 and 5. Experimental results are in section 6. We conclude in section 7.

2 Related Work

There is extensive previous work on unsupervised or weakly-supervised learning of object categories:

Topic models([2], [3]) Topic models from statistical text analysis (LDA [4], p-LSA [5]) use unordered “bag of words“ representation of documents to automatically discover topics in large text corpora. In the visual domain, usually an image corresponds to a document and a local patch descriptor to a visual 'word'. Much of previous work is devoted in imposing spatial coherence between the visual words. Authors of [1] propose uniformity of topic assignments to the words belonging to the same superpixel. Work of [6] uses multiple segmentations of images and models each segment as a document. Segments well corresponding to topics are expected to have more peaked topic distributions than wrong (leaking) segments. In [7] a fixed outline of the object is used as extra input to guide learning.

Discriminative models Part of previous work ([8], [9]) takes a discriminative approach having a negative collection of images (not containing the common object) as additional input for detection of the common object. Works of [10], [11] and [12] model weakly-supervised learning as multiple instance learning (MIL), using MILboosting for object or part detection. Recently, authors of [13] used discriminative clustering to assign figure-ground labels to image segments in the image collection such that figure and ground classes are best separated. However, their formulation, does not take into account image saliency of foreground.

In [14] a hierarchical model representation is built from a few training examples. Plausible feature groupings are discovered iteratively based on the principles of suspicious coincidence and competitive exclusion. Authors of [15] attempt to segment a pair of images containing a common object. The problem is formulated as an MRF with a global constraint about appearance histogram matching of the corresponding parts from the two images. Authors of [16] learn a generative model for segmentation of a collection of images combining appearence with object shape and pose.

Our model exploits an informative topic prior based on image figure-ground cues and maximizes a conditional likelihood of the image collection given that prior. In this way, it is more suitable for the problem of weakly-supervised learning than pure generative models. We can tolerate larger amount of intraclass variability with smaller amount of training data. Shape is not provided as input but is recovered along the way. Discrimination is built within the image, by trying to discriminate the common object from its background. We anticipate that figure-ground information would be useful in learning most of the representations that appear in previous work, especially for objects with large intra-class variability.

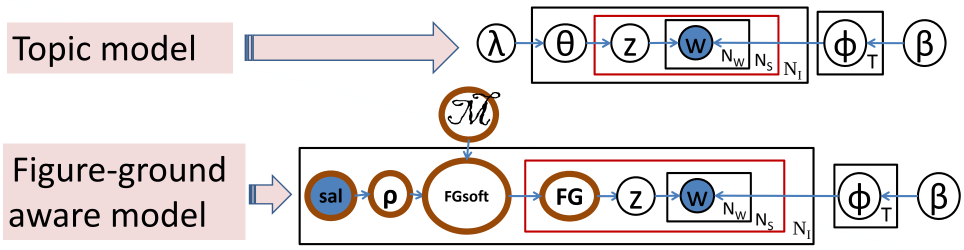

Fig. 2. Shading indicates observed variables and no shading indicates latent variables. N_i denotes number of images ($|\mathcal{I}|$), N_S number of segments and N_w number of words. *Top*: Topic model from [1]. θ^i is a multinomial distribution over topics for image $I^i$and λ is the parameter of a uniform Dirichlet prior over distributions $\theta^i, \;\; i = 1 \cdots |\mathcal{I}|$. *Bottom*: Shape and figure-ground aware model. The topic prior tightly depends on image figure-ground cues, as expressed by variable sal. Given the observed w, information flows from feature re-occurrence as expressed by ϕ back to the scores of figure-ground cues ρ, realizing the *feedback loop* from similarity across images to image figure-ground labelling. See text for the rest of notation.

3 Shape and figure-ground aware model

Adopting the terminology of topic models we claim that in images *topics are not created equal*. Extensive literature on bottom-up image saliency tells us that topics do not have uniform prior distribution given an image: Foreground topics tend to occupy salient image locations, while background topics less salient ones. Our model proposes a topic prior tightly depending on bottom-up image figure-ground cues.

Let $T = \{t_1, t_2, \cdots t_{|T|}\}$ be the topics in which to organize the image collection, t_1 denotes the single foreground topic and $t_2 \cdots t_{|T|}$ the background topics. Let s^i_k be the kth segment of image I^i and $\mathcal{S}^i$ be the set of segments of that image: $\mathcal{S}^i = \{s^i_1, s^i_2, \cdots s^i_{|\mathcal{S}^i|}\}$. Let z^i_k denote the topic of s^i_k. Let W be the word codebook and w^i_{kl} be the lth word of s^i_k (see also section 3.1). Let ϕ^z be a multinomial distribution over words given topic z and β be the parameter of a uniform Dirichlet prior over $\phi^z, \;\; z = t_1 \cdots t_{|T|}$.

Let FG^i_k to be the figure-ground label of segment $s^i_k \in \mathcal{S}^i$ in image I^i:

$$\mathrm{FG}^i_k \in \{0, 1\}$$

$\mathrm{FG}^i_k = 1$ if the segment s^i_k belongs to the common object in $\mathcal{I}$. Note that each image may have multiple foreground objects. $\mathrm{FG} = 1$ refers to the presence of the common object (common figure) that is of interest to us. We abuse language and call it figure-ground label for brevity.

Let FGmap_j^i be the jth soft figure-ground map as found by bottom-up figure-ground image organization and let $\mathcal{R}^i$ be the set of these maps in image I^i (see section 3.2):

$$\mathrm{FGmap}_j^i: \;\; \mathcal{S}^i \longrightarrow [0,1], \quad \mathcal{R}^i = \{\mathrm{FGmap}_j^i, \;\; j = 1 \cdots |\mathcal{R}^i|\} \tag{1}$$

FGmap_{jk}^i represents the probability of segment s_k^i to be part of foreground *given* FGmap_j^i. According to our formulation, segments have different probabilities of foreground given different maps.

We define sal_j^i to be the saliency score of map FGmap_j^i as computed by image saliency scoring (see section 3.2).

$$\mathrm{sal}_j^i \in \;[0,1] \quad , \qquad \textstyle\sum_{j=1}^{\mathcal{R}^i} \mathrm{sal}_j^i = 1$$

Let $\mathbf{sal}^{\boldsymbol{i}}$ be the saliency values of maps in image I^i.

We define ρ_j^i to be the trust score of map FGmap_j^i:

$$\rho_j^i \in \;[0,1] \quad , \qquad \textstyle\sum_{j=1}^{\mathcal{R}^i} \rho_j^i = 1$$

In contrast to the saliency score sal_j^i, the trust score of a map depends on both bottom-up saliency of each image in isolation, as well as feature re-occurrence across images. It realizes the *feedback loop* from feature re-occurrence in $\mathcal{I}$ back to image figure-ground segmentation. Intuitively, the trust score of a map is high if it maps the segments occupied by the common object to high foreground probabilities and the rest of the segments to low foreground probabilities. Let $\boldsymbol{\rho^i}$ be trust scores of maps in image I^i.Let FGsoft^i to be the map with the highest trust score in image I^i:

$$\mathrm{FGsoft}^i = \mathrm{FGmap}_\ell^i, \quad \text{where} \quad \ell = \arg\max_{j=1 \cdots |\mathcal{R}^i|} \rho_j^i \tag{2}$$

Let $\mathcal{M}$ denote the shape of the common object represented by a mixture of K Gaussian distributions over vectors of real values representing shape descriptors attached to binary shape masks of the object. Let L be the dimension of our shape descriptor:

$$\mathcal{M} = \{\omega_l, \mu_l, v_l, \;\; 0 < \omega_l < 1, \;\; \textstyle\sum_{l=1}^{K} \omega_l = 1, \;\; \mu_l \in \mathbb{R}^L, v_l \in \mathbb{R}, \;\; l = 1 \cdots K\},$$

Naturally, the probability of a shape descriptor $\mathrm{sc} \in \mathbb{R}^L$ given shape model $\mathcal{M}$ is:

$$P(\mathrm{sc}|\mathcal{M}) = \sum_{l=1}^{K} \omega_l \cdot \exp(-\frac{||\mu_l - \mathrm{sc}||^2}{2v_l^2}) \tag{3}$$

Our input is a set of soft figure-ground maps along with a distribution of saliency scores over them. During learning we alter the initial score distribution taking into account feature re-occurrence across images. Intuitively, we assign topics to segments such that segments found to have different figure-ground labels by maps of high saliency value are mapped to different topics and segments belonging to the same topic are most similar.

Our model parameters are $\mathcal{M}$ and ϕ, w and sal are observed variables and FG, z, ρ and FGsoft are latent variables. Learning of our model amounts to optimizing the following conditional likelihood of the image collection:

$$\max_{\mathcal{M},\boldsymbol{\phi}} P(\mathbf{FG}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\rho}, \mathbf{FGsoft}|\mathbf{sal}, \beta) \tag{4}$$

$$= \max_{\mathcal{M},\boldsymbol{\phi}} \prod_{i=1}^{|\mathcal{I}|} P(\boldsymbol{\rho^i}|\mathbf{sal}^i, \mathcal{M}) \cdot P(\mathrm{FGsoft}^i|\boldsymbol{\rho^i}) \cdot \prod_{k=1}^{|S^i|} P(z_k^i|\mathrm{FG}_k^i) \cdot P(\mathrm{FG}_k^i|\mathrm{FGsoft}^i) \cdot \prod_{l=1}^{|W_k^i|} P(w_{kl}^i|z_k^i, \phi^{z_k^i})$$

We optimize a conditional likelihood of topic assignments given bottom-up saliency information of the images in $\mathcal{I}$. We call $\mathbf{sal}^{\boldsymbol{i}}$, $i = 1 \cdots |\mathcal{I}|$ a prior, since saliency values are computed from each image in isolation, without taking into account the image collection $\mathcal{I}$ and re-occurrence of features, that is without seeing all the data.

A byproduct of our model is the organization of backgrounds into visually coherent groups. The performance of our model in learning the common object is not sensitive to the total number of topics used, a single background topic would do. However, by increasing the number of topics, we additionaly get meaningful models for background clusters as in the topic model literature.

Our model exploits effectively the rich figure-ground information present in images to guide the topic discovery process in our weakly-supervised framework.

3.1 Image representation

We use image segments as our basic units. Each image is described by a set of overlapping segments, obtained from multiscale segmentation. We used the multiscale normalized cut code [17] and discretized the eigenvectors using different number of segments. Within each image segment we find a number of interest points using the scale invariant saliency detector [18]. Each interest point is described by a SIFT descriptor[19]. We discretized the space of SIFT descriptors using unsupervised k-means clustering. Each segment is further described by a texture word and color word, each resulting from quantization of texton and color histograms using k-means. For ease of presentation we will refer to the description of each image segment by a bag of visual words, without discriminating among SIFT or color/texture words.

3.2 Figure-ground image organization

Figure-ground labelling is a step of perceptual organization which assigns a contour to one of the two abutting regions. There is experimental evidence that rich figure-ground information is available in images much before any of their content is recognized ([20], [21]). We assume saliency and figure-ground organization are related, that is *most salient image parts tend to belong to foreground (common object) while less salient ones to background* ([22]).

We note two important properties of figure-ground organization and image saliency that violate image segment independence: 1) Competition among different image parts for visual attention allocation (illustrated in the literature through normalization of saliency scores across image locations). 2) Convexity, connectedness of foreground, center-surround figure-ground competition. We take the above into account and choose to represent bottom-up figure-ground information with a score distribution over *multiple* segment foreground probability maps (soft figure-ground maps) (see eq. 1). This suits well our unsupervised learning framework: each one of these maps proposes image figure-ground labelling and learning choses the correct one by altering their score distribution.

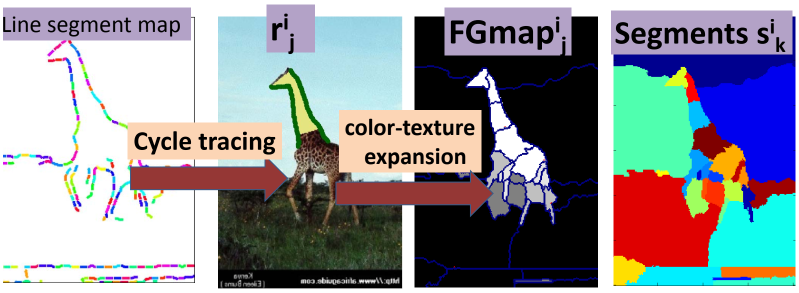

Fig. 3. *Ribbon extraction using cycle tracing*. a) Fitting straight line segments to a contour map. b) Extracting ribbons using cycle tracing, by piecing pairs of line segments in a graph partitioning framework. Yellow indicates ribbon interior. c)Figure-ground map (FGmap) obtained from color and texture expansion of the ribbon. White indicates high probability of foreground and black low probability. d) Segmentation map. Ribbons prevent over-fragmentation and achieve *scale invariance*. On the contrary, in segmentation we get very different image groupings for different numbers of segments. Here for illustration purposes we show segmentation of the finest scale (superpixels) although we used multiple segmentation scales.

Multiple soft figure-ground maps

Figure-ground organization is a mid-level process and mid-level grouping is required to provide information about figure-ground labelling. Our approach involves the following steps:

- We piece together over-fragmented segmentation boundaries to recover large (possible overlapping) foreground image structures. This can be done using multiple segmentations or greedy segment extension based on continuity of segment boundaries. We call these structures *ribbons* to distinguish them from segments and indicate that they can be obtained from different (not necessarily segment based) computational procedures. Later we present a *globally optimal* way for piecing segment boundaries for ribbon extraction using contour continuity.
- For each ribbon a segment foreground probability map is calculated: The interior of the ribbon is sent to foreground and surrounding highly contrasting segments to background. This is extended to a full segment foreground probability map by classifying each of the remaining image segments as foreground or background using color and texture features. Let r^i_j denote the jth ribbon of image I^i and $|\mathcal{R}^i|$ the number of ribbons in image I^i. The corresponding foreground probability map FGmap^i_j(see eq. 1) represents the probability of each segment s^i_k to be part of foreground *given* ribbon r^i_j. For each map FGmap^i_j, $j = 1 \cdots |\mathcal{R}^i|$ we define the following sets of segments:

$$S^i_{in_j} = \{s \in \mathcal{S}^i \ \ s.t \ \text{FGmap}^i_j(s) > l_1\}, \quad S^i_{out_j} = \{s \in \mathcal{S}^i \ \ s.t \ \text{FGmap}^i_j(s) < l_2\}$$
$$S^i_{dont\text{-}know_j} = \{s \in \mathcal{S}^i \ \ s.t. \ s \notin S^i_{in_j}, s \notin S^i_{out_j}\}$$

where $\mathcal{S}^i$ is the set of segments of image I^i and $l_2 < 0.5 < l_1$ (we chose $l_1 = 0.6$ and $l_2 = 0.4$). So, naturally, each FGmap^i_j constraints the figure-ground labelling of the segments of image I^i, sending $S^i_{in_j}$ to the foreground and $S^i_{out_j}$ to the background.

We define a shape mask $mask^i_j$: $mask^i_j(p) = \begin{cases} 1 & \text{if } \exists\, s \in S_{in^i_j} \text{ covering } p \\ 0 & \text{otherwise} \end{cases}$

describing the foreground FGmap^i_j selects.

To each mask^i_j we attach a grid shape feature sc^i_j of dimensions 6×6 and with 6 angular bins in each spatial cell to describe its shape.

- Maps are scored using saliency cues (see eq. 5) and scores are normalized to create a dictribution.We used $100 - 150$ figure-ground maps (FGmap) per image.

Cycle tracing for ribbon extraction We present here a novel approach for ribbon extraction which we used along with the multiple segmentation approach: We piece together over-fragmented segmentation boundaries in a *globally optimal way* based on good continuity of the boundary contour, generalizing the tool for cycle tracing for contour extraction of [23]. More precisely, we threshold the output of Probability of boundary detector [24] and fit line segments in a greedy way. We build a graph **W** whose nodes correspond to *pairs of roughly parallel line segments* and edge weights e_{ij} reflect the bending energy of the side contours from pair i into pair j. We have high affinity between two pairs of line segments when the one naturally extends into the other. We discretized the complex eigenvectors of the Laplacian of **W** corresponding to complex eigenvalues with large norm. For discretization we used the shortest path algorithm to recover the cycle enclosing the largest area in the embedding space. For further details refer to [23]. Ribbons obtained this way provide scale (distance between the two parallel contours) and orientation (orientation of the symmetry axis) helping alignment and recognition of shape.

Image saliency scoring

Saliency is the property of some parts of the image popping out and being well separated from their surrounding. Image saliency has been extensively studied in the literature ([25], [26],[27],[28]) and is related to properties such as local contrast, global exception in the image, centrality of location.

We score our figure-ground maps using image saliency cues. In each image I^i we define the saliency value sal^i_j of each FGmap^i_j:

$$\text{sal}^i_j = \frac{1}{Z} \cdot \text{FGcontrast}(\text{FGmap}^i_j) \cdot \text{Uniqueness}(\text{FGmap}^i_j) \tag{5}$$

- $\text{FGContrast}(\text{FGmap}^i_j)$ measures feature dissimilarity between the figure and ground that FGmap^i_j defines: $\text{FGcontrast}(\text{FGmap}^i) = \frac{1}{Z}\text{D}_{\text{KL}}\{\text{f}(p, p \in S^i_{in_j})||\text{f}(p, p \in S^i_{out_j})\}$

where $\mathrm{D_{KL}}$ denotes KL-divergence and f(pixel-set) denotes feature distribution with support in pixel-set. We used textons and quantized RGB intensity values as our features.

- $\mathrm{Uniqueness(FGmap}_j^i)$ measures dissimilarity between features of the figure of FGmap_j^i and the rest of the image I^i: $\mathrm{Uniqueness(FGmap}_j^i) = \frac{1}{Z} \cdot \frac{\sum_{p \in S^i_{in_j}} \sum_{l \in I^i} d_{pl} \cdot \mathrm{D_{KL}}(\mathrm{f}_p^{s(p)}, \mathrm{f}_l^i)}{|p \in S^i_{in_j}|}$ where $s(p)$ is the superpixel containing pixel p and f_l^i is the feature distribution of superpixel containing pixel l. We take into account the distances d_{pl} of pixels: high similarity found in large distances is worse that high similarity in small distances since it indicates concavity, a property of background.

In summary, in each image I^i our figure-ground representation is a set of segment foreground probability maps $\mathrm{FGmap}_j^i, \quad j = 1 \cdots |\mathcal{R}^i|, \quad i = 1 \cdots |\mathcal{I}|$, with a distribution $\mathbf{sal}^i$ of saliency scores over them.

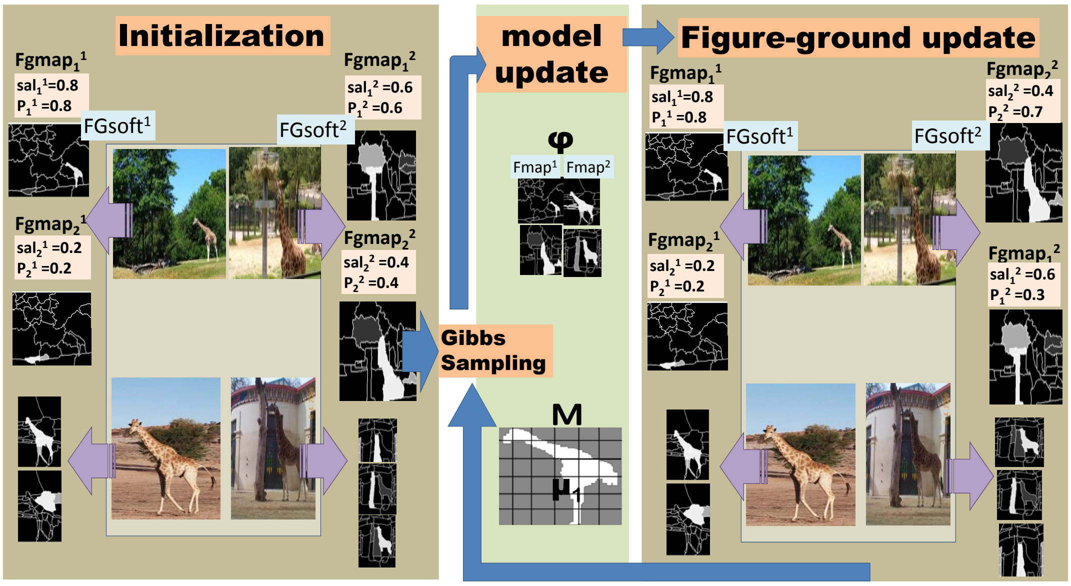

Fig. 4. *Learning a shape and figure-ground aware model.* White indicates high probability of foreground and black low probability. For each image we show the corresponding figure-ground maps ordered by their ρ scores. Notice the changing of ρ scores of the maps of the left pair of images. The presence of multiple foreground objects is not a problem in our model. A framework with fixed saliency scores would not be flexible enough to deal with multiple foreground objects present in images. For illustration purposes we use $K = 1$ for the shape model $\mathcal{M}$.

4 Learning

We use a type of EM procedure to estimate the parameters of our model. We use Gibbs sampling to get the expected conditional likelihood of latent given the observed variables at E step. $iter_{out}$ denotes the iteration counter for the EM algorithm and $iter_{in}$

the iteration counter for Gibbs sampling. We initialize the model parameters $\mathcal{M}$ and ϕ to the uniform distributions over the corresponding domains, that is: $\phi_w^{t_i} = \frac{1}{|W|},\ \ w \in W,\ \ i = 1 \cdots |T|$ and $v_l = \infty,\ \ l = 1 \cdots K$.

E step: From model parameters to figure-ground constraints

Sampling of figure-ground trust scores ρ
`Initial iteration` $(iter_{out} = 1)$: Initially, since ϕ and $\mathcal{M}$ are non informative (uniform) we have: $\rho_{\mathbf{j}}^{\mathbf{i}} = \mathrm{sal}_j^i,\ \ j = 1 \cdots \mathcal{R}^i,\ \ i = 1 \cdots |\mathcal{I}|, \quad iter_{out} = 1.$
`Later iterations` $(iter_{out} > 1)$: Given the bag of words representation $(\phi^z,\ \ z = t_1 \cdots t_{|T|})$ we compute for each image I^i a pixel foreground probability map Fmap^i. We assign to each image pixel p the mean foreground probability of the segments containing it:

$$\mathrm{Fmap}(p)^i = \frac{\sum_{k=1}^{|S_p|} P(z_k^i = t_1|\phi)}{|S_p|} = \frac{\sum_{k=1}^{|S_p|} \prod_{l=1}^{|W_k^i|} \phi_{w_{kl}^i}^{t_1}}{|S_p|} \tag{6}$$

where $\phi_{w_{kl}^i}^{t_1}$ is the probability of word w_{kl}^i given topic t_1, p is a pixel of image I^i and S^p is the set of segments containing it. We update the scores ρ_j^i of all the figure-ground maps FGmap_j^i in the image collection:

$$\rho_{\mathbf{j}}^{\mathbf{i}_{new}} = \frac{1}{Z} \cdot \mathrm{sal}_j^i \cdot \frac{1_{\mathrm{Fmap}^i} \cap 1_{\mathrm{mask}_j^i}}{1_{\mathrm{Fmap}^i} \cup 1_{\mathrm{mask}_j^i}} \cdot P(\mathrm{sc}_j^i|\mathcal{M}) \tag{7}$$

$j = 1 \cdots \mathcal{R}^i, \quad i = 1 \cdots |\mathcal{I}|, \quad iter_{out} > 1$
where $1_{\mathrm{Fmap}^i} = \{p,\ \ \mathrm{Fmap}^i(p) > \frac{1}{2}\}$, $1_{\mathrm{mask}_j^i} = \{p,\ \ \mathrm{mask}_j^i(p) = 1\}$ and $P(\mathrm{sc}_j^i|\mathcal{M})$ is given by equation 3.
Intuitively, figure-ground maps with high bottom-up saliency values that propose foreground agreeing with $\mathcal{M}$ and the corresponding Fmap get higher trust scores.

Determining FGsoft*:* For each image I^i we keep the highest scoring figure-ground map applying a *winner take all* strategy. Different maps may be competing with each other so averaging (marginalizing) them would not be meaningful. See equation 2.

Sampling segment figure-ground labels FG*:* $P(FG_k^i = 1|\mathrm{FGsoft}^i) = \mathrm{FGsoft}_k^i, \quad k = 1 \cdots |\mathcal{S}^i|,\ \ i = 1 \cdots |\mathcal{I}|$
Sampling segment topics z: Denote by W the word vocabulary, by $W(s_k^i)$ the words of segment s_k^i, by $n_{t_l}^w$ the number of assignments of word w to topic t_l, by n_{t_l} the total number of word assignments to topic t_l and by $n_{-s_k^i}$ the count of word assignments excluding words belonging to the segment s_k^i.
We have: $z_{s_k^i} = \begin{cases} t_1 & \text{if } \mathrm{FG}_k^i = 1 \\ \sim P'(z|\mathbf{z}_{-\mathbf{s}_\mathbf{k}^\mathbf{i}}, \mathbf{w}) & \text{if } \mathrm{FG}_k^i = 0 \end{cases}$
with :

$$P(z_{s_k^i} = t_l|\mathbf{z}_{-\mathbf{s}_\mathbf{k}^\mathbf{i}}, \mathbf{w}) \propto \prod_{w \in W(s_k^i)} \left(\frac{n_{-s_k^i, t_l}^w + \beta}{n_{-s_k^i, t_l}^{(\cdot)} + |W| \cdot \beta}\right) \tag{8}$$

where $\sim$ denotes sample from distribution and $P^{'}(z|\mathbf{z_{-i}},\mathbf{w})$ is the distribution over background topics: we exclude topic t_1 from T, compute $P(z|\mathbf{z_{-i}},\mathbf{w})$ for $z = t_2 \cdots |T|$ using equation 8 and normalize. We perform 500 iterations of figure-ground segment labels updates and segment topic assignments over all segments of $\mathcal{I}$ in random order.

M step : From figure-ground constraints to model parameters

Updating multinomial distributions of words given topics $\phi^z, \;\; z = t_1 \cdots t_{|T|}$*:* We update $\phi^z, \;\; z = t_1 \cdots t_{|T|}$ be counting word assignments to topics during all the iterations of Gibbs sampling of the previous E step.

Updating shape model $\mathcal{M}$*:* Let $\mathbf{FGsoft}$ be the set of highest scoring figure-ground maps during the previous E step: $\mathbf{FG_{soft}} = \{\mathrm{FGsoft}^i, \;\; i = 1 \cdots |\mathcal{I}|\}$. We compute all pair shape affinities between the corresponding shape features $\mathrm{sc}_k, \;\; k = 1 \cdots |\mathcal{I}|$, obtaining affinity matrix $\mathbf{A}$: $\mathbf{A}_{kl} = \exp(-\frac{||\mathrm{sc}_k - \mathrm{sc}_l||^2}{2d^2}), \;\; k, \; l \; = 1 \cdots |\mathcal{I}|$ Since we do not expect all shape masks to be correct, we aim at extracting compact clusters in this shape feature set. We zero out pairwise affinities with values below a threshold as indicating disagreement in shape. In the remaining shape affinity set, we order our features based on the number of neighbors. Large number of neighbors indicates high probability of exhibiting the common shape. Let $sc_k^{\mathrm{best}}, \;\; k \cdots K$ denote the K shape context features with the highest number of neighbors and n_k^{best} denote the corresponding number of neighbors. Then:

$$\mathcal{M} = \{\mu_l = \mathrm{sc}_l^{\mathrm{best}}, v_l = d, \;\; \omega_l = \tfrac{1}{Z} \cdot n_l^{\mathrm{best}}, \;\; l = 1 \cdots K\} \quad , \quad Z = \textstyle\sum_{l=1}^{K} n_l^{\mathrm{best}}$$

That is, the weights and centers of the mixtures are updated, while the variances are kept fixed and equal to constant d (same for all datasets used).

5 Inference

We used two different kinds of inference to score the performance of our model at the end of training and at test time:

- Inference using shape and figure-ground aware model (**sFGmodel**). We compute the segmentation labelling for image I^i: $\mathrm{label}^i(p) = \frac{(\mathrm{Fmap}^i(p) + \mathrm{FGmap}^i_{\mathrm{best}}(p))}{2}, \;\; p \in I^i$ where: $\mathrm{best} = \arg\max_{j=1 \cdots |\mathcal{R}^i|} \; \mathrm{sal}^i_j \cdot P(\mathrm{sc}^i_j|\mathcal{M})$, $\mathrm{FGmap}^i_{\mathrm{best}}(p) = \frac{\sum_{k=1}^{|\mathcal{S}_p|} \mathrm{FGmap}^i_{\mathrm{best}k}}{|\mathcal{S}_p|}$ where $\mathcal{S}_p$ the set of segments containing pixel p. We threshold label^i to get binary pixel labels.
- Inference using only the bag of words representation learnt from the shape and figure-ground aware model(**bagFGmodel**). We compute the segmentation labelling for image I^i: $\mathrm{label}^i(p) = \mathrm{Fmap}^i(p), \;\; p \in I^i$. We threshold label^i to get binary pixel labels. In **bagFGmodel** figure-ground and shape information are used for learning but only the bag of words representation $\phi^z, \;\; z = t_1 \cdots t_{|T|}$ is used to infer image labelling when scoring performance of our model.

6 Experiments

We use various datasets with different levels of difficulty to test our algorithm: Caltech 101:1) 81 images of Airplanes; MSRC: 2) 70 images of Cars, 3) 84 images of Cows;

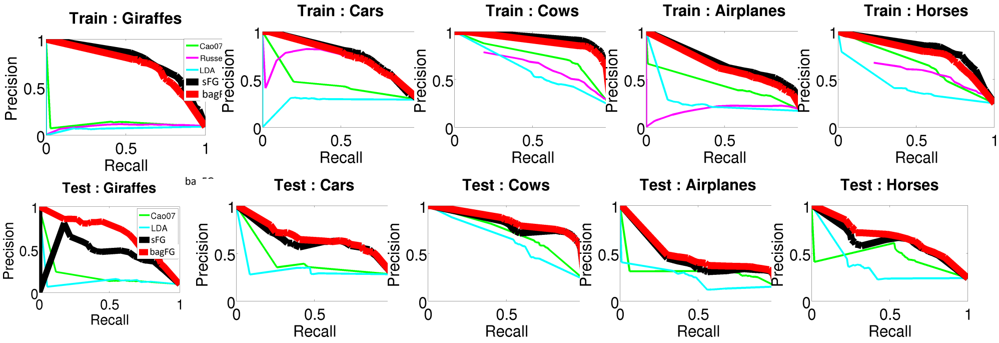

Fig. 5. *Precision-Recall curves for training and testing for 5 out of the 7 datasets.* Our models, sFG and bagFG, outperform all baseline methods.

ETH: 4) 48 images of Bottles, 5) 29 images of Swans, 6) 85 images of Giraffes; WeizmannHorses:7) 80 images In the cases where the whole dataset is not included, images were picked at random.

In each dataset we randomly picked $2/3$ of images for training and $1/3$ for testing. In the datasets where ground truth segmentation is not provided we labeled it by hand by marking superpixels. We score the performance of our model using pixel precision and recall. We do not use segmentation accuracy since many times the object of interest captures a small part in the image and thus an algorithm with very low precision and high recall can get very high scores for segmentation accuracy by getting the background correctly.

We compare against 3 baseline models: 1) Standard LDA model. We used code provide in Topic Modelling Toolbox([29]). 2) Cao et al 07 ([1]) (SpatialLTM model) . In SpatialLTM words belonging to the same superpixel are assigned to the same topic. (see also sections 2). 3) Russel et al 06 ([6]). Each segment is treated as a document and segment based uniformity of words is exploited (see also section 2). We use the code provided online.

For each baseline method, we use the same features and word vocabularies as our model for a fair comparison. Since our baseline models do not discriminate between the foreground and background topics, for each topic we compute the average precision and choose the one with the highest value as the foreground topic. That is we compare against the best scoring topic found by each baseline model. For LDA and Cao et al 07 it is obvious how to get a pixel probability map from the multinomial distributions learned (see also eq. 6), which we threshold to compute our PR curves. For Russel et al 06 we sum the KL divergence scores of all segments to get a pixel score map which we threshold to obtain similar curves. The model Russel et al 07 ([6]) aims at organizing the segments of the training dataset into topics, and does not have a test component, so this method is not used at test time.

We tested both versions of our model: $\mathbf{sFGmodel}$ and $\mathbf{bagFGmodel}$. By using the $\mathbf{bagFGmodel}$ we show how figure-ground information can improve learning of even a simple representation. We believe it provides a fairer comparison with our baselines since same model representation is used to segment a new image.

The results show that using figure-ground information substantially improves the performance of even a simple bag of words represenation. We notice that in some object categories such as Giraffes or Airplanes, the best topic chosen by baseline methods, did not find similar shape across images. In these categories, the bag of feature representations is not strong enough to lead to clustering of the foreground features. In easier datasets like Cows and Horses, we see the baseline topic models to have reasonable performance. The shape and figure-ground aware model outperforms the baseline methods in all datasets.

Training	Giraffes	Cars	Cows	Airplanes	Horses	Bottles	Swans
Cao et al 07	0.124	0.460	0.738	0.436	0.651	0.274	**0.488**
Russell et al 06	0.100	0.672	0.479	0.181	0.404	0.323	0.287
LDA	0.157	0.358	0.595	0.268	0.428	0.297	0.420
sFG	**0.774**	**0.757**	**0.925**	**0.668**	**0.809**	**0.692**	0.487
bagFG	0.729	0.744	0.893	0.632	0.764	0.617	0.456
Testing							
Cao et al 07	0.208	0.423	0.706	0.315	0.448	0.244	0.593
LDA	0.144	0.331	0.627	0.241	0.368	0.229	0.538
sFG	**0.524**	0.638	0.812	0.492	0.624	0.236	0.693
bagFG	0.508	**0.702**	**0.879**	**0.544**	**0.710**	**0.239**	**0.706**

Fig. 6. *Average Precision at train and test time.* The results show that image figure-ground information is useful during training to learn the model, but at test time the representation learned is enough, using saliency in the new image does not offer more in most of the cases.

7 Conclusion

We presented a shape and figure-ground aware model for weakly-supervised detection and segmentation of objects and their backgrounds. We show that by exploiting figure-ground information in images, we learn to segment the foreground object in challenging datasets. Our model uses a prior depending on image figure-ground cues and optimizes a discriminative cost function, which suits well our task of weakly-supervised image segmentation. We use a flexible representation of figure-ground, where figure-ground cues are influenced by feature re-occurrence in the image collection. Our model can tolerate multiple foreground objects in images and still be guided to the correct common figure, it does not make unnatural assumptions and is suitable for a wide variety of datasets. We will submit code for learning and inference for our model.

References

1. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: ICCV. (2007)
2. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: ICCV. (2005) 370–377
3. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: ICCV, Washington, DC, USA (2005)
4. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 2003

5. Hofmann, T.: Probabilistic latent semantic analysis. In: In Proc. of Uncertainty in Artificial Intelligence, UAI?99. (1999) 289–296
6. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR. (2006)
7. Abhinav Gupta, Jianbo Shi, L.D.: A 'shape aware' model for semi-supervised learning of objects and its context. In: NIPS. (2008)
8. Nguyen, M., Torresani, L., de la Torre, F., Rother, C.: (Weakly supervised discriminative localization and classification: A joint learning process)
9. An Exemplar Model for Learning Object Classes. (In: CVPR07)
10. Carolina Galleguillos, Boris Barenko, A.R.S.B.: Weakly supervised object localization with stable segmentations. In: ECCV. (2008)
11. Viola, P., Platt, J., Zhang, C.: Multiple instance boosting for object detection. In: NIPS. (2006)
12. Dollár, P., Babenko, B., Belongie, S., Perona, P., Tu, Z.: Multiple component learning for object detection. In: ECCV '08. (2008)
13. Armand Joulin, F.B., Ponce, J.: Discriminative clustering for image co-segmentation. (In: CVPR10)
14. Zhu, L.L., Lin, C., Huang, H., Chen, Y., Yuille, A.L.: Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In: ECCV 08. (2008)
15. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In: CVPR '06. (2006)
16. J. Winn, N.: Locus: Learning object classes with unsupervised segmentation. (In: ICCV05)
17. Cour, T., Benezit, F., Shi, J.: Spectral segmentation with multiscale graph decomposition. In: CVPR '05, Washington, DC, USA, IEEE Computer Society (2005) 1124–1131
18. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **V45** (2001) 83–105
19. Lowe, D.: Distinctive image features from scale-invariant key-points. Intl. Journal of Computer Vision **60** (2004) 91–110
20. Ren, X., Fowlkes, C.C., Malik, J.: Figure/ground assignment in natural images. In: ECCV. Volume 2. (2006) 614–627
21. Hoiem, D., Stein, A.N., Efros, A.A., Hebert, M.: Recovering occlusion boundaries from a single image. ICCV **0** (2007) 1–8
22. Stas Goferman, Ayellet Tal, L.Z.M.: Puzzle-like collage. In: Computer Graphics Forum (EUROGRAPHICS). (2010)
23. Zhu, Q., Song, G., Shi, J.: Untangling cycles for contour grouping. In: ICCV '07. (2007)
24. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. PAMI **26** (2004) 530–549
25. Itti, L.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40** (2000) 1489–1506
26. Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): Meaningful attention using stochastic image modeling. In: PAMI. (2007)
27. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. J. Vis. **8** (2008) 1–20
28. Seo, H.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. J. Vis. **9** (2009) 1–27
29. Griffiths, T.L., S.M..T.: Finding scientific topics. In: National Academy of sciences, IEEE Computer Society (2007)