Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation

Jian Yao
TTI Chicago
yaojian@ttic.edu

Sanja Fidler
University of Toronto
fidler@cs.toronto.edu

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

Abstract

In this paper we propose an approach to holistic scene understanding that reasons jointly about regions, location, class and spatial extent of objects, presence of a class in the image, as well as the scene type. Learning and inference in our model are efficient as we reason at the segment level, and introduce auxiliary variables that allow us to decompose the inherent high-order potentials into pairwise potentials between a few variables with small number of states (at most the number of classes). Inference is done via a convergent message-passing algorithm, which, unlike graph-cuts inference, has no submodularity restrictions and does not require potential specific moves. We believe this is very important, as it allows us to encode our ideas and prior knowledge about the problem without the need to change the inference engine every time we introduce a new potential. Our approach outperforms the state-of-the-art on the MSRC-21 benchmark, while being much faster. Importantly, our holistic model is able to improve performance in all tasks.

1. Introduction

While there has been significant progress in solving tasks such as image labeling [14], object detection [5] and scene classification [26], existing approaches could benefit from solving these problems jointly [9]. For example, segmentation should be easier if we know where the object of interest is. Similarly, if we know the type of the scene, we can narrow down the classes we are expected to see, e.g., if we are looking at the sea, we are more likely to see a boat than a cow. Conversely, if we know which semantic regions (e.g., sky, road) and which objects are present in the scene, we can more accurately infer the scene type. Holistic scene understanding aims at recovering multiple related aspects of a scene so as to provide a deeper understanding of the scene as a whole.

Most existing approaches to image labeling formulate the problem as inference in a conditional random field (CRF). Nodes in the graph represent the semantic label associated with a pixel or super-pixel, and potentials are created to define the energy of the system. Image evidence is incorporated via local potentials, while smoothness is defined via pairwise potentials between (neighboring) nodes in the graph. Knowledge about other tasks is typically incorporated as image evidence in the form of unitary potentials on the (super-) pixels, encouraging the segmentation to agree with the other task (e.g., object detection) [2]. While effective, this paradigm is suboptimal, as errors in the auxiliary tasks will be propagated to the segmentation. Some of the most recent work has taken a more holistic approach to the problem. However, they either tackle only a subset of tasks [24, 6, 25, 15, 27], treat different components as black boxes [9], and/or rely on complex inference algorithms which are difficult to extend to incorporate other types of potentials/tasks [15].

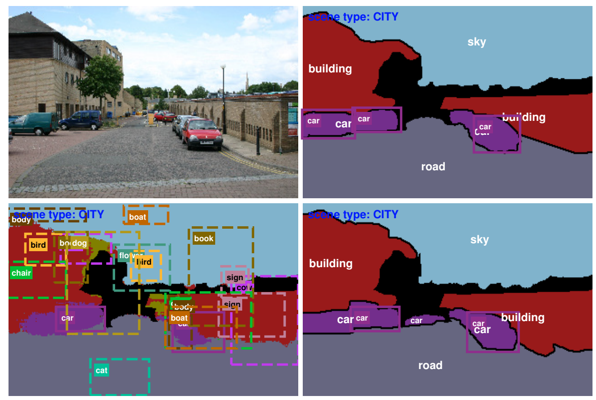


Figure 1. **Top left**: original image, **top right**: groundtruth, **bottom left**: output of individual tasks, **bottom right**: our approach.

In this paper, we propose an approach to holistic scene understanding that simultaneously reasons about regions, location, class and spatial extent of objects, as well as the type of scene (see Fig. 1 for an illustration). We frame the holistic problem as a structure prediction problem in a graphical model defined over hierarchies of regions of different sizes, as well as auxiliary variables encoding the scene type, the presence of a given class in the scene, and the correctness of the bounding boxes output by an object detector. For objects with well-defined shape (e.g., cow, car), we additionally incorporate a shape prior that takes the form of a soft mask learned from training examples.

Unlike existing approaches that reason at the (super-) pixel level, we employ [20] to obtain (typically large) regions which respect boundaries well. This enables us to represent the problem using only a small number of variables. Learning and inference are efficient in our model as the auxiliary variables we utilize allow us to decompose the inherent high-order potentials into pairwise potentials between a few variables with small number of states (at most the number of classes). We take advantage of the structure prediction framework of [8] to learn the importance of each of these potentials. Joint inference is performed using a message-passing scheme which is guaranteed to converge [21]. Unlike existing approaches which employ graph-cuts inference [14, 15], we have no submodularity restriction, and we do not require to construct potential-specific moves. This is adventageous as it allows us to develop our holistic approach relating all tasks with very different types of potentials. Importantly, it enables us to encode new ideas and prior knowledge about the problem without the need to change the inference engine every time we introduce a new potential.

We demonstrate the effectiveness of our approach on the MSRC-21 benchmark and show that our joint model improves segmentation, object detection as well as scene classification accuracy. Our learning and inference algorithms are very efficient; our approach achieves state-of-the-art performance after only 60 seconds of training and an average of 0.02 seconds per image for inference. To guarantee reproducibility of our results, the code as well as annotations and potentials are available online: `http://ttic.uchicago.edu/~yaojian/HolisticSceneUnderstanding.html`.

2. Related Work

In this section we briefly review work on holistic scene understanding. Most existing approaches that combine object detection and segmentation incorporate the output of an object detector to the segmentation as image evidence [2]. In [18, 16], the segmentation is defined to refine the region within ground-truth bounding boxes. Unfortunately, these approaches rely on the correctness of the detector and/or are limited to classes with well defined shape. Conversely, Gu et al [7] uses regions to perform object detection, relying on a single image over-segmentation.

Combining contextual scene information, object detection and segmentation has also been tackled in the past. Torralba et al [24] incorporates contextual information into a CRF, boosting object detection. Sudderth et al [23] relate scenes object and parts, but not segmentation, in a generative model. In [13], context is modeled via spatial interactions between objects using a hierarchical model. In [2], a global image node was introduced. However, their potentials do not decompose, modeling the full power set. Moreover, they bias all pixels towards a particular set of labels. This is suboptimal as we would like to encourage only parts of the image to have certain labels. In [14], co-occurrence potentials were developed to enforce consistency among region labels. In our approach, this is handled by augmenting the graph with two additional layers: a class layer that enforces consistent region labeling and bounding box activations, and a scene layer that ensures that classes consistent with the scene type are active.

In [9], state-of-the-art approaches for each task are used as black boxes, feeding their output as input to the task of interest. While effective, this is suboptimal, as one cannot correct for mistakes. A more holistic approach to combine detection and segmentation was proposed in [6, 25], defining the joint inference within a CRF framework. However, their joint inference does not improve performance significantly and/or rely on complex merging/splitting moves. In [1], a part-based poselet detector was used to perform segmentation by non-rigidly aligning the detections to potential object contours. Our approach also employs probability of boundary contours [20], but only to define the set of segments which compose the first two levels of the hierarchy. Moreover, we model interactions with a CRF.

Probably the most similar approach to ours is [15], which tackles the problem of joint image segmentation and object detection within a CRF. We differ from their approach in several important points: our approach is more holistic as we reason jointly about regions, objects (their location, class and spatial extent), which classes are present in the scene, as well as the scene type. We explicitly show that such an approach not only boosts segmentation but also object detection and scene classification. Moreover, they do not use a shape prior when segmenting an object within the bounding box which can lead to incoherent labelings. Importantly, unlike [15], which relies on graph-cut inference, we do not have submodularity restrictions, and no problem dependent moves are necessary, facilitating the incorporation of very different types of potentials within a unified graphical model. Furthermore, instead of reasoning at the pixel level, we employ the recently developed region segmentation [20] to obtain regions which respect boundaries well. This enables us to do estimation only over a small number of variables making inference much more efficient.

3. Holistic Scene Understanding

In this section we describe our approach to holistic scene understanding. We formulate the problem as the one of inference in a CRF. The random field contains variables representing the class labels of image segments at two levels in a segmentation hierarchy (segments and larger super-segments) as well as binary variables indicating the correctness of candidate object detections. In addition, a multi-

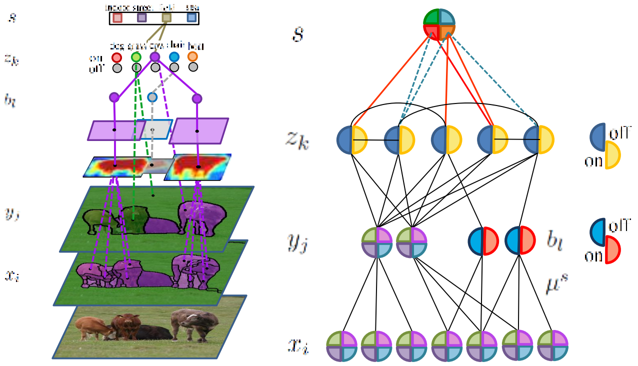


Figure 2. Overview of the model

labeled variable represents the scene type and binary variables encode the presence/absence of a class in the scene. Fig. 2 gives an overview of our model.

The segments and super-segments reason about the semantic class labels to be assigned to each pixel in the image. We employ super-segments to create longer range dependencies as well as for computational reasons, i.e., as they are fewer in number we can connect them more densely to other parts of the model. The binary variables corresponding to each candidate bounding box generated by an object detector allow the model to accept or reject these detections. We associate a shape prior with these nodes to encourage segments to take the class label of the detector. The binary class variables reason about which classes are present in an image. This allows for a natural way to model class co-occurrences, scene-class affinities as well as to ensure that segments and class nodes agree. The scene node encourages certain classes to be present/absent in certain scene types.

More formally, let $x_i \in \{1, \cdots, C\}$ be a random variable representing the class label of the i-th segment in the lower level of the hierarchy, while $y_j \in \{1, \cdots, C\}$ is a random variable associated with the class label of the j-th segment of the second level of the hierarchy. Following recent approaches [14, 17], we represent the detections with a set of candidate bounding boxes. Let $b_l \in \{0, 1\}$ be a binary random variable associated with a candidate detection, taking value 0 when the detection is a false detection. We use the detector of [5] to generate candidate detections, which provides us with an object class (c_l), a score (r_l), the location and aspect ratio of the bounding box, as well as the root mixture component ID that has generated the detection (m_l). The latter gives us information about the expected shape of the object. Let β_l be an observed random variable that encompasses all the information returned by the detector. Let $z_k \in \{0, 1\}$ be a random variable which takes value 1 if class k is present in the image, and let $\mathbf{s} \in \{1, \dots, C_l\}$ be a random variable representing the scene type.

We define our *holistic conditional random field* as

$$p(\mathbf{a}) = p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s}) = \frac{1}{Z} \prod_{type} \prod_{\alpha} \psi_{\alpha}^{type}(\mathbf{a}_\alpha) \quad (1)$$

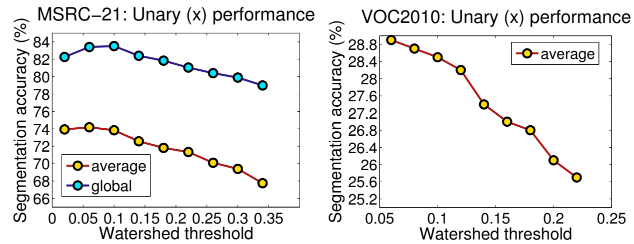


Figure 4. Accuracy vs watershed thresh. for unary region potential.

where $\mathbf{a} = (\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s})$ represents the set of all segmentation random variables, $\mathbf{x}$ and $\mathbf{y}$, the set of C binary random variables $\mathbf{z}$ representing the presence of the different classes in the scene, the set of all candidate detections $\mathbf{b}$, and ψ_{α}^{type} encodes potential functions over sets of variables. Note that the variables in a clique α can be of the same task (e.g., two segments) or different tasks (e.g., detection and segmentation). Our joint inference is then performed by computing the MAP estimate of the random field defined in Eq. 1.

In the following, we describe the different potentials that define our model. For clarity, we describe the potentials in the log domain, i.e., $w_{type}\phi_{\alpha}^{type} = \log(\psi_\alpha)$. We employ a different weight for each type of potential, and share the weights across cliques. We learn the weights from training data using the structure prediction framework of [8] by defining appropriate loss functions for the holistic task.

3.1. Segmentation Potentials

Unary potential: We compute the unary potential for each region at segment and super-segment level by averaging the TextonBoost [14] pixel potentials inside each region.

Segment-SuperSegment compatibility: We use the P^n potentials of [11], which can be written as

$$\phi_{i,j}(x_i, y_j) = \begin{cases} -\gamma & \text{if } x_i \neq y_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that since we learn the weight associated with this potential, we are implicitly learning γ.

3.2. Object Reasoning Potentials

Object Detection Probability: We define a unitary potential on b_l which is a function of the detector's score r_l as

$$\phi_{cls}^{BBox}(b_l, \beta_l) = \begin{cases} \sigma(r_l - \lambda_{cls}) & \text{if } b_l = 1 \wedge c_l = cls \\ 0 & \text{otherwise.} \end{cases}$$

Here λ_{cls} is the detector's threshold, c_l is the detector's class, and $\sigma(x) = 1/(1 + \exp(-1.5\,x))$ is a logistic function. We employ a different weight for each class in order to perform "context re-scoring" when doing inference.

Shape prior: The mixture components of a part-based model typically reflect the pose and shape of the object. We exploit this by defining new potentials which utilize a shape

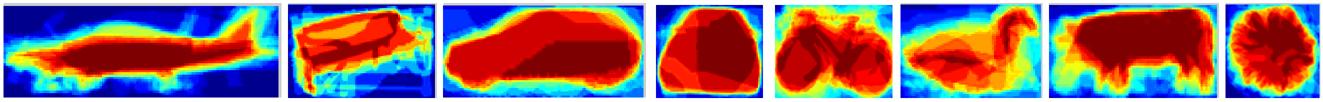


Figure 3. Example of shape prior masks learned on the MSRC dataset [22]. Note that each component of a class detector has its own mask.

prior for each component. Fig. 3 depicts examples of shape masks. We incorporate this information in our model by encouraging x_i that lie inside the detected bounding boxes to take the class of the detector, with strength proportional to the shape mask overlaid over x_i. This encourages the labeling inside the box to be consistent with the average object shape. We learn a different weight per-class, as the shape prior is more reliable for some classes than others. We thus define

$$\phi^{sh}_{cls}(x_i, b_l, \beta_l) = \begin{cases} \mu(x_i, \beta_l) & \text{if } x_i = c_l = cls \wedge b_l = 1 \\ 0 & \text{otherwise.} \end{cases}$$

where $\mu(x_i, \beta_l) = \frac{1}{|A_i|} \sum_{p \in A_i} \mu(p, m_l)$, A_i is the set of pixels in the i-th segment, $|A_i|$ is the cardinality of this set, and $\mu(p, m_l)$ is the value of the mean mask for component m_l at pixel p. Note that this feature only affects the segments inside the bounding box as μ is zero outside.

Class-detection compatibility: This term allows the bounding box to be on only when the class label of that object is also detected as present in the scene. Thus

$$\phi^{BClass}_{l,k}(\beta_l, b_l, z_k) = \begin{cases} -\alpha & \text{if } z_k = 0 \wedge c_l = k \wedge b_l = 1 \\ 0 & \text{otherwise.} \end{cases}$$

where α is a very large number estimated during learning.

3.3. Class Presence Potentials

Class co-occurrences: As we do not have enough training data to estimate the co-occurrence of all pairs of labels, we use the Chow-Liu algorithm [3] to learn a tree-structure linking the z_k. This algorithm is guaranteed to yield the best tree which explains the data statistics. Given the tree-structure, we compute pairwise potentials between nodes in the tree by computing their frequencies empirically.

Class-Segment compatibility: This potential ensures that the classes that are inferred to be present in the scene are compatible with the classes that are chosen at the segment level. Thus

$$\phi_{j,k}(y_j, z_k) = \begin{cases} -\eta & \text{if } y_j = k \wedge z_k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

with η an arbitrarily large number, which is also learned.

3.4. Scene Potentials

Scene probability: In order to incorporate global information about the scene without making hard decisions, we define the unary potential over the scene node as

$$\phi^{Scene}(s = u) = \sigma(t_u)$$

where t_u denotes the classifier score for scene class u and σ is again the logistic function.

Scene-class compatibility: We define a pairwise compatibility potential between the scene and the class labels as

$$\phi^{SC}(s, z_k) = \begin{cases} f_{s,z_k} & \text{if } z_k = 1 \wedge f_{s,z_k} > 0 \\ -\tau & \text{if } z_k = 1 \wedge f_{s,z_k} = 0 \\ 0 & \text{otherwise.} \end{cases}$$

where f_{s,z_k} represents the probability of occurrence of class z_k for scene type s, which is estimated empirically from the training data. This potential "boosts" classes that frequently co-occur with a scene type, and "suppresses" the classes that never co-occur with it, e.g., given a *water* scene we will positively boost *boat* and *water* but suppress *car*.

3.5. Learning with a Holistic Loss

Learning approaches for structured problems require the specification of a task loss, which scores a hypothesis with respect to the ground truth. To deal with our holistic setting, we employ a holistic loss which takes into account all tasks. We define it to be a weighted sum of losses, each one designed for a particular task, e.g., detection, segmentation. In order to do efficient learning, it is important that the losses decompose as a sum of functions on small subsets of variables. Here, we define loss functions which decompose into unitary terms. In particular, we define the segmentation loss at each level of the hierarchy to be the percentage of wrongly predicted pixels. This decomposes as sums of unitary terms (one for each segment). We utilize a 0-1 loss for the variables encoding the classes that are present in the scene, which also decomposes as the sum of unitary 0-1 losses on each z_k. We define a 0-1 loss over the scene type, and a PASCAL loss over the detections which decomposes as the sum of losses for each detection

$$\Delta_B(b_l, \hat{b}_l) = \begin{cases} 1 - \frac{intersection}{union} & \text{if } b_l = 1 \\ \frac{intersection}{union} & \text{otherwise} \end{cases}$$

Note that these are unitary potentials on the b_l.

We take advantage of the family of structure prediction problems developed in [8] and employ a CRF with ℓ_2 regularization (i.e., $p = 2$, $\epsilon = 1$). We use the message-passing algorithm of [21] with $\epsilon = 1$ for inference.

4. Experimental Evaluation

We test our approach on the tasks of semantic segmentation on the MSRC-21 dataset [22] as well as object segmentation on the PASCAL VOC2010 [4] benchmark. We employ [20] to obtain regions which respect boundaries well.

	building	grass	tree	cow	sheep	sky	aeropl.	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**	**global**
	cleanMSRC dataset																						
HCRF+Coocc. [15]	**73**	**93**	**82**	81	**91**	**98**	81	83	**88**	74	85	**97**	79	**38**	96	61	**90**	69	48	**67**	18	75.8	**85.0**
x unary	69	92	**82**	81	87	**98**	84	83	83	77	91	94	80	23	97	57	86	79	54	61	0	74.2	83.8
x, y, z	68	92	81	81	87	**98**	84	**84**	79	78	93	96	79	30	**98**	63	86	76	54	60	0	74.7	83.9
$x, y, z, s, \tau = 1$	67	92	81	81	89	97	84	84	83	78	93	96	79	30	98	66	88	77	55	60	0	75.1	84.0
x, y, z, b, 2 classes	68	92	80	81	86	98	88	84	79	75	93	96	78	30	98	65	87	76	74	63	23	75.9	84.0
x, y, z, b, 10 classes	68	92	80	82	87	98	86	83	79	79	93	96	92	31	98	70	87	78	55	64	23	76.7	84.3
x, y, z, b, 15 classes	67	92	80	**82**	89	97	**86**	83	86	**79**	**94**	96	85	35	**98**	**70**	86	78	**55**	62	**23**	**77.4**	84.4
full model, $\tau = 1$	68	92	80	82	90	97	86	83	87	79	94	96	82	36	98	68	86	82	55	62	18	77.1	84.3
	origMSRC dataset																						
Shotton et al. [22]	49	88	79	**97**	**97**	78	82	54	87	74	72	74	36	24	93	51	78	75	35	66	18	67	72
Jiang and Tu [10]	53	97	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	**66**	64	13	68	78
Harmony potential [2]	60	78	77	91	68	88	87	76	73	77	93	97	73	**57**	95	**81**	76	81	46	56	**46**	75	77
HCRF+Coocc. [15]	74	98	90	75	86	**99**	81	84	**90**	83	91	**98**	75	49	95	63	**91**	71	49	**72**	18	77.8	**86.5**
Dense CRF [12]	75	**99**	**91**	84	82	95	82	71	89	**90**	**94**	95	**77**	48	96	61	90	78	48	80	22	78.3	86.0
x unary	72	97	90	78	85	96	84	84	83	81	91	97	69	49	95	59	90	81	53	65	0	76.1	85.2
x, y, z	72	98	**91**	77	82	93	86	**86**	82	82	93	97	71	50	96	59	88	78	51	67	0	76.2	85.1
x, y, z, b, 15 classes	70	98	88	78	86	92	**88**	84	**90**	84	**94**	**98**	75	51	**97**	72	87	**83**	53	67	7	78.2	85.5
full model, $\tau = 20$	71	98	90	79	86	93	**88**	**86**	**90**	84	**94**	**98**	76	53	**97**	71	89	**83**	55	68	17	**79.3**	86.2

Table 1. MSRC-21 segmentation results

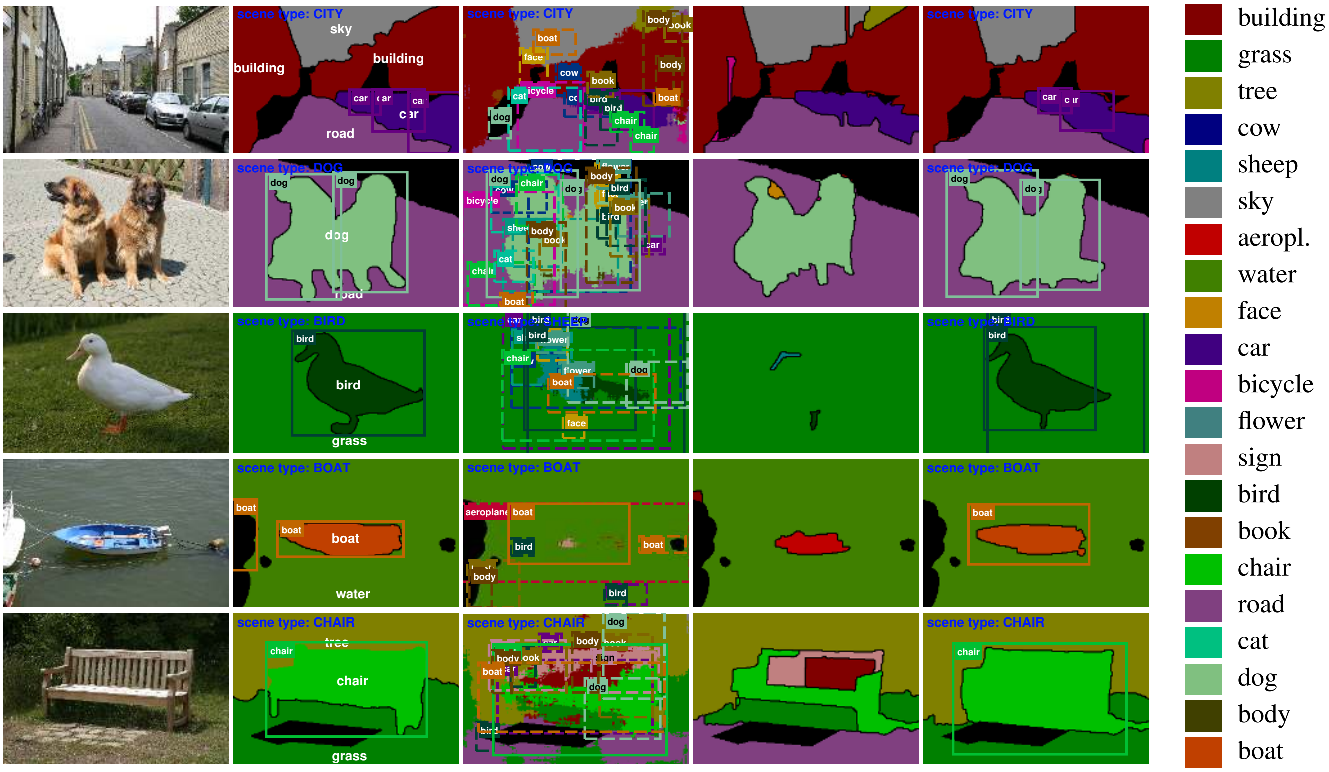
	cow	sheep	aeropl.	face	car	bicycle	flower	sign	bird	book	chair	cat	dog	body	boat	**avg.**
	Recall at equal FPPI															
FPPI rate	0.03	0.02	0.00	0.01	0.05	0.03	0.04	0.02	0.02	0.01	0.00	0.02	0.04	0.04	0.02	0.02
LSVM [5]	84.6	73.9	84.6	**59.4**	51.2	63.6	16.9	40.0	18.9	23.7	50.0	20.0	20.0	43.2	**18.8**	44.6
context LSVM [5]	76.9	**84.6**	36.6	40.0	16.2	**68.8**	**30.0**	20.0	**34.1**	18.8	45.5	39.1	21.1	**59.4**	13.6	40.3
x, y, z, b	**88.5**	78.3	**100.0**	43.8	**53.7**	63.6	20.3	**53.3**	16.2	42.1	**62.5**	40.0	**26.7**	36.4	6.2	48.8
full model, tau=20	**88.5**	82.6	**100.0**	43.8	**53.7**	63.6	20.3	**53.3**	16.2	**44.7**	**62.5**	**50.0**	**26.7**	38.6	15.6	**50.7**
	Average Precision															
LSVM [5]	**78.6**	76.5	96.2	**56.4**	**54.1**	**61.7**	19.9	45.0	18.5	30.0	59.2	31.4	28.0	45.5	**22.1**	48.2
context LSVM [5]	77.5	**93.1**	52.3	41.0	16.1	58.1	**30.2**	32.0	**43.4**	24.5	61.4	45.7	**30.8**	**59.4**	19.4	45.7
x, y, z, b	76.1	72.7	**100.0**	45.5	53.1	60.9	22.9	**48.5**	18.2	42.9	**63.6**	45.3	27.3	34.3	9.1	48.0
full model, tau=20	78.1	81.8	**100.0**	45.5	53.1	60.9	22.9	48.5	18.2	**44.4**	**63.6**	45.6	27.3	34.8	14.8	**49.3**

Table 2. MSRC-21 object detection results (our models were trained no origMSRC)

The number and size of the output regions depends on the watershed threshold. Fig. 4 shows the segmentation accuracy based only on the region unary potentials at different thresholds, $\mathcal{K}_0$. In our experiments we set the threshold to be 0.08 and 0.16 for the two layers in the hierarchy for MSRC-21 and 0.14 and 0.18 for PASCAL. This gives us on average 65 and 19 regions per image at the segment and super-segment level for MSRC-21 and 49 and 32 for PASCAL. To create the unitary potentials for the scenes, we use a standard bag-of-words spatial pyramid with $1, 2$ and 4 levels over a 1024 sparse coding dictionary on SIFT features, colorSIFT, RGB histograms and color moment invariants, and train a linear one-vs-all SVM classifier.

We use the detector of [5] to generate candidate detections. We trained a few models with different number of parts and mixture components and selected the best performing one for each class. For each detector we also lowered the threshold to produce over-detections. We follow Felzenswalb et al.'s entry in PASCAL'09 to compute the soft shape masks. For each class we ran the detector on the training images and chose those that overlaped with groundtruh more than 0.5 in the intersection over union measure. For each positive detection we also recorded the winning component. We compute the mask for each component by simply averaging the groundtruth class regions inside the assigned groundtruth boxes. Prior to averaging, all bounding boxes were warped to the same size, i.e., the size of the root filter of the component. To get the shape mask μ for each detection we warped the average mask of the detected component to the predicted bounding box.

We first evaluate our approach on MSRC-21 [22], which will be referred to as *origMSRC*, as well as on the more accurately annotated version [19], referred to as *cleanMSRC*. The dataset contains classes such as *sky, water*, as well as more shape-defined classes such as *cow, car*. We manually annotated bounding boxes for the latter classes, with a total of 15 classes and 934 annotations. We also annotated 21 scenes, taking the label of the salient object in the image, if there is one, or a more general label such as "city" or "water" otherwise. Note that other annotations are possible, as



(a) image (b) groundtruth (c) indep. tasks (d) x, y, z (e) full model

Figure 5. **cleanMRSC-21**: Results of joint image segmentation and object class detection.

	classifier	full, $\tau = 1$	full, $\tau = 3$	full, $\tau = 20$
accuracy	79.5	79.3	80.4	**80.6**

Table 3. origMSRC scene classification

there is no clear unique scene labels for this dataset.

We follow the standard error measure of average per-class accuracy as well as average per-pixel accuracy, denoted as global [15]. We used the standard train/test split [22] to train the full model, the pixel unary potential, object detector and scene classifier. Different pixel classifiers were trained for clean and origMSRC. Table 1 reports the accuracy on cleanMSRC, where we evaluated the segmentation performance of our approach when incorporating different components in our model. We refer to these with the name of the variables we used in each experiment. We compare to ALE [15] by running their code. Note that our approach significantly outperforms [15] in the average accuracy, while it is a little behind in global. This is expected as we focus on objects, which are smaller than classes such as "sky" or "grass". Importantly, we show that (i) we significantly improve over the pixel unary, (ii) our scene potential outperforms [15]'s coocurrence potential, (iii) performance of the joint model increases as a function of the number of class detectors used. We believe that for segmentation a comparison on cleanMSRC is more important than on the origMSRC, as better labeling makes possible to make finer distinctions between the methods. Table 1 also reports the

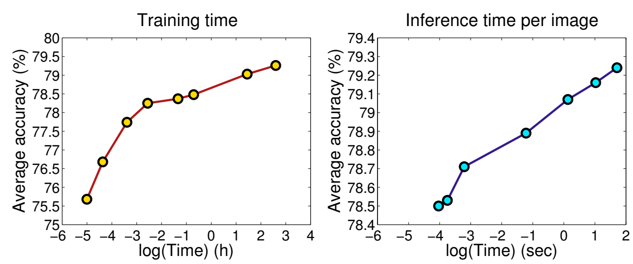


Figure 8. origMRSC: Segmentation accuracy as a function of **(left)** training time, **(right)** inference time.

segmentation accuracy on the origMSRC, along with the comparisons with the existing state-of-the-art. Similarly, ALE was run by us. Our joint model outperforms all models with less components, and achieves the highest average accuracy reported on this dataset to date. Furthermore, we next show that the joint model not only improves segmentation accuracy but also significantly boosts object detection and scene classification.

Fig. 5 shows the results of our method compared to groundtruth and independent inference over each task. Note that our approach (last column) is able to reliably detect the objects of interest from a large pull of candidate detections, and our segmentations accurately follow the boundaries. Failure cases are depicted in Fig. 7. Some of the main sources of error are bad unary potentials (e.g., *boat* unary has 0% accuracy), region under-segmentations (i.e., failures of [20]), or false negative detections.

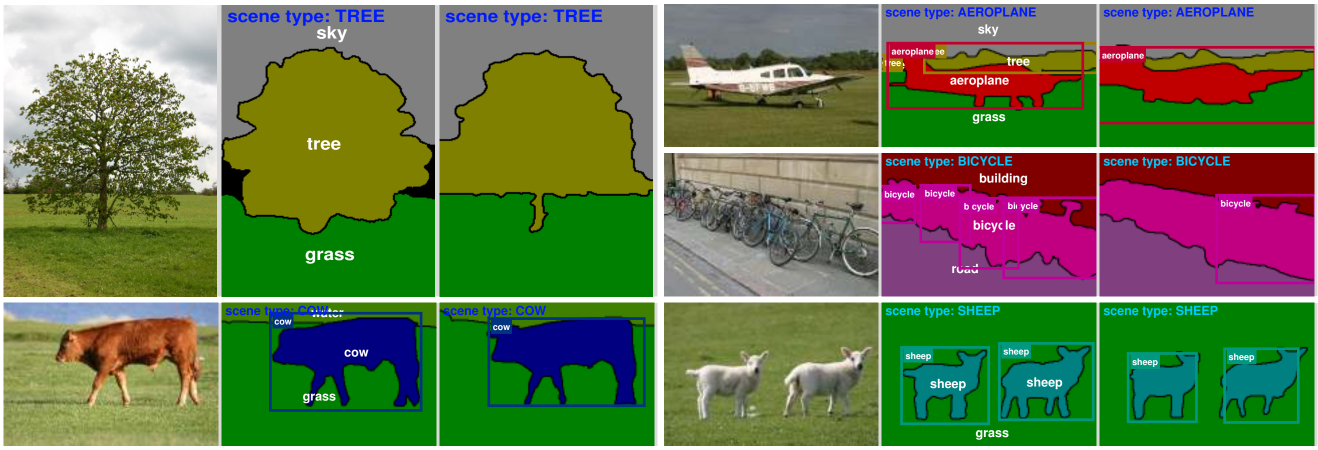


Figure 6. origMRSC: Segmentation examples: (image, groundtruth, our holistic scene model)

Figure 7. origMRSC: Examples of failure modes.

Table 2 shows detection results for our approach as well as the original detector [5] (refered to as LSVM). Since original LSVM assumes single-class evaluation, we also trained and ran its multi-class extension, named *context re-scoring* [5]. The goal of the latter is the same as ours, context based multi-class object detection, and is thus more directly comparable. Note, that due to the lack of training examples in this dataset context re-scoring performs worse than the original detector. Since our approach only marks a detection as "on" or "off" and does not assign a score, we evaluated the performance as **recall** at the false-positive rate of our detector (which is always smaller or equal to that of the original one). We show that our model without scenes outperforms LSVM by 4.2%, which is boosted to 6.1% when using the scene potentials as well. Our result is by 10.4% better than context-LSVM. In Table 2 we also evaluate performance in terms of average precision (AP), which we compute for our approach by taking only the boxes marked as "on" but using the original score for evaluation. Note that this is somewhat artificial as our holistic approach only returns a point in the precision-recall curve. However, this enables us to compare the approaches under the standard PASCAL's performance measure. We improve 1.1% over LSVM and 3.6% over context-LSVM.

Table 3 shows scene classification accuracy as a function of the parameter τ. Note that once more our full model is able to improve performance.

Finally, we analyze the complexity of our model. Fig. 8 shows performance as a function of training and inference time (controlled by the number of iterations in [8] and in the massage-passing algorithm of [21]). All experiments were performed on an Intel i7-2700K 3.4GHz processor. Both training and inference were performed using 4 cores. The times and accuracies are reported for the full, best performing, model ($\tau = 20$) on origMSRC. Note that it takes only 59.6 seconds to train the model to get accuracy 78.3% (state-of-the-art), while it takes 3.5 hours to get the full performance of 79.3%. In inference, it takes only 0.02 seconds per image to reach accuracy of 78.5% (higher than state-of-the-art), and 7.3 seconds to get the full 79.3% accuracy.

We also performed initial experiments on the VOC2010 dataset, which we report in Table 4. Segmentation accuracy was computed using the standard VOC measure [4]. Results are reported on the validation set of comp5. The unary pixel potentials [12] (trained on 40% of the training set) on their own achieve 27.6%. Our x unary gets 27.4%. We trained the (context re-scored) LSVM detector on all VOC2010 excluding val images of comp5. Following Felzenswalb et al.'s PASCAL entry, the detector alone achieves 26% segmentation accuracy (we used the code kindly provided by the authors to compute the segmentation). Our model with no scene type classifier results in 31.2% accuracy, comparing favorably to 30.2% of [12]. In future work, we plan to augment PASCAL with scene classification. Segmentation examples are shown in Fig. 9.

	[5]	x unary	[12]	ours (x, y, z, b)
average	26.0	27.4	30.2	**31.2**

Table 4. VOC2010 segmentation results. All model trained on train and evaluated on val of comp5.

5. Conclusion

We presented a holistic approach to scene understanding which reasons jointly about segmentation, objects, presence

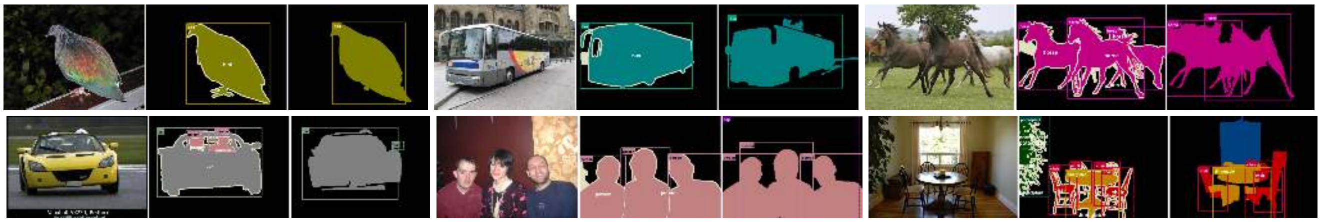


Figure 9. PASCAL VOC2010: Segmentation examples (image, groundtruth, our holistic model)

of a class in the image, as well as scene type. Our approach achieves state-of-the-art performance in the popular MSRC-21 benchmark, while being much faster than existing approaches. More importantly, our holistic model is able to improve performance in all tasks. We plan to extend our model to reason about other tasks such as activities, and to incorporate other sources of information such as location priors and other notions of context.

Acknowledgments. S.F. has been supported by DARPA, contract number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government.

References

[1] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011. 2

[2] G. Cardinal, X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010. 1, 2, 5

[3] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462467, 1968. 4

[4] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 4, 7

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010. 1, 3, 5, 7

[6] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009. 1, 2

[7] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using region. In *CVPR*, 2009. 2

[8] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 2, 3, 4, 7

[9] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008. 1, 2

[10] J. Jiang and Z. Tu. Efficient scale space auto-context for image segmentation and labeling. In *CVPR*, 2009. 5

[11] P. Kohli, M. P. Kumar, and P. H. S. Torr. p^3 and beyond: Solving energies with higher order cliquess. In *CVPR*, 2007. 3

[12] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 5, 7

[13] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005. 2

[14] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 1, 2, 3

[15] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010. 1, 2, 5, 6

[16] D. Larlus and F. Jurie. Combining appearance models and markov random fields for category level object segmentation. In *CVPR*, 2008. 2

[17] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010. 3

[18] V. Lempitsky, P. Kohli, C. Rother, and B. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009. 2

[19] T. Malisiewicz and A. A.Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007. 5

[20] P.Arbelaez, M.Maire, C.Fowlkes, and J.Malik. Contour detection and hierarchical image segmentation. In *PAMI*, 2011. 2, 4, 6

[21] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 2, 4, 7

[22] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 4, 5, 6

[23] E. Sudderth, A. Torralba, W. T. Freeman, and A. Wilsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005. 2

[24] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, pages 1401–1408, 2005. 1, 2

[25] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, volume 4, pages 733–747, 2008. 1, 2

[26] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1

[27] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010. 1